IBM SPSS Modeler Text Analytics 18.5
Guia do Usuário

Nota

Antes de usar estas informações e o produto suportado por elas, leia as informações nos “Avisos” na página 227.

Informações do produto

Esta edição aplica-se à versão do 18.5.0 de IBM® SPSS Modeler Text Analytics e a todos os releases e modificações subseqüentes até que indicado de outra forma em novas edições.

Índice

Prefácio

IBM SPSS Modeler Text Analytics oferece poderosas capacidades de analítica de texto que usam tecnologias linguísticas avançadas e Processamento de Linguagem Natural (NLP) para processar rapidamente uma grande variedade de dados de texto não estruturado e, a partir desse texto, extrair e organizar conceitos chave. Além disso, o IBM SPSS Modeler Text Analytics pode agrupar esses conceitos em categorias.

Cerca de 80% dos dados mantidos dentro de uma organização estão em forma de documentos de texto - por exemplo, relatórios, páginas da web, emails e notas da central de atendimento. O texto é um fator chave na ativação de uma organização tenha um melhor entendimento do comportamento de seus clientes. Um sistema que incorpora o NLP pode extrair conceitos de maneira inteligente, incluindo frases compostas. Além disso, o conhecimento do idioma subjacente permite a classificação de termos em grupos relacionados, como produtos, organizações ou pessoas, usando significado e contexto. Como resultado, é possível determinar rapidamente a relevância das informações para suas necessidades. Essas categorias e conceitos extraídos podem ser combinados com dados estruturados existentes, como demográficos, e aplicados à modelagem no conjunto completo do IBM SPSS Modeler de ferramentas de mineração de dados para gerar decisões melhores e mais focadas.

Os sistemas linguísticos são sensíveis ao conhecimento — quanto mais informações estão contidas em seus dicionários, mais alta será a qualidade dos resultados. IBM SPSS Modeler Text Analytics é entregue com um conjunto de recursos linguísticos, como dicionários para termos e sinônimos, bibliotecas e modelos. Esse produto ainda permite desenvolver e refinar esses recursos linguísticos para seu contexto. O ajuste dos recursos linguísticos geralmente é um processo interativo e é necessário para a categorização e a recuperação de um conceito preciso. Modelos customizados, bibliotecas e dicionários para domínios específicos, como CRM e genoma, também estão incluídos.

Sobre o IBM Business Analytics

O software IBM Business Analytics fornece informações completas, consistentes e exatas nas quais os tomadores de decisão confiam para melhorar o desempenho de negócios. Um portfólio abrangente de inteligência de negócios, analise preditiva, gerenciamento de desempenho financeiro e estratégiae aplicativos analíticos fornecem insight claro, imediato e prático sobre o desempenho atual e a capacidade de prever resultados futuros. Combinado com soluções para segmentos do mercado, práticas comprovadas e serviços profissionais completos, organizações de qualquer tamanho poderão conduzir maior produtividade, automatizar as decisões de modo confiável e entregar melhores resultados.

Como parte deste dossier, o software IBM SPSS Predictive Analytics ajuda as organizações a prever futuros eventos e agir proativamente com esse insight para melhores resultados de negócios. Clientes comerciais, governamentais e acadêmicos do mundo todo confiam na tecnologia IBM SPSS como uma vantagem competitiva para atrair, reter e aumentar clientes, enquanto reduz a fraude e minimiza riscos. Ao incorporar o software IBM SPSS em suas operações diárias, as organizações se tornam empresas preditivas, ou seja, capazes de direcionar e de automatizar as decisões para atender às metas de negócios e obter vantagem competitiva mensuráveis. Para obter mais informações ou entrar em contato com um representante, visite http://www.ibm.com/spss.

Suporte técnico

O suporte técnico está disponível para manutenção dos clientes. Os clientes podem entrar em contato com o Suporte Técnico para obter assistência no uso de produtos IBM Corp. ou para obter ajuda na instalação em um dos ambientes de hardware suportados. Para entrar em contato o Suporte Técnico, consulte o website em IBM Corp. http://www.ibm.com/support. Esteja preparado para se identificar, sua organização e seu contrato de suporte ao solicitar assistência.

Capítulo 1. Sobre o IBM SPSS Modeler Text Analytics

IBM SPSS Modeler Text Analytics oferece poderosas capacidades de analítica de texto que usam tecnologias linguísticas avançadas e Processamento de Linguagem Natural (NLP) para processar rapidamente uma grande variedade de dados de texto não estruturado e, a partir desse texto, extrair e organizar conceitos chave. Além disso, o IBM SPSS Modeler Text Analytics pode agrupar esses conceitos em categorias.

Cerca de 80% dos dados mantidos dentro de uma organização estão em forma de documentos de texto - por exemplo, relatórios, páginas da web, emails e notas da central de atendimento. O texto é um fator chave na ativação de uma organização tenha um melhor entendimento do comportamento de seus clientes. Um sistema que incorpora o NLP pode extrair conceitos de maneira inteligente, incluindo frases compostas. Além disso, o conhecimento do idioma subjacente permite a classificação de termos em grupos relacionados, como produtos, organizações ou pessoas, usando significado e contexto. Como resultado, é possível determinar rapidamente a relevância das informações para suas necessidades. Essas categorias e conceitos extraídos podem ser combinados com dados estruturados existentes, como demográficos, e aplicados à modelagem no conjunto completo do IBM SPSS Modeler de ferramentas de mineração de dados para gerar decisões melhores e mais focadas.

Os sistemas linguísticos são sensíveis ao conhecimento — quanto mais informações estão contidas em seus dicionários, mais alta será a qualidade dos resultados. IBM SPSS Modeler Text Analytics é entregue com um conjunto de recursos linguísticos, como dicionários para termos e sinônimos, bibliotecas e modelos. Esse produto ainda permite desenvolver e refinar esses recursos linguísticos para seu contexto. O ajuste dos recursos linguísticos geralmente é um processo interativo e é necessário para a categorização e a recuperação de um conceito preciso. Modelos customizados, bibliotecas e dicionários para domínios específicos, como CRM e genoma, também estão incluídos.

Implementação. É possível implementar fluxos de mineração de texto usando IBM SPSS Modeler Solution Publisher para escoragem em tempo real de dados não estruturados. A capacidade de implementar esses fluxos assegura implementações de mineração de texto de loop fechado bem-sucedidas. Por exemplo, agora sua organização pode analisar notas da área de rascunho de responsáveis pela chamada de entrada ou saída ao aplicar seus modelos preditivos para aumentar a precisão da sua mensagem de marketing em tempo real.

Para executar IBM SPSS Modeler Text Analytics com IBM SPSS Modeler Solution Publisher, inclua o diretório `<install_directory>/ext/bin/spss.TMWBServer` na variável de ambiente `$LD_LIBRARY_PATH` .

Nota: O adaptador japonês para IBM SPSS Modeler Text Analytics foi descontinuado a partir da versão 18.1.

Fazendo upgrade do IBM SPSS Modeler Text Analytics

Antes de instalar IBM SPSS Modeler Text Analytics, você deve salvar e exportar quaisquer TAPs, templates e bibliotecas a partir da sua versão atual que você deseja usar na nova versão. É recomendado salvar esses arquivos em um diretório que não seja excluído ou sobrescrito quando você instalar a versão mais recente.

Depois de instalar a versão mais recente de IBM SPSS Modeler Text Analytics, você pode carregar o arquivo TAP salvo, adicionar quaisquer bibliotecas salvas ou importar e carregar quaisquer modelos salvos para usá-los na versão mais recente.

Importante: Se você desinstalar sua versão atual sem salvar e exportar os arquivos necessários primeiro, qualquer trabalho de TAP, modelo e biblioteca pública realizado na versão anterior será perdido e incapaz de ser usado na versão mais recente de IBM SPSS Modeler Text Analytics.

Sobre a mineração de texto

Hoje uma quantidade cada vez maior de informações está sendo realizada em formatos não estruturados e semistruturados, como e-mails de clientes, notas de call center, respostas de pesquisa em aberto, feeds de notícias, formulários Web, etc. Essa abundância de informações representa um problema para muitas organizações que se perguntam: "Como podemos coletar, explorar e potencializar essas informações?"

Mineração de texto é o processo de análise de coleções de materiais textuais para capturar os principais conceitos e temas e descobrir relacionamentos ocultos e tendências, sem exigir que você conheça as palavras ou termos exatos que os autores usaram para expressar tais conceitos. Embora elas sejam bastante diferentes, a mineração de texto às vezes é confundida com recuperação de informações. Embora a recuperação precisa e o armazenamento de informações seja um enorme desafio, a extração e o gerenciamento de conteúdo de qualidade, terminologia e relacionamentos contidos nas informações são processos essenciais e críticos.

Mineração de texto e mineração de dados

Para cada artigo de texto, a mineração de texto baseada em linguística retorna um índice de conceitos, bem como informações sobre tais conceitos. Estas informações extraída e estruturadas podem ser combinada com outras origens de dados para abordar questões como:

- Quais conceitos ocorrem juntos?
- Ao que mais eles são vinculados?
- Quais categorias de nível superior podem ser criadas a partir das informações extraídas?
- O que os conceitos ou as categorias preveem?
- Como os conceitos ou as categorias preveem o comportamento?

Combinar a mineração de texto com a mineração de dados oferece um insight maior do que está disponível a partir dos dados estruturados ou desestruturados sozinhos. Esse processo normalmente inclui as etapas a seguir:

1. **Identifique o texto a ser minerado.** Prepare o texto para mineração. Se o texto existir em diversos arquivos, salve os arquivos em um único local. Para bancos de dados, determine o campo contendo o texto.
2. **Minere o texto e extraia os dados estruturados.** Aplique os algoritmos de mineração de texto no texto de origem.
3. **Construir modelos de conceito e categoria.** Identifique os principais conceitos e/ou crie categorias. O número de conceitos retornados dos dados não estruturados geralmente é muito grande. Identifique os melhores conceitos e categorias para escoragem.
4. **Analise os dados estruturados.** Utilize técnicas de mineração de dados tradicionais, tais como armazenamento em cluster, classificação e modelagem preditiva para descobrir relacionamentos entre os conceitos. Mescle os conceitos extraídos com outros dados estruturados para prever o comportamento futuro com base nos conceitos.

Análise e categorização de texto

Análise de texto, um formulário de análise qualitativa, é a extração de informações úteis do texto para que as ideias principais ou conceitos contidos neste texto possam ser agrupados em um número apropriado de categorias. A análise de texto pode ser executada em todos os tipos e comprimentos de texto, embora a abordagem para a análise irá variar um pouco.

Menos registros ou documentos são categorizados mais facilmente já que eles não são tão complexos e, geralmente, contêm menos palavras e respostas ambíguas. Por exemplo, com perguntas de pesquisa de opinião curtas e em aberto, se solicitarmos que as pessoas citem suas três atividades de férias favoritas, podemos esperar ver muitas respostas curtas, como *ir à praia*, *visitar parques nacionais* ou *fazer nada*. Respostas em aberto mais longas, por outro lado, podem ser bastante complexas e muito alongadas, especialmente se os respondentes forem instruídos, motivados e tiverem tempo suficiente

para preencher um questionário. Se solicitarmos às pessoas que nos informem sobre suas convicções políticas em uma pesquisa de opinião ou tivermos um feed de blog sobre política, podemos esperar alguns comentários alongados sobre todos os tipos de questões e posições.

A capacidade de extrair conceitos principais e criar categorias intuitiva a partir dessas origens de texto mais longas em um período muito curto de tempo é uma vantagem principal do uso do IBM SPSS Modeler Text Analytics. Essa vantagem é obtida através da combinação de técnicas linguísticas e estatísticas automatizadas para produzir os resultados mais confiáveis para cada estágio do processo de análise de texto.

Processamento linguístico e NLP

O problema principal com o gerenciamento de todos esses dados de texto não estruturados é que não há regras padrão para gravação de texto para que um computador possa compreendê-los. O texto e, consequentemente, o significado, varia para cada documento e cada parte do texto. A única maneira de recuperar e organizar os dados não estruturados com precisão é analisar o texto e, assim, descobrir o seu significado. Há várias abordagens automatizadas diferentes para a extração de conceitos a partir de informações não estruturadas. Essas abordagens podem ser divididos em dois tipos, linguístico e não linguístico.

Algumas organizações tentaram utilizar soluções não linguísticas automatizadas com base em estatísticas e redes neurais. Usando a tecnologia de computador, essas soluções podem varrer e categorizar conceitos principais mais rapidamente do que os leitores humanos podem. Infelizmente, a precisão de tais soluções é razoavelmente baixa. A maioria dos sistemas baseados em estatísticas simplesmente contam o número de vezes que as palavras ocorrem e calculam suas proximidade estatística com os conceitos relacionados. Eles produzem muitos resultados irrelevantes, ou ruído, e perdem os resultados que eles deveriam ter encontrado, referidos como silêncio.

Para compensar a sua precisão limitada, algumas soluções incorporam regras não linguísticas complexas que o ajudam a distinguir entre resultados relevantes e irrelevantes. Isso é referido como *mineração de texto baseada em regra*.

Mineração de texto baseada em linguística, por outro lado, aplica os princípios de processamento de idioma natural (NLP)– a análise assistida por computador das linguagens humanas - na análise de palavras, frases e sintaxe ou estrutura do texto. Um sistema que incorpora o NLP pode extrair conceitos de maneira inteligente, incluindo frases compostas. Além disso, o conhecimento do texto subjacente permite a classificação de conceitos em grupos relacionados, tais como produtos, organizações ou pessoas, usando o significado e o contexto.

A mineração de texto baseada em linguística localiza significado no texto muito como as pessoas fazem - reconhecendo uma variedade de formatos da palavra com significados semelhantes e analisando a estrutura da sentença para fornecer uma estrutura para o entendimento do texto. Esta abordagem oferece a velocidade e a efetividade em custo dos sistemas baseados em estatísticas, mas oferece um grau muito maior de precisão enquanto requer muito menos intervenção humana.

Para ilustrar a diferença entre abordagens baseadas em estatística e linguística durante o processo de extração, considere como cada uma responderia a uma consulta sobre `reproduction of documents`. Ambas as soluções baseadas em estatística e linguística teriam que expandir a palavra `reproduction` para incluir sinônimos, como `copy` e `duplication`. Caso contrário, informações relevantes serão omitidas. Mas se uma solução baseada em estatística tentar fazer esse tipo de sinônimo -procurando outros termos com o mesmo significado - é provável que inclua o termo `birth` também, gerando uma série de resultados irrelevantes. O entendimento do idioma evita a ambiguidade do texto, tornando a mineração de textos baseada em linguística, por definição, a abordagem mais confiável.

Entender como o processo de extração funciona pode ajudá-lo a tomar decisões sobre quando usar o ajuste fino nos seus recursos linguísticos (bibliotecas, tipos, sinônimos e mais). Etapas no processo de extração incluem:

- Conversão da origem de dados em um formato padrão
- Identificação de termos candidatos

- Identificação de classes de equivalência e integração de sinônimos
- Designação de um tipo
- Indexação e, quando solicitado, correspondência de padrões com um analisador secundário

Etapa 1. Convertendo dados de origem em um formato padrão

Nesta primeira etapa, os dados importados são convertidos para um formato uniforme que pode ser usado para análise adicional. Esta conversão é executada internamente e não muda seus dados originais.

Etapa 2. Identificando os termos candidatos

É importante entender o papel dos recursos linguísticos na identificação de termos candidatos durante a extração linguística. Recursos linguísticos são usados a cada vez que uma extração é executada. Eles existem no formato de modelos, bibliotecas e recursos compilados. As bibliotecas incluem listas de palavras, relacionamentos e outras informações usadas para especificar ou ajustar a extração. Os recursos compilados não podem ser visualizados ou editados. Entretanto, os recursos restantes podem ser editados no Editor de Template ou se você estiver em uma sessão de ambiente de trabalho interativa, no Editor de Recursos.

Os recursos compilados são componentes internos principais do mecanismo de extração dentro do IBM SPSS Modeler Text Analytics . Estes recursos incluem um dicionário geral contendo uma lista de formulários base com um código de parte do discurso (substantivo, verbo, adjetivo e assim por diante).

Além desses recursos compilados, várias bibliotecas são entregues com o produto e podem ser usadas para complementar os tipos e as definições de conceito nos recursos compilados, bem como para oferecer sinônimos. Estas bibliotecas — e quaisquer umas criadas — são compostas de diversos dicionários. Estas incluem dicionários de tipo, dicionários de sinônimos e excluem dicionários.

Assim que os dados foram importados e convertidos, o mecanismo de extração começará a identificação de termos candidatos para extração. Os termos candidatos são palavras ou grupos de palavras que são usados para identificar conceitos no texto. Durante o processamento do texto, palavras únicas (*unitermos*) e palavras compostas (*multitermos*) são identificadas usando extratores de padrão de classe gramatical. Então, as palavras-chave de sentimento do candidato são identificadas usando a análise de link de texto de sentimento.

Nota: Os termos do dicionário geral compilado supramencionado acima representam uma lista de todas as palavras que provavelmente não têm interesse ou são linguisticamente ambíguas como unitermos. Estas palavras são excluídas da extração quando você está identificando os unitermos. Entretanto, elas são reavaliadas quando você está determinando parte do discurso ou verificando palavras compostas candidatas (multitermos).

Etapa 3. Identificação de classes de equivalência e integração de sinônimos

Após os unitermos e multitermos candidatos serem identificados, o software usa um dicionário de normalização para identificar classes de equivalência. Uma classe de equivalência é uma forma base de uma frase ou uma forma única de duas variantes da mesma frase. Para determinar qual conceito utilizar para a classe de equivalência o motor de extração aplica as seguintes regras no pedido listado:

- O formulário especificado do usuário em uma biblioteca.
- O formulário mais frequente, conforme definido pelos recursos pré-compilados.

Etapa 4. Tipo de designação

A seguir, os tipos são designados para conceitos extraídos. Um tipo é um agrupamento semântico de conceitos. Ambos os recursos compilados e as bibliotecas são usados nesta etapa. Os tipos incluem coisas como conceitos de alto nível, palavras positivas e negativas, nomes, locais, organizações e mais. Consulte o tópico “Dicionários de tipo” na página 183 para obter informações adicionais.

Os sistemas linguísticos são sensíveis ao conhecimento — quanto mais informações estão contidas em seus dicionários, mais alta será a qualidade dos resultados. A modificação do conteúdo do dicionário, tais como definições de sinônimo, pode simplificar as informações resultantes. Esse geralmente é um processo iterativo e é necessário para a recuperação de conceito precisa. O NLP é um elemento principal do IBM SPSS Modeler Text Analytics.

Como funciona a extração

Durante a extração de conceitos-chave e ideias de suas respostas, o IBM SPSS Modeler Text Analytics conta com análise de texto baseada em linguística. Esta abordagem oferece a velocidade e a efetividade em custo dos sistemas baseados em estatísticas. Mas, ela oferece um grau muito maior de precisão, enquanto requer muito menos intervenção humana. A análise de texto baseada em linguística é baseada no campo de estudo conhecido como processamento de linguagem natural, também conhecido como linguística computacional.

Entender como o processo de extração funciona pode ajudá-lo a tomar decisões sobre quando usar o ajuste fino nos seus recursos linguísticos (bibliotecas, tipos, sinônimos e mais). Etapas no processo de extração incluem:

- Conversão da origem de dados em um formato padrão
- Identificação de termos candidatos
- Identificação de classes de equivalência e integração de sinônimos
- Designação de um tipo
- Indexação
- Correspondência de extração de padrões e eventos

Etapa 1. Conversão da origem de dados em um formato padrão

Nesta primeira etapa, os dados importados são convertidos para um formato uniforme que pode ser usado para análise adicional. Esta conversão é executada internamente e não muda seus dados originais.

Etapa 2. Identificação de termos candidatos

É importante entender o papel dos recursos linguísticos na identificação de termos candidatos durante a extração linguística. Recursos linguísticos são usados a cada vez que uma extração é executada. Eles existem no formato de modelos, bibliotecas e recursos compilados. As bibliotecas incluem listas de palavras, relacionamentos e outras informações usadas para especificar ou ajustar a extração. Os recursos compilados não podem ser visualizados ou editados. Entretanto, os recursos restantes (modelos) podem ser editados no Editor de Template ou se você estiver em uma sessão interativa do ambiente de trabalho, no Editor de Recursos.

Os recursos compilados são componentes principais internos do mecanismo de extração dentro do IBM SPSS Modeler Text Analytics. Estes recursos incluem um dicionário geral contendo uma lista de formas originais com um código de parte do discurso (substantivo, verbo, adjetivo, advérbio, particípio, coordenador, determinador ou preposição). Os recursos também incluem tipos integrados reservados usados para designar muitos termos extraídos para os tipos a seguir, `<Location>`, `<Organization>` ou `<Person>`. Veja o tópico “Tipos integrados” na página 184 para obter mais informações.

Além desses recursos compilados, várias bibliotecas são fornecidas com o produto e podem ser usadas para complementar os tipos e as definições de conceito nos recursos compilados, bem como para oferecer outros tipos e sinônimos. Estas bibliotecas — e quaisquer umas criadas — são compostas de diversos dicionários. Esses incluem dicionários de tipos, dicionários de substituições (sinônimos e elementos opcionais) e dicionários de exclusão. Consulte o tópico Capítulo 15, “Trabalhando com bibliotecas”, na página 175 para obter informações adicionais.

Assim que os dados foram importados e convertidos, o mecanismo de extração começará a identificação de termos candidatos para extração. Os termos candidatos são palavras ou grupos de palavras que são usados para identificar conceitos no texto. Durante o processamento do texto, palavras simples (*unitermos*) que não estão no recursos compilados são consideradas como extrações de termo candidatas. Palavras compostas candidatas (*multitermos*) são identificadas usando extratores de padrão de parte do discurso. Por exemplo, o multitermo `sports car`, que segue o padrão de parte do discurso "adjetivo substantivo", tem dois componentes. O multiterm `fast sports car`, que segue o padrão "adjetivo adjetivo noun" parte-de-fala, tem três componentes.

Nota: Os termos do dicionário geral compilado supramencionado acima representam uma lista de todas as palavras que provavelmente não têm interesse ou são linguisticamente ambíguas como unitermos. Estas palavras são excluídas da extração quando você está identificando os unitermos. Entretanto, elas são reavaliadas quando você está determinando parte do discurso ou verificando palavras compostas candidatas (multitermos).

Finalmente, um algoritmo especial é usada para manipular sequências de letras maiúsculas, tais como títulos de trabalho, de modo que esses padrões especiais possam ser extraídos.

Etapa 3. Identificação de classes de equivalência e integração de sinônimos

Após unitermos e multitermos candidatos serem identificados, o software usa um conjunto de algoritmos para compará-los e identificar classes de equivalência. Uma classe de equivalência é uma forma original de uma frase ou uma forma única de duas variantes da mesma frase. O objetivo de atribuir frases a classes de equivalência é garantir que, por exemplo, `president of the company` e `company president` não sejam tratados como conceitos separados. Para determinar qual conceito utilizar para a classe de equivalência-ou seja, se `president of the company` ou `company president` é usado como termo de chumbo, o motor de extração aplica as seguintes regras na ordem listada:

- O formulário especificado do usuário em uma biblioteca.
- A forma mais frequente no corpo completo do texto.
- A forma mais curta no corpo completo do texto (que geralmente corresponde à forma original).

Etapa 4. Tipo de designação

A seguir, os tipos são designados para conceitos extraídos. Um tipo é um agrupamento semântico de conceitos. Ambos os recursos compilados e as bibliotecas são usados nesta etapa. Os tipos incluem coisas como conceitos de alto nível, palavras positivas e negativas, nomes, locais, organizações e mais. Tipos adicionais podem ser definidos pelo usuário. Consulte o tópico “Dicionários de tipo” na página 183 para obter informações adicionais.

Etapa 5. Indexação

O conjunto inteiro de registros ou documentos é indexado ao estabelecer um ponteiro entre uma posição de texto e o termo representativo para cada classe de equivalência. Isto assume que todas as instâncias de forma flexionada de um conceito candidato são indexadas como uma forma original candidata. A frequência global é calculada para cada forma original.

Etapa 6. Correspondência de extração de padrões e eventos

O IBM SPSS Modeler Text Analytics pode descobrir não apenas tipos e conceitos, mas também relacionamentos entre eles. Diversos algoritmos e bibliotecas estão disponíveis com este produto e fornecem a capacidade de extrair padrões de relacionamento entre tipos e conceitos. Eles são especialmente úteis ao tentar descobrir opiniões específicas (por exemplo, reações do produto) ou os links relacionais entre pessoas ou objetos (por exemplo, links entre grupos políticos ou genomas).

Como funciona a categorização

Ao criar modelos de categoria no IBM SPSS Modeler Text Analytics, diversas técnicas diferentes que você pode escolher para criar categorias. Como cada conjunto de dados é exclusivo, o número de técnicas e a ordem na qual você as aplica pode mudar. Uma vez que sua interpretação dos resultados pode ser diferente da de outra pessoa, é possível precisar experimentar as diferentes técnicas para ver qual produz os melhores resultados para seus dados de texto. No IBM SPSS Modeler Text Analytics, é possível criar modelos de categoria em uma sessão de ambiente de trabalho no qual você pode explorar e fazer ajuste fino adicional em suas categorias.

Neste guia, **construção de categoria** refere-se à geração de definições de categoria e classificação através do uso de uma ou mais técnicas integradas e **categorização** refere-se ao processo de escoragem

ou rotulação, em que identificadores exclusivos (nome/ID/valor) são designados às definições de categoria para cada registro ou documento.

Durante a construção da categoria, os conceitos e tipos que foram extraídos são usados como os blocos de construção para suas categorias. Ao construir categorias, os registros ou documentos são automaticamente designados para categorias se eles contêm texto que corresponde a um elemento de uma definição de categoria.

O IBM SPSS Modeler Text Analytics oferece diversas técnicas de construção de categoria automatizadas para ajudá-lo a categorizar seus documentos ou registros rapidamente.

Técnicas de Agrupamento

Cada uma das técnicas disponíveis é adequada para certos tipos de dados e situações, mas geralmente é útil combinar técnicas na mesma análise para capturar o intervalo completo de documentos ou registros. Você pode ver um conceito em diversas categorias ou localizar categorias redundantes.

Derivação de Raiz de Conceito. Esta técnica cria categorias pegando um conceito e localizando outros conceitos relacionados a ele analisando se algum dos componentes do conceito está morfologicamente relacionado ou raízes de compartilhamento. Essa técnica é bastante útil para identificar conceitos de palavras compostas sinônimas, já que os conceitos em cada categoria gerada são sinônimos ou estão fortemente relacionados em termos de significado. Isso funciona com dados de vários comprimentos e gera um número menor de categorias compactas. Por exemplo, o conceito `opportunities to advance` seria agrupado com os conceitos `opportunity for advancement` e `advancement opportunity`. Veja o tópico "Derivação de raiz de conceito" na página 106 para obter mais informações.

Rede Semântica. Esta técnica começa identificando os possíveis sentidos de cada conceito abrangente a partir de seu índice extensivo de relacionamentos de palavras e, em seguida, cria categorias ao agrupar conceitos relacionados. Esta técnica é melhor quando os conceitos são conhecidos para a rede semântica e não são ambíguos. Ela é menos útil quando o texto contém terminologia especializada ou jargão desconhecido para a rede. Em um exemplo, o conceito `granny smith apple` poderia ser agrupado com `gala apple` e `winesap apple` já que eles são irmãos da graninha smith. Em outro exemplo, o conceito `animal` pode ser agrupado com `cat` e `kangaroo` já que eles são hiponyms de `animal`. Esta técnica está disponível somente para texto em inglês nesta liberação. Consulte o tópico "Redes semânticas" na página 108 para obter mais informações.

Inclusão de conceito. Esta técnica constrói categorias agrupando conceitos multitermos (palavras compostas), dependendo se elas contêm palavras que são subconjuntos ou superconjuntos de uma palavra na outra. Por exemplo, o conceito `seat` seria agrupado com `safety seat`, `seat belte seat belt buckle`. Consulte o tópico "Inclusão de conceito" na página 108 para obter mais informações.

Coocorrência. Esta técnica cria categorias a partir de coocorrências localizadas no texto. A ideia é que, quando conceitos ou padrões de conceito geralmente são localizados juntos em documentos e registros, essa coocorrência reflete um relacionamento subjacente que provavelmente tem valor em suas definições de categoria. Quando palavras coocorrem significativamente, uma regra de coocorrência é criada e pode ser usada como um descritor de categoria para uma nova subcategoria. Por exemplo, se muitos registros contêm as palavras `price` e `availability` (mas poucos registros contêm um sem o outro), então esses conceitos poderiam ser agrupados em uma regra de co-ocorrência, (`price` & `available`) e atribuídos a uma subcategoria da categoria `price` por exemplo. Consulte o tópico "Regras de coocorrência" na página 109 para obter mais informações.

Número mínimo de documentos. Para ajudá-lo a determinar quão interessantes são as coocorrências, defina o número mínimo de documentos ou registros que devem conter uma determinada coocorrência para que ela possa ser usada como um descritor em uma categoria.

IBM SPSS Modeler Text Analytics Nós

Junto com muitos nós padrão entregues com o IBM SPSS Modeler, também é possível trabalhar com os nós de mineração de texto para incorporar o poder da análise de texto nos fluxos. IBM SPSS Modeler Text Analytics oferece diversos nós de mineração de texto para fazer isso. Esses nós são armazenados na guia IBM SPSS Modeler Text Analytics da paleta do nó.

Os nós a seguir estão incluídos:

- O **Nó de Origem Lista de Arquivos** gera uma lista de nomes de documentos como entrada para o processo de mineração de texto. Isso é útil quando o texto reside em documentos externos em vez de no banco de dados ou outro arquivo estruturado. O nó emite um único campo com um registro para cada documento ou pasta listados, que podem ser selecionados como entrada em um nó Mineração de Texto subsequente. Veja o tópico "nó de lista de arquivos" na página 11 para obter mais informações.
- O **Nó de Origem Web Feed** possibilita a leitura de textos em Web feeds, como blogs ou feeds de notícias, nos formatos RSS ou HTML e usa esses dados no processo de mineração de texto. O nó emite um ou mais campos para cada registro localizado nos feeds, que podem ser selecionados como entrada em um nó Mineração de Texto subsequente. Veja o tópico "Nó Web Feed" na página 13 para obter mais informações.
- O nó **Language Identifier** é um nó de processo que varre o texto fonte para determinar qual linguagem humana ela está escrita e, em seguida, marca isso em um novo campo. Primariamente projetado para ser usado com grandes quantidades de dados, este nódulo é particularmente útil quando você tem mais de um idioma em suas fontes de dados e quer processar apenas uma língua. Veja o tópico "nó de linguagem" na página 17 para obter mais informações.
- O **nó de mineração de texto** usa métodos linguísticos para extrair conceitos-chave do texto, permite criar categorias com esses conceitos e outros dados e oferece a capacidade de identificar relacionamentos e associações entre conceitos com base em padrões conhecidos (chamada análise de link de texto). O nó pode ser usado para explorar o conteúdo dos dados de texto ou para produzir um modelo de conceito ou um modelo de categoria. Os conceitos e as categorias podem ser combinados com dados estruturados existentes, como demográficos, e aplicados à modelagem. Veja o tópico "nó de modelagem de mineração de texto" na página 20 para obter mais informações.
- O **nó Análise de Link de Texto** extrai conceitos e também identifica relacionamentos entre conceitos com base em padrões conhecidos no texto. A extração de padrão pode ser usada para descobrir relacionamentos entre seus conceitos, bem como quaisquer opiniões e qualificadores conectados a esses conceitos. O nó Análise de Ligação de Texto oferece uma maneira mais direta de identificar e extrair padrões do seu texto e incluir os resultados do padrão no conjunto de dados no fluxo. Mas também é possível executar TLA usando uma sessão de ambiente de trabalho interativa no nó Modelagem de Mineração de Texto. Veja o tópico "Nó de análise de link de texto" na página 45 para obter mais informações.
- Durante a mineração de texto a partir de documentos externos, o **Nó de Origem Mineração de Texto** pode ser usado para gerar uma página HTML contendo links para os documentos dos quais conceitos foram extraídos. Veja o tópico "Nó Visualizador de Arquivo" na página 51 para obter mais informações.

Aplicativos

Em geral, qualquer pessoa que precise revisar diariamente grandes volumes de documentos para identificar elementos chave para uma exploração adicional pode ser beneficiada pelo IBM SPSS Modeler Text Analytics.

Alguns aplicativos específicos incluem:

- **Pesquisa científica e médica.** Explore materiais de pesquisa secundários, como relatórios de patentes, artigos de diários e publicações de protocolos. Identifique associações anteriormente desconhecidas (como um médico associado a um determinado produto) apresentando possibilidades para uma exploração adicional. Minimize o tempo gasto no processo de descoberta de drogas. Use-o como um auxílio em pesquisas genômicas.
- **Pesquisa de investimento.** Revise relatórios de analistas diários, artigos de notícias e press releases da empresa para identificar os principais pontos estratégicos ou mudanças de mercado. A análise de tendência dessas informações revela problemas emergentes ou oportunidades para uma firma ou indústria em um período de tempo.
- **Detecção de fraude.** Use em fraudes financeiras e de assistência médica para detectar anomalias e descobrir sinalizadores vermelhos em grandes quantidades de texto.

- **Pesquisa de mercado.** Use em esforços de pesquisa de mercado para identificar os principais tópicos nas respostas às pesquisas de opinião abertas ilimitadas.
- **Análise de blogs e feeds da web.** Explore e construa modelos usando as principais ideias localizadas em feeds de notícias, blogs, etc.
- **CRM.** Construa modelos usando dados de todos os pontos de contato do cliente, como email, transações e pesquisas de opinião.

Capítulo 2. Lendo do texto de origem

Dados para mineração de texto podem estar em qualquer um dos formatos padrão que são usados por IBM SPSS Modeler, incluindo bancos de dados ou outros formatos "retangulares" que representam dados em linhas e colunas, ou em formatos de documentos, como Microsoft Word, Adobe PDFou HTML, que não estão em conformidade com esta estrutura.

- Para ler em texto a partir de documentos que não se adequam à estrutura de dados padrão, incluindo Microsoft Word, Microsoft Excele Microsoft PowerPoint, além de Adobe PDF, XML, HTML e outros, o nó da Lista de Arquivos pode ser usado para gerar uma lista de documentos ou pastas como entrada para o processo de mineração de texto. Para obter informações adicionais, consulte “nó de lista de arquivos” na página 11.
- Para ler em texto a partir de web feed, tais como blogs ou feeds de notícias nos formatos RSS ou HTML, o nó Web Feed pode ser usado para formatar dados do web feed para no processo de mineração de texto. Para obter mais informações, consulte “Nó Web Feed” na página 13.
- Para ler em texto a partir de qualquer um dos formatos de dados padrão utilizados por SPSS Modeler, como um banco de dados com um ou mais campos de texto para comentários do cliente, é possível utilizar qualquer um dos nós de origem SPSS Modeler . Para obter mais informações, consulte a documentação do nó SPSS Modeler .
- Quando você estiver processando grandes quantidades de dados, o que pode incluir texto em vários idiomas diferentes, use o nó Idioma para identificar o idioma usado em um campo específico. Para obter mais informações, consulte “nó de linguagem” na página 17.

nó de lista de arquivos

Para ler texto de documentos não estruturados salvos em formatos como Microsoft Word, Microsoft Excel, e Microsoft PowerPoint, bem como Adobe PDF, XML, HTML e outros, o nó Lista de Arquivos pode ser usado para gerar uma lista de documentos ou pastas como entrada para o processo de mineração de texto. Isso é necessário porque documentos de texto não estruturados não podem ser representados por campos e registros — linhas e colunas — da mesma maneira que outros dados usados pelo IBM SPSS Modeler.

O nó da Lista de Arquivos funciona como um nó de origem.

É possível localizar esse nó na guia IBM SPSS Modeler Text Analytics da paleta de nós na parte inferior da janela IBM SPSS Modeler. Consulte o tópico “IBM SPSS Modeler Text Analytics Nós” na página 7 para obter mais informações.

Importante: Quaisquer nomes de diretórios e nomes de arquivo contendo caracteres que não estão incluídos na codificação local da máquina não são suportados. Ao tentar executar um fluxo contendo um nó Lista de Arquivos, qualquer arquivo ou nomes de diretórios contendo esses caracteres fará com que a execução do fluxo falhe. Isso pode acontecer com nomes de diretórios de idiomas estrangeiros ou nomes de arquivo, como um nome de arquivo alemão em um locale francês.

Suporte de dados locais. Se você estiver conectado a um remoto IBM SPSS Modeler Text Analytics Server e tiver um fluxo com um nó da Lista de Arquivos, os dados devem residir na mesma máquina que o IBM SPSS Modeler Text Analytics Server -ou garantir que a máquina servidor tenha acesso à pasta onde os dados de origem no nó da Lista de Arquivos são armazenados.

Nota: Não é possível usar o nó Lista de Arquivos para escoragem dentro de uma configuração do IBM SPSS Collaboration and Deployment Services-Pontuação.

Nó da lista de arquivos: guia Configurações

Nesta aba você define os diretórios, extensões de arquivo e entrada para este nó.

Nota: A extração de mineração de texto não pode processar Microsoft Office e Adobe PDF arquivos sob as plataformas nãoMicrosoft Windows . Entretanto, arquivos XML, HTML ou de texto sempre podem ser processados.

Quaisquer nomes de diretórios e nomes de arquivo contendo caracteres que não estão incluídos na codificação local da máquina não são suportados. Ao tentar executar um fluxo contendo um nó Lista de Arquivos, qualquer arquivo ou nomes de diretórios contendo esses caracteres fará com que a execução do fluxo falhe. Isso pode acontecer com nomes de diretórios de idiomas estrangeiros ou nomes de arquivo, como um nome de arquivo alemão em um locale francês.

:NONE. Especifica a pasta raiz contendo os documentos que você deseja listar.

- **Incluir subdiretórios.** Especifica que os subdiretórios também devem ser digitalizados.

Tipo(s) de arquivos a incluir na lista: é possível selecionar ou cancelar a seleção dos tipos e extensões de arquivo que você deseja usar. Ao cancelar a seleção de uma extensão de arquivo, os arquivos com tal extensão são ignorados. É possível filtrar pelas extensões a seguir:

Tabela 1. Filtros de tipo de arquivo por extensão de arquivo

• `.rtf`, `.doc`, `.docx`, `.docm`	• `.xls`, `.xlsx`, `.xlsm`	• `.ppt`, `.pptx`, `.ppt m`	• `.txt`, `.text`
• `.htm`, `.html`, `.shtml`	• `.xml`	• `.pdf`	• `.$`

Nota: Para obter mais informações, consulte “nó de lista de arquivos” na página 11.

Se você tiver arquivos com nenhuma extensão, ou uma extensão de ponto rasteiro (por exemplo `File01` ou `File01.`), use a opção **Sem extensão** para selecioná-los.

Apenas oupõe os nomes de nomes de documentos. Selecione esta opção se o campo de saída contará com um ou mais nomes de nomes para o (s) local (s) de onde os documentos residem.

Codificação de entrada. Se o campo de saída conterá texto exato, escolha o valor relevante a partir da seguinte lista:

- Automática (Europeia)
- UTF-8
- UTF-16
- ISO-8859-1
- ISO-8859-2
- Windows-1250
- ascii EUA

A saída é mostrada como texto do documento UTF-8 .

Nó Lista de Arquivos: outras guias

A guia Tipos é uma guia padrão nos nós IBM SPSS Modeler, assim como é a guia Anotações.

Usando o nó Lista de Arquivos na mineração de texto

O nó Lista de Arquivos é usado quando os dados de dados residem em documentos externos não estruturados em formatos como Microsoft Word, Microsoft Excel e Microsoft PowerPoint, bem como Adobe PDF, XML, HTML e outros.

Como exemplo, suponha que conectamos um nó Lista de Arquivos a um nó Mineração de Texto para fornecer texto que reside em documentos externos:

1. **Nó Lista de Arquivos (guia Configurações).** Primeiro, incluímos esse nó no fluxo para especificar onde os documentos de texto serão armazenados. Nós selecionamos o diretório contendo todos os documentos nos quais desejamos executar mineração de texto.
2. **Nó de Mineração de Texto (guia Campos).** Em seguida, incluímos e conectamos um nó Mineração de Texto no nó Lista de Arquivos. Nesse nó, definimos nosso formato de entrada, modelo de recurso e formato de saída. Selecionamos o nome de campo produzido a partir do nó da Lista de Arquivos, do campo de texto e de outras configurações. Veja o tópico "Usando o nó Mineração de Texto em um fluxo" na página 29 para obter mais informações.

Para obter mais informações sobre como usar o nó Mineração de Texto, consulte "nó de modelagem de mineração de texto" na página 20.

Nó Web Feed

O nó Web Feed pode ser usado para preparar dados de texto a partir de web feeds para o processo de mineração de texto. Esse nó aceita web feeds em dois formatos:

- Formato RSS. RSS é um formato padronizado simples baseado em XML para conteúdo da web. A URL para esse formato aponta para uma página que tem um conjunto de artigos vinculados, como fontes de notícias organizadas e blogs. Como RSS é um formato padronizado, cada artigo vinculado é automaticamente identificado e tratado como um registro separado no fluxo de dados resultante. Não é necessária mais nenhuma entrada adicional para que seja possível identificar dados de texto importantes e os registros do feed, a menos que você queira aplicar uma técnica de filtragem ao texto.
- Formato HTML. É possível definir uma ou mais URLs para páginas HTML na guia Entrada. Então, na guia Registros, defina a tag de início do registro, bem como identifique as tags que delimitam o conteúdo da resposta e designe essas tags aos campos de saída de sua escolha (descrição, título, data de modificação, entre outros). Veja o tópico "Nó Web Feed: Guia Registros" na página 14 para obter mais informações.

Importante! Se você estiver tentando recuperar informações sobre a web através de um servidor proxy, você deve ativar o servidor proxy no arquivo `net.properties` para ambos o IBM SPSS Modeler Text Analytics Client e Server. Siga as instruções detalhadas dentro desse arquivo. Isto é aplicável ao acessar a web através do nó Feed da Web ou recuperar uma licença de Software como Serviço (SaaS) SDL já que estas conexões percorrem o Java™. Este arquivo está localizado no *C:\Program Files\IBM\SPSS\Modeler \ 18.5.0\jre\lib\net.properties* por padrão.

A saída desse nó é um conjunto de campos usado para descrever registros. O campo **Descrição** é usado mais comumente por conter grande parte do conteúdo de texto. No entanto, você também pode estar interessado em outros campos, como a descrição simples de um registro (campo **Desc. Simp.**) ou o título do registro (campo **Título**). Quaisquer campos de saída podem ser selecionados como entrada para um nó Mineração de Texto subsequente.

Nota: Não é possível usar o nó Web Feed para escoragem em uma configuração do IBM SPSS Collaboration and Deployment Services-Pontuação.

É possível localizar esse nó na guia IBM SPSS Modeler Text Analytics da paleta de nós na parte inferior da janela IBM SPSS Modeler. Consulte o tópico "IBM SPSS Modeler Text Analytics Nós" na página 7 para obter mais informações.

Nó Web Feed: guia Entrada

A guia Entrada é usada para especificar um ou mais endereços da web, ou URLs, para capturar dados de texto. No contexto de mineração de texto, seria possível especificar URLs para feeds que contêm dados de texto.

Importante: Durante o trabalho com dados não RSS, talvez você prefira usar uma ferramenta de web scraping, como WebQL®, para automatizar a reunião de conteúdo e depois consultar a saída da ferramenta usando um nó de origem diferente.

É possível configurar os parâmetros a seguir:

Inserir ou colar URLs. Neste campo, é possível digitar ou colar uma ou mais URLs. Se você estiver inserindo mais de uma, insira somente uma por linha e use a tecla **Enter/Return** para separar linhas. Insira o caminho da URL completo para o arquivo. Essas URLs podem ser para feeds em um dos dois formatos:

- `RSS format.` RSS é um formato padronizado simples baseado em XML para conteúdo da web. A URL para esse formato aponta para uma página que tem um conjunto de artigos vinculados, como fontes de notícias organizadas e blogs. Como RSS é um formato padronizado, cada artigo vinculado é automaticamente identificado e tratado como um registro separado no fluxo de dados resultante. Não é necessária mais nenhuma entrada adicional para que seja possível identificar dados de texto importantes e os registros do feed, a menos que você queira aplicar uma técnica de filtragem ao texto.
- `HTML format.` É possível definir uma ou mais URLs para páginas HTML na guia Entrada. Então, na guia Registros, defina a tag de início do registro, bem como identifique as tags que delimitam o conteúdo da resposta e designe essas tags aos campos de saída de sua escolha (descrição, título, data de modificação, entre outros). Durante o trabalho com dados não RSS, talvez você prefira usar uma ferramenta de web scraping, como WebQL®, para automatizar a reunião de conteúdo e depois consultar a saída da ferramenta usando um nó de origem diferente. Veja o tópico "Nó Web Feed: Guia Registros" na página 14 para obter mais informações.

Número de entradas mais recentes para ler por URL. Este campo especifica o número máximo de registros para ler para cada URL listada no campo, começando com o primeiro registro localizado no feed. A quantidade de texto afeta a velocidade de processamento durante o recebimento de dados de extração em um nó Mineração de Texto ou nó Análise de Ligação de Texto.

Salvar e reutilizar feeds web anteriores quando possível. Com esta opção, web feeds são digitalizados e os resultados processados são armazenados em cache. Em seguida, em execuções de fluxo subsequentes, se o conteúdo de um determinado feed não tiver mudado, ou se o feed estiver inacessível (uma indisponibilidade da Internet, por exemplo), a versão em cache será usada para acelerar o tempo de processamento. Qualquer novo conteúdo descoberto nesses feeds também será armazenado em cache na próxima vez que você executar o nó.

- **Rótulo**. Se você selecionar **Salvar e reutilizar web feeds anteriores quando possível**, deve-se especificar um nome de rótulo para os resultados. Esse rótulo é usado para descrever os feeds em cache no servidor. Se nenhum rótulo for especificado ou se ele for desconhecido, a reutilização não será possível.

Nó Web Feed: Guia Registros

A guia Registros é usada para especificar o conteúdo do texto de feeds não RSS identificando onde cada novo registro começa, bem como outras informações relevantes referentes a cada registro. Se você souber que um feed não RSS (HTML) contém texto que está em vários registros, deve-se identificar a tag de início do registro ou o texto será tratado como um registro. Embora os feeds RSS sejam padronizados e não requeiram a especificação de nenhuma tag nessa guia, ainda é possível visualizar o conteúdo na guia Visualização.

Importante: Durante o trabalho com dados não RSS, talvez você prefira usar uma ferramenta de web scraping, como WebQL®, para automatizar a reunião de conteúdo e depois consultar a saída da ferramenta usando um nó de origem diferente.

URL. Essa lista suspensa contém uma lista de URLs inseridas na guia Entrada. Ambos os feeds formatados em HTML e RSS estão presentes. Se o endereço da URL for muito longo para a lista suspensa, ele será automaticamente cortado ao meio usando uma reticências para substituir o texto cortado, como *http://www.ibm.com/example/start-of-address...rest-of-address/path.htm*.

- Com **feeds formatados em HTML**, se o feed contiver mais de um registro (ou entrada), é possível definir quais tags HTML contêm os dados correspondentes ao campo mostrado na tabela. Por exemplo, é possível definir a tag de início que indica que um novo registro foi iniciado, uma tag de data de modificação ou um nome de autor.

- Com **feeds formatados em RSS**, não será solicitado que você insira nenhuma tag, já que o RSS é um formato padronizado. No entanto, é possível visualizar resultados de amostra na guia Visualização, se desejado. Todos os feeds RSS reconhecidos são precedidos pela imagem do logotipo do RSS.

Guia de origem. Nesta guia, é possível visualizar o código de origem para quaisquer feeds HTML. Esse código não é editável. É possível usar o campo Localizar para localizar tags específicas ou informações sobre essa página que podem ser copiadas e coladas na tabela abaixo. O campo Localizar não faz distinção entre maiúsculas e minúsculas e corresponderá às sequências de caracteres parciais.

Guia de visualização. Nesta guia, é possível visualizar como um registro será lido no nó Web Feed. Isso é útil principalmente para feeds HTML, já que é possível mudar como um registro será lido definindo tags HTML na tabela abaixo da guia Visualização.

Tag de início de registro não RSS. Esta opção se aplica apenas a feeds não RSS. Se seu feed HTML contiver diversos textos que você deseja dividir em vários registros, especifique a tag HTML que sinaliza o início de um registro (como um artigo ou uma entrada de blog) aqui. Se você não definir um para um feed não RSS, o Modeler tentará adivinhar o formato XML e retornar registros correspondentes. Se o Modelador não conseguir adivinhar o formato XML, nada será devolvido. Se o seu objetivo é importar todo o conteúdo de uma página e depois processá-la depois, recomendamos utilizar leitores XML separados com funcionalidades mais poderosas e, em seguida, importar o resultado para o Modeler Text Analytics.

Tabela de campo. Esta opção se aplica apenas a feeds não RSS. Nesta tabela, é possível dividir o conteúdo de texto em campos de saída específicos inserindo uma tag de início para qualquer um dos campos de saída predefinidos. Insira somente a tag de início. Todas as correspondências são feitas analisando o HTML e correspondendo o conteúdo da tabela aos nomes de tag e atributos localizados no HTML. É possível usar os botões na parte inferior para copiar as tags que você definiu e reutilizá-los para outros feeds.

Tabela 2. Possíveis campos de saída para feeds não RSS (formatos HTML)

Nome do campo de saída	**Conteúdo da Tag Esperado**
Título	A tag delimitando o título do registro. (opcional)
Descrição Simples	A tag delimitando a descrição simples ou rótulo. (opcional)
Descrição	A tag delimitando o texto principal. Se for deixado em branco, este campo conterá todos os outros conteúdos na tag <body> (se houver um registro único) ou o conteúdo localizado dentro do registro atual (quando um delimitador de registro é especificado).
Autor	A tag delimitando o autor do texto. (opcional)
Contribuidores	A tag delimitando os nomes dos contribuidores. (opcional)
Data de Publicação	A tag delimitando a data em que o texto foi publicado. Se for deixado em branco, este campo conterá a data em que o nó lê os dados.
Data de Modificação	A tag delimitando a data em que o texto foi modificado. Se for deixado em branco, este campo conterá a data em que o nó lê os dados.

Quando você insere uma tag na tabela, o feed é digitalizado usando essa tag como a tag mínima para correspondência em vez de correspondência exata. Ou seja, se você inserisse para o campo Título, isso corresponderia a qualquer tag no feed, incluindo aquelas com atributos especificados (como), de modo que é igual à tag raiz ( ) e quaisquer derivados que incluem um atributo e usam tal conteúdo para o campo de saída Título. Se você inserir uma tag raiz, quaisquer atributos adicionais também serão incluídos.

Tabela 3. Exemplos de tags HTML usadas para identificar o texto para os campos de saída			
Se você inserisse:	**Isso corresponderia a:**	**E também corresponderia a:**	**Mas não corresponderia a:**
`<div>`	`<div>`	`<div class="post">`	qualquer outra tag
`<p class="auth">`	`<p class="auth">`	`<p color="black" class="auth" id="85643">`	`<p color="black">`

Nó Web Feed: guia Filtro de Conteúdo

A guia Filtro de Conteúdo é usada para aplicar uma técnica de filtro ao conteúdo do feed RSS. Essa guia não se aplica a feeds HTML. Talvez você queira filtrar se o feed contém muito texto em forma de cabeçalhos, rodapés, menis, publicidade, entre outros. É possível usar essa guia para remover marcas HTML indesejadas, JavaScript e palavras ou linhas curtas do conteúdo.

Filtragem de Conteúdo. Se não quiser aplicar uma técnica de limpeza, selecione **Nenhum**. Caso contrário, selecione **Limpador de Conteúdo RSS**.

Opções de Limpador de Conteúdo RSS. Se você selecionar **Limpador de Conteúdo RSS**, é possível escolher descartar linhas com base em determinados critérios. Uma linha é delimitada por uma marca HTML, como `<p>` e `<li>`, mas excluindo tags sequenciais, como `<span>`, `<b>` e `<font>`. Observe que tags `<br>` são processadas como quebras de linha.

- **Descartar linhas curtas**. Esta opção ignora linhas que não contêm o **número mínimo de palavras** definido aqui.
- **Descartar linhas com palavras curtas**. Esta opção ignora linhas que têm mais do que o **comprimento médio mínimo de palavras** definido aqui.
- **Descartar linhas com muitas palavras de um único caractere**. Esta opção ignora linhas que contêm mais do que que certa **proporção de palavras de um único caractere**.
- **Descartar linhas contendo tags específicas**. Esta opção ignora texto em linhas contendo qualquer uma das tags especificadas no campo.
- **Descartar linhas contendo texto específico**. Esta opção ignora linhas contendo qualquer texto especificado no campo.

Usando o nó Web Feed Node na mineração de texto

O nó Web Feed pode ser usado para preparar dados de texto a partir de web feeds da Internet para o processo de mineração de texto. Esse nó aceita web feeds em formato HTML ou RSS. Esses feeds servem de entrada para o processo de mineração de texto (um nó Mineração de Texto ou Análise de Ligação de Texto subsequente).

Se usar o nó Web Feed, você deve se certificar de especificar que o campo Texto representa **texto real** no nó Mineração de Texto ou Análise de Ligação de Texto para indicar se esses feeds são ligados diretamente a cada entrada do artigo ou blog.

Importante! Se você estiver tentando recuperar informações sobre a web através de um servidor proxy, você deve ativar o servidor proxy no arquivo `net.properties` para ambos o IBM SPSS Modeler Text Analytics Client e Server. Siga as instruções detalhadas dentro desse arquivo. Isto é aplicável ao acessar a web através do nó Feed da Web ou recuperar uma licença de Software como Serviço (SaaS) SDL já que estas conexões percorrem o Java. Este arquivo está localizado no *C:\Program Files\IBM\SPSS\Modeler \ 18.5.0\jre\lib\net.properties* por padrão.

Exemplo: nó Web Feed (Feed RSS) com o nó de modelagem Mineração de Texto

Como um exemplo, suponha que conectamos um nó Web Feed a um nó Mineração de Texto para fornecer dados de texto de um feed RSS para um processo de mineração de texto.

1. **Nó Web Feed (guia Entrada).** Primeiro, incluímos esse nó no fluxo para especificar onde o conteúdo do feed está localizado e para verificar a estrutura do conteúdo. Na primeira guia, fornecemos a

URL para um feed RSS. Como nosso exemplo é para um feed RSS, a formatação já está definida, e não é necessário fazer nenhuma mudança na guia Registros. Um algoritmo de filtragem de conteúdo opcional está disponível para feeds RSS, no entanto, nesse caso, ele não foi aplicado.

2. **Nó de Mineração de Texto (guia Campos).** Em seguida, incluímos e conectamos um nó Mineração de Texto ao nó Web Feed. Nessa guia, definimos a saída do campo de texto pelo nó Web Feed. Nesse caso, queríamos usar o campo **Descrição**. Também selecionamos que a opção de campo Texto representa o **texto real**, bem como outras configurações.
3. **Nó Mineração de Texto (guia Modelo).** Depois, na guia Modelo, escolhemos o modo de construção e os recursos. Nesse exemplo, escolhemos construir um modelo de conceito diretamente a partir desse nó usando o modelo de recurso padrão.

Para obter mais informações sobre como usar o nó Mineração de Texto, consulte "nó de modelagem de mineração de texto" na página 20.

nó de linguagem

Você pode usar o nó Idioma para identificar a linguagem natural de um campo de texto dentro de seus dados de origem.

A saída desse nó é um campo derivado que contém o código de linguagem detectado.

Nota: Não é possível utilizar o nó Idioma para pontuar dentro de uma configuração IBM SPSS Collaboration and Deployment Services-Pontuação .

É possível localizar esse nó na guia IBM SPSS Modeler Text Analytics da paleta de nós na parte inferior da janela IBM SPSS Modeler. Consulte o tópico "IBM SPSS Modeler Text Analytics Nós" na página 7 para obter mais informações.

Nó da Idioma: Tab de Configurações

Nesta aba você especifica como se deve a saída dos detalhes do idioma para um campo de texto selecionado.

Campo de texto Selecione o campo de texto para o qual você deseja identificar o idioma.

Derivar nome do campo Digite um nome para o campo derivado que conterá o código de idioma detectado. O valor padrão é *Idioma*.

Valor padrão para quando a linguagem não pode ser identificada Especificar o nome do campo a ser criado se o idioma não puder ser identificado. As escolhas disponíveis são:

- **Indefinido** Se selecionado, o campo derivado contém valores nulos.
- **Suportado** Se selecionado, você pode escolher a partir de uma das seguintes linguagens ISO suportadas:
  - Inglês (EN)
  - Alemão (DE)
  - Espanhol (ES)
  - Francês (FR)
  - Italiano (IT)
  - Holandês (NL)
  - Português (PT)
- **Custom** Se nenhuma linguagem suportada for adequada, use esta opção para especificar que um valor personalizado deve ser usado. Geralmente este pode ser um código de idioma ISO de 2 letras, mas pode ser qualquer sequência de texto que você exigir.

Capítulo 3. Mineração para conceitos e categorias

O nó de modelagem de Mineração de Texto é usado para gerar um dos dois nuggets do modelo de mineração de texto:

- Os *nuggets do modelo de conceito* revelam e extraem conceitos importantes de seus dados de texto estruturados ou não estruturados.
- Os *nuggets do modelo de categoria* pontuam e atribuem documentos e registros a categorias, que são formadas pelos conceitos (e padrões) extraídos.

Os conceitos e padrões extraídos, bem como as categorias dos seus nuggets do modelo, podem ser todos combinados com dados estruturados existentes, tais como demográficos, e aplicados usando o conjunto inteiro de ferramentas do IBM SPSS Modeler para produzir decisões melhores e mais focadas. Por exemplo, se os clientes frequentemente listarem problemas de login como o impedimento principal para concluírem as tarefas de gerenciamento de conta online, você poderá desejar incorporar "problemas de login" em seus modelos.

Além disso, o nó de modelagem Mineração de Texto é totalmente integrado no IBM SPSS Modeler, para que seja possível implementar fluxos de mineração via IBM SPSS Modeler Solution Publisher para escoragem em tempo real de dados não estruturados em aplicativos, tais como PredictiveCallCenter. A capacidade de implementar esses fluxos assegura implementações de mineração de texto de loop fechado com sucesso. Por exemplo, agora sua organização pode analisar notas da área de rascunho de responsáveis pela chamada de entrada ou saída ao aplicar seus modelos preditivos para aumentar a precisão da sua mensagem de marketing em tempo real. O uso do modelo de mineração de texto resulta em fluxos mostrados para melhorar a precisão dos modelos de dados preditivos.

Para executar IBM SPSS Modeler Text Analytics com IBM SPSS Modeler Solution Publisher, inclua o diretório `<install_directory>/ext/bin/spss.TMWBServer` na variável de ambiente `$LD_LIBRARY_PATH` .

No IBM SPSS Modeler Text Analytics, nós sempre nos referimos a conceitos e categorias extraídos. É importante entender o significado de conceitos e categorias, pois eles podem ajudá-lo a tomar decisões mais informadas durante seu trabalho exploratório e construção de modelo.

Conceitos e Nuggets do Modelo de Conceito

Durante o processo de extração, os dados de texto são varridos e analisados para identificar palavras únicas interessantes ou relevantes, tais como `election` ou `peace` e frase, tais como `presidential election`, `election of the president`ou `peace treaties`. Essas palavras e frases são chamadas coletivamente de *termos*. Usando os recursos linguísticos, os termos relevantes são extraídos e termos semelhantes são agrupados em um termo principal denominado **conceito**.

Dessa forma, um conceito pode representar diversos termos subjacentes, dependendo de seu texto e do conjunto de recursos linguísticos que você está usando. Por exemplo, digamos que temos uma pesquisa de satisfação de funcionários e o conceito `salary` foi extraído. Digamos também que quando você examinou os registros associados a `salary`, observou que `salary` não está sempre presente no texto mas, em vez disso, determinados registros continham algo semelhante, tais como os termos `wage`, `wages` e `salaries`. Estes termos são agrupados sob `salary` já que o mecanismo de extração os considerou como semelhantes ou determinou que eles eram sinônimos com base nas regras de processamento ou nos recursos linguísticos. Neste caso, quaisquer documentos ou registros contendo quaisquer desses termos seriam tratados como se contivessem a palavra `salary`.

Se você desejar ver quais termos estão agrupados sob um conceito, é possível explorar o conceito dentro de um ambiente de trabalho interativo ou examinar quais sinônimos são mostrados no modelo de conceito. Consulte o tópico "Termos Subjacentes nos Modelos de Conceito" na página 32 para obter informações adicionais.

Um **nugget do modelo de conceito** contém um conjunto de conceitos que pode ser usado para identificar registros ou documentos que também contêm o conceito (incluindo qualquer um de seus sinônimos

ou termos agrupados). Um modelo de conceito pode ser usado de duas maneiras. A primeira seria explorar e analisar os conceitos que foram descobertos no texto originar ou identificar rapidamente os documentos de interesse. A segunda seria aplicar esse modelo a novos registros ou documentos de texto para identificar rapidamente os mesmos conceitos principais nos novos documentos/registros, tal como a descoberta em tempo real dos conceitos principais nos dados da área de rascunho a partir de uma central de atendimento.

Consulte o tópico "Nugget de mineração de texto: modelo de conceito" na página 30 para obter informações adicionais.

Categorias e Nuggets o Modelo de Categoria

Você pode criar **categorias** que representam, em essência, conceitos ou tópicos de nível superior para capturar as principais ideias, conhecimentos e atitudes expressas no texto. As categorias são compostas pelo conjunto de descritores, tais como *conceitos*, *tipos* e *regras*. Juntos, esses descritores são usados para identificar se um registro ou documento pertence ou não a uma determinada categoria. Um documento ou registro pode ser varrido para ver se alguma parte de seu texto corresponde a um descritor. Se uma correspondência for localizada, o documento/registro será designado a essa categoria. Esse processo é chamado de **categorização**.

As categorias podem ser construídas automaticamente usando o conjunto robusto de técnicas automatizadas do produto, usando insight adicional manualmente que você possa ter com relação aos dados ou uma combinação de ambos. Também é possível carregar um conjunto de categorias pré-construídas a partir de um pacote de análise de texto através da guia Modelo desse nó. A criação manual de categorias ou o refinamento de categorias pode ser feito somente através do ambiente de trabalho interativo. Consulte o tópico "Nó Mineração de Texto: guia Modelo" na página 23 para obter informações adicionais.

Um **nugget do modelo de categoria** contém um conjunto de categorias junto com seus descritores. O modelo pode ser usado para categorizar um conjunto de documentos ou registros com base no texto em cada documento/registro. Cada documento ou registro é lido e, em seguida, designado a cada categoria para a qual uma correspondência do descritor foi localizada. Dessa forma, um documento ou registro poderia ser designado a mais de uma categoria. É possível usar nuggets do modelo de categoria para ver as ideias essenciais em respostas de pesquisa de opinião em aberto ou em um conjunto de entradas de blog, por exemplo.

Consulte o tópico "Nugget de mineração de texto: modelo de categoria" na página 38 para obter informações adicionais.

nó de modelagem de mineração de texto

O nó Mineração de Texto usa técnicas linguísticas e de frequência para extrair conceitos-chave do texto e criar categorias com esses conceitos e outros dados. O nó pode ser usado para explorar o conteúdo dos dados de texto ou para produzir um nugget do modelo de conceito ou um nugget do modelo de categoria. Ao executar esse nó de modelagem, um mecanismo de extração linguística extrai e organiza os conceitos, padrões e/ou categorias usando métodos de processamento de idioma natural.

É possível executar o nó Mineração de Texto e automaticamente produzir um nugget do modelo de conceito ou categoria usando a opção **Gerar diretamente**. Como alternativa, é possível usar uma abordagem mais prática e exploratória usando o modo **Construir interativamente**, no qual não apenas pode extrair conceitos, criar categorias e refinar seus recursos linguísticos, mas também realizar análises de links de texto e explorar clusters. Consulte o tópico "Nó Mineração de Texto: guia Modelo" na página 23 para obter mais informações.

É possível localizar esse nó na guia IBM SPSS Modeler Text Analytics da paleta de nós na parte inferior da janela IBM SPSS Modeler. Consulte o tópico "IBM SPSS Modeler Text Analytics Nós" na página 7 para obter mais informações.

Requisitos. Os nós de modelagem Mineração de Texto aceitam dados de texto de um nó Feed da Web, um nó Lista de Arquivos ou qualquer um dos nós de origem padrão. Esse nó é instalado com o IBM SPSS Modeler Text Analytics e pode ser acessado na paleta do IBM SPSS Modeler Text Analytics.

Nota: Este nó substitui o nó de Extração de Texto, que foi oferecido em versões antigas do produto. Se você tiver fluxos mais antigos que usem os nós antigos ou nuggets de modelo, você deve reconstruir seus fluxos usando o nó Mining de Texto.

Nó Mineração de Texto: guia Campos

Utilize a aba Campos para especificar as configurações de campo para os dados a partir dos quais você estará extraindo conceitos. Considere usar um envio de dados do nó Amostra a partir deste nó ao trabalhar com conjuntos de dados maiores para acelerar os tempos de processamento. Consulte o tópico "Envio de dados de amostra para economizar tempo" na página 29 para obter informações adicionais.

É possível configurar os parâmetros a seguir:

campo ID Selecione o campo contendo o identificador para os registros de texto. Identificadores devem ser números inteiros. O campo de ID serve de índice para os registros de texto individuais. Use um campo de ID se o campo de texto representar o texto a ser minado.

Campo de texto. Selecione o campo contendo o texto a ser extraído. Esse campo depende da origem de dados.

Campo de linguagem Selecione o campo que contém o identificador de linguagem ISO de duas letras. Se você não selecionar um campo, a linguagem de cada documento será assumida como a do gabarito fornecido.

Tipo de Documento. O tipo de documento especifica a estrutura do texto. Selecione um dos seguintes tipos:

- **Texto completo.** Use para a maioria dos documentos ou fontes de texto. O conjunto inteiro de texto é digitalizado para extração. Ao contrário de outras opções, não há configurações adicionais para essa opção.
- **Texto estruturado.** Use para formulários bibliográficos, patentes e quaisquer arquivos que contenham estruturas regulares que possam ser identificadas e analisadas. Esse tipo de documento é usado para ignorar todo ou parte do processo de extração. Ele permite definir separadores de termo, designar tipos e impor um valor de frequência mínimo. Se você selecionar esta opção, você deve clicar no botão **Configurações** e inserir separadores de texto na **Formatação de Texto Estruturado.** área da caixa de diálogo Configurações do Documento. Consulte o tópico "Configurações do documento para a guia Campos" na página 22 para obter mais informações.

Unidade textual. Selecione o modo de extração a seguir:

- **Modo de documento.** Use para documentos que são curtos e semanticamente homogêneos, como artigos de agências de notícias.
- **Modo de parágrafo.** Use para páginas da web e documentos sem tags. O processo de extração divide os documentos semanticamente, aproveitando a vantagem de características como tags internas e sintaxe. Se esse modo for selecionado, a escoragem será aplicada parágrafo por parágrafo. Portanto, por exemplo, a regra `apple & orange` será verdadeira se `apple` e `orange` estiverem localizadas no mesmo parágrafo.

  Nota: Devido à maneira como o texto é extraído de documentos PDF, o **Modo de Parágrafo** não funciona nesses documentos. Isso porque a extração suprime o marcador de retorno de linha.

Configurações do modo de parágrafo. Esta opção está disponível apenas se você configurar a opção de unidade textual para o **Modo de parágrafo**. Especifique os limites de caractere a serem usados em qualquer extração. O tamanho real é arredondado para cima ou para baixo para o período mais próximo. Para assegurar que as associações de palavras produzidas a partir do texto da coleção de documentos sejam representantes, evite especificar um tamanho de extração muito pequeno.

- **mínimo.** Especifique o número mínimo de caracteres a ser usado em qualquer extração.
- **Máximo.** Especifique o número máximo de caracteres a ser usado em qualquer extração.

Modo Partição Use o modo de partição para escolher se partitura com base nas configurações do nó do tipo ou para selecionar outra partição. O particionamento separa os dados em amostras de treinamento e teste.

Configurações do documento para a guia Campos

Formatação de Texto Estruturado

Se desejar ignorar todo ou parte do processo de extração por você ter dados estruturados ou desejar impor regras na manipulação do texto, use a opção de tipo de documento **Texto Estruturado** e declare os campos ou tags contendo o texto na seção **Formatação de Texto Estruturado** da caixa de diálogo Configurações de Documento. Os termos extraídos são derivados somente do texto contido nos campos ou tags declarados (e tags filhas). Qualquer campo ou tag não declarado será ignorado.

Em certos contextos, o processamento linguístico não é necessário e o mecanismo de extração linguístico pode ser substituído por declarações explícitas. Em um arquivo bibliográfico em que os campos de palavras-chave são separados por separadores, como ponto e vírgula (;) ou vírgula (,), é suficiente extrair a sequência de caracteres entre os dois separadores. Por esse motivo, é possível suspender o processo de extração total e definir regras de manipulação especial para declarar separadores de termos, designar tipos ao texto extraído ou impor uma contagem de frequência mínima para extração.

Use as regras a seguir ao declarar elementos de texto estruturado:

- Somente um campo, uma tag ou um elemento por linha pode ser declarado. Eles não precisam estar presentes nos dados.
- Declarações fazem distinção entre maiúsculas e minúsculas.
- Se declarar uma tag que tenha atributos, como `<title id="1234">`, e você quiser incluir todas as variações ou, neste caso, todos os IDs, inclua a tag sem o atributo ou o suporte de ângulo final (>), como `<title`
- Inclua dois pontos após o nome do campo ou tag para indicar que ele é um texto estruturado. Inclua os dois pontos diretamente após o campo ou tag, mas antes de quaisquer separadores, tipos ou valores de frequência, como `author:` ou `<place>:`.
- Para indicar que diversos termos estão contidos no campo ou tag e que um separador está sendo usado para designar os termos individuais, declare o separador após os dois pontos, como `author:,` ou `<section>:;`.
- Para designar um tipo ao conteúdo localizado na tag, declare o nome do tipo após os dois pontos e um separador, como `author:,Person` ou `<place>:;Location`. Declare o tipo usando os nomes conforme eles aparecem no Editor de Recurso.
- Para definir uma contagem de frequência mínima para um campo ou tag, declare um número no final da linha, como `author:,Person1` ou `<place>:;Location5`. Em que n é a contagem de frequência definida; os termos localizados no campo ou tag devem ocorrer pelo menos n vezes no conjunto inteiro de documentos ou registros a ser extraído. Isso também requer a definição de um separador.
- Se você tiver uma tag que contém dois pontos, deve-se preceder os dois pontos com um caractere de barra invertida para que a declaração não seja ignorada. Por exemplo, se você tiver um campo chamado `<topic:source>`, insira-o como `<topic\:source>`.

Para ilustrar a sintaxe, vamos supor que você tenha os seguintes campos bibliográficos recorrentes:

```
author:Morel, Kawashima
abstract:This article describes how fields are declared.
publication:Text Mining Documentation
datepub:March 2010
```

Para esse exemplo, se quiséssemos que o processo de extração focasse no autor e no resumo, mas ignorasse o restante do conteúdo, declararíamos somente os campos a seguir:

```
author:,Person1
abstract:
```

Nesse exemplo, a declaração do campo `author:,Person1` diz que o processamento linguístico foi suspenso nos conteúdos do campo. Em vez disso, ele diz que o campo de autor contém mais de um nome, que é separado do próximo por um separador de vírgula, e esses nomes devem ser designados ao tipo Pessoa; se esse nome ocorrer pelo menos uma vez no conjunto inteiro de documentos ou registros,

ele deverá ser extraído. Como o campo `abstract:` é listado sem quaisquer outras declarações, o campo será digitalizado durante a extração e o processamento linguístico padrão e a tipificação serão aplicados.

Formatação de Texto XML

Se desejar limitar o processo de extração a somente o texto dentro das tags XML específicas, use a opção de tipo de documento **Texto XML** e declare as tags contendo o texto na seção **Formatação de Texto XML** da caixa de diálogo Configurações de Documento. Os termos extraídos são derivados somente do texto contido dentro dessas tags ou suas tags filhas.

Importante! Se desejar ignorar o processo de extração e impor regras nos separadores de termos, designar os tipos ao texto extraído ou impor uma contagem de frequências para os termos extraídos, use a opção **Texto Estruturado** descrita a seguir.

Use as regras a seguir ao declarar tags para a formatação de texto XML:

- Somente uma tag XML por linha pode ser declarada.
- Elementos de tag fazem distinção entre maiúsculas e minúsculas.
- Se uma tag possui atributos, como `<title id="1234">`, e você deseja incluir todas as variações ou, neste caso, todos os IDs, inclua a tag sem o atributo ou o suporte de ângulo final (>), tais como `<title`

Para ilustrar a sintaxe, vamos supor que você tenha o seguinte documento XML:

```
<section>Rules of the Road
     <title id="01234">Traffic Signals</title>
     <p>Road signs are helpful.</p>
</section>
<p>Learning the rules is important.</p>
```

Para esse exemplo, declararemos as tags a seguir:

```
<section>
<title
```

Neste exemplo, uma vez que você declarou a tag `<section>`, o texto nesta tag e suas tags aninhadas, `Traffic Signals` e `Road signs are helpful`, são digitalizados durante o processo de extração. No entanto, `Learning the rules is important` será ignorado, já que a tag <p> não foi declarada explicitamente, nem a tag aninhada dentro de uma tag declarada.

Nó Mineração de Texto: guia Modelo

Use a guia Modelo para especificar o método de construção e configurações gerais do modelo para a saída do nó.

É possível configurar os parâmetros a seguir:

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Modo de construção. Especifica como os nuggets do modelo serão produzidos quando um fluxo com este nó de Mineração de texto for executado. Como alternativa, você pode usar uma abordagem mais prática e exploratória usando o modo **Construir interativamente** no qual é possível não somente extrair conceitos, criar categorias e refinar seus recursos linguísticos, mas também executar análise de link de texto e explorar clusters.

- **Construir interativamente.** Quando um fluxo é executado, esta opção lança uma interface interativa na qual é possível extrair conceitos e padrões, explorar e afinar os resultados extraídos, construir e refinar categorias, afinar os recursos linguísticos (templates, sinônimos, tipos, bibliotecas, etc.), e construir

nuggets de modelo de categoria. Consulte o "Construir interativamente" na página 24 para obter mais informações.

- **Gerar diretamente.** Esta opção indica que, quando o fluxo for executado, um modelo deverá ser automaticamente criado e adicionado à paleta de Modelos. Ao contrário do ambiente de trabalho interativo, nenhuma manipulação adicional é necessária a partir de você no tempo de execução de lado a partir das configurações definidas no nó. Se você selecionar essa opção, aparecerão opções específicas do modelo com as quais é possível definir o tipo de modelo que deseja produzir. Consulte o "Gerar diretamente" na página 25 para obter mais informações.

Loja grandes modelos em AS. Se você tiver uma conexão com IBM SPSS Analytic Server, selecione esta opção para armazenar seus modelos remotamente no servidor.

Nota: Qualquer modelo que seja construído e armazenado em um servidor só pode ser pontuado nesse servidor. Para retomar uma sessão interativa de ambiente de trabalho que contém tal modelo, você precisa de uma conexão com o servidor original que foi usado para criar a sessão.

Copiar recursos de. Ao minerar texto, a extração é baseada não apenas nas configurações na guia Especialista, mas também nos recursos linguísticos. Esses recursos servem como a base para como manipular e processar o texto durante a extração para obter os conceitos, tipos e às vezes padrões. Você pode copiar recursos para este nó a partir de um modelo de recurso, um pacote de análise de texto (`.tap`) ou um arquivo de projeto SPSS Analítica de Texto para Pesquisas de Opinião (`.tas`). Faça sua seleção e, em seguida, clique em **Carregar** para definir o modelo, pacote ou projeto a partir do qual os recursos serão copiados. No momento em que você carrega, uma cópia dos recursos é armazenada no nó. Portanto, se você alguma vez quiser usar um recurso atualizado, você deve recarregá-lo aqui ou em uma sessão interativa de ambiente de trabalho. Para sua conveniência, a data e hora em que os recursos foram copiados e carregados são mostradas no nó. Consulte "Copiando recursos dos modelos e TAPs" na página 26 para obter informações adicionais.

Idioma do texto. Identifica o idioma do texto da mineração. Os recursos copiados no nó controlam as opções de idioma apresentadas. Selecione o idioma para o qual os recursos foram ajustados.

Construir interativamente

Na guia Modelo do nó de modelagem de mineração de texto, é possível escolher um modo de construção para seus nuggets do modelo. Se você escolher **Construir interativamente**, então, uma interface interativa será aberta ao executar o fluxo. Neste ambiente de trabalho interativo, é possível:

- Extrair e explorar os resultados da extração, incluindo conceitos e tipificação para descobrir as ideias importantes em seus dados de texto.
- Usar uma variedade de métodos para construir e estender as categorias de conceitos, tipos, padrões de TLA padrões e regras para que você possa escorar seus documentos e registros nessas categorias.
- Refinar seus recursos linguísticos (modelos de recursos, bibliotecas, dicionários, sinônimos e mais) para que você possa melhorar seus resultados por meio de um processo interativo no qual os conceitos são extraídos, examinados e refinados.
- Executar análise de link de texto (TLA) e usar os padrões de TLA descobertos para criar melhores nuggets do modelo de categoria. O nó de Análise de Texto de Link não oferece as mesmas opções exploratórias ou recursos de modelagem.
- Gerar clusters para descobrir novos relacionamentos e explorar relacionamentos entre conceitos, tipos, padrões e categorias na área de janela Visualização.
- Gerar nuggets do modelo de categoria refinados para a paleta Modelos no IBM SPSS Modeler e usá-los em outros fluxos.

Nota: Não é possível construir um modelo interativo se você estiver criando uma tarefa IBM SPSS Collaboration and Deployment Services.

Usar trabalho de sessão (categorias, TLA, recursos, etc.) do última atualização de nó. Ao trabalhar em uma sessão interativa de ambiente de trabalho, é possível atualizar o nó com dados de sessão (parâmetros de extração, recursos, definições de categoria, etc.). A opção **Usar trabalho de sessão** permite que você reative o ambiente de trabalho interativo usando os dados de sessão salvos. Esta opção

estará desativada na primeira vez que você usar esse nó, já que nenhum dado da sessão poderia ter sido salvo. Para aprender como atualizar o nó com dados da sessão para que você possa usar essa opção, veja "Atualizando nós de modelagem e salvando" na página 75.

Se você ativar uma sessão *com* esta opção, então, as configurações de extração, categorias, recursos e qualquer outro trabalho da última vez em que você executou uma atualização de nó a partir de um ambiente interativo estarão disponíveis quando você ativar uma sessão na próxima vez. Como os dados de sessão salvos são usados com essa opção, determinado conteúdo, tais como os recursos copiados a partir do modelo abaixo e outras guias são desativas e ignorado. Mas, se você ativar uma sessão *sem* essa opção, apenas o conteúdo do nó conforme eles são definidos são usados agora, o que significa que qualquer trabalho anterior que você tenha executado no ambiente de trabalho não estará disponível.

Nota: Se você alterar o nó de origem para o seu fluxo após resultados de extração ter sido armazenado em cache com o **Use session work ...** opção, você precisará executar uma nova extração uma vez que a sessão interativa do ambiente de trabalho seja lançada se você quiser obter resultados de extração atualizados.

Ignorar de extração e reutilizar os dados e resultados em cache. É possível reutilizar qualquer resultado e dados da extração em cache na sessão de ambiente de trabalho interativo. Essa opção é particularmente útil quando você deseja economizar tempo e reutilizar os resultados da extração, em vez de aguardar uma extração completamente nova ser executada quando a sessão é ativada. Para usar essa opção, você deve ter atualizado este nó anteriormente a partir dentro de uma sessão de ambiente de trabalho interativo e escolheu a opção de **Manter o trabalho de sessão e os dados de texto com resultados da extração em cache para reutilização**. Para aprender como atualizar o nó com dados da sessão para que você possa usar essa opção, veja "Atualizando nós de modelagem e salvando" na página 75.

Iniciar sessão por. Selecione a opção indicando a visualização e ação que deseja que ocorra primeiro a ativação da sessão de ambiente de trabalho interativo. Independentemente da visualização na qual você inicia, você pode alternar para qualquer visualização uma vez na sessão.

- **Usando resultados da extração para construir categorias.** Esta opção ativa o ambiente de trabalho interativo na visualização Categorias e Conceitos e, se aplicável, executa uma extração. Nessa visualização, é possível criar categorias e gerar um modelo de categoria. Também é possível alternar para outra visualização. Veja o tópico Capítulo 7, "Modo de ambiente de trabalho interativo", na página 65 para obter mais informações.
- **Explorando resultados de análise de link de texto (TLA).** Esta opção ativa e começa extraindo e identificando relacionamentos entre conceitos dentro do texto, tais como opiniões ou outros links na visualização Análise de Link de Texto. Deve-se selecionar um modelo ou pacote de análise de texto que contém regras de padrão de TLA para usar esta opção e obter resultados. Se você estiver trabalhando com conjuntos de dados maiores, a extração de TLA pode levar algum tempo. Nesse caso, talvez você queira considerar o uso de um envio de dados do nó de Amostra. Veja o tópico Capítulo 11, "Explorando a análise de ligação de texto", na página 145 para obter mais informações.
- **Analisando clusters de co-word.** Esta opção é ativada na visualização Clusters e atualiza quaisquer resultados da extração desatualizados. Nesta visualização, é possível executar análise de cluster de co-word, que produz um conjunto de clusters. O armazenamento em cluster de co-word é um processo que começa ao avaliar a força do valor de link entre dois conceitos com base em sua coocorrência em um determinado registro ou documento e termina com o agrupamento de conceitos fortemente vinculado em clusters. Veja o tópico Capítulo 7, "Modo de ambiente de trabalho interativo", na página 65 para obter mais informações.

Gerar diretamente

Na guia Modelo do nó de modelagem de mineração de texto, é possível escolher um modo de construção para seus nuggets do modelo. Se você escolher **Gerar diretamente**, será possível configurar as opções no nó e, em seguida, apenas executar seu fluxo. A saída é um nugget do modelo de conceito que foi colocado diretamente na paleta Modelos. Diferentemente do ambiente de trabalho interativo, nenhuma manipulação adicional é necessária por sua parte no tempo de execução além das configurações de frequência definidas para esta opção no nó.

Número máximo de conceitos a incluir no modelo. Esta opção, que se aplica apenas ao construir um modelo automaticamente (não interativo), indica que você deseja criar um modelo de conceito. Ela também determina que este modelo deve conter não mais que o número especificado de conceitos.

- **Marque os conceitos com base na frequência mais alta. Número máximo de conceitos.** Iniciando com o conceito com a frequência mais alta, este é o número de conceitos que será selecionado. Aqui, frequência se refere ao número de vezes que um conceito (e todos os seus termos subjacentes) aparece no conjunto inteiro de documentos/registros. Este número poderia ser maior que a contagem de registros, já que um conceito pode aparecer diversas vezes em um registro.
- **Desmarque os conceitos que ocorrem em muitos registros. Porcentagem de registros.** Desmarca conceitos com uma porcentagem de contagem de registros superior ao número especificado. Esta opção é útil para excluir conceitos que ocorrem frequentemente em seu texto ou em cada registro, mas não possuem significado em sua análise.

Otimizar para velocidade de escoragem. Selecionada por padrão, essa opção assegura que o modelo criado seja compacto e obtenha um escore na mais alta velocidade. Desmarcar essa opção cria um modelo muito maior que escora mais lentamente. Entretanto, o modelo maior assegura que as escoragens exibidas inicialmente no modelo de conceito gerado são as mesmas que aquelas obtidas ao escorar o mesmo texto com o nugget do modelo.

Copiando recursos dos modelos e TAPs

Ao minerar texto, a extração é baseada não apenas nas configurações na guia Especialista, mas também nos recursos linguísticos. Esses recursos servem como a base para como tratar e processar o texto durante a extração para obter os conceitos, tipos e às vezes padrões. Você pode copiar recursos para este nó a partir de um *resource template*, e se você estiver no nó de Mineração de Texto, você também pode selecionar um *pacote de análise de texto* (TAP) ou um projeto SPSS Analítica de Texto para Pesquisas de Opinião (`.tas`).

Por padrão, os recursos são copiados no nó a partir do modelo básico para linguagens licenciadas para o seu produto quando você adicionar o nó na tela. Se você possui licenças para vários idiomas, o primeiro idioma selecionado é usado para determinar o modelo a carregar automaticamente.

No momento em que você carrega, uma cópia dos recursos selecionados é armazenada no nó. Apenas os conteúdos do template, TAP ou SPSS Analítica de Texto para Pesquisas de Opinião recursos do projeto são copiados enquanto o template, TAP ou SPSS Analítica de Texto para Pesquisas de Opinião em si não está vinculado ao nó. Isso significa que, se os recursos forem posteriormente atualizados, essas atualizações não estarão automaticamente disponíveis no nó. Resumindo, os recursos carregados no nó são sempre usados, a menos que você recarregue uma nova cópia dos recursos, ou a menos que você atualize um nó de Mineração de Texto e selecione a opção **Usar trabalho de sessão** . Para obter mais informações sobre **Usar trabalho de sessão**, veja mais adiante nesta seção.

Ao selecionar um recurso, escolha um com a mesma linguagem que seus dados de texto. Você só pode usar recursos nas línguas para as quais está licenciado. Se você desejar executar análise de link de texto, deve-se selecionar um modelo que contém padrões de TLA. Se um modelo contém padrões de TLA, um ícone aparecerá na coluna TLA da caixa de diálogo Carregar Recursos do Modelo.

Nota: Não é possível carregar TAPs ou SPSS Analítica de Texto para Pesquisas de Opinião projetos no nó de Análise de Link de Texto.

Modelos de recursos

Um modelo de recurso é um conjunto predefinido de bibliotecas e recursos linguísticos e não linguísticos avançados que foram ajustados para um determinado domínio ou uso. No nó de modelagem de mineração de texto, uma cópia dos recursos de um modelo básico já estão carregados no nó quando você inclui o nó no fluxo, mas você pode mudar modelos ou carregar um pacote de análise de texto ao selecionar **Modelo de recurso** ou **Pacote de análise de texto** e, em seguida, clicar em **Carregar**. Para modelos, você pode, então, selecionar o modelo na caixa de diálogo Carregar Modelo de Recurso.

Nota: Se você não ver o template que deseja na lista mas você tem uma cópia exportada em sua máquina, você pode importá-lo agora. Também é possível exportar a partir desta caixa de diálogo para

compartilhar com outros usuários. Veja “Importando e exportando modelos” na página 171 para obter mais informações.

Pacotes de análise de texto (TAPs) e Análise de Texto para Projetos de Pesquisa de Opines (TAS)

Um pacote de análise de texto (TAP) é um conjunto predefinido de bibliotecas e recursos linguísticos e não linguísticos avançados empacotados com um ou mais conjuntos de categorias predefinidas. IBM SPSS Modeler Text Analytics oferece vários TAPs pré-construídos, que é afinado para um domínio específico. Você pode editar esses TAPs e salvá-los a um diretório diferente para usá-los para saltar iniciar seu prédio de modelo de categoria. Também é possível criar seus próprios TAPs na sessão interativa. Consulte o “Carregando pacotes de análise de texto” na página 132 para obter mais informações.

Se você optar por importar um projeto SPSS Analítica de Texto para Pesquisas de Opinião (`.tas`), ele será convertido em um TAP.

Nota: Não é possível carregar TAPs ou SPSS Analítica de Texto para Pesquisas de Opinião projetos no nó de Análise de Link de Texto.

Usando a opção "Usar Trabalho de Sessão" (guia Modelo)

Embora os recursos sejam copiados no nó na guia Modelo, você também pode fazer mudanças posteriormente nos recursos em uma sessão interativa e desejar atualizar o nó de modelagem de mineração de texto com essas mudanças mais recentes. Nesse caso, você deve selecionar a opção **Usar trabalho de sessão** na guia Modelo do nó de modelagem de mineração de texto.

Se você selecionar o botão **Usar trabalho de sessão**, o botão **Carregar** estará desativado no nó para indicar que tais recursos originários do ambiente de trabalho interativo serão usados em vez de os recursos que foram carregados anteriormente aqui.

Para fazer mudanças nos recursos assim que tiver selecionado a opção **Usar trabalho de sessão**, você pode editar ou alternar seus recursos diretamente dentro da sessão de ambiente de trabalho interativo através da visualização Editor de Recursos. Consulte“Atualizando recursos do nó após o carregamento” na página 170 para obter mais informações.

Nó Mineração de Texto: guia Especialista

A guia Especialista contém determinados parâmetros avançados que impactam como o texto é extraído e manipulado. Os parâmetros nesta caixa de diálogo controlam o comportamento básico, bem como alguns comportamentos avançados, do processo de extração. Entretanto, eles representam apenas uma parte das opções disponíveis para você. Há também um número de recursos e opções linguísticos que impactam os resultados da extração, que são controlados pelo modelo de recurso que você seleciona na guia Modelo. Consulte o tópico “Nó Mineração de Texto: guia Modelo” na página 23 para obter informações adicionais.

Nota: Esta guia inteira será desativado se você tiver selecionado o modo **Construir interativamente** usando informações do ambiente de trabalho interativo na guia Modelo, caso no qual as configurações de extração são obtidas da última sessão de ambiente de trabalho salva.

Você pode definir os seguintes parâmetros sempre que extrair:

Limite a extração a conceitos com uma frequência global de pelo menos [n]. Especifica o número mínimo de vezes que uma palavra ou frase deve ocorrer no texto para que ela seja extraída. Dessa maneira, um valor de 5 limita a extração àquelas palavras ou frases que ocorrem pelo menos cinco vezes no conjunto inteiro de registros ou documentos.

Em alguns casos, a mudança desse limite pode fazer uma enorme diferença nos resultados da extração e, consequentemente, em suas categorias. Digamos que você esteja trabalhando com alguns dados de um restaurante e não deseja aumentar o limite para acima de 1 para essa opção. Nesse caso, você pode localizar *(1), pizza fina (2), pizza de espinafre (2) e pizza favorita (2)* e *pizza favorita (2)* em seus resultados de extração. Entretanto, se limitasse a extração a uma frequência global de 5 ou mais e

refizesse a extração, você não obteria mais três desses conceitos. Em vez disso, você obteria *pizza (7)*, já que *pizza* é a forma mais simples e também essa palavra já existia como uma possível candidata. E dependendo do restante de seu texto, talvez você tenha de fato uma frequência superior a sete se ainda houver outras frases com pizza no texto. Além disso, se a *pizza de espinafre* já for um descritor de categoria, pode ser necessário incluir *pizza* como um descritor, em vez de capturar todos os registros. Por esse motivo, mude esse limite com cuidado sempre que as categorias já tiverem sido criadas.

Observe que essa é uma variável somente de extração; se seu modelo contiver termos (geralmente contém), e um termo para o modelo for localizado no texto, o termo será indexado, independentemente de sua frequência.

Por exemplo, suponhamos que você use um modelo de Recursos Básicos que inclua "los angeles" sob o tipo `<Location>` na biblioteca Core; se o seu documento contém Los Angeles apenas uma vez, então Los Angeles fará parte da lista de conceitos. Para evitar isso, você terá que configurar um filtro para exibir conceitos que ocorrem pelo menos o mesmo número de vezes que o valor inserido no campo **Limitar extração a conceitos com uma frequência global de pelo menos [n]**.

Acomodar erros de pontuação. Esta opção normaliza temporariamente o texto contendo erros de pontuação (por exemplo, uso incorreto) durante a extração para melhorar a extractabilidade de conceitos. Esta opção é muito útil quando o texto é curto e de qualidade ruim (como, por exemplo, em respostas de pesquisa sem estrutura, e-mail e dados CRM) ou quando o texto contém muitas abreviações.

Acomodar a ortografia para um comprimento mínimo de caracteres de palavra de [n] Esta opção aplica uma técnica de agrupamento fuzzy que ajuda a agrupar palavras comumente digitadas ou palavras bem escritas sob um só conceito. O algoritmo de agrupamento difuso temporariamente remove todas as vogais (exceto a primeira) e remove consoantes duplas/triplas das palavras extraída e, em seguida, as compara para ver se elas são a mesma, desta forma `modeling` e `modelling` seriam agrupadas. Entretanto, se cada termo for designado a um tipo diferente, excluindo o tipo `<Unknown>`, a técnica de agrupamento difuso não será aplicada.

Também é possível definir o número mínimo de caracteres *root* necessários antes que o agrupamento difuso seja usado. O número de caracteres raiz em um termo é calculado ao totalizar todos os caracteres e subtrair quaisquer caracteres que formam sufixos de inflexão e, no caso de termos com palavras compostas, determinadores e preposições. Por exemplo, o termo `exercises` seria contado como 8 caracteres raiz no formato “exercise,” já que a letra *s* no final da palavra é uma inflexão (forma plural). De maneira semelhante, `apple sauce` é contado como 10 caracteres raiz (“apple sauce”) e `manufacturing of cars` é contado como 16 caracteres raiz (“manufacturing car”). Este método de contagem é usado apenas para verificar se o agrupamento difuso deve ser aplicado, mas não influencia como as palavras são correspondidas.

Nota: Se você descobrir que certas palavras são mais tarde agrupadas incorretamente, você pode excluir pares de palavras desta técnica declarando-as explicitamente na seção **Agrupamento Fuzzy: Excepções** na guia Recursos Avançados. Veja o tópico “Agrupamento difuso” na página 197 para obter mais informações.

Extrair uniterms Esta opção extrai palavras únicas (uniterms) desde que a palavra não seja já parte de uma palavra composta e se for um substantivo ou uma parte não reconhecida da fala.

Extrair entidades não lingüísticas Esta opção extrai entidades não linguísticas, como números de telefone, números de previdência social, horários, datas, moedas, dígitos, porcentagens, endereços de e-mail e endereços HTTP. Você pode incluir e excluir determinados tipos de entidades não linguísticas na seção **Entidades Não Linguísticas: Configuração** da guia Recursos Avançados. Ao desativar qualquer entidades desnecessárias, o mecanismo de extração não desperdiçará tempo de processamento. Veja o tópico “Configuração” na página 201 para obter mais informações.

Algoritmos Uppercase Esta opção extrai termos simples e compostos que não estão nos dicionários embutidos desde que a primeira letra do termo esteja em maiúseada. Esta opção oferece uma boa maneira de extrair substantivos mais adequados.

Grupos parciais e completos de pessoa juntos quando possível Esta opção agrupa nomes que aparecem de forma diferente no texto em conjunto. Esse recurso é útil já que os nomes são frequentemente referidos em sua forma completa no início do texto e, então, apenas por uma versão mais curta. Esta opção tenta corresponder qualquer unitermo com o tipo`<Unknown>` com a última

palavra de qualquer um dos termos compostos que é digitado como <Person>. Por exemplo, se *doe* for localizado e inicialmente digitado como <Unknown>, o mecanismo de extração verifica se quaisquer termos compostos no tipo <Person> incluem *doe* como a última palavras, tal como *john doe*. Esta opção não se aplica a nomes já que a maioria nunca é extraída como unitermos.

Permutação de palavra não função máxima Esta opção especifica o número máximo de palavras sem função que podem estar presentes ao aplicar a técnica de permutação. Essa técnica permutação agrupa frases semelhantes que diferem uma da outra apenas pelas palavras sem função contidas (por exemplo, de e o), independentemente da inflexão. Por exemplo, digamos que você configure este valor para no máximo duas palavras e ambos `company officials` e `officials of the company` foram extraídas. Nesse caso, ambos os termos extraídos seriam agrupados na lista de conceito final já que ambos os termos são considerados o mesmo quando `of the` é ignorado.

Usar derivação ao agrupar multitermos Ao processar Big Data, selecione esta opção para agrupar multitermos usando regras de derivação.

Nota: Para ativar a extração de resultados de Análise de Texto, você deve iniciar a sessão com a opção **Explorando resultados da análise de link de texto** e também escolher recursos que contêm definições de TLA. Você sempre pode extrair resultados de TLA posteriormente durante uma sessão de ambiente interativo através do diálogo Configurações da Extração. Veja o tópico "Extraindo dados" na página 80 para obter mais informações.

Envio de dados de amostra para economizar tempo

Quando você tem uma grande quantia de dados, os tempos de processamento podem levar de minutos a horas, especialmente ao usar a sessão de ambiente de trabalho interativo. Quanto maior o tamanho dos dados, mais tempo os processos de extração e categorização levará. Para trabalhar de forma mais eficiente, você pode adicionar umIBM SPSS Modeler nós de Amostra a montante a partir do seu nó Texto Mining. Use este nó de Amostra para obter uma amostra aleatória usando um subconjunto menor de documentos ou registros para fazer as primeiras transmissões.

Uma amostra menor geralmente é perfeitamente adequada para decidir como editar seus recursos e até mesmo criar a maioria, se não todas as categorias. E depois de ter executado no conjunto de dados menor e estiver satisfeito com os resultados, é possível aplicar a mesma técnica para criar categorias para o conjunto de dados inteiro. Em seguida, é possível procurar documentos ou registros que não se ajustam às categorias que você criou e fazer ajustes, conforme necessário.

Nota: O nó de Amostra é um nó IBM SPSS Modeler padrão.

Usando o nó Mineração de Texto em um fluxo

O nó de modelagem Mineração de Texto é usado para acessar dados e extrair conceitos em um fluxo. É possível usar qualquer nó de origem para acessar dados, tais como um nó Banco de Dados, um nó Arquivo de Var. um nó Feed da Web ou um nó Arquivo Corrigido. Para texto que reside em documentos externos, um nó Lista de Arquivos pode ser usado.

Exemplo 1: nó Lista de Arquivos e nó Mineração de Texto para construir um nugget do modelo de conceito diretamente

O exemplo a seguir mostra como usar o nó Lista de Arquivos juntamente com o nó de modelagem Mineração de Texto para gerar o nugget do modelo de conceito. Para obter mais informações sobre o uso do nó Lista de Arquivos, veja"nó de lista de arquivos" na página 11.

1. **Nó Lista de Arquivos (guia Configurações).** Primeiro, incluímos esse nó no fluxo para especificar onde os documentos de texto serão armazenados. Nós selecionamos o diretório contendo todos os documentos nos quais desejamos executar mineração de texto.
2. **Nó de Mineração de Texto (guia Campos).** Em seguida, incluímos e conectamos um nó Mineração de Texto no nó Lista de Arquivos. Nesse nó, definimos nosso formato de entrada, modelo de recurso e formato de saída. Selecionamos o nome de campo produzido a partir do nó da Lista de Arquivos

e selecionamos o campo de texto, assim como outras configurações. Veja o tópico "Usando o nó Mineração de Texto em um fluxo" na página 29 para obter mais informações.

3. **Nó Mineração de Texto (guia Modelo)**. Em seguida, na guia Modelo, selecionamos o modo de construção para gerar um nugget do modelo de conceito diretamente a partir desse nó. É possível selecionar um modelo de recurso diferente ou manter os recursos básicos.

Exemplo 2: nós Arquivo Excel e Mineração de Texto para construir um modelo de categoria interativamente

Este exemplo mostra como o nó Mineração de Texto também pode ativar uma sessão de ambiente de trabalho interativo. Para obter mais informações sobre o ambiente de trabalho interativo, veja Capítulo 7, "Modo de ambiente de trabalho interativo", na página 65.

1. **Nó de origem Excel (guia Dados).** Primeiro, incluímos esse nó no fluxo para especificar onde o texto será armazenado.
2. **Nó de Mineração de Texto (guia Campos).** Em seguida, incluímos e conectamos um nó Mineração de Texto. Na primeira guia, definimos nosso formato de entrada. Selecionamos um nome de campo a partir do nó de origem.
3. **Nó Mineração de Texto (guia Modelo).** Em seguida, na guia Modelo, selecionamos para construir um nugget do modelo de categoria interativamente e usamos os resultados da extração para criar categorias automaticamente. Neste exemplo, carregamos uma cópia de recursos e um conjunto de categorias a partir de um pacote de análise de texto.
4. **Sessão de Ambiente de Trabalho Interativo**. Em seguida, executamos o fluxo e a interface de ambiente de trabalho interativo aberta. Após uma extração ter sido executada, começamos a explorar nossos dados e melhorar nossas categorias.

Nugget de mineração de texto: modelo de conceito

Um nugget do modelo de conceito de Mineração de Texto é criado sempre que você executa com sucesso um nó de modelo de Mineração de Texto onde você selecionou a opção para **Gerar um modelo diretamente** na guia Modelo. Um nugget do modelo de conceito de mineração de texto é usado para a descoberta em tempo real dos conceitos-chave em outros dados de texto, tais como dados da área de rascunho a partir de uma central de atendimento.

O nugget do modelo de conceito em si consiste em uma lista de conceitos, que foram designados a tipos. É possível selecionar qualquer ou todos os conceitos nesse modelo para escoragem em relação a outros dados. Ao executar um fluxo contendo um nugget do modelo Mineração de Texto, novos campos são incluídos nos dados de acordo com o modo de construção selecionado na guia Modelo do nó de modelagem Mineração de Texto antes de construir o modelo. Consulte o tópico "Modelo de conceito: guia Modelo" na página 31 para obter informações adicionais.

Se o nugget do modelo foi gerado usando documentos traduzidos, a escoragem será executada no idioma traduzido. Da mesma forma, se o nugget do modelo foi gerado usando inglês como o idioma, é possível especificar um idioma de tradução no nugget do modelo, já que os documentos serão traduzidos para o inglês.

Os nuggets do modelo de Mineração de Texto são colocados na paleta do nugget do modelo (localizada na guia Modelos no lado superior direito da janela IBM SPSS Modeler) quando são gerados.

Visualizando Resultados

Para ver informações sobre o nugget do modelo, clique com o botão direito no nó na paleta de nuggets do modelo e escolha **Navegar** no menu de contexto (ou **Editar** para nós em um fluxo).

Incluindo modelos no fluxo

Para incluir um nugget do modelo em seu fluxo, clique no ícone na paleta de nuggets do modelo e clique na tela de fluxo onde você deseja colocar o nó. Ou clique com o botão direito no ícone e escolha **Incluir no Fluxo** do menu de contexto. Em seguida, conecte o fluxo ao nó e você estará pronto para passar dados para gerar predições.

Cuidado: Se você desejar usar um nugget de escoragem para gerar novamente um nó de modelagem que contém ambos o modelo de categoria e o modelo usado, recomendamos que você crie um TAP e use-o em uma sessão interativa, no lugar do nó de modelagem, antes de gerar o nugget de escoragem.

Modelo de conceito: guia Modelo

Em modelos de conceito, a guia Modelo exibe o conjunto de conceitos que foram extraídos. Os conceitos são apresentados em um formato de tabela com uma linha para cada conceito. O objetivo nessa guia é selecionar qual dos conceitos será usado para escoragem.

Nota: se você gerou um nugget do modelo de conceito em vez disso, essa guia apresentará informações diferentes. Veja o tópico "Nugget do modelo de categoria: guia Modelo" na página 39 para obter mais informações.

Todos os conceitos são selecionados para escoragem por padrão, conforme mostrado nas caixas de seleção na coluna mais à esquerda. Uma caixa selecionada significa que o conceito será usado para escoragem. Uma caixa não selecionada significa que o conceito será excluído da escoragem. É possível marcar diversas linhas ao selecioná-las e clicar em uma das caixas de seleção em sua seleção.

Para saber mais sobre cada conceito, é possível examinar as informações adicionais fornecidas em cada uma das colunas:

Conceito. Esta é palavra ou frase principal que foi extraída. Em alguns casos, esse conceito representa o nome do conceito, bem como alguns outros termos subjacente associados a esse conceito. Para ver quais termos subjacentes fazem parte de um conceito, exiba a área de janela Termos Subjacentes dentro dessa guia e selecione o conceito para ver os termos correspondentes na parte inferior da caixa de diálogo. Veja o tópico "Termos Subjacentes nos Modelos de Conceito" na página 32 para obter mais informações.

Global. Aqui, global (frequência) refere-se ao número de vezes que um conceito (e todos os seus termos subjacentes) aparece no conjunto inteiro de documentos/registros.

- **Gráfico de barras.** A frequência global desse conceito nos dados de texto, apresentada como um gráfico de barras. A barra assume a cor do tipo para o qual o conceito é designado para visualmente distinguir os tipos.
- **%.** A frequência global deste conceito nos dados de texto apresentados como porcentagem.
- **N.** O número real de ocorrências deste conceito nos dados do texto.

Documentos. Aqui, Docs refere-se à contagem de documentos, significando o número de documentos ou registros nos quais o conceito (e todos os seus termos subjacentes) aparece.

- **Gráfico de barras.** A contagem de documentos para esse conceito, apresentada como um gráfico de barras. A barra assume a cor do tipo para o qual o conceito é designado para visualmente distinguir os tipos.
- **%.** A contagem de documentos para este conceito apresentada como porcentagem.
- **N.** O número real de documentos ou registros contendo este conceito.

Tipo. O tipo ao qual o conceito é designado. Para cada conceito, as colunas Global e Docs aparecem em uma cor para denotar o tipo para o qual esse conceito é designado. Um **tipo** é um agrupamento semântico de conceitos. Veja o tópico "Dicionários de tipo" na página 183 para obter mais informações.

Trabalhando com Conceitos

Ao clicar com o botão direito em uma célula na tabela, é possível exibir um menu de contexto no qual você pode:

- **Selecionar todos.** Todas as linhas na tabela serão selecionadas.
- **Copiar.** O(s) conceito(s) selecionado(s) são copiado(s) na área de transferência.
- **Copiar com Campos** Os conceitos selecionados são copiados na área de transferência junto com o título da coluna.
- **Marcar selecionado.** Verifica todas as caixas de seleção para as linhas selecionadas na tabela, portanto, incluindo aqueles conceitos para pontuação.

- **Desmarcar selecionado.** Desmarca todas as caixas de seleção para as linhas selecionadas na tabela.
- **Marcar todos.** Marca todas as caixas de seleção na tabela. Isto resulta em todos os conceitos sendo usados na saída final.
- **Desmarcar todos.** Desmarca todas as caixas de seleção na tabela. Desmarcar um conceito significa que ele não será usado na saída final.
- **Incluir Conceitos.** Exibe a caixa de diálogo Incluir Conceitos. Veja o tópico "Opções para Inclusão de Conceitos para Escoragem" na página 32 para obter mais informações.

Opções para Inclusão de Conceitos para Escoragem

Para marcar ou desmarcar rapidamente tais conceitos que serão usados para escoragem, clique no botão da barra de ferramentas para **Incluir Conceitos.**.

Figura 1. Botão da barra de ferramentas Incluir Conceitos

Clicar neste botão da barra de ferramentas abrirá a caixa de diálogo Incluir Conceitos para permitir que você selecione conceitos baseados em regras. Todos os conceitos que possuem uma marca de seleção na guia Modelo serão incluídos para escoragem. Aplique uma regra neste subdiálogo para mudar quais conceitos serão usados para escoragem.

É possível escolher entre as seguintes opções:

Marque os conceitos com base na frequência mais alta. Número máximo de conceitos. Iniciando com o conceito com a mais alta frequência global, este é o número de conceitos que serão marcados. Aqui, frequência se refere ao número de vezes que um conceito (e todos os seus termos subjacentes) aparece no conjunto inteiro de documentos/registros. Este número poderia ser maior que a contagem de registros, já que um conceito pode aparecer diversas vezes em um registro.

Marcar conceitos com base na contagem de documentos. Contagem mínima. Esta é a contagem de documentos mais baixa necessária para os conceitos a serem marcados. Aqui, contagem de documentos refere-se ao número de documentos/registros nos quais o conceito (e todos os seus termos subjacentes) aparece.

Marcar conceitos designados ao tipo. Selecione um tipo na lista suspensa para marcar todos os conceitos que são designados a esse tipo. Os conceitos são automaticamente designados a tipos durante o processo de extração. Um **tipo** é um agrupamento semântico de conceitos. Os tipos incluem coisas como conceitos de nível superior, palavras e qualificadores positivos e negativos, qualificadores contextuais, nomes, locais, organizações e mais. Veja o tópico "Dicionários de tipo" na página 183 para obter mais informações.

Desmarque os conceitos que ocorrem em muitos registros. Porcentagem de registros. Desmarca conceitos com uma porcentagem de contagem de registros superior ao número especificado. Esta opção é útil para excluir conceitos que ocorrem frequentemente em seu texto ou em cada registro, mas não possuem significado em sua análise.

Desmarcar conceitos designados ao tipo. Desmarca conceitos correspondendo ao tipo selecionado na lista suspensa.

Termos Subjacentes nos Modelos de Conceito

E possível ver os termos subjacentes que são definidos para os conceitos que você selecionou na tabela. Ao clicar no botão de alternância dos termos subjacentes na barra de ferramentas, você pode exibir a tabela de termos subjacente em uma área de janela dividida na parte inferior do diálogo.

Estes termos subjacentes incluem os sinônimos definidos nos recursos linguísticos (independentemente se eles estavam localizados no texto ou não), bem como quaisquer termos plurais/singulares extraídos localizados no texto usado para gerar o nugget do modelo, termos permutados, termos de agrupamento difuso e assim por diante.

Figura 2. Botão da barra de ferramentas Exibir Termos Subjacentes

Nota: não é possível editar a lista de termos subjacentes. Essa lista é gerada através de substituições, definições de sinônimos (no dicionário de substituições), agrupamento difuso e mais – todos os quais são definidos nos recursos linguísticos. Para fazer mudanças em como termos são agrupados sob um conceito ou como eles são manipulados, você deve fazer mudanças diretamente nos recursos (editável no Editor de Recursos no ambiente de trabalho interativo ou no Editor de Template e, em seguida, recarregar no nó) e, em seguida, execute novamente o fluxo para obter um novo nugget do modelo com os resultados atualizados.

Ao clicar com o botão direito na célula que contém um termo ou conceito subjacente, você pode exibir um menu de contexto no qual você pode:

- **Copiar.** A célula selecionada é copiada na área de transferência.
- **Copiar com campos.** A célula selecionada é copiada na área de transferência junto com os títulos de colunas.
- **Selecionar todos.** Todas as células na tabela serão selecionadas.

Modelo de conceito: guia Configurações

A guia Configurações é usada para definir o valor do campo de texto para os novos dados de entrada, se necessário. Ela também é o local onde você define o modelo de dados para sua saída (modo de escoragem).

Nota: Essa guia aparece apenas quando o nugget do modelo é colocado na tela. Ela não existe quando você está acessando esta caixa de diálogo diretamente na paleta Modelos.

Modo de escoragem: conceitos como registros

Com esse modo de escoragem, um novo registro é criado para cada par `concept`/`document`. Tipicamente, há mais registros na saída do que havia na entrada.

Além dos campos de entrada, os novos campos a seguir são incluídos nos dados:

Tabela 4. Campos de saída para "Conceitos como registros"

Campo	Descrição
`Concept`	Contém o nome de conceito extraído localizado no campo de dados de texto.
`Type`	Armazena o tipo de conceito como um nome de tipo completo, como *Localização* ou *Pessoa*. Um **tipo** é um agrupamento semântico de conceitos. Consulte o tópico "Dicionários de tipo" na página 183 para obter mais informações.
`Count`	Exibe o número de ocorrências para esse conceito (e seus termos subjacentes) no corpo do texto (registro/documento).

Quando você seleciona essa opção, todas as outras opções, exceto **Acomodar erros de pontuação** são desativados.

Modo de escoragem: Conceitos como campos

Nos modelos de conceito, para cada registro de entrada, um novo registro é criado para cada conceito localizado em um determinado documento. Portanto, há tantos registros de saída quanto havia na entrada. Entretanto, cada registro (linha) agora contém um novo campo (coluna) para cada conceito que foi selecionado (usando a marca de seleção) na guia Modelo. O valor para cada campo de conceito depende de você selecionar **Sinalizações** ou **Contagens** como seu valor de campo nessa guia.

Nota: Se você estiver usando conjuntos de dados muito grandes, por exemplo, com um banco de dados DB2, usar **Conceitos como campos** pode encontrar problemas de processamento devido à quantia de dados. Neste caso, recomendamos usar **Conceitos como registros** em vez disso.

Valores de Campo. Escolha se o novo campo para cada conceito conterá uma contagem ou um valor de sinalização.

- **Sinalizações.** Esta opção é usada para obter sinalizadores com dois valores distintos na saída, tais como *Sim*/*Não*, *True*/*False*, *T*/*F*, ou *1* e *2*. Os tipos de armazenamento são configurados automaticamente para refletir os valores escolhidos. Por exemplo, se você inserir valores numéricos para as sinalizações, ele serão automaticamente manipulados como um valor de número inteiro. Os tipos de armazenamento para flags podem ser sequência de caracteres, número inteiro, número real ou data/hora. Insira um valor de sinalização para **True** e para **False**.
- **contagens.** Usadas para obter uma contagem de quantas vezes o conceito ocorreu em um determinado registro.

Extensão do nome do campo. Especifique uma extensão para o nome do campo. Nomes do campo são gerados usando o nome do conceito mais esta extensão.

- **Adicionar como.** Especifique onde a extensão deve ser incluída no nome do campo. Escolha o **Prefixo** para incluir a extensão no início da sequência. Escolha o **Sufixo** para incluir a extensão no final da sequência.

Acomodar erros de pontuação. Esta opção normaliza temporariamente o texto contendo erros de pontuação (por exemplo, uso incorreto) durante a extração para melhorar a extractabilidade de conceitos. Esta opção é muito útil quando o texto é curto e de qualidade ruim (como, por exemplo, em respostas de pesquisa sem estrutura, e-mail e dados CRM) ou quando o texto contém muitas abreviações.

Modelo de conceito: guia Campos

A guia Campos define o valor de campo de texto para os novos dados de entrada, se necessário.

Nota: Esta guia aparece apenas quando o nugget modelo é colocado no fluxo. Ela não existe quando você está acessando esta saída diretamente na paleta Modelos.

Campo de texto. Selecione o campo contendo o texto a ser extraído. Esse campo depende da origem de dados.

Tipo de Documento. O tipo de documento especifica a estrutura do texto. Selecione um dos seguintes tipos:

- **Texto completo.** Use para a maioria dos documentos ou fontes de texto. O conjunto inteiro de texto é digitalizado para extração. Ao contrário de outras opções, não há configurações adicionais para essa opção.
- **Texto estruturado.** Use para formulários bibliográficos, patentes e quaisquer arquivos que contenham estruturas regulares que possam ser identificadas e analisadas. Esse tipo de documento é usado para ignorar todo ou parte do processo de extração. Ele permite definir separadores de termo, designar tipos e impor um valor de frequência mínimo. Se você selecionar esta opção, você deve clicar no botão **Configurações** e inserir separadores de texto na **Formatação de Texto Estruturado.** área da caixa de diálogo Configurações do Documento. Consulte o tópico "Configurações do documento para a guia Campos" na página 22 para obter mais informações.

Codificação de entrada. Esta opção estará disponível somente se você indicou que o campo de texto representa **Nomes de Caminho para Documentos**. Isso especifica a codificação de texto padrão. Uma conversão é feita a partir da codificação especificada ou reconhecida para `IS0-8859-1`. Portanto, mesmo se você especificar outra codificação, o mecanismo de extração a converterá em `IS0-8859-1` antes de ela ser processada. Quaisquer caracteres que não se encaixem na definição de codificação `IS0-8859-1` serão convertidos em espaços.

Idioma do texto. Identifica o idioma do texto que está sendo minerado; esse é o idioma principal detectado durante a extração. Entre em contato com o representante de vendas se você estiver

interessado em adquirir uma licença para um idioma suportado para o qual você não tem acesso atualmente.

Modelo de conceito: guia Resumo

A guia Resumo apresenta informações sobre o modelo em si (pasta *Análise*), os campos usados no modelo (pasta *Campos*), as configurações usadas ao construir o modelo (pasta *Configurações da Construção*) e sobre o treinamento do modelo (pasta *Resumo de Treinamento*).

Quando você procura um nó de modelagem pela primeira vez, as pastas na guia Resumo são reduzidas. Para ver os resultados de interesse, use o controle expansor à esquerda da pasta para mostrar os resultados ou clique no botão **Expandir Todos** para mostrar todos os resultados. Para ocultar os resultados após visualizá-los, use o controle expansor para reduzir a pasta específica que você deseja ocultar ou clique no botão **Reduzir Todos** para reduzir todas as pastas.

Usando nuggets do modelo de conceito em um fluxo

Ao usar um nó de modelagem Mineração de Texto, é possível gerar um nugget do modelo de conceito ou um nugget do modelo de categoria (por meio de uma sessão de ambiente de trabalho interativo). O exemplo a seguir mostra como usar um modelo de conceito em um fluxo simples.

Exemplo: o nó Arquivo de Estatísticas com o nugget do modelo de conceito

O exemplo a seguir mostra como usar o nugget do modelo de conceito de Mineração de Texto.

Figura 3. Fluxo de exemplo: nó Arquivo de Estatísticas com um nugget do modelo de conceito de Mineração de Texto

1. **Nó Arquivo de Estatísticas (guia Dados).** Primeiro, incluímos esse nó no fluxo para especificar onde os documentos de texto serão armazenados.

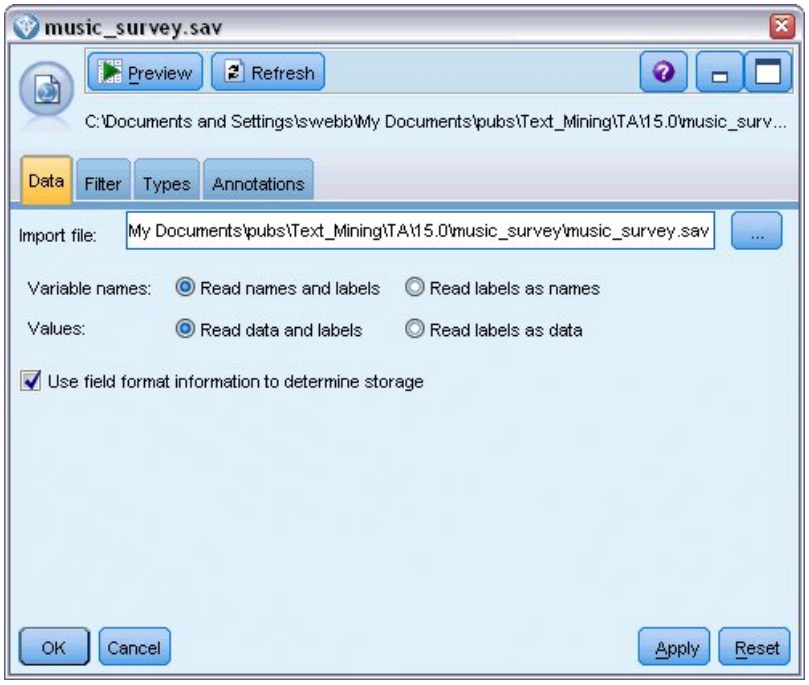


Figura 4. Caixa de diálogo do nó Arquivo de Estatísticas: guia Dados

2. **Nugget do modelo de conceito de Mineração de Texto (guia Modelo).** Em seguida, incluímos e conectamos um nugget do modelo de conceito no nó Arquivo de Estatísticas. Selecionamos os conceitos que desejávamos usar para escorar nossos dados.

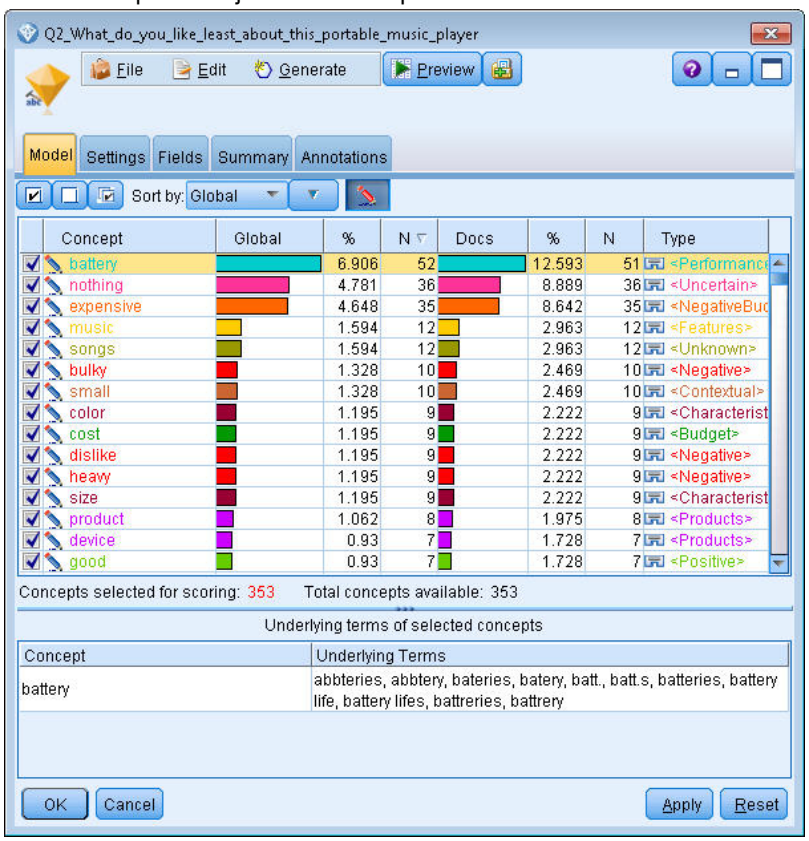


Figura 5. Caixa de diálogo do nugget do modelo de Mineração de Texto: guia Modelo

3. **Nugget do modelo de conceito de Mineração de Texto (guia Configurações).** Em seguida, definimos o formato de saída e selecionamos *Conceitos como campos*. Um novo campo será criado na saída para

cada conceito selecionado na guia Modelo. Cada nome do campo será composto do nome do conceito e do prefixo "*Concept_*"

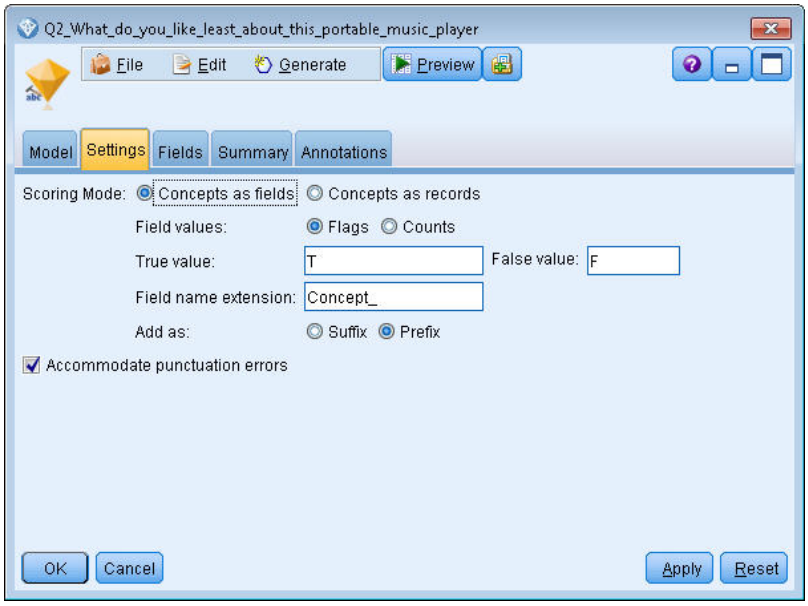


Figura 6. Caixa de diálogo do nugget do modelo de conceito de Mineração de Texto: guia Configurações

4. **Nugget do modelo de conceito de Mineração de Texto (guia Campos).** Em seguida, selecionamos o campo de texto, **Q2_What_do_you_like_least_about_this_portable_music_player**, que é o nome do campo proveniente do nó Arquivo de Estatísticas. Também selecionamos a opção **O campo Texto representa: texto Real**.

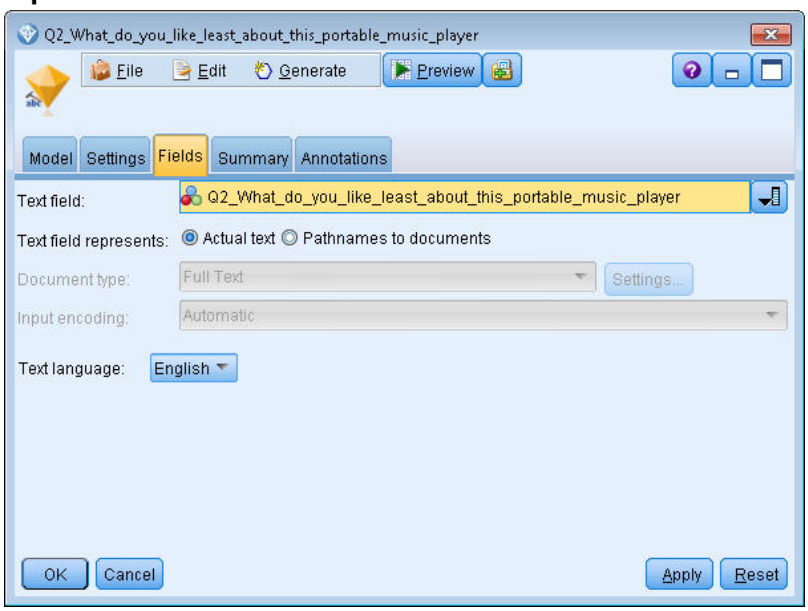
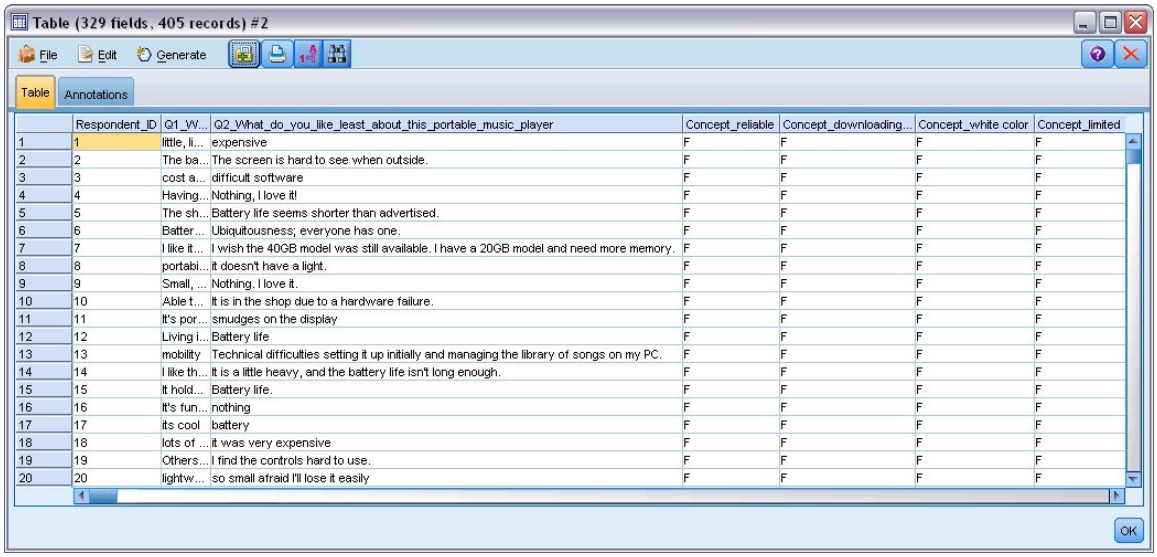


Figura 7. Caixa de diálogo do nugget do modelo de conceito de Mineração de Texto: guia Campos

5. **Nó de Tabela.** Em seguida, nós anexamos um nó de tabela para ver os resultados e executamos o fluxo. A saída de tabela é aberta na tela.



	Respondent_ID	Q1_W...	Q2_What_do_you_like_least_about_this_portable_music_player	Concept_reliable	Concept_downloading...	Concept_white color	Concept_limited
1	1	little, li...	expensive	F	F	F	F
2	2	The ba...	The screen is hard to see when outside.	F	F	F	F
3	3	cost a...	difficult software	F	F	F	F
4	4	Having...	Nothing, I love it!	F	F	F	F
5	5	The sh...	Battery life seems shorter than advertised.	F	F	F	F
6	6	Batter...	Ubiquitousness; everyone has one.	F	F	F	F
7	7	I like it...	I wish the 40GB model was still available. I have a 20GB model and need more memory.	F	F	F	F
8	8	portabi...	it doesn't have a light.	F	F	F	F
9	9	Small, ...	Nothing. I love it.	F	F	F	F
10	10	Able t...	It is in the shop due to a hardware failure.	F	F	F	F
11	11	It's por...	smudges on the display	F	F	F	F
12	12	Living i...	Battery life	F	F	F	F
13	13	mobility	Technical difficulties setting it up initially and managing the library of songs on my PC.	F	F	F	F
14	14	I like th...	It is a little heavy, and the battery life isn't long enough.	F	F	F	F
15	15	It hold...	Battery life.	F	F	F	F
16	16	It's fun...	nothing	F	F	F	F
17	17	its cool	battery	F	F	F	F
18	18	lots of ...	it was very expensive	F	F	F	F
19	19	Others...	I find the controls hard to use.	F	F	F	F
20	20	lightw...	so small afraid I'll lose it easily	F	F	F	F

Figura 8. Saída de tabela rolada para mostrar sinalizações de conceito

Nugget de mineração de texto: modelo de categoria

Um nugget do modelo de categoria de Mineração de Texto é criado sempre que você gera um modelo de categoria a partir do ambiente de trabalho interativo. Este de nugget de modelagem contém um conjunto de categorias, cuja definição é composta de conceitos, tipos, padrões de TLA e/ou regras de categoria. O nugget é usado para categorizar respostas de pesquisa, entradas de blog, outras alimentações da web e quaisquer outros dados de texto.

Se você ativar uma sessão do ambiente de trabalho interativo no nó de modelagem, é possível explorar os resultados da extração, refinar os recursos, realizar ajuste fino nas suas categorias antes de gerar modelos de categoria. Ao executar um fluxo contendo um nugget do modelo Mineração de Texto, novos campos são incluídos nos dados de acordo com o modo de construção selecionado na guia Modelo do nó de modelagem Mineração de Texto antes de construir o modelo. Consulte o tópico "Nugget do modelo de categoria: guia Modelo" na página 39 para obter informações adicionais.

Se o nugget do modelo foi gerado usando documentos traduzidos, a escoragem será executada no idioma traduzido. Da mesma forma, se o nugget do modelo foi gerado usando inglês como o idioma, é possível especificar um idioma de tradução no nugget do modelo, já que os documentos serão traduzidos para o inglês.

Os nuggets do modelo de Mineração de Texto são colocados na paleta do nugget do modelo (localizada na guia Modelos no lado superior direito da janela IBM SPSS Modeler) quando são gerados.

Visualizando Resultados

Para ver informações sobre o nugget do modelo, clique com o botão direito no nó na paleta de nuggets do modelo e escolha **Navegar** no menu de contexto (ou **Editar** para nós em um fluxo).

Incluindo modelos no fluxo

Para incluir um nugget do modelo em seu fluxo, clique no ícone na paleta de nuggets do modelo e clique na tela de fluxo onde você deseja colocar o nó. Ou clique com o botão direito no ícone e escolha **Incluir no Fluxo** do menu de contexto. Em seguida, conecte o fluxo ao nó e você estará pronto para passar dados para gerar predições.

Cuidado: Se você desejar usar um nugget de escoragem para gerar novamente um nó de modelagem que contém ambos o modelo de categoria e o modelo usado, recomendamos que você crie um TAP e use-o em uma sessão interativa, no lugar do nó de modelagem, antes de gerar o nugget de escoragem.

Nugget do modelo de categoria: guia Modelo

Para modelos de categoria, a guia modelo exibe a lista de categorias no modelo de categoria à esquerda e os descritores para uma categoria selecionada à direita. Cada categoria é composta por um número de descritores. Para cada categoria que você seleciona, os descritores associados aparecem na tabela. Esses descritores podem incluir conceitos, regras de categoria, tipos e padrões de TLA. O tipo de cada descritor, bem como alguns exemplo do que cada descritor representa, também é mostrado.

Nesta guia, o objetivo é selecionar as categorias que você deseja usar para escoragem. Para um modelo de categoria, os documentos e registros são escorados em categorias. Se um documento ou registro contém um ou mais dos descritores no seu texto ou quaisquer termos subjacentes, então, esse documento ou registro é designado à categoria à qual o descritor pertence. Esses termos subjacentes incluem os sinônimos definidos nos recursos linguísticos (independentemente se eles estavam localizados no texto ou não), bem como quaisquer termos plurais/singulares extraídos localizados no texto usado para gerar o nugget do modelo, termos permutados, termos de agrupamento difuso e assim por diante.

Nota: se você gerou um nugget do modelo de conceito em vez disso, essa guia conterá resultados diferentes. Veja o tópico "Modelo de conceito: guia Modelo" na página 31 para obter mais informações.

Árvore de Categorias

Para aprender mais sobre cada categoria, selecione essa categoria e revise as informações que aparecem para os descritores nessa categoria. Para cada descritor, é possível revisar as informações a seguir:

- Nome do **descritor**. Este campo contém um ícone representando qual tipo de descritor ele é, bem como o nome do descritor.

Tabela 5. Ícones do Descritor

	Conceitos		Padrões de TLA
	Tipos		Regras de Categoria

- **Tipo.** Este campo contém o nome do tipo para o descritor. Tipos são coleções de conceitos semelhantes (agrupamentos semânticos), tais como nomes de organizações, produtos ou opiniões positivas. As regras não são designadas a tipos.
- **Detalhes.** Este campo contém uma lista do que está incluído em tal descritor. Dependendo do número de correspondências, você pode não ver a lista inteira para cada descritor devido a limitações de tamanho na caixa de diálogo.

Selecionando e Copiando Categorias

Todas as categorias principais são selecionadas para escoragem por padrão, conforme mostrado nas caixas de seleção na área de janela à esquerda. Uma caixa marcada significa que a categoria será usada para escoragem. Uma caixa desmarcada significa que a categoria será excluída da escoragem. É possível marcar diversas linhas ao selecioná-las e clicar em uma das caixas de seleção em sua seleção. Além disso, se uma categoria ou subcategoria for selecionada, mas uma de suas subcategorias não for selecionada, então, a caixa de seleção mostra um plano de fundo azul para indicar que há apenas uma seleção parcial nos filhos da categoria selecionada.

Ao clicar com o botão direito em uma categoria na árvore, é possível exibir um menu de contexto a partir do qual você pode:

- **Marcar selecionado.** Marca todas as caixas de seleção para as linhas selecionadas na tabela.
- **Desmarcar selecionado.** Desmarca todas as caixas de seleção para as linhas selecionadas na tabela.
- **Marcar todos.** Marca todas as caixas de seleção na tabela. Isto resulta em todas as categorias sendo usadas na saída final. Também é possível usar o ícone de caixa de seleção correspondente na barra de ferramentas.

- **Desmarcar todos.** Desmarca todas as caixas de seleção na tabela. Desmarcar uma categoria significa que ela não será usada na saída final. Também é possível usar o ícone de caixa de seleção vazia correspondente na barra de ferramentas.

Ao clicar com o botão direito em uma célula na tabela do descritor, é possível exibir um menu de contexto no qual você pode:

- **Copiar.** O(s) conceito(s) selecionado(s) são copiado(s) na área de transferência.
- **Copiar com campos.** O descritor selecionado é copiado na área de transferência junto com os títulos da coluna.
- **Selecionar todos.** Todas as linhas na tabela serão selecionadas.

Nugget do modelo de categoria: guia Configurações

A guia Configurações é usada para definir o valor do campo de texto para os novos dados de entrada, se necessário. Ela também é o local onde você define o modelo de dados para sua saída (modo de escoragem).

Nota: Esta guia aparece na caixa de diálogo do nó apenas quando o nugget do modelo é colocado na tela ou em um fluxo. Ela não existe quando você está acessando este nugget diretamente na paleta Modelos.

Modo de escoragem: categorias como campos

Com esta opção, há tantos registros de saída quanto havia na entrada. Entretanto, cada registro agora contém um novo campo para cada categoria que foi selecionada (utilizando a marca de seleção) na guia Modelo. Para cada campo, digite um valor de sinalização para **True** e para **False**, como *Yes*/*No*, *True*/*False*, *T*/*F*, ou *1* e *2*. Os tipos de armazenamento são configurados automaticamente para refletir os valores escolhidos. Por exemplo, se você inserir valores numéricos para as sinalizações, ele serão automaticamente manipulados como um valor de número inteiro. Os tipos de armazenamento para flags podem ser sequência de caracteres, número inteiro, número real ou data/hora.

Nota: Se você estiver usando conjuntos de dados muito grandes, por exemplo com um banco de dados DB2, usar **Categorias como campos** pode encontrar problemas de processamento devido à quantia de dados. Neste caso, recomendamos usar **Categorias como registros** em vez disso.

Extensão do nome do campo. É possível escolher especificar um prefixo/sufixo de extensão para o nome do campo ou é possível escolher usar os códigos de categoria. Nomes do campo são gerados usando o nome da categoria mais esta extensão.

- **Adicionar como.** Especifique onde a extensão deve ser incluída no nome do campo. Escolha o **Prefixo** para incluir a extensão no início da sequência. Escolha o **Sufixo** para incluir a extensão no final da sequência.

Se uma subcategoria estiver não selecionada. Essa opção permite especificar como os descritores pertencentes a subcategorias que não foram selecionadas para escoragem serão manipulados. Há duas opções.

- A opção **Excluir completamente seus descritores da escoragem** fará com que os descritores de subcategorias sem marcas de verificação (não selecionados) sejam ignorados e não usados durante a escoragem.
- A opção **Agregar descritores com aqueles na categoria-pai** fará com que os descritores de subcategorias sem marcas de verificação (não selecionados) sejam usados como descritores para a categoria-pai (a categoria acima dessa subcategoria). Se vários níveis de subcategorias estiverem não selecionados, os descritores serão movidos para cima sob a primeira categoria-pai disponível.

Pontuar apenas categoria de correspondência de nível inferior. Use esta opção para saída da categoria apenas em uma única linha (por exemplo, se a categoria for `GeneralSatisfaction/Pos`, selecionar esta opção resulta em `GeneralSatisfaction/Pos`. Sem essa opção, você obterá duas linhas: `GeneralSatisfaction` e `GeneralSatisfaction/Pos`).

Acomodar erros de pontuação. Esta opção normaliza temporariamente o texto contendo erros de pontuação (por exemplo, uso incorreto) durante a extração para melhorar a extractabilidade de conceitos.

Esta opção é muito útil quando o texto é curto e de qualidade ruim (como, por exemplo, em respostas de pesquisa sem estrutura, e-mail e dados CRM) ou quando o texto contém muitas abreviações.

Modo de escoragem: Categorias como registros

Com esta opção, um novo registro é criado para cada par `category`, `document`. Tipicamente, há mais registros na saída do que havia na entrada. Além dos campos de entrada, novos campos também são incluídos nos dados, dependendo de qual tipo de modelo ele é.

Tabela 6. Campos de saída para "Categorias como registros"

Novo Campo de Saída	Descrição
`Category`	Contém o nome da categoria à qual o documento de texto foi designado. Se a categoria for uma subcategoria de outra, então o caminho completo para o nome da categoria será controlado pelo valor que você escolheu neste diálogo.

Valores para categorias hierárquicas. Esta opção controla como os nomes de subcategorias são exibidos na saída.

- **Caminho completo da categoria.** Esta opção produzirá o nome da categoria e o caminho completo das categorias pai, se aplicável, usando barras para separar os nomes de categoria dos nomes de subcategoria.
- **Caminho abreviado da categoria.** Esta opção produzirá apenas o nome da categoria, mas usará reticências para mostrar o número de categorias pai para a categoria em questão.
- **Categoria de nível inferior.** Esta opção produzirá apenas o nome da categoria sem o caminho completo ou as categorias pai mostradas.

Se uma subcategoria estiver não selecionada. Essa opção permite especificar como os descritores pertencentes a subcategorias que não foram selecionadas para escoragem serão manipulados. Há duas opções.

- A opção **Excluir completamente seus descritores da escoragem** fará com que os descritores de subcategorias sem marcas de verificação (não selecionados) sejam ignorados e não usados durante a escoragem.
- A opção **Agregar descritores com aqueles na categoria-pai** fará com que os descritores de subcategorias sem marcas de verificação (não selecionados) sejam usados como descritores para a categoria-pai (a categoria acima dessa subcategoria). Se vários níveis de subcategorias estiverem não selecionados, os descritores serão movidos para cima sob a primeira categoria-pai disponível.

Acomodar erros de pontuação. Esta opção normaliza temporariamente o texto contendo erros de pontuação (por exemplo, uso incorreto) durante a extração para melhorar a extractabilidade de conceitos. Esta opção é muito útil quando o texto é curto e de qualidade ruim (como, por exemplo, em respostas de pesquisa sem estrutura, e-mail e dados CRM) ou quando o texto contém muitas abreviações.

Nugget do modelo de categoria: outras guias

A guia Campos e a guia Configurações para o nugget do modelo de categoria são as mesmas que para o nugget do modelo de conceito.

- Guia Campos. Consulte o tópico “Modelo de conceito: guia Campos” na página 34 para obter informações adicionais.
- Guia Resumo. Consulte o tópico “Modelo de conceito: guia Resumo” na página 35 para obter informações adicionais.

Usando nuggets do modelo de categoria em um fluxo

O nugget do modelo de categoria de Mineração de Texto é gerado a partir de uma sessão de ambiente de trabalho interativo. É possível usar esse nugget do modelo em um fluxo.

Exemplo: nó Arquivo de Estatísticas com o nugget do modelo de categoria

O exemplo a seguir mostra como usar o nugget do modelo de Mineração de Texto.

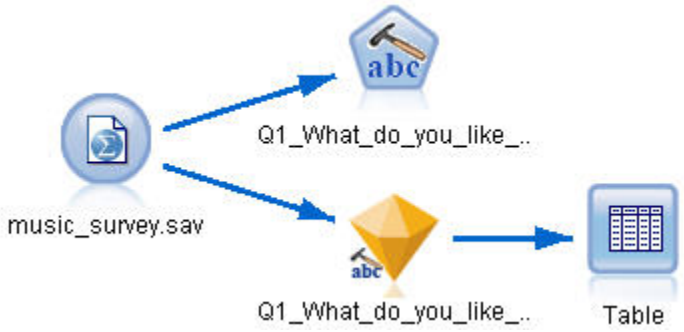


Figura 9. Fluxo de exemplo: nó Arquivo de Estatísticas com um nugget do modelo de categoria de Mineração de Texto

1. **Nó Arquivo de Estatísticas (guia Dados).** Primeiro, incluímos esse nó no fluxo para especificar onde os documentos de texto serão armazenados.

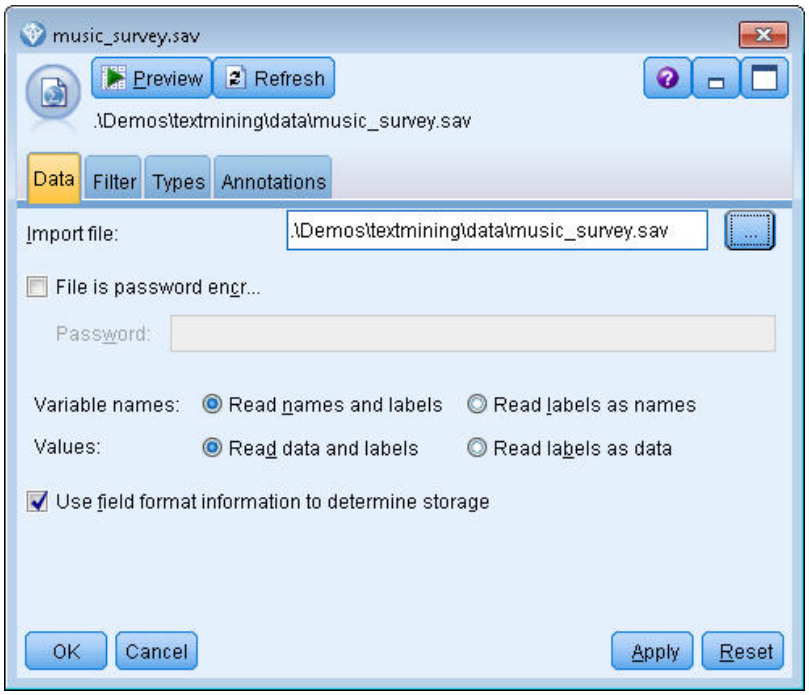


Figura 10. Caixa de diálogo do nó Arquivo de Estatísticas: guia Dados

2. **Nugget do modelo de categoria de Mineração de Texto (guia Modelo).** Em seguida, incluímos e conectamos um nugget do modelo de categoria no nó Arquivo de Estatísticas. Selecionamos as categorias que desejávamos usar para escorar nossos dados.

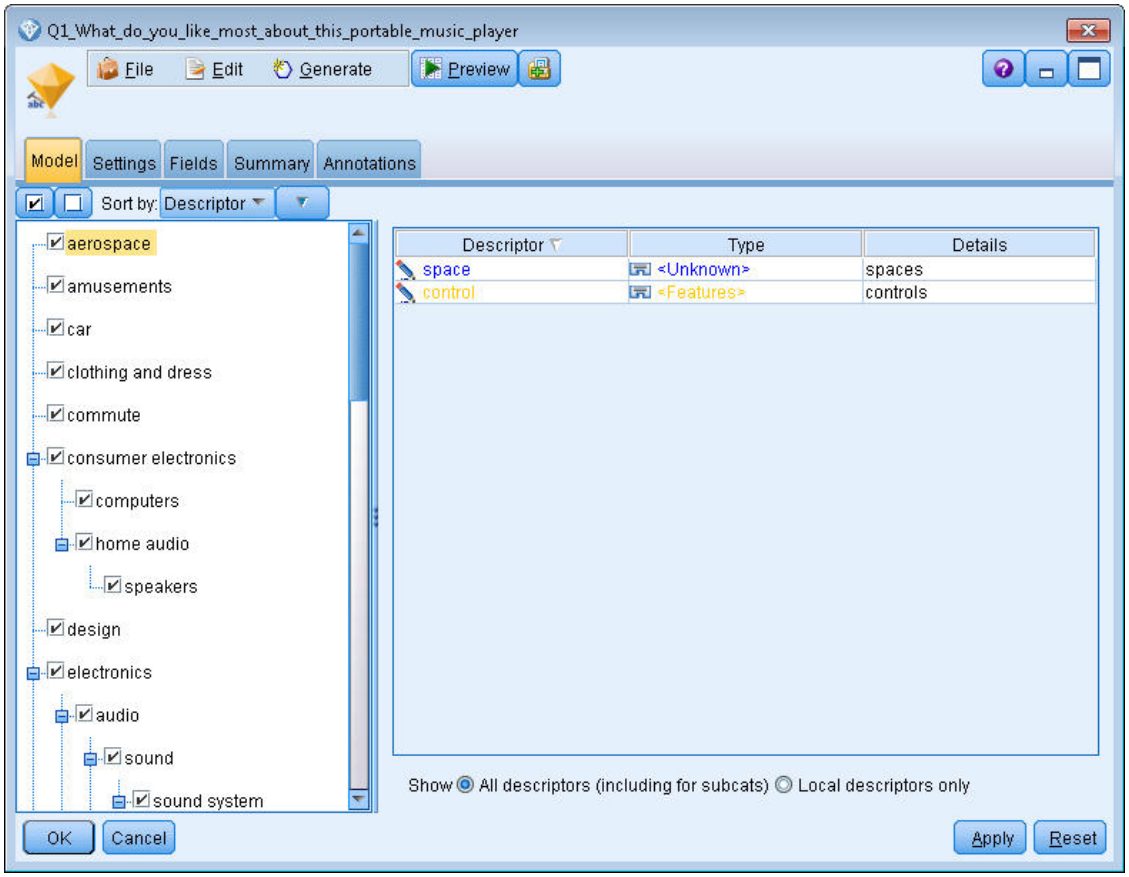


Figura 11. Caixa de diálogo do nugget do modelo de Mineração de Texto: guia Modelo

3. **Nugget do modelo de Mineração de Texto (guia Configurações).** Em seguida, definimos o formato da saída **Categorias como campos**.

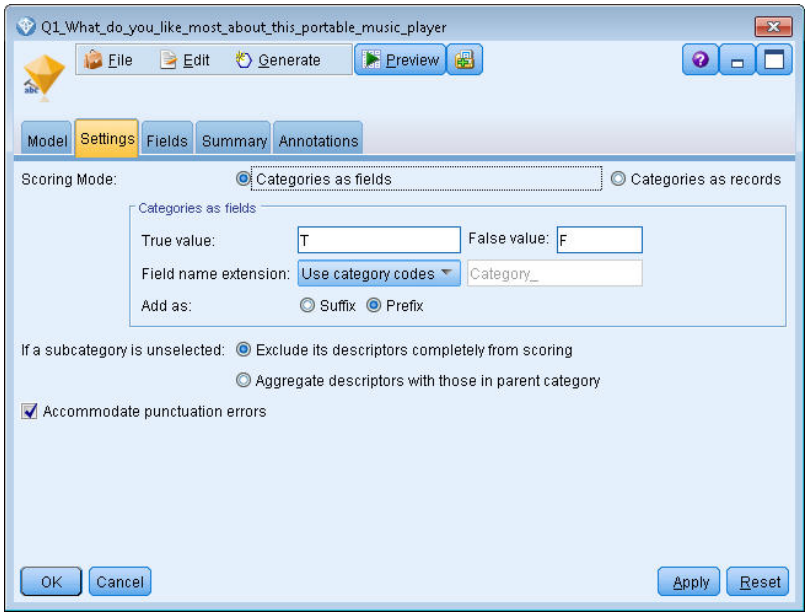


Figura 12. Caixa de diálogo do nugget do modelo de categoria: guia Configurações

4. **Nugget do modelo de categoria de Mineração de Texto (guia Campos).** Em seguida, selecionamos a variável de campo de texto, que é o nome do campo proveniente do nó Arquivo de Estatísticas e selecionou a opção O campo Texto representa **texto Real**, bem como outras configurações.

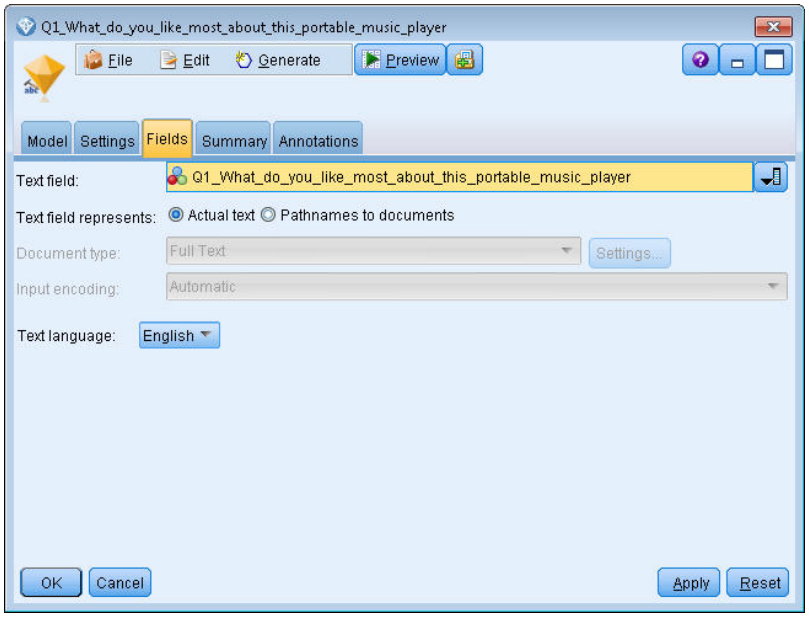
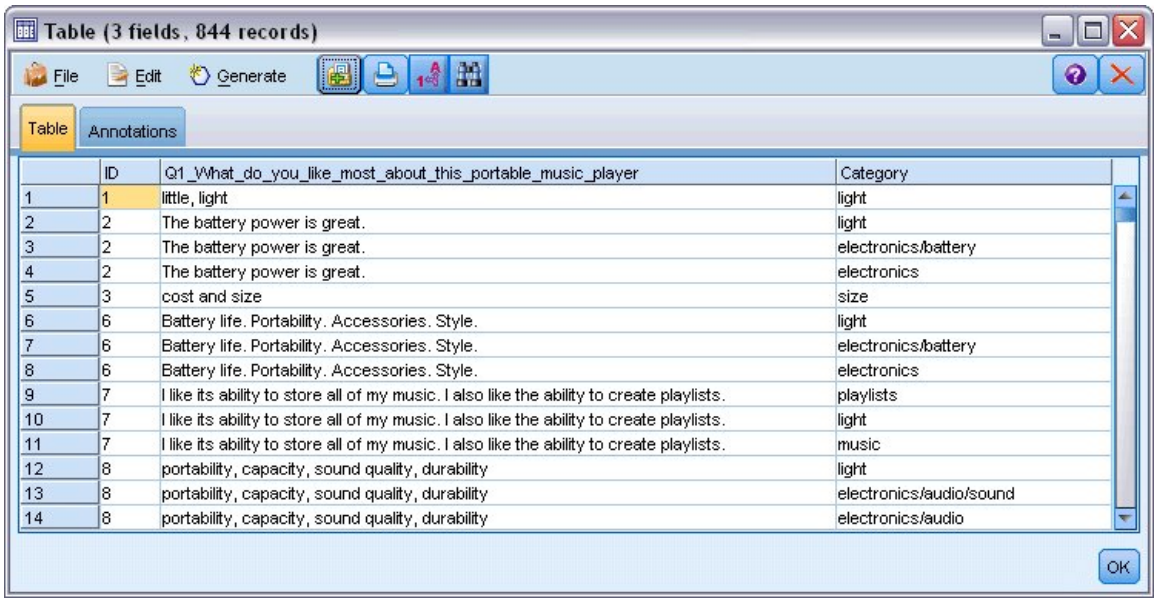


Figura 13. Caixa de diálogo do nugget do modelo de Mineração de Texto: guia Campos

5. **Nó de Tabela.** Em seguida, nós anexamos um nó de tabela para ver os resultados e executamos o fluxo.



	ID	Q1_What_do_you_like_most_about_this_portable_music_player	Category
1	1	little, light	light
2	2	The battery power is great.	light
3	2	The battery power is great.	electronics/battery
4	2	The battery power is great.	electronics
5	3	cost and size	size
6	6	Battery life. Portability. Accessories. Style.	light
7	6	Battery life. Portability. Accessories. Style.	electronics/battery
8	6	Battery life. Portability. Accessories. Style.	electronics
9	7	I like its ability to store all of my music. I also like the ability to create playlists.	playlists
10	7	I like its ability to store all of my music. I also like the ability to create playlists.	light
11	7	I like its ability to store all of my music. I also like the ability to create playlists.	music
12	8	portability, capacity, sound quality, durability	light
13	8	portability, capacity, sound quality, durability	electronics/audio/sound
14	8	portability, capacity, sound quality, durability	electronics/audio

Figura 14. Resultado da tabela

Capítulo 4. Minerando para links de texto

Nó de análise de link de texto

O nó Análise de Ligação de Texto (TLA) inclui uma tecnologia de correspondência de padrões na extração do conceito de mineração de texto para identificar relacionamentos entre os conceitos nos dados de texto baseados em padrões conhecidos. Esses relacionamentos podem descrever como um cliente se sente em relação a um produto, quais empresas estão fazendo negócios juntas ou até mesmo os relacionamentos entre genes e agentes farmacêuticos.

Por exemplo, a extração do nome do produto do seu concorrente pode não ser interessante o suficiente para você. Usando esse nó, seria possível saber também como as pessoas se sentem em relação a esse produto, caso existam opiniões nos dados. Relacionamentos e associações são identificados e extraídos correspondendo padrões conhecidos com seus dados de texto.

É possível usar as regras do padrão TLA dentro de certos modelos de recursos fornecidos com IBM SPSS Modeler Text Analytics ou criar/editar os seus. Regras de padrão são compostas por macros, listas de palavras e diferenças de palavras para formar um query booleano, ou regra, que seja comparado com seu texto de entrada. Sempre que uma regra do padrão TLA corresponde a um texto, esse texto pode ser extraído como um resultado de TLA e reestruturado como dados de saída. Veja o tópico Capítulo 18, “Sobre regras de ligação de texto”, na página 207 para obter mais informações.

O nó Análise de Ligação de Texto oferece uma maneira mais direta de identificar e extrair resultados do padrão TLA do seu texto e incluir os resultados no conjunto de dados no fluxo. Mas o nó Análise de Ligação de Texto não é a única maneira de executar uma análise de ligação de texto. Também é possível usar uma sessão de ambiente de trabalho interativa no nó Modelagem de Mineração de Texto.

No ambiente de trabalho interativo, é possível explorar os resultados do padrão TLA e usá-los como descritores de categoria e/ou aprender mais sobre os resultados usando drill down e gráficos. Consulte o tópico Capítulo 11, “Explorando a análise de ligação de texto”, na página 145 para obter informações adicionais. De fato, o uso do nó Mineração de Texto para extrair resultados de TLA é uma grande maneira de explorar e ajustar modelos aos seus dados para usar depois diretamente no nó TLA.

A saída pode ser representada em até seis slots ou partes. Consulte o tópico “Saída do nó TLA” na página 48 para obter mais informações.

É possível localizar esse nó na guia IBM SPSS Modeler Text Analytics da paleta de nós na parte inferior da janela IBM SPSS Modeler. Consulte o tópico “IBM SPSS Modeler Text Analytics Nós” na página 7 para obter mais informações.

Requisitos. O nó Análise de Ligação de Texto aceita dados de texto lidos em um campo usando qualquer um dos nós de origem padrão (nó Banco de Dados, nó Arquivo Simples, etc.) ou lidos em um campo listando caminhos para documentos externos gerados por um nó Arquivo Simples ou nó Web Feed.

Intensidades. O nó Análise de Ligação de Texto vai além da extração de conceitos básicos para fornecer informações sobre os relacionamentos *entre* conceitos, bem como opiniões relacionadas ou qualificadores que podem ser revelados nos dados.

Nó Análise de Ligação de Texto: Guia Campos

Utilize a aba Campos para especificar as configurações de campo para os dados a partir dos quais você estará extraindo conceitos. É possível configurar os parâmetros a seguir:

campo ID. Selecione o campo contendo o identificador para os registros de texto. Identificadores devem ser números inteiros. O campo de ID serve de índice para os registros de texto individuais. Use um campo de ID se o campo de texto representar o texto a ser minado.

Campo de texto. Selecione o campo contendo o texto a ser extraído. Esse campo depende da origem de dados.

Campo de linguagem. Selecione o campo que contém o identificador de linguagem ISO de duas letras. Se você não selecionar um campo, a linguagem de cada documento será assumida como a do gabarito fornecido.

Tipo de Documento. O tipo de documento especifica a estrutura do texto. Selecione um dos seguintes tipos:

- **Texto completo.** Use para a maioria dos documentos ou fontes de texto. O conjunto inteiro de texto é digitalizado para extração. Ao contrário de outras opções, não há configurações adicionais para essa opção.
- **Texto estruturado.** Use para formulários bibliográficos, patentes e quaisquer arquivos que contenham estruturas regulares que possam ser identificadas e analisadas. Esse tipo de documento é usado para ignorar todo ou parte do processo de extração. Ele permite definir separadores de termo, designar tipos e impor um valor de frequência mínimo. Se você selecionar esta opção, você deve clicar no botão **Configurações** e inserir separadores de texto na **Formatação de Texto Estruturado.** área da caixa de diálogo Configurações do Documento. Consulte o tópico "Configurações do documento para a guia Campos" na página 22 para obter mais informações.

Unidade textual. Selecione o modo de extração a seguir:

- **Modo de documento.** Use para documentos que são curtos e semanticamente homogêneos, como artigos de agências de notícias.
- **Modo de parágrafo.** Use para páginas da web e documentos sem tags. O processo de extração divide os documentos semanticamente, aproveitando a vantagem de características como tags internas e sintaxe. Se esse modo for selecionado, a escoragem será aplicada parágrafo por parágrafo. Portanto, por exemplo, a regra `apple & orange` será verdadeira se `apple` e `orange` estiverem localizadas no mesmo parágrafo.

  Nota: Devido à maneira como o texto é extraído de documentos PDF, o **Modo de Parágrafo** não funciona nesses documentos. Isso porque a extração suprime o marcador de retorno de linha.

Configurações do modo de parágrafo. Esta opção está disponível apenas se você configurar a opção de unidade textual para o **Modo de parágrafo**. Especifique os limites de caractere a serem usados em qualquer extração. O tamanho real é arredondado para cima ou para baixo para o período mais próximo. Para assegurar que as associações de palavras produzidas a partir do texto da coleção de documentos sejam representantes, evite especificar um tamanho de extração muito pequeno.

- **mínimo.** Especifique o número mínimo de caracteres a ser usado em qualquer extração.
- **Máximo.** Especifique o número máximo de caracteres a ser usado em qualquer extração.

Copiar recursos de. Ao minerar texto, a extração é baseada não apenas nas configurações na guia Especialista, mas também nos recursos linguísticos. Esses recursos servem de base para a forma de manipular e processar o texto durante a extração para obter os conceitos, tipos e padrões de TLA. É possível copiar recursos nesse nó de um modelo de recurso.

Um modelo de recurso é um conjunto predefinido de bibliotecas e recursos linguísticos e não linguísticos avançados que foram ajustados para um determinado domínio ou uso. Esses recursos servem de base para a forma de manipular e processar os dados durante a extração. Clique em **Carregar** e selecione o modelo do qual deseja copiar seus recursos.

Modelos são carregados quando você os seleciona, e não quando o fluxo é executado. No momento do carregamento, uma cópia dos recursos é armazenada no nó. Portanto, se alguma vez você quisesse usar um modelo atualizado, seria necessário recarregá-lo aqui. Consulte o tópico "Copiando recursos dos modelos e TAPs" na página 26 para obter informações adicionais.

Idioma do texto. Identifica o idioma do texto da mineração. Os recursos copiados no nó controlam as opções de idioma apresentadas. Selecione o idioma para o qual os recursos foram ajustados.

Nó Análise de Ligação de Texto: Guia Especialista

Neste nó, a extração dos resultados do padrão de análise de ligação de texto (TLA) é ativada automaticamente. A guia Especialista contém certos parâmetros adicionais que afetam como o texto

é extraído e manipulado. Os parâmetros nesta caixa de diálogo controlam o comportamento básico, bem como alguns comportamentos avançados, do processo de extração. Há também inúmeras opções e recursos linguísticos que também afetam os resultados da extração, os quais são controlados pelo modelo de recurso selecionado.

Limite a extração a conceitos com uma frequência global de pelo menos [n]. Especifica o número mínimo de vezes que uma palavra ou frase deve ocorrer no texto para que ela seja extraída. Dessa maneira, um valor de 5 limita a extração àquelas palavras ou frases que ocorrem pelo menos cinco vezes no conjunto inteiro de registros ou documentos.

Em alguns casos, a mudança desse limite pode fazer uma enorme diferença nos resultados da extração e, consequentemente, em suas categorias. Digamos que você esteja trabalhando com alguns dados de um restaurante e não deseja aumentar o limite para acima de 1 para essa opção. Nesse caso, você pode localizar *(1), pizza fina (2), pizza de espinafre (2) e pizza favorita (2)* e *pizza favorita (2)* em seus resultados de extração. Entretanto, se limitasse a extração a uma frequência global de 5 ou mais e refizesse a extração, você não obteria mais três desses conceitos. Em vez disso, você obteria *pizza (7)*, já que *pizza* é a forma mais simples e também essa palavra já existia como uma possível candidata. E dependendo do restante de seu texto, talvez você tenha de fato uma frequência superior a sete se ainda houver outras frases com pizza no texto. Além disso, se a *pizza de espinafre* já for um descritor de categoria, pode ser necessário incluir *pizza* como um descritor, em vez de capturar todos os registros. Por esse motivo, mude esse limite com cuidado sempre que as categorias já tiverem sido criadas.

Observe que essa é uma variável somente de extração; se seu modelo contiver termos (geralmente contém), e um termo para o modelo for localizado no texto, o termo será indexado, independentemente de sua frequência.

Por exemplo, suponhamos que você use um modelo de Recursos Básicos que inclua "los angeles" sob o tipo `<Location>` na biblioteca Core; se o seu documento contém Los Angeles apenas uma vez, então Los Angeles fará parte da lista de conceitos. Para evitar isso, você terá que configurar um filtro para exibir conceitos que ocorrem pelo menos o mesmo número de vezes que o valor inserido no campo **Limitar extração a conceitos com uma frequência global de pelo menos [n]**.

Acomodar erros de pontuação. Esta opção normaliza temporariamente o texto contendo erros de pontuação (por exemplo, uso incorreto) durante a extração para melhorar a extractabilidade de conceitos. Esta opção é muito útil quando o texto é curto e de qualidade ruim (como, por exemplo, em respostas de pesquisa sem estrutura, e-mail e dados CRM) ou quando o texto contém muitas abreviações.

Acomodar a ortografia para um comprimento mínimo de caracteres de palavra de [n] Esta opção aplica uma técnica de agrupamento fuzzy que ajuda a agrupar palavras comumente digitadas ou palavras bem escritas sob um só conceito. O algoritmo de agrupamento difuso temporariamente remove todas as vogais (exceto a primeira) e remove consoantes duplas/triplas das palavras extraída e, em seguida, as compara para ver se elas são a mesma, desta forma `modeling` e `modelling` seriam agrupadas. Entretanto, se cada termo for designado a um tipo diferente, excluindo o tipo `<Unknown>`, a técnica de agrupamento difuso não será aplicada.

Também é possível definir o número mínimo de caracteres *root* necessários antes que o agrupamento difuso seja usado. O número de caracteres raiz em um termo é calculado ao totalizar todos os caracteres e subtrair quaisquer caracteres que formam sufixos de inflexão e, no caso de termos com palavras compostas, determinadores e preposições. Por exemplo, o termo `exercises` seria contado como 8 caracteres raiz no formato “exercise,” já que a letra *s* no final da palavra é uma inflexão (forma plural). De maneira semelhante, `apple sauce` é contado como 10 caracteres raiz (“apple sauce”) e `manufacturing of cars` é contado como 16 caracteres raiz (“manufacturing car”). Este método de contagem é usado apenas para verificar se o agrupamento difuso deve ser aplicado, mas não influencia como as palavras são correspondidas.

Nota: Se você descobrir que certas palavras são mais tarde agrupadas incorretamente, você pode excluir pares de palavras desta técnica declarando-as explicitamente na seção **Agrupamento Fuzzy: Excepções** na guia Recursos Avançados. Veja o tópico “Agrupamento difuso” na página 197 para obter mais informações.

Extrair uniterms Esta opção extrai palavras únicas (uniterms) desde que a palavra não seja já parte de uma palavra composta e se for um substantivo ou uma parte não reconhecida da fala.

Extrair entidades não lingüísticas Esta opção extrai entidades não linguísticas, como números de telefone, números de previdência social, horários, datas, moedas, dígitos, porcentagens, endereços de e-mail e endereços HTTP. Você pode incluir e excluir determinados tipos de entidades não linguísticas na seção **Entidades Não Linguísticas: Configuração** da guia Recursos Avançados. Ao desativar qualquer entidades desnecessárias, o mecanismo de extração não desperdiçará tempo de processamento. Veja o tópico "Configuração" na página 201 para obter mais informações.

Algoritmos Uppercase Esta opção extrai termos simples e compostos que não estão nos dicionários embutidos desde que a primeira letra do termo esteja em maiúseada. Esta opção oferece uma boa maneira de extrair substantivos mais adequados.

Grupos parciais e completos de pessoa juntos quando possível Esta opção agrupa nomes que aparecem de forma diferente no texto em conjunto. Esse recurso é útil já que os nomes são frequentemente referidos em sua forma completa no início do texto e, então, apenas por uma versão mais curta. Esta opção tenta corresponder qualquer unitermo com o tipo<Unknown> com a última palavra de qualquer um dos termos compostos que é digitado como <Person>. Por exemplo, se *doe* for localizado e inicialmente digitado como <Unknown>, o mecanismo de extração verifica se quaisquer termos compostos no tipo <Person> incluem *doe* como a última palavras, tal como *john doe*. Esta opção não se aplica a nomes já que a maioria nunca é extraída como unitermos.

Permutação de palavra não função máxima Esta opção especifica o número máximo de palavras sem função que podem estar presentes ao aplicar a técnica de permutação. Essa técnica permutação agrupa frases semelhantes que diferem uma da outra apenas pelas palavras sem função contidas (por exemplo, de e o), independentemente da inflexão. Por exemplo, digamos que você configure este valor para no máximo duas palavras e ambos `company officials` e `officials of the company` foram extraídas. Nesse caso, ambos os termos extraídos seriam agrupados na lista de conceito final já que ambos os termos são considerados o mesmo quando `of the` é ignorado.

Usar derivação ao agrupar multitermos Ao processar Big Data, selecione esta opção para agrupar multitermos usando regras de derivação.

Saída do nó TLA

Após a execução do nó Análise de Ligação de Texto, os dados são reestruturados. É importante entender a maneira como a mineração de texto reestrutura seus dados. Se você desejar uma estrutura diferente para mineração de dados, é possível usar nós na paleta Operações de Campo para conseguir isso. Por exemplo, se você estivesse trabalhando com dados nos quais cada linha representasse um registro de texto, uma linha seria criada para cada padrão descoberto nos dados de texto de origem. Para cada linha na saída, há 15 campos:

- Seis campos (**Concept#**, tais como **Concept1**, **Concept2**, ..., e **Concept6**) representam quaisquer conceitos encontrados na correspondência de padrões.
- Seis campos (**Type#**, tais como **Type1**, **Type2**, ..., e **Type6**) representam o tipo para cada conceito.
- **Nome da Regra** representa o nome da regra de ligação de texto usada para corresponder ao texto e produzir a saída.
- Um campo usando o nome do campo de ID que você especificou no nó e representando o ID do documento ou registro como ele estava nos dados de entrada
- **Texto Correspondido** representa a parte dos dados de texto no documento ou registro original correspondente ao padrão TLA.

Nota: Quaisquer fluxos pré-existentes contendo um nó de Análise de Link de Texto de uma liberação anterior à 5.0 podem não ser totalmente executáveis até que você atualize os nós. Certas melhorias em versões mais recentes do IBM SPSS Modeler requerem que os nós mais antigos sejam substituídos pelas versões mais novas, que são mais implementáveis e poderosas.

Também é possível executar uma tradução automática de certos idiomas. Essa variável permite minar documentos em um idioma que talvez você não leia ou fale. Se desejar usar a variável de tradução, deve-se ter acesso ao Software como Serviço (SaaS) SDL.

Armazenando resultados do TLA em cache

Se você armazená-los em cache, os resultados da análise de ligação de texto ficarão no fluxo. Para evitar a repetição da extração de resultados de análise de link de texto cada vez que o fluxo é executado, selecione o nó de Análise de Link de Texto e dos menus escolha, **Editar> Nó> Cache> Ativar**. Na próxima vez que o fluxo for executado, a saída será armazenada em cache no nó. O ícone de nó exibe um pequeno gráfico de "documento" que muda de branco para verde quando o cache é preenchido. O cache é preservado enquanto durar a sessão. Para preservar o cache por mais um dia (após o fluxo ser fechado e reaberto), selecione o nó e a partir dos menus escolha, **Editar> Nó> Cache> Salvar Cache**. Na próxima vez que você abrir o fluxo, é possível recarregar o cache salvo em vez de executar a conversão novamente.

Alternativamente, é possível salvar ou ativar um cache de nó clicando no nó com o botão direito e escolhendo **Cache** no menu de contexto.

Usando o nó Análise de Ligação de Texto em um fluxo

O nó Análise de Ligação de Texto é usado para acessar dados e extrair conceitos em um fluxo. É possível usar qualquer nó de origem para acessar dados.

Exemplo: nó Arquivo de Estatísticas com o nó Análise de Ligação de Texto

O exemplo a seguir mostra como usar o nó Análise de Ligação de Texto.

Figura 15. Exemplo: nó Arquivo de Estatísticas com o nó Análise de Ligação de Texto

1. **Nó Arquivo de Estatísticas (guia Dados).** Primeiro, incluímos esse nó no fluxo para especificar onde o texto será armazenado.

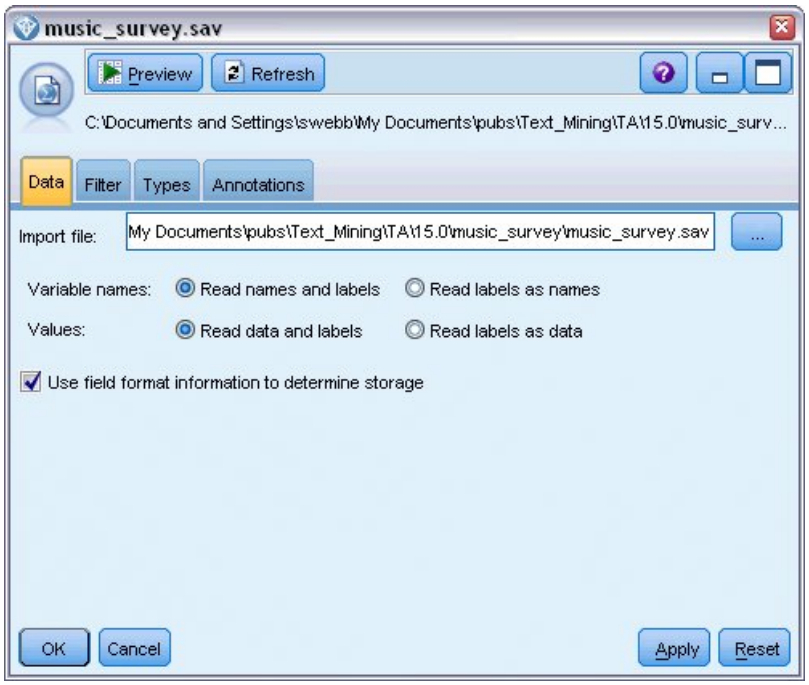


Figura 16. Caixa de diálogo do nó Arquivo de Estatísticas: guia Dados

2. **Nó Análise de Ligação de Texto (guia Campos).** Em seguida, anexamos esse nó ao fluxo para extrair conceitos para visualização ou modelagem de recebimento de dados. Especificamos o nome do campo de ID e do campo de texto contendo os dados, bem como outras configurações.

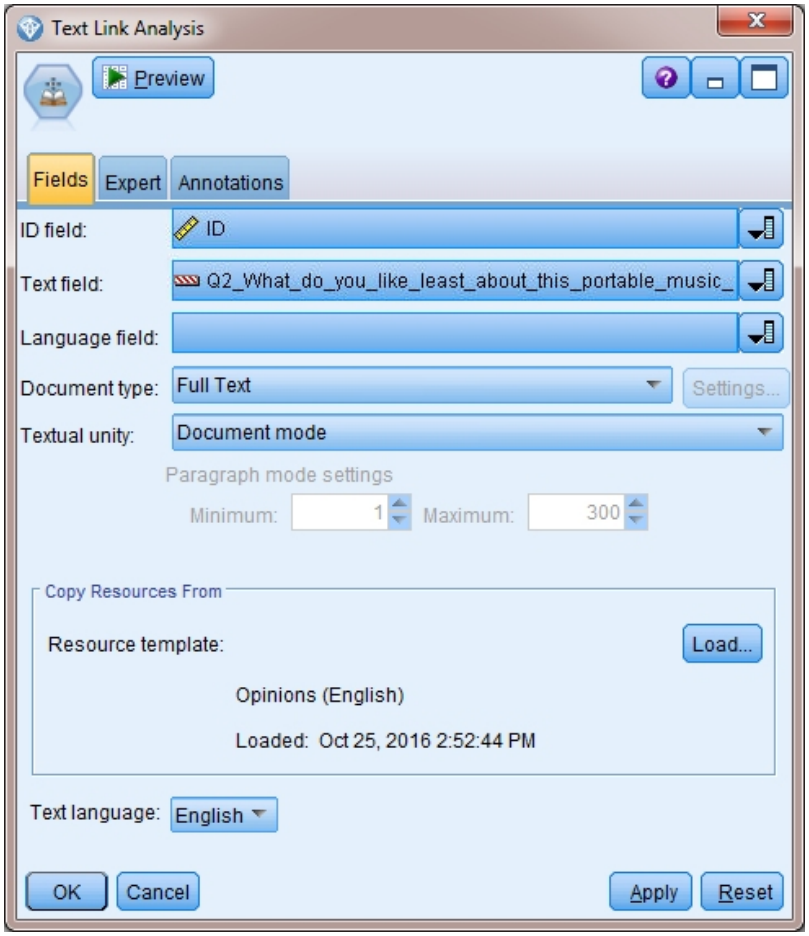
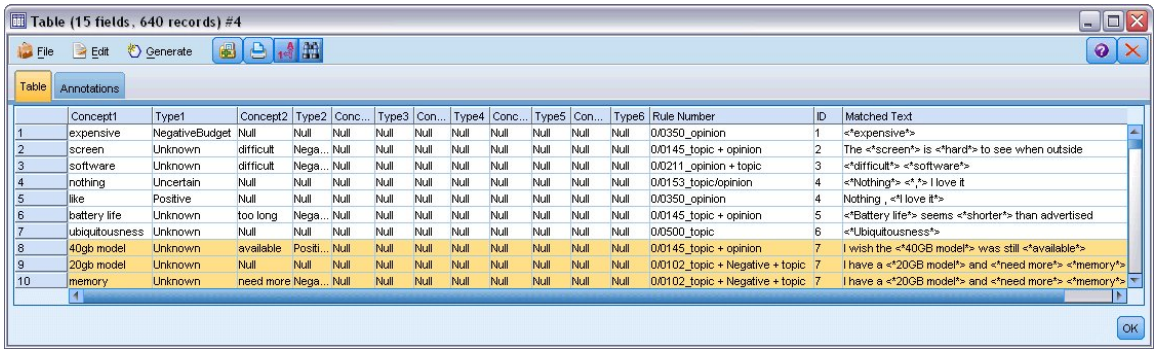


Figura 17. Caixa de diálogo do nó Análise de Ligação de Texto: guia Campos

3. **Nó da tabela.** Por fim, anexamos um nó Tabela para visualizar os conceitos que foram extraídos de nossos documentos de texto. Na saída de tabela mostrada, é possível ver os resultados do padrão de TLA localizados nos dados após esse fluxo ter sido executado com um nó Análise de Ligação de Texto. Alguns resultados mostram que apenas um conceito/tipo era correspondente. Em outros, os resultados são mais complexos e contêm vários tipos e conceitos. Além disso, como resultado da execução de dados por meio do nó Análise de Ligação de Texto e da extração de conceitos, vários aspectos dos dados mudam. Os dados originais em nosso exemplo continham 8 campos e 405 registros. Após a execução do nó Análise de Ligação de Texto, agora há 15 campos e 640 registros. Agora há uma linha para cada resultado do padrão de TLA localizado. Por exemplo, `ID 7` se tornou três linhas a partir do original, pois três resultados do padrão do TLA foram extraídos. É possível usar um nó Mesclagem se você desejar mesclar esses dados de saída de volta com seus dados originais.



	Concept1	Type1	Concept2	Type2	Conc...	Type3	Con...	Type4	Conc...	Type5	Con...	Type6	Rule Number	ID	Matched Text
1	expensive	NegativeBudget	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	1	<*expensive*>
2	screen	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	2	The <*screen*> is <*hard*> to see when outside
3	software	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0211_opinion + topic	3	<*difficult*> <*software*>
4	nothing	Uncertain	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0153_topic/opinion	4	<*Nothing*> <*,*> I love it
5	like	Positive	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	4	Nothing , <*I love it*>
6	battery life	Unknown	too long	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	5	<*Battery life*> seems <*shorter*> than advertised
7	ubiquitousness	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0500_topic	6	<*Ubiquitousness*>
8	40gb model	Unknown	available	Positi...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	7	I wish the <*40GB model*> was still <*available*>
9	20gb model	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>
10	memory	Unknown	need more	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>

Figura 18. nó Saída de tabela

Capítulo 5. Navegando no texto de origem externa

Nó Visualizador de Arquivo

Quando você está minerando uma coleção de documentos, você pode especificar os nomes de caminhos completos de arquivos diretamente em seus nós de modelagem de Mineração de Texto. No entanto, ao emitir para um nó Tabela, você verá somente o nome do caminho completo de um documento, e não o texto dentro dele. O nó Visualizador de Arquivo pode ser usado como um analógico do nó Tabela, e isso permite que você acesse o texto real dentro de cada um dos documentos sem precisar mesclá-los em um único arquivo.

O nó Visualizador de Arquivo pode ajudá-lo a entender melhor os resultados da extração de texto fornecendo acesso ao texto de origem, ou não traduzido, do qual conceitos foram extraídos, já que, de outra forma, ele seria inacessível no fluxo. Esse nó é incluído no fluxo após um nó Lista de Arquivos para se obter uma lista de ligações para todos os arquivos.

O resultado desse nó é uma janela mostrando todos os elementos do documento que foram lidos e usados para extrair conceitos. Nessa janela, é possível clicar em um ícone de barra de ferramentas para ativar o relatório em um navegador externo listando nomes de documentos como hyperlinks. É possível clicar em uma ligação para abrir o documento correspondente na coleção. Consulte o tópico "Usando o nó Visualizador de Arquivo" na página 51 para obter informações adicionais.

É possível localizar esse nó na guia IBM SPSS Modeler Text Analytics da paleta de nós na parte inferior da janela IBM SPSS Modeler. Consulte o tópico "IBM SPSS Modeler Text Analytics Nós" na página 7 para obter mais informações.

Nota: Quando você estiver trabalhando no modo cliente-servidor e os nós do File Viewer fizerem parte do fluxo, as coleções de documentos deverão ser armazenadas em um diretório do servidor da Web no servidor Como o nó de saída Mineração de Texto produz uma lista de documentos armazenados no diretório do servidor da web, as configurações de segurança do servidor da web gerenciam as permissões para esses documentos.

Configurações do nó Visualizador de Arquivo

É possível especificar as configurações a seguir para o nó Visualizador de Arquivo.

Campo Documento. Selecione o campo a partir de seus dados que contenha o nome completo e o caminho dos documentos a serem exibidos.

Título para página HTML gerada. Crie um título para aparecer na parte superior da página que contém a lista de documentos.

Usando o nó Visualizador de Arquivo

O exemplo a seguir mostra como usar o nó Visualizador de Arquivo.

Exemplo: nó Lista de Arquivos e nó Visualizador de Arquivo

Figura 19. Fluxo ilustrando o uso de um nó Visualizador de Arquivo

1. **Nó Lista de Arquivos (guia Configurações).** Primeiro, incluímos esse nó para especificar onde os documentos estão localizados.

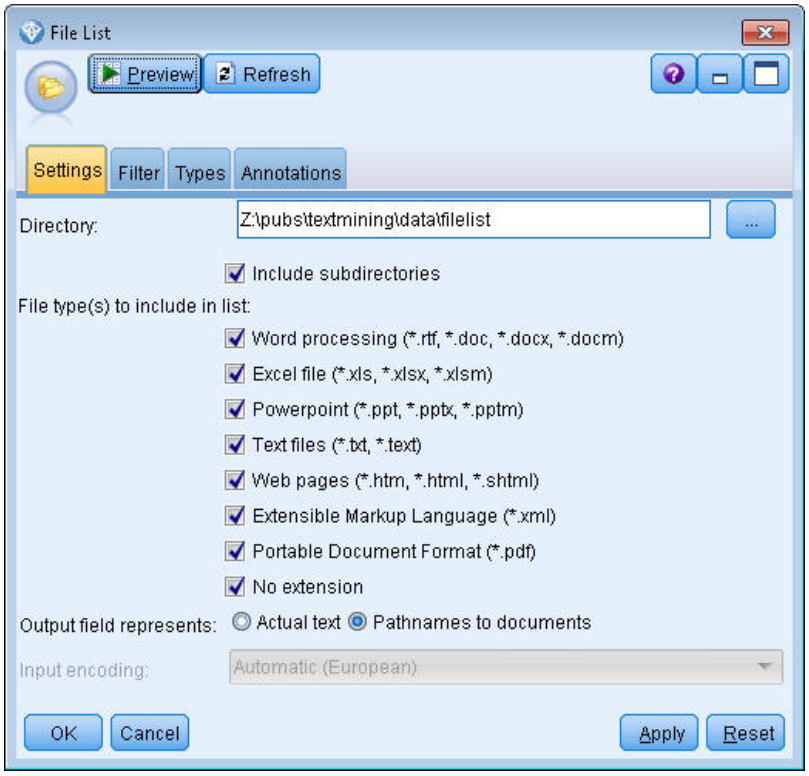


Figura 20. Caixa de diálogo do nó Lista de Arquivos: guia Configurações

2. **Nó Visualizador de Arquivo (guia Configurações).** Em seguida, anexamos o nó Visualizador de Arquivo para produzir uma lista HTML de documentos.

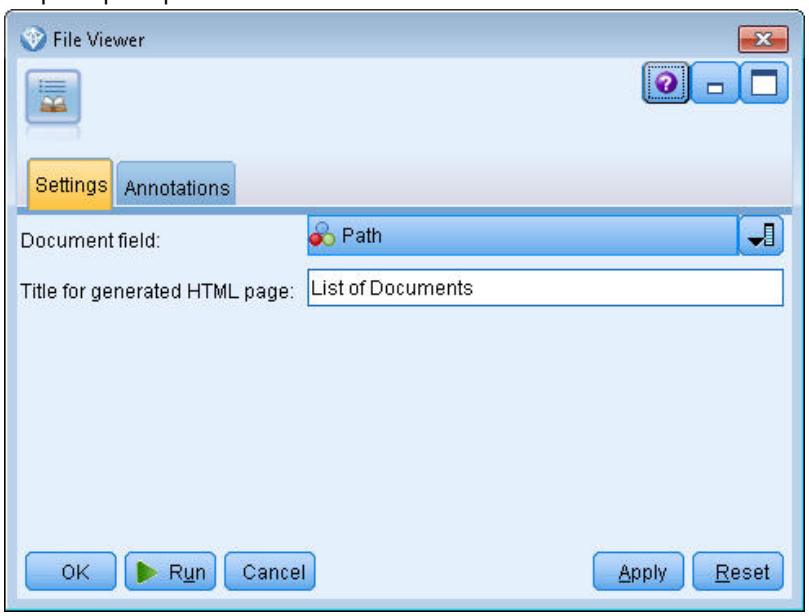


Figura 21. Caixa de diálogo do nó Visualizador de Arquivo: guia Configurações

3. **Diálogo Saída do Visualizador de Arquivo.** Em seguida, executamos o fluxo que emite a lista de documentos em uma nova janela.

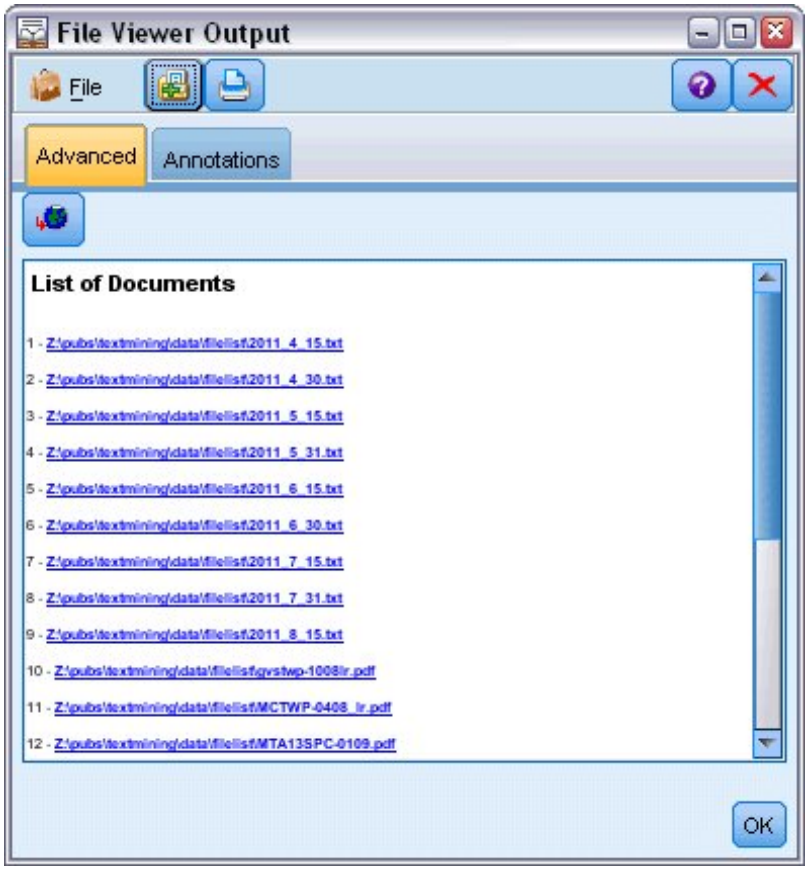


Figura 22. Saída do Visualizador de Arquivo

4. Para ver os documentos, clicamos no botão da barra de ferramentas mostrando um globo com uma seta vermelha. Isso abriu uma lista de hiperlinks de documentos em nosso navegador.

Capítulo 6. Propriedades do nó para script

O IBM SPSS Modeler possui um idioma de script para permitir executar fluxos a partir da linha de comandos. Aqui, é possível aprender sobre propriedades do nó que são específicas para cada um dos nós fornecidos com o IBM SPSS Modeler Text Analytics. Para obter mais informações sobre o conjunto padrão de nós fornecidos com o IBM SPSS Modeler, consulte o Guia de Script e Automação.

Nó da lista de arquivos: filelistnode

É possível usar as propriedades na tabela a seguir para script. O nó em si é denominado `filelistnode`.

Tabela 7. Propriedade de script do nó Lista de Arquivos

Propriedades de script	**Tipo de dados**
`path`	*sequência*
`recurse`	*sinalização*
`word_processing`	*sinalização*
`excel_file`	*sinalização*
`powerpoint_file`	*sinalização*
`text_file`	*sinalização*
`web_page`	*sinalização*
`xml_file`	*sinalização*
`pdf_file`	*sinalização*
`no_extension`	*sinalização*

Nota: o parâmetro 'Create list' não está mais disponível e quaisquer scripts contendo tal opção serão automaticamente convertidos em uma saída 'Files'.

Nó Web Feed: webfeednode

É possível usar as propriedades na tabela a seguir para script. O nó em si é chamado `webfeednode`.

Tabela 8. Propriedades de script do nó Web Feed

Propriedades de script	**Tipo de dados**	**Descrição da propriedade**
`urls`	*string1 string2 ... stringn*	Cada URL é especificada na estrutura de lista. A lista de URLs é separada por “\n”
`recent_entries`	*sinalização*	
`limit_entries`	*Número inteiro*	Número das entradas mais recentes para ler por URL.
`use_previous`	*sinalização*	Para salvar e reutilizar o cache Web Feed.
`use_previous_label`	*sequência*	Nome do cache da web salvo.
`start_record`	*sequência*	Tag de início não RSS.

Tabela 8. Propriedades de script do nó Web Feed (continuação)		
Propriedades de script	**Tipo de dados**	**Descrição da propriedade**
`url` *n* `.title`	*sequência*	Para cada URL na lista, deve-se definir uma também. A primeira será `url1.title`, em que o número corresponde a seu ranqueamento na lista de URLs. Essa é a tag de início contendo o título do conteúdo.
`url` *n* `.short_description`	*sequência*	Mesmo que para `url` *n* `.title`.
`url` *n* `.description`	*sequência*	Mesmo que para `url` *n* `.title`.
`url` *n* `.authors`	*sequência*	Mesmo que para `url` *n* `.title`.
`url` *n* `.contributors`	*sequência*	Mesmo que para `url` *n* `.title`.
`url` *n* `.published_date`	*sequência*	Mesmo que para `url` *n* `.title`.
`url` *n* `.modified_date`	*sequência*	Mesmo que para `url` *n* `.title`.

`html_alg`	`None`
	
	`HTMLCleaner`
		Método de filtragem de conteúdo.

`discard_lines`	*sinalização*	Descartar linhas curtas. Usado com `min_words`
`min_words`	*Número inteiro*	Número mínimo de palavras.
`discard_words`	*sinalização*	Descartar linhas curtas. Usado com `min_avg_len`
`min_avg_len`	*Número inteiro*	
`discard_scw`	*sinalização*	Descartar linhas com muitas palavras de caractere único. Usado com `max_scw`
`max_scw`	*Número inteiro*	Porcentagem de proporção máxima de 0-100 de palavras de caractere único em uma linha
`discard_tags`	*sinalização*	Descartar linhas contendo determinadas tags.
`tags`	*sequência*	Caracteres especiais devem ser escapados com um caractere de barra invertida \.
`discard_spec_words`	*sinalização*	Descartar linhas contendo sequências de caracteres específicas.
`words`	*sequência*	Caracteres especiais devem ser escapados com um caractere de barra invertida \.

Nó da Linguagem: Languageidentifier

É possível usar as propriedades na tabela a seguir para script. O próprio nó é chamado `languageidentifier`.

Tabela 9. Propriedades de scripting do nó da linguagem		
Propriedades de script	**Tipo de dados**	**Descrição da propriedade**
`text`	*campo*	
`language_field_name`	*sequência*	O nome de campo que é gerado como saída.

Tabela 9. Propriedades de scripting do nó da linguagem (continuação)		
Propriedades de script	**Tipo de dados**	**Descrição da propriedade**

`unidentified_language_value`	`Undefined`
	
	`Supported`
	
	`Custom`
		Valor padrão a ser usado quando o idioma não pode ser identificado.

`unidentified_language_supported`	`en`
	
	`de`
	
	`es`
	
	`fr`
	
	`it`
	
	`ja`
	
	`nl`
	
	`pt`
		Código Iso. Só disponível se `unidentified_language_value` for `Supported`.

`unidentified_language_custom`	*sequência*	Só disponível se `unidentified_language_value` for `Custom`.

Nó Mineração de Texto: TextMiningWorkbench

É possível usar os parâmetros a seguir para definir ou atualizar um nó por meio de scripts. O nó em si é chamado `TextMiningWorkbench`.

Importante: Não é possível especificar um modelo de recurso diferente via script. Se você julgar que precisa de um modelo, deve-se selecioná-lo na caixa de diálogo do nó.

Tabela 10. Propriedades de script do nó de modelagem de Mineração de Texto		
Propriedades de script	**Tipo de dados**	**Descrição da propriedade**
`text`	*campo*	

`method`	`ReadText`
	
	`ReadPath`	

`docType`	*Número inteiro*	Com possíveis valores (0,1, 2) em que 0 = `Full Text`, 1 = `Structured Text`, e 2 = `XML`

Tabela 10. Propriedades de script do nó de modelagem de Mineração de Texto (continuação)

Propriedades de script	Tipo de dados	Descrição da propriedade
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Observe que os valores com caracteres especiais, como "UTF-8", devem ser colocados entre aspas para evitar confusão com um operador matemático.
unity	*Número inteiro*	Com possíveis valores (0, 1) em que 0 = Paragraph e 1 = Document
para_min	*Número inteiro*	
para_max	*Número inteiro*	
mtag	*sequência*	Contém todas as configurações de mtag (da caixa de diálogo Configurações para arquivos XML)
mclef	*sequência*	Contém todas as configurações de mclef (da caixa de diálogo Configurações para arquivos Texto Estruturado)
partition	*campo*	
custom_field	*sinalização*	Indica se um campo de partição será ou não especificado.
use_model_name	*sinalização*	
model_name	*sequência*	
use_partitioned_data	*sinalização*	Se um campo de partição for definido, apenas os dados de treinamento serão usados para construção de modelo.
model_output_type	Interactive Model	Interactive resulta em um modelo de categoria. Model resulta em um modelo de conceito.
use_interactive_info	*sinalização*	Para construir interativamente em uma sessão de ambiente de trabalho apenas.

Tabela 10. Propriedades de script do nó de modelagem de Mineração de Texto (continuação)		
Propriedades de script	**Tipo de dados**	**Descrição da propriedade**
reuse_extraction_results	*sinalização*	Para construir interativamente em uma sessão de ambiente de trabalho apenas.
interactive_view	Categories TLA Clusters	Para construir interativamente em uma sessão de ambiente de trabalho apenas.
extract_top	*Número inteiro*	Este parâmetro é usado quando model_type = Concept
use_check_top	*sinalização*	
check_top	*Número inteiro*	
use_uncheck_top	*sinalização*	
uncheck_top	*Número inteiro*	
language	de en es fr it ja nl pt	
frequency_limit	*Número inteiro*	Descontinuado na 14.0.
concept_count_limit	*Número inteiro*	Limitar extração a conceitos com uma frequência global de pelo menos este valor.
fix_punctuation	*sinalização*	
fix_spelling	*sinalização*	
spelling_limit	*Número inteiro*	
extract_uniterm	*sinalização*	
extract_nonlinguistic	*sinalização*	
upper_case	*sinalização*	
group_names	*sinalização*	
permutation	*Número inteiro*	Permutação máxima de palavra sem função (o padrão é 3).

Nugget do modelo de mineração de texto: TMWBModelApplier

É possível usar as propriedades na tabela a seguir para script. O nugget é si é chamado `TMWBModelApplier`.

Tabela 11. Propriedades do nugget do modelo de Mineração de Texto

Propriedades de script	**Tipo de dados**	**Descrição da propriedade**

`scoring_mode`	`Fields`
	
	`Records`	

`field_values`	`Flags`
	
	`Counts`
		Esta opção não está disponível no nugget do modelo Categoria. Para `Flags`, configure como TRUE ou FALSE

`true_value`	*sequência*	Com `Flags`, defina o valor para true.
`false_value`	*sequência*	Com `Flags`, defina o valor para false.
`extension_concept`	*sequência*	Especifique uma extensão para o nome do campo. Nomes do campo são gerados usando o nome do conceito mais esta extensão. Especifique onde colocar esta extensão usando o valor `add_as`.
`extension_category`	*sequência*	Extensão do nome do campo. É possível escolher especificar um prefixo/sufixo de extensão para o nome do campo ou é possível escolher usar os códigos de categoria. Nomes do campo são gerados usando o nome da categoria mais esta extensão. Especifique onde colocar esta extensão usando o valor `add_as`.

`add_as`	`Suffix`
	
	`Prefix`	

`fix_punctuation`	*sinalização*	

Tabela 11. Propriedades do nugget do modelo de Mineração de Texto (continuação)

Propriedades de script	Tipo de dados	Descrição da propriedade

`excluded_subcategories_descriptors`	`RollUpToParent`
	
	`Ignore`
		Apenas para modelos de categoria. Se uma subcategoria estiver desmarcada. Essa opção permite especificar como os descritores pertencentes a subcategorias que não foram selecionadas para escoragem serão manipulados. Há duas opções.
		• `Ignore`. A opção Exclude seus descritores completamente de pontuação fará com que os descritores de subcategorias que não possuem marcas de verificação (não selecionados) sejam ignorados e não utilizados durante a pontuação.
		• `RollUpToParent`. A opção Agregar descritores com aqueles em categoria pai fará com que os descritores de subcategorias que não possuam marcas de verificação (não selecionados) sejam utilizados como descritores para a categoria pai (a categoria acima desta subcategoria). Se vários níveis de subcategorias estiverem desmarcados, os descritores serão revertidos sob a primeira categoria pai disponível

`check_model`	*sinalização*	Descontinuado na versão 14
`text`	*campo*	

`method`	`ReadText`
	
	`ReadPath`	

`docType`	*Número inteiro*	Com possíveis valores (0,1, 2) em que 0 = `Full Text`, 1 = `Structured Text`, e 2 = `XML`

Tabela 11. Propriedades do nugget do modelo de Mineração de Texto (continuação)		
Propriedades de script	**Tipo de dados**	**Descrição da propriedade**

`encoding`	`Automatic`
	
	`"UTF-8"`
	
	`"UTF-16"`
	
	`"ISO-8859-1"`
	
	`"US-ASCII"`
	
	`"CP850"`
	
	`"EUC-JP"`
	
	`"SHIFT-JIS"`
	
	`"ISO2022-JP"`
		Observe que os valores com caracteres especiais, como `"UTF-8"`, devem ser colocados entre aspas para evitar confusão com um operador matemático.

`language`	`de`
	
	`en`
	
	`es`
	
	`fr`
	
	`it`
	
	`ja`
	
	`nl`
	
	`pt`	

Nó Análise de Ligação de Texto: textlinkanalysis

É possível usar os parâmetros na tabela a seguir para definir ou atualizar um nó por meio de script. O nó em si é chamado `textlinkanalysis`.

Importante: Não é possível especificar um modelo de recurso via script. Para selecionar um modelo, você deve fazer isso de dentro da caixa de diálogo do nó.

Tabela 12. Propriedades de script do nó Análise de Ligação de Texto (TLA)		
Propriedades de script	**Tipo de dados**	**Descrição da propriedade**
`id_field`	*campo*	
`text`	*campo*	

Tabela 12. Propriedades de script do nó Análise de Ligação de Texto (TLA) (continuação)		
Propriedades de script	**Tipo de dados**	**Descrição da propriedade**
method	ReadText ReadPath	
docType	*Número inteiro*	Com possíveis valores (0,1, 2) em que 0 = Full Text, 1 = Structured Text, e 2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Observe que os valores com caracteres especiais, como "UTF-8", devem ser colocados entre aspas para evitar confusão com um operador matemático.
unity	*Número inteiro*	Com possíveis valores (0, 1) em que 0 = Paragraph e 1 = Document
para_min	*Número inteiro*	
para_max	*Número inteiro*	
mtag	*sequência*	Contém todas as configurações de mtag (da caixa de diálogo Configurações para arquivos XML)
mclef	*sequência*	Contém todas as configurações de mclef (da caixa de diálogo Configurações para arquivos Texto Estruturado)

Tabela 12. Propriedades de script do nó Análise de Ligação de Texto (TLA) (continuação)

Propriedades de script	Tipo de dados	Descrição da propriedade
language	de en es fr it ja nl pt	
concept_count_limit	*Número inteiro*	Limitar extração a conceitos com uma frequência global de pelo menos este valor.
fix_punctuation	*sinalização*	
fix_spelling	*sinalização*	
spelling_limit	*Número inteiro*	
extract_uniterm	*sinalização*	
extract_nonlinguistic	*sinalização*	
upper_case	*sinalização*	
group_names	*sinalização*	
permutation	*Número inteiro*	Permutação máxima de palavra sem função (o padrão é 3).

Capítulo 7. Modo de ambiente de trabalho interativo

A partir de um nó de modelagem de mineração de texto, é possível escolher ativar uma sessão de ambiente de trabalho interativa durante uma execução de fluxo. Nesse ambiente de trabalho, é possível extrair conceitos chave de seus dados de texto, construir categorias e explorar padrões e clusters de análise de ligação de texto, além de gerar modelos de categoria. Neste capítulo , discutiremos a interface de ambiente de trabalho a partir de uma perspectiva de alto nível, junto com os principais elementos com os quais você irá trabalhar, incluindo:

- **Resultados de extração.** Após uma extração ser executada, esses são as palavras-chave e frases identificadas e extraídas de seus dados de texto, também chamadas de *conceitos*. Esses conceitos são agrupados em *tipos*. Usando esses conceitos e tipos, é possível explorar seus dados, bem como criar suas categorias. Estes são gerenciados na visualização **Categorias e conceitos**.
- **Categorias.** Usando descritores (como resultados da extração, padrões e regras) como uma definição, é possível criar manualmente ou automaticamente um conjunto de categorias ao qual documentos e registros são designados, dependendo se eles contiverem ou não uma parte da definição de categoria. Estes são gerenciados na visualização **Categorias e conceitos**.
- **Clusters.** *Clusters* são um agrupamento de conceitos entre os quais links foram descobertos que indicam um relacionamento entre eles. Os conceitos são agrupados usando um algoritmo complexo que usa, entre outros fatores, a frequência com que dois conceitos aparecem juntos em comparação com a frequência com que eles aparecem separadamente. Estes são gerenciados na visualização **Clusters**. Também é possível incluir os conceitos que compõem um cluster nas categorias.
- **Padrões de análise de link de texto.** Se você tiver regras de padrão de análise de ligação de texto (TLA) em seus recursos linguísticos ou estiver usando um modelo de recurso que já tenha algumas regras de TLA, é possível extrair padrões de seus dados de texto. Esses padrões podem ajudar a descobrir relacionamentos interessantes entre conceitos em seus dados. Também é possível usar esses padrões como descritores em suas categorias. Estes são gerenciados na visualização **Análise de link de texto**.
- **Recursos linguísticos.** O processo de extração conta com um conjunto de parâmetros e definições linguísticas para governar como o texto é extraído e manipulado. Eles são gerenciados na forma de modelos e bibliotecas na visualização do **Editor de recursos**.

Questões Potenciais do Ambiente de Trabalho Interativo

- Várias Sessões do Ambiente Interativo podem causar um comportamento lento. SPSS Análise de Texto do Modeler e SPSS Modeler compartilham um mecanismo de tempo de execução Java comum quando uma sessão interativa de ambiente de trabalho é lançada. Dependendo do número de sessões do Workbench Interativo que você chama durante uma sessão SPSS Modeler , a memória do sistema pode fazer com que o aplicativo se torne lento, mesmo se abrindo e fechando a mesma sessão. Este efeito pode ser especialmente pronunciado se você estiver trabalhando com dados grandes ou ter uma máquina com menos do que a configuração de RAM recomendada de 4GB. Se você notar que sua máquina está lenta para responder, é recomendável que você salve todo o seu trabalho, desligue SPSS Modelere relembre o aplicativo. Executar SPSS Análise de Texto do Modeler em uma máquina com menos do que a memória recomendada, particularmente ao trabalhar com grandes conjuntos de dados ou por períodos prolongados de tempo, pode fazer com que o Java fique sem memória e encerrado. É fortemente sugerida você se atualizar para a configuração de memória recomendada ou maior (ou use SPSS Análise de Texto do Modeler Server) se você trabalha com dados grandes.
- SPSS Modeler O cliente pode ficar sem memória após várias SPSS Análise de Texto do Modeler sessões do Workbench Interativo são executadas sem recomeçar o aplicativo. Monitore o uso de memória na linha de status e, se estiver rodando baixo, feche e reabra o Cliente SPSS Modeler .

A visualização Categorias e Conceitos

A interface de aplicativo é composta por diversas visualizações. A visualização Categorias e Conceitos é a janela onde é possível criar e explorar categorias, bem como explorar e ajustar os resultados da extração. *Categorias* referem-se a um grupo de ideias e padrões intimamente relacionados aos quais documentos e registros são designados por meio de um processo de pontuação. Já *conceitos* referem-se ao nível mais básico de resultados de extração disponíveis para serem usados como blocos de construção, chamados descritores, para suas categorias.

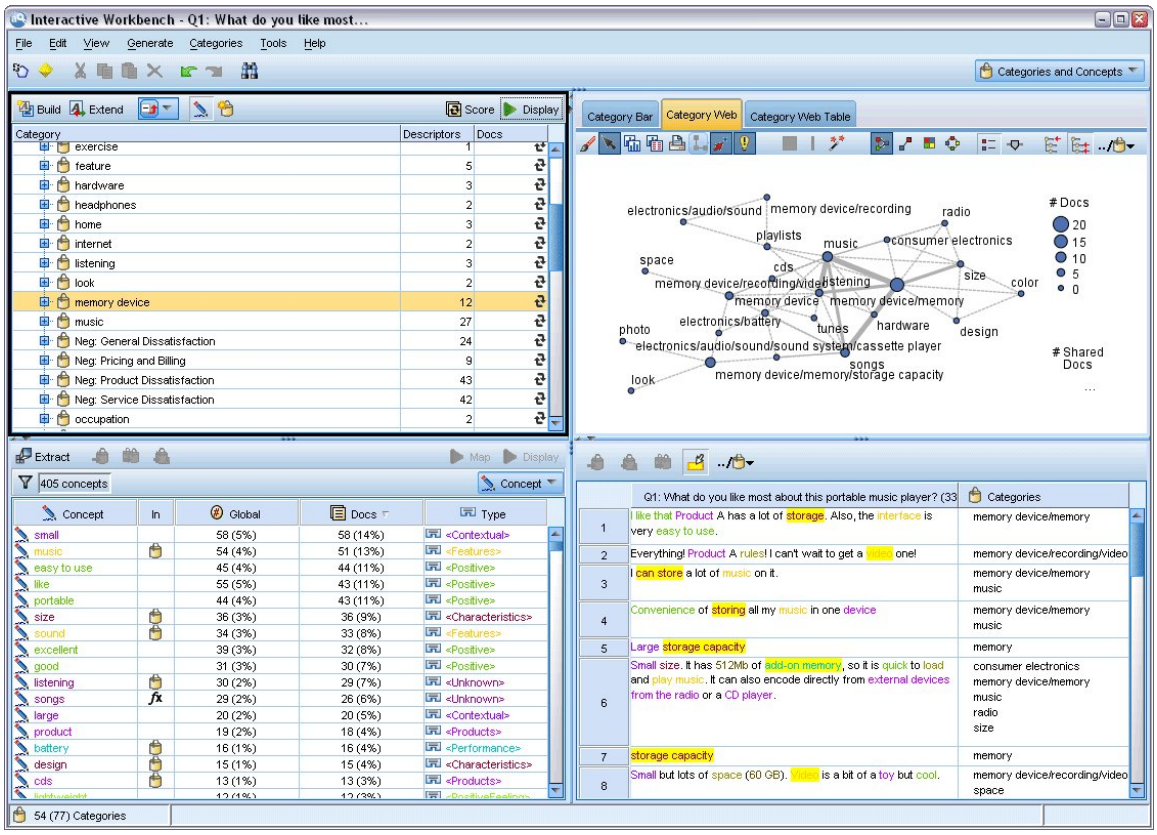


Figura 23. Visualização Categorias e Conceitos

A visualização Categorias e Conceitos é organizada em quatro áreas de janela, sendo que cada uma pode ser oculta ou mostrada selecionando seu nome no menu Visualizar. Consulte o tópico Capítulo 9, “Categorizando dados de texto”, na página 91 para obter informações adicionais.

Área de janela Categorias

Localizada no canto superior esquerdo, esta área apresenta uma tabela na qual é possível gerenciar quaisquer categorias que você construir. Após você extrair os conceitos e os tipos de seus dados de texto, é possível começar a construir categorias usando técnicas como redes semânticas e inclusão de conceito ou criando-as manualmente. Se você der um clique duplo nome de uma categoria, a caixa de diálogo Definições de Categoria será aberta e exibirá todos os descritores que compõem sua definição, como conceitos, tipos e regras. Consulte o tópico Capítulo 9, “Categorizando dados de texto”, na página 91 para obter informações adicionais. Nem todas as técnicas automáticas estão disponíveis para todos os idiomas.

Quando você seleciona uma linha na área de janela, é possível exibir informações sobre documentos/registros correspondentes ou descritores nas áreas de janela Dados e Visualização.

Área de janela Resultados da Extração

Localizada no canto inferior esquerdo, esta área apresenta os resultados da extração. Quando você executa uma extração, o mecanismo de extração lê os dados de texto, identifica os conceitos relevantes e designa um tipo a cada um. *Conceitos* são palavras ou frases extraídas de seus dados de texto. *Tipos* são agrupamentos semânticos de conceitos armazenados na forma de dicionários de tipos. Quando a extração é concluída, conceitos e tipos aparecem com codificação de cor na área de janela Resultados de Extração. Consulte o tópico "Resultados da extração: conceitos e tipos" na página 79 para obter mais informações.

É possível ver o conjunto de termos subjacentes para um conceito ao passar o mouse sobre o nome do conceito. Fazer isso exibirá uma dica de ferramenta mostrando o nome do conceito e várias linhas de termos que são agrupados sob tal conceito. Esses termos subjacentes incluem os sinônimos definidos nos recursos linguísticos (independentemente se eles estavam localizados no texto ou não), bem como quaisquer termos plurais/singulares extraídos, termos permutados, termos de agrupamento difuso e assim por diante. É possível copiar esses termos ou ver o conjunto completo de termos subjacentes ao clicar com o botão direito no nome do conceito e escolher a opção de menu de contexto.

A mineração de texto é um processo interativo no qual os resultados da extração são revisados de acordo com o contexto dos dados de texto, ajustados para produzir novos resultados e depois reavaliados. Os resultados da extração podem ser refinados modificando recursos linguísticos. Esse ajuste pode ser feito em parte diretamente a partir da área de janela Resultados da Extração ou Dados, mas também diretamente na visualização Editor de Recurso. Consulte o tópico "A visualização do Editor de recursos" na página 72 para obter informações adicionais.

Nota: Se houver mais resultados que possam caber na pane visível, você pode usar os controles na parte inferior da pane para mover para frente e para trás através dos resultados, ou inserir um número de página para ir até.

Área de janela Visualização

Localizada no canto superior direito, esta área apresenta diversas perspectivas nos compartilhamentos na categorização documento/registro. Cada gráfico ou diagrama fornece informações semelhantes, mas as apresenta de uma maneira diferente ou com um nível de detalhes diferente. Esses gráficos e diagramas podem ser usados para analisar seus resultados de categorização e ajudar no ajuste de categorias e em relatórios. Por exemplo, em um gráfico, você pode descobrir categorias que são muito semelhantes (por exemplo, que compartilham mais de 75% de seus registros) ou muito distintas. Os conteúdos em um gráfico ou diagrama correspondem à seleção nas outras áreas de janela. Consulte o tópico "Gráficos e diagramas de categoria" na página 155 para obter informações adicionais.

Área de janela Dados

A área de janela Dados está localizada no canto inferior direito. Essa área de janela apresenta uma tabela contendo os documentos ou registros correspondentes a uma seleção em outra área da visualização. Dependendo do que for selecionado, somente o texto correspondente aparecerá na área de janela Dados. Após você fazer uma seleção, clique no botão **Exibir** para preencher a área de janela Dados com o texto correspondente.

Se você tiver uma seleção em outra área de janela, os documentos ou registros correspondentes mostrarão os conceitos destacados em cores para ajudá-lo a identificá-los facilmente no texto. Também é possível passar o mouse sobre os itens codificados por cores para exibir uma dica de ferramenta mostrando o nome do conceito sob o qual eles foram extraído e o tipo ao qual eles foram designados. Consulte o tópico "A área de janela Dados" na página 99 para obter informações adicionais.

Procurando e localizando na visualização Categorias e Conceitos

Em alguns casos, talvez seja necessário localizar informações rapidamente em uma determinada seção. Usando a barra de ferramentas Localizar, é possível inserir a sequência de caracteres que você deseja procurar e definir outros critérios de procura, como distinção entre maiúsculas e minúsculas ou direção da procura. Então é possível escolher a área de janela onde deseja procurar.

Para usar a variável Localizar

1. Na visualização de Categorias E Conceitos, escolha **Editar> Localizar** a partir dos menus. A barra de ferramentas Localizar aparece acima da área de janela Categorias e da área de janela Visualização.
2. Insira a sequência de palavras que deseja procurar na caixa de texto. É possível usar os botões da barra de ferramentas para controlar a distinção entre maiúsculas e minúsculas, correspondência parcial e direção da procura.
3. Na barra de ferramentas, clique no nome da área de janela em que deseja procurar. Se uma correspondência for localizada, o texto será destacado na janela.
4. Para procurar a próxima correspondência, clique no nome da área de janela novamente.

A visualização Clusters

Na visualização Clusters, é possível construir e explorar resultados de clusters localizados em seus dados de texto. *Clusters* são agrupamentos de conceitos gerados por algoritmos de armazenamento em cluster com base na frequência com que os conceitos ocorrem e na frequência com que aparecem juntos. O objetivo dos clusters é agrupar conceitos que coocorram juntos, enquanto o objetivo das categorias é agrupar documentos ou registros com base em como o texto que eles contêm corresponde aos descritores (conceitos, regras, padrões) para cada categoria.

Quanto mais vezes os conceitos dentro de um cluster ocorrerem juntos acoplados a uma menor frequência em que eles ocorrem com outros conceitos, melhor o cluster identificará relacionamentos de conceitos interessantes. Dois conceitos coocorrem quando ambos aparecem (ou um de seus sinônimos ou termos aparecem) no mesmo documento ou registro. Consulte o tópico Capítulo 10, “Analisando clusters”, na página 139 para obter informações adicionais.

É possível construir clusters e explorá-los em um conjunto de diagramas e gráficos que poderiam lhe ajudar a descobrir relacionamentos entre conceitos que, de outra forma, levariam muito tempo para serem localizados. Embora não seja possível incluir clusters em suas categorias, é possível incluir os conceitos em um cluster em uma categoria por meio da caixa de diálogo Definições de Cluster. Consulte o tópico “Definições de cluster” na página 143 para obter informações adicionais.

É possível fazer mudanças nas configurações para o armazenamento em cluster influenciar nos resultados. Consulte o tópico “Construindo clusters” na página 140 para obter informações adicionais.

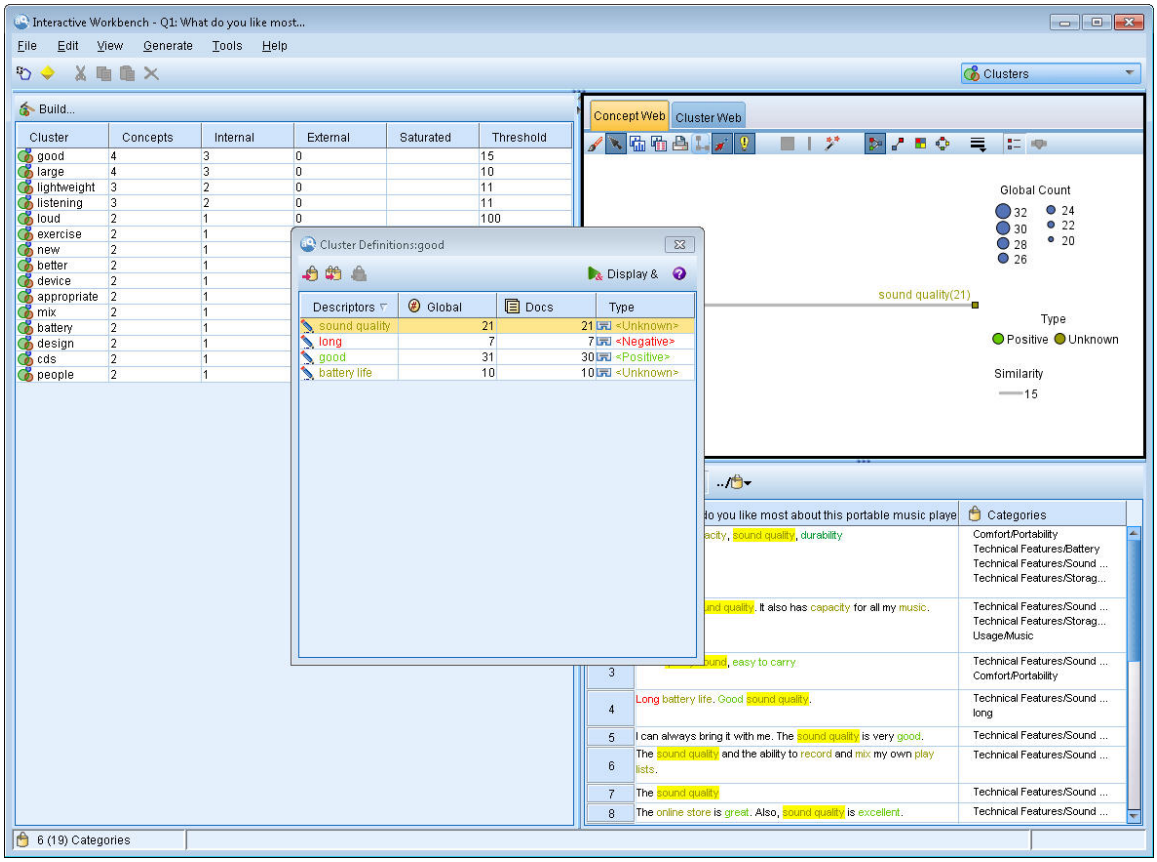


Figura 24. Visualização Clusters

A visualização Clusters é organizada em três áreas de janela, sendo que cada uma pode ser oculta ou mostrada selecionando seu nome no menu Visualizar. Normalmente, somente a área de janela Clusters e a área de janela Visualização ficam visíveis.

Área de janela Clusters

Localizada no lado esquerdo, esta área de janela apresenta os clusters que foram descobertos nos dados de texto. É possível criar resultados de armazenamento em cluster clicando no botão **Construir**. Clusters são formados por um algoritmo de clusterização, que tenta identificar conceitos que ocorrem juntos com frequência.

Sempre que acontece uma nova extração, os resultados do cluster são limpos e você precisa reconstruir os clusters para obter os resultados mais recentes. Durante a construção de seus clusters, é possível mudar algumas configurações, como o número máximo de clusters para criar, o número máximo de conceitos que ele pode conter ou o número máximo de ligações com conceitos externos que ele pode ter. Consulte o tópico “Explorando clusters” na página 142 para obter informações adicionais.

Área de janela Visualização

Localizada no canto superior direito, esta área de janela oferece duas perspectivas sobre o armazenamento em cluster: um gráfico Web de Conceito e um gráfico Web de Cluster. Se não visível, você pode acessar esta pane no menu Visualizar (**Visualizar> Visualização**). Dependendo do que estiver selecionado na área de janela de clusters, é possível visualizar as interações correspondentes entre ou dentro de clusters. Os resultados são apresentados em vários formatos:

- **Web de Conceito.** Gráfico da web mostrando todos os conceitos dentro do(s) cluster(s) selecionado(s), bem como conceitos vinculados fora do cluster.
- **Cluster Web.** Gráfico da web mostrando as ligações do(s) cluster(s) selecionado(s) para outros clusters, bem como quaisquer ligações entre esses outros clusters.

Nota: Para exibir um gráfico Web de Cluster, deve-se ter construído clusters com ligações externas. Ligações externas são aquelas entre pares de conceitos em clusters separados (um conceito dentro de um cluster e um conceito fora de outro cluster). Consulte o tópico "Gráficos de cluster" na página 157 para obter informações adicionais.

Área de janela Dados

A área de janela Dados está localizada no canto inferior direito e fica oculta por padrão. Não é possível exibir nenhum resultado da área de janela Dados a partir da área de janela Clusters, já que esses clusters abrangem vários documentos/registros, tornando os resultados dos dados desinteressantes. No entanto, é possível ver os dados correspondentes a uma seleção dentro da caixa de diálogo Definições de Cluster. Dependendo do que estiver selecionado nessa caixa de diálogo, somente o texto correspondente aparecerá na área de janela Dados. Uma vez que você faça uma seleção, clique no botão **Exibir &** para preencher o painel de Dados com os documentos ou registros que contêm todos os conceitos juntos.

Os documentos ou registros correspondentes mostram os conceitos destacados em cores para ajudá-lo a identificá-los facilmente no texto. Também é possível passar o mouse sobre os itens codificados por cores para exibir o conceito sob o qual eles foram extraído e o tipo ao qual eles foram designados. A área de janela Dados pode conter diversas colunas, mas a coluna do campo de texto será sempre mostrada. Ela carrega o nome do campo de texto que foi usado durante a extração ou um nome de documento, caso os dados de texto estejam em vários arquivos diferentes. Há outras colunas disponíveis. Consulte o tópico "A área de janela Dados" na página 99 para obter informações adicionais.

A visualização de análise de link de texto

Na visualização Análise de Ligação de Texto, é possível construir e explorar os padrões de análise de ligação de texto localizados em seus dados de texto. A análise de ligação de texto (TLA) é uma tecnologia de correspondência de padrões que permite definir regras de TLA e compará-las com os conceitos e relacionamentos reais extraídos localizados em seu texto.

Padrões são mais úteis quando você está tentando descobrir relacionamentos entre conceitos ou pareceres sobre um determinado assunto. Alguns exemplos incluem querer extrair pareceres sobre produtos de dados de pesquisas de opinião, relacionamentos genômicos de dentro de documentos de pesquisas médicas ou relacionamentos entre pessoas ou locais de dados de inteligência.

Após você ter extraído alguns padrões de TLA, é possível explorá-los nas áreas de janela Dados ou Visualização e até incluí-los em categorias na visualização Categorias e Conceitos. Deve haver algumas regras de TLA definidas no modelo de recurso ou bibliotecas que você está usando para extrair resultados de TLA. Consulte o tópico Capítulo 18, "Sobre regras de ligação de texto", na página 207 para obter informações adicionais.

Se você escolheu extrair resultados do padrão de TLA, os resultados serão apresentados nessa visualização. Se não tiver escolhido isso, você terá que usar o botão **Extrair** e escolher a opção para ativar a extração de padrões .

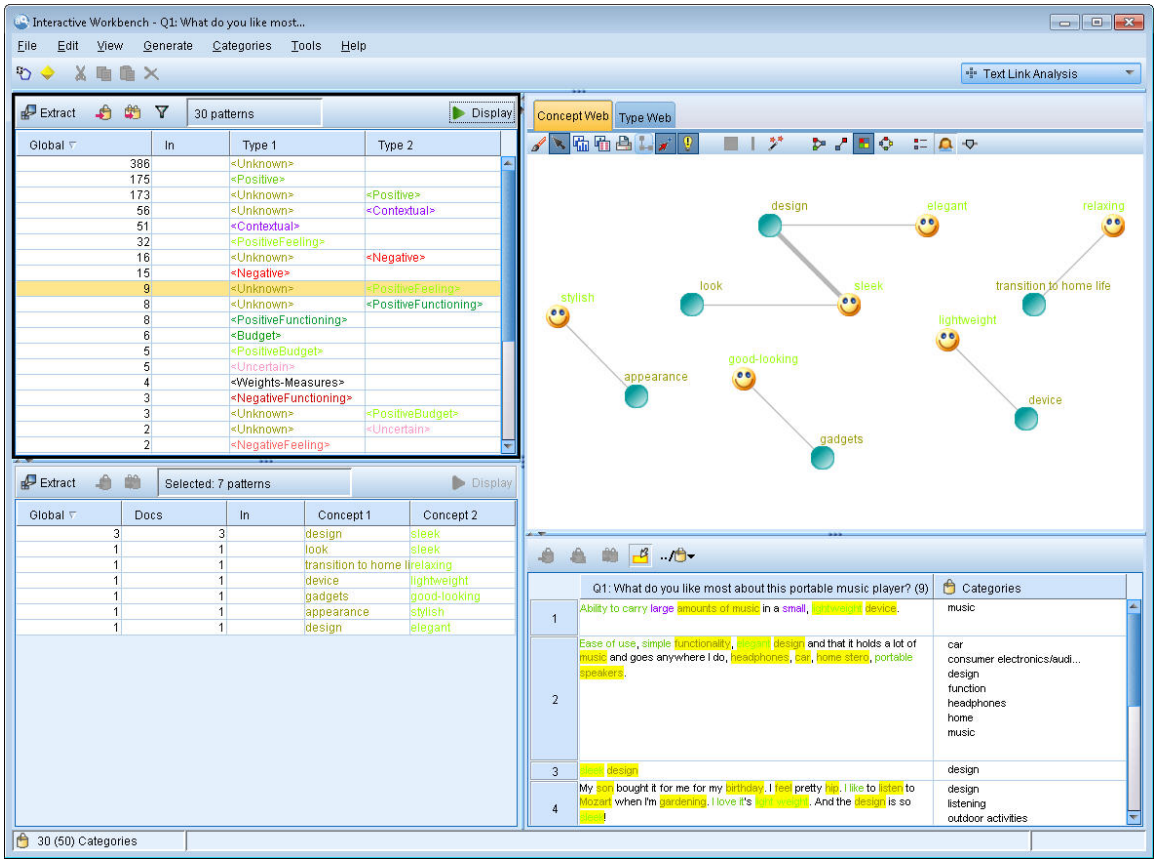


Figura 25. Visualização Análise de Ligação de Texto

A visualização Análise de Ligação de Texto é organizada em quatro áreas de janela, sendo que cada uma pode ser oculta ou mostrada selecionando seu nome no menu Visualizar. Consulte o tópico Capítulo 11, “Explorando a análise de ligação de texto”, na página 145 para obter informações adicionais.

Áreas de janela Padrões de Tipo e Padrões de Conceito

Localizadas no lado esquerdo, as áreas de janela Padrões de Tipo e Padrões de Conceitos são duas áreas de janela interconectadas nas quais é possível explorar e selecionar resultados de seu padrão de TLA. Padrões são compostos por uma série de até seis tipos ou seis conceitos. A regra de padrão de TLA, conforme definido nos recursos linguísticos, dita a complexidade dos resultados do padrão. Veja o tópico Capítulo 18, “Sobre regras de ligação de texto”, na página 207 para obter mais informações.

Os resultados do padrão são agrupados primeiro em nível de tipo e depois divididos em padrões de conceito. Por esse motivo, há duas áreas de janela de resultado diferentes: Padrões de Tipo (superior esquerda) e Padrões de Conceito (inferior esquerda).

- **Padrões de tipo.** A área de janela Padrões de Tipo apresenta padrões extraídos que consistem em dois ou mais tipos relacionados correspondentes a uma regra de padrão de TLA. Os padrões do tipo são mostrados como `<Organization>` + `<Location>` + `<Positive>`, o que pode fornecer feedback positivo sobre uma organização em um local específico.
- **Padrões de conceito.** A área de janela Padrões de Conceito apresenta os padrões extraídos em nível de conceito para todos os padrões de tipo atualmente selecionados na área de janela Padrões de Tipo acima dela. Os padrões de conceito seguem uma estrutura como `hotel + paris + wonderful`.

Assim como acontece com os resultados da extração na visualização Categorias e Conceitos, é possível visualizar os resultados aqui. Se vir algum refinamento que gostaria de fazer nos tipos e conceitos que compõem esses padrões, você os fará na área de janela Resultados da Extração na visualização Categorias e Conceitos ou diretamente no Editor de Recurso e extrairá seus padrões novamente.

Área de janela Visualização

Localizada no canto superior direito da visualização Análise de Ligação de Texto, esta área de janela apresenta um gráfico da web dos padrões selecionados, como padrões de tipo ou padrões de conceito. Se não visível, você pode acessar esta pane no menu Visualizar (**View** > **Visualization**). Dependendo do que estiver selecionado em outras áreas de janela, é possível visualizar as interações correspondentes entre documentos/registros e padrões.

Os resultados são apresentados em vários formatos:

- **Gráfico de conceito.** Este gráfico apresenta todos os conceitos no(s) padrão(ões) selecionado(s). A largura da linha e os tamanhos de nó (se os ícones de tipo não forem mostrados) em um gráfico de conceito mostram o número de ocorrências globais na tabela selecionada.
- **Gráfico de tipo.** Este gráfico apresenta todos os tipos no(s) padrão(ões) selecionado(s). A largura da linha e os tamanhos de nó (se os ícones de tipo não forem mostrados) no gráfico mostram o número de ocorrências globais na tabela selecionada. Nós são representados por uma cor de tipo ou por um ícone.

Consulte o tópico "Gráficos Análise de Ligação de Texto" na página 158 para obter informações adicionais.

Área de janela Dados

A área de janela Dados está localizada no canto inferior direito. Essa área de janela apresenta uma tabela contendo os documentos ou registros correspondentes a uma seleção em outra área da visualização. Dependendo do que for selecionado, somente o texto correspondente aparecerá na área de janela Dados. Após você fazer uma seleção, clique no botão **Exibir** para preencher a área de janela Dados com o texto correspondente.

Se você tiver uma seleção em outra área de janela, os documentos ou registros correspondentes mostrarão os conceitos destacados em cores para ajudá-lo a identificá-los facilmente no texto. Também é possível passar o mouse sobre os itens codificados por cores para exibir uma dica de ferramenta mostrando o nome do conceito sob o qual eles foram extraído e o tipo ao qual eles foram designados. Consulte o tópico "A área de janela Dados" na página 99 para obter informações adicionais.

A visualização do Editor de recursos

O IBM SPSS Modeler Text Analytics captura conceitos-chave de maneira rápida e precisa a partir de dados de texto usando um mecanismo de extração robusto. Este mecanismo conta fortemente com recursos linguísticos para ditar como grandes quantidades de dados não estruturados e textuais devem ser analisadas e interpretadas.

A visualização do Editor de Recursos é onde você pode visualizar e fazer ajuste fino nos recursos linguísticos usados para extrair conceitos, agrupá-los sob tipos, descobrir padrões nos dados de texto e muito mais.IBM SPSS Modeler Text Analytics oferece vários modelos de recursos pré-configurados. Além disso, em alguns idiomas, também é possível usar os recursos em um pacote de análise de texto. Consulte o tópico "Usando pacotes de análise de texto" na página 131 para obter mais informações.

Como esses recursos nem sempre podem ser perfeitamente adaptados ao contexto de seus dados, é possível criar, editar e gerenciar seus próprios recursos para um determinado contexto ou domínio no Editor de Recursos. Consulte o tópico Capítulo 15, "Trabalhando com bibliotecas", na página 175 para obter mais informações.

Para simplificar o processo de ajuste fino dos seus recursos linguísticos, você pode executar tarefas de dicionário comuns diretamente a partir da visualização Categorias e Conceitos através de menus de contexto nas áreas de janela Resultados e Dados da Extração. Veja o tópico "refinando resultados da extração" na página 86 para obter mais informações.

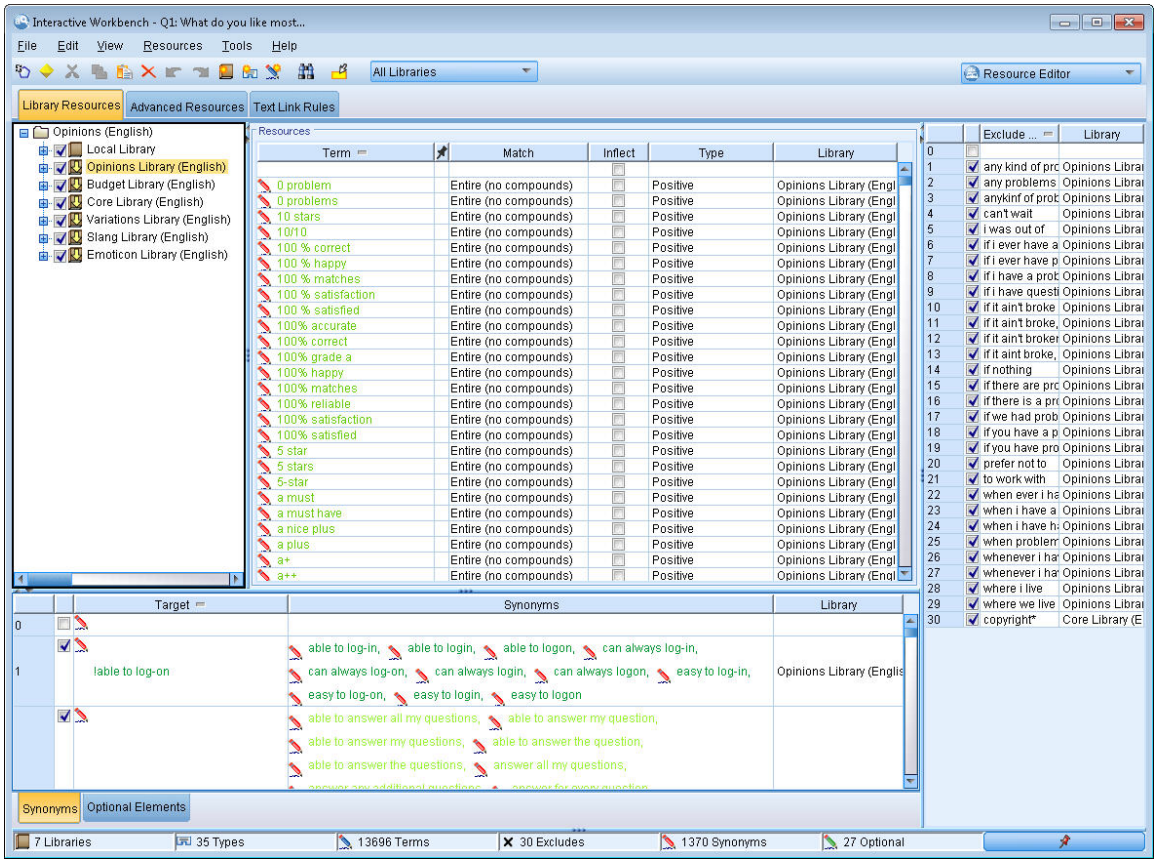


Figura 26. Visualização Editor de Recursos

As operações que você executa na visualização Editor de Recursos giram em torno do gerenciamento e ajuste fino dos recursos linguísticos. Esses recursos são armazenados na forma de modelos e bibliotecas. A visualização Editor de Recursos é organizada em quatro partes: área de janela Árvore de Bibliotecas, área de janela Dicionário de Tipos, área de janela Dicionário de Substituições e área de janela Dicionário de Exclusões.

Nota: Consulte o tópico "A interface do editor" na página 166 para obter informações adicionais.

Configurando opções

É possível configurar opções gerais para o IBM SPSS Modeler Text Analytics na caixa de diálogo Opções. Essa caixa de diálogo contém as seguintes guias:

- **Sessão.** Esta guia contém opções gerais e delimitadores.
- **Exibição.** Esta guia contém opções para as cores usadas na interface.
- **Sons.** Esta guia contém opções para dicas de som.

Para Editar Opções

1. A partir dos menus, escolha **Ferramentas> Opções**. A caixa de diálogo Opções se abre.
2. Selecione a guia contendo as informações que você deseja mudar.
3. Mude quaisquer opções.
4. Clique em **OK** para salvar as alterações.

Opções: guia Sessão

Nesta guia, é possível definir algumas configurações básicas.

Área de janela Dados e exibição do gráfico Categoria. Essas opções afetam como dados são apresentados na área de janela Dados e na área de janela Visualização na visualização Categorias e Conceitos.

- **Exibir limite para a área de janela Dados e Web de Categoria.** Esta opção configura o número máximo de documentos para mostrar ou usar para preencher as áreas de janela Dados ou gráficos e diagramas na visualização Categorias e Conceitos.
- **Mostrar categorias para documentos/registros no momento da exibição.** Se selecionada, os documentos ou registros serão escorados sempre que você clicar em Exibir para que quaisquer categorias às quais eles pertencem possam ser exibidas na coluna Categorias na área de janela Dados, bem como nos gráficos da categoria. Em alguns casos, principalmente com conjuntos de dados maiores, talvez você queira desativar essa opção para que dados e gráficos sejam exibidos mais rápido.

Incluir na categoria da área de janela Dados. Estas opções afetam o que é incluído nas categorias quando documentos e registros são incluídos a partir da área de janela Dados.

- **Na visualização Categorias e Conceitos, copiar.** A inclusão de um documento ou registro a partir da área de janela Dados nessa visualização copiará sobre **Somente Conceitos** ou **Conceitos e Padrões**.
- **Na visualização Análise de Ligação de Texto, copiar.** A inclusão de um documento ou registro a partir da área de janela Dados nessa visualização copiará sobre **Somente Padrões** ou **Conceitos e Padrões**.

Delimitador do Editor de Recurso. Selecione o caractere a ser usado como um delimitador quando você inserir elementos, como conceitos, sinônimos e elementos opcionais, na visualização Editor de Recurso.

Opções: guia Exibir

Nesta guia, é possível editar as opções que afetam a aparência geral do aplicativo e as cores usadas para distinguir elementos.

Nota: Para alternar a aparência do produto para uma aparência clássica ou de uma liberação anterior, abra o diálogo Opções do Usuário no menu Ferramentas na janela principal do IBM SPSS Modeler.

Cores Customizadas. Edite as cores para os elementos que aparecem na tela. Para cada elemento na tabela, é possível mudar a cor. Para especificar uma cor customizada, clique na área colorida à direita do elemento que deseja mudar e escolha uma cor da lista de cores suspensa.

- **Texto não extraído.** Os dados de texto que não foram extraídos, embora visíveis na área de janela Dados.
- **Segundo plano destacado.** Cor do plano de fundo da seleção de texto ao selecionar elementos nas áreas de janela ou texto na área de janela Dados.
- **Segundo plano com extração necessária.** A cor do plano de fundo das áreas de janela Resultados da Extração, Padrões e Clusters indicando que foram feitas mudanças nas bibliotecas e uma extração é necessária.
- **Segundo plano com feedback de categoria.** A cor do plano de fundo da categoria que aparece após uma operação.
- **Tipo padrão.** A cor padrão para os tipos e conceitos que aparecem na área de janela Dados e na área de janela Resultados da Extração. Essa cor se aplicará a quaisquer tipos customizados que você criar no Editor de Recurso. É possível substituir essa cor padrão por seus dicionários de tipo customizados editando as propriedades desses dicionários de tipo no Editor de Recursos. Consulte o tópico “Criando tipos” na página 185 para obter mais informações.
- **Tabela listrada 1.** Primeira das duas cores usadas de maneira alternada na tabela na caixa de diálogo Edit Forced concepts, a fim de diferenciar cada conjunto de linhas.
- **Tabela listrada 2.** Segunda das duas cores usadas de maneira alternada na tabela na caixa de diálogo Edit Forced concepts, a fim de diferenciar cada conjunto de linhas.

Nota: Se você clicar no botão **Reconfigurar para Padrões**, todas as opções nessa caixa de diálogo serão reconfiguradas para os valores que elas tinham quando você instalou o produto pela primeira vez.

Opções: guia Sons

Nesta guia, é possível editar as opções que afetam os sons. Sob Eventos de Som, é possível especificar um som para ser usado para notificá-lo quando ocorrer um evento. Há inúmeros sons disponíveis. Use o botão de reticências **(...)** para navegar e selecionar um som. Os arquivos *.wav* usados para criar sons para IBM SPSS Modeler Text Analytics estão armazenados no subdiretório *media* do diretório de instalação. Se não quiser que os sons sejam reproduzidos, selecione **Silenciar Todos os Sons**. Os sons ficam silenciados por padrão.

Nota: Se você clicar no botão **Reconfigurar para Padrões**, todas as opções nessa caixa de diálogo serão reconfiguradas para os valores que elas tinham quando você instalou o produto pela primeira vez.

Configurações do Microsoft Internet Explorer para ajuda

Configurações do Microsoft Internet Explorer

A maioria dos recursos de ajuda neste aplicativo usa tecnologia baseada no Microsoft Internet Explorer. Algumas versões do Internet Explorer (incluindo a versão fornecida com o Microsoft Windows XP, Service Pack 2) bloquearão, por padrão, aquilo que considerarem "conteúdo ativo" nas janelas do Internet Explorer em seu computador local. A configuração padrão pode resultar em conteúdo bloqueado nos recursos de ajuda. Para ver todo o conteúdo de ajuda, é possível mudar o comportamento padrão do Internet Explorer.

1. Nos menus do Internet Explorer, escolha:

   Ferramentas > **Opções da Internet ...**
2. Clique na guia **Avançado**.
3. Role para baixo para a seção **Segurança**.
4. Selecione **Permitir conteúdo ativo para execução em arquivos no Meu Computador**.

Gerando nuggets do modelo e nós de modelagem

Quando está em uma sessão interativa, talvez você queira usar o trabalho que fez para gerar:

- **Um nó de modelagem de mineração de texto.** Um nó de modelagem gerado a partir de uma sessão de ambiente de trabalho interativa é um nó Mineração de Texto cujas configurações e opções refletem aquelas armazenadas na sessão interativa aberta. Isso pode ser útil quando você não tem mais o nó Mineração de Texto original ou quando você deseja criar uma nova versão. Veja o tópico Capítulo 3, “Mineração para conceitos e categorias”, na página 19 para obter mais informações.
- **Um nugget do modelo de categoria.** Um nugget do modelo gerado a partir de uma sessão de ambiente de trabalho interativa é um nugget do modelo de categoria. Deve-se ter pelo menos uma categoria na visualização Categorias e Conceitos para gerar um nugget do modelo de categoria. Veja o tópico “Nugget de mineração de texto: modelo de categoria” na página 38 para obter mais informações.

Para Gerar um Nó de Modelagem Mineração de Texto

1. A partir dos menus, escolha **Gerar> Gerar Nó de Modelagem**. Um nó de modelagem Mineração de Texto é incluído na tela de trabalho usando todas as configurações atualmente na sessão de ambiente de trabalho. O nó é nomeado após o campo de texto.

Para Gerar um Nugget do Modelo de Categoria

1. A partir dos menus, escolha **Gerar> Gerar Modelo**. Um nugget do modelo é gerado diretamente na paleta Modelo com o nome padrão.

Atualizando nós de modelagem e salvando

Enquanto você está trabalhando em uma sessão interativa, é recomendado atualizar o nó de modelagem de tempos em tempos para salvar suas mudanças. Também é necessário atualizar seu nó de modelagem

sempre que você concluir o trabalho na sessão do ambiente de trabalho interativa e desejar salvá-lo. Quando você atualiza o nó de modelagem, o conteúdo da sessão do ambiente de trabalho é salvo novamente no nó Mineração de Texto que originou a sessão do ambiente de trabalho interativa. Isso não fecha a janela de saída.

Importante! Essa atualização não salvará seu fluxo. Para salvar seu fluxo, faça isso na janela principal do IBM SPSS Modeler após atualizar o nó de modelagem.

Para atualizar um nó de modelagem

1. A partir dos menus, escolha **Arquivo> Atualizar Nó de Modelagem**. O nó de modelagem é atualizado com as configurações de construção e extração junto com quaisquer opções e categorias que você tiver.

Fechando e terminando sessões

Quando você tiver concluído o trabalho em sua sessão, é possível deixar a sessão de três maneiras diferentes:

- **Salvar.** Esta opção permite que, primeiro, você salve seu trabalho de volta no nó de modelagem original para futuras sessões, bem como publique quaisquer bibliotecas para reutilizar em outras sessões. Veja o tópico "Compartilhando bibliotecas" na página 180 para obter mais informações. Após você ter salvado, a janela da sessão será fechada e a sessão será excluída do gerenciador de Saída na janela IBM SPSS Modeler.
- **Sair.** Esta opção descartará qualquer trabalho não salvo, fechará a janela da sessão e excluirá a sessão do gerenciador de Saída na janela IBM SPSS Modeler. Para liberar memória, recomendamos salvar todos os trabalhos importantes e sair da sessão.
- **Fechar.** Esta opção não salva ou descarta nenhum trabalho. Esta opção fecha a janela da sessão, mas a sessão continua sendo executada. É possível abrir a janela da sessão novamente selecionando essa sessão no gerenciador de Saída na janela IBM SPSS Modeler.

Para fechar uma sessão de ambiente de trabalho

1. A partir dos menus, escolha **Arquivo> Fechar**.

Teclado de acessibilidade

A interface do ambiente de trabalho interativa oferece atalhos de tecla para deixar a funcionalidade do produto mais acessível. No nível mais básico, é possível pressionar a tecla ALT mais a tecla apropriada para ativar os menus da janela (por exemplo, Alt+F para acessar o menu Arquivo) ou a tecla Tab para rolar pelos controles da caixa de diálogo. Esta seção cobre os atalhos de teclado para navegação alternativa. Há outros atalhos de teclado para a interface do IBM SPSS Modeler.

Tabela 13. Atalhos de teclado genéricos

Tecla de atalho	**Função**
Ctrl+1	Exibe a primeira guia em uma área de janela com guias.
Ctrl+2	Exibe a segunda guia em uma área de janela com guias.
Ctrl+A	Seleciona todos os elementos da área de janela que tem o foco.
Ctrl+C	Copia o texto selecionado para a área de transferência.
Ctrl+E	Ativa a extração nas visualizações Categorias e Conceitos e Análise de Ligação de Texto.
Ctrl+F	Exibe a barra de ferramentas Localizar no Editor de Recursos/Editor de Template, caso ela ainda não esteja visível, e a coloca em foco.

Tabela 13. Atalhos de teclado genéricos (continuação)	
Tecla de atalho	**Função**
Ctrl + I	Na visualização Categorias e Conceitos, ativa a caixa de diálogo Definições de Categoria para a categoria selecionada. Na visualização Cluster, ativa a caixa de diálogo Definições de Cluster para o cluster selecionado.
Ctrl+R	Abre a caixa de diálogo Incluir Termos no Editor de Recursos/Editor de Template.
Ctrl+T	Abre a caixa de diálogo Propriedades de Tipo para criar um novo tipo no Editor de Recursos/Editor de Template.
Ctrl+V	Cola conteúdo da área de transferência.
Ctrl+X	Corta os itens selecionados do Editor de Recursos/Editor de Template.
Ctrl+Y	Refaz a última ação na visualização.
Ctrl+Z	Desfaz a última ação na visualização.
F1	Exibe Ajuda ou, quando em uma caixa de diálogo, exibe Ajuda de Contexto para um item.
F2	Alterna a entrada e a saída do modo de edição nas células da tabela.
F6	Move o foco entre as áreas de janela principais na visualização ativa.
F8	Move o foco para as barras divisoras da área de janela para redimensionamento.
F10	Expande o menu Arquivo principal.
seta para cima, seta para baixo	Redimensiona a área de janela verticalmente quando a barra divisora é selecionada.
seta para esquerda, seta para direita	Redimensiona a área de janela horizontalmente quando a barra divisora é selecionada.
Home, End	Redimensiona as áreas de janela para os tamanhos mínimo e máximo quando a barra divisora é selecionada.
Guia	Move para frente pelos itens na janela, área de janela ou caixa de diálogo.
Shift+F10	Exibe o menu de contexto para um item.
Shift+Tab	Move para trás pelos itens na janela ou caixa de diálogo.
Shift+seta	Seleciona caracteres no campo de edição quando no campo de edição (F2).
Ctrl+Tab	Move o foco para frente para a próxima área principal na janela.
Shift+Ctrl+Tab	Move o foco para trás para a área principal anterior na janela.

Atalhos para caixas de diálogo

Várias teclas de atalho e de leitor de tela são úteis quando você está trabalhando com caixas de diálogo. Ao entrar em uma caixa de diálogo, você pode precisar pressionar a tecla Tab para colocar o foco no primeiro controle e para iniciar o leitor de tela. Uma lista completa de atalhos de teclado especiais e de leitor de tela é fornecida na tabela a seguir.

Tabela 14. Atalhos da caixa de diálogo	
Tecla de atalho	**Função**
Guia	Move para frente pelos itens na janela ou caixa de diálogo.

Tabela 14. Atalhos da caixa de diálogo (continuação)	
Tecla de atalho	**Função**
Ctrl+Tab	Move para frente de uma caixa de texto para o próximo item.
Shift+Tab	Move para trás pelos itens na janela ou caixa de diálogo.
Shift+Ctrl+Tab	Move para trás de uma caixa de texto para o item anterior.
barra de espaço	Selecione o controle ou botão que tem o foco.
Esc	Cancelar mudanças e fechar a caixa de diálogo.
Inserir	Validar mudanças e fechar a caixa de diálogo (equivalente ao botão OK). Se você estiver em uma caixa de texto, primeiro deve-se pressionar Ctrl+Tab para sair da caixa de texto.

Capítulo 8. Extraindo conceitos e tipos

Sempre que você executa um fluxo que lança o ambiente de trabalho interativo, uma extração é realizada automaticamente nos dados do texto no fluxo. O resultado final dessa extração é um conjunto de conceitos, tipos e, no caso em que existem padrões TLA nos recursos linguísticos, padrões. É possível visualizar e trabalhar com conceitos e tipos na área de janela Resultados da Extração. Veja "Como funciona a extração" na página 5 para obter mais informações.

Se você deseja afinar os resultados da extração, você pode modificar os recursos linguísticos e re-extrair. Veja "refinando resultados da extração" na página 86 para obter mais informações. O processo de extração conta com os recursos e quaisquer parâmetros na caixa de diálogo Extração para determinar como extrair e organizar os resultados. É possível usar os resultados da extração para definir a melhor parte, se não todas, de suas definições de categoria.

Nota: A partir da versão 18.2, os resultados do conceito extraído foram melhorados (agora eles são semelhantes aos resultados do conceito extraído em IBM SPSS Analítica de Texto para Pesquisas de Opinião)..

Resultados da extração: conceitos e tipos

Durante o processo de extração, todos os dados de texto são varridos e os conceitos relevantes são identificados, extraídos e designados aos tipos. Quando a extração é concluída, os resultados aparecem na área de janela Resultados da Extração localizada no canto inferior esquerdo da visualização Categorias e Conceitos. A primeira vez que você ativar a sessão, o modelo de recursos linguísticos que você selecionou no nó é usado para extrair e organizar esses conceitos e tipos.

Nota: Se houver mais resultados que possam caber na pane visível, você pode usar os controles na parte inferior da pane para mover para frente e para trás através dos resultados, ou inserir um número de página para ir até.

Os conceitos, tipos e padrões de TLA que são extraídos são coletivamente referidos como **resultados da extração** e eles servem como os descritores ou blocos de construção para suas categorias. Também é possível usar conceitos, tipos e padrões em suas regras de categoria. Além disso, as técnicas automáticas usam conceitos e tipos para construir a categorias.

A de mineração de texto é um processo iterativo no qual os resultados da extração são revisados de acordo com o contexto dos dados de texto, passam por ajuste fino para produzir novos resultados e, então, são reavaliados. Após a extração, você deve revisar os resultados e fazer quaisquer mudanças que você julgue necessário ao modificar os recursos linguísticos. É possível fazer ajuste fino nos recursos, em parte, diretamente da área de janela Resultados da Extração, da área de janela Dados, da caixa de diálogo Definições de Categoria ou da caixa de diálogo Definições do Cluster. Consulte o tópico "refinando resultados da extração" na página 86 para obter mais informações. Também é possível fazer isso diretamente na visualização Editor de Recursos. Veja o tópico "A visualização do Editor de recursos" na página 72 para obter mais informações.

Após o ajuste fino, é possível então extrair novamente para ver os novos resultados. Ao fazer ajuste fino nos resultados da sua extração a partir do início, você pode ter certeza que cada vez que extrair novamente, obterá resultados idênticos em suas definições de categoria, perfeitamente adaptados para ao contexto dos dados. Desta maneira, os documentos/registros serão designados para suas definições de categoria de uma maneira mais precisa e repetida.

Conceitos

Durante o processo de extração, os dados de texto são varridos e analisados para identificar palavras únicas interessantes e relevantes (tais como `election` ou `peace`) e frases (tais como `presidential election`, `election of the president` ou `peace treaties`) no texto. Essas palavras e frases são chamadas coletivamente de *termos*. Usando os recursos linguísticos, os termos relevantes são extraídos e, então, termos semelhantes são agrupados sob um termo principal denominado **conceito**.

É possível ver o conjunto de termos subjacentes para um conceito ao passar o mouse sobre o nome do conceito. Fazer isso exibirá uma dica de ferramenta mostrando o nome do conceito e várias linhas de termos que são agrupados sob tal conceito. Esses termos subjacentes incluem os sinônimos definidos nos recursos linguísticos (independentemente se eles estavam localizados no texto ou não), bem como quaisquer termos plurais/singulares extraídos, termos permutados, termos de agrupamento difuso e assim por diante. É possível copiar esses termos ou ver o conjunto completo de termos subjacentes ao clicar com o botão direito no nome do conceito e escolher a opção de menu de contexto.

Por padrão, os conceitos são mostrados em letras minúsculas e classificados em ordem decrescente, de acordo com a contagem do documento (coluna Doc.) . Quando os conceitos são extraídos, eles são designados a um tipo para ajudar a agrupar conceitos semelhantes. Eles são codificados com cores de acordo com este tipo. As cores são definidas nas propriedades de tipo no Editor de Recursos. Consulte o tópico "Dicionários de tipo" na página 183 para obter mais informações.

Sempre que um conceito, um tipo ou padrão está sendo usado em uma definição de categoria, um ícone aparece na coluna **In** classificável .

Tipos

Tipos são agrupamentos semânticos de conceitos. Quando os conceitos são extraídos, eles são designados a um tipo para ajudar a agrupar conceitos semelhantes. Diversos tipos integrados são entregues com o IBM SPSS Modeler Text Analytics , tal como `<Location>`, `<Organization>`, `<Person>`, `<Positive>`, `<Negative>` e assim por diante. Por exemplo, o tipo `<Location>` agrupa palavras-chave e locais geográficos. Este tipo seria designado a conceitos como `chicago`, `paris` e `tokyo`. Consulte o tópico Para a maioria das línguas, conceitos que não são encontrados em nenhum dicionário do tipo mas são extraídos do texto são digitados automaticamente como `<Unknown>` para obter mais informações. "Tipos integrados" na página 184

Quando você seleciona a visualização Tipo, os tipos extraídos aparecem por padrão em ordem decrescente pela frequência global. Também é possível ver que os tipos são codificados por cores para ajudar a distingui-los. As cores fazem parte das propriedades de tipo. Veja o tópico "Criando tipos" na página 185 para obter mais informações. Também é possível criar seus próprios tipos.

Padrões

Os padrões também podem ser extraídos dos seus dados de texto. Entretanto, deve-se ter uma biblioteca que contém algumas regras de padrão de Análise de Texto do Link (TLA) no Editor de Recursos. Também se deve escolher extrair esses padrões na configuração do nó doIBM SPSS Modeler Text Analytics ou na caixa de diálogo Extrair usando a opção **Ativar extração do padrão de Análise de Link de Texto**. Veja o tópico Capítulo 11, "Explorando a análise de ligação de texto", na página 145 para obter mais informações.

Extraindo dados

Sempre que uma extração é necessária, a área de janela Resultados da Extração fica com a cor amarela e a mensagem **Pressione o Botão Extrair para Extrair Conceitos** aparece abaixo da barra de ferramentas nessa área de janela.

Você pode precisar extrair se não tiver nenhum resultado de extração ainda, ter feito alterações nos recursos linguísticos e necessidade de atualizar os resultados da extração, ou tenha reaberto um session no qual você não salvou os resultados de extração (**Ferramentas** > **Opções**).

Nota: Se você alterar o nó de origem para o seu fluxo após resultados de extração ter sido armazenado em cache com o **Use session work ...** opção, você precisará executar uma nova extração uma vez que a sessão interativa do ambiente de trabalho seja lançada se você quiser obter resultados de extração atualizados.

Ao executar uma extração, um indicador de progresso aparece para fornecer feedback sobre o status da extração. Durante este tempo, o mecanismo de extração lê através de todos os dados de texto e identifica os termos e padrões relevantes e os extrai e atribui para um tipo. Em seguida, o mecanismo tenta agrupar termos sinônimos sob um termo principal, chamado de conceito. Quando o processo for concluído, os conceitos, tipos e padrões resultantes aparecerão na área de janela Resultados da Extração.

O processo de extração resulta em um conjunto de conceitos e tipos, bem como padrões de Análise de Link de Texto (TLA), se ativados. É possível visualizar e trabalhar com esses conceitos e tipos na área de janela Resultados da Extração na visualização Categorias e Conceitos. Se você extraiu padrões de TLA, poderá vê-los na visualização Análise de Link de Texto.

Nota: Há uma relação entre o tamanho do seu dataset e o tempo que ele leva para concluir o processo de extração. Você pode sempre considerar inserir um envio de dados de nó de Amostra ou otimizar a configuração da sua máquina.

Para Extrair Dados

1. A partir dos menus, escolha **Ferramentas** > **Extração**. Como alternativa, clique no botão da barra de ferramentas **Extrair**.
2. Se você escolheu sempre exibir o diálogo Configurações de Extração, ele aparecerá para que você possa fazer quaisquer mudanças. Veja mais neste tópico para os descritores de cada uma das configurações.
3. Clique em **Extrair** para começar o processo de extração. Assim que a extração começar, a caixa de diálogo será aberta. Após a extração, os resultados aparecem na área de janela Resultados da Extração. Por padrão, os conceitos são mostrados em letras minúsculas e classificados em ordem decrescente, de acordo com a contagem do documento (coluna Doc.) .

É possível revisar os resultados usando as opções da barra de ferramentas para classificar os resultados de forma diferente, para filtrar os resultados ou para alternar para uma visualização diferente (conceitos ou tipos). Também é possível refinar seus resultados de extração ao trabalhar com os recursos linguísticos. Veja o tópico "refinando resultados da extração" na página 86 para obter mais informações.

Questões potenciais de extração

Várias Sessões do Ambiente Interativo podem causar um comportamento lento. SPSS Análise de Texto do Modeler e SPSS Modeler compartilham um mecanismo de tempo de execução Java comum quando uma sessão interativa de ambiente de trabalho é lançada. Dependendo do número de sessões do Workbench Interativo você chama durante uma sessão SPSS Modeler -mesmo se abrindo e fechando a mesma memória do sistema de sessão pode fazer com que o aplicativo fique lento. Este efeito pode ser especialmente pronunciado se você estiver trabalhando com dados grandes ou ter uma máquina com menos do que a configuração de RAM recomendada de 4GB. Se você notar que sua máquina está lenta para responder, é recomendável que você salve todo o seu trabalho, desligue SPSS Modelere relembre o aplicativo. Executar SPSS Análise de Texto do Modeler em uma máquina com menos do que a memória recomendada, particularmente ao trabalhar com grandes conjuntos de dados ou por períodos prolongados de tempo, pode fazer com que o Java fique sem memória e encerrado. É fortemente sugerida você se atualizar para a configuração de memória recomendada ou maior (ou use SPSS Análise de Texto do Modeler Server) se você trabalha com dados grandes.

Para texto em holandês, inglês, francês, alemão, italiano, português e espanhol

A caixa de diálogo Configurações de Extração contém algumas opções básicas de extração.

Ativar extração de padrão de Análise de Link de Texto. Especifica se você deseja extrair padrões de TLA dos seus dados de texto. Também assume que você tem regras do padrão de TLA em uma de suas bibliotecas no Editor de Recursos. Esta opção pode aumentar significativamente o tempo da extração. Veja o tópico Capítulo 11, "Explorando a análise de ligação de texto", na página 145 para obter mais informações.

Acomodar erros de pontuação. Esta opção normaliza temporariamente o texto contendo erros de pontuação (por exemplo, uso incorreto) durante a extração para melhorar a extractabilidade de conceitos. Esta opção é muito útil quando o texto é curto e de qualidade ruim (como, por exemplo, em respostas de pesquisa sem estrutura, e-mail e dados CRM) ou quando o texto contém muitas abreviações.

Acomodar a ortografia para um comprimento mínimo de caracteres de palavra de [n] Esta opção aplica uma técnica de agrupamento fuzzy que ajuda a agrupar palavras comumente digitadas ou palavras bem escritas sob um só conceito. O algoritmo de agrupamento difuso temporariamente remove todas

as vogais (exceto a primeira) e remove consoantes duplas/triplas das palavras extraída e, em seguida, as compara para ver se elas são a mesma, desta forma `modeling` e `modelling` seriam agrupadas. Entretanto, se cada termo for designado a um tipo diferente, excluindo o tipo <Unknown>, a técnica de agrupamento difuso não será aplicada.

Também é possível definir o número mínimo de caracteres *root* necessários antes que o agrupamento difuso seja usado. O número de caracteres raiz em um termo é calculado ao totalizar todos os caracteres e subtrair quaisquer caracteres que formam sufixos de inflexão e, no caso de termos com palavras compostas, determinadores e preposições. Por exemplo, o termo `exercises` seria contado como 8 caracteres raiz no formato "exercise," já que a letra *s* no final da palavra é uma inflexão (forma plural). De maneira semelhante, `apple sauce` é contado como 10 caracteres raiz ("apple sauce") e `manufacturing of cars` é contado como 16 caracteres raiz ("manufacturing car"). Este método de contagem é usado apenas para verificar se o agrupamento difuso deve ser aplicado, mas não influencia como as palavras são correspondidas.

Nota: Se você descobrir que certas palavras são mais tarde agrupadas incorretamente, você pode excluir pares de palavras desta técnica declarando-as explicitamente na seção **Agrupamento Fuzzy: Excepções** na guia Recursos Avançados. Veja o tópico "Agrupamento difuso" na página 197 para obter mais informações.

Extrair uniterms Esta opção extrai palavras únicas (uniterms) desde que a palavra não seja já parte de uma palavra composta e se for um substantivo ou uma parte não reconhecida da fala.

Extrair entidades não lingüísticas Esta opção extrai entidades não linguísticas, como números de telefone, números de previdência social, horários, datas, moedas, dígitos, porcentagens, endereços de e-mail e endereços HTTP. Você pode incluir e excluir determinados tipos de entidades não linguísticas na seção **Entidades Não Linguísticas: Configuração** da guia Recursos Avançados. Ao desativar qualquer entidades desnecessárias, o mecanismo de extração não desperdiçará tempo de processamento. Veja o tópico "Configuração" na página 201 para obter mais informações.

Algoritmos Uppercase Esta opção extrai termos simples e compostos que não estão nos dicionários embutidos desde que a primeira letra do termo esteja em maiúseada. Esta opção oferece uma boa maneira de extrair substantivos mais adequados.

Grupos parciais e completos de pessoa juntos quando possível Esta opção agrupa nomes que aparecem de forma diferente no texto em conjunto. Esse recurso é útil já que os nomes são frequentemente referidos em sua forma completa no início do texto e, então, apenas por uma versão mais curta. Esta opção tenta corresponder qualquer unitermo com o tipo<Unknown> com a última palavra de qualquer um dos termos compostos que é digitado como <Person>. Por exemplo, se *doe* for localizado e inicialmente digitado como <Unknown>, o mecanismo de extração verifica se quaisquer termos compostos no tipo <Person> incluem *doe* como a última palavras, tal como *john doe*. Esta opção não se aplica a nomes já que a maioria nunca é extraída como unitermos.

Permutação de palavra não função máxima Esta opção especifica o número máximo de palavras sem função que podem estar presentes ao aplicar a técnica de permutação. Essa técnica permutação agrupa frases semelhantes que diferem uma da outra apenas pelas palavras sem função contidas (por exemplo, de e o), independentemente da inflexão. Por exemplo, digamos que você configure este valor para no máximo duas palavras e ambos `company officials` e `officials of the company` foram extraídas. Nesse caso, ambos os termos extraídos seriam agrupados na lista de conceito final já que ambos os termos são considerados o mesmo quando `of the` é ignorado.

Usar derivação ao agrupar multitermos Ao processar Big Data, selecione esta opção para agrupar multitermos usando regras de derivação.

Opção de Índice de Mapa Conceito Especifica que você deseja construir o índice de mapa no momento da extração para que os mapas de conceito possam ser rapidamente desenhados posteriormente. Para editar as configurações de índice, clique em **Configurações**. Veja o tópico "Construindo índices de mapa de conceito" na página 86 para obter mais informações.

Sempre mostre este diálogo antes de iniciar uma extração Especificar se deseja ver o diálogo Configurações de Extração cada vez que você extrair, se você nunca quiser vê-lo a menos que você vá para o menu Ferramentas, ou se deseja ser perguntado cada vez que você extraia se deseja editar qualquer configurações de extração.

Filtrando resultados da extração

Quando você está trabalhando com conjuntos de dados muito grandes, o processo de extração pode produzir milhões de resultados. Para muitos usuários, essa quantidade pode dificultar a revisão efetiva dos resultados. Portanto, para aumentar o zoom naqueles que são mais interessantes, é possível filtrar esses resultados por meio do diálogo Filtro disponível na área de janela Resultados da Extração.

Lembre-se de que todas as configurações nesse diálogo Filtro são usadas juntas para filtrar os resultados da extração disponíveis para categorias.

Filtrar por Frequência Você pode filtrar para exibir apenas esses resultados com um determinado valor de frequência global ou de documento.

- **Frequência Global** é o número total de vezes que um conceito aparece no conjunto inteiro de documentos ou registros e é mostrado na coluna **Global**.
- **Frequência de Documento** é o número total de documentos ou registros em que um conceito aparece e é mostrado na coluna **Docs**.

Por exemplo, se um conceito `nato` apareceu 800 vezes em 500 registros, poderíamos dizer que ele tem uma frequência global de 800 e uma frequência de documento de 500.

E por Tipo Você pode filtrar para exibir apenas aqueles resultados pertencentes a determinados tipos. É possível escolher todos os tipos ou somente tipos específicos.

E pelo Match Text Você também pode filtrar para exibir apenas aqueles resultados que correspondem à regra que você define aqui. Insira o conjunto de caracteres a ser correspondido no campo **Texto de Correspondência** e depois selecione a condição à qual aplicar a correspondência.

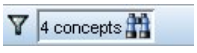
Tabela 15. Condições do texto de correspondência

Condição	Descrição
Contém	Texto é correspondido se a sequência de caracteres ocorrer em qualquer lugar. (Opção padrão)
Começa com	Texto é correspondido somente se o conceito ou tipo começar com o texto especificado.
Termina com	Texto é correspondido somente se o conceito ou tipo terminar com o texto especificado.
Correspondência exata	A sequência de caracteres inteira deve corresponder ao nome do conceito ou tipo.

Resultados exibidos na área de janela Resultado da Extração

Aqui estão alguns exemplos de como os resultados podem ser exibidos, em inglês, na barra de ferramentas da área de janela Resultado da Extração baseada em filtros.

Tabela 16. Exemplos de feedback de filtro

Feedback de filtro	Descrição
	A barra de ferramentas mostra o número de resultados. Como não havia um filtro correspondente ao texto e o máximo não foi atingido, nenhum ícone adicional é mostrado.
	A barra de ferramentas mostra resultados que foram limitados ao máximo especificado no filtro, que era 300. Se um ícone púrpura estiver presente, isso significa que o número máximo de conceitos foi atingido. Passe o mouse sobre o ícone para obter mais informações.
	A barra de ferramentas mostra resultados que foram limitados usando um filtro de texto correspondente. Isso é mostrado pelo ícone de lupa.

Para filtrar os resultados

1. A partir dos menus, escolha **Ferramentas** > **Filtro**. A caixa de diálogo Filtro é aberta.
2. Selecione e refine os filtros que deseja usar.
3. Clique em **OK** para aplicar os filtros e ver os novos resultados na área de janela Resultado da Extração.

Explorando mapas de conceito

É possível criar um mapa de conceito para explorar como os conceitos estão interrelacionados. Selecionando um único conceito e clicando em **Mapa**, uma janela de mapa de conceito será aberta para que seja possível explorar o conjunto de conceitos relacionado ao conceito selecionado. É possível filtrar quais conceitos serão exibidos editando configurações como tipos para incluir, tipos de relacionamento para procurar, entre outros.

Importante: Para que um mapa possa ser criado, um índice deve ser gerado. Isto pode levar alguns minutos. No entanto, após você ter gerado o índice, não será necessário gerá-lo novamente até fazer uma nova extração. Se quiser que o índice seja gerado automaticamente cada vez que você extrair, selecione essa opção nas configurações de extração. Consulte o tópico "Extraindo dados" na página 80 para obter informações adicionais.

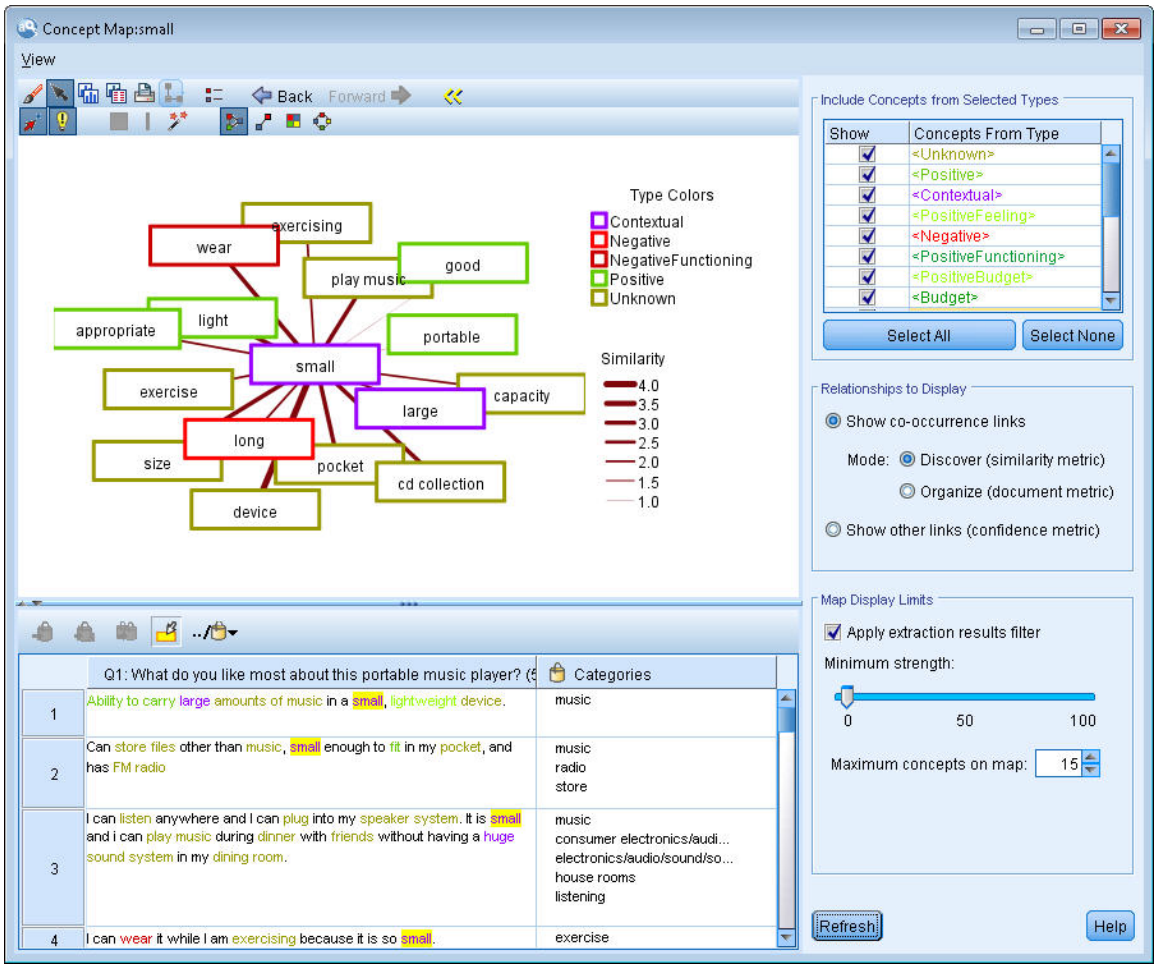


Figura 27. Um mapa de conceito para o conceito selecionado

Para Visualizar um Mapa de Conceito

1. Na área de janela Resultados da Extração, selecione um conceito único.
2. Na barra de ferramentas dessa área de janela, clique no botão **Mapa**. Se o índice do mapa já tiver sido gerado, o mapa de conceito será aberto em um diálogo separado. Se o índice do mapa ainda não

foi gerado ou estiver desatualizado, o índice deverá ser reconstruído. Este processo pode levar vários minutos.

3. Clique ao redor do mapa para explorá-lo. Se você der um clique duplo em um conceito vinculado, o mapa se redesenhará e mostrará os conceitos vinculados para o conceito no qual você acabou de dar um clique duplo.
4. A barra de ferramentas superior oferece algumas ferramentas de mapa básicas, como se mover de volta para um mapa anterior, filtrar ligações de acordo com a intensidade de um relacionamento e também abrir o diálogo de filtro para controlar os tipos de conceitos que aparecem, bem como os tipos de relacionamentos a serem representados. Uma segunda linha da barra de ferramentas contém ferramentas de edição de gráfico. Consulte o tópico "Usando barras de ferramentas e paletas de gráfico" na página 159 para obter informações adicionais.
5. Se você estiver insatisfeito com os tipos de links sendo localizados, revise as configurações para esse mapa mostradas do lado direito do mapa.

Configurações de Mapa: Incluir Conceitos dos Tipos Selecionados

Somente os conceitos que pertencem aos tipos selecionados na tabela são mostrados no mapa. Para ocultar conceitos de um determinado tipo, cancele a seleção desse tipo na tabela.

Configurações de Mapa: Relacionamentos para Exibir

Mostrar links de co-ocorrência Se você quiser mostrar links de co-ocorrência, escolha o modo. O modo afeta como a intensidade da ligação foi calculada.

- *Descobrir (métrica de similaridade)*. Com esta métrica, a intensidade da ligação é calculada usando um cálculo mais complexo que leva em conta a frequência com que dois conceitos aparecem separados, bem como a frequência com que aparecem juntos. Um alto valor de intensidade significa que um par de conceitos tende a aparecer com mais frequência junto do que separado. Com a fórmula a seguir, quaisquer valores de ponto flutuante são convertidos em números inteiros.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Figura 28. Fórmula de coeficiente de similaridade

Nesta fórmula, C_I é o número de documentos ou registros nos quais ocorre o conceito `I`.

C_J é o número de documentos ou registros nos quais ocorre o conceito `J`.

C_{IJ} é o número de documentos ou registros nos quais há coocorrência de par de conceitos `I` e `J` no conjunto de documentos.

- *Organizar (métrica de documento)*. A intensidade das ligações com essa métrica é determinada pela contagem bruta de coocorrências. Em geral, quanto mais frequentes os dois conceitos forem, maior a probabilidade de eles ocorrerem juntos às vezes. Um alto valor de intensidade significa que um par de conceitos aparece junto com mais frequência.

Mostrar outras ligações (métrica de confiança). É possível escolher outras ligações para exibir; elas podem ser semânticas, derivação (morfológica) ou inclusão (sintática) e estão relacionadas com quantas etapas removidas está um conceito a partir do conceito com o qual elas estão ligadas. Isso pode ajudá-lo a ajustar recursos, particularmente sinônimos ou a disambiguate.For descrições simples de cada uma dessas técnicas de agrupamento, consulte "Configurações linguísticas avançadas" na página 104

Nota: Tenha em mente que se estes não foram selecionados quando o índice foi construído ou se nenhum relacionamento foi encontrado, então nenhum será exibido. Consulte o tópico "Construindo índices de mapa de conceito" na página 86 para obter informações adicionais.

Configurações de Mapa: Limites de Exibição de Mapa

Aplicar filtros de resultados de extração. Se você não desejar usar todos os conceitos, é possível usar o filtro na área de janela de resultados de extração para limitar o que é mostrado. Em seguida, selecione essa opção e o IBM SPSS Modeler Text Analytics procurará conceitos relacionados usando esse conjunto filtrado. Consulte o tópico "Filtrando resultados da extração" na página 83 para obter mais informações.

Intensidade mínima. Configure a intensidade de ligação mínima aqui. Quaisquer conceitos relacionados com uma intensidade de relacionamento inferior a esse limite serão ocultos no mapa.

Conceitos máximos no mapa. Especifique o número máximo de relacionamentos para mostrar no mapa.

Construindo índices de mapa de conceito

Para que um mapa possa ser criado, um índice de relacionamentos de conceito deve ser gerado. Sempre que você criar um mapa de conceito, IBM SPSS Modeler Text Analytics refere-se a este índice. É possível escolher quais relacionamentos indexar selecionando as técnicas nesse diálogo.

Técnicas de agrupamento. Escolha um ou mais técnicas. Para obter descrições simples de cada uma dessas técnicas, consulte "Sobre técnicas de linguística" na página 106 Nem todas as técnicas estão disponíveis para todos os idiomas de texto.

Evitar emparelhamento de conceitos específicos. Selecione esta caixa de seleção para impedir que o processo agrupe ou pareie dois conceitos na saída. Para criar ou gerenciar pares de conceitos, clique em **Gerenciar pares**. Consulte o tópico "Gerenciando pares de exceção de link" na página 106 para obter mais informações.

A construção do índice pode demorar vários minutos. No entanto, após você ter gerado o índice, não será necessário gerá-lo novamente até fazer uma nova extração ou a menos que você queira mudar as configurações para incluir mais relacionamentos. Se desejar gerar um índice sempre que você extrair, é possível selecionar essa opção nas configurações de extração. Consulte o tópico "Extraindo dados" na página 80 para obter informações adicionais.

refinando resultados da extração

A extração é um processo iterativo pelo qual você pode extrair, revisar os resultados, fazer alterações em eles e, em seguida, re-extrair para atualizar os resultados. Uma vez que a precisão e a continuidade são essenciais para a mineração e categorização de texto bem-sucedidos, afinando seus resultados de extração desde o início garante que cada vez que você reextrai, você obterá precisamente os mesmos resultados em suas definições de categoria. Desta maneira, os registros e documentos serão designados às suas categorias de uma maneira mais precisa e repetida.

Os resultados da extração servem como os blocos de construção para as categorias. Ao criar categorias usando esses resultados da extração, os registros e documentos são automaticamente designados às categorias se eles contiverem texto que corresponda a um ou mais descritores de categoria. Embora você possa começar a categorização antes de fazer quaisquer refinamentos nos recursos linguísticos, é útil revisar seus resultados da extração pelo menos uma vez antes de começar.

Conforme revisa seus resultados, você pode localizar elementos que deseja que os mecanismo de extração manipule de maneira diferente. Considere os exemplos a seguir:

- **Sinônimos não reconhecidos.** Suponha que você localize vários conceitos que considere serem sinônimos, tais como `smart`, `intelligent`, `bright` e `knowledgeable` e todos eles apareçam como conceitos individuais nos resultados da extração. Você poderia criar uma definição de sinônimo na qual `intelligent`, `bright` e `knowledgeable` são todos agrupados sob o conceito de destino `smart`. Fazer isso agruparia todos esses com `smart` e a contagem de frequência global também seria maior. Consulte o tópico "Incluindo sinônimos" na página 87 para obter mais informações.
- **Conceitos com tipo incorreto.** Suponha que os conceitos em seus resultados da extração apareçam em um tipo e você gostaria que eles fossem designados a outro. Em outro exemplo, imagine que você localize 15 conceitos de vegetal em seus resultados da extração e deseje que todos eles sejam incluído em um novo tipo chamado `<Vegetable>`. Para a maioria das línguas, conceitos que não são

encontrados em nenhum dicionário do tipo mas são extraídos do texto são digitados automaticamente como <Unknown> É possível incluir conceitos nos tipos. Consulte o tópico "incluindo conceitos nos tipos" na página 88 para obter mais informações.

- **Conceitos não significativos.** Suponha que você localize um conceito que foi extraído e possui uma contagem de frequência muito alta - ou seja, ele está localizado em muitos registros ou documentos. Entretanto, você considera este conceito como não significativo para sua análise. É possível excluí-lo da extração. Consulte o tópico "Excluindo conceitos da extração" na página 89 para obter mais informações.
- **Correspondências incorretas.** Suponha que na revisão dos registros ou documentos que contêm um determinado conceito, você descubra que duas palavras foram incorretamente agrupadas, tais como `faculty` e `facility`. Essa correspondência pode ser devido a um algoritmo interno, referido como agrupamento difuso, que ignora temporariamente consoantes e vogais duplas ou triplas para agrupar erros de ortografia comuns. É possível incluir essas palavras em uma lista de pares de palavras que não devem ser agrupadas. Veja o tópico "Agrupamento difuso" na página 197 para obter mais informações.
- **Conceitos não extraídos.** Suponha que você espere localizar determinados conceitos extraídos, mas observe que algumas palavras ou frases não foram extraídas ao revisar o texto do registro ou documento. Muitas vezes, essas são verbos ou adjetivos nos quais você não está interessado. Entretanto, algumas vezes, você deseja usar uma palavra ou frase que não foi extraída como parte de uma definição de categoria. Para extrair o conceito, é possível forçar um termo em um dicionário de tipos. Consulte o tópico "Forçando palavras na extração" na página 90 para obter mais informações.

Muitas dessas mudanças podem ser executadas diretamente a partir da área de janela Resultados da Extração , da área de janela Dados, da caixa de diálogo Definições de Categoria ou da caixa de diálogo Definições de Cluster ao selecionar um ou mais elementos e clicar com o botão direito do seu mouse para acessar menus de contexto.

Depois de fazer suas alterações, a cor de pano de fundo muda para mostrar que você precisa re-extrair para visualizar suas alterações. Consulte o tópico "Extraindo dados" na página 80 para obter informações adicionais. Se você estiver trabalhando com conjuntos de dados maiores, pode ser mais eficiente re-extrair depois de fazer várias alterações em vez de depois de cada mudança.

Nota: É possível visualizar todo o conjunto de recursos linguísticos editáveis utilizados para produzir os resultados de extração na visualização Editor de Recursos (Visualizar> Editor de Recursos). Esses recursos aparecem na forma de bibliotecas e dicionários nesta visão. É possível customizar os conceitos e tipos diretamente nas bibliotecas e dicionários. Veja o tópico Capítulo 15, "Trabalhando com bibliotecas", na página 175 para obter mais informações.

Incluindo sinônimos

Sinônimos associam duas ou mais palavras que possuem o mesmo significado. Os sinônimos também são frequentemente usados para agrupar termos com suas abreviações ou para agrupar palavras comumente digitadas incorretamente com a ortografia correta. Ao usar sinônimos, a frequência para o conceito de destino é maior, o que torna muito mais fácil descobrir informações semelhantes que são apresentadas em formas diferentes em seus dados de texto.

Os modelos e as bibliotecas de recursos linguísticos fornecidos com o produto contêm muitos sinônimos predefinidos. Entretanto, se você descobrir sinônimos não reconhecidos, é possível defini-los para que eles sejam reconhecidos na próxima vez que você extrair.

A primeira etapa é decidir qual será o conceito de destino ou principal. O *conceito de destino* é a palavra ou frase sob a qual você deseja agrupar todos os termos sinônimos nos resultados finais. Durante a extração, os sinônimos são agrupados sob esse conceito de destino. A segunda etapa é identificar todos os sinônimos para esse conceito. O conceito de destino é substituído por todos os sinônimos na última extração. Um termo deve ser extraído para ser um sinônimo. Entretanto, o conceito de destino não precisa ser extraídos para que a substituição ocorra. Por exemplo, se você desejar que `intelligent` seja substituído por `smart`, então `intelligent` é o sinônimo e `smart` é o conceito de destino.

Se você criar uma nova definição de sinônimo, um novo conceito de destino será incluído no dicionário. Deve-se, então, incluir sinônimos em tal conceito de destino. Sempre que você criar ou editar sinônimos, essas mudanças serão registradas em dicionários de sinônimos no Editor de Recursos. Se você desejar

visualizar o conteúdo inteiro desses dicionários de sinônimos ou se desejar criar um número substancial de mudanças, poderá preferir trabalhar diretamente no Editor de Recursos. Veja o tópico "Dicionários de substituição/sinônimo" na página 190 para obter mais informações.

Quaisquer novos sinônimos serão automaticamente armazenados na primeira biblioteca listada na árvore de bibliotecas na visualização Editor de Recursos — por padrão, essa é a **Biblioteca Local**.

Nota: Se você procurar uma definição de sinônimo e não puder localizá-lo através dos menus de contexto ou diretamente no Editor de Recursos, uma correspondência pode ter resultado de uma técnica de agrupamento difuso interna. Veja o tópico "Agrupamento difuso" na página 197 para obter mais informações.

Para Criar um Novo Sinônimo

1. Na área de janela Resultados da Extração , na área de janela Dados, na caixa de diálogo Definições de Categoria ou na caixa de diálogo Definições de Cluster, selecione o(s) conceito(s) para o(s) qual(is) você deseja criar um novo sinônimo.
2. A partir dos menus, escolha **Editar** > **Adicionar ao Sinônimo** > **Novo**. A caixa de diálogo Criar Sinônimo é aberta.
3. Insira um conceito de destino na caixa de texto Destino. Este é o conceito sob o qual todos os sinônimos serão agrupados.
4. Se você desejar incluir mais sinônimos, insira-os na caixa de listagem Sinônimos. Use o separador global para separar cada termo sinônimo. Consulte o tópico "Opções: guia Sessão" na página 73 para obter mais informações.
5. Clique em **OK** para aplicar suas mudanças. A caixa de diálogo é fechada e a cor de plano de fundo da área de janela Resultados da Extração muda, indicando que você precisa extrair novamente para ver suas mudanças. Se você tiver diversas mudanças, faça-as antes de extrair novamente.

Para adicionar um sinônimo

1. Na área de janela Resultados da Extração , na área de janela Dados, na caixa de diálogo Definições de Categoria ou na caixa de diálogo Definições de Cluster, selecione o(s) conceito(s) que você deseja incluir em uma definição de sinônimo existente.
2. A partir dos menus, escolha **Editar** > **Adicionar ao Sinônimo**. O menu exibe um conjunto de sinônimos com os mais recentemente criados no topo da lista. Selecione o nome do sinônimo ao qual deseja inclui o(s) conceito(s) selecionado(s). Se você vir o sinônimo que está procurando, selecione-o e o(s) conceito(s) selecionado(s) será(ão) selecionado(s) tem tal definição de sinônimo. Se você não o vir, selecione **Mais** para exibir a caixa de diálogo Todos os Sinônimos.
3. Na caixa de diálogo Todos os Sistemas, é possível classificar a lista pela ordem de classificação natural (origem de criação) ou na ordem crescente ou decrescente. Selecione o nome do sinônimo ao qual você deseja incluir o(s) conceito(s) selecionado(s) e clique em **OK**. A caixa de diálogo é fechada e os conceitos são incluídos na definição de sinônimo.

incluindo conceitos nos tipos

Sempre que uma extração é executada, os conceitos extraídos são designados a tipos em um esforço para agrupar termos que têm algo em comum. IBM SPSS Modeler Text Analytics é entregue com muitos tipos embutidos. Veja o tópico "Tipos integrados" na página 184 para obter mais informações. Para a maioria das línguas, conceitos que não são encontrados em nenhum dicionário do tipo mas são extraídos do texto são digitados automaticamente como <Unknown>

Ao revisar seus resultados, você pode descobrir que alguns conceitos que aparecem em um tipo que deseja designado para outro ou pode descobrir que um grupo de palavras realmente pertence a um novo tipo por si mesmo. Nesses casos, você desejaria redesignar os conceitos para outro tipo ou criar um novo tipo por completo.

Por exemplo, suponha que você esteja trabalhando com dados de pesquisa de opinião relativos a automóveis e esteja interessado em categorizar ao concentrar-se em diferentes áreas dos veículos.

Você poderia criar um tipo chamado <Dashboard> para agrupar todos os conceitos relacionados a medidores e botões localizado no painel de veículos. Em seguida, você poderia designar conceitos como gas gauge, heater, radio e odometer para esse novo tipo.

Em outro exemplo, suponha que você esteja trabalhando com dados de pesquisa de opinião relativos a universidades e colégios e a extração tipificou Johns Hopkins (a universidade) como um tipo <Person> em vez de como um tipo <Organization>. Nesse caso, você poderia incluir esse conceito no tipo <Organization>.

Sempre que você criar um tipo ou incluir conceitos em uma lista de termos do tipo, essas mudanças são registradas em dicionários de tipos dentro das suas bibliotecas de recursos linguísticos no Editor de Recursos. Se você desejar visualizar o conteúdo dessas bibliotecas ou fazer um número substancial de mudanças, pode preferir trabalhar diretamente no Editor de Recursos. Veja o tópico "incluindo termos" na página 186 para obter mais informações.

Para Incluir um Conceito em um Tipo

1. Na área de janela Resultados da Extração , na área de janela Dados, na caixa de diálogo Definições de Categoria ou na caixa de diálogo Definições de Cluster, selecione o(s)t conceito(s) que você deseja incluir em um tipo existente.
2. Clique com o botão direito para abrir o menu de contexto.
3. A partir dos menus, escolha **Editar** > **Adicionar ao Tipo**. O menu exibe um conjunto de tipos com os mais recentemente criados no topo da lista. Selecione o nome do tipo ao qual deseja incluir o(s) conceito(s) selecionado(s). Se você vir o nome do tipo que está procurando, selecione-o e o(s) conceito(s) selecionado(s) será(ão) incluído(s) em tal tipo. Se você não o vir, selecione **Mais** para exibir a caixa de diálogo Todos os Tipos.
4. Na caixa de diálogo Todos os Tipos, é possível classificar a lista pela classificação natural (ordem de criação) ou na ordem crescente ou decrescente. Selecione o nome do tipo no qual você deseja incluir o(s) conceito(s) e clique em **OK**. A caixa de diálogo é fechada e eles são incluídos como termos no tipo.

Para Criar um Novo Tipo

1. Na área de janela Resultados da Extração , na área de janela Dados, na caixa de diálogo Definições de Categoria ou na caixa de diálogo Definições de Cluster, selecione os conceitos para os quais você deseja criar um novo tipo.
2. A partir dos menus, escolha **Editar** > **Adicionar ao Tipo** > **Novo**. A caixa de diálogo Propriedades de Tipo é aberta.
3. Insira um novo nome para esse tipo na caixa de texto Nome e faça quaisquer mudanças nos outros campos. Consulte o tópico "Criando tipos" na página 185 para obter informações adicionais.
4. Clique em **OK** para aplicar suas mudanças. A caixa de diálogo fecha e a cor de trecho Resultados da Extração muda, indicando que você precisa re-extrair para ver suas alterações. Se você tiver várias alterações, faça-as antes de se reextrair.

Excluindo conceitos da extração

Ao revisar seus resultados, ocasionalmente você pode localizar conceitos que não desejava extraídos ou usados por quaisquer técnicas automáticas de construção de categorias. Em alguns casos, esses conceitos possuem uma contagem de frequência muito alta e são completamente insignificantes para sua análise. Nesse caso, você pode marcar um conceito a ser excluído da extração final. Geralmente, os conceitos que você inclui nesta lista são palavras ou frases de preenchimento usados no texto para continuidade, mas que não acrescentam nada importante e podem confundir os resultados da extração. Ao incluir conceitos no dicionário de exclusões, você pode certificar-se de que eles nunca serão extraídos.

Ao excluir conceitos, todas as variações do conceito excluído desaparecerão dos resultados da extração na próxima vez que você extrair. Se esse conceito já aparecer como um descritor em uma categoria, ele permanecerá na categoria com contagem de zero após a reextração.

Ao excluir, essas mudanças são registradas em um dicionário de exclusões no Editor de Recursos. Se você desejar visualizar todas as definições de exclusão e editá-las diretamente, pode preferir trabalhar diretamente no Editor de Recursos. Veja o tópico “Dicionários de exclusão” na página 193 para obter mais informações.

Para Excluir Conceitos

1. Na área de janela Resultados da Extração , na área de janela Dados, na caixa de diálogo Definições de Categoria ou na caixa de diálogo Definições de Cluster, selecione o(s) conceito(s) que você deseja excluir da extração.
2. Clique com o botão direito para abrir o menu de contexto.
3. Selecione **Excluir da Extração**. O conceito é adicionado ao dicionário de exclusão no Editor de Recursos e nas alterações de cor de fundo de janela de Extração de Extração, indicando que você precisa re-extrair para ver suas alterações. Se você tiver várias alterações, faça-as antes de se reextrair.

Nota: Quaisquer palavras que você excluir serão automaticamente armazenadas na primeira biblioteca listada na árvore de bibliotecas no Editor de Recursos – por padrão, essa é a **Biblioteca Local**.

Forçando palavras na extração

Ao revisar os dados de texto na área de janela Dados após a extração, você pode descobrir que algumas palavras ou frases não foram extraídas. Muitas vezes, essas são palavras são verbos ou adjetivos nos quais você não está interessado. Entretanto, algumas vezes, você deseja usar uma palavra ou frase que não foi extraída como parte de uma definição de categoria.

Se você gostaria de ter essas palavras e frases extraídas, pode forçar um termo em uma biblioteca de tipos. Consulte o tópico “Forçando termos” na página 188 para obter informações adicionais.

Importante! Marcar um termo em um dicionário como forçado não é infalível. Por isto, queremos dizer que embora você tenha explicitamente incluído um termo em um dicionário, há situações em que ele pode não estar presente na área de janela Resultados da Extração após ter extraído novamente ou ele aparece, mas não exatamente como você o declarou. Embora essa ocorrência seja rara, ela pode acontecer quando uma palavra ou frase já foi extraída como parte de uma frase mais longa. Para evitar isso, aplique a opção de correspondência **Inteiro (não compostos)** a este termo no dicionário de tipos. Veja o tópico “incluindo termos” na página 186 para obter mais informações.

Capítulo 9. Categorizando dados de texto

Na visualização Categorias e Conceitos , é possível criar *categorias* que representam, em essência, conceitos ou tópicos de nível superior, que capturarão as ideias-chave, os conceitos e as atitudes expressos no texto.

A partir da liberação do IBM SPSS Modeler Text Analytics 14, as categorias também podem ter uma estrutura hierárquica, o que significa que elas contêm subcategorias e essas também podem ter suas próprias subcategorias e assim por diante. É possível importar estruturas de categoria predefinidas, anteriormente denominadas estruturas de código, com categorias hierárquicas, bem como construir estas categorias hierárquicas dentro do produto.

Essencialmente, as categorias hierárquicas permitem que você construa uma estrutura em árvore com uma ou mais subcategorias para agrupar itens como áreas de conceito ou tópico com mais precisão. Um exemplo simples pode ser relacionado a atividades de lazer; ao responder a uma pergunta como *Qual atividade que você gostaria de fazer se tivesse mais tempo?* você pode ter categorias principais como *sports*, *art and craft*, *fishing* e assim por diante; um nível abaixo, sob *sports*, você pode ter subcategorias para ver se isso é *ball games*, *water-related* e assim por diante.

As categorias são compostas de um conjunto de descritores, tais como *conceitos*, *tipos*, *padrões* e *regras de categoria*. Juntos, esses descritores são usados para identificar se um documento ou registro pertence ou não a uma determinada categoria. O texto em um documento ou registro pode ser varrido para ver se algum texto corresponde a um descritor. Se uma correspondência for localizada, o documento/registro será designado a tal categoria. Esse processo é chamado de *categorização*.

Você pode trabalhar com, construir e explorar visualmente suas categorias usando os dados apresentados nas quatro áreas de janela da visualização Categorias, cada uma das quais pode ser oculta ou mostrada ao selecionar seu nome no menu Visualizar.

- **Categorias de janela.** Construa e gerencie suas categorias nessa área de janela. Veja o tópico “A área de janela Categorias” na página 92 para obter mais informações.
- **Área de janela Resultados da Extração.** Explore e trabalhe com os conceitos e tipos extraídos nessa área de janela. Veja o tópico “Resultados da extração: conceitos e tipos” na página 79 para obter mais informações.
- **Painel de visualização.** Explore visualmente suas categorias e como elas interagem nessa área de janela. Veja o tópico “Gráficos e diagramas de categoria” na página 155 para obter mais informações.
- **Painel de dados.** Explore e revise o texto contido nos documentos e registros que correspondem às seleções nessa área de janela. Veja o tópico “A área de janela Dados” na página 99 para obter mais informações.

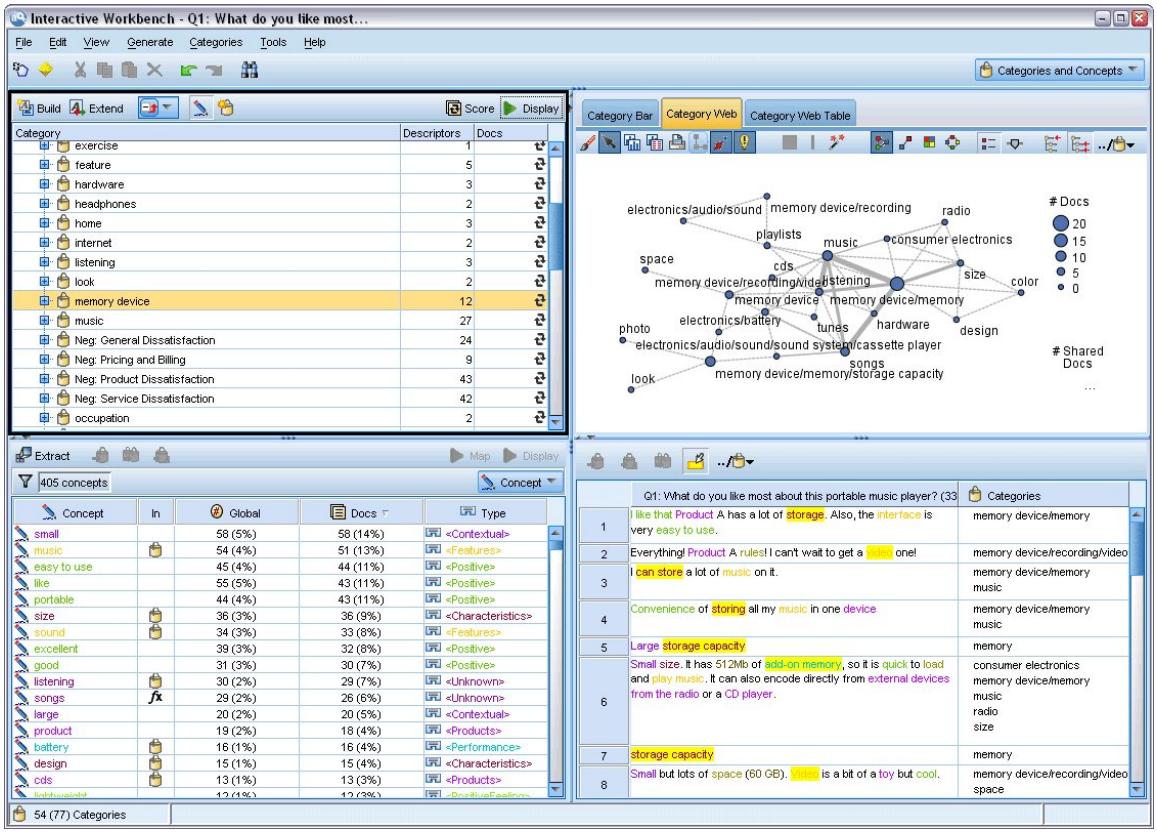


Figura 29. Visualização Categorias e Conceitos

Embora você possa iniciar com um conjunto de categorias a partir de um pacote de análise de texto (TAP) ou importar a partir de um arquivo de categoria predefinido, também pode ser necessário criar suas próprias. As categorias podem ser criadas automaticamente usando o conjunto robusto de técnicas automatizadas do produto que usam os resultados da extração (conceitos, tipos e padrões) para gerar categorias e seus descritores. As categorias também pode ser criadas manualmente usando insight adicional que você pode ter sobre os dados. Entretanto, você pode apenas criar categorias manualmente ou realizar ajuste fino nelas através do ambiente de trabalho interativo. Veja o tópico "Nó Mineração de Texto: guia Modelo" na página 23 para obter mais informações. É possível criar definições de categoria manualmente ao arrastar e soltar os resultados da extração nas categorias. É possível enriquecer essas categorias ou qualquer categoria vazia ao incluir regras de categoria em uma categoria, usar suas próprias categorias predefinidas ou uma combinação.

Cada uma das técnicas é bem adequada para certos tipos de dados e situações, mas muitas vezes será útil combinar técnicas na mesma análise para capturar o intervalo completo de documentos ou registros. E durante a categorização, você pode ver outras mudanças a fazer nos recursos linguísticos.

A área de janela Categorias

A área de janela Categorias é a área na qual você pode construir e gerenciar suas categorias. Essa área de janela está localizada no canto superior esquerdo da visualização Categorias e Conceitos. Depois de extrair os conceitos e tipos a partir de seus dados de texto, você pode começar a construir categorias automaticamente usando técnicas como inclusão de conceito, co-ocorrência, e assim por diante ou manualmente. Consulte o tópico "Construindo categorias" na página 102 para obter mais informações.

Cada vez que uma categoria é criada ou atualizada, os documentos ou registros podem ser escorados ao clicar no botão **Escorar** ver se algum texto corresponde a um descritor em uma determinada categoria. Se uma correspondência for localizada, o documento ou registro será designado a tal categoria. O resultado final é que a maioria, se não todos os documentos ou registros são designados a categorias com base nos descritores nas categorias.

Nota: Se houver mais categorias que podem caber no painel visível, você pode usar os controles na parte inferior da pane para mover para frente e para trás através das categorias ou inserir um número de página para ir.

Tabela em Árvore de Categorias

A tabela em árvore nesta área de janela apresenta o conjunto de categorias, subcategorias e descritores. A árvore também tem várias colunas que apresentam informações para cada item da árvore. As colunas a seguir podem estar disponíveis para exibição:

- **Código** Lista o valor de código para cada categoria. Esta coluna está oculta por padrão. Você pode exibir esta coluna através dos menus: **View** > **Categorias Painel**.
- **Categoria.** Contém a árvore de categorias mostrando o nome da categoria e das subcategorias. Além disso, se o ícone da barra de ferramentas dos descritores for clicado, o conjunto de descritores também será exibido.
- **Descritores.** Fornece o número de descritores que compõem sua definição. Esta contagem não inclui o número de descritores na subcategorias. Nenhuma contagem é fornecida quando um nome de descritor é mostrado na coluna **Categorias**. Você pode exibir ou ocultar os próprios descritores na árvore através dos menus: **View** > **Categorias Painel** > **Todos Descritores**.
- **Docs** Após a pontuação, esta coluna fornece o número de documentos ou registros que são categorizados em uma categoria e todas as suas subcategorias. Portanto, se 5 registros corresponderem à sua categoria principal em seus descritores e 7 registros diferentes corresponderem a uma subcategoria com base em seus descritores, a contagem total de documentos para a categoria principal será a soma dos dois-- nesse caso, ele seria 12. Entretanto, se o mesmo registro correspondeu a categoria principal e sua subcategoria, então a contagem seria 11.

Quando não existem categorias, a tabela ainda conterá duas linhas. A linha superior, chamada **Todos os Documentos**, é o número total de documentos ou registros. Uma segunda linha, chamada **Não categorizados**, mostra o número de documentos/registros que ainda precisam ser categorizados.

Para cada categoria na área de janela, um pequeno ícone de lixeira amarela precede o nome da categoria. Se você clicar duas vezes em uma categoria , ou escolher **Visualizar> Definições de categoria** nos menus, a caixa de diálogo Definições de Categoria abre e apresenta todos os elementos, chamados de *descritores*, que compõem sua definição, tais como conceitos, tipos, padrões e regras de categoria. Consulte o tópico “Sobre Categorias” na página 98 para obter mais informações. Por padrão, a tabela em árvores de categorias não mostra os descritores nas categorias. Se você desejar ver os descritores diretamente na árvore em vez de na caixa de diálogo Definições de Categoria, clique no botão de alternância com o ícone de lápis na barra de ferramentas. Quando este botão de alternância é selecionado, é possível expandir sua árvore para ver também os descritores.

Escorando Categorias

A coluna **Docs**. na tabela em árvores de categorias exibe o número de documentos ou registros que são categorizados em tal categoria específica. Se os números estiverem fora de data ou não forem calculados, um ícone aparece nessa coluna. Você pode clicar em **Score** na barra de ferramentas da pane para recalcular o número de documentos. Tenha em mente que o processo de escoragem pode levar algum tempo quando você está trabalhando com conjuntos de dados maiores.

Selecionando Categorias na Árvore

Ao fazer seleções na árvore, é possível selecionar apenas categorias irmãs -- ou seja, se você selecionar categorias de nível superior, não será possível também selecionar uma subcategoria. Ou se você selecionar 2 subcategorias de uma determinada categoria, não será possível selecionar simultaneamente uma subcategoria de outra categoria. Selecionar uma categoria não contígua resultará na perda da seleção anterior.

Exibindo nas áreas de janela Dados e Visualização

Ao selecionar uma linha na tabela, você pode clicar no botão **Exibir** para atualizar as áreas de janela Visualização e Dados com informações correspondentes à sua seleção. Se uma área de janela não estiver visível, clicar em **Exibir** fará com que a área de janela apareça.

Refinando suas Categorias

A categorização pode não produzir resultados perfeitos para seus dados na primeira tentativa e também pode haver categorias que você deseja excluir ou combinar com outras categorias. Também é possível descobrir, através de uma revisão dos resultados da extração, que existem algumas categorias que não foram criadas que você pode achar útil. Se for assim, você pode fazer mudanças manuais nos resultados para realizar ajuste fino neles para seu contexto específico. Veja o tópico "Editando e refinando categorias" na página 134 para obter mais informações.

Métodos e estratégias para criar categorias

Se você ainda não tiver extraído ou se os resultados da extração estiverem desatualizados, o uso de uma das técnicas de construção ou extensão da categoria solicitará uma extração automaticamente. Depois de ter aplicado uma técnica, os conceitos e tipos que foram agrupados em uma categoria ainda estarão disponíveis para construção de categoria com outras técnicas. Isso significa que você pode ver um conceito em diversas categorias, a menos que você escolha não reutilizá-los.

Na para ajudá-lo a criar as melhores categorias, revise o seguinte:

- **Métodos para criação de categorias**
- **Estratégicas para criar categorias**
- **Dicas para criar categorias**

Métodos para criação de categorias

Como cada conjunto de dados é exclusivo, o número de métodos de criação de categoria e a ordem na qual você os aplica pode mudar ao longo do tempo. Além disso, como seus objetos de mineração de texto podem ser diferentes de um conjunto de dados para o próximo, você pode precisar experimentar com os diferentes métodos para ver qual deles produz os melhores resultados para os dados de texto determinados.Nenhuma das técnicas automáticas categorizará perfeitamente seus dados; portanto, recomendamos localizar e aplicar uma ou mais técnicas automáticas que funcionem bem com seus dados.

Além de usar os pacotes de análise de texto (TAPs, **.tap*) com conjuntos de categorias pré-construídos, você também pode categorizar suas respostas usando qualquer combinação dos métodos a seguir:

- **Técnicas de construção automática.** Várias opções de categoria baseadas em linguística e baseadas em frequência estão disponíveis para automaticamente construir categorias para você. Veja o tópico "Construindo categorias" na página 102 para obter mais informações.
- **Técnicas de extensão automáticas.** Várias técnicas linguísticas estão disponíveis para estender as categorias existentes ao incluir descritores de aprimoramento para que eles capturem mais registros. Veja o tópico "Estendendo categorias" na página 111 para obter mais informações.
- **Técnicas manuais.** Há diversos métodos manuais, tais como arrastar e soltar. Veja o tópico "Criando categorias manualmente" na página 114 para obter mais informações.

Estratégias para criação de categorias

A lista de estratégias a seguir não é, de modo algum, completa, mas pode fornecer algumas ideias sobre como abordar a construção de suas categorias.

- Ao definir o nó Mineração de Texto, selecione um conjunto de categorias em um pacote de análise de texto (TAP) para que você inicie sua análise com algumas categorias pré-construídas. Essas categorias podem categorizar suficientemente o seu texto desde o início. No entanto, se você quiser adicionar

mais categorias, você pode editar as configurações do Build Categories (**Categorias> Configurações de Construção**). Abra o diálogo **Configurações Avançadas: Linguísticas** e escolha a opção de entrada da Categoria **Resultados de extração não usados** e construa as categorias adicionais.

- Ao definir o nó, selecione um conjunto de categorias a partir de um TAP na visualização Categorias e Conceitos no Ambiente de Trabalho Interativo. Em seguida, arraste e solte os conceitos ou padrões não usados nas categorias que você julgar apropriadas. Em seguida, estenda as categorias existentes que você acabou de editar (**Categorias> Categorias de Extensão**) para obter mais descritores que estejam relacionados com os descritores de categoria existentes.
- Construir categorias automaticamente usando as configurações linguísticas avançadas (**Categorias> Categorias de Construção**). Em seguida, refine as categorias manualmente ao excluir descritores, excluir categorias ou mesclar categorias semelhantes até que você esteja satisfeito com as categorias resultantes. Além disso, se você originalmente construiu categorias **sem** usar a opção **Generalizar com curingas onde possível**, também é possível tentar simplificar as categorias automaticamente usando Estender Categorias usando a opção **Generalizar**.
- Importe um arquivo de categoria predefinido com nomes de categoria e/ou anotações bastante descritivos. Além disso, se você importou originalmente **sem** escolher a opção para importar ou gerar descritores a partir de nomes de categorias, será possível usar o diálogo Estender Categorias posteriormente e escolher a opção **Estender categorias vazias com descritores gerados a partir do nome da categoria** opção. Em seguida, estenda tais categorias uma segunda vez para usar as técnicas de agrupamento essa vez.
- Crie manualmente um primeiro conjunto de categorias ao classificar conceitos ou padrões de conceito por frequência e, em seguida, arrastar e soltar os mais interessantes na área de janela Categorias. Uma vez que você tenha esse conjunto inicial de categorias, use o recurso de Extensão (**Categorias> Categorias de Extensão**) para expandir e refinar todas as categorias selecionadas para que elas incluam outros descritores relacionados e, assim, correspondam a mais registros.

Após aplicar essas técnicas, recomendamos que você revise as categorias resultantes e use técnicas manuais para fazer pequenos ajustes, remover quaisquer classificações incorretas ou incluir registros ou palavras que podem ter sido perdidas. Além disso, como o uso de técnicas diferentes pode produzir categorias redundantes, você também pode mesclar ou excluir categorias, conforme necessário. Veja o tópico "Editando e refinando categorias" na página 134 para obter mais informações.

Dicas para criar categorias

Para ajudar a criar melhores categorias, é possível revisar algumas dicas que podem ajudá-lo a tomar decisões sobre sua abordagem.

Dicas sobre a Razão Categoria para Documento

As categorias nas quais os documentos e registros são designados não são frequentemente mutuamente exclusivas na análise de texto qualitativa por pelo menos por duas razões:

- Primeiro, uma regra básica geral diz que quanto mais longo o texto documento ou registro, mais distintas serão as ideias e opiniões expressas. Assim, a probabilidade de que um documento ou registro possa ser designado a diversas categorias aumenta muito.
- Em segundo lugar, frequentemente, há várias maneiras de agrupar e interpretar texto documentos ou registros que não são logicamente separados. No caso de uma pesquisa de opinião com uma pergunta em aberto sobre crenças políticas do respondente, poderíamos criar categorias, tais como *Liberal* e *Conservative* ou *Republican* e *Democrat*, bem como categorias mais específicas, tais como *Socially Liberal*, *Fiscally Conservative* e assim por diante. Estas categorias não precisam ser mutuamente exclusivas e completas.

Dicas sobre o Número de Categorias a Criar

A criação da categoria deve fluir diretamente a partir dos dados — conforme você vê algo interessante com relação aos seus dados, pode criar uma categoria para representar tais informações. Em geral, não há limite superior recomendado no número de categorias que você cria. Entretanto, certamente é possível criar muitas categorias para serem gerenciáveis. Dois princípios se aplicam:

- **Frequência de categoria.** Para uma categoria ser útil, ela deve conter um número mínimo de documentos ou registros. Um ou dois documentos podem incluir algo bastante intrigante, mas se eles são um ou dois de 1.000 documentos , as informações que eles contêm podem não ser frequentes o suficiente na população para serem úteis na prática.
- **Complexidade.** Quanto mais categorias você cria, mais informações terá para revisar e resumir após a conclusão da análise. Entretanto, muitas categorias, embora incluam complexidade, podem não incluir detalhes úteis.

Infelizmente, não há regras para determinar quantas categorias são consideradas muitas ou para determinar o número mínimo de registros por categoria. Você terá que fazer tais determinações com base nas demandas da sua situação específica.

Entretanto, podemos oferecer aconselhamento sobre onde iniciar. Embora o número de categorias não deva ser excessivo, nos primeiros estágios da análise é melhor ter muitas em vez de poucas categorias. É mais fácil agrupar categorias que são relativamente semelhantes do que separar casos em novas categorias, portanto, uma estratégia de trabalhar a partir de mais para menos categorias é geralmente a melhor prática. Dada a natureza iterativa de mineração de texto e a facilidade com a qual ela pode ser conseguida com esse programa de software, a construção de mais categorias é aceitável no início.

Escolhendo os melhores descritores

As informações a seguir contêm algumas diretrizes para escolher ou tornar os melhores descritores (conceitos, tipos, padrões de TLA e regras de categoria) para suas categorias. Os descritores são os blocos de construção das categorias. Quando algum ou todo o texto em um documento ou registro corresponde a um descritor, o documento ou registro é correspondido com a categoria.

A menos que um descritor contenha ou corresponda a um conceito ou padrão extraído, ele não será correspondido a quaisquer documentos ou registros. Portanto, use conceitos, tipos, padrões e regras de categoria, conforme descrito nos parágrafos a seguir.

Como os conceitos representam não apenas eles próprios, mas também um conjunto de termos subjacentes que podem variar de formas plurais/singulares, a sinônimos, até variações de ortografia, apenas o conceito em si deve ser usado como um descritor ou como parte de um descritor. Para saber mais sobre os termos subjacentes para qualquer conceito determinado, clique no nome do conceito na área de janela Resultados da Extração da visualização Categorias e Conceitos. Ao passar o mouse sobre o nome do conceito, uma dica de ferramenta aparece e exibe qualquer um dos termos subjacentes localizados em seu texto durante a última extração. Nem todos os conceitos possuem termos subjacentes. Por exemplo, se `car` e `vehicle` forem sinônimos, mas `car` foi extraído como o conceito com `vehicle` como um termo subjacente, então você desejar usar apenas `car` em um descritor já que ele corresponderá automaticamente documentos ou registros com `vehicle`.

Conceitos e Tipos como Descritores

Use um conceito como um descritor quando você desejar localizar todos os documentos ou registros contendo tal conceito (ou qualquer um de seus termos subjacentes). Nesse caso, o uso de uma regra de categoria mais complexa não é necessário já que o nome do conceito exato é suficiente. Tenha em mente que ao usar recursos que extraem opiniões, algumas vezes, os conceitos podem mudar durante a extração de padrão de TLA para capturar o sentido mais verdadeiro da sentença (consulte o exemplo na próxima seção sobre TLA).

Por exemplo, uma resposta de pesquisa de opinião indicando as frutas favoritas de cada pessoa, tais como "*Maçã e abacaxi são as melhores*" poderia resultar na extração de `apple` e `pineapple`. Ao incluir o conceito `apple` como um descritor para sua categoria, todas as respostas que contêm o conceito `apple` (ou qualquer um de seus termos subjacentes) são correspondidas com tal categoria.

No entanto, se você estiver interessado em simplesmente saber quais respostas mencionam *apple* de qualquer maneira, você pode escrever uma regra de categoria como `* apple *` e você também irá capturar respostas que contêm conceitos como `apple`, `apple sauce`ou `french apple tart`.

Também é possível capturar todos os documentos ou registros que contêm conceitos que foram digitados da mesma maneira, usando um tipo como um descritor diretamente, tal como <`Fruit`>. Observe que não é possível usar * com tipos.

Veja o tópico “Resultados da extração: conceitos e tipos” na página 79 para obter mais informações.

Padrões de Análise de Link de Text (TLA) como Descritores

Use um resultado de padrão de TLA como um descritor quando desejar capturar ideias mais refinadas e com nuances. Quando o texto é analisado durante a extração de TLA, o texto é processado uma sentença ou cláusula de cada vez, em vez de examinar o texto inteiro (o documento ou registro). Ao considerar todas as partes de uma única sentença, a TLA pode identificar opiniões, relacionamentos entre dois elementos ou uma negação, por exemplo, e entender o sentido mais verdadeiro. É possível usar padrões de conceito ou padrões de tipo como descritores. Veja o tópico “Padrões de Tipo e Conceito” na página 147 para obter mais informações.

Por exemplo, se tínhamos o texto "*a sala não estava tão limpa*", os conceitos a seguir poderiam ser extraídos: `room` e `clean`. Entretanto, se a extração de TLA não foi ativada a configuração de extração, o TLA poderia detectar que `clean` foi usado de uma maneira negativa e, na verdade, corresponde a `not clean`, que é um sinônimo do conceito `dirty`. Aqui, você pode ver que o uso do conceito `clean` como um descritor por si só iria corresponder esse texto, mas também poderia capturar outros documentos ou registros mencionando limpeza. Portanto, pode ser melhor usar o padrão de conceito de TLA com `dirty` como conceito de saída, já que ele corresponderia esse texto e provavelmente seria um descritor mais apropriado.

Regras de Negócios de Categoria como Descritores

Regras de categoria são instruções que automaticamente classificam documentos ou registos em uma categoria, com base em uma expressão lógica usando conceitos, tipos e padrões extraídos, bem como operadores booleanos. Por exemplo, você poderia gravar uma expressão que significa *incluir todos os registros que contêm o conceito extraído* `embassy` *, mas não* `argentina` *, nesta categoria*.

Você pode escrever e usar regras de categoria como descritores em suas categorias para expressar várias ideias diferentes usando `&`, `|` e `!()` Booleanos. Para obter informações detalhadas sobre a sintaxe dessas regras e como gravar e editá-las, veja “Usando regras de categoria” na página 115.

- Use uma regra de categoria com o operador booleano `&` (AND) para ajudá-lo a localizar documentos ou registros nos quais ocorrem 2 ou mais conceitos. Os 2 ou mais conceitos conectados pelos operadores `&` não precisam ocorrer na mesma sentença ou frase, mas podem ocorrer em qualquer lugar no mesmo documento ou registro para serem considerados uma correspondência para a categoria. Por exemplo, se você criar a regra de categoria `food & cheap` como um descritor, ela corresponderia a um registro contendo o texto, "*a comida foi bastante cara, mas os quartos foram baratos*" apesar do fato de que `food` não era o substantivo sendo chamado com `cheap`, já que o texto continha ambos `food` e `cheap`.
- Use uma regra de categoria com o operador booleano `!()` (NOT) como um descritor para ajudá-lo a encontrar documentos ou registros em que algumas coisas ocorrem mas outras não. Isso pode ajudar a evitar o agrupamento de informações que podem parecer relacionadas com base em palavras, mas não no contexto. Por exemplo, se você criar a regra da categoria `<Organization> & !(ibm)` como um descritor, ela corresponderia ao texto a seguir *SPSS Inc. foi uma empresa fundada em 1967* e não corresponde ao texto a seguir *a empresa de software foi adquirida pela IBM.*.
- Use uma regra de categoria com o operador booleano `|` (OR) como um descritor para ajudá-lo a localizar documentos ou registros contendo um dos vários conceitos ou tipos. Por exemplo, se você criar a regra da categoria `(personnel|staff|team|coworkers) & bad` como um descritor, ela corresponderia a quaisquer documentos ou registros em que qualquer um desses substantivos seja encontrado com o conceito `bad`.
- Use tipos nas regras de categoria para torná-los mais genéricos e possivelmente mais implementáveis. Por exemplo, se você estava trabalhando com dados de hotel, você pode estar muito interessado em saber o que os clientes pensam sobre a equipe do hotel. Termos relacionados podem incluir palavras como recepcionista, garçom, garçonete, balcão de recepção, recepção e assim por diante. Você poderia, neste caso, criar um novo tipo chamado `<HotelStaff>` e incluir todos os anteriores termos nesse tipo. Enquanto é possível criar uma regra de categoria para todo tipo de equipe como `[* waitress * & nice]`, `[* desk * & friendly]`, `[* receptionist * & accommodating]`, você poderia criar uma regra de categoria única e mais genérica usando o tipo `<HotelStaff>` para capturar todas as respostas que tenham opiniões favoráveis da equipe do hotel na forma de `[<HotelStaff> & <Positive>]`.

Nota: é possível usar ambos + e & nas regras de categoria ao incluir padrões de TLA em tais valores. Consulte o tópico "Usando padrões de TLA nas regras de categoria" na página 118 para obter mais informações.

Exemplo de como conceitos, TLA ou regras de categoria como descritores correspondem de maneira diferente

O exemplo a seguir demonstra como usar um conceito como um descritor, uma regra de categoria como um descritor ou usar um padrão de TLA como um descritor afeta como os documentos ou registros são categorizados. Digamos que você tinha o 5 registros a seguir.

- A: *"fantástica equipe do restaurante, comida excelente e quartos confortáveis e limpos."*
- B: *"a equipe do pessoal foi horrível, mas os quartos estavam limpos."*
- C: *"Quartos confortáveis e limpos."*
- D: *"Meu quarto não estava tão limpo."*
- E: *"Limpo."*

Como os registros incluem a palavra *limpo* e você deseja capturar essas informações, poderia cria um dos descritores mostrados na tabela a seguir. Com base na essência que você está tentando capturar, você pode ver como usar um tipo de descritor sobre outro pode produzir resultados diferentes.

Tabela 17. Como registros de exemplo corresponderam descritores

Descritor	A	B	C	D	E	Explicação
`clean`	*match*	*match*	*match*	*match*	*match*	O descritor é um conceito extraído. Cada registro continha o conceito `clean`, até mesmo o registro D, já que que sem a TLA, não se sabe automaticamente que "*not clean*" significa `dirty` de acordo com as regras de TLA.
`clean + .`	-	-	-	-	*match*	Descritor é um padrão de TLA que representa `clean` por si só. Correspondido apenas com o registro no qual `clean` foi excluído sem nenhum conceito associado durante a extração de TLA.
`[clean]`	*match*	*match*	*match*	-	*match*	Descritor é uma regra de categoria que procura uma regra de TLA que contém `clean` por si só ou com outra coisa. Correspondido com todos os registros nos quais uma saída de TLA contendo `clean` foi localizada, independente do fato de que `clean` estava vinculado a outro conceito, tal como `room` e em qualquer posição de slot.

Sobre Categorias

Categorias referem-se a um grupo de conceitos, opiniões ou atitudes estritamente relacionados. Para ser útil, uma categoria também deve ser facilmente descrita por uma frase curto ou rótulo que captura seu significado essencial.

Por exemplo, se você estiver analisando respostas de pesquisa de consumidores sobre um novo sabão para roupa, é possível criar uma categoria rotulada *odor* que contém todas as respostas que descrevem o cheiro do produto. No entanto, tal categoria não pode diferenciar entre aqueles que julgaram o cheiro

agradável e aqueles que julgaram desagradável. Como o IBM SPSS Modeler Text Analytics é capaz de extrair opiniões ao usar os recursos apropriados, você pode, então, criar duas categorias para identificar respondentes que *gostaram do odor* e respondentes que *não gostaram do odor*.

É possível criar e trabalhar com suas categorias na área de janela Categorias na área de janela superior esquerda da janela visualização Categorias e Conceitos . Cada categoria é definida por um ou mais descritores. **Descritores** são conceitos, tipos e padrões, bem como regras de categoria que foram usados para definir uma categoria.

Se você desejar os descritores que compõem uma determinada categoria, é possível clicar no ícone de lápis na barra de ferramentas da área de janela Categorias e, então, expandir a árvore para ver os descritores. Alternativamente, selecione a categoria e abra a caixa de diálogo Definições de Categoria (**Visualizar> Definições de Categoria**).

Ao construir categorias automaticamente utilizando técnicas de construção de categoria como inclusão de conceito, as técnicas utilizarão conceitos e tipos como os descritores para criar suas categorias. Se você extrair padrões de TLA, y ou também pode adicionar padrões ou partes desses padrões como descritores de categoria. Veja o tópico Capítulo 11, “Explorando a análise de ligação de texto”, na página 145 para obter mais informações. E se você construir clusters, poderá incluir os conceitos em um cluster para categorias novas ou existentes. Por último, é possível criar manualmente as regras de categoria para usar como descritores em suas categorias. Veja o tópico “Usando regras de categoria” na página 115 para obter mais informações.

Propriedades da categoria

Além de descritores, as categorias também têm propriedades que você pode editar para renomear categorias, incluir um rótulo ou incluir uma anotação.

As propriedades a seguir existem:

- **NOME.** Este nome aparece na árvore por padrão. Quando uma categoria é criada usando uma técnica automatizada, ela recebe um nome automaticamente.
- **Rótulo.** O uso de rótulos é útil na criação de descrições de categoria mais significativas para uso em outros produtos ou em outras tabelas ou gráficos. Se você escolher a opção para exibir o rótulo, então, o rótulo será usado na interface para identificar a categoria.
- **Código.** O número de código corresponde ao valor de código para essa categoria. .
- **Anotação.** Você pode incluir uma descrição simples para cada categoria neste campo. Quando uma categoria é gerada pelo diálogo Construir Categorias, uma nota é incluída nesta anotação automaticamente. Você também pode adicionar texto de amostra a uma anotação diretamente da pane de Dados, selecionando o texto e escolhendo **Categorias> Adicionar à Anotação** a partir dos menus.

A área de janela Dados

Conforme você cria categorias, pode haver momentos em que você deseja revisar alguns dados de texto com os quais está trabalhando. Por exemplo, se você criar uma categoria na qual 640 documentos são categorizados, talvez você queira consultar alguns ou todos os documentos para ver qual texto foi realmente escrito. É possível revisar registros ou documentos na área de janela Dados, que está localizada no lado inferior direito. Se não visível por padrão, escolha **Visualizar> Panes> Dados** a partir dos menus.

A área de janela Dados apresenta uma linha por documento ou registro correspondente à seleção na área de janela Categorias, área de janela Resultados da Extração ou caixa de diálogo Definições de Categoria até certo limite de exibição. Por padrão, o número de documentos ou registros mostrados na área de janela Dados é limitado para permitir que você veja seus dados mais rapidamente. No entanto, é possível ajustar isso na caixa de diálogo Opções. Se você estiver lidando com datasets muito grandes, a velocidade de exibição pode ser melhorada, desligando a opção para mostrar categorias. Consulte o tópico “Opções: guia Sessão” na página 73 para obter mais informações.

Nota: Se houver mais registros que podem caber na pane visível, você pode usar os controles na parte inferior da pane para mover para frente e para trás através dos registros, ou inserir um número de página para ir.

Exibindo e Atualizando a Área de Janela Dados

A área de janela Dados não atualiza sua exibição automaticamente, pois com dados automáticos de conjuntos de dados grandes, a atualização levaria muito tempo para ser concluída. Portanto, sempre que você fizer uma seleção em outra área de janela nessa visualização ou na caixa de diálogo Definições de Categoria, clique em **Exibir** para atualizar o conteúdo da área de janela Dados.

Documentos de texto ou registros

Se seus dados de texto estiverem no formato de registros e o texto for relativamente pequeno, o campo de texto na área de janela Dados exibirá os dados de texto como um todo. No entanto, quando você trabalha com registros e conjuntos de dados maiores, a coluna do campo de texto mostra um pequeno pedaço do texto e abre uma área de janela Visualização de Texto à direita para exibir mais ou todo o texto do registro que você selecionou na tabela. Se seus dados de texto estiverem no formato de documentos individuais, a área de janela Dados mostrará o nome do arquivo do documento. Quando você seleciona um documento, a área de janela Visualização de Texto é aberta com o texto do documento selecionado.

Cores e destaque

Sempre que você exibe dados, os conceitos e descritores localizados nesses documentos ou registros são destacados em cores para ajudá-lo a identificá-los facilmente no texto. A codificação de cor corresponde aos tipos aos quais os conceitos pertencem. Também é possível passar o mouse sobre os itens codificados por cores para exibir o conceito sob o qual eles foram extraído e o tipo ao qual eles foram designados. Qualquer texto não extraído aparece em preto. Normalmente, essas palavras não extraídas costumam ser conectores (*e* ou *com*), pronomes (*mim* ou *eles*) e verbos (*ser*, *ter* ou *tomar*).

Colunas da área de janela Dados

Embora a coluna do campo de texto esteja sempre visível, também é possível exibir outras colunas. Para exibir outras colunas, escolha **Visualizar** > **Pane de Dados** a partir dos menus e, em seguida, selecione a coluna que deseja exibir no painel de Dados. As colunas a seguir podem estar disponíveis para exibição:

- **"Nome do campo de texto" (#)/Documentos.** Inclui uma coluna para os dados de texto dos quais conceitos e tipos foram extraídos. Se seus dados estiverem em documentos, a coluna será chamada Documentos e somente o nome do arquivo ou caminho completo do documento ficará visível. Para ver o texto para esses documentos, deve-se consultar a área de janela Visualização de Texto. O número de linhas na área de janela Dados é mostrado entre parênteses após o nome dessa coluna. Pode haver momentos em que nem todos os documentos ou registros são mostrados devido a um limite no diálogo Opções usado para aumentar a velocidade do carregamento. Se o máximo for atingido, o número será seguido por **-Max**. Consulte “Opções: guia Sessão” na página 73 para obter mais informações.
- **Categorias.** Lista cada uma das categorias às quais um registro pertence. Sempre que essa coluna é mostrada, a atualização da área de janela Dados pode demorar um pouco mais para mostrar as informações mais atuais.
- **Força Em.** Lista as categorias em que você forçou um documento. Os documentos podem ser forçados na categoria por meio da seleção de menu **Editar** > **Força In** . Consulte “Forçando documentos em categorias” na página 136 para obter mais informações.
- **Força Para Fora.** Lista as categorias a partir das quais você removeu um documento. Os documentos podem ser forçados a sair de uma categoria através da seleção de menu **Edit** > **Force Out** . Por exemplo, isso pode ser usado quando o sarcasmo de um respondente faz com que uma resposta seja mal categorizada. Consulte “Forçando documentos em categorias” na página 136 para obter mais informações.
- **Conta de categoria.** Lista o número de categorias às quais o registro pertence.
- **Classificações De Relevância.** Fornece um ranqueamento para cada registro em uma única categoria. Esse ranqueamento mostra o grau de ajuste do registro à categoria em comparação com os outros registros nessa categoria. Selecione uma categoria na área de janela Categorias (área de janela superior

esquerda) para ver o ranqueamento. Consulte o “Relevância da categoria” na página 101 para obter mais informações.

- **Sinalizadores De Resposta.** Adicio uma coluna que mostra quaisquer sinalizadores que você pode estar usando. Clique dentro desta coluna para alterar o tipo de sinalização que você atribui aos documentos. Você pode sinalizar documentos com uma bandeira "completa" ou uma sinalização "importante", ou remover bandeiras. Isso é útil para revisar a integralidade de um modelo de categoria. Consulte o “Respostas sinalizadoras” na página 101 para obter mais informações.

Relevância da categoria

Para ajudá-lo a construir categorias melhores, você pode revisar a relevância dos documentos ou registros em cada categoria, bem como a relevância de todas as categorias às quais um documento ou registro pertence.

Relevância de uma Categoria para um Registro

Sempre que um documento ou registro aparece na área de janela Dados, todas as categorias às quais ele pertence são listadas na coluna Categorias. Quando um documento ou registro pertence a diversas categorias, as categorias nesta coluna aparecem na ordem da correspondência mais relevante para a menos relevante. A categoria lista primeiro é considerada como a melhor correspondência para este documento ou registro. Veja o tópico “A área de janela Dados” na página 99 para obter mais informações.

Relevância de um Registro para uma Categoria

Ao selecionar uma categoria, você pode revisar a relevância de cada um de seus registros na coluna Classificação de Relevância na área de janela Dados. Esta classificação de relevância indica quão bem o documento ou registro se encaixa na categoria selecionada em comparação aos outros registros nessa categoria. Para ver a classificação dos registros para uma única categoria, selecione essa categoria na área de janela Categorias (área de janela superior esquerda) e a classificação para o documento ou registro aparece na coluna. Essa coluna não está visível por padrão, mas é possível escolher exibi-la. Veja o tópico “A área de janela Dados” na página 99 para obter mais informações.

Quanto menor o número da classificação do registro, melhor o ajuste ou mais relevante é esse registro para a categoria selecionada, de modo que 1 é o melhor ajuste. Se mais de um registro possuir a mesma relevância, cada um aparecerá com a mesma classificação seguida de um sinal de igual (=) para denotar que eles têm a mesma relevância. Por exemplo, você pode ter as classificações a seguir 1=, 1=, 3, 4 e assim por diante, o que significa que há dois registros que são igualmente considerados como as melhores correspondências para essa categoria.

Dica: você poderia incluir o texto do registro mais relevante na anotação da categoria para ajudar a fornecer uma melhor descrição da categoria. Adicio o texto diretamente da pane de Dados, selecionando o texto e escolhendo **Categorias> Adicionar à Anotação** dos menus.

Respostas sinalizadoras

Para ajudá-lo a monitorar o seu progresso, você pode marcar documentos usando sinalizadores no painel de Dados. Este recurso só está disponível se o documento de origem contiver um ID exclusivo. Se o documento de origem não contiver um ID exclusivo, você poderá adicionar um nó de Derivação entre o documento de origem e o nó Mining de Texto.

Há muitas razões pelas quais você pode querer marcar um documento, incluindo:

- Para marcar os documentos que você revisou manualmente para que você saiba onde buscar mais tarde
- Para marcar de fora um documento que você está incerto sobre como tratar

Uma vez que você marca um documento com uma bandeira, você pode continuar a trabalhar com os documentos. Eles são puramente para a sua própria gravadora. Você pode escolher entre as seguintes bandeiras:

Tabela 18. Descrições de sinalização

Sinalização	Descrição
	Sinalização completa para denotar documentos que você julgar terminado.
	Bandeira importante para denotar documentos que julgar importantes.

Para marcar um documento com uma bandeira:

1. De dentro do painel de Dados, clique com o botão direito do mouse no documento que você deseja marcar.
2. No menu de contexto, escolha **Visualizar** > **Painel de Dados** > **Sinalizadores de Resposta** e, em seguida, selecione o tipo de sinalização que você deseja usar (Importante Flag ou Complete Flag). A sinalização selecionada é atribuída. Se a coluna Bandeira no painel de Dados não estiver visível, ela aparece.

Para limpar bandeiras:

1. De dentro do painel de Dados, clique com o botão direito do mouse sobre os documentos para os quais deseja remover uma bandeira.
2. No menu de contexto, escolha **Respostas Mark With** > **Clear Flags**. As bandeiras selecionadas são removidas.

Construindo categorias

Embora você possa ter categorias de um pacote de análise de texto, também é possível construir categorias automaticamente usando uma série de técnicas linguísticas e de frequência. Através da caixa de diálogo Construir Configurações de Categorias, você pode aplicar as técnicas automatizadas de linguística e de frequência para produzir categorias a partir de conceitos ou padrões de conceito.

Em geral, as categorias podem ser compostas de diferentes tipos de descritores (tipos, conceitos, padrões de TLA, regras de categoria). Ao construir categorias usando as técnicas de construção de categoria automatizadas, as categorias resultante recebem o nome de um conceito ou padrão de conceito (dependendo da entrada selecionada) e cada uma contém um conjunto de descritores. Esses descritores pode estar na forma de regras de categoria ou conceitos e incluem todos os conceitos relacionados descobertos pelas técnicas.

Depois de construção de categorias, você pode aprender muito sobre as categorias ao revisá-las na área de janela Categorias ou explorá-las por meio de gráficos e diagramas. Você pode, então, usar técnicas manuais para fazer pequenos ajustes, remova quaisquer classificações incorretas ou incluir registros ou palavras que podem ter sido perdidos. Depois de ter aplicado uma técnica, os conceitos, tipos e padrões que foram agrupados em uma categoria ainda estarão disponíveis para outras técnicas. Além disso, como o uso de técnicas diferentes também pode produzir categorias redundantes ou inapropriadas, também é possível mesclar ou excluir categorias. Veja o tópico “Editando e refinando categorias” na página 134 para obter mais informações.

Importante! Em liberações anteriores, regras de coocorrência e sinônimos eram cercadas por colchetes. Nesta liberação, os colchetes agora indicam um resultado de padrão análise de link de texto. Em vez disso, as regras de co-ocorrência e de sinônimo serão encapsuladas por parênteses como `(speaker systems|speakers)`.

Para Construir Categorias

1. A partir dos menus, escolha **Categorias> Categorias Construir**. A menos que você tenha escolhido nunca avisar, será exibida uma caixa de mensagens.
2. Escolha se você deseja construir agora ou editar as configurações primeiro.

- Clique em **Construir Agora** para começar a construir categorias usando as configurações atuais. As configurações selecionadas por padrão são geralmente suficientes para começar o processo de categorização. O processo de construção de categoria inicia e um diálogo de progresso é exibido.
- Clique em **Editar** para revisar e modificar as configurações de construção.

Nota: O número máximo de categorias que podem ser exibidas é de 10.000. Um aviso é exibido se este número for atingido ou excedido. Se isto acontecer, você deve mudar suas opções Construir ou Estender Categorias para reduzir o número de categorias construídas.

Entradas

As categorias são construídas a partir de descritores derivados de padrões de tipo ou tipos. Na tabela, você pode selecionar os tipos ou padrões individuais para incluir no processo de construção da categoria.

Padrões de tipos. Se você selecionar padrões de tipo, as categorias são construídas a partir de padrões em vez de tipos e conceitos por conta própria. Desta forma, quaisquer registros ou documentos contendo um padrão de conceito pertencente ao padrão de tipo selecionado são categorizados. Assim, se você selecionar o padrão do tipo <Budget> e <Positive> na tabela, categorias como `cost & <Positive>` ou `rates & excellent` poderão ser produzidas.

Ao usar padrões de tipo como entrada para construção de categoria automatizada, há momentos em que as técnicas identificam diversas maneiras para formar a estrutura da categoria. Tecnicamente, não há uma única maneira certa de produzir as categorias; entretanto, você pode localizar uma estrutura mais adequada para sua análise do que outra. Para ajudar a customizar a saída neste caso, é possível designar um tipo como o foco preferencial. Todas as categorias de nível superior produzidas virá de um conceito do tipo selecionado aqui (e nenhum outro tipo). Cada subcategoria conterá um padrão de link de texto deste tipo. Escolha esse tipo no campo **Estruturar categorias por tipo de padrão:** e a tabela será atualizada para mostrar somente os padrões aplicáveis que contêm o tipo selecionado. Muito frequentemente, <Unknown> será pré-selecionado para você. Isso resulta em todos os padrões contendo o tipo <Unknown> sendo selecionados. A tabela exibe os tipos em ordem decrescente começando por aquele com o maior número de registros ou documentos (**Doc.**).

Tipos. Se você selecionar tipos, as categorias serão construídas a partir de conceitos pertencentes aos tipos selecionados. Portanto, se você selecionar o tipo <Budget> na tabela, categorias como `cost` ou `price` podem não ser produzidas, já que `cost` e `price` são conceitos designados ao tipo <Budget>.

Por padrão, apenas os tipos que capturam a maioria dos registros ou documentos são selecionados. Esta pré-seleção permite que você foque rapidamente nos tipos mais interessantes e evite construir categorias desinteressantes. A tabela exibe os tipos na ordem decrescente, iniciando com o tipo com o maior número de registros ou documentos (contagem **Doc.**). Os tipos da biblioteca `Opinions` são desmarcados por padrão na tabela de tipos.

A entrada escolhida afeta as categorias obtidas. Quando você escolhe usar Tipos como entrada, é possível ver os conceitos claramente relacionados mais facilmente. For example, if you build categories using Types as input, you could obtain a category `Fruit` with concepts such as `apple`, `pear`, `citrus fruits`, `orange` and so on. Se você escolher Padrões de Tipo como entrada em vez disso e selecionar o padrão <Unknown> + <Positive>, por exemplo, poderá obter uma categoria `fruit + <Positive>` com um ou dois tipos de frutas, tais como `fruit + tasty` e `apple + good`. Este segundo resultado mostra apenas 2 padrões de conceito, pois as outras ocorrências de frutas não são necessariamente positivamente qualificadas. E, embora isso possa ser bom o suficiente para seus dados de texto atuais, em estudos longitudinais nos quais você usa diferentes conjuntos de documentos, você pode desejar incluir manualmente em outros descritores, como `citrus fruit + positive` ou tipos de uso Usar tipos sozinhos como entrada o ajudará a localizar todas as frutas possíveis.

Técnicas

Como cada conjunto de dados é exclusivo, o número de métodos e a ordem na qual você os aplica pode mudar ao longo do tempo. Como seus objetivos de mineração de texto podem ser diferentes de um conjunto de dados para o próximo, pode ser necessário experimentar com diferentes técnicas para ver qual produz os melhores resultados para os determinados dados de texto.

Você não precisa ser um especialista nestas configurações para usá-las. Por padrão, as configurações médias mais comuns já estão selecionadas. Portanto, é possível efetuar bypass dos diálogos de configuração avançada e ir direto para a construção de suas categorias. Da mesma forma, se você fizer mudanças aqui, não é necessário voltar para o diálogo de configurações toda vez, já que as configurações mais recentes sempre são retidas.

Selecione as técnicas de linguística ou frequência e clique no botão Configurações Avançadas para exibir as configurações para as técnicas selecionadas. Nenhuma das técnicas automáticas categorizará perfeitamente seus dados; portanto, recomendamos localizar e aplicar uma ou mais técnicas automáticas que funcionem bem com seus dados. Não é possível construir usando técnicas linguísticas e de frequência simultaneamente.

- **Técnicas de linguística avançada**. Para obter mais informações, consulte "Configurações linguísticas avançadas" na página 104.
- **Técnicas de frequência avançada**. Para obter mais informações, consulte "Configurações de frequência avançadas" na página 110.

Configurações linguísticas avançadas

Ao construir categorias, é possível selecionar a partir de diversas técnicas avançadas de construção de categoria linguística como *inclusão de conceito* e *redes semânticas* (apenas texto em inglês). Estas técnicas podem ser usadas individualmente ou em combinação entre si para criar categorias.

Tenha em mente que, porque cada conjunto de dados é exclusivo, o número de métodos e a ordem na qual você os aplica pode mudar ao longo do tempo. Como seus objetivos de mineração de texto podem ser diferentes de um conjunto de dados para o próximo, pode ser necessário experimentar com diferentes técnicas para ver qual produz os melhores resultados para os determinados dados de texto.Nenhuma das técnicas automáticas categorizará perfeitamente seus dados; portanto, recomendamos localizar e aplicar uma ou mais técnicas automáticas que funcionem bem com seus dados.

As áreas e os campos a seguir estão disponíveis na caixa de diálogo Configurações Avançadas: Linguísticas:

Entrada e Saída

Entrada de categoria Selecione a partir do que as categorias serão construídas:

- **Resultados de extração não usados**. Esta opção permite que categorias sejam construídas a partir de resultados de extração que não são usado em nenhuma categoria existente. Isso minimiza a tendência dos registros corresponderem diversas categorias e limita o número de categorias produzidas.
- **Todos os resultados da extração**. Esta opção permite que as categorias sejam construídas usando qualquer um dos resultados da extração. Isso é mais útil quando nenhuma ou poucas categorias já existem.

Saída da categoria Selecione a estrutura geral para as categorias que serão construídas:

- **Hierárquica com subcategorias**. Esta opção permite a criação de subcategorias e sub-subcategorias. É possível configurar a profundidade de suas categorias ao escolher o número máximo de níveis (campo **Máximo de níveis criados**) que pode ser criado. Se você escolher 3, as categorias poderiam conter subcategorias e tais subcategorias também poderiam ter subcategorias.
- **Categorias simples (apenas nível único)**. Esta opção ativa apenas um nível de categoria a ser construído, significando que nenhuma subcategoria será gerada.

Técnicas de Agrupamento

Cada uma das técnicas disponíveis é adequada para certos tipos de dados e situações, mas geralmente é útil combinar técnicas na mesma análise para capturar o intervalo completo de documentos ou registros. Você pode ver um conceito em diversas categorias ou localizar categorias redundantes.

Inclusão de conceito. Esta técnica constrói categorias agrupando conceitos multitermos (palavras compostas), dependendo se elas contêm palavras que são subconjuntos ou superconjuntos de uma

palavra na outra. Por exemplo, o conceito `seat` seria agrupado com `safety seat`, `seat belte` `seat belt buckle`. Consulte o tópico "Inclusão de conceito" na página 108 para obter mais informações.

Rede Semântica. Esta técnica começa identificando os possíveis sentidos de cada conceito abrangente a partir de seu índice extensivo de relacionamentos de palavras e, em seguida, cria categorias ao agrupar conceitos relacionados. Esta técnica é melhor quando os conceitos são conhecidos para a rede semântica e não são ambíguos. Ela é menos útil quando o texto contém terminologia especializada ou jargão desconhecido para a rede. Em um exemplo, o conceito `granny smith apple` poderia ser agrupado com `gala apple` e `winesap apple` já que eles são irmãos da graninha smith. Em outro exemplo, o conceito `animal` pode ser agrupado com `cat` e `kangaroo` já que eles são hiponyms de `animal`. Esta técnica está disponível somente para texto em inglês nesta liberação. Consulte o tópico "Redes semânticas" na página 108 para obter mais informações.

Nota: A opção **distância máxima de busca** só está disponível se você selecionar **Rede Semanética**.

distância máxima de pesquisa Selecione o quão longe você quer que as técnicas pesquisem antes de produzir categorias. Quanto menor o valor, menos resultados você obterá - entretanto, esses resultados serão menos ruidosos e terão mais probabilidade de estarem significativamente vinculados ou associados entre si. Quanto mais alto o valor, mais resultados você pode obter - entretanto, esses resultados podem ser menos confiáveis ou relevantes. Embora esta opção seja globalmente aplicada em todas as técnicas, seu efeito é maior em coocorrências e redes semânticas.

Evitar emparelhamento de conceitos específicos. Selecione esta caixa de seleção para impedir que o processo agrupe ou pareie dois conceitos na saída. Para criar ou gerenciar pares de conceito, clique em **Gerenciar Pairs ...**. Consulte o tópico "Gerenciando pares de exceção de link" na página 106 para obter mais informações.

Geralize com curingas onde possível Selecione esta opção para permitir que o produto gere regras genéricas em categorias usando o curinga asterisco. Por exemplo, em vez de produzir vários descritores como `[apple tart + .]` e `[apple sauce + .]`, usar curingas pode produzir `[apple * + .]`. Se você generalizar com curingas, com frequência obterá exatamente o mesmo número de registros ou documentos como anteriormente. Entretanto, esta opção tem a vantagem de reduzir o número e simplificar descritores de categoria. Além disso, esta opção aumenta a capacidade de categorizar mais registros ou documentos usando estas categorias nos novos dados de texto (por exemplo, em estudos longitudinais/de onda).

Outras Opções para Construir Categorias

Além de selecionar as técnicas de agrupamento a aplicar, você pode editar várias outras opções de construção, conforme a seguir:

Número máximo de nível superior categorias criadas. Use esta opção para limitar o número de categorias que podem ser geradas quando você clica no botão próximo a Construir Categorias. Em alguns casos, é possível obter melhores resultados ao configurar esse valor alto e, em seguida, excluir qualquer uma das categorias não interessantes.

Número mínimo de descritores e / ou subcategorias por categoria. Use esta opção para definir o número mínimo de descritores e subcategorias que uma categoria precisa conter para ser criada. Esta opção ajuda a limitar a criação de categorias que não capturam um número significativo de registros ou documentos.

Permitir que descritores apareçam em mais de uma categoria Quando selecionado, esta opção permite que os descritores sejam usados em mais de uma das categorias que serão construídas em seguida. Esta opção geralmente é selecionada, pois os itens comumente ou "naturalmente" se enquadram em duas ou mais categorias e permitir que eles façam isso normalmente leva a categorias de qualidade superior. Se você não selecionar essa opção, reduz a sobreposição de registros em diversas categorias e dependendo do tipo de dados que você tem, isso pode ser desejável. Entretanto, com a maioria dos tipos de dados, restringir descritores para uma única categoria geralmente resulta em uma perda de qualidade ou cobertura da categoria. Por exemplo, digamos que você tivesse o conceito `car seat manufacturer`. Com esta opção, esse conceito poderia aparecer em uma categoria com base no texto `car seat` e em outra com base em `manufacturer`. Mas se essa opção não for selecionada, embora você ainda possa obter ambas as categorias, o conceito `car seat manufacturer` só aparecerá como

um descritor na categoria ele melhor corresponde com base em vários fatores incluindo o número de registros em que `car seat` e `manufacturer` ocorrem cada um deles.

Resolva nomes de categoria duplicados por Selecione como manipular quaisquer novas categorias ou subcategorias cujos nomes seriam os mesmos das categorias existentes. É possível mesclar as novas categorias (e seus descritores) com as categorias existentes com o mesmo nome. Alternativamente, é possível escolher ignorar a criação de qualquer categorias se um nome duplicado for localizado nas categorias existentes.

Gerenciando pares de exceção de link

Durante a construção da categoria, armazenamento em cluster e o mapeamento de conceito, os algoritmos internos agrupam palavras por associações conhecidas. Para evitar que dois conceitos sejam pareados ou vinculados, é possível ativar esse recurso no diálogo **Construir Configurações Avançadas de Categorias**, o diálogo **Construir Clusters** e o diálogo **Configurações de Índice de Mapa de Conceito** e clique no botão **Gerenciar Pares**.

No diálogo **Gerenciar exceções de link** resultante, é possível incluir, editar ou excluir pares de conceitos. Insira um par por linha. Inserir pares aqui impedirá que o pareamento ocorra ao construir ou estender as categorias, armazenamento em cluster e mapeamento de conceito. Insira palavras exatamente como você as deseja, por exemplo, a versão acentuada da palavra não é igual à versão não acentuada da palavra.

Por exemplo, se você quisesse ter certeza de que `hot dog` e `dog` não estão agrupados, você poderia adicionar a dupla como uma linha separada na tabela.

Sobre técnicas de linguística

Quando você constrói ou te estende categorias, você pode selecionar a partir de várias técnicas avançadas de construção de categoria linguística incluindo *inclusão de conceito* e *redes semânticas* (somente em inglês). Estas técnicas podem ser usadas individualmente ou em combinação entre si para criar categorias.

Você não precisa ser um especialista nestas configurações para usá-las. Por padrão, as configurações médias mais comuns já estão selecionadas. Se você desejar, é possível efetuar o bypass deste diálogo de configuração avançada e ir direto para a construção ou extensão de suas categorias. Da mesma forma, se você fizer mudanças aqui, não é necessário voltar para o diálogo de configurações toda vez já que ele lembrará o que você usou pela última vez.

Entretanto, tenha em mente que, porque cada conjunto de dados é exclusivo, o número de métodos e a ordem na qual você os aplica pode mudar ao longo do tempo. Como seus objetivos de mineração de texto podem ser diferentes de um conjunto de dados para o próximo, pode ser necessário experimentar com diferentes técnicas para ver qual produz os melhores resultados para os determinados dados de texto.Nenhuma das técnicas automáticas categorizará perfeitamente seus dados; portanto, recomendamos localizar e aplicar uma ou mais técnicas automáticas que funcionem bem com seus dados.

As técnicas linguísticas automatizadas principais para construção de categorias são:

- **Inclusão de conceito.** Esta técnica cria categorias ao obter um conceito e localizar outros conceitos que o incluem. Veja o tópico "Inclusão de conceito" na página 108 para obter mais informações.
- **Rede semântica.** Esta técnica começa identificando os possíveis sentidos de cada conceito abrangente a partir de seu índice extensivo de relacionamentos de palavras e, em seguida, cria categorias ao agrupar conceitos relacionados. Veja o tópico "Redes semânticas" na página 108 para obter mais informações. Esta opção está disponível apenas para texto em inglês.

Derivação de raiz de conceito

A técnica de derivação de raiz de conceito cria categorias ao obter um conceito e localizar outros conceitos que estão relacionados a ele ao analisar se algum dos componentes do conceito estão morfologicamente relacionados. Um componente é uma palavra. A técnica tenta agrupar conceitos ao

examinar as terminações (sufixos) de cada componente em um conceito e localizar outros conceitos que poderiam ser derivados deles. A ideia é que quando as palavras são derivadas umas das outras, ela provavelmente compartilham ou têm um significado próximo. Para identificar as terminações, regras específicas de idioma são usadas. Por exemplo, o conceito `opportunities to advance` seria agrupado com os conceitos `opportunity for advancement` e `advancement opportunity`.

É possível usar a derivação de raiz de conceito em qualquer tipo de texto. Por si só, ela produz relativamente poucas categorias e cada uma delas tende a conter poucos conceitos. Os conceitos em cada categoria são sinônimos ou são situacionalmente relacionados. Você pode achar útil usar este algoritmo, mesmo se você estiver construindo categorias manualmente; os sinônimos localizados podem ser sinônimos de tais conceitos nos quais você está particularmente interessado.

Nota: Você pode evitar que conceitos sejam agrupados, especificando-os explicitamente. Consulte o tópico "Gerenciando pares de exceção de link" na página 106 para obter informações adicionais.

Componentização e remoção de inflexão de termos

Quando as técnicas de derivação de raiz de conceito ou de inclusão de conceito são aplicadas, os termos são, primeiro, quebrados em componentes (palavras) e, em seguida, a flexão dos componentes é removida. Quando uma técnica é aplicada, os conceitos e seus termos associados são carregados e divididos em componentes com base em separadores, tais como espaços, hifens e apóstrofos. Por exemplo, o termo `system administrator` está dividido em componentes como `{administrator, system}`.

Entretanto, algumas partes do termo original não podem ser usadas e são referidas como palavras comuns. Em inglês, alguns destes componentes que podem ser ignorados podem incluir `a`, `and`, `as`, `by`, `for`, `from`, `in`, `of`, `on`, `or`, `the`, `toe` `with`.

Por exemplo, o termo `examination of the data` tem o conjunto de componentes `{data, examination}`, e ambos `of` e `the` são considerados ignoráveis. Além disso, a ordem do componente não está em um conjunto de componentes. Dessa forma, os três termos seguintes poderiam ser equivalentes: `cough relief for child`, `child relief from a coughe` `relief of child cough` já que todos possuem o mesmo conjunto de componentes `{child, cough, relief}`. Cada vez que um par de termos é identificado como sendo equivalente, os conceitos correspondentes são mesclados para formar um novo conceito que faz referência a todos os termos.

Além disso, como os componentes de um termo podem ser flexionados, as regras específicas de idioma são aplicadas internamente para identificar termos equivalentes, independentemente da variação da inflexão, tais como formas plurais. Dessa forma, os termos `level of support` e `support levels` podem ser identificados como equivalentes, uma vez que a forma singular de de-inflected seria `level`.

Como a Derivação de Raiz de Conceito Funciona

Após os termos terem sido componentizados e a inflexão removida (veja a seção anterior), o algoritmo de derivação de raiz de conceito analisa as terminações, ou sufixos, do componente, para localizar a raiz do componente e, em seguida, agrupa os conceitos com outros conceitos que têm as mesmas raízes ou raízes semelhantes. As terminações são identificadas usando um conjunto de regras de derivação linguística específico para o idioma do texto. Por exemplo, existe uma regra de derivação para texto em inglês que indica que um componente de conceito terminando com o sufixo `ical` pode ser derivado de um conceito que tenha o mesmo radical e terminando com o sufixo `ic`. Usando esta regra (e a desinflexão), o algoritmo seria capaz de agrupar os conceitos `epidemiologic study` e `epidemiological studies`.

Uma vez que os termos já são componentizados e os componentes ignoráveis (por exemplo, `in` e `of`) foram identificados, o algoritmo de derivação de raiz conceito também seria capaz de agrupar o conceito `studies in epidemiology` com `epidemiological studies`.

O conjunto de regras de derivação de componentes foi escolhido de modo que a maioria dos conceitos agrupados por este algoritmo são sinônimos: os conceitos `epidemiologic studies`, `epidemiological studies`, `studies in epidemiology` são todos os termos equivalentes. Para aumentar a abrangência, há algumas regras de derivação que permitem que o algoritmo agrupe conceitos

que são situacionalmente relacionados. Por exemplo, o algoritmo pode agrupar conceitos, tais como `empire builder` e `empire building`.

Inclusão de conceito

A técnica de inclusão de conceito constrói categorias ao obter um conceito e usar algoritmos de série léxica, identifica os conceitos incluídos em outros conceitos. A ideia é que quando palavras em um conceito são um subconjunto de outro conceito, ele reflete um relacionamento semântico subjacente. A inclusão é uma técnica poderosa que pode ser usada com qualquer tipo de texto.

Essa técnica funciona bem em combinação com redes semântico, mas pode ser usada separadamente. A inclusão de conceito também pode fornecer melhores resultados quando os documentos ou registros contêm uma grande quantidade de terminologia ou jargão específica de domínio. Isso é especialmente verdadeiro se você tiver ajustado os dicionários antecipadamente para que os termos especiais sejam extraídos e agrupados de forma apropriada (com sinônimos).

Como a Inclusão de Conceito Funciona

Antes do algoritmo de inclusão de conceito ser aplicado, os termos são componentizados e a sua inflexão removida. Consulte o tópico "Derivação de raiz de conceito" na página 106 para obter informações adicionais. Em seguida, o algoritmo de inclusão de conceito analisa os conjuntos de componentes. Para cada conjunto de componentes, o algoritmo procura outro conjunto de componentes que é um subconjunto do primeiro conjunto de componentes.

Por exemplo, se você tiver o conceito `continental breakfast`, que tem o conjunto de componentes `{breakfast, continental}` e tiver o conceito `breakfast`, que possui o conjunto de componentes `{breakfast}`, o algoritmo poderia concluir que `continental breakfast` é um tipo de `breakfast` e agrupá-los.

Em um exemplo maior, se você tiver o conceito `seat` no painel Resultados da Extração e aplicar este algoritmo, então conceitos como `safety seat`, `leather seat`, `seat belt`, `seat belt buckle`, `infant seat carriere` `car seat laws` também seriam agrupados nessa categoria.

Como os termos já estão componentizados e os componentes que podem ser ignorados (por exemplo, `in` e `of`) foram identificados, o algoritmo de inclusão de conceito reconheceria que o conceito `advanced spanish course` inclui o conceito `course in spanish`.

Nota: É possível evitar que conceitos sejam agrupados ao especificá-los explicitamente. Veja o tópico "Gerenciando pares de exceção de link" na página 106 para obter mais informações.

Redes semânticas

Nesta liberação, a técnica de redes semânticas está disponível apenas para o texto no idioma inglês.

Esta técnica constrói categorias usando uma rede integrada de relacionamentos de palavras. Por esse motivo, essa técnica pode produzir resultados muito bons quando os termos são concretos e não são muito ambíguos. Entretanto, você não deve esperar que a técnica localize muitos links entre conceitos altamente técnicos/especializados. Ao lidar com tais conceitos, é possível constatar que as técnicas de inclusão de conceito e de derivação de raiz de conceito são mais úteis.

Como a Rede Semântica Funciona

A ideia por trás da técnica de rede técnica semântica é alavancar relacionamentos de palavras conhecidos para criar categorias de sinônimos ou hipônimos. Um **hipônimo** é quando um conceito é um tipo do segundo conceito, de tal forma que há um relacionamento hierárquico, também conhecido como um relacionamento ISA. Por exemplo, se `animal` for um conceito, então `cat` e `kangaroo` são hipônimos de `animal` já que eles são tipos de animais.

Além de relacionamentos de sinônimo e hipônimo, a técnica de rede semântica também examina links parciais e totais entre quaisquer conceitos a partir do tipo `<Location>`. Por exemplo, a técnica agrupará os conceitos `normandy`, `provence` e `france` em uma categoria, já que Normandia e Provença são partes da França.

Redes semânticas começam identificando os possíveis sentidos de cada conceito na rede semântica. Quando os conceitos são identificados como sinônimos ou hipônimos, eles são agrupados em uma única categoria. Por exemplo, a técnica criaria uma única categoria contendo esses três conceitos: `eating apple`, `dessert applee granny smith` já que a rede semântica contém a informação de que: 1) `dessert apple` é um sinônimo de um `eating apple`, e 2) `granny smith` é uma espécie de `eating apple` (significando que é um hiponym de `eating apple`).

Usados individualmente, muitos conceitos, especialmente unitermos, são ambíguos. Por exemplo, o conceito `buffet` pode denotar um tipo de refeição ou uma peça de móvel. Se o conjunto de conceitos inclui `meal`, `furniture` e `buffet`, então, o algoritmo será forçado a escolher entre agrupar `buffet` com `meal` ou com `furniture`. Esteja ciente de que, em alguns casos, as opções feitas pelo algoritmo não podem ser apropriadas no contexto de um conjunto específico de registros ou documentos.

A técnica de rede semântico pode superar a inclusão de conceito com determinados tipos de dados. Enquanto tanto a rede semântica quanto a inclusão conceito reconhecem que `apple pie` é uma espécie de `pie`, apenas a rede semântica reconhece que `tart` também é uma espécie de `pie`.

As redes semânticas trabalharão em conjunto com as outras técnicas. Por exemplo, suponha que você tenha selecionado ambas as técnicas de rede semântica e inclusão e que a rede semântica agrupou o conceito `teacher` com o conceito `tutor` (porque um tutor é um tipo de professor). O algoritmo de inclusão pode agrupar o conceito `graduate tutor` com `tutor` e, como resultado, os dois algoritmos colaboram para a produção de uma categoria de saída contendo todos os três conceitos: `tutor`, `graduate tutore teacher`.

Opções para Rede Semântica

Há um número de configurações adicionais que podem ser interessantes com essa técnica.

- Mude a **Distância máxima da procura**. Selecione até onde você deseja que as técnicas procurem antes de produzir categorias. Quanto menor o valor, menos resultados são produzidos — entretanto, estes resultados serão menos ruidosos e terão mais probabilidade de estarem significativamente vinculados ou associados entre si. Quanto mais alto o valor, mais resultados você obterá - entretanto, esses resultados podem ser menos confiáveis ou relevantes.

Por exemplo, dependendo da distância, o algoritmo procura desde `Danish pastry` até `coffee roll` (seu pai) e, então bun (avô) e para cima até `bread`.

Ao reduzir a distância da procura, essa técnica produz categorias menores que podem ser mais fáceis de trabalhar se você sentir que as categorias sendo produzidas são muito grandes ou agrupam muitas coisas.

Importante! Além disso, recomendamos que você não aplique a opção **Acomodar erros de ortografia para um limite mínimo de caracteres raiz de** (definido na guia Especialista do nó ou na caixa de diálogo Extrair) para agrupamento difuso ao usar essa técnica, já que alguns agrupamentos falsos podem ter um impacto muito negativo nos resultados.

Regras de coocorrência

As regras de coocorrência permitem que você descubra e agrupe conceitos que estão fortemente relacionadas dentro do conjunto de documentos ou registros. A ideia é que quando os conceitos são frequentemente encontrados juntos em documentos e os registros, essa coocorrência reflete um relacionamento subjacente que é provavelmente de valor em suas definições de categoria. Esta técnica cria regras de coocorrência que podem ser usadas para criar uma nova categoria, estender uma categoria ou como entrada para outra técnica de categoria. Dois conceitos coocorrem fortemente se eles frequentemente aparecem juntos em um conjunto de registros e raramente separadamente em qualquer um dos outros registros. Essa técnica pode produzir bons resultados com conjuntos de dados maiores com pelo menos algumas centenas de documentos ou registros.

Por exemplo, se muitos registros contêm as palavras `price` e `availability`, esses conceitos poderiam ser agrupados em uma regra de co-ocorrência, `(price & available)`. Em outro exemplo, se os conceitos `peanut butter`, `jelly`, `sandwich` e aparecer com mais frequência juntos do que separados, eles seriam agrupados em uma regra de co-ocorrência de conceito `(peanut butter & jelly & sandwich)`.

Importante! Em liberações anteriores, regras de coocorrência e sinônimos eram cercadas por colchetes. Nesta liberação, os colchetes agora indicam um resultado de padrão análise de link de texto. Em vez disso, as regras de co-ocorrência e de sinônimo serão encapsuladas por parênteses como `(speaker systems|speakers)`.

Como as Regras de Coocorrência Funcionam

Esta técnica varre os documentos ou registros procurando dois ou mais conceitos que tendem a aparecer juntos. Dois ou mais conceitos coocorrem fortemente se eles frequentemente aparecem juntos em um conjunto de documentos ou registros e se eles raramente aparecem separadamente em quaisquer dos outros documentos ou registros.

Quando conceitos de coocorrência são localizados, uma regra de categoria é formada. Essas regras consistem em dois ou mais conceitos conectados usando o operador booleano &. Essas regras são instruções lógicas que classificarão automaticamente um documento ou registro em uma categoria se o conjunto de conceitos na regra ocorrer todo nesse documento ou registro.

Opções para Regras de Coocorrência

Se você estiver usando a técnica de regra de coocorrência, poderá realizar ajuste fino nas várias configurações que influenciam as regras resultantes:

- Mude a **Distância máxima da procura**. Selecione até que você deseja que a técnica procure coocorrências. Conforme você aumenta a distância de procura, o valor de similaridade mínima necessária para cada coocorrência é diminuído; como resultado, muitas regras de coocorrência podem ser produzidas, mas apenas aquelas que têm um valor de baixa similaridade frequentemente terão pouco significado. Conforme você reduz a distância da procura, o valor mínimo requerido de similaridade aumenta; como resultado, menos regras de coocorrência são produzidas, mas eles tenderão a ser mais significativas (mais fortes).
- **Número mínimo de documentos.** O número mínimo de registros ou documentos que devem conter um determinado par de conceitos para serem considerados como uma coocorrência; quanto menor o valor configurado para essa opção, mais fácil será localizar coocorrências. Aumentar o valor resulta em um número de menor de ocorrências, mas elas serão mais significativas. Como exemplo, suponha que os conceitos "apple" e "pear" sejam localizados juntos em 2 registros (e que nenhum dos dois conceitos ocorram em quaisquer outros registros). Como o **Número mínimo de documentos.** configurados como 2 (o padrão), a técnica de coocorrência criará uma regra de categoria (apple and pear). Se o valor for aumentado para 3, a regra não será mais criada.

Nota: Com pequenos datasets (< 1000 respostas) você pode não encontrar nenhuma co-ocorrências com as configurações padrão. Se sim, tente aumentar o valor da distância da procura.

Nota: É possível evitar que conceitos sejam agrupados ao especificá-los explicitamente. Veja o tópico “Gerenciando pares de exceção de link” na página 106 para obter mais informações.

Configurações de frequência avançadas

Você pode construir categorias com base em uma técnica de frequência direta e mecânica. Com essa técnica, você pode construir uma categoria para cada item (tipo, conceito ou padrão) que foi localizado acima da contagem de um determinado registro ou documento. Além disso, é possível construir uma única categoria para todos os itens que ocorrem menos frequentemente. Por contagem, nos referimos ao número de registros ou documentos que contém o conceito (e qualquer um de seus sinônimos), tipo ou padrão extraído em questão, em oposição ao número total de ocorrências no texto inteiro.

O agrupamento de itens que ocorrem frequentemente pode produzir resultados interessantes, já que ele pode indicar uma resposta comum ou significativa. A técnica é muito útil nos resultados da extração não usados após outras técnicas terem sido aplicadas. Outra aplicação é executar essa técnica imediatamente após a extração, quando não existem outras categorias, editar os resultados para excluir categorias não interessantes, e em seguida, estender essas categorias para que elas correspondam a ainda mais registros ou documentos. Veja o tópico “Estendendo categorias” na página 111 para obter mais informações.

Em vez de usar essa técnica, você poderia classificar os conceitos ou padrões de conceito pelo número decrescente de registros ou documentos na área de janela Resultados da Extração e, em seguida, arrastar e soltar os principais na área de janela Categorias para criar as categorias correspondentes.

Os campos a seguir estão disponíveis na caixa de diálogo Configurações Avançadas: Frequências:

Gerar descritores de categoria em. Selecione o tipo de entrada para os descritores. Veja o tópico "Construindo categorias" na página 102 para obter mais informações.

- **Nível de conceitos**. Selecionar esta opção significa que as frequências de conceitos ou padrões de conceito serão usadas. Os conceitos serão usados se foram selecionados tipos como entrada para construção de categorias e padrões de conceito são utilizados se foram selecionados padrões de tipo. Em geral, a aplicação dessa técnica no nível de conceito produzirá resultados mais específicos, já que que conceitos e padrões de conceito representam um nível inferior de medida.
- **Nível de tipos**. Selecionar esta opção significa que as frequências de tipo ou padrões de tipo serão usadas. Os tipos serão usados se foram selecionados tipos como entrada para construção de categorias e padrões de tipo são usados se foram selecionados padrões de tipo. A aplicação dessa técnica no nível de tipo permite que você obtenha uma visualização rápida sobre o tipo de informações presente fornecido.

Contagem mínima de doc. conte com os itens para ter sua própria categoria. Esta opção permite que você construa as categorias a partir de itens que ocorrem frequentemente. Essa opção restringe a saída para somente aquelas categorias que contêm um descritor que ocorreu em pelo menos no número X de registros ou documentos, em que X é o valor a inserir para essa opção.

Agrupar todos os itens restantes em uma categoria chamada. Esta opção permite agrupar todos os tipos ou conceitos que não ocorrem com frequência em uma única categoria 'catch-all' com o nome de sua escolha. Por padrão, esta categoria é denominada *Outro*.

Entrada da categoria. Selecione o grupo ao qual aplicar as técnicas:

- **Resultados de extração não usados**. Esta opção permite que categorias sejam construídas a partir de resultados de extração que não são usado em nenhuma categoria existente. Isso minimiza a tendência dos registros corresponderem diversas categorias e limita o número de categorias produzidas.
- **Todos os resultados da extração**. Esta opção permite que as categorias sejam construídas usando qualquer um dos resultados da extração. Isso é mais útil quando nenhuma ou poucas categorias já existem.

Resolver nomes de categoria duplicados por. Selecione como manipular quaisquer novas categorias ou subcategorias cujos nomes seriam os mesmo que os das categorias existentes. É possível mesclar as novas categorias (e seus descritores) com as categorias existentes com o mesmo nome. Alternativamente, é possível escolher ignorar a criação de qualquer categorias se um nome duplicado for localizado nas categorias existentes.

Estendendo categorias

Extensão é um processo por meio do qual os descritores são incluídos ou aprimorados automaticamente para 'aumentar' categorias existentes. O objeto é produzir uma categoria melhor que captura registos ou documentos relacionados que não foram originalmente designados a tal categoria.

As técnicas de agrupamento automático que você selecionar tentarão identificar conceitos, padrões de TLA e regras de categoria relacionados aos descritores existentes da categoria. Esses novos conceitos, padrões e regras de categoria são, então, incluídos como novos descritores ou incluídos nos descritores existentes. As técnicas de agrupamento para extensão incluem *derivação de raiz conceito*, *inclusão de conceito*, *redes semânticas* (somente em inglês), e *regras de co-ocorrência*. O método **Estender categorias vazias com descritores gerados a partir do nome da categoria** gera descritores usando as palavras nos nomes de categoria, portanto, quanto mais descritivos forem os nomes das categorias, melhor serão os resultados.

Nota: As técnicas de frequência não estão disponíveis ao estender categorias.

A extensão é uma excelente maneira de melhorar as categorias interativamente. Aqui estão alguns exemplos de quando você pode estender uma categoria:

- Após arrastar/soltar padrões de conceito para criar categorias na área de janela Categorias
- Após criar categorias à mão e incluir regras e descritores de categoria simples
- Após importar um arquivo de categoria predefinido no qual as categorias tinham nomes muitos descritivos
- Após refinar as categorias que vieram do TAP, você escolheu

É possível estender uma categoria diversas vezes. Por exemplo, se você importou um arquivo de categoria predefinido com nomes muito descritivo, você poderia estender usando a opção **Estender categorias vazias com descritores gerados a partir do nome de categoria** para obter um primeiro conjunto de descritores e, em seguida, estender tais categorias novamente. No entanto, em outros casos, estender diversas vezes pode resultar em uma categoria muito genérica se os descritores forem estendidos mais e mais. Como as técnicas de agrupamento de construção e extensão usam algoritmos subjacente semelhantes, é improvável que a extensão diretamente após a construção de categorias produza resultados mais interessantes.

Sugestão:

- Se você tentar se estender e não quiser usar os resultados, você pode sempre desfazer a operação (**Editar> Undo**) imediatamente após ter estendido.
- A extensão pode produzir duas ou mais regras de categoria em uma categoria que correspondam exatamente ao mesmo conjunto de documentos já que as regras são construídas independentemente durante o processo. Se desejado, você pode revisar as categorias e remover redundâncias ao editar manualmente a descrição da categoria. Consulte o tópico "Editando descritores de categoria" na página 135 para obter informações adicionais.

Para Estender Categorias

1. Na área de janela Categorias, selecione as categorias que você deseja estender.
2. A partir dos menus, escolha **Categorias> Categorias de Extensão**. A menos que você tenha escolhido a opção para nunca avisar, será exibida uma caixa de mensagens.
3. Escolha se você deseja construir agora ou editar as configurações primeiro.

- Clique em **Estender Agora** para começar a estender categorias usando as configurações atuais. O processo começa e um diálogo de progresso é exibido.
- Clique em **Editar** para revisar e modificar as configurações.

Depois de tentar estender, quaisquer categorias para as quais foram localizados novos descritores são sinalizados pela palavra **Estendido** na área de janela Categorias, para que você possa identificá-los rapidamente. O texto Estendido permanece até que você estenda novamente, edite a categoria de outra forma ou limpe-o através do menu de contexto.

Nota: O número máximo de categorias que podem ser exibidas é de 10.000. Um aviso é exibido se este número for atingido ou excedido. Se isto acontecer, você deve mudar suas opções Construir ou Estender Categorias para reduzir o número de categorias construídas.

Cada uma das técnicas disponível ao construir ou estender categorias é bem adequada a determinados tipos de dados e situações, mas frequentemente, é útil combinar técnicas na mesma análise para capturar a amplitude completa dos documentos ou registros. No ambiente de trabalho interativo, os conceitos e tipos que foram agrupados em uma categoria ainda estarão disponíveis na próxima vez que você construir categorias. Isto significa que você pode ver um conceito em diversas categorias ou localizar categorias redundantes.

As áreas e campos a seguir estão disponíveis na caixa de diálogo Estender Categorias: Configurações:

Estender com. Selecione qual entrada será usada para estender as categorias:

- **Resultados de extração não usados**. Esta opção permite que categorias sejam construídas a partir de resultados de extração que não são usado em nenhuma categoria existente. Isso minimiza a tendência dos registros corresponderem diversas categorias e limita o número de categorias produzidas.
- **Todos os resultados da extração**. Esta opção permite que as categorias sejam construídas usando qualquer um dos resultados da extração. Isso é mais útil quando nenhuma ou poucas categorias já existem.

Técnicas de agrupamento

Para descrições simples de cada uma dessas técnicas, veja "Configurações linguísticas avançadas" na página 104. Essas técnicas incluem:

- **Conceito raiz derivada**
- **Rede Semântica** (apenas texto em inglês, e não usado se a opção Generalizar apenas for selecionada.)
- **Inclusão de conceito**
- **Co-ocorrência** e **Número mínimo de docs** subopção.

Um número de tipos é excluído permanentemente da técnica de redes semânticas uma vez que esses tipos não produzirão resultados relevantes. Eles incluem `<Positive>`, `<Negative>`, `<IP>`, outros tipos não linguísticos, etc.

distância máxima de pesquisa Selecione o quão longe você quer que as técnicas pesquisem antes de produzir categorias. Quanto menor o valor, menos resultados você obterá - entretanto, esses resultados serão menos ruidosos e terão mais probabilidade de estarem significativamente vinculados ou associados entre si. Quanto mais alto o valor, mais resultados você pode obter - entretanto, esses resultados podem ser menos confiáveis ou relevantes. Embora esta opção seja globalmente aplicada em todas as técnicas, seu efeito é maior em coocorrências e redes semânticas.

Evitar emparelhamento de conceitos específicos. Selecione esta caixa de seleção para impedir que o processo agrupe ou pareie dois conceitos na saída. Para criar ou gerenciar pares de conceito, clique em **Gerenciar Pairs ...**. Consulte o tópico "Gerenciando pares de exceção de link" na página 106 para obter mais informações.

Onde possível: Escolha se irá simplesmente estender, generalizar os descritos usando curinga ou ambos.

- **Estender e generalizar.** Esta opção irá estender as categorias selecionadas e, em seguida, generalizar os descritores. Ao escolher generalizar, o produto criará regras de categoria genéricas nas categorias usando o caractere curinga asterisco. Por exemplo, em vez de produzir vários descritores como `[apple tart + .]` e `[apple sauce + .]`, usar curingas pode produzir `[apple * + .]`. Se você generalizar com curingas, com frequência obterá exatamente o mesmo número de registros ou documentos como anteriormente. Entretanto, esta opção tem a vantagem de reduzir o número e simplificar descritores de categoria. Além disso, esta opção aumenta a capacidade de categorizar mais registros ou documentos usando estas categorias nos novos dados de texto (por exemplo, em estudos longitudinais/de onda).
- **Apenas estender**. Esta opção irá estender suas categorias sem generalizar. Pode ser útil escolher primeiro a opção **Estender apenas** para categorias criadas manualmente e, em seguida, estender as mesmas categorias novamente usando a opção **Estender e generalizar** .
- **Generalizar apenas.** Esta opção irá generalizar os descritores sem estender suas categorias de qualquer outra maneira.

  Nota: A seleção desta opção desabilita a opção **Semantic network** ; isto é porque a opção **Rede Semantic** só está disponível quando uma descrição deve ser estendida.

Outras Opções para Estender Categorias

Além de selecionar as técnicas a aplicar, é possível editar qualquer uma das opções a seguir:

Número máximo de itens para estender um descritor por. Ao estender um descritor com itens (conceitos, tipos e outras expressões), defina o número máximo de itens que podem ser incluídos em um único descritor. Se você configurar esse limite para 10, então não mais de 10 itens adicionais podem

ser incluídos em um descritor existente. Se houver mais de 10 itens a serem incluídos, as técnicas param de incluir novos itens após o décimo ser incluído. Fazer isso pode tornar uma lista de descritores mais curta, mas não garante que os itens mais interessantes foram usados primeiro. Você pode preferir diminuir o tamanho da extensão sem penalizar a qualidade usando a opção **Generalizar com curingas onde possível**. Esta opção só se aplica a descritores que contêm os Booleanos & (AND) ou ! (NOT).

Também estender subcategorias. Esta opção também estende quaisquer subcategorias abaixo das categorias selecionadas.

Estender categorias vazias com descritores gerados a partir do nome da categoria. Este método aplica-se apenas às categorias vazias, que possuem descritores 0. Se uma categoria já contiver descritores, ela não será estendida desta maneira. Esta opção tenta criar automaticamente descritores para cada categoria com base nas palavras que compõem o nome da categoria. O nome da categoria é verificado para ver se palavras no nome correspondem a quaisquer conceitos extraídos. Se um conceito for reconhecido, ele será usado para localizar padrões de conceito correspondentes e ambos serão usados para formar descritores para a categoria. Esta opção produz os melhores resultados quando os nomes das categorias são ambos longos e descritivos. Este é um método rápido para gerar descritores de categoria, que, por sua vez, ativa a categoria para capturar registros que contêm esses descritores. Esta opção é mais útil ao importar categorias de algum outro lugar ou ao criar categorias manualmente com nomes descritivos longos.

Gerar descritores como. Esta opção é aplicável apenas se a opção anterior for selecionada.

- **Conceitos.** Escolha esta opção para produzir os descritores resultantes na forma de conceitos, independentemente se eles foram extraídos do texto de origem.
- **Padrões.** Escolha esta opção para produzir os descritos resultantes na forma de padrões, independentemente se os padrões resultantes ou quaisquer padrões foram extraídos.

Criando categorias manualmente

Além de criar categorias usando técnicas de construção de categoria automatizadas e o editor de regras, também é possível criar categorias manualmente. Os métodos manuais a seguir existem:

- Criar uma categoria vazia na qual você incluirá elementos um a um. Consulte o tópico “Criando novas categorias ou renomeando categorias” na página 114 para obter informações adicionais.
- Arrastar termos, tipos e padrões na área de janela de categorias. Consulte o tópico “Criando categorias ao arrastar e soltar” na página 115 para obter informações adicionais.

Criando novas categorias ou renomeando categorias

É possível criar categorias vazias para incluir conceitos e tipos nelas. Também é possível renomear suas categorias.

Para criar uma nova categoria vazia

1. Acesse a área de janela Categorias.
2. A partir dos menus, escolha **Categorias> Criar Categoria Vazia**. A caixa de diálogo Propriedades da Categoria é aberta.
3. Insira um nome para esta categoria no campo Nome.
4. Clique em **OK** para aceitar o nome e fechar a caixa de diálogo. A caixa de diálogo é fechada e um novo nome de categoria aparece na área de janela.

Agora é possível começar a incluir nesta categoria. Consulte o tópico “Incluindo descritores nas categorias” na página 134 para obter informações adicionais.

Para Renomear uma Categoria

1. Selecione uma categoria e escolha **Categorias> Categoria Rename**. A caixa de diálogo Propriedades da Categoria é aberta.
2. Insira um novo nome para essa categoria no campo Nome.

3. Clique em **OK** para aceitar o nome e fechar a caixa de diálogo. A caixa de diálogo é fechada e um novo nome de categoria aparece na área de janela.

Criando categorias ao arrastar e soltar

A técnica arrastar e soltar é manual e não é baseada em algoritmos. Você pode criar categorias na área de janela Categorias ao arrastar:

- Conceitos, tipos ou padrões extraído a partir da área de Resultados de Extração na área de janela Categorias.
- Conceitos extraído da área de janela Dados na área de janela Categorias.
- Linhas inteiras da área de janela Dados na área de janela Categorias. Isto irá criar uma categoria composta de todos os conceitos e padrões extraído contidos em tal linha.

Nota: a área de janela Resultados da Extração suporta diversas seleções para facilitar o ação de arrastar e soltar diversos elementos.

Importante! Não é possível arrastar e soltar conceitos a partir da área de janela Dados que não foram extraídos do texto. Se você desejar forçar a extração de um conceito que você localizou em seus dados, deve-se incluir esse conceito em um tipo. Em seguida, execute a extração novamente. Os resultados da extração conterão o conceito que você acabou de incluir. Você pode, então, usá-lo em sua categoria. Veja o tópico "incluindo conceitos nos tipos" na página 88 para obter mais informações.

Para criar categorias usando arrastar e soltar:

1. Na área de janela Resultados da Extração ou na área de janela Dados, selecione um ou mais conceitos, padrões, tipos, registros ou registros parciais.
2. Mantendo o botão do mouse pressionado, arraste o elemento para uma categoria existente ou para a área de janela para criar uma nova categoria.
3. Quando você tiver atingido a área em que você gostaria de soltar o elemento, solte o botão do mouse. O elemento é incluído na área de janela Categorias. As categorias que foram modificadas aparecem com uma cor de plano de fundo especial. Esta cor é chamado de **plano de fundo de feedback da categoria**. Veja o tópico "Configurando opções" na página 73 para obter mais informações.

Nota: a categoria resultante foi nomeada automaticamente. Se você desejar mudar um nome, é possível renomeá-lo. Veja o tópico "Criando novas categorias ou renomeando categorias" na página 114 para obter mais informações.

Se você desejar ver quais registros são designados a uma categoria, selecione essa categoria na área de janela Categorias. A área de janela de dados é automaticamente atualizada e exibe todos os registros para essa categoria.

Usando regras de categoria

É possível criar categorias de muitas maneiras. Uma dessas maneiras é definir regras de categoria para expressar ideias. Regras de categoria são instruções que automaticamente classificam documentos ou registos em uma categoria, com base em uma expressão lógica usando conceitos, tipos e padrões extraídos, bem como operadores booleanos. Por exemplo, você poderia gravar uma expressão que significa *incluir todos os registros que contêm o conceito extraído* `embassy` *, mas não* `argentina` *, nesta categoria*.

Enquanto algumas regras de categoria são produzidas automaticamente ao construir categorias usando técnicas de agrupamento como *co-ocorrência* e *derivação de raiz conceito* (**Categorias> Configurações de Construção> Configurações Avançadas: Linguísticas**), você também pode criar regras de categoria manualmente no editor de regras usando a sua compreensão de categoria dos dados e do contexto. Cada regra é conectada a uma única categoria, de forma que cada documento ou registros correspondendo à regra seja, então, escorado em tal categoria.

As regras de categoria ajudam a melhorar a qualidade e a produtividade dos resultados de mineração de texto e de análise quantitativa adicionais, permitindo que você categorize respostas com maior especificidade. Sua experiência e conhecimento de negócios pode fornecer a você um entendimento

específicos de seus dados e contexto. Você pode alavancar esse entendimento para converter tal conhecimento em regras de categorias para categorizar seus documentos ou registros de maneira ainda mais eficiente e precisa ao combinar elementos extraídos com lógica booleana.

A capacidade de criar essas regras aprimora a precisão, eficiência e produtividade da codificação ao permitir que você propague camadas do conhecimento de negócios na tecnologia de extração do produto.

Nota: Para exemplos de como regras correspondem a texto, veja "Exemplos da regra de categoria" na página 123

Sintaxe de regra de categoria

Enquanto algumas regras de categoria são produzidas automaticamente ao construir categorias usando técnicas de agrupamento como *co-ocorrência* e *derivação de raiz conceito* (**Categorias> Configurações de Construção> Configurações Avançadas: Linguísticas**), você também pode criar regras de categoria manualmente no editor de regras. Cada regra é um descritor de uma categoria única; portanto, cada documento ou de registro correspondente à regra é automaticamente escorado em tal categoria.

Nota: Para exemplos de como regras correspondem a texto, veja "Exemplos da regra de categoria" na página 123

Ao criar ou editar uma regra, deve-se ter a regra aberta no editor de regras. É possível incluir conceitos, tipos ou padrões, bem como usar curingas para estender as correspondências. Ao usar conceitos, tipos e padrões extraídos, você se beneficia da localização de todos os conceitos relacionados.

Importante! Para evitar erros comuns, recomendamos arrastar e soltar conceitos diretamente da área de janela Resultados da Extração, das áreas de janela Análise de Link de Texto ou da área de janela Dados no editor de regras ou incluir neles por meio dos menus de contexto sempre que possível.

Quando os conceitos, tipos e padrões são reconhecidos, um ícone aparece próximo ao texto.

Tabela 19. Ícones de extração

Ícone	Descrição
	Conceito extraído
	Tipo extraído
	Padrão extraído

Sintaxe e Operadores de Regra

A tabela a seguir contém os caracteres com os quais você definirá sua sintaxe de regras. Use esses caracteres juntamente com os conceitos, tipos e padrões para criar sua regra.

Tabela 20. Sintaxe suportada

Caractere	Descrição
&	O booleano "and". Por exemplo, a & b contém ambos a *e* b tais como: - `invasion & united states` - `2016 & olympics` - `good & apple`

Tabela 20. Sintaxe suportada (continuação)	
Caractere	**Descrição**

|	O booleano "or" é inclusivo, o que significa que se algum ou todos os elementos são localizados, uma correspondência é feita. Por exemplo, a | b contém a *ou* b, tal como:
	
	`- attack | france`
	
	`- condominium | apartment`

!()	O booleano "not". Por exemplo, `!(a)` não contém a. tais como, `!(good & hotel)`, `assassination & !(austria)`ou `!(gold) & !(copper)`
*	Um curinga representando qualquer coisa, desde um único caractere até uma palavra inteira, dependendo de como ele é usado. Consulte o tópico "Usando curingas nas regras de categoria" na página 120 para obter informações adicionais.
()	Um delimitador de expressão. Qualquer expressão dentro dos parênteses é avaliada primeiro.
+	O conector padrão usado para formar um padrão específico de ordem. Quando presentes, os colchetes devem ser usados. Consulte o tópico "Usando padrões de TLA nas regras de categoria" na página 118 para obter informações adicionais.
[]	O delimitador padrão é necessário se você estiver procurando corresponder com base em um padrão de TLA extraído dentro de uma regra de categoria. O conteúdo dentro dos colchetes refere-se a padrões de TLA e nunca corresponderá a conceitos ou tipos com base na coocorrência simples. Se você não extraiu este padrão de TLA, então, nenhuma correspondência será possível. Consulte o tópico "Usando padrões de TLA nas regras de categoria" na página 118 para obter informações adicionais. Não use colchetes se você estiver procurando corresponder conceitos e tipos em vez de padrões. *Nota*: em versões mais antigas, regras de coocorrência e de sinônimo geradas pelas técnicas de construção de categoria costumavam estar entre colchetes. Em todas as novas versões, os colchetes agora indicam a presença de um padrão de TLA. Em vez disso, as regras produzidas pela técnica de coocorrência e os sinônimos serão colocados entre parênteses, tal como `(speaker systems|speakers)`.

Os operadores & e | são comuttativos tais que a & b = b & a e a | b = b | a.

Escapando Caracteres com Barra Invertida

Se você tiver um conceito que contém qualquer caractere que também é um caractere de sintaxe, deve colocar uma barra invertida na frente de tal caractere para que a regra seja corretamente interpretada. O caractere de barra invertida (\) é usado para caracteres de escape que de outro modo têm um significado especial. Ao arrastar e soltar no editor, a aplicação de barras invertidas é feita para você automaticamente.

Os caracteres de sintaxe de regra a seguir devem ser precedidos por uma barra invertida se você deseja que eles sejam tratados como estão, em vez de como a sintaxe de regra:

```
& ! | + < > ( ) [ ] *
```

Por exemplo, como o conceito `r&d` contém o operador "and" (&), a barra invertida é requerida quando ele é digitado no editor de regra, tal como: `r\&d`.

Usando padrões de TLA nas regras de categoria

Padrões de análise de link de texto podem ser explicitamente definidos em regras de categoria para permitir que você obtenha resultados ainda mais específicos e contextuais. Ao definir um padrão em uma regra de categoria, você está ignorando os resultados de extração de conceito mais simples e apenas correspondendo documentos e registros com base nos resultados do padrão de análise de link de texto extraídos.

Importante! Para corresponder documentos usando padrões de TLA em suas regras de categoria, você deve ter executado uma extração com a análise de link de texto ativada. A regra de categoria irá verificar as correspondências localizadas durante esse processo. Se você não escolheu explorar os resultados de TLA na guia Modelo do seu nó de Mineração Texto, poderá escolher ativar a extração de TLA nas configurações de extração na sessão interativa e, em seguida, extrair novamente. Veja o tópico "Extraindo dados" na página 80 para obter mais informações.

Delimitando com colchetes. Um padrão de TLA devem ser colocado entre colchetes `[ ]` se você estiver utilizando-o dentro de uma regra de categoria. O delimitador de padrão é necessário se você estiver procurando a correspondência com base em um padrão de TLA extraído. Como as regras de categoria podem conter tipos, conceitos ou padrões, os colchetes esclarecem para a regra que o conteúdo dentro dos colchetes se refere ao padrão de TLA extraído. Se você não extraiu este padrão de TLA, então, nenhuma correspondência será possível. Se você vir um padrão sem colchetes como `apple + good` na área de janela Categorias, isso provavelmente significa que o padrão foi incluído diretamente na categoria fora do editor de regras de categoria. Por exemplo, se você incluir um padrão de conceito diretamente na categoria a partir da visualização de análise de texto de link análise , ele não aparecerá com colchetes. No entanto, ao usar um padrão dentro de uma regra da categoria, você deve encapsular o padrão dentro dos colchetes dentro da regra da categoria como `[banana + !(good)]`.

Usando o sinal + nos padrões. No IBM SPSS Modeler Text Analytics, é possível ter um padrões de até 6 partes ou slots. Para indicar que a ordem é importante, use o sinal + para conectar cada elemento, como `[company1 + acquired + company2]`. Aqui, a ordem é importante, pois ela alteraria o significado do que a empresa estava adquirindo. A ordem não é determinada pela estrutura da sentença, mas sim pelo modo como a saída do padrão de TLA é estruturada. Por exemplo, se você tem o texto "*Eu amo Paris*" e você quer extrair essa ideia, o padrão TLA provavelmente será `[paris + like]` ou `[<Location> + <Positive>]` em vez de `[<Positive> + <Location>]` já que os recursos de opinião padrão geralmente colocam opiniões na segunda posição em 2 padrões de peça. Portanto, pode ser útil usar o padrão diretamente como um descritor em sua categoria para evitar problemas. Entretanto, se você precisar usar um padrão como parte de uma instrução mais complexa, preste atenção especial à ordem dos elementos dentro dos padrões apresentados na visualização Análise de Texto de Link, já que a ordem desempenha uma função grande no fato de uma correspondência poder ser localizada.

Por exemplo, digamos que você tinha as duas seguintes expressões de amostra: "I *like pineapple*" e "I hate *pineapple*. However, I *like* strawberries". A expressão `like & pineapple` corresponderia a ambos os textos já que ela é uma expressão de conceito e não uma regra de link de texto (não contida entre colchetes). A expressão `pineapple + like` corresponde apenas "I *like pineapple*" desde que no segundo texto, a palavra *like* está associada a *strawberries* em vez disso.

Agrupando com padrões. É possível simplificar suas regras com seus próprios padrões. Digamos que você deseje capturar as três expressões a seguir, `cayenne peppers + like`, `chili peppers + like` e `peppers + like`. É possível agrupá-las em uma única regra de categoria, tal como `[* peppers & like]`. Se você tinha outra expressão `hot peppers + good`, é possível agrupar essas quatro com uma regra, tal como `[* peppers + <Positive>]`.

Ordem nos padrões. Para organizar melhor a saída, as regras de análise de link de texto fornecidas nos modelos instalados com seu produto tentam produzir padrões básicos na mesma ordem, independente da ordem da palavra na sentença. Por exemplo, se você tinha um registro contendo o texto "*Boas apresentações.*" e outro registro contendo "*as apresentações eram boas*", ambos os textos são correspondidos pela mesma regra e saída na mesma ordem que `presentation + good` nos resultados do padrão conceito em vez de `presentation + good` e também `good + presentation`. E em padrões com dois slots, tais como aqueles no exemplo, os conceitos designados os tipos na biblioteca Opiniões serão apresentados, por padrão, por último na saída, tal como `apple + bad`.

Tabela 21. Sintaxe padrão e uso de booleano

Expressão	Corresponde um documento ou registro que
`[ ]`	Contém qualquer padrão de TLA. O delimitador de padrão é requerido *em regras de categoria* se você estiver procurando corresponder com base em um padrão de TLA extraído. O conteúdo dentro dos colchetes refere-se a padrões de TLA, não simplesmente a conceitos e tipos. Se você não extraiu este padrão de TLA, então, nenhuma correspondência será possível. Se você quisesse criar uma regra que não incluía nenhum padrão, você poderia usar `!([ ])`.
`[a]`	Contém um padrão de que, pelo menos, um elemento é a, independentemente de sua posição no padrão. Por exemplo, `[deal]` pode combinar `[deal + good]` ou apenas `[deal + .]`
`[a + b]`	Contém um padrão de conceito. Por exemplo, `[deal + good]`. *Nota:* Se você desejar apenas capturar este padrão sem incluir qualquer outro elemento, nós recomendados incluir o padrão diretamente em sua categoria em vez de criar uma regra com ele.
`[a + b + c]`	Contém um padrão de conceito. O sinal + denota que a ordem dos elementos correspondentes é importante. Por exemplo, `[company1 + acquired + company2]`.
`[<A> + <B>]`	Contém qualquer padrão com o tipo <A> no primeiro lote e o tipo <B> no segundo slot e há exatamente dois slots. O sinal + denota que a ordem dos elementos correspondentes é importante. Por exemplo, `[<Budget> + <Negative>]`. *Nota:* Se você desejar apenas capturar este padrão sem incluir qualquer outro elemento, nós recomendados incluir o padrão diretamente em sua categoria em vez de criar uma regra com ele.
`[<A> & <B>]`	Contém qualquer padrão de tipo com o tipo <A> e o tipo<B>. Por exemplo, `[<Budget> & <Negative>]`. Este padrão TLA nunca será extraído; no entanto, quando escrito como tal ele é realmente igual a `[<Budget> + <Negative>]|[<Negative> + <Budget>]`. A ordem dos elementos de correspondência não é importante. Além disso, outros elementos podem estar no padrão, mas ele deve ter pelo menos <Budget> e <Negative>.
`[a + .]`	Contém um padrão no qual a é o único conceito e não há nada em qualquer outro slot para tal padrão. Por exemplo, `[deal + .]` corresponde ao padrão de conceito no qual a única saída é o conceito `deal`. Se você incluiu o conceito `deal` como um descritor de categoria, obterá todos os registros com negociação como um conceito, incluindo declarações positivas sobre uma negociação. No entanto, usando `[deal + .]` corresponderá apenas aqueles registros padrão resultados representando `deal` e nenhum outro relacionamento ou opiniões e não corresponderia `deal + fantastic`. *Nota:* Se você desejar apenas capturar este padrão sem incluir qualquer outro elemento, nós recomendados incluir o padrão diretamente em sua categoria em vez de criar uma regra com ele.

Tabela 21. Sintaxe padrão e uso de booleano (continuação)	
Expressão	**Corresponde um documento ou registro que**
`[<A> + <>]`	Contém um padrão no qual <A> é o único tipo. Por exemplo, `[<Budget> + <>]` corresponde ao padrão em que a única saída é um conceito do tipo <Budget>. *Nota:* Você pode usar o <> para denotar um tipo vazio apenas ao colocá-lo após o padrão + símbolo em padrão de tipo como `[<Budget> + <>]` mas não `[price + <>]`. *Nota:* Se você desejar apenas capturar este padrão sem incluir qualquer outro elemento, nós recomendados incluir o padrão diretamente em sua categoria em vez de criar uma regra com ele.
`[a + !(b)]`	Contém pelo menos um padrão que inclui o conceito a mas não inclui o conceito b. Deve incluir pelo menos um padrão. Por exemplo,`[price + !(high)]` ou para tipos, `[!(<Fruit>|<Vegetable>) + <Positive>]`
`!([<A> & <B>])`	Não contém um padrão específico. Por exemplo, `!([<Budget> & <Negative>])`.

Nota: Para exemplos de como regras correspondem a texto, veja "Exemplos da regra de categoria" na página 123

Usando curingas nas regras de categoria

Curingas podem ser incluídos nos conceitos nas regras para estender os recursos correspondentes. O curinga asterisco * pode ser colocado antes e/ou após uma palavra para indicar como os conceitos podem ser correspondidos. Há dois tipos de curinga usados:

- **Curingas de afixação**. Esses curingas prefixam ou sufixam imediatamente sem qualquer espaço separando a cadeia e o asterisco. Por exemplo, `operat*` poderia corresponder *operat*, *operate*, *operates*, *operations*, *operational* e assim por diante.
- **Curingas de palavra**. Esses curingas prefixam ou sufixam um conceito com um espaço entre o conceito e o asterisco. Por exemplo, `* operation` poderia corresponder*operation*, *surgical operation*, *post operation* e assim por diante. Além disso, um curinga de palavra pode ser usado juntamente com um curinga de afixação, tal como `* operat* *`, que poderia corresponder *operation*, *surgical operation*, *telephone operator*, *operatic aria*, e assim por diante. Como você pode ver neste último exemplo, recomendamos que curingas sejam usados com cuidado de modo a não lançar a rede muito longe e capturar correspondências indesejadas.

Exceções!

- Um curinga nunca pode ser independente. Por exemplo, `(apple | * )` não seria aceito.
- Um curinga nunca pode ser usado para corresponder nomes de tipos. `<Negative*>` não corresponderá a nenhum nome de tipo.
- Não é possível usar um filtro para impedir que certos tipos sejam correspondidos com conceitos localizados através de curingas. O tipo ao qual o conceito é designado é usado automaticamente.
- Um curinga nunca pode estar no meio de uma sequência de palavras, quer ele esteja no final ou no início de uma palavra (open* account) ou seja um componente independente (open * account). Também não é possível usar curingas em nomes de tipo. Por exemplo, `word* word`, como `apple* recipe`, não combinará com receita de applesauce ou qualquer outra coisa em nada. No entanto, `apple* *` corresponderia a *receita de applesauce*, *pie pie*, *apple* e assim por diante. Em outro exemplo, `word *`

word, tal como `apple * toast`, não corresponderá a *apple cinnamon toast* ou a qualquer outra coisa, já que o asterisco aparece entre duas outras palavras. Entretanto, `apple *` corresponderia a *apple cinnamon toast*, *apple*, *apple pie* e assim por diante.

Tabela 22. Uso de curinga

Expressão	**Corresponde um documento ou registro que**
`*apple`	Contém um conceito que termina com letras escritas, mas pode ter qualquer número de letras como um prefixo. Por exemplo: `*apple` termina com as letras *apple*, mas pode assumir um prefixo, tal como: - `apple` - `pineapple` - `crabapple`
`apple*`	Contém um conceito que inicia com letras escritas, mas pode ter qualquer número de letras como um sufixo. Por exemplo: `apple*` inicia com as letras *apple*, mas pode assumir um sufixo ou nenhum sufixo, tal como: - `apple` - `applesauce` - `applejack` Por exemplo, `apple* & !(pear* | quince)`, que contém um conceito que começa com as letras `apple` mas não um conceito começando com as letras *pear* ou o conceito `quince`, NÃO combinaria: `apple & quince` mas, corresponderia a: - `applesauce` - `apple & orange`
`*product*`	Contém um conceito que contém as letras escritas `product`, mas pode ter qualquer número de letras como um prefixo, um sufixo ou ambos. Por exemplo: `*product*` poderia corresponder a: - `product` - `byproduct` - `unproductive`

Tabela 22. Uso de curinga (continuação)

Expressão	Corresponde um documento ou registro que
`* loan`	Contém um conceito que contém a palavra `loan`, mas pode ser uma composição com outra palavra colocada antes dela. Por exemplo, `* loan` poderia corresponder a: - `loan` - `car loan` - `home equity loan` Por exemplo, `[* delivery + <Negative>]` contém um conceito que termina na palavra `delivery` na primeira posição e contém um tipo `<Negative>` na segunda posição poderia corresponder aos padrões de conceito a seguir: - `package delivery + slow` - `overnight delivery + late`
`event *`	Contém um conceito que contém a palavra `event`, mas pode ser uma composição acompanhada por outra palavra. Por exemplo, `event *` poderia corresponder a: - `event` - `event location` - `event planning committee`
`* apple *`	Contém um conceito que pode iniciar com qualquer palavra acompanhada pela palavra `apple`, possivelmente acompanhada por outra palavra. `*` significa 0 ou n, para que ele corresponde a `apple`. Por exemplo, `* apple *` poderia corresponder a: - `gala applesauce` - `granny smith apple crumble` - `famous apple pie` - `apple` Por exemplo, `[* reservation* * + <Positive>]`, que contém um conceito com a palavra `reservation` (independente de onde ela esteja no conceito) na primeira posição e contém um tipo `<Positive>` na segunda posição, corresponderia aos padrões de conceito: - `reservation system + good` - `online reservation + good`

Nota: Para exemplos de como regras correspondem a texto, veja "Exemplos da regra de categoria" na página 123

Exemplos da regra de categoria

Para ajudar a demonstrar como as regras são correspondidas a registros de forma diferente com base na sintaxe usada para expressá-las, considere o exemplo a seguir.

Registros de Exemplo

Imagine que você tivesse dois registros:

- **Registro A**: *"when I checked my wallet, I saw I was missing 5 dollars."*
- **Registro B**: *"$5 was found at the picnic area, but the blanket was missing."*

As duas tabelas a seguir mostram o que pode ser extraído para conceitos e tipos, bem como padrões de conceito e padrões de tipo.

Conceitos e Tipos Extraídos do Exemplo

Tabela 23. Conceitos e tipos extraídos de exemplo

Conceito Extraído	**Conceitos Tipificados Como**
`wallet`	`<Unknown>`
`missing`	`<Negative>`
`USD5`	`<Currency>`
`blanket`	`<Unknown>`
`picnic area`	`<Unknown>`

Padrões de TLA Extraídos do Exemplo

Tabela 24. Saída do padrão de TLA extraída de exemplo

Padrões de Conceito Extraídos	**Padrões de Tipo Extraídos**	**Do Registro**
`picnic area + .`	`<Unknown> + <>`	Registro B
`wallet + .`	`<Unknown> + <>`	Registro A
`blanket + missing`	`<Unknown> + <Negative>`	Registro B
`USD5 + .`	`<Currency> + <>`	Registro B
`USD5 + missing`	`<Currency> + <Negative>`	Registro A

Como Possíveis Regras de Categoria Correspondem

A tabela a seguir contém alguma sintaxe que pode não poderia ser inserida no editor de regras da categoria. Nem todas as regras aqui funcionam e nem todas correspondem aos mesmos registros. Veja como a sintaxe diferente afeta os registros correspondidos.

Tabela 25. Regras de amostra

Sintaxe de Regra	**Resultado**
`USD5 & missing`	Corresponde a ambos os registros A e B, já que ambos contêm o conceito extraído `missing` e o conceito extraído USD5. Isso é equivalente a: `(USD5 & missing)`

Tabela 25. Regras de amostra (continuação)

Sintaxe de Regra	**Resultado**
`missing & USD5`	Corresponde a ambos os registros A e B, já que ambos contêm o conceito extraído `missing` e o conceito extraído USD5. Isso é equivalente a: `(missing & USD5)`
`missing & <Currency>`	Corresponde a ambos os registros A e B já que ambos contêm o conceito extraído `missing` e um conceito correspondendo ao tipo `<Currency>`. Isso é equivalente a: `(missing & <Currency>)`
`<Currency> & missing`	Corresponde a ambos os registros A e B já que ambos contêm o conceito extraído `missing` e um conceito correspondendo ao tipo `<Currency>`. Isso é equivalente a: `(<Currency> & missing)`
`[USD5 + missing]`	Corresponde A, mas **não** B, já que o registro B não produziu nenhuma saída de padrão de TLA contendo `USD5 + missing` (veja a tabela anterior). Isto é equivalente à saída do padrão de TLA: `USD5 + missing`
`[missing + USD5]`	Não corresponde nem A nem B, já que nenhum padrão de TLA extraído (veja a tabela anterior) corresponde à ordem expressa aqui com `missing` na primeira posição. Isto é equivalente à saída do padrão de TLA: `USD5 + missing`
`[missing & USD5]`	As correspondências A mas **não** B desde que nenhum padrão de TLA foi extraído do registro B. O uso do caractere & indica que a ordem não é importante ao combinar; portanto, essa regra procura uma correspondência de padrão para `[missing + USD5]` ou `[USD5 + missing]`. Apenas `[USD5 + missing]` do registro A tem uma correspondência.
`[missing + <Currency>]`	Não corresponde nem ao registro A nem ao B já que nenhum padrão de TLA extraído correspondeu esta ordem. Isso não tem equivalente, já que uma saída de TLA é baseada apenas nos termos (`USD5 + missing`) ou nos tipos de (`<Currency> + <Negative>`), mas não combina conceitos e tipos.
`[<Currency> + <Negative>]`	Corresponde o registro A mas não B desde que nenhum padrão TLA foi extraído do registro B. Isso é equivalente à saída TLA: `<Currency> + <Negative>`

Tabela 25. Regras de amostra (continuação)	
Sintaxe de Regra	**Resultado**
`[<Negative> + <Currency>]`	Não corresponde nem ao registro A nem ao B já que nenhum padrão de TLA extraído correspondeu esta ordem. No modelo `Opinions` , por padrão, quando um *tópico* é encontrado com um *opinião*, o *tópico* (`<Currency>`) ocupa a primeira posição de slot e *opinião* (`<Negative>`) ocupa a segunda posição de caça-níqueis.

Criando regras de categoria

Ao criar ou editar uma regra, deve-se ter a regra aberta no editor de regras. É possível incluir conceitos, tipos ou padrões, bem como usar curingas para estender as correspondências. Ao usar conceitos, tipos e padrões reconhecidos, você se beneficia pois eles localizarão todos os conceitos relacionados. Por exemplo, ao usar um conceito, todos os seus termos associados, formas no plural e sinônimos também são correspondidos com a regra. Da mesma forma, quando você usa um tipo, todos os seus conceitos também são capturadas pela regra.

É possível abrir o editor de regras ao editar uma regra existente ou ao clicar com o botão direito no nome da categoria e escolher **Criar Regra**.

Você pode usar os menus de contexto, arrastar e soltar ou inserir manualmente os conceitos, tipos e padrões no editor. Em seguida, combine estes com operadores booleanos (`&`, `!`, `()`, `|`) e suportes para formar suas expressões de regra. Para evitar erros comuns, recomendamos arrastar e soltar os conceitos diretamente a partir da área de janela Resultados de Extração ou da área de janela Dados no editor de regras. Preste bastante atenção à sintaxe das regras para evitar erros. Consulte o tópico "Sintaxe de regra de categoria" na página 116 para obter mais informações.

Nota: para obter exemplos de como as regras correspondem ao texto, veja "Exemplos da regra de categoria" na página 123.

Para criar uma regra

1. Se você ainda não extraiu nenhum dado ou sua extração estiver desatualizada, faça isso agora. Consulte o tópico "Extraindo dados" na página 80 para obter informações adicionais.

   Nota: se você filtrar uma extração de tal forma que não haja mais nenhum conceito visível, uma mensagem de erro será exibida ao tentar criar ou editar uma regra de categoria. Para evitar isso, modifique seu filtro de extração de modo que os conceitos estejam disponíveis.
2. Na área de janela Categorias, selecione a categoria na qual você deseja incluir sua regra.
3. A partir dos menus escolha **Categorias> Criar Regra**. A área de janela do editor de regras de categoria é aberta na janela.
4. No campo Nome da Regra, insira um nome para sua regra. Se você não fornecer um nome, a expressão será usada como o nome automaticamente. É possível renomear essa regra posteriormente.
5. No campo de texto de expressão maior, é possível:
   - Inserir texto diretamente no campo ou arrastar e soltar a partir de outra área de janela. Use apenas conceitos, tipos e padrões extraídos. Por exemplo, se você inserir a palavra `cats`, mas apenas na forma singular, `cat`, aparecerá em sua área de janela Resultados da Extração, o editor não será capaz de reconhecer `cats`. Neste último caso, a forma singular pode incluir automaticamente o plural, caso contrário, você poderia usar um curinga. Consulte o tópico "Sintaxe de regra de categoria" na página 116 para obter mais informações.
   - Selecione os conceitos, tipos ou padrões que você deseja incluir nas regras e use os menus.
   - Inclua operadores booleanos para vincular elementos em sua regra. Use os botões da barra de ferramentas para adicionar o "e" Boolean **&**, o "ou" Boolean **|**, o "not" Boolean **! ()**, parênteses **()**, e colchetes para padrões **[]** à sua regra.

6. Clique no botão **Testar Regra** para verificar se sua regra é bem-formada. Veja o tópico "Sintaxe de regra de categoria" na página 116 para obter mais informações. O número de documentos ou registros localizados aparece entre parênteses próximo ao texto **Resultado do teste**. À direita deste texto, você pode ver os elementos em sua regra que foram reconhecidos ou quaisquer mensagens de erro. Se o gráfico próximo ao tipo, padrão ou conceito aparecer com um ponto de interrogação vermelho, isso indica que o elemento não corresponde a nenhuma extração conhecida. Se ele não corresponder, então, a regra não localizará nenhum registro.
7. Para testar uma parte da sua regra, selecione tal parte e clique em **Testar Seleção**.
8. Faça quaisquer mudanças necessárias e teste novamente sua regra se você localizou problemas.
9. Quando finalizado, clique em **Salvar & Fechar** para salvar sua regra novamente e fechar o editor. O novo nome da regra aparece na categoria.

Editando e excluindo regras

Após você ter criado e salvo uma regra, é possível editar tal regra a qualquer momento. Consulte o tópico "Sintaxe de regra de categoria" na página 116 para obter informações adicionais.

Se você não quiser mais uma regra, será possível excluí-la.

Para editar regras

1. Na tabela Descritores na caixa de diálogo Definições de Categoria, selecione a regra.
2. A partir dos menus escolha **Categorias> Editar Regra** ou clique duas vezes no nome da regra. O editor é aberto com a regra selecionada.
3. Faça quaisquer mudanças na regra usando resultados da extração e os botões da barra de ferramentas.
4. Teste novamente sua regra para certificar-se de que ela retorne os resultados esperados.
5. Clique em **Salvar & Fechar** para salvar sua regra novamente e fechar o editor.

Para excluir uma regra

1. Na tabela Descritores na caixa de diálogo Definições de Categoria, selecione a regra.
2. A partir dos menus, escolha **Editar> Excluir**. A regra é excluída da categoria.

Importando e exportando categorias predefinidas

Se você tiver suas próprias categorias armazenadas em um arquivo Microsoft Excel (**.xls*, **.xlsx*), será possível importá-lo no IBM SPSS Modeler Text Analytics .

Também é possível exportar as categorias que você tem em uma sessão de ambiente de trabalho interativo aberta para um Microsoft Excel (**.xls*, **.xlsx*). Ao exportar suas categorias, é possível escolher incluir ou excluir algumas informações adicionais, tais como descritores e escoragens. Consulte o tópico "Exportando categorias" na página 130 para obter mais informações.

Se as suas categorias predefinidas não tiverem códigos ou você quiser novos códigos, você poderá gerar automaticamente um novo conjunto de códigos para o conjunto de categorias na área de janela de categorias escolhendo **Categorias> Gerenciar Categorias> Códigos de Geração Automático** a partir dos menus. Isto removerá quaisquer códigos existentes e os renomeará automaticamente.

Importando categorias predefinidas

É possível importar suas categorias predefinidas no IBM SPSS Modeler Text Analytics . Antes de importar, certifique-se de que o arquivo de categoria predefinido esteja em um arquivo Microsoft Excel (**.xls*, **.xlsx*) e esteja estruturado em um dos formatos suportados. Também é possível escolher que o produto automaticamente detecte o formato para você. Os seguintes formatos são suportados:

- **Formato de lista simples**: Veja o tópico "Formato de lista simples" na página 128 para obter mais informações.
- **Formato compacto**: Veja o tópico "Formato compacto" na página 128 para obter mais informações.

- **Formato indentado**: Veja o tópico "Formato indentado" na página 129 para obter mais informações.

Para importar categorias predefinidas

1. A partir dos menus do ambiente de trabalho interativo, escolha **Categorias> Gerenciar Categorias> Importar Categorias Predefinidas**. Um assistente Importar Categorias Predefinidas é exibido.
2. Na lista suspensa Verificar em, selecione a unidade e a pasta na qual o arquivo está localizado.
3. Selecione o arquivo na lista. O nome do arquivo aparece na caixa de texto Nome do Arquivo.
4. Selecione a planilha contendo as categorias predefinidas na lista. O nome da planilha aparece no campo Planilha.
5. Para começar a escolher o formato de dados, clique em **Avançar**.
6. Escolha o formato para seu arquivo ou escolha a opção para permitir que o produto tente detectar o formato automaticamente. A detecção automática funciona melhor na maioria dos formatos comuns.
   - **Formato de lista simples**: Veja o tópico "Formato de lista simples" na página 128 para obter mais informações.
   - **Formato compacto**: Veja o tópico "Formato compacto" na página 128 para obter mais informações.
   - **Formato indentado**: Veja o tópico "Formato indentado" na página 129 para obter mais informações.
7. Para definir as opções de importação adicionais, clique em **Avançar**. Se você escolher ter o formato detectado automaticamente, será direcionado para a etapa final.
8. Se uma ou mais linhas contêm cabeçalhos de coluna ou outras informações externas, selecione o número da linha a partir da qual você deseja iniciar a importação na opção **Iniciar importação na linha**. Por exemplo, se seus nomes de categoria começam na linha 7, deve-se inserir o número 7 para esta opção para importar o arquivo corretamente.
9. Se o seu arquivo contém códigos de categoria, escolha a opção **Contém códigos de categoria**. Fazer isso ajuda o assistente a reconhecer seus dados corretamente.
10. Revise as células codificadas por cor e a legenda para certificar-se de que os dados tenham sido corretamente identificados. Quaisquer erros detectados no arquivo são mostrados em vermelho e referenciados abaixo da tabela de visualização de formato. Se o formato incorreto foi selecionado, volte e escolha outro. Se você precisar fazer correções em seu arquivo, faça essas mudanças e reinicie o assistente ao selecionar o arquivo novamente. Deve-se corrigir todos os erros antes de poder concluir o assistente.
11. Para revisar o conjunto de categorias e subcategorias que serão importados e para definir como criar descritores para essas categorias, clique em **Avançar**.
12. Revise o conjunto de categorias que serão importados na tabela. Se você não verá as palavras-chave que você esperava ver como descritores, pode ser que eles não foram reconhecidos durante a importação. Certifique-se de que eles estejam adequadamente prefixados e apareça na célula correta.
13. Escolha como você deseja manipular quaisquer categorias pré-existentes em sua sessão.
    - **Substituir todas as categorias existentes**. Esta opção limpa todas as categorias existentes e, em seguida, as categorias recém-importadas são usadas sozinhas em seu lugar.
    - **Anexar às categorias existentes**. Esta opção importará as categorias e mesclar quaisquer categorias comuns com as categorias existentes. Ao incluir em categorias existentes, você precisa determinar como deseja que quaisquer duplicatas sejam manipuladas. Uma opção (opção: **Mesclar**) é mesclar quaisquer categorias sendo importadas com categorias existentes se elas compartilharem um nome de categoria. Outra opção (opção: **Excluir da importação**) é proibir a importação de categorias se uma com o mesmo nome existir.
14. **Importar palavras-chave como descritores** é uma opção para importar as palavras-chave identificadas em seus dados como descritores para a categoria associada.
15. **Estender categorias ao derivar descritores** é uma opção que gerará descritores a partir das palavras que representam o nome da categoria ou subcategoria e/ou das palavras que compõe a

anotação. Se as palavras corresponderem aos resultados extraído, então, tais palavras são incluídas como descritores para a categoria. Esta opção produz os melhores resultados quando os nomes ou anotações das categorias são ambos longos e descritivos. Este é um método rápido para gerar os descritores de categorias que permitem que a categoria capture registros que contêm tais descritores.

- O campo **A partir de** permite que você selecione a partir de qual texto os descritores serão derivados, os nomes ou as categorias e subcategorias, as palavras nas anotações ou ambos.
- O campo **Como** permite que você escolha criar esses descritores na forma de conceitos ou padrões de TLA. Se a extração de TLA não ocorreu, as opções de **padrões** estarão desativadas neste assistente.

16. Para importar as categorias predefinidas na área de janela Categorias, clique em **Concluir**.

Formato de lista simples

No formato de lista simples, há apenas um nível superior de categorias sem qualquer hierarquia, significando que não há subcategorias ou sub-redes. Os nomes de categorias estão em uma única coluna.

As informações a seguir podem ser contidas em um arquivo deste formato:

- A coluna **Códigos** opcionais contém valores numéricos que identificam exclusivamente cada categoria. Se você especificar que o arquivo de dados contém códigos (opção **Contém códigos de categoria** na etapa **Configurações de Conteúdo**), então, uma coluna contendo códigos exclusivos para cada categoria deve existir na célula diretamente à esquerda do nome da categoria. Se seus dados não contêm códigos, mas você deseja criar alguns códigos mais tarde, você sempre pode gerar códigos posteriormente (**Categorias> Gerenciar Categorias> Códigos de Geração Automático**).
- Uma coluna **Nomes de categorias** *requerida* contém todos os nomes das categorias. Esta coluna é requerida para importar usando este formato.
- **Anotações** opcionais na célula imediatamente à direita do nome da categoria. Esta anotação consiste em texto que descreve suas categorias/subcategorias.
- **Palavras-chave** opcionais podem ser importadas como descritores para categorias. Para serem reconhecidas, estas palavras-chave devem existir na célula diretamente abaixo do nome da categoria/subcategoria associada e a lista de palavras-chave deve ser prefixada pelo caractere de sublinhado (_), tal como `_firearms, weapons / guns`. A célula da palavra-chave pode conter uma ou mais palavras usadas para descrever cada categoria. Estas palavras serão importadas como descritores ou ignoradas, dependendo do que você especificar na última etapa do assistente. Posteriormente, os descritores são comparados aos resultados extraídos do texto. Se uma correspondência for localizada, então, tal registro ou documento será armazenado na categoria contendo este descritor.

Tabela 26. Formato de lista simples com códigos, palavras-chave e anotações

Coluna A	**Coluna B**	**Coluna C**
Código de categoria (*opcional*)	Nome da categoria	Anotação
	Lista _Descriptor/keyword (*opcional*)	

Formato compacto

O formato compacto é estruturado de forma semelhante ao formato de lista simples, exceto que o formato compacto é usado com categorias hierárquicas. Portanto, uma coluna no nível de código é necessária para definir o nível hierárquico de cada categoria e subcategoria.

As informações a seguir podem ser contidas em um arquivo deste formato:

- Uma coluna **no nível de código** *requerida* contém números que indicam a posição hierárquica para as informações subsequentes em tal linha. Por exemplo, se os valores 1, 2 ou 3 forem especificados e você tiver ambas as categorias e subcategorias, então, 1 é para categorias, 2 é para subcategorias e 3 é para

sub-subcategorias. Se você tiver apenas categorias e subcategorias, então, 1 é para categorias e 2 é para subcategorias. E assim por diante, até a profundidade desejada da categoria.

- A coluna de **códigos** opcional contém valores que identificam exclusivamente cada categoria. Se você especificar que o arquivo de dados contém códigos (opção **Contém códigos de categoria** na etapa **Configurações de Conteúdo**), então, uma coluna contendo códigos exclusivos para cada categoria deve existir na célula diretamente à esquerda do nome da categoria. Se seus dados não contêm códigos, mas você deseja criar alguns códigos mais tarde, você sempre pode gerar códigos posteriormente (**Categorias> Gerenciar Categorias> Códigos de Geração Automático**).
- Uma colunas **nomes de categorias** *requerida* contém todos os nomes das categorias e subcategorias. Esta coluna é requerida para importar usando este formato.
- **Anotações** opcionais na célula imediatamente à direita do nome da categoria. Esta anotação consiste em texto que descreve suas categorias/subcategorias.
- **Palavras-chave** opcionais podem ser importadas como descritores para categorias. Para serem reconhecidas, estas palavras-chave devem existir na célula diretamente abaixo do nome da categoria/subcategoria associada e a lista de palavras-chave deve ser prefixada pelo caractere de sublinhado (_), tal como `_firearms, weapons / guns`. A célula da palavra-chave pode conter uma ou mais palavras usadas para descrever cada categoria. Estas palavras serão importadas como descritores ou ignoradas, dependendo do que você especificar na última etapa do assistente. Posteriormente, os descritores são comparados aos resultados extraídos do texto. Se uma correspondência for localizada, então, tal registro ou documento será armazenado na categoria contendo este descritor.

Tabela 27. Exemplo de formato compacto com códigos

Coluna A	**Coluna B**	**Coluna C**
Nível de código hierárquico	Código de categoria (*opcional*)	Nome da categoria
Nível de código hierárquico	Código de subcategoria (*opcional*)	Nome da subcategoria

Tabela 28. Exemplo de formato compacto sem códigos

Coluna A	**Coluna B**
Nível de código hierárquico	Nome da categoria
Nível de código hierárquico	Nome da subcategoria

Formato indentado

No formato de arquivo indentado, o conteúdo é hierárquico, que significa que ele contém categorias e um ou mais níveis de subcategorias. Além disso, sua estrutura é indentada para denotar essa hierarquia. Cada linha no arquivo contém uma categoria ou subcategoria, mas as subcategorias são indentadas das categorias e quaisquer sub-subcategorias são indentadas das subcategorias e assim por diante. É possível criar manualmente essa estrutura no Microsoft Excel ou usar uma que foi exportada de outro produto e salva em um formato Microsoft Excel.

- **Os códigos de categoria de nível superior e os nomes de categorias** ocupam as colunas A e B, respectivamente. Ou, se nenhum código estiver presente, então, o nome da categoria está na coluna A.
- **Códigos de subcategoria e nomes de subcategoria** ocupam as colunas B e C, respectivamente. Ou, se nenhum código estiver presente, então, o nome da subcategoria está na coluna B. A subcategoria é um membro de uma categoria. Não é possível ter subcategorias se você não tiver categorias de nível superior.

Tabela 29. Estrutura indentada com códigos			
Coluna A	**Coluna B**	**Coluna C**	**Coluna D**
Código de categoria (*opcional*)	Nome da categoria		
	Código de subcategoria (*opcional*)	Nome da subcategoria	
		Código da sub-subcategoria (*opcional*)	Nome da sub-subcategoria

Tabela 30. Estrutura indentada sem códigos		
Coluna A	**Coluna B**	**Coluna C**
Nome da categoria		
	Nome da subcategoria	
		Nome da sub-subcategoria

As informações a seguir podem ser contidas em um arquivo deste formato:

- **Códigos** opcionais devem ser valores que identificam exclusivamente cada categoria ou subcategoria. Se você especificar que o arquivo de dados contém códigos (opção **Contém códigos de categoria** na etapa **Configurações de Conteúdo**), então, um código exclusivo para cada categoria ou subcategoria deve existir na célula diretamente à esquerda do nome da categoria/subcategoria. Se seus dados não contêm códigos, mas você deseja criar alguns códigos mais tarde, você sempre pode gerar códigos posteriormente (**Categorias> Gerenciar Categorias> Códigos de Geração Automático**).
- Um **nome** *requerido* para cada categoria e subcategoria. As subcategorias devem ser indentadas das categorias por uma célula à direita em uma linha separada.
- **Anotações** opcionais na célula imediatamente à direita do nome da categoria. Esta anotação consiste em texto que descreve suas categorias/subcategorias.
- **Palavras-chave** opcionais podem ser importadas como descritores para categorias. Para serem reconhecidas, estas palavras-chave devem existir na célula diretamente abaixo do nome da categoria/subcategoria associada e a lista de palavras-chave deve ser prefixada pelo caractere de sublinhado (_), tal como `_firearms, weapons / guns`. A célula da palavra-chave pode conter uma ou mais palavras usadas para descrever cada categoria. Estas palavras serão importadas como descritores ou ignoradas, dependendo do que você especificar na última etapa do assistente. Posteriormente, os descritores são comparados aos resultados extraídos do texto. Se uma correspondência for localizada, então, tal registro ou documento será armazenado na categoria contendo este descritor.

Importante! Se você usar um código em um nível, deve-se incluir um código para cada categoria e subcategoria. Caso contrário, o processo de importação falhará.

Exportando categorias

Também é possível exportar as categorias que você tem em uma sessão de ambiente de trabalho interativo em um formato de arquivo Microsoft Excel (**.xls*, **.xlsx*). Os dados que serão exportados são originários em grande parte do conteúdo atual da área de janela Categorias ou a partir das propriedades da categoria. Portanto, recomendamos que você escore novamente se planeja também exportar o valo de escore **Docs.** .

Tabela 31. Opções de exportação da categoria	
Sempre é exportado...	**Exportado opcionalmente...**
• Códigos de categorias, se presentes • Nomes de categoria (e subcategoria) • Níveis de código, se presentes (formato *Simples*/*Compacto*) • Títulos da coluna (formato *Simples*/*Compacto*)	• **Documentos.** pontuações • Anotações de categorias • Nomes de descritores • Contagens de descritores

Importante! Ao exportar descritores, eles são convertidos em sequência de caracteres de texto e prefixados por um sublinhado. Se você reimportar nesse produto, a capacidade de distinguir entre os descritores que são padrões, aqueles que são regras de categoria e aqueles que são conceitos simples serão perdidos. Se você pretende reutilizar essas categorias nesse produto, é altamente recomendado criar um arquivo de pacote de análise de texto (TAP) em vez disso, uma vez que o formato o TAP preservará todos os descritores como eles estão atualmente definidos, bem como todas as suas categorias, códigos e também os recursos linguísticos usados. Os arquivos TAP podem ser usados em ambos IBM SPSS Modeler Text Analytics e IBM SPSS Analítica de Texto para Pesquisas de Opinião . Consulte o tópico "Usando pacotes de análise de texto" na página 131 para obter mais informações.

Para Exportar Categorias Predefinidas

1. A partir dos menus do ambiente de trabalho interativo, escolha **Categorias> Gerenciar Categorias> Categorias de Exportação**. Um assistente Exportar Categorias é exibido.
2. Escolha o local e insira o nome do arquivo que será exportado.
3. Insira um nome para o arquivo de saída na caixa de texto Nome do Arquivo.
4. Para escolher o formato no qual você irá exportar seus dados da categoria, clique em **Avançar**.
5. Escolha o formato a partir do seguinte:
   - **Formato de lista Simples ou Compacta**: Veja o tópico "Formato de lista simples" na página 128 para obter mais informações. A lista Simples não contém subcategorias. Consulte o tópico "Formato compacto" na página 128 para obter mais informações. O formato de lista Compacta contém categorias hierárquicas.
   - **Formato indentado**: Veja o tópico "Formato indentado" na página 129 para obter mais informações.
6. Para começar escolhendo o conteúdo a ser exportado e para revisar os dados propostos, clique em **Avançar**.
7. Revise o conteúdo para o arquivo exportado.
8. Selecione ou desmarque as configurações de conteúdo adicionais a serem exportadas, tais como **Anotações** ou **Nomes de descritores**.
9. Para exportar as categorias, clique em **Concluir**.

Usando pacotes de análise de texto

Um pacote de análise de texto, também chamado de TAP, serve como um gabarito para categorização de resposta de texto. Usar um TAP é uma maneira fácil para você categorizar seus dados de texto com uma intervenção mínima já que ele contém o conjuntos de categorias pré-construídos e os recursos linguísticos que são necessários para codificar um vasto número de registros de forma rápida e automática. Usando os recursos linguísticos, os dados de texto são analisados e minerados para extrair os conceitos-chave. Com base em conceitos-chave e padrões encontrados no texto, os registros podem ser categorizados no conjunto de categorias que você selecionou no TAP. Você pode criar seu próprio TAP ou atualizar um.

Um TAP é composto pelos elementos a seguir:

- **Conjunto(s) de Categorias**. Um conjunto de categoria é constituído basicamente por categorias predefinidas, códigos de categoria, descritores para cada categoria e, por último, um nome para todo o conjunto da categoria. Descritores são elementos linguísticos (conceitos, tipos, padrões e regras) como

o termo *barato* ou o padrão *bom preço*. Os descritores são usados para definir uma categoria de modo que quando o texto corresponde a qualquer descritor de categoria, o documento ou registro é colocado na categoria.

- **Recursos Linguísticos**. Recursos linguísticos são um conjunto de bibliotecas e recursos avançados que são ajustados para extrair os principais conceitos e padrões. Esses conceitos e padrões de extração, por sua vez, são usados como os descritores que permitem que os registros sejam colocados em uma categoria no conjunto de categorias.

É possível criar seu próprio TAP, atualizar um ou carregar pacotes de análise de texto.

Depois de selecionar o TAP e escolher um conjunto de categorias, SPSS Análise de Texto do Modeler pode extrair e categorizar seus registros.

Nota: Os TAPs podem ser criados e utilizados de forma intercambiável entre SPSS Analítica de Texto para Pesquisas de Opinião e SPSS Análise de Texto do Modeler . No entanto, note que marcar em regras pode ser diferente em SPSS Análise de Texto do Modeler dependendo se você carregar um pacote de análise de texto (TAP) a partir de SPSS Análise de Texto do Modeler diretamente ou se você carrega um TAP de IBM SPSS Analítica de Texto para Pesquisas de Opinião . Recomendamos que você use TAPs que são feitos dentro de SPSS Análise de Texto do Modeler ; isto porque TAPs que são feitos em IBM SPSS Analítica de Texto para Pesquisas de Opinião podem ser criados usando uma versão diferente dos recursos linguísticos.

Criando Pacotes de Análise de Texto

Sempre que você tiver uma sessão com pelo menos uma categoria e alguns recursos, é possível criar um pacote de análise de texto (TAP) a partir do conteúdo da sessão de ambiente de trabalho interativo aberta. O conjunto de categorias e descritores (conceitos, tipos, regras ou saídas de padrão de TLA) pode ser criado em um TAP juntamente com todos os recursos linguísticos abertos no editor de recursos.

É possível ver o idioma para o qual os recursos foram criados. O idioma é configurado na guia Recursos Avançados do Editor de Template ou Editor de Recursos.

Para Criar um Pacote de Análise de Texto

1. A partir dos menus, escolha **Arquivo> Pacotes de Análise de Texto> Fazer Pacote**. O diálogo Criar Pacote aparece.
2. Navegue para o diretório no qual você salvará o TAP. Por padrão, os TAPs são salvos no subdiretório \TAP do diretório de instalação do produto.
3. Insira um nome para o TAP no campo **Nome do Arquivo**.
4. Insira um rótulo no campo **Rótulo do Pacote** . Ao inserir um nome de arquivo, esse nome aparecerá automaticamente como o rótulo, mas você pode mudar esse rótulo.
5. Para excluir um conjunto de categorias do TAP, desmarque a caixa de seleção **Incluir**. Fazer isso assegurará que ele não seja incluído em seu pacote. Por padrão, um conjunto de categorias por questão é incluído no TAP. Sempre deve haver pelo menos uma categoria configurada no TAP.
6. Renomear quaisquer conjuntos de categorias. A coluna **Novo Conjunto de Categorias** contém nomes genéricos por padrão, que são gerados ao incluir o prefixo `Cat_` no nome da variável de texto. Um único clique na célula torna o nome editável. Enter ou um clique em qualquer lugar aplica a renomeação. Se você renomear um conjunto de categorias, o nome muda apenas no TAP e não muda o nome da variável na sessão aberta.
7. Reordene os conjuntos de categoria se desejado usando as teclas de seta para a direita da tabela do conjunto de categorias.
8. Clique em **Salvar** para criar o pacote de análise de texto. A caixa de diálogo é fechada.

Carregando pacotes de análise de texto

Ao configurar um nó de modelagem de mineração de texto, deve-se especificar os recursos que serão usados durante a extração. Em vez de escolher um modelo de recurso, você pode selecionar um pacote de análise de texto (TAP) ou um projeto SPSS Analítica de Texto para Pesquisas de Opinião (`.tas`) a fim

de copiar não apenas seus recursos, mas também uma categoria definida no nó. Se você selecionar um arquivo `.tas` , ele será convertido em um TAP.

Os TAPs são mais interessantes ao criar um modelo de categoria de forma interativa, já que é possível usar a categoria definida como um ponto de início para categorização. Ao executar o fluxo, os lançamentos da sessão interativa do ambiente de trabalho e este conjunto de categorias aparecem no painel de categorias. Dessa forma, você escora seus documentos e registros imediatamente usando essas categorias e, em seguida, continua a refinar, construir e estender essas categorias até que elas satisfaçam as suas necessidades. Veja "Métodos e estratégias para criar categorias" na página 94 para obter mais informações.

A partir da versão 14, você também pode ver o idioma para o qual os recursos no TAP foram definidos quando você clica em **Carregar** e escolher o TAP.

Como carregar um TAP ou um TAS

1. Edite o nó de modelagem Mineração de Texto.
2. Na guia Models, escolha *Pacote de análise de texto* na seção **Copiar Recursos de** .
3. Clique em **Carregar**. O diálogo Carregar Pacote de Análise de Texto é aberto.
4. Navegue até a localização do TAP ou do projeto SPSS Analítica de Texto para Pesquisas de Opinião (`.tas`) contendo os recursos e o conjunto de categoria desejado para copiar no nó. Por padrão, eles são salvos no subdiretório \TAP do seu diretório de instalação do produto.
5. Insira um nome para o TAP no campo **Nome do Arquivo**. A etiqueta é exibida automaticamente.
6. Selecione o conjunto de categorias que você deseja usar. Este é o conjunto de categorias que aparecerá na sessão de ambiente de trabalho interativo. Você pode, então, ajustar e melhorar estas categorias manualmente ou usando as opções das categorias Construir ou Estender.
7. Clique em **Carregar** para copiar o conteúdo do pacote de análise de texto ou do projeto SPSS Analítica de Texto para Pesquisas de Opinião no nó. A caixa de diálogo é fechada. Quando os conteúdos são carregados, eles são copiados no nó; portanto, quaisquer alterações que você fizer aos recursos e categorias externas não serão refletidas a menos que você o atualize explicitamente e o recarregue.

Atualizando Pacotes de Análise de Texto

Se você fizer melhorias em um conjunto de categorias, recursos linguísticos ou criar um conjunto de categorias totalmente novo, será possível atualizar um pacote de análise de texto (TAP) para tornar mais fácil de reutilizar essas melhorias posteriormente. Para fazer isso, deve-se estar na sessão aberta contendo as informações que você deseja colocar no TAP. Ao atualizar, é possível escolher anexar conjuntos de categoria, substituir recursos, mudar o rótulo do pacote ou renomear/reordenar conjuntos de categorias.

Para Atualizar um Pacote de Análise de Texto

1. A partir dos menus, escolha **Arquivo> Pacotes de Análise de Texto> Pacote de Atualização**. O diálogo Atualizar Pacote aparece.
2. Navegue para o diretório que contém o pacote de análise de texto que você deseja atualizar.
3. Insira um nome para o TAP no campo **Nome do Arquivo**.
4. Para substituir os recursos linguísticos dentro do TAP por aqueles na sessão atual, selecione a opção **Substituir os recursos neste pacote por aqueles na sessão aberta**. Geralmente faz sentido atualizar os recursos linguísticos, já que eles foram usados para extrair os principais conceitos e padrões usados para criar as definições de categoria. Ter os recursos linguísticos mais recentes assegura que você obtenha os melhores resultados na categorização de seus registros. Se você não selecionar essa opção, os recursos linguísticos que já estavam no pacote são mantidos inalterados.
5. Para atualizar apenas os recursos linguísticos, certifique-se de selecionar a opção **Substituir os recursos neste pacote por aqueles na sessão aberta** e selecione apenas os conjuntos de categorias atuais que já estavam no TAP.

6. Para incluir o novo conjunto de categorias do aberto sessão no TAP, marque a caixa de seleção para cada conjunto de categorias a ser incluído. É possível incluir um, diversos ou nenhum dos conjuntos de categorias.
7. Para remover conjuntos de categorias do TAP, desmarque a caixa de seleção correspondente **Include** . Você pode escolher remover um conjunto de categorias que já estava no TAP já que você está incluindo um melhorado. Para isso, desmarque a caixa de seleção **Include** para o conjunto de categoria correspondente na coluna Conjunto de Categoria Atual. Sempre deve haver pelo menos uma categoria configurada no TAP.
8. Renomeie os conjuntos de categorias, se necessário. Um único clique na célula torna o nome editável. Enter ou um clique em qualquer lugar aplica a renomeação. Se você renomear um conjunto de categorias, o nome muda apenas no TAP e não muda o nome da variável na sessão aberta. Se dois conjuntos de categoria possuem o mesmo nome, os nomes aparecerão em vermelho até que você corrija a duplicata.
9. Para criar um novo pacote com o conteúdo da sessão mesclado com o conteúdo do TAP selecionado, clique em **Salvar Como Novo**. O diálogo Salvar como Pacote de Análise de Texto aparece. Veja as instruções a seguir.
10. Clique em **Atualizar** para salvar as mudanças feitas no TAP selecionado.

Para Salvar um Pacote de Análise de Texto

1. Navegue para o diretório no qual você salvará o arquivo do TAP. Por padrão, os arquivos TAP são salvos no subdiretório \TAP do diretório de instalação.
2. Digite um nome para o arquivo TAP no campo **Nome do arquivo** .
3. Insira um rótulo no campo **Rótulo do Pacote** . Ao inserir um nome de arquivo, esse nome é automaticamente usado como o rótulo. Entretanto, é possível renomear esse rótulo. Deve-se ter um rótulo.
4. Clique em **Salvar** para criar o novo pacote.

Editando e refinando categorias

Depois de criar algumas categorias, você invariavelmente desejará examiná-las e fazer alguns ajustes. Além de refinar os recursos linguísticos, você deve revisar suas categorias procurando maneiras de combinar ou limpar suas definições, bem como verificar alguns dos documentos ou registros categorizados. Também é possível revisar os documentos ou registros em uma categoria e fazer ajustes para que as categorias sejam definidas de tal forma que nuances e distinções sejam capturadas.

Você pode usar técnicas integradas e automatizadas de construção de categoria para criar suas categorias; entretanto, você provavelmente deseja executar alguns ajustes nessas categorias. Após usar uma ou mais técnicas, um número de novas categorias aparece na janela. É possível, então, revisar os dados em uma categoria e fazer ajustes até que você esteja confortável com suas definições de categoria. Consulte o tópico “Sobre Categorias” na página 98 para obter informações adicionais.

Aqui estão algumas opções para refinar suas categorias, a maioria das quais são descritas nas páginas a seguir:

Incluindo descritores nas categorias

Após usar técnicas automatizadas, você provavelmente ainda terá resultados de extração que não foram usados em qualquer uma das definições de categoria. Deve-se revisar esta lista na área de janela de Resultados da Extração. Se você localizar elementos que gostaria de mover em uma categoria, é possível incluí-las em uma categoria existente ou nova.

Para Incluir um Conceito ou Tipo em uma Categoria

1. A partir das áreas de janela Resultados da Extração e Dados, selecione os elementos que você deseja incluir em uma categoria nova ou existente.

2. A partir dos menus, escolha **Categorias> Adicionar à Categoria**. A caixa de diálogo Todas as Categorias exibe o conjunto de categorias. Selecione a categoria na qual você deseja incluir os elementos selecionados. Se você desejar incluir os elementos em uma nova categoria, selecione **Nova Categoria**. Uma nova categoria aparece na área de janela Categorias, usando o nome do primeiro elemento selecionado.

Editando descritores de categoria

Assim que você tiver criado algumas categorias, poderá abrir cada categoria para ver todos os descritores que compõem sua definição. Dentro da caixa de diálogo Definições de Categoria, você pode fazer várias edições em seus descritores de categoria. Além disso, se forem mostradas categorias na árvore de categorias, você também pode trabalhar com elas ali.

Para Editar uma Categoria

1. Selecione a categoria que você deseja editar na área de janela Categorias.
2. A partir dos menus, escolha **Visualizar> Definições de Categoria**. A caixa de diálogo Definições de Categoria é aberta.
3. Selecione o descritor que deseja editar e clique no botão da barra de ferramentas correspondente.

A tabela a seguir descreve cada botão da barra de ferramentas que você pode usar para editar suas definições de categoria.

Tabela 32. Botões e descrições da barra de ferramentas

ícones	**Descrição**
	Exclui os descritores selecionados da categoria.
	Move os descritores selecionados para uma categoria nova ou existente.
	Move os descritores selecionados na forma de uma regra de categoria & para uma categoria. Consulte o tópico “Usando regras de categoria” na página 115 para obter informações adicionais.
	Move cada um dos descritores selecionados como sua própria nova categoria
 **Exibir**	Atualiza o que é exibido na área de janela Dados e na área de janela Visualização de acordo com os descritores selecionados

Movendo categorias

Se você desejar colocar uma categoria em outra categoria existente ou mover descritores para outra categoria, você poderá movê-los.

Para Mover uma Categoria

1. Na área de janela Categorias, selecione as categorias que gostaria de mover para outra categoria.
2. A partir dos menus, escolha **Categorias> Mover-se para a Categoria**. O menu apresenta um conjunto de categorias com a categoria mais recentemente criada no topo da lista. Selecione o nome da categoria para a qual deseja mover os conceitos selecionados.

- Se você vir o nome que está procurando, selecione-o e os elementos selecionados serão incluídos em tal categoria.
- Se você não o vir, selecione **Mais** para exibir a caixa de diálogo Todas as Categorias e selecione a categoria na lista.

Comprimindo categorias

Quando você tem uma estrutura de categorias hierárquicas com categorias e subcategorias, é possível comprimir sua estrutura. Ao comprimir uma categoria, todos os descritores nas subcategorias de tal categoria são movidos na categoria selecionada e as então subcategorias vazias são excluídas. Desta maneira, todos os documentos que costumavam corresponder às subcategorias agora são categorizados na categoria selecionada.

Para Comprimir uma Categoria

1. Na área de janela Categorias, selecione uma categoria (de nível superior ou subcategoria) que você gostaria de comprimir.
2. A partir dos menus, escolha **Categorias> Categorias Flatten**. As subcategorias são movidas e os descritores são mesclados na categoria selecionada.

Mesclando ou combinando categorias

Se você quiser combinar duas ou mais categorias existentes em uma nova categoria, é possível mesclá-las. Ao mesclar categorias, uma nova categoria é criada com um nome genérico. Todos os conceitos, tipos e padrões usados nos descritores de categoria são movidos nesta nova categoria. Posteriormente, você pode renomear essa categoria ao editar as propriedades da categoria.

Para Mesclar uma Categoria ou Parte de uma Categoria

1. Na área de janela Categorias, selecione os elementos que você gostaria de mesclar.
2. A partir dos menus, escolha **Categorias> Mesclar Categorias**. A caixa de diálogo Propriedades da Categoria é exibida, na qual você insere um nome para a categoria recém-criada. As categorias selecionadas são mescladas na nova categoria como subcategorias.

Forçando documentos em categorias

Forçando documentos dentro e fora de categorias possibilita substituir as definições de categoria criadas pelas técnicas de construção automática da categoria sem alterar a definição de categoria real. Você pode achar que, embora o documento contenha termos que são usados para definir uma determinada categoria, o documento em si não deve estar nessa categoria. Nesse caso, você pode forçar o documento a sair dessa categoria sem precisar remover os termos da definição da categoria.

Forçar é usado em casos especiais onde um documento se encaixa (ou não se encaixa) uma categoria, mas por uma razão ou outra (por exemplo, ele contém um determinado termo) é atribuído (ou não atribuído) a essa categoria. Por exemplo, isso pode ocorrer quando um respondente usa sarcasmo em sua resposta, como "*A pizza foi ótima. Tenho certeza que todo mundo adora pizza queimada, fria.*" Vamos supor que você tenha tido uma categoria chamada `Pos: [<Food> + <Positive>]` para capturar opiniões positivas relativas à comida que um restaurante serve, e esta resposta é atribuída a essa categoria. Neste caso, você pode querer forçar essa resposta para fora da categoria.

Para forçar em nossa fora de categorias

1. De dentro do painel de Dados, selecione o documento que deseja forçar para dentro ou para fora de uma determinada categoria.
2. A partir dos menus, escolha **Categorias** > **Força Em** ou **Categorias** > **Força Fora**. Um submenu exibe a lista de categorias a partir das quais você pode selecionar.
3. Selecione a categoria para a qual ou a partir da qual você deseja forçar este documento. Se você tiver criado muitas categorias, algumas podem não estar visíveis no submenu.

- Neste caso, escolha **Mais** na parte inferior do submenu. A Caixa de diálogo de todas as categorias é aberta, na qual é possível selecionar a categoria e clicar em **OK** para aplicar a mudança.
- Se você deseja forçar o documento em uma nova categoria, selecione **Criar categoria vazia**. Uma nova categoria aparece na árvore de categoria usando um nome genérico.

Sempre que uma categoria contém um ou mais documentos forçados, uma pseudo-categoria chamada **Force In** ou **Force Out** é exibida abaixo do nome da categoria na árvore.

Para limpar um estado forçado

1. De dentro do painel de Dados, selecione o documento que você não deseja mais forçar para dentro ou para fora de uma categoria.
2. A partir dos menus, escolha **Categorias> Força Em** para forçar em, ou escolha **Categorias> Força Fora** para forçar a saída. As categorias em que o documento é forçado a sair ou para dentro são precedidas de uma marca de checo.
3. Selecione a categoria no submenu que é verificado e para o qual deseja remover a força. A marca de verificação é removida e o documento não é mais forçado.

Para limpar todos os estados forçados

1. De dentro do painel de Dados, selecione um registro contendo uma **Force In** ou **Force Out**.
2. A partir dos menus, escolha **Categorias** > **Limpar Tudo** > **Força Ins** ou **Categorias** > **Limpar Tudo** > **Força Outs**. O estado forçado nos documentos é apurado e eles não são mais forçados dentro ou fora das categorias.

Nota: Este recurso só está disponível se o seu texto de origem contiver um ID exclusivo. Se o texto-fonte não tiver um ID exclusivo, você poderá adicionar um nó de Derivação entre o documento de origem e o nó Mining de Texto. Esse recurso só tem um impacto ao executar uma sessão interativa. Ao implantar o modelo de categoria para pontuação não interativa, esta peça de informação não é preservada ou usada, uma vez que é baseada em um ID de documento.

Excluindo categorias

Se você não quiser mais manter uma categoria, é possível excluí-la.

Para Excluir uma Categoria

1. Na área de janela Categorias, selecione a categoria ou categorias que você gostaria de excluir.
2. A partir dos menus, escolha **Editar> Excluir**.

Capítulo 10. Analisando clusters

Você pode construir e explorar clusters de conceito na visualização de Clusters (**View** > **Clusters**). Um *cluster* é um agrupamento de conceitos relacionados gerado pelos algoritmos de clusterização com base na frequência com que esses conceitos ocorrem no documento/registro configurado e na frequência com que eles aparecem juntos no mesmo documento, também conhecido como *coocorrência*. Cada conceito em um cluster coocorre com pelo menos um outro conceito no cluster. O objetivo dos clusters é agrupar conceitos que coocorram juntos, enquanto o objetivo das categorias é agrupar documentos ou registros com base em como o texto que eles contêm corresponde aos descritores (conceitos, regras, padrões) para cada categoria.

Um bom cluster é aquele com conceitos fortemente ligados, que coocorre frequentemente e com algumas ligações com conceitos em outros clusters. Durante o trabalho com conjuntos de dados maiores, essa técnica pode resultar em tempos de processamento significativamente mais longos.

O armazenamento em cluster é um processo que começa analisando um conjunto de conceitos e procurando conceitos que coocorrem com frequência nos documentos. Dois conceitos que coocorrem em um documento são considerados um par de conceitos. Em seguida, o processo de armazenamento em cluster avalia o *valor de similaridade* de cada par de conceitos comparando o número de documentos em que o par ocorre junto com o número de documentos em que ocorre cada conceito. Veja o tópico “Calculando valores de ligação de similaridade” na página 141 para obter mais informações.

Por último, o processo de armazenamento em cluster agrupa conceitos semelhantes em clusters por agregação e leva em consideração seus valores de ligação e as configurações definidas na caixa de diálogo Construir Clusters. Com agregação, queremos dizer que conceitos são incluídos ou clusters menores são mesclados a um cluster maior até o cluster ser saturado. Um cluster é *saturado* quando uma mesclagem adicional de conceitos ou clusters menores o faria exceder as configurações na caixa de diálogo Construir Clusters (número de conceitos, ligações internas ou ligações externas). Um cluster leva o nome do conceito dentro do cluster que tem o número geral mais alto de ligações com outros conceitos dentro do cluster.

No final, nem todos os pares de conceitos terminam juntos no mesmo cluster, já que pode haver uma forte ligação em outro cluster ou a saturação pode evitar a mesclagem dos clusters nos quais eles ocorrem. Por esse motivo, há ligações internas e externas.

- *Ligações internas* são aquelas entre pares de conceitos dentro de um cluster. Nem todos os conceitos são vinculados uns aos outros em um cluster. No entanto, cada conceito é vinculado a pelo menos um outro conceito dentro do cluster.
- *Ligações externas* são aquelas entre pares de conceitos em clusters separados (um conceito dentro de um cluster e um conceito fora de outro cluster).

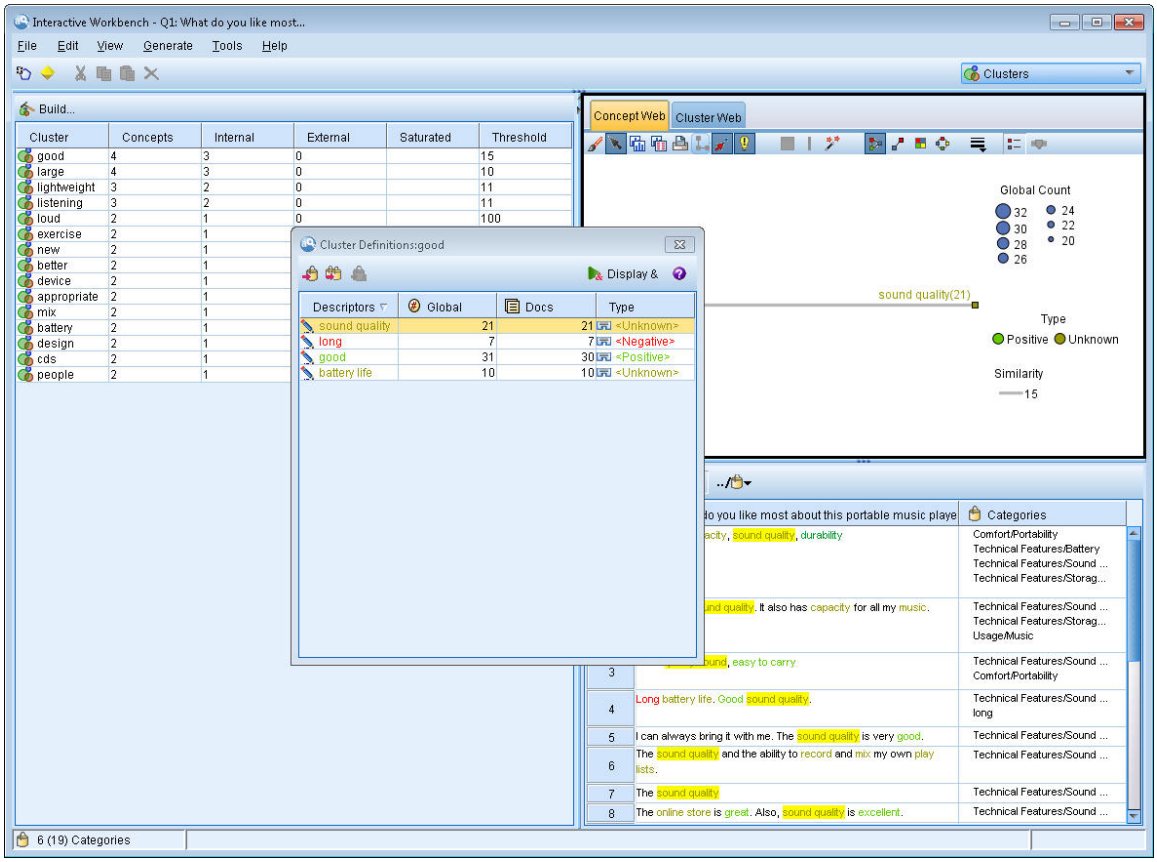


Figura 30. Visualização Clusters

A visualização Clusters é organizada em três áreas de janela, sendo que cada uma pode ser oculta ou mostrada selecionando seu nome no menu Visualizar:

- **Pane em clusters** Você pode construir e gerenciar seus clusters neste painel. Veja o tópico "Explorando clusters" na página 142 para obter mais informações.
- **Painel de visualização** Você pode explorar visualmente seus clusters e como eles interagem nesta pane. Veja o tópico "Gráficos de cluster" na página 157 para obter mais informações.
- **Painel de dados** Você pode explorar e revisar o texto contido dentro de documentos e registros que correspondem a seleções na caixa de diálogo Definições de Cluster. Veja o tópico "Definições de cluster" na página 143 para obter mais informações.

Construindo clusters

Quando você acessa a visualização Clusters pela primeira vez, não há nenhum cluster visível. Você pode construir os clusters através dos menus (**Ferramentas> Construir Clusters**) ou clicando na **Construir ...** na barra de ferramentas. Essa ação abre a caixa de diálogo Construir Clusters na qual é possível definir as configurações e os limites para a construção de seus clusters.

Nota: Sempre que os resultados da extração não correspondem mais aos recursos, essa área de janela fica amarela, assim como a área de janela Resultados da Extração. É possível fazer a extração novamente para obter os resultados da extração mais recentes e a coloração amarela desaparecerá. No entanto, cada vez que uma extração é executada, a área de janela Clusters é limpa e você precisa reconstruir seus clusters. Da mesma forma, os clusters não são salvos de uma sessão para outra.

As áreas e os campos a seguir estão disponíveis na caixa de diálogo Construir Clusters:

Entradas

Tabela de entradas Clusters são construídos a partir de descritores derivados de determinados tipos. Na tabela, é possível selecionar os tipos para incluir no processo de construção. Os tipos que capturam a maioria dos registros ou documentos são pré-selecionados por padrão.

Conceitos para cluster: Escolha o método para selecionar conceitos que você deseja usar para armazenamento em cluster. Reduzindo o número de conceitos, é possível acelerar o processo de armazenamento em cluster. É possível armazenar em cluster usando um número máximo de conceitos, uma porcentagem máxima de conceitos ou usando todos os conceitos:

- **Número baseado no doc. count** Quando você selecionar **Número superior de conceitos**, digite o número de conceitos a serem considerados para clustering. Os conceitos são escolhidos com base naqueles que têm o valor de contagem de doc mais alto. A contagem de doc é o número de documentos ou registros em que o conceito aparece. O valor máximo é de 150.000.
- **Percentual baseado no doc. count** Quando você selecionar **Principal porcentagem de conceitos**, digite a porcentagem de conceitos a serem considerados para clustering. Os conceitos são escolhidos com base nessa porcentagem de conceitos com o valor de contagem de doc mais alto.

Limites de Saída

Número máximo de clusters a criar Este valor é o número máximo de clusters para gerar e exibir na pane de Clusters. Durante o processo de armazenamento em cluster, clusters saturados são apresentados antes dos não saturados e, portanto, muitos dos clusters resultantes serão saturados. Para ver mais clusters não saturados, é possível mudar essa configuração para um valor maior que o número de clusters saturados.

Conceitos máximos em um cluster Este valor é o número máximo de conceitos que um cluster pode conter.

Conceitos mínimos em um cluster Este valor é o número mínimo de conceitos que devem ser vinculados a fim de criar um cluster.

Número máximo de links internos Este valor é o número máximo de links internos que um cluster pode conter. Ligações internas são aquelas entre pares de conceitos dentro de um cluster.

Número máximo de links externos Este valor é o número máximo de links para conceitos fora do cluster. Ligações externas são aquelas entre pares de conceitos em clusters separados.

Valor do link mínimo Este valor é o menor valor de link aceito para um par de conceito a ser considerado para clustering. O valor da ligação é calculado usando uma fórmula de similaridade. Veja o tópico “Calculando valores de ligação de similaridade” na página 141 para obter mais informações.

Evitar emparelhamento de conceitos específicos. Selecione esta caixa de seleção para impedir que o processo agrupe ou pareie dois conceitos na saída. Para criar ou gerenciar pares de conceitos, clique em **Gerenciar pares**. Consulte o tópico “Gerenciando pares de exceção de link” na página 106 para obter mais informações.

Calculando valores de ligação de similaridade

Saber apenas o número de documentos em que um par de conceitos coocorre não lhe diz quão semelhantes os dois conceitos são. Nesses casos, o valor de similaridade pode ser útil. O valor de ligação de similaridade é medido usando a contagem de documentos de coocorrência comparada com as contagens de documento individuais para cada conceito no relacionamento. Durante o cálculo de similaridade, a unidade de medida é o número de documentos (contagem de doc) em que um conceito ou par de conceitos é localizado. Um conceito ou par de conceitos é "localizado" em um documento, caso ocorra *pelo menos* uma vez no documento. É possível escolher ter a espessura da linha no gráfico Conceito representando o valor de ligação de similaridade nos gráficos.

O algoritmo revela aqueles relacionamentos que são mais fortes, o que significa que a tendência para os conceitos aparecerem juntos nos dados de texto é muito maior do que sua tendência de ocorrer independentemente. Internamente, o algoritmo gera um coeficiente de similaridade que varia de 0 a

1, em que um valor de 1 significa que dois conceitos sempre aparecem juntos e nunca separados. O resultado do coeficiente de similaridade é então multiplicado por 100 e arredondado para o número inteiro mais próximo. O coeficiente de similaridade é calculado usando a fórmula mostrada na figura a seguir.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Figura 31. Fórmula de coeficiente de similaridade

em que:

- C_I é o número de documentos ou registros em que o conceito `I` ocorre.
- C_J é o número de documentos ou registros nos quais ocorre o conceito `J`.
- C_{IJ} é o número de documentos ou registros em que o par de conceitos `I` e `J` coocorre no conjunto de documentos.

Por exemplo, suponha que você tenha 5.000 documentos. Deixe `I` e `J` sejam extraídos conceitos e deixe `IJ` ser um par de conceito coocorrência de `I` e `J`. A tabela a seguir propõe dois cenários para demonstrar como o valor do coeficiente e do link são calculados.

Tabela 33. Exemplo de frequências de conceito

Conceito/Par	**Cenário A**	**Cenário B**
Conceito: `I`	Ocorre em 20 docs	Ocorre em 30 docs
Conceito: `J`	Ocorre em 20 docs	Ocorre em 60 docs
Par de Conceitos: `IJ`	Coocorre em 20 docs	Coocorre em 20 docs
Coeficiente de similaridade	1	0.22222
Valor de ligação de similaridade	100	22

No cenário A, os conceitos `I` e `J`, bem como o par `IJ`, ocorrem em 20 documentos, gerando um coeficiente de similaridade de 1, o que significa que os conceitos sempre ocorrerão juntos. O valor de ligação de similaridade para esse par seria 100.

No cenário B, o conceito `I` ocorre em 30 documentos e o conceito `J` ocorre em 60 documentos, mas o par `IJ` ocorre em somente 20 documentos. Como resultado, o coeficiente de similaridade é 0,22222. O valor de ligação de similaridade para esse par seria arredondado para 22.

Explorando clusters

Após você construir clusters, é possível ver um conjunto de resultados na área de janela Clusters. Para cada cluster, as seguintes informações estão disponíveis na tabela:

- **Cluster.** Este é o nome do cluster. Clusters são nomeados após o conceito com o número mais alto de ligações internas.
- **Conceitos.** Este é o número de conceitos no cluster. Veja o tópico “Definições de cluster” na página 143 para obter mais informações.
- **Interno.** Este é o número de ligações internas no cluster. Ligações internas são aquelas entre pares de conceitos dentro de um cluster.
- **Externo.** Este é o número de ligações externas no cluster. Ligações externas são aquelas entre pares de conceitos quando um conceito está em um cluster e o outro conceito está em outro cluster.
- **Sat.** Se houver um símbolo presente, ele indica que este cluster poderia ter sido maior, mas um ou mais limites seriam excedidos e, portanto, o processo de armazenamento em cluster seria finalizado para esse cluster e considerado *saturado*. No final do processo de armazenamento em cluster, os clusters saturados são apresentados antes dos não saturados e, portanto, muitos dos clusters resultantes serão

saturados. Para ver mais clusters não saturados, é possível mudar a configuração **Número máximo de clusters para criar** para um valor maior que o número de clusters saturados ou diminuir o **Valor de ligação mínimo**. Veja o tópico "Construindo clusters" na página 140 para obter mais informações.

- **Limite.** Para todos os pares de conceitos de coocorrência no cluster, este é o valor de ligação de similaridade mais baixo de todos no cluster. Veja o tópico "Calculando valores de ligação de similaridade" na página 141 para obter mais informações. Um cluster com um valor limite alto significa que os conceitos nesse cluster têm uma similaridade geral maior e estão mais fortemente relacionados do que aqueles em um cluster cujo valor limite é menor.

Para saber mais sobre um determinado clusters, é possível selecioná-lo e a área de janela de visualização à direita mostrará dois gráficos para ajudá-lo a explorar o(s) cluster(s). Consulte o tópico "Gráficos de cluster" na página 157 para obter informações adicionais. Também é possível cortar e colar o conteúdo da tabela em outros aplicativos.

Sempre que os resultados da extração não correspondem mais aos recursos, essa área de janela fica amarela, assim como a área de janela Resultados da Extração. É possível fazer a extração novamente para obter os resultados da extração mais recentes e a coloração amarela desaparecerá. No entanto, cada vez que uma extração é executada, a área de janela Clusters é limpa e você precisa reconstruir seus clusters. Da mesma forma, os clusters não são salvos de uma sessão para outra.

Definições de cluster

É possível ver todos os conceitos dentro de um cluster, selecionando-o na área de janela Clusters e abrindo a caixa de diálogo Definições de Cluster (**View> Definições de Cluster**).

Todos os conceitos no cluster selecionado aparecem na caixa de diálogo Definições de Cluster. Se você selecionar um ou mais conceitos na caixa de diálogo de Definições de Cluster e clicar em **Exibir &**, o painel de dados exibirá todos os registros ou documentos em que *todos os conceitos selecionados aparecerão juntos*. No entanto, a área de janela Dados não exibe nenhum registro de texto ou documento quando você seleciona um cluster na área de janela Clusters. Para obter informações gerais sobre a área de janela Dados, consulte em.

A seleção de conceitos nessa caixa de diálogo também muda o gráfico da web de conceito. Consulte o tópico "Gráficos de cluster" na página 157 para obter informações adicionais. Da mesma forma, quando você seleciona um ou mais conceitos na caixa de diálogo Definições de Cluster, a área de janela Visualização mostra todas as ligações externas e internas a partir desses conceitos.

Descrições de coluna

Ícones são mostrados para que você identifique facilmente cada descritor.

Tabela 34. Ícones de colunas e descritor

Colunas	**Descrição**
Descritores	O nome do conceito.
 **Global**	Mostra o número de vezes que este descritor aparece no conjunto de dados inteiro, conhecido como frequência global.
 **Docs**	Mostra o número de documentos ou registros em que este descritor aparece, conhecido como frequência de documento.
Tipo	Mostra o tipo ou tipos aos quais o descritor pertence. Se o descritor for uma regra de categoria, nenhum nome de tipo será mostrado nessa coluna.

Ações da barra de ferramentas

Nesta caixa de diálogo, é possível selecionar um ou mais conceitos para usar em uma categoria. Há diversas maneiras de se fazer isso, mas a mais interessante é selecionar conceitos que acompanham um cluster e incluí-los como uma regra de categoria. Consulte o tópico "Regras de coocorrência" na página

109 para obter informações adicionais. É possível usar os botões da barra de ferramentas para incluir os conceitos nas categorias.

Tabela 35. Botões da barra de ferramentas para incluir conceitos em categorias

ícones	**Descrição**
	Inclui os conceitos selecionados em uma categoria nova ou existente
	Inclui os conceitos selecionados em forma de uma regra de categoria & em uma categoria nova ou existente. Consulte o tópico “Usando regras de categoria” na página 115 para obter informações adicionais.
	Inclui cada um dos conceitos selecionados como sua própria nova categoria
	Atualiza o que é exibido na área de janela Dados e na área de janela Visualização de acordo com os descritores selecionados

Nota: Também é possível incluir conceitos em um tipo, como sinônimos, ou excluir itens usando os menus de contexto.

Capítulo 11. Explorando a análise de ligação de texto

Na visualização Análise de Ligação de Texto (TLA), é possível explorar os resultados do padrão de análise de ligação de texto. A análise de ligação de texto é uma tecnologia de correspondência de padrões que permite definir regras padrão e compará-las com os conceitos e relacionamentos reais extraídos localizados em seu texto.

Por exemplo, a extração de ideias sobre uma organização pode não ser interessante o suficiente para você. Usando TLA, você também poderia aprender sobre as ligações entre essa organização e as outras organizações ou as pessoas dentro de uma organização. Também é possível usar TLA para extrair opiniões sobre produtos ou, para alguns idiomas, os relacionamentos entre genes.

Após a extração de alguns resultados do padrão TLA, é possível revisá-los nas áreas de janela Padrões de Tipo e Padrões de Conceito da visualização Análise de Ligação de Texto. Consulte o tópico “Padrões de Tipo e Conceito” na página 147 para obter informações adicionais. É possível explorá-los ainda mais nas áreas de janela Dados ou Visualização nessa visualização. Provavelmente o mais importante, é possível incluí-los em categorias.

Se você ainda não tiver escolhido fazer isso, é possível clicar em **Extrair** e escolher **Ativar extração do padrão Análise de Ligação de Texto** na caixa de diálogo Configurações de Extração. Consulte o tópico “Extraindo resultados do padrão de TLA” na página 146 para obter mais informações.

Deve haver algumas regras do padrão TLA definidas no modelo de recurso ou nas bibliotecas sendo usadas para extrair resultados do padrão TLA. É possível usar os padrões TLA em certos modelos de recurso fornecidos com IBM SPSS Modeler Text Analytics. O tipo dos relacionamentos e padrões que podem ser extraídos dependem totalmente das regras TLA definidas em seus recursos. Você pode definir suas próprias regras de TLA. Padrões são compostos por macros, listas de palavras e diferenças de palavras para formar um query booleano, ou regra, que seja comparado com seu texto de entrada. Consulte o tópico Capítulo 18, “Sobre regras de ligação de texto”, na página 207 para obter mais informações.

Sempre que uma regra do padrão TLA corresponde a um texto, esse texto pode ser extraído como um padrão e reestruturado como dados de saída. Os resultados ficam então visíveis nas áreas de janela da visualização Análise de Ligação de Texto. Cada área de janela pode ficar oculta ou ser mostrada selecionando seu nome no menu Visualização:

- **Áreas de janela Padrões de Tipo e Padrões de Conceito.** É possível construir e explorar seus padrões nessas duas áreas de janela. Veja o tópico “Padrões de Tipo e Conceito” na página 147 para obter mais informações.
- **Painel de visualização.** É possível explorar visualmente como os conceitos e os tipos em seus padrões interagem nessa área de janela. Veja o tópico “Gráficos Análise de Ligação de Texto” na página 158 para obter mais informações.
- **Painel de dados.** É possível explorar e revisar o texto contido dentro dos documentos e registros que correspondem às seleções em outra área de janela. Veja o tópico “Área de janela Dados” na página 149 para obter mais informações.

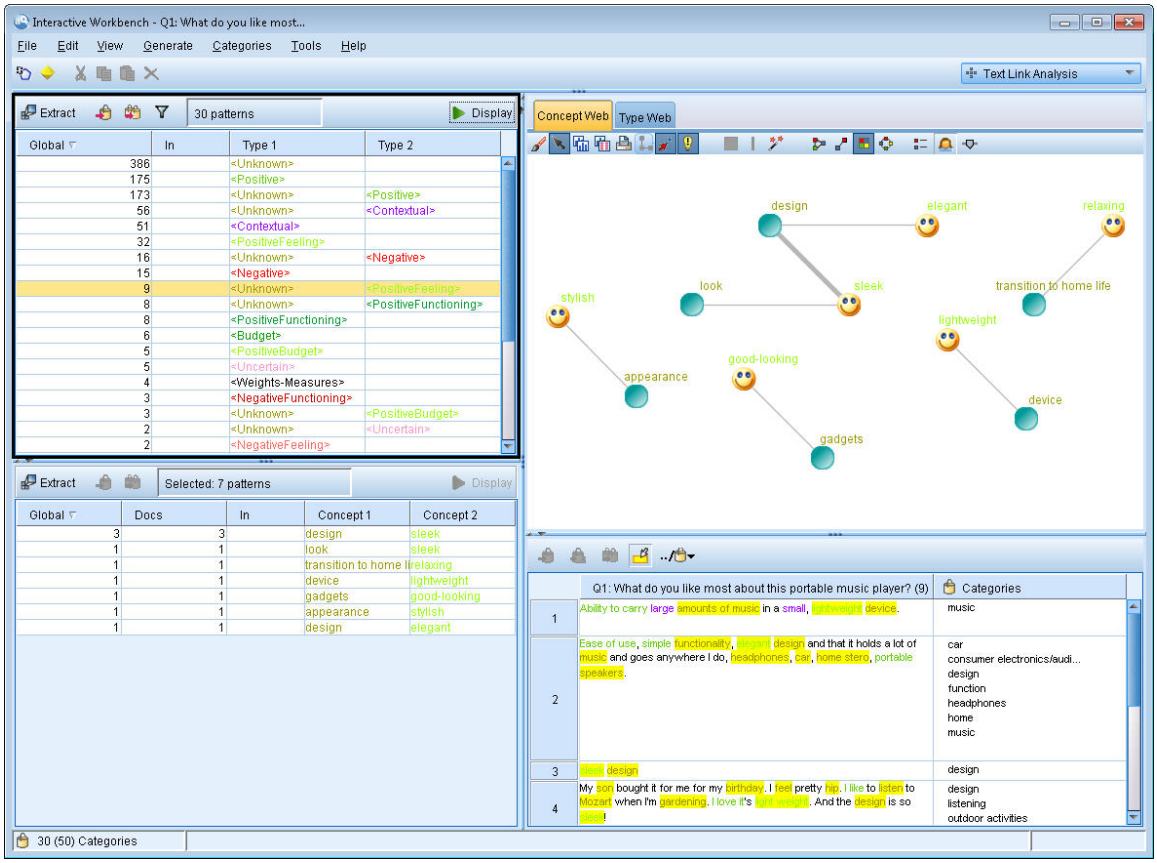


Figura 32. Visualização Análise de Ligação de Texto

Extraindo resultados do padrão de TLA

O processo de extração resulta em um conjunto de conceitos e tipos, bem como padrões de Análise de Link de Texto (TLA), se ativados. Se você extraiu padrões de TLA, é possível vê-los na visualização Análise de Ligação de Texto. Sempre que os resultados da extração não estiverem em sincronização com os recursos, as áreas de janela Padrões ficarão amarelas para indicar que uma nova extração produziria resultados diferentes.

Você precisa escolher extrair esses padrões na configuração do nó ou na caixa de diálogo Extrair usando a opção **Ativar extração do padrão de Análise de Ligação de Texto**. Veja o tópico “Extraindo dados” na página 80 para obter mais informações.

Nota: Há um relacionamento entre o tamanho do seu conjunto de dados e o tempo levado para se concluir o processo de extração. Consulte as instruções de instalação para conhecer as recomendações e as estatísticas de desempenho. Também é possível considerar inserir um envio de dados do nó Amostra ou otimizar a configuração de sua máquina.

Para Extrair Dados

1. A partir dos menus, escolha **Ferramentas> Extração**. Como alternativa, clique no botão da barra de ferramentas **Extrair**.
2. Mude qualquer uma das opções que deseja usar. Lembre-se de que a opção **Ativar extração do padrão de Análise de Ligação de Texto** deve estar selecionada nesta guia, bem como as regras de TLA em seu modelo para extrair os resultados do padrão de TLA. Veja o tópico “Extraindo dados” na página 80 para obter mais informações.
3. Clique em **Extrair** para começar o processo de extração.

Assim que a extração começar, a caixa de diálogo será aberta. Se desejar interromper a extração, clique em **Cancelar**. Quando a extração for concluída, a caixa de diálogo fechará e os resultados aparecerão na área de janela. Veja o tópico “Padrões de Tipo e Conceito” na página 147 para obter mais informações.

Padrões de Tipo e Conceito

Padrões são compostos por duas partes, uma combinação de conceitos e tipos. Padrões são mais úteis quando você está tentando descobrir pareceres sobre um determinado assunto ou relacionamentos entre conceitos. Por exemplo, a extração do nome do produto do seu concorrente pode não ser interessante o suficiente para você. Nesse caso, é possível consultar os padrões extraídos para ver se é possível localizar exemplos nos quais um documento ou registro contém texto expressando que o produto é bom, ruim ou caro.

Padrões podem consistir em até seis tipos ou seis conceitos. Por esse motivo, as linhas em ambas as áreas de janela de padrão contêm até seis slots ou ranqueamentos. Cada slot corresponde ao ranqueamento específico de um elemento na regra de padrão de TLA, conforme definido nos recursos linguísticos. No ambiente de trabalho interativo, se um slot não contiver valores, ele não será mostrado na tabela. Por exemplo, se os resultados do padrão mais longo contiverem no máximo quatro slots, os dois últimos não serão mostrados. Consulte o tópico Capítulo 18, “Sobre regras de ligação de texto”, na página 207 para obter informações adicionais.

Quando você extrai resultados de padrão, eles são agrupados primeiro em nível de tipo e depois divididos em padrões de conceito. Por esse motivo, há duas áreas de janela de resultado diferentes: **Padrões de Tipo** (superior esquerda) e **Padrões de Conceito** (inferior esquerda). Para ver todos os padrões de conceito retornados, selecione todos os padrões de tipo. A área de janela de padrões de conceito inferior exibirão todos os padrões de conceito até o valor de ranqueamento máximo (conforme definido na caixa de diálogo Filtro).

Padrões do tipo Este pano de janela apresenta resultados de padrões constituídos por um ou mais tipos relacionados que correspondem a uma regra de padrão TLA. Os padrões do tipo são mostrados como `<Organization>` + `<Location>` + `<Positive>`, o que pode fornecer feedback positivo sobre uma organização em um local específico. A sintaxe e a seguinte:

```
<Type1> + <Type2> + <Type3> + <Type4> + <Type5> + <Type6>
```

Padrões de Conceito Este pano de janela apresenta os resultados do padrão no nível de conceito para todos os (s) padrão (s) tipo (s) atualmente selecionados no painel Type Patterns acima dele. Os padrões de conceito seguem uma estrutura como `hotel` + `paris` + `wonderful`. A sintaxe é a seguinte:

```
concept1 + concept2 + concept3 + concept4 + concept5 + concept6
```

Quando resultados do padrão usam menos do que o máximo de seis slots, somente o número necessário de slots (ou colunas) é exibido. Quaisquer slots vazios localizados entre dois slots preenchidos são descartados, de modo que o padrão `<Type1>+<>+<Type2>+<>+<>+<>` possa ser representado por `<Type1>+<Type3>`. Para um padrão de conceito, este seria `concept1+.+concept2` (onde `.` representa um valor nulo).

Assim como acontece com os resultados da extração na visualização Categorias e Conceitos, é possível visualizar os resultados aqui. Se vir algum refinamento que gostaria de fazer nos tipos e conceitos que compõem esses padrões, você os fará na área de janela Resultados da Extração na visualização Categorias e Conceitos ou diretamente no Editor de Recurso e extrairá seus padrões novamente. Sempre que um conceito, tipo ou padrão é usado em uma definição de categoria no estado em que se encontra ou como parte de uma regra, um ícone de categoria ou regra aparece na coluna **Em** na tabela Resultados da Extração ou Padrão.

Nota: Se houver mais resultados que possam caber na pane visível, você pode usar os controles na parte inferior da pane para mover para frente e para trás através dos resultados, ou inserir um número de página para ir até.

Filtrando resultados de TLA

Quando você está trabalhando com conjuntos de dados muito grandes, o processo de extração pode produzir milhões de resultados. Para muitos usuários, essa quantidade pode dificultar a revisão efetiva dos resultados. É possível, no entanto, filtrar esses resultados para aumentar o zoom nos mais

interessantes. É possível mudar as configurações na caixa de diálogo Filtro para limitar quais padrões serão mostrados. Todas essas configurações são usadas juntas.

Na visualização TLA, a caixa de diálogo Filtro contém as seguintes áreas e campos.

Filtrar por Frequência Você pode filtrar para exibir apenas esses resultados com um determinado valor de frequência global ou de documento.

- **Frequência Global** é o número total de vezes que um padrão aparece no conjunto inteiro de documentos ou registros e é mostrado na coluna **Global**.
- **Frequência de Documento** é o número total de documentos ou registros em que um padrão aparece e é mostrado na coluna **Docs**.

Por exemplo, se um padrão apareceu 300 vezes em 500 registros, poderíamos dizer que esse padrão tem uma frequência global de 300 e uma frequência de documento de 500.

E pelo Match Text Você também pode filtrar para exibir apenas aqueles resultados que correspondem à regra que você define aqui. Insira o conjunto de caracteres a ser correspondido no campo **Texto de Correspondência** e selecione se deseja procurar esse texto dentro dos nomes de conceito ou tipo, identificando o número do slot ou todos. Em seguida, selecione a condição à qual aplicar a correspondência (não é necessário usar sinal de maior e menor para denotar o início ou o fim de um nome de tipo). Selecione **And** ou **Or** da lista suspensa para que a regra corresponda a ambas as instruções ou a apenas uma delas e defina a segunda instrução correspondente ao texto da mesma maneira que a primeira.

Tabela 36. Condições do texto de correspondência

Condição	**Descrição**
Contém	Texto é correspondido se a sequência de caracteres ocorrer em qualquer lugar. (Opção padrão)
Inicia com	Texto é correspondido somente se o conceito ou tipo começar com o texto especificado.
Termina com	Texto é correspondido somente se o conceito ou tipo terminar com o texto especificado.
Correspondência exata	A sequência de caracteres inteira deve corresponder ao nome do conceito ou tipo.

Resultados exibidos na área de janela Padrões

Suponha que você esteja usando uma versão em inglês do software; aqui estão alguns exemplos de como os resultados podem ser exibidos na barra de ferramentas da área de janela Padrões baseada em filtros.

Figura 33. Exemplo dos resultados do filtro 1

Neste exemplo, a barra de ferramentas mostra que o número de padrões retornado foi limitado devido ao ranqueamento máximo especificado no filtro. Se um ícone púrpura estiver presente, isso significa que o número máximo de padrões foi atendido. Passe o mouse sobre o ícone para obter mais informações. Consulte a explicação precedente do filtro **E por Ranqueamento**.

Figura 34. Exemplo dos resultados do filtro 2

Neste exemplo, a barra de ferramentas mostra que os resultados foram limitados usando um filtro de texto de correspondência (veja o ícone de lupa). É possível passar o mouse sobre o ícone para ver qual é o texto de correspondência.

Para filtrar os resultados

1. A partir dos menus, escolha **Ferramentas** > **Filtro**. A caixa de diálogo Filtro é aberta.
2. Selecione e refine os filtros que deseja usar.
3. Clique em **OK** para aplicar os filtros e ver os novos resultados.

Área de janela Dados

Conforme extrai e explora os padrões de análise de ligação de texto, talvez você queira revisar alguns dos dados com os quais está trabalhando. Por exemplo, talvez você queira ver os registros reais nos quais um grupo de padrões foi descoberto. É possível revisar registros ou documentos na área de janela Dados, que está localizada no lado inferior direito. Se não visível por padrão, escolha **Visualizar** > **Panes** > **Dados** a partir dos menus.

A área de janela Dados apresenta uma linha por documento ou registro correspondente a uma seleção na visualização até certo limite de exibição. Por padrão, o número de documentos ou registros mostrados na área de janela Dados é limitado para deixar sua visualização dos dados mais rápida. No entanto, é possível ajustar isso na caixa de diálogo Opções. Veja “Opções: guia Sessão” na página 73 para obter mais informações.

Nota: Se houver mais resultados que possam caber na pane visível, você pode usar os controles na parte inferior da pane para mover para frente e para trás através dos resultados, ou inserir um número de página para ir até.

Exibindo e Atualizando a Área de Janela Dados

A área de janela Dados não atualiza sua exibição automaticamente, pois com dados automáticos de conjuntos de dados grandes, a atualização levaria muito tempo para ser concluída. Portanto, sempre que você seleciona padrões de tipo ou conceito nessa visualização, é possível clicar em **Exibir** para atualizar o conteúdo da área de janela Dados.

Documentos de texto ou registros

Se seus dados de texto estiverem no formato de registros e o texto for relativamente pequeno, o campo de texto na área de janela Dados exibirá os dados de texto como um todo. No entanto, quando você trabalha com registros e conjuntos de dados maiores, a coluna do campo de texto mostra um pequeno pedaço do texto e abre uma área de janela Visualização de Texto à direita para exibir mais ou todo o texto do registro que você selecionou na tabela. Se seus dados de texto estiverem no formato de documentos individuais, a área de janela Dados mostrará o nome do arquivo do documento. Quando você seleciona um documento, a área de janela Visualização de Texto é aberta com o texto do documento selecionado.

Cores e destaque

Sempre que você exibe dados, os conceitos e descritores localizados nesses documentos ou registros são destacados em cores para ajudá-lo a identificá-los facilmente no texto. A codificação de cor corresponde aos tipos aos quais os conceitos pertencem. Também é possível passar o mouse sobre os itens codificados por cores para exibir o conceito sob o qual eles foram extraído e o tipo ao qual eles foram designados. Qualquer texto não extraído aparece em preto. Normalmente, essas palavras não extraídas costumam ser conectores (*e* ou *com*), pronomes (*mim* ou *eles*) e verbos (*ser*, *ter* ou *tomar*).

Colunas da área de janela Dados

Embora a coluna do campo de texto esteja sempre visível, também é possível exibir outras colunas. Para exibir outras colunas, escolha **Visualizar** > **Pane de Dados** a partir dos menus e, em seguida, selecione a coluna que deseja exibir no painel de Dados. As colunas a seguir podem estar disponíveis para exibição:

- **"Nome do campo de texto" (#)/Documentos.** Inclui uma coluna para os dados de texto dos quais conceitos e tipos foram extraídos. Se seus dados estiverem em documentos, a coluna será chamada Documentos e somente o nome do arquivo ou caminho completo do documento ficará visível. Para ver o texto para esses documentos, deve-se consultar a área de janela Visualização de Texto. O número de linhas na área de janela Dados é mostrado entre parênteses após o nome dessa coluna. Pode haver momentos em que nem todos os documentos ou registros são mostrados devido a um limite no diálogo

Opções usado para aumentar a velocidade do carregamento. Se o máximo for atingido, o número será seguido por **-Max**. Consulte “Opções: guia Sessão” na página 73 para obter mais informações.

- **Categorias.** Lista cada uma das categorias às quais um registro pertence. Sempre que essa coluna é mostrada, a atualização da área de janela Dados pode demorar um pouco mais para mostrar as informações mais atuais.
- **Força Em.** Lista as categorias em que você forçou um documento. Os documentos podem ser forçados na categoria por meio da seleção de menu **Editar** > **Força In** . Consulte “Forçando documentos em categorias” na página 136 para obter mais informações.
- **Força Para Fora.** Lista as categorias a partir das quais você removeu um documento. Os documentos podem ser forçados a sair de uma categoria através da seleção de menu **Edit** > **Force Out** . Por exemplo, isso pode ser usado quando o sarcasmo de um respondente faz com que uma resposta seja mal categorizada. Consulte “Forçando documentos em categorias” na página 136 para obter mais informações.
- **Conta de categoria.** Lista o número de categorias às quais o registro pertence.
- **Classificações De Relevância.** Fornece um ranqueamento para cada registro em uma única categoria. Esse ranqueamento mostra o grau de ajuste do registro à categoria em comparação com os outros registros nessa categoria. Selecione uma categoria na área de janela Categorias (área de janela superior esquerda) para ver o ranqueamento. Consulte o “Relevância da categoria” na página 101 para obter mais informações.
- **Sinalizadores De Resposta.** Adicio uma coluna que mostra quaisquer sinalizadores que você pode estar usando. Clique dentro desta coluna para alterar o tipo de sinalização que você atribui aos documentos. Você pode sinalizar documentos com uma bandeira "completa" ou uma sinalização "importante", ou remover bandeiras. Isso é útil para revisar a integralidade de um modelo de categoria. Consulte o “Respostas sinalizadoras” na página 101 para obter mais informações.

Respostas sinalizadoras

Para ajudá-lo a monitorar o seu progresso, você pode marcar documentos usando sinalizadores no painel de Dados. Este recurso só está disponível se o documento de origem contiver um ID exclusivo. Se o documento de origem não contiver um ID exclusivo, você poderá adicionar um nó de Derivação entre o documento de origem e o nó Mining de Texto.

Há muitas razões pelas quais você pode querer marcar um documento, incluindo:

- Para marcar os documentos que você revisou manualmente para que você saiba onde buscar mais tarde
- Para marcar de fora um documento que você está incerto sobre como tratar

Uma vez que você marca um documento com uma bandeira, você pode continuar a trabalhar com os documentos. Eles são puramente para a sua própria gravadora. Você pode escolher entre as seguintes bandeiras:

Tabela 37. Descrições de sinalização

Sinalização	Descrição
	Sinalização completa para denotar documentos que você julgar terminado.
	Bandeira importante para denotar documentos que julgar importantes.

Para marcar um documento com uma bandeira:

1. De dentro do painel de Dados, clique com o botão direito do mouse no documento que você deseja marcar.
2. No menu de contexto, escolha **Visualizar** > **Painel de Dados** > **Sinalizadores de Resposta** e, em seguida, selecione o tipo de sinalização que você deseja usar (Importante Flag ou Complete Flag). A

sinalização selecionada é atribuída. Se a coluna Bandeira no painel de Dados não estiver visível, ela aparece.

Para limpar bandeiras:

1. De dentro do painel de Dados, clique com o botão direito do mouse sobre os documentos para os quais deseja remover uma bandeira.
2. No menu de contexto, escolha **Respostas Mark With** > **Clear Flags**. As bandeiras selecionadas são removidas.

Regras de redesignação de tipo

As Regras de Redesignação de Tipo (TRRs) visam transformar uma sequência de tipos, macros e / ou tokens em um novo conceito com um tipo específico. Especificamente, eles são usados em modelos de Opinião para pegar opiniões com uma mudança na polaridade. Por exemplo, na frase "`not that bad`," a palavra "bad" é uma opinião *negativa* . Mas, neste contexto, o significado real é "`not bad`"-que é um *positivo*.

Até a versão 18.2, essa mudança na polaridade era gerenciada pelas regras específicas do Text Link Analysis (TLA):

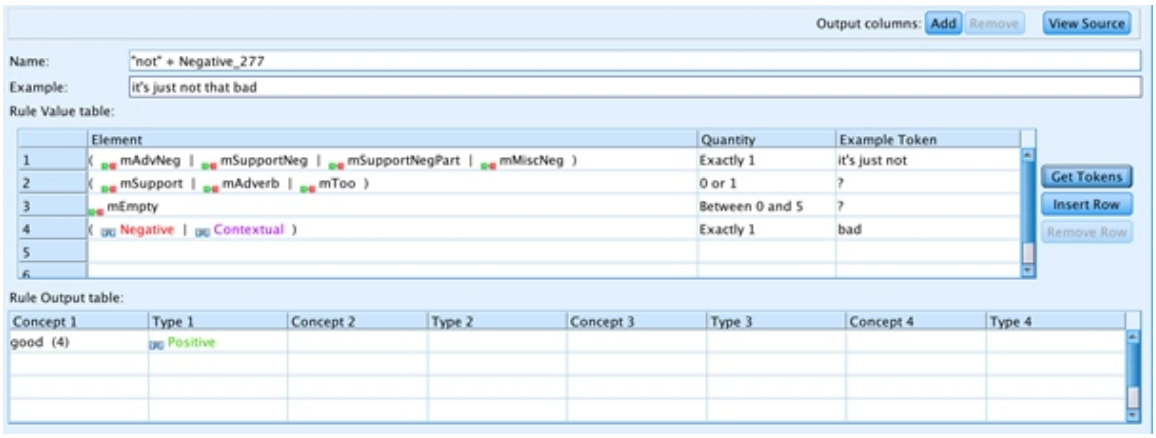


Figura 35. Regras do TLA

Because of the different opinion types (`Positive`, `PositiveAttitude`, `PositiveBudget`, `PositiveCompetence`, `PositiveFeeling`, `PositiveFunctioning`, `PositiveRecommendation`, `Negative`, `NegativeAttitude`, `NegativeBudget`, `NegativeCompetence`, `NegativeFeeling`, `NegativeFunctioning`, `NegativeRecommendation`, and `Contextual`), it involved writing specific TLA rules:

- Para cada tipo. Por exemplo:

```
"not + xxx + <NegativeBudget>" => "<PositiveBudget>"
```

ou

```
"not + xxx + <PositiveAttitude>" => "<NegativeAttitude>"
```

- Para muitos contextos sintáticos. Por exemplo:

```
* topic + negation + opinion (“hotel wasn’t good”)
* negation + opinion + topic (“it was not a good hotel”)
* negation + opinion (“not very good”)
* topic + opinion + negation + opinion (“hotel was well-located but  not that good”)
* 2 topics + negation + opinion (“room and swimming pool weren’t always clean”)
* …
```

Iniciando com a versão 18.2, a nova abordagem é "capturar" tais sequências (qualquer negação + qualquer palavra vazia + uma opinião específica), selecionar as palavras para aparecer no novo conceito

(uma negação padronizada-por exemplo, "não"-e a opinião) e definir um tipo para esse novo conceito (aka, "pseudotermo"). Este novo conceito pode, então, ser usado em regras da TLA.

Como consequência, a regra a seguir corresponderá a qualquer sequência contendo um tópico seguido de um parecer, se o parecer é um termo (`comfortable`) ou um psuedo-termo (`not economical`), independentemente de subtipo de opinião específico (Atitude, Budget, etc).

```
#@# Bed was extremely comfortable
[pattern(190)]
name=topic + opinion_190
value=$mTopic ($mEmpty|$mToo){0,3} ($mOpinionPos|$mOpinionNeg|$Contextual)
output(1)=$1\t#1\t$3\t#3
```

Junto com a mudança de polaridade para opiniões, você também pode usar o TRRs para ajudar a afinar o seu dicionário. Por exemplo, vamos supor que você tenha um tipo chamado `Anatomy` com partes do corpo tais como `heart`, `chest`, `breaste adrenal gland` -e outro tipo chamado `MedicalProcedures` com procedimentos como `biopsy`, `needle biopsy`, `MRIe CT scan`. Seria quase impossível listar todos os procedimentos médicos corretamente associados a um órgão. Assim, você poderia criar dois TRRs para identificar possíveis procedimentos médicos como visto nos números a seguir. Em seguida, uma vez que a extração é realizada, é possível adicionar um filtro no tipo `PotentialMedicalProcedures`, revisar os termos do candidato e, em seguida, incluí-los no tipo `MedicalProcedures` .

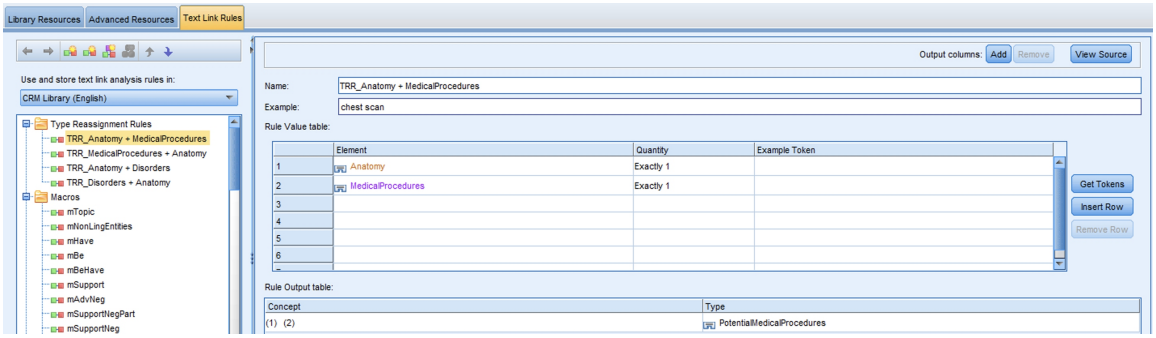


Figura 36. TRR para anatomia + procedimentos médicos

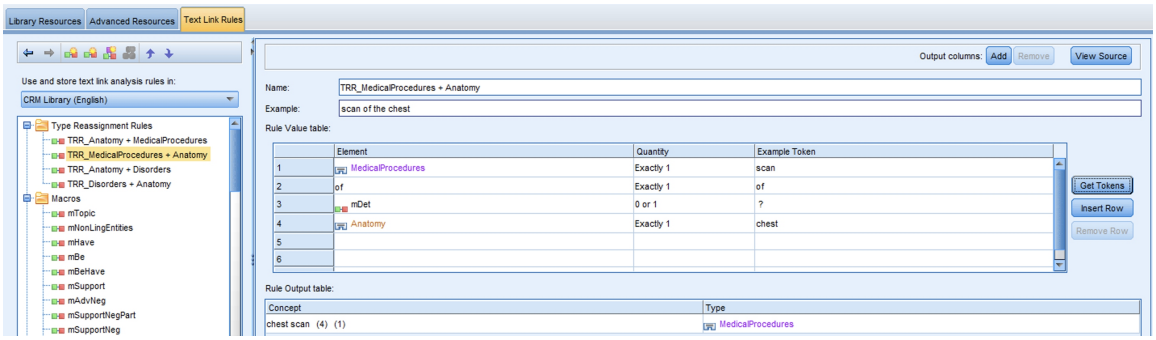


Figura 37. TRR para procedimentos médicos + anatomia

Sintaxe

```
#@# not that expensive
[typeReassignmentRule]
name=TRR_"not" NegativeBudget
value=$mAllNeg ($mAdverb|$mBe|$mHave|$mSupport|$mDet|that|more|$mQuant){0,3} $NegativeBudget
output=not $3\tPositiveBudget
```

- "`name`" deve ser único (`TRR_"not" NegativeBudget`). Ele não pode ser usado em uma regra de macro ou em uma TLA. Apenas o tipo definido na saída pode ser utilizado.
- "`value`" é uma sequência de elementos para combinar. Os elementos podem ser tipos (`$NegativeBudget`), macros (`$mAllNeg`) ou tokens (`more`). Alguns elementos podem ser necessários, opcionais, ou ter uma quantidade específica.

- "output" é um par ÚNICO de conceito + tipo (not $3\tPositiveBudget). Note que na saída você pode usar um tipo disponível (um que é definido no template) ou você pode criar um novo tipo.

  O tipo de saída também pode fazer referência a um elemento correspondido (por exemplo, #2). Esse recurso é especialmente útil quando não há alteração no tipo entre o valor e a saída. Por exemplo:

```
#@# could not have been any more pleased
[typeReassignmentRule]
name=TRR_"couldn't be more" opinion
value=$mNotNeg ($mOpinionPos|$mOpinionNeg|$Contextual)
output=$2\t#2
```

Como nas regras da TLA, um TRR mais específico deve ser definido antes de um mais genérico. Para garantir que todos os TRRs sejam definidos na ordem correta, você pode utilizar o recurso Obter Tokens para testar cada TRR em sequência. Se um TRR não combinar, mas corresponde a outra definição, você pode motá-lo para baixo (ou para cima).

Casos especiais

Em alguns casos, é necessário ainda ter acesso aos elementos individuais da sequência e não a um TRR. Isso tipicamente diz respeito à *coordenação* sobre *negação*. Na frase "not that fashionable or eyecatching," a coordenação "or" não permite descobrir que, neste contexto, "eyecatching" realmente significa "not eyecatching."

Neste caso, recomendamos o uso de uma regra específica como:

```
#@# not that fashionable or eyecatching
[pattern(263)]
name="not" + 2 Positive_263
value=($mAdvNeg|$mSupportNeg|$mMiscNeg) @{0,1} $PositiveFeeling or $PositiveFeeling
output(1)=not $3\tNegativeFeeling
output(2)=not $5\tNegativeFeeling
```

Apesar de a primeira parte da regra (($mAdvNeg|$mSupportNeg|$mMiscNeg) @{0,1} $PositiveFeeling) poder corresponder a um TRR, a regra da TLA terá prioridade.

Se você gravar uma regra mais genérica, como o exemplo a seguir, as mesmas restrições existentes na versão 18.1.1 e anteriores ainda serão aplicadas. O novo conceito criado (pseudo-conceito) pode ter um tipo incorreto (<Negative> ao invés de <NegativeFeeling>), e você pode acabar com um conceito TLA com dois tipos diferentes. Um workaround é para criar o termo correspondente (não xxx) com o tipo correto.

```
#@# not that fashionable or eyecatching
[pattern(263)]
name="not" + 2 Positive_263
value=($mAdvNeg|$mSupportNeg|$mMiscNeg) @{0,1} $mPos or $mPos
output(1)=not $3\tNegative
output(2)=not $5\tNegative
```

Vantagens

- A principal vantagem de usar o TRRs é ter menos regras de TLA.
- Uma vantagem menos óbvia é que as TRRs garantem que um pseudo-termo resultará principalmente no tipo correto (mas tenha em mente a restrição mencionada anteriormente). No passado, alguns "not + positiveXXX" foram digitados como Negative em vez de NegativeXXX, porque faltava algumas regras específicas da TLA.
- Se um usuário adicionar um tipo de opinião específico (por exemplo, NegativeNoise), não há necessidade de replicar regras de TLA específicas para inverter a polaridade. O usuário só precisa criar o TRR relevante.

Capítulo 12. Visualizando gráficos

A visualização Categorias e Conceitos, a visualização Clusters e a visualização Análise de Ligação de Texto têm uma área de janela de visualização no canto superior direito da janela. É possível usar essa área de janela para explorar seus dados visualmente. Os gráficos e diagramas a seguir estão disponíveis.

- **Visualização Categorias e Conceitos.** Esta visualização tem três gráficos e diagramas: *Barra de Categoria*, *Web de Categoria* e *Tabela da Web de Categoria*. Nessa visualização, os gráficos são atualizados somente quando você clica em **Exibir**. Consulte o tópico "Gráficos e diagramas de categoria" na página 155 para obter mais informações.
- **Visualização Clusters.** Esta visualização tem dois gráficos da web: *Gráfico Web de Conceito* e *Gráfico Web de Cluster*. Consulte o tópico "Gráficos de cluster" na página 157 para obter mais informações.
- **Visualização Análise de Ligação de Texto.** Esta visualização tem dois gráficos da web: *Gráfico Web de Conceito* e *Gráfico Web de Tipo*. Consulte o tópico "Gráficos Análise de Ligação de Texto" na página 158 para obter mais informações.

Para obter mais informações sobre todas as barras de ferramentas e paletas gerais usadas para editar gráficos, consulte a seção em Editando Gráficos na ajuda on-line ou no arquivo *ModelerSPOnodes.pdf*, que está disponível como parte do download do produto.

Gráficos e diagramas de categoria

Ao construir suas categorias, é importante reservar o tempo para revisar as definições de categoria, as de documentos ou registros que elas contêm e como as categorias são sobrepostas. A área de janela de visualização oferece diversas perspectivas em suas categorias. A área de janela Visualização está localizada no canto superior direito da visualização Categorias e Conceitos . Se ele ainda não estiver visível, você pode acessar esta pane a partir do menu Exibir (**Visualizar> Panes> Visualização**).

Nessa visualização, a área de janela de visualização oferece três perspectivas sobre os pontos em comum na categorização da do documento ou registro. Os diagramas e gráficos nesta área de janela podem ser usados para analisar seus resultados de categorização e auxiliar nas categorias ou nos relatórios de ajuste fino. Ao refinar categorias, é possível usar esta área de janela para revisar suas definições de categoria para descobrir categorias que são muito semelhantes (por exemplo, elas compartilham mais de 75% de suas de documentos ou registros) ou muito distintas. Se duas categorias são muito semelhantes, isso pode ajudá-lo a decidir combinar as duas categorias. Como alternativa, você pode decidir refinar as definições de categoria ao remover determinados descritores de uma categoria ou outra.

Dependendo do que está selecionado na área de janela Resultados de Extração, na área de janela Categorias ou na caixa de diálogo Definições de Categoria, você pode visualizar as interações correspondentes entre e categorias de documentos/registros em cada uma das guias nessa área de janela. Cada uma apresenta informações semelhantes, mas de uma maneira diferente ou com um nível diferente de detalhe. Entretanto, para atualizar um gráfico para a seleção atual, clique em **Exibir** na barra de ferramentas da área de janela ou caixa de diálogo na qual você fez sua seleção.

A área de janela Visualização na visualização Categorias e Conceitos oferece os gráficos e diagramas a seguir:

- **Gráfico de barras Categoria.** Uma tabela e um gráfico de barras apresentam a sobreposição entre as de documentos/registros correspondentes à sua seleção e as categorias associadas. O gráfico de barras também apresenta razões das de documentos/registros nas categorias para o número total de documentos/registros. Consulte o tópico "Gráfico de barras de Categoria" na página 156 para obter mais informações.
- **Gráfico de Categoria da Web.** Este gráfico apresenta a sobreposição de do documento/registro para as categorias às quais as de documentos/registros pertencem, de acordo com a seleção nas outras áreas de janela. Consulte o tópico "Gráfico de categoria da web" na página 156 para obter mais informações.

- **Tabela de Categoria da Web.** Esta tabela apresenta as mesmas informações que a guia Categoria da Web, mas em um formato de tabela. A tabela contém três colunas que podem ser classificas ao clicar nos cabeçalhos das colunas. Consulte o tópico "Tabela de categoria da web" na página 157 para obter mais informações.

Consulte o tópico Capítulo 9, "Categorizando dados de texto", na página 91 para obter mais informações.

Gráfico de barras de Categoria

Esta guia exibe uma tabela e um gráfico de barras mostrando a sobreposição entre os documentos/registros correspondentes à sua seleção e as categorias associadas. O gráfico de barras também apresenta razões dos documentos/registros nas categorias para o número total de documentos ou registros . Não é possível editar o layout deste gráfico. Entretanto, é possível classificar as colunas ao clicar nos títulos das colunas.

O gráfico contém as colunas a seguir:

- **Categoria.** Esta coluna apresenta o nome das categorias em sua seleção. Por padrão, a categoria mais comum em sua seleção é listada primeiro.
- **Barra.** Esta coluna apresenta, de uma maneira visual, a razão dos documentos ou registos em uma determinada categoria para o número total de documentos ou registros.
- **% de seleção.** Esta coluna apresenta uma porcentagem baseada na razão do número total de documentos ou registros para uma categoria para o número total de documentos ou registros representados na seleção.
- **Documentos.** Esta coluna apresenta o número de documentos ou registros em uma seleção para a determinada categoria.

Gráfico de categoria da web

Esta guia exibe um gráfico de categoria da web. A web apresenta a sobreposição dos documentos ou registros para as categorias às quais os documentos ou registros pertencem, de acordo com a seleção nas outras áreas de janela. Se os rótulos de categoria existirem, esses rótulos aparecem no gráfico. É possível escolher um layout de gráfico (rede, círculo, direcionado ou grade) usando os botões da barra de ferramentas nesta área de janela.

Na web, cada nó representa uma categoria. Usando o mouse, é possível selecionar e mover os nós dentro da área de janela. O tamanho do nó representas o tamanho relativo com base no número de documentos ou registros para tal categoria em sua seleção. A espessura e a cor da linha entre duas categorias denota o número de documentos ou registros comuns que elas possuem. Se você passar o mouse sobre um nó no modo Explorar, uma Dica de Ferramenta exibe o nome (ou rótulo) da categoria e o número geral de documentos ou registros na categoria.

Nota: Por padrão, o modo Explore em ativado para os gráficos nos quais você pode mover nós. Entretanto, é possível alternar para o modo Editar para editar seus layouts de gráfico, incluindo cores, fontes, legendas e mais. Para obter mais informações, consulte "Usando barras de ferramentas e paletas de gráfico" na página 159.

Se você copiar os dados do gráfico, usando o botão **Copiar Dados de Visualização** , e cole-o em uma planilha ou editor de texto, você verá que os dados recebem cabeçalhos de coluna como V1, V2, através do V7. Essas colunas contêm as seguintes informações:

- **V1, V2** Estes valores correspondem às coordenadas da tela (X e Y, respectivamente).
- **V3, V5** Lista o conceito de categoria.
- **Tamanho, V6** Mostra o número de documentos que os conceitos foram encontrados em.
- **V7** Atualmente não utilizados.

Tabela de categoria da web

Esta guia exibe as mesmas informações que a guia Categoria da Web, mas em um formato de tabela. A tabela contém três colunas que podem ser classificadas ao clicar nos cabeçalhos da coluna:

- **Contagem.** Esta coluna apresenta o número de documentos ou registros compartilhados ou comuns entre as duas categorias.
- **Categoria 1.** Esta coluna apresenta o nome da primeira categoria seguida do número total de documentos ou registros que ele contém, mostrado entre parênteses.
- **Categoria 2.** Esta coluna apresenta o nome da segunda categoria seguida do número total de documentos ou registros que ele contém, mostrado entre parênteses.

Gráficos de cluster

Após a construção de seus clusters, é possível explorá-los visualmente nos gráficos da web na área de janela Visualização. A área de janela de visualização oferece duas perspectivas sobre armazenamento em cluster: um gráfico Web de Conceito e um gráfico Web de Cluster. Os gráficos da web nessa área de janela podem ser usados para analisar seus resultados de armazenamento em cluster e ajudar a descobrir alguns conceitos e regras que você deseja incluir em suas categorias. A área de janela Visualização está localizada no canto superior direito da visualização Clusters. Se ele ainda não estiver visível, você pode acessar esta pane no menu Exibir (**View> Panes> Visualização**). Selecionando um cluster na área de janela Clusters, é possível exibir automaticamente os gráficos correspondentes na área de janela Visualização.

Nota: Por padrão, os gráficos estão no modo interativo/seleção no qual é possível mover nós. No entanto, é possível editar os layouts de seus gráficos no modo Editar, incluindo cores e fontes, legendas e mais. Veja o tópico "Usando barras de ferramentas e paletas de gráfico" na página 159 para obter mais informações.

A visualização Clusters tem dois gráficos da web.

- **Gráfico Web de Conceito.** Este gráfico apresenta todos os conceitos dentro do(s) cluster(s) selecionado(s), bem como conceitos vinculados fora do cluster. Este gráfico pode ajudá-lo a ver como os conceitos dentro de um cluster estão vinculados e quaisquer outras ligações externas. Veja o tópico "Gráfico Web de Conceito" na página 157 para obter mais informações.
- **Gráfico Web de Cluster.** Este gráfico apresenta o(s) cluster(s) selecionado(s) com todas as ligações externas entre os clusters selecionados, conforme mostrado nas linhas pontilhadas. Veja o tópico "Gráfico Web de Cluster" na página 158 para obter mais informações.

Consulte o tópico Capítulo 10, "Analisando clusters", na página 139 para obter informações adicionais.

Gráfico Web de Conceito

Esta guia exibe um gráfico da web mostrando todos os conceitos dentro do(s) cluster(s) selecionado(s), bem como conceitos vinculados fora do cluster. Este gráfico pode ajudá-lo a ver como os conceitos dentro de um cluster estão vinculados e quaisquer outras ligações externas. Cada conceito em um cluster é representado como um nó, que é codificado por cor de acordo com a cor do tipo. Consulte o tópico "Criando tipos" na página 185 para obter informações adicionais.

As ligações internas entre os conceitos dentro de um cluster são desenhadas e a espessura da linha de cada ligação está diretamente relacionada à contagem de docs para cada coocorrência do par de conceitos ou valor da ligação de similaridade, dependendo de sua escolha na barra de ferramentas do gráfico. As ligações externas entre os conceitos de um cluster e os conceitos fora do cluster também são mostradas.

Se conceitos forem selecionados na caixa de diálogo Definições de Cluster, o gráfico Web de Conceito exibirá esses conceitos e quaisquer ligações internas e externas associadas a eles. Nenhuma ligação entre outros conceitos que não inclua um dos conceitos selecionados aparecerá no gráfico.

Nota: Por padrão, os gráficos estão no modo interativo / seleção no qual você pode mover nós. No entanto, é possível editar seus layouts de gráfico no modo de edição, incluindo cores e fontes, legendas

e mais. Para obter mais informações, consulte “Usando barras de ferramentas e paletas de gráfico” na página 159.

Se você copiar os dados do gráfico, usando o botão **Copiar Dados de Visualização** , e cole-o em uma planilha ou editor de texto, você verá que os dados recebem cabeçalhos de coluna como V1, V2, através do V7. Essas colunas contêm as seguintes informações:

- **V1, V2** Estes valores correspondem às coordenadas da tela (X e Y, respectivamente).
- **V3, V6** Lista o tipo de conceito.
- **V4, V5** Mostra o rótulo de conceito.
- **V7** Atualmente não utilizados.

Gráfico Web de Cluster

Esta guia exibe um gráfico da web mostrando o(s) cluster(s) selecionado(s). As ligações externas entre os clusters selecionados, bem como quaisquer ligações entre outros clusters, são todas mostradas como linhas pontilhadas. Em um gráfico Web de Cluster, cada nó representa um cluster inteiro e a espessura das linhas desenhadas entre eles representa o número de ligações externas entre dois clusters.

Importante! Para exibir um gráfico Web de Cluster, deve-se ter construído clusters com ligações externas. Ligações externas são aquelas entre pares de conceitos em clusters separados (um conceito dentro de um cluster e um conceito fora de outro cluster).

Por exemplo, digamos que temos dois clusters. `Cluster A` tem três conceitos: `A1`, `A2` e `A3`. `Cluster B` tem dois conceitos: `B1` e `B2`. Os conceitos a seguir são vinculados: `A1-A2`, `A1-A3`, `A2-B1` (Externo), `A2-B2` (Externo), `A1-B2` (Externo) e `B1-B2`. Isso significa que, no gráfico Web de Cluster, a espessura da linha representaria as três ligações externas.

Nota: Por padrão, os gráficos estão no modo interativo/seleção no qual é possível mover nós. No entanto, é possível editar seus layouts de gráfico no modo de edição, incluindo cores e fontes, legendas e mais. Veja o tópico “Usando barras de ferramentas e paletas de gráfico” na página 159 para obter mais informações.

Gráficos Análise de Ligação de Texto

Após a extração de seus padrões de Análise de Ligação de Texto (TLA), é possível explorá-los visualmente nos gráficos da web na área de janela Visualização. A área de janela Visualização oferece duas perspectivas nos padrões de TLA: um gráfico web de conceito (padrão) e um gráfico web de tipo (padrão). Os gráficos da web nessa área de janela pode ser usados para representar padrões visualmente. A área de janela Visualização está localizada no canto superior direito da Análise de Ligação de Texto. Se ele ainda não estiver visível, você pode acessar esta pane no menu Exibir (**View> Panes> Visualização**). Se não houver uma seleção, a área do gráfico estará vazia.

Nota: Por padrão, os gráficos estão no modo interativo/seleção no qual é possível mover nós. No entanto, é possível editar seus layouts de gráfico no modo de edição, incluindo cores e fontes, legendas e mais. Veja o tópico “Usando barras de ferramentas e paletas de gráfico” na página 159 para obter mais informações.

A visualização Análise de Ligação de Texto tem dois gráficos da web.

- **Gráfico Web de Conceito.** Este gráfico apresenta todos os conceitos no(s) padrão(ões) selecionado(s). A largura da linha e os tamanhos de nó (se os ícones de tipo não forem mostrados) em um gráfico de conceito mostram o número de ocorrências globais na tabela selecionada. Veja o tópico “Gráfico Web de Conceito” na página 159 para obter mais informações.
- **Gráfico Web de Tipo.** Este gráfico apresenta todos os tipos no(s) padrão(ões) selecionado(s). A largura da linha e os tamanhos de nó (se os ícones de tipo não forem mostrados) no gráfico mostram o número de ocorrências globais na tabela selecionada. Nós são representados por uma cor de tipo ou por um ícone. Veja o tópico “Gráfico Web de Tipo” na página 159 para obter mais informações.

Consulte o tópico Capítulo 11, “Explorando a análise de ligação de texto”, na página 145 para obter informações adicionais.

Gráfico Web de Conceito

Esse gráfico da web apresenta todos os conceitos representados na seleção atual. Por exemplo, se você selecionasse um padrão de tipo que tivesse três padrões de conceito correspondentes, esse gráfico mostraria três conjuntos de conceitos vinculados. A largura da linha e os tamanhos de nó em um gráfico de conceito representam as contagens de frequência globais. O gráfico representa visualmente as mesmas informações que as selecionadas nas áreas de janela de padrão. Os tipos de cada conceito são apresentados por uma cor ou por um ícone, dependendo do que você seleciona na barra de ferramentas do gráfico. Consulte o tópico "Usando barras de ferramentas e paletas de gráfico" na página 159 para obter informações adicionais.

Gráfico Web de Tipo

Este gráfico da web apresenta cada padrão de tipo para a seleção atual. Por exemplo, se você selecionasse dois padrões de conceito, esse gráfico mostraria um nó por tipo nos padrões selecionados e as ligações entre aqueles localizados no mesmo padrão. A largura da linha e os tamanhos de nó representam as contagens de frequência globais para o conjunto. O gráfico representa visualmente as mesmas informações que as selecionadas nas áreas de janela de padrão. Além dos nomes de tipo que aparecem no gráfico, os tipos também são identificados por sua cor ou por um ícone de tipo, dependendo do que você selecionar na barra de ferramentas do gráfico. Consulte o tópico "Usando barras de ferramentas e paletas de gráfico" na página 159 para obter informações adicionais.

Usando barras de ferramentas e paletas de gráfico

Para cada gráfico, há uma barra de ferramentas que fornece acesso rápido a algumas paletas comuns a partir das quais você pode executar uma série de ações com seus gráficos. Cada visualização (Categorias e Conceitos, Clusters e Análise de Link de Texto) possui uma barra de ferramentas ligeiramente diferente. É possível escolher entre o modo de visualização *Explorar* ou o modo de visualização *Editar*.

Ao passo que o modo Explorar permite explorar analiticamente os dados e os valores representados pela visualização, o modo de Edição permite alterar o layout e a aparência da visualização. Por exemplo, você pode mudar as fontes e cores para corresponder ao guia de estilo da sua organização. Para selecionar este modo, escolha **Visualizar** > **Pane de Visualização** > **Modo de Edição** a partir dos menus (ou clique no ícone da barra de ferramentas).

No modo de Edição, existem várias barras de ferramentas que afetam os diferentes aspectos do layout da visualização. Se houver algumas que você não utiliza, será possível ocultá-las para aumentar a quantidade de espaço na caixa de diálogo na qual o gráfico é exibido. Para selecionar ou cancelar a seleção das barras de ferramentas, clique na barra de ferramentas relevante ou no nome da paleta no menu Visualizar.

Para obter mais informações sobre todas as barras de ferramentas e paletas gerais usadas para editar gráficos, consulte a seção Editando Visualizações na ajuda on-line ou no arquivo *ModelerSPOnodes.pdf* , que está disponível como parte do download do produto.

Tabela 38. Botões da barra de ferramentas de Analítica de Texto

Botão/Lista	Descrição
	Ativa o modo de Edição. Alterne para o modo de Edição para mudar a aparência do gráfico, tal como alargamento da fonte, mudanças da cores para corresponder ao guia de estilo da sua empresa ou remoção de rótulos e legendas.
	Ativa o modo de Exploração. Por padrão, o modo Explorar está ativado, o que significa que você pode mover e arrastar nós ao redor do gráfico, bem como passar o mouse sobre objetos do gráfico para exibir informações de Dicas da Ferramenta adicionais.

Tabela 38. Botões da barra de ferramentas de Analítica de Texto (continuação)	
Botão/Lista	**Descrição**
	Selecione um tipo de exibição da web para os gráficos na visualização Categorias e Conceitos, bem como na visualização Análise de Link de Texto. • **Layout Círculo** Um layout geral que pode ser aplicado em qualquer gráfico. Ele define o layout de um gráfico assumindo que os links são não direcionados e trata todos os nós da mesma maneira. Os nós são colocados apenas ao redor do perímetro de um círculo. • **Layout de Rede** Um layout geral que pode ser aplicado em qualquer gráfico. Ele define o layout de um gráfico assumindo que os links são não direcionados e trata todos os nós da mesma maneira. Os nós são colocados livremente no layout. • **Layout Direcionado** Um layout que só deve ser usado para gráficos direcionados. Esse layout produz estruturas semelhantes a uma árvore, desde nós raiz até nós folha e organiza por cores. Os dados hierárquicos tendem a ser exibidos perfeitamente com esse layout. • **Layout da grade** Um layout geral que pode ser aplicado em qualquer gráfico. Ele define o layout de um gráfico assumindo que os links são não direcionados e trata todos os nós da mesma maneira. Os nós são colocados apenas nos pontos de grade no espaço.
	Representação de tamanho de link. Escolha qual a espessura da linha é representada no gráfico. Isso se aplica apenas à visualização Clusters. O gráfico da web Clusters mostra apenas o número de links externos entre os clusters. É possível escolher entre: • **Similaridade** Espessura indica o número de links externos entre dois clusters • **Co-ocorrência** Espessura indica o número de documentos em que ocorre uma co-ocorrência de descritores.
	Um botão de alternância que, quando pressionado, exibe a legenda. Quando o botão não é pressionado, a legenda não é mostrada.
	Um botão de alternância que, quando pressionado, exibe os ícones de tipo no gráfico em vez de cores de tipo. Isto se aplica apenas à visualização Análise de Link de Texto.
	Um botão de alternância que, quando pressionado, exibe a Régua de Controle de Links abaixo do gráfico. Você pode filtrar os resultados ao deslizar a seta.
	Exibirá o gráfico para o nível mais alto de categorias selecionadas em vez de para suas subcategorias.
	Exibirá o gráfico para o nível mais baixo de categorias selecionadas.
	Esta opção controla como os nomes de subcategorias são exibidos na saída. • **Caminho da categoria Completo** Esta opção irá saída o nome da categoria e o caminho completo de categorias pai se aplicável usando barras para separar nomes de categoria de nomes de subcategoria. • **Caminho da categoria Curta** Esta opção irá saída apenas o nome da categoria mas use elipses para mostrar o número de categorias pai para a categoria em questão. • **Categoria nível inferior** Esta opção irá saída apenas o nome da categoria sem o caminho completo ou as categorias pai mostradas.

Capítulo 13. Editor de recurso de sessão

IBM SPSS Modeler Text Analytics captura e extrai de maneira rápida e precisa os conceitos chave dos dados de texto. Esse processo de extração é altamente baseado em recursos linguísticos para ditar como extrair informações dos dados de texto. Por padrão, esses recursos vêm de modelos de recurso.

IBM SPSS Modeler Text Analytics é fornecido com um conjunto de **modelos de recurso** especializados que contêm um conjunto de recursos linguísticos e não linguísticos, em forma de bibliotecas e recursos avançados, para ajudar a definir como seus dados serão manipulados e extraídos. Consulte o tópico Capítulo 14, “Modelos e recursos”, na página 165 para obter mais informações.

Na caixa de diálogo do nó, é possível carregar uma cópia dos recursos do modelo no nó. Uma vez dentro de uma sessão do ambiente de trabalho interativa, é possível customizar esses recursos especificamente para os dados desse nó, se você desejar. Durante uma sessão do ambiente de trabalho interativa, é possível trabalhar com recursos na visualização Editor de Recursos. Sempre que uma sessão interativa é ativada, uma extração é executada usando os recursos carregados na caixa de diálogo do nó, a menos que você tenha armazenado seus dados e os resultados da extração em cache em seu nó.

Editando recursos no editor de recurso

O Editor de Recursos oferece acesso ao conjunto de recursos usado para produzir resultados de extração (conceitos, tipos e padrões) para uma sessão do ambiente de trabalho interativa. Esse editor é bastante semelhante ao Editor de Template, exceto que no Editor de Recursos, você está editando os recursos para essa sessão. Ao concluir o trabalho em seus recursos e qualquer outro, será possível atualizar o nó de modelagem para salvar esse trabalho para que ele possa ser restaurado em uma sessão do ambiente de trabalho interativa subsequente. Consulte o tópico “Atualizando nós de modelagem e salvando” na página 75 para obter mais informações.

Se desejar trabalhar diretamente nos modelos usados para carregar recursos em nós, é recomendado usar o Editor de Template. Muitas das tarefas que podem ser executadas dentro do Editor de Recursos são executadas exatamente como são no Editor de Template, como:

- **Trabalhando com bibliotecas**. Veja o tópico Capítulo 15, “Trabalhando com bibliotecas”, na página 175 para obter mais informações.
- **Criando dicionários de tipo**. Veja o tópico “Criando tipos” na página 185 para obter mais informações.
- **Incluindo termos nos dicionários**. Veja o tópico “incluindo termos” na página 186 para obter mais informações.
- **Criando sinônimos**. Veja o tópico “Definindo sinônimos” na página 190 para obter mais informações.
- **Importando e exportando modelos**. Veja o tópico “Importando e exportando modelos” na página 171 para obter mais informações.
- **Publicando bibliotecas**. Veja o tópico “Publicando bibliotecas” na página 181 para obter mais informações.

Para texto em holandês, inglês, francês, alemão, italiano, português e espanhol

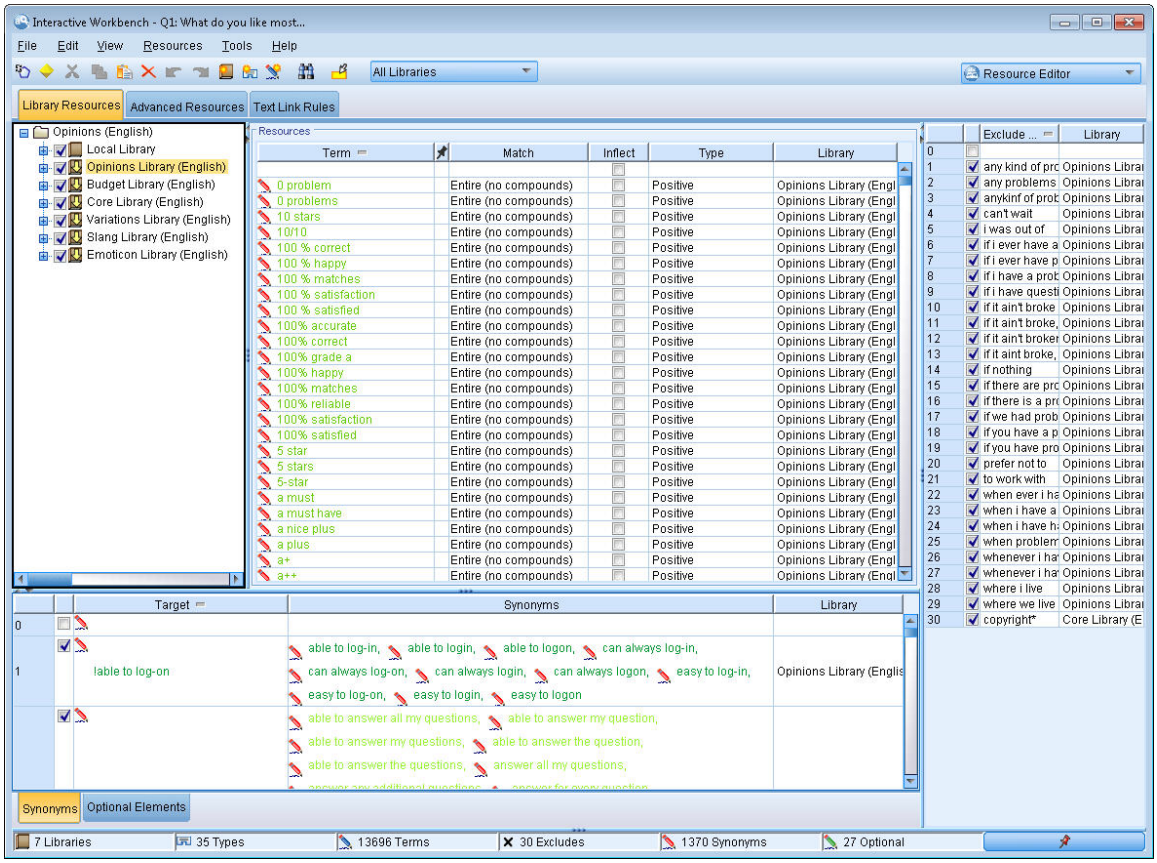


Figura 38. Visualização Editor de Recursos

Criando e atualizando modelos

Sempre que você fizer mudanças em seus recursos e desejar reutilizá-los no futuro, é possível salvar os recursos como um modelo. Ao fazer isso, você pode escolher salvar usando um nome de modelo existente ou fornecendo um novo nome. Em seguida, sempre que você carregar esse modelo no futuro, será capaz de obter os mesmos recursos. Veja o tópico "Copiando recursos dos modelos e TAPs" na página 26 para obter mais informações.

Nota: também é possível publicar e compartilhar suas bibliotecas. Veja o tópico "Compartilhando bibliotecas" na página 180 para obter mais informações.

Para Criar (ou Atualizar) um Modelo

1. A partir dos menus na visualização Editor de Recursos , escolha **Recursos> Fazer Modelo de Recursos**. A caixa de diálogo Criar Modelo de Recurso é aberta.
2. Insira um novo nome no campo Nome do Modelo, se você desejar criar um novo modelo. Selecione um modelo na tabela se desejar sobrescrever um modelo existente com os recursos carregados atualmente.
3. Clique em **Salvar** para criar o modelo.

Importante! Como os modelos são carregados quando você os seleciona no nó e não quando o fluxo é executado, certifique-se de recarregar o modelo de recurso em qualquer outro nó no qual ele é usado se você desejar obter as mudanças mais recentes. Veja o tópico "Atualizando recursos do nó após o carregamento" na página 170 para obter mais informações.

Alternando modelos de recursos

Se você deseja substituir os recursos atualmente carregados na sessão com uma cópia desses a partir de outro modelo, é possível alternar para esses recursos. Fazer isso sobrescreverá quaisquer recursos

atualmente carregados na sessão. Se você estiver alternando recursos para ter algumas regras de padrão de Análise de Link de Texto (TLA) predefinidas, certifique-se de selecionar um modelo que as tem marcadas na coluna TLA.

Trocar recursos é particularmente útil quando você deseja restaurar o trabalho de sessão (categorias, padrões e recursos) mas deseja carregar uma cópia atualizada dos recursos a partir de um modelo sem perder o seu outro trabalho de sessão.Você pode selecionar o modelo cujo conteúdo deseja copiar no Editor de Recursos e clicar em **OK.** Isso substitui os recursos que você tem nessa sessão. Certifique-se de atualizar o nó de modelagem no final de sua sessão se quiser manter essas mudanças na próxima vez que você ativar a sessão de ambiente de trabalho interativo.

Nota: Se você alternar para o conteúdo de outro template durante uma sessão interativa, o nome do template listado no nó ainda será o nome do último modelo carregado e copiado. Para se beneficiar destes recursos ou trabalho de outra sessão, atualize o nó de modelagem antes de sair da sessão e selecione a opção **Usar trabalho de sessão** no nó. Veja o tópico "Atualizando nós de modelagem e salvando" na página 75 para obter mais informações.

Para Alternar Recursos

1. A partir dos menus na visualização Editor de Recursos , escolha **Recursos> Switch Resource Templates**. A caixa de diálogo Alternar Recursos é aberta.
2. Selecione o modelo que deseja usar entre os mostrados na tabela.
3. Clique em **OK** para abandonar esses recursos atualmente carregados e carregar uma cópia desses no modelo selecionado em seu lugar. Se você tiver feito mudanças em seus recursos e desejar salvar suas bibliotecas para uso futuro, é possível publicar, atualizar e compartilhá-los antes de alternar. Veja o tópico "Compartilhando bibliotecas" na página 180 para obter mais informações.

Capítulo 14. Modelos e recursos

IBM SPSS Modeler Text Analytics captura e extrai de maneira rápida e precisa os conceitos chave dos dados de texto. Esse processo de extração é altamente baseado em recursos linguísticos para ditar como extrair informações dos dados de texto. Veja o tópico “Como funciona a extração” na página 5 para obter mais informações. É possível ajustar esses recursos na visualização Editor de Recursos.

Ao instalar o software, você também obtém um conjunto de recursos especializados. Esses recursos fornecidos permitem que você se beneficie de anos de pesquisas e ajustes para idiomas específicos e aplicativos específicos. Como nem sempre os recursos fornecidos podem ser perfeitamente adaptados ao contexto de seus dados, é possível editar esses modelos de recurso ou até criar e usar bibliotecas customizadas ajustadas exclusivamente aos dados de sua organização. Esses recursos vêm de várias formas, e cada um pode ser usado em sua sessão. Os recursos podem ser localizados no seguinte:

- **Modelos de recurso**. Modelos são compostos por um conjunto de bibliotecas, tipos e alguns recursos avançados que, juntos, formam um conjunto especializado de recursos adaptados a um determinado domínio ou contexto, como opiniões sobre um produto.
- **Pacotes de análise de texto (TAP)**. Além dos recursos armazenados em um modelo, TAPs também empacotam um ou mais conjuntos de categorias especializados gerados usando esses recursos, de modo que categorias e recursos sejam armazenados juntos e reutilizáveis. Veja o tópico “Usando pacotes de análise de texto” na página 131 para obter mais informações.
- **Bibliotecas**. Bibliotecas são usadas como blocos de construção para TAPs e modelos. Elas também podem ser incluídos individualmente nos recursos em sua sessão. Cada biblioteca é composta por vários dicionários usados para definir e gerenciar tipos, sinônimos e listas de exclusão. Embora as bibliotecas também sejam entregues individualmente, elas são pré-empacotadas em modelos e TAPs. Consulte o tópico Capítulo 15, “Trabalhando com bibliotecas”, na página 175 para obter mais informações.

Nota: Durante a extração, alguns recursos internos compilados também são usados. Esses recursos compilados contêm um grande número de definições complementando os tipos na biblioteca Principal. Esses recursos compilados não podem ser editados.

O Editor de Recursos oferece acesso ao conjunto de recursos usado para produzir resultados de extração (conceitos, tipos e padrões). Há inúmeras tarefas que podem ser executadas no Editor de Recursos; elas incluem:

- **Trabalhando com bibliotecas**. Veja o tópico Capítulo 15, “Trabalhando com bibliotecas”, na página 175 para obter mais informações.
- **Criando dicionários de tipo**. Veja o tópico “Criando tipos” na página 185 para obter mais informações.
- **Incluindo termos nos dicionários**. Veja o tópico “incluindo termos” na página 186 para obter mais informações.
- **Criando sinônimos**. Veja o tópico “Definindo sinônimos” na página 190 para obter mais informações.
- **Atualizando os recursos em TAPs**. Veja o tópico “Atualizando Pacotes de Análise de Texto” na página 133 para obter mais informações.
- **Criando modelos**. Veja o tópico “Criando e atualizando modelos” na página 162 para obter mais informações.
- **Importando e exportando modelos**. Veja o tópico “Importando e exportando modelos” na página 171 para obter mais informações.
- **Publicando bibliotecas**. Veja o tópico “Publicando bibliotecas” na página 181 para obter mais informações.

Editor do Modelo vs. Editor de Recurso

Há dois métodos principais com os quais trabalhar e editar seus modelos, bibliotecas e recursos. É possível trabalhar em recursos linguísticos no Editor de Template ou Editor de Recursos.

Editor de Template

O Editor de Template permite criar e editar modelos de recurso sem uma sessão de ambiente de trabalho interativa e independente de um nó ou fluxo específico. É possível usar esse editor para criar ou editar modelos de recurso antes de carregá-los no nó Análise de Ligação de Texto e no nó de modelagem Mineração de Texto.

O Editor de Template é acessível através da barra de ferramentas principal IBM SPSS Modeler a partir do menu **Ferramentas> Editor de Template Analytics Editor** .

Editor de Recursos

O Editor de Recursos, que está acessível dentro de uma sessão de ambiente de trabalho interativa, permite trabalhar com os recursos no contexto de um nó ou conjunto de dados específico. Quando você inclui um nó de modelagem Mineração de Texto em um fluxo, é possível carregar uma cópia do conteúdo de um modelo de recurso ou uma cópia de um pacote de análise de texto (conjuntos de categorias *e* recursos) para controlar como o texto é extraído para mineração de texto. Quando você ativa uma sessão de ambiente de trabalho interativa, além de criar categorias, extrair padrões de análise de ligação de texto e criar modelos de categoria, também é possível ajustar os recursos para os dados da sessão na visualização do Editor de Recursos integrada. Consulte o tópico “Editando recursos no editor de recurso” na página 161 para obter mais informações.

Sempre que você trabalha nos recursos em uma sessão de ambiente de trabalho interativa, essas mudanças se aplicam somente a essa sessão. Se você desejar salvar seu trabalho (recursos, categorias, padrões, etc.) para que seja possível continuar em uma sessão subsequente, deve-se atualizar o nó de modelagem. Consulte o tópico “Atualizando nós de modelagem e salvando” na página 75 para obter informações adicionais.

Se você desejar salvar suas mudanças de volta no modelo original, cujo conteúdo foi copiado no nó de modelagem, de modo que esse modelo atualizado possa ser carregado em outros nós, é possível criar um modelo a partir dos recursos. Consulte o tópico “Criando e atualizando modelos” na página 162 para obter informações adicionais.

Nota: Se você fizer alterações em modelos ou bibliotecas e salvá-las em um diretório de backup e, em seguida, atualizar sua versão de IBM SPSS Modeler Text Analytics, você receberá a opção de importar seus modelos e bibliotecas personalizados. A primeira vez que você executar um fluxo SPSS Análise de Texto do Modeler ou abrir o Editor de Recursos após um upgrade, modelos padrão e bibliotecas são copiados para a sua máquina. Um aviso **Modelos Saved** ou um aviso **Saved bibliotecas** (ou ambos) é exibido com uma lista dos templates e / ou bibliotecas que são atualizados como parte do upgrade do produto, e você recebe a opção de importar seus modelos personalizados e bibliotecas do diretório em que os salvou. Depois de clicar em **OK** na mensagem de aviso, você pode abrir o diálogo **Gerenciar modelos de recursos** ou o diálogo **Gerenciar bibliotecas** a qualquer momento para escolher quais modelos personalizados ou bibliotecas você deseja importar.

A interface do editor

As operações que você executa no Editor de Modelo ou Editor de Recursos são revolvidas em torno do gerenciamento e do ajuste de recursos linguísticos. Esses recursos são armazenados na forma de modelos e bibliotecas. Consulte o tópico “Dicionários de tipo” na página 183 para obter mais informações.

Guia Recursos de Biblioteca

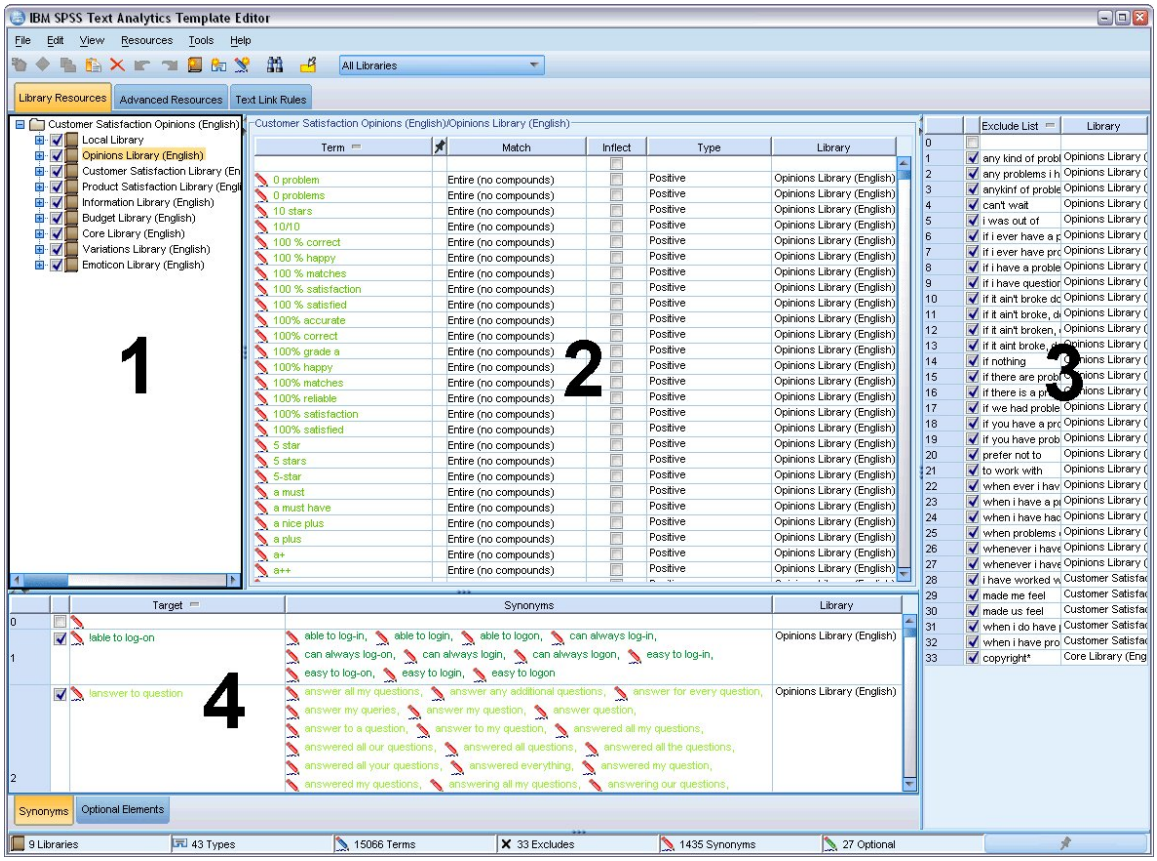


Figura 39. Editor de modelo de mineração de texto

A interface é organizada em quatro partes, conforme a seguir:

1. Área de janela Árvore de Bibliotecas. Localizado no canto superior esquerdo, esta área de janela exibe uma árvore das bibliotecas. É possível ativar e desativar as bibliotecas nessa árvore, assim como filtrar as visualizações nas outras áreas de janela ao selecionar uma biblioteca na árvore. É possível executar muitas operações nessa árvore usando os menus de contexto. Se você expandir uma biblioteca na árvore, é possível ver o conjunto de tipos de que ela contém. Também é possível filtrar essa lista por meio do menu **Visualizar** se você desejar manter o foco em uma biblioteca específica.

2. Termo Listas da área de janela do Type Dictionaries. Localizada à direita da árvore de bibliotecas, esta área de janela exibe as listas de termos dos dicionários de tipo para as bibliotecas selecionadas na árvore. Um **dicionário de tipos** é uma coleção de termos a serem agrupados sob um nome de rótulo ou tipo. Quando o mecanismo de extração lê os dados de texto, ele compara as palavras localizadas no texto para os termos nos dicionários de tipos. Se um conceito extraído aparece como um termo em um dicionário de tipos, então, esse nome de tipo é designado. Você pode pensar no dicionário de tipos como um dicionário distinto de termos que possuem algo em comum. Por exemplo, o tipo `<Location>` na biblioteca Core contém conceitos como `new orleans`, `great britaine` `new york`. Estes termos todos representam locais geográficos. Uma biblioteca pode conter um ou mais dicionários de tipos. Veja o tópico "Dicionários de tipo" na página 183 para obter mais informações.

3. Painel Excluir Dicionário. Localizada no lado direito, esta área de janela exibe a coleção de termos que serão excluídos dos resultados finais da extração. Os termos que aparecem nesse dicionário de exclusões não aparecem na área de janela Resultados da Extração. Os termos excluídos podem ser armazenados na biblioteca de sua escolha. Entretanto, a área de janela Dicionário de Exclusões exibe todos os termos excluídos para todas as bibliotecas visíveis na árvores de bibliotecas. Veja o tópico "Dicionários de exclusão" na página 193 para obter mais informações.

4. Área de janela Dicionário de Substituições. Localizada na parte inferior esquerda, esta área de janela exibe sinônimos e elementos opcionais, cada um em sua própria guia. Sinônimos e elementos opcionais ajudam a agrupar termos semelhantes sob um conceito principal ou destino nos resultados

finais da extração. Esse dicionário pode conter sinônimos conhecidos e sinônimos definidos pelo usuário, bem como erros ortográficos comuns pareados com a ortografia correta. Definições de sinônimo e elementos opcionais podem ser armazenados na biblioteca de suas escolha. Entretanto, a área de janela do dicionário de substituições exibe todo o conteúdo para todas as bibliotecas visíveis na árvore de bibliotecas. Embora esta área de janela exiba todos os sinônimos ou elementos opcionais de todas as bibliotecas, as substituições para todas as bibliotecas na árvore são mostradas juntas nessa área de janela. Uma biblioteca pode conter apenas um dicionário de substituições. Consulte o tópico "Dicionários de substituição/sinônimo" na página 190 para obter mais informações.

Notas:

- Se você deseja filtrar para ver apenas as informações relativas a uma única biblioteca, é possível mudar a visualização de biblioteca usando a lista suspensa na barra de ferramentas. Ela contém a entrada de nível superior denominada **Todas as Bibliotecas**, bem como uma entrada adicional para cada biblioteca individual. Consulte o tópico "Visualizando bibliotecas" na página 177 para obter mais informações.

Guia Recursos Avançados

Os recursos avançados estão disponíveis a partir da segunda guia da visualização do editor. É possível revisar e editar os recursos avançados nessa guia. Consulte o tópico Capítulo 17, "Sobre recursos avançados", na página 195 para obter informações adicionais.

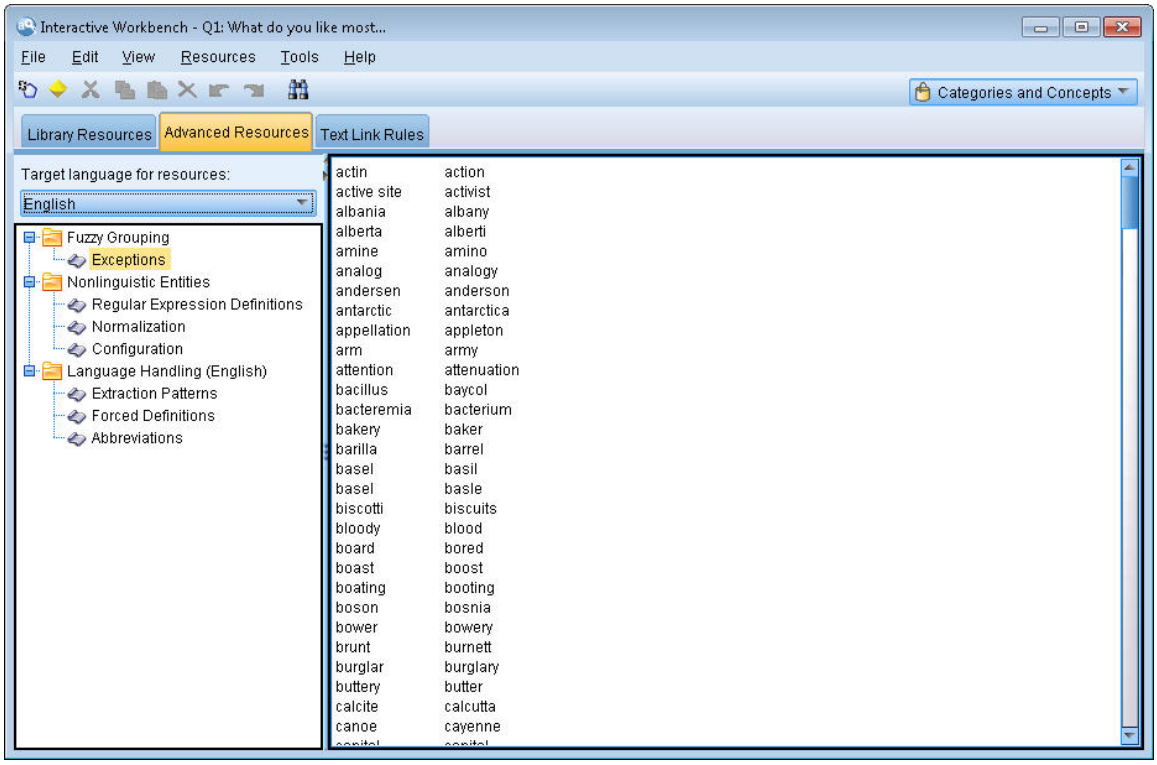


Figura 40. Editor de modelo de mineração de texto - guia recursos avançados

Guia Regras de Ligação de Texto

Desde a versão 14, as regras de análise de ligação de texto são editáveis em sua própria guia da visualização do editor. É possível trabalhar no editor de regras e até executar simulações para ver como suas regras afetam os resultados da TLA. Consulte o tópico Capítulo 18, "Sobre regras de ligação de texto", na página 207 para obter informações adicionais.

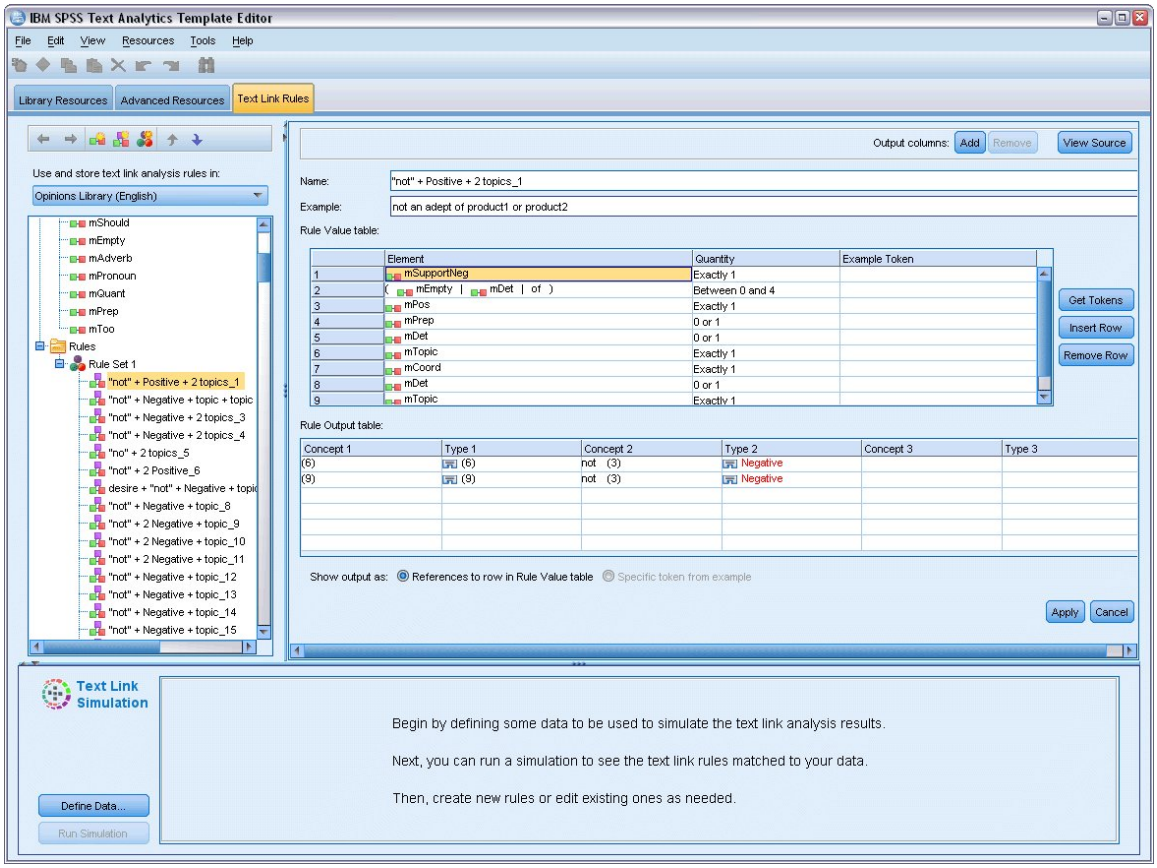


Figura 41. Editor de modelo de mineração de texto - guia Regras de Ligação de Texto

Abrindo modelos

Ao ativar o Editor de Template, é solicitado que você abra um modelo. Da mesma forma, é possível abrir um modelo a partir do menu Arquivo. Se desejar que seu modelo contenha algumas regras de Análise de Ligação de Texto (TLA), certifique-se de selecionar um modelo que tenha um ícone na coluna TLA. O idioma para o qual um modelo foi criado é mostrado na coluna Idioma.

Se desejar importar um modelo que não é mostrado na tabela ou desejar exportar um modelo, é possível usar os botões na caixa de diálogo Abrir Modelo. Consulte o tópico "Importando e exportando modelos" na página 171 para obter informações adicionais.

Para abrir um modelo

1. A partir dos menus no Editor de Template, escolha **Arquivo> Modelo de Recursos Abertas**. A caixa de diálogo Abrir Modelo de Recurso é aberta.
2. Selecione o modelo que deseja usar entre os mostrados na tabela.
3. Clique em **OK** para abrir esse modelo. Se você tiver outro modelo aberto atualmente no editor, um clique em OK abandonará esse modelo e exibirá o modelo que você selecionou aqui. Se você tiver feito mudanças em seus recursos e desejar salvar suas bibliotecas para usar no futuro, é possível publicar, atualizar e compartilhar todas elas antes de abrir outra. Veja o tópico "Compartilhando bibliotecas" na página 180 para obter mais informações.

Salvando modelos

No Editor de Template, é possível salvar as mudanças feitas no modelo. É possível escolher salvar usando um nome de modelo existente ou fornecendo um novo nome.

Se fizer mudanças em um modelo que já foi carregado em um nó anteriormente, você terá que recarregar o conteúdo do modelo no nó para obter as mudanças mais recentes. Consulte o tópico "Copiando recursos dos modelos e TAPs" na página 26 para obter informações adicionais.

Ou, se estiver usando a opção **Usar Trabalho Interativo Salvo** na guia Modelo do nó Mineração de Texto, o que significa que você está usando recursos de uma sessão de ambiente de trabalho interativa anterior, será necessário alternar para os recursos desse modelo de dentro da sessão de ambiente de trabalho interativa. Veja o tópico "Alternando modelos de recursos" na página 162 para obter mais informações.

Nota: também é possível publicar e compartilhar suas bibliotecas. Veja o tópico "Compartilhando bibliotecas" na página 180 para obter mais informações.

Para Salvar um Modelo

1. A partir dos menus no Editor de Template, escolha **Arquivo> Salvar Modelo de Recursos**. A caixa de diálogo Salvar Modelo de Recurso é aberta.
2. Insira um novo nome no campo Nome do Modelo, caso queira salvar esse modelo como um novo nome. Selecione um modelo na tabela se desejar sobrescrever um modelo existente com os recursos carregados atualmente.
3. Se desejar, insira uma descrição para exibir um comentário ou uma anotação na tabela.
4. Clique em **Salvar** para salvar o modelo.

Importante! Como os recursos dos modelos ou TAPs são carregados/copiados no nó, deve-se atualizar os recursos recarregando-os se você fizer mudanças em um modelo e desejar se beneficiar dessas mudanças em um fluxo existente. Veja o tópico "Atualizando recursos do nó após o carregamento" na página 170 para obter mais informações.

Atualizando recursos do nó após o carregamento

Por padrão, quando você inclui um nó em um fluxo, um conjunto de recursos de um modelo padrão é carregado e integrado ao seu nó. E se você muda os modelos ou usa uma TAP, ao carregá-los, uma cópia desses recursos sobrescreve os recursos. Como modelos e TAPs não são vinculados diretamente ao nó, nenhuma mudança feita em um modelo ou TAP fica automaticamente disponível em um nó pré-existente. Para se beneficiar dessas mudanças, você teria que atualizar os recursos nesse nó. Os recursos podem ser atualizados de uma de duas formas.

Método 1: Recarregando recursos na guia Modelo

Se você desejar atualizar os recursos no nó usando um modelo ou TAP novo ou atualizado, é possível recarregá-los na guia Modelo do nó. Ao recarregar, você substituirá a cópia dos recursos no nó pela cópia mais atual. Para sua comodidade, a data e a hora atualizadas aparecerão na guia Modelo junto com o nome do modelo de origem. Consulte o tópico "Copiando recursos dos modelos e TAPs" na página 26 para obter informações adicionais.

No entanto, se você estiver trabalhando com dados de uma sessão interativa em um nó de modelagem Mineração de Texto e tiver selecionado a opção **Usar Trabalho da Sessão** na guia Modelo, o trabalho da sessão salvo e os recursos serão usados e o botão **Carregar** será desativado. Ele é desativado porque, em um determinado momento na sessão de ambiente de trabalho interativa, você escolheu a opção **Atualizar Nó de Modelagem** e manteve as categorias, recursos e outros trabalhos da sessão. Nesse caso, se desejar mudar ou atualizar esses recursos, você poderá tentar o método seguinte de alternância de recursos no Editor de Recursos.

Método 2: Switching Resources no Editor de Recursos

Sempre que você quiser usar diferentes recursos durante uma sessão interativa, é possível trocar esses recursos usando a caixa de diálogo Alternar Recursos. Isso é útil principalmente quando você deseja reutilizar um trabalho de categoria existente, mas substituir os recursos. Nesse caso, é possível selecionar a opção **Usar Trabalho da Sessão** na guia Modelo de um nó de modelagem Mineração de Texto. Isso desativará a capacidade de recarregar de um modelo por meio da caixa de diálogo do nó e manterá as configurações e as mudanças que foram feitas durante sua sessão. Então você poderá ativar a sessão de ambiente de trabalho interativa executando o fluxo e alternar os recursos no Editor

de Recursos. Consulte o tópico “Alternando modelos de recursos” na página 162 para obter mais informações.

Para manter o trabalho da sessão para sessões subsequentes, incluindo os recursos, você precisa atualizar o nó de modelagem de dentro da sessão de ambiente de trabalho interativa para que os recursos (e outros dados) sejam salvos de volta no nó. Consulte o tópico “Atualizando nós de modelagem e salvando” na página 75 para obter informações adicionais.

Nota: Se você alternar para o conteúdo de outro modelo durante uma sessão interativa, o nome do modelo listado no nó continuará sendo o nome do último modelo carregado e copiado. Para se beneficiar desses recursos ou outro trabalho da sessão, atualize seu nó de modelagem antes de sair da sessão.

Gerenciando modelos

Há também algumas tarefas básicas de gerenciamento que você pode querer executar de tempos em tempos em seus modelos, como renomear seus modelos, importar e exportar modelos ou excluir os modelos obsoletos. Essas tarefas são executadas na caixa de diálogo Gerenciar Modelos. A importação e a exportação de modelos permite o compartilhamento de modelos com outros usuários. Consulte o tópico “Importando e exportando modelos” na página 171 para obter informações adicionais.

Nota: Não é possível renomear ou excluir os modelos que são instalados (ou fornecidos) com este produto. Em vez disso, se desejar renomear, é possível abrir o modelo instalado e criar um novo com o nome de sua escolha. É possível excluir seus modelos customizados; no entanto, se você tentar excluir um modelo fornecido, ele será reconfigurado para a versão instalada originalmente.

Para Renomear um Modelo

1. A partir dos menus, escolha **Recursos> Gerenciar Templos de Recursos**. A caixa de diálogo Gerenciar Modelos é aberta.
2. Selecione o modelo que você deseja renomear e clique em **Renomear**. A caixa de nome se torna um campo editável na tabela.
3. Digite um novo nome e pressione a tecla Enter. Uma caixa de diálogo de confirmação é aberta.
4. Se você estiver satisfeito com a mudança de nome, clique em **Sim**. Se não, clique em **Não**.

Para Excluir um Modelo

1. A partir dos menus, escolha **Recursos> Gerenciar Templos de Recursos**. A caixa de diálogo Gerenciar Modelos é aberta.
2. Na caixa de diálogo Gerenciar Modelos, selecione o modelo que deseja excluir.
3. Clique em **Excluir**. Uma caixa de diálogo de confirmação é aberta.
4. Clique em **Sim** para excluir ou clique em **Não** para cancelar a solicitação. Se você clicar em **Sim**, o modelo será excluído.

Importando e exportando modelos

É possível compartilhar modelos com outros usuários ou máquinas importando-os e exportando-os. Modelos são armazenados em um banco de dados interno, mas podem ser exportados como arquivos **.lrt* para seu disco rígido.

Como há circunstâncias nas quais você pode querer importar ou exportar modelos, há diversas caixas de diálogo que oferecem essas capacidades.

- Caixa de diálogo open Template no Editor de Template
- Carregue a caixa de diálogo Recursos no nó de modelagem Mineração de Texto e no nó Análise de Ligação de Texto.
- Caixa de diálogo Gerenciar Modelos no Editor de Template e Editor de Recursos.

Para Importar um Modelo

1. Na caixa de diálogo, clique em **Importar**. A caixa de diálogo Importar Modelo é aberta.

2. Selecione o arquivo de modelo de recurso (*.*lrt*) para importar e clique em **Importar**. É possível salvar o modelo sendo importado com outro nome ou sobrescrever o existente. A caixa de diálogo é fechada e agora o modelo aparece na tabela.

Para Exportar um Modelo

1. Na caixa de diálogo, selecione o modelo que deseja exportar e clique em **Exportar**. A caixa de diálogo Selecionar Diretório é aberta.
2. Selecione o diretório para o qual deseja exportar e clique em **Exportar**. Essa caixa de diálogo é fechada e o modelo é exportado e carrega a extensão de arquivo (*.*lrt*)

Saindo do Editor de Template

Ao concluir o trabalho no Editor de Template, é possível salvá-lo e sair do editor.

Para Sair do Editor de Template

1. A partir dos menus, escolha **Arquivo> Fechar**. A caixa de diálogo Salvar e Fechar é aberta.
2. Selecione **Salvar Mudanças no Modelo** para salvar o modelo aberto antes de fechar o editor.
3. Selecione **Publicar Bibliotecas** se desejar publicar alguma das bibliotecas no modelo aberto antes de fechar o editor. Se você selecionar essa opção, será solicitado que você selecione as bibliotecas para publicar. Veja o tópico "Publicando bibliotecas" na página 181 para obter mais informações.

Fazendo backup de recursos

Talvez você queira fazer backup de seus recursos de tempos em tempos como uma medida de segurança.

Importante! Quando você restaura, o conteúdo inteiro dos seus recursos é limpo e somente o conteúdo do arquivo de backup fica acessível no produto. Isso inclui qualquer trabalho aberto.

Nota: Só é possível fazer backup e restaurar a mesma versão principal do seu software. Por exemplo, se você fizer backup da versão 15, não será possível restaurá-lo para a versão 16.

Para fazer backup de recursos

1. A partir dos menus, escolha **Recursos> Ferramentas de Backup> Recursos de Backup**. A caixa de diálogo Backup é aberta.
2. Insira um nome para seu arquivo de backup e clique em **Salvar**. A caixa de diálogo é fechada e o arquivo de backup é criado.

Para restaurar os recursos

1. A partir dos menus, escolha **Recursos> Ferramentas de Backup> Restaurar Recursos**. Um alerta avisa que a restauração sobrescreverá o conteúdo atual do seu banco de dados.
2. Clique em **Sim** para continuar. A caixa de diálogo é aberta.
3. Selecione o arquivo de backup que deseja restaurar e clique em **Abrir**. A caixa de diálogo é fechada e os recursos são restaurados no aplicativo.

Importando arquivos de recursos

Se você tiver feito mudanças diretamente nos arquivos de recurso fora deste produto, é possível importá-las em uma biblioteca selecionada selecionando-a e prosseguindo com a importação. Quando você importa um diretório, é possível importar todos os arquivos suportados em uma biblioteca aberta específica também. Só é possível importar arquivos `*.txt`.

Cada arquivo importado deve conter somente uma entrada por linha e se o conteúdo estiver estruturado como:

- Uma lista de palavras ou frases (uma por linha). O arquivo é importado como uma lista de termos para um dicionário de tipo, em que o dicionário de tipo usa o nome do arquivo menos a extensão.

- Uma lista de entradas como `term1` <TAB> `term2`, então ela é importada como uma lista de sinônimos, em que `term1` é o conjunto do termo subjacente e `term2` é o termo de destino.

Para Importar um Único Arquivo de Recursos

1. A partir dos menus, escolha **Recursos> Importação de Arquivos> Arquivo Único de Importação**. A caixa de diálogo Arquivo de Importação é aberta.
2. Selecione o arquivo que deseja importar e clique em **Importar**. O conteúdo do arquivo é transformado em um formato interno e incluído em sua biblioteca.

Como importar todos os arquivos em um diretório

1. A partir dos menus, escolha **Recursos> Arquivos de Importação> Diretório de Encanas de Importação**. A caixa de diálogo Importar Diretório é aberta.
2. Selecione a biblioteca na qual deseja que todos os arquivos de recursos sejam importados da lista **Importar**. Se você selecionar a opção **Padrão**, uma nova biblioteca será criada usando o nome do diretório como seu nome.
3. Selecione o diretório do qual deseja importar arquivos. Subdiretórios não serão lidos.
4. Clique em **Importar**. A caixa de diálogo é fechada e agora o conteúdo desses arquivos de recursos importados aparece no editor em forma de dicionários e arquivos de recursos avançados.

Capítulo 15. Trabalhando com bibliotecas

Os recursos usados pelo mecanismo de extração para extrair e agrupar termos a partir dos dados de texto sempre contêm uma ou mais bibliotecas. É possível ver o conjunto de bibliotecas na árvore de biblioteca localizada na parte superior esquerda do Editor de Template e Editor de Recursos. As bibliotecas são compostas por três tipos de dicionários: Tipo, Substituição e Exclusão. Consulte o tópico Capítulo 16, “Sobre dicionários de biblioteca”, na página 183 para obter mais informações.

O modelo de recurso ou os recursos de TAP que você escolheu incluem várias bibliotecas para permitir que você comece a extrair conceitos imediatamente dos seus dados de texto. No entanto, é possível criar suas próprias bibliotecas, bem como publicá-las para poder reutilizá-las. Consulte o tópico “Publicando bibliotecas” na página 181 para obter informações adicionais.

Por exemplo, suponha que você trabalhe frequentemente com dados de texto relacionados à indústria automotiva. Após analisar seus dados, você decide que gostaria de criar alguns recursos customizados para manipular um jargão ou um vocabulário específico da indústria. Usando Editor de Template , é possível criar um novo modelo e, nele, uma biblioteca para extrair e agrupar termos automotivos. Como você precisará de informações nessa biblioteca novamente, você publica sua biblioteca em um repositório central, acessível na caixa de diálogo **Gerenciar Bibliotecas**, para que ela possa ser reutilizada independentemente em diferentes sessões de fluxo .

Suponha que você também esteja interessado em agrupar termos específicos de diferentes subindústrias, como dispositivos eletrônicos, mecanismos, sistemas de resfriamento ou até um determinado fabricante ou mercado. É possível criar uma biblioteca para cada grupo e depois publicá-las para que possam ser usadas com vários conjuntos de dados de texto. Dessa forma, é possível incluir as bibliotecas que melhor correspondem ao contexto dos seus dados de texto.

Nota: Recursos adicionais podem ser configurados e gerenciados na guia Recursos Avançados. Alguns se aplicam a todas as bibliotecas e gerenciam entidades não linguísticas, exceções de agrupamento difuso, entre outros. Além disso, é possível editar regras de padrão de análise de ligação de texto, que são específicas da biblioteca, na guia Regras de Ligação de Texto também. Consulte o tópico Capítulo 17, “Sobre recursos avançados”, na página 195 para obter mais informações.

Bibliotecas enviadas

Por padrão, várias bibliotecas são instaladas com IBM SPSS Modeler Text Analytics. É possível usar essas bibliotecas pré-formatadas para acessar milhares de termos e sinônimos predefinidos, bem como muitos tipos diferentes. Essas bibliotecas enviadas são ajustadas para diversos domínios diferentes e estão disponíveis em vários idiomas diferentes.

Há inúmeras bibliotecas, mas as mais usadas são as seguintes:

- **Biblioteca local.** Usada para armazenar dicionários definidos pelo usuário. Ela é uma biblioteca vazia incluída por padrão para todos os recursos. Ela também contém um dicionário de tipo vazio. Ela é mais útil quando você faz mudanças ou refinamentos diretamente (como incluir uma palavra para um tipo) a partir da visualização Categorias e Conceitos, visualização Clusters e visualização Análise de Ligação de Texto . Nesse caso, essas mudanças e refinamentos são armazenados automaticamente na primeira biblioteca listada na árvore de biblioteca no Editor de Recursos; por padrão, essa é a *biblioteca local*. Não é possível publicar essa biblioteca porque ela é específica para os dados de sessão . Se desejar publicar seu conteúdo, deve-se renomear a biblioteca primeiro.
- **Biblioteca de núcleo.** Usada na maioria dos casos, já que abrange os cinco tipos básicos integrados que representam: pessoas, localizações, organizações, produtos e desconhecido. Embora você possa ver somente alguns termos listados em um de seus dicionários de tipo, os tipos representados na biblioteca Núcleo são, de fato, complementos para os tipos robustos localizados nos recursos compilados internos entregues com seu produto de mineração de texto. Esses recursos compilados internos contêm milhares de termos para cada tipo. Por esse motivo, embora você não possa ver um termo na lista de termos do dicionário de tipo, ele ainda pode ser extraído e tipificado com um

tipo Núcleo. Isso explica como nomes como *George* podem ser extraídos e digitados como <Person> quando apenas *John* aparece no dicionário do tipo <Person> na biblioteca Core. Da mesma forma, se não incluir a biblioteca Núcleo, você ainda poderá ver esses tipos em seus resultados de extração, já que os recursos compilados contendo esses tipos ainda serão usados pelo mecanismo de extração.

- **Biblioteca de pareceres.** Usado mais comumente para extrair pareceres e impressões dos dados de texto. Essa biblioteca inclui milhares de palavras representando atitudes, qualificadores e preferências que - quando usados em conjunto com outros termos — indicam um parecer sobre um assunto. Essa biblioteca inclui inúmeros tipos, sinônimos e exclusões integrados. Ela também inclui um grande conjunto de regras de padrão para análise de ligação de texto. Para se beneficiar das regras de análise de ligação de texto e dos resultados de padrão que elas produzem, essa biblioteca deve ser especificada na guia Regras de Ligação de Texto. Consulte o tópico Capítulo 18, “Sobre regras de ligação de texto”, na página 207 para obter mais informações.
- **Biblioteca de orçamento.** Usada para extrair termos que se referem ao custo de algo. Essa biblioteca inclui muitas palavras e frases que representam adjetivos, qualificadores e julgamentos referentes ao preço ou à qualidade de algo.
- **Biblioteca de variações.** Usada para incluir casos em que certas variações de idioma requerem definições de sinônimo para agrupá-las adequadamente. Essa biblioteca inclui somente definições de sinônimo.

Embora algumas das bibliotecas enviadas fora dos modelo lembrem os conteúdos em alguns modelos, os modelos foram ajustados especificamente para determinados aplicativos e contêm recursos avançados adicionais. É recomendado tentar usar um modelo que foi projetado para o tipo de dados de texto com o qual você está trabalhando e fazer mudanças nesses recursos em vez de apenas incluir bibliotecas individuais em um modelo mais genérico.

Os recursos compilados também são entregues com IBM SPSS Modeler Text Analytics. Eles são sempre usados durante o processo de extração e contêm um grande número de definições complementares para os dicionários de tipo integrados nas bibliotecas padrão. Como esses recursos são compilados, eles não podem ser visualizados ou editados. No entanto, é possível forçar um termo que foi tipificado por esses recursos compilados para qualquer outro dicionário. Veja o tópico “Forçando termos” na página 188 para obter mais informações.

Criando bibliotecas

É possível criar qualquer número de bibliotecas. Após criar uma nova biblioteca, é possível começar a criar dicionários de tipo nessa biblioteca e inserir termos, sinônimos e exclusões.

Para criar uma biblioteca

1. A partir dos menus, escolha **Recursos> Nova Biblioteca**. O diálogo Propriedades da Biblioteca é aberto.
2. Insira um nome para a biblioteca na caixa de texto Nome.
3. Se desejar, insira um comentário na caixa de texto Anotação.
4. Clique em **Publicar** se desejar publicar essa biblioteca agora antes de inserir qualquer coisa nela. Veja o tópico “Compartilhando bibliotecas” na página 180 para obter mais informações. Também é possível publicar a qualquer momento depois.
5. Clique em **OK** para criar a biblioteca. A caixa de diálogo é fechada e a biblioteca aparece na visualização em árvore. Se expandir as bibliotecas na árvore, você verá que um dicionário de tipo vazio foi incluído automaticamente na biblioteca. É possível começar a incluir termos nele imediatamente. Veja o tópico “incluindo termos” na página 186 para obter mais informações.

Incluindo bibliotecas públicas

Se desejar reutilizar uma biblioteca de outros dados de sessão , é possível incluí-la nos atuais recursos, contanto que ela seja biblioteca pública. Uma *biblioteca pública* é uma biblioteca que foi publicada. Consulte o tópico “Publicando bibliotecas” na página 181 para obter mais informações.

Quando você inclui uma biblioteca pública, uma cópia *local* é integrada aos dados de sessão . É possível fazer mudanças nessa biblioteca; no entanto, deve-se publicar novamente a versão pública da biblioteca se você desejar compartilhar as mudanças.

Quando você inclui uma biblioteca pública, uma caixa de diálogo Resolver Conflitos pode aparecer se algum conflito for descoberto entre os termos e os tipos em uma biblioteca e as outras bibliotecas locais. Deve-se resolver esses conflitos ou aceitar as resoluções propostas para concluir essa operação. Consulte o tópico "Resolvendo conflitos" na página 181 para obter informações adicionais.

Nota: Se você sempre atualizar suas bibliotecas quando você lançar uma sessão de ambiente de trabalho interativa ou publicar quando fechar uma , é menos provável que você tenha bibliotecas que estejam fora de sincronia. Consulte o tópico "Compartilhando bibliotecas" na página 180 para obter mais informações.

Para Incluir uma Biblioteca

1. A partir dos menus, escolha **Recursos> Adicionar Biblioteca**. A caixa de diálogo Incluir Biblioteca é aberta.
2. Selecione a biblioteca ou bibliotecas na lista.
3. Clique em **Incluir**. Se ocorrer algum conflito entre as bibliotecas recém-incluídas e aquelas que já estavam lá, será solicitado que você verifique as resoluções de conflito ou mude-as antes de concluir a operação. Veja o tópico "Resolvendo conflitos" na página 181 para obter mais informações.

Localizando termos e tipos

É possível procurar nas várias áreas de janela no editor usando a variável Localizar. No editor, você pode escolher **Editar> Localizar** a partir dos menus e a barra de ferramentas Localizar aparecer. É possível usar essa barra de ferramentas para localizar um ocorrência por vez. Clicando em **Localizar** novamente, é possível localizar ocorrências subsequentes de seu termo de procura.

Durante a procura, o editor procura somente na biblioteca ou bibliotecas listadas na lista suspensa na barra de ferramentas Localizar. Se **Todas as Bibliotecas** for selecionada, o programa procurará tudo no editor.

Quando você inicia uma procura, ela começa na área que tem o foco. A procura continua em cada seção, fazendo um loopback até retornar à célula ativa. É possível inverter a ordem da procura usando as setas direcionais. Também é possível escolher se a procura faz ou não distinção entre maiúsculas e minúsculas.

Para Localizar Sequências de Caracteres na Visualização

1. A partir dos menus, escolha **Editar** > **Localizar**. A barra de ferramentas Localiza é exibida.
2. Insira a sequência de caracteres que deseja procurar.
3. Clique no botão **Localizar** para iniciar a procura. A próxima ocorrência do termo ou tipo é destacada.
4. Clique no botão novamente para se mover de uma ocorrência para outra.

Usando um asterisco em termos

Usar um asterisco (*) em termos é especialmente útil se você estiver lidando com uma linguagem aglutinativa que cria novas palavras compondo outras palavras juntas sem espaços interditados. Por exemplo, a palavra alemã *Übernachtungspreis*, que é composta por: *Übernachtung* + *s* + *Preis*.

Como exemplo, se você pesquisar em termos para `preis*` no tipo `Budget`, ele corresponderá a conceitos extraídos como `preiserhöhung`. Da mesma forma, `*preis` corresponderá a `Übernachtung` e `*preis*` corresponderá a `Übernachtungspreiserhöhung`.

Visualizando bibliotecas

É possível exibir o conteúdo de uma determinada biblioteca ou de todas as bibliotecas. Isso pode ser útil quando você estiver lidando com várias bibliotecas ou quando quiser revisar o conteúdo de uma

biblioteca específica antes de publicá-la. A mudança da visualização impacta somente aquilo que você vê nessa guia Recursos de Biblioteca, mas não impede que nenhuma biblioteca seja usada durante a extração. Consulte o tópico "Desativando biblioteca locais" na página 178 para obter informações adicionais.

A visualização padrão é **Todas as Bibliotecas**, que mostra todas as bibliotecas na árvore e seu conteúdo em outras áreas de janela. Você pode alterar esta seleção usando a lista suspensa na barra de ferramentas ou através de uma seleção de menu (**View> Libraries**) Quando uma única biblioteca está sendo visualizada, todos os itens em outras bibliotecas desaparecem da visualização mas ainda são lidos durante a extração.

Para mudar a visualização Biblioteca

1. A partir dos menus na aba Recursos da Biblioteca, escolha **Visualizar> Bibliotecas**. Um menu com todas as bibliotecas locais é aberto.
2. Selecione a biblioteca que deseja ver ou selecione a opção **Todas as Bibliotecas** para ver o conteúdo de todas as bibliotecas. O conteúdo da visualização é filtrado de acordo com sua seleção.

Gerenciando bibliotecas locais

Bibliotecas locais são as bibliotecas dentro de sua sessão do ambiente de trabalho interativa ou dentro de um modelo, ao contrário de bibliotecas públicas. Consulte o tópico "Gerenciando bibliotecas públicas" na página 179 para obter mais informações. Também existem algumas tarefas de gerenciamento de biblioteca local básicas que você pode querer executar, incluindo: renomear, desativar ou excluir uma biblioteca local.

Renomeando bibliotecas locais

É possível renomear bibliotecas locais. Se você renomear uma biblioteca local, ela será desassociada da versão pública, caso uma versão pública exista. Isso significa que mudanças subsequentes não poderão mais ser compartilhadas com a versão pública. É possível publicar novamente essa biblioteca local sob seu novo nome. Isso também significa que não será possível atualizar a versão pública original com nenhuma mudança feita nessa versão local.

Nota: Não é possível renomear uma biblioteca pública.

1. A partir dos menus, escolha **Editar> Propriedades da Biblioteca**. A caixa de diálogo Propriedades da Biblioteca é aberta.

Para Renomear uma Biblioteca Local

1. Na visualização em árvore, selecione a biblioteca que deseja renomear.
2. Insira um novo nome para a biblioteca na caixa de texto Nome.
3. Clique em **OK** para aceitar o novo nome para a biblioteca. A caixa de diálogo fecha e o nome da biblioteca é atualizado na visualização em árvore.

Desativando biblioteca locais

Se desejar excluir provisoriamente uma biblioteca do processo de extração, é possível cancelar a seleção da caixa à esquerda do nome da biblioteca na visualização em árvore. Isso sinaliza que você deseja manter a biblioteca, mas quer que seu conteúdo seja ignorado durante a verificação de conflitos e durante a extração.

Para Desativar uma Biblioteca

1. Na área de janela em árvore de biblioteca, selecione a biblioteca que deseja desativar.
2. Clique na barra de espaço. A caixa de seleção à esquerda do nome é limpa.

Excluindo bibliotecas locais

É possível remover uma biblioteca sem excluir a versão pública da biblioteca e vice-versa. A exclusão de uma biblioteca local excluirá a biblioteca e todo o seu conteúdo somente da sessão . A exclusão de uma versão local de uma biblioteca não remove essa biblioteca de outras sessões ou da versão pública. Veja o tópico “Gerenciando bibliotecas públicas” na página 179 para obter mais informações.

Para Excluir uma Biblioteca Local

1. Na visualização em árvore, selecione a biblioteca que deseja excluir.
2. A partir dos menus, escolha **Editar> Excluir** para excluir a biblioteca. A biblioteca é removida.
3. Se você nunca tiver publicado essa biblioteca antes, uma mensagem perguntando se você gostaria de excluir ou manter essa biblioteca será aberta. Clique em **Excluir** para continuar ou em **Manter** se quiser manter essa biblioteca.

Nota: Uma biblioteca deve sempre permanecer.

Gerenciando bibliotecas públicas

Para reutilizar as bibliotecas locais, é possível publicá-las e depois trabalhar com elas e vê-las através da caixa de diálogo Gerenciar Bibliotecas (**Recursos> Gerenciar Bibliotecas**). Veja o tópico “Compartilhando bibliotecas” na página 180 para obter mais informações. Algumas tarefas de gerenciamento de biblioteca pública básicas que você pode querer executar incluem importar, exportar ou excluir uma biblioteca pública. Não é possível renomear uma biblioteca pública.

Importando bibliotecas públicas

1. Na caixa de diálogo Gerenciar Bibliotecas, clique em **Importar...**. A caixa de diálogo Importar Biblioteca é aberta.
2. Selecione o arquivo de biblioteca (**.lib*) que deseja importar e, se também quiser incluir essa biblioteca localmente, selecione **Incluir Biblioteca no Projeto Atual**.
3. Clique em **Importar**. A caixa de diálogo é fechada. Se uma biblioteca pública com o mesmo nome já existir, será solicitado que você renomeie a biblioteca sendo importada ou sobrescreva a atual biblioteca pública.

Exportando bibliotecas públicas

É possível exportar bibliotecas públicas no formato *.lib* para que seja possível compartilhá-las.

1. Na caixa de diálogo Gerenciar Biblioteca, selecione a biblioteca que deseja exportar na lista.
2. Clique em **Exportar**. A caixa de diálogo Selecionar Diretório é aberta.
3. Selecione o diretório para o qual deseja exportar e clique em **Exportar**. A caixa de diálogo fecha e o arquivo de biblioteca (**.lib*) é exportado.

Excluindo bibliotecas públicas

É possível remover uma biblioteca local sem excluir sua versão pública e vice-versa. No entanto, se a biblioteca for excluída da caixa de diálogo, ela não poderá mais ser incluída em nenhum recurso da sessão até que uma versão local seja publicada novamente.

Se você excluir uma biblioteca que foi instalada com o produto, a versão instalada originalmente será restaurada.

1. Na caixa de diálogo Gerenciar Bibliotecas, selecione a biblioteca que deseja excluir. É possível ordenar a lista clicando no cabeçalho apropriado.
2. Clique em **Excluir** para excluir a biblioteca. IBM SPSS Modeler Text Analytics verifica se a versão local da biblioteca é a mesma da biblioteca pública. Se sim, a biblioteca será removida sem alerta. Se as versões da biblioteca forem diferentes, um alerta será aberto perguntando se você deseja manter ou remover a versão pública.

Compartilhando bibliotecas

Bibliotecas permitem trabalhar com recursos de uma maneira fácil de compartilhar entre várias sessões de ambiente de trabalho interativas . Bibliotecas podem existir em dois estados ou versões. Bibliotecas que são editáveis no editor e parte de uma sessão de ambiente de trabalho interativa são chamadas **bibliotecas locais**. Enquanto trabalha com uma sessão de ambiente de trabalho interativa , você pode fazer várias mudanças na biblioteca *Vegetais*, por exemplo. Se suas mudanças pudessem ser úteis com outros dados, seria possível disponibilizar esses recursos criando uma versão de **biblioteca pública** da biblioteca *Vegetais*. Uma biblioteca pública, como o nome mostra, está disponível para quaisquer outros recursos em qualquer sessão de ambiente de trabalho interativa.

É possível ver as bibliotecas públicas na caixa de diálogo Gerenciar Bibliotecas. Uma vez que essa versão da biblioteca pública existe, é possível incluí-la nos recursos em outros contextos para que esses recursos linguísticos customizados possam ser compartilhados.

As bibliotecas fornecidas são bibliotecas públicas inicialmente. É possível editar os recursos nessas bibliotecas e depois criar uma nova versão pública. Essas novas versões ficariam então acessíveis em outras sessões de ambiente de trabalho interativas.

Conforme você continua trabalhando com suas bibliotecas e fazendo mudanças, suas versões de biblioteca ficam dessincronizadas. Em alguns casos, uma versão local pode ser mais recente que a versão pública, e em outros casos, a versão pública pode ser mais recente que a versão local. Também é possível que as versões pública e local contenham mudanças que as outras não contêm se a versão pública tiver sido atualizada de dentro de outra sessão de ambiente de trabalho interativa. Se as versões de sua biblioteca ficarem dessincronizadas, é possível sincronizá-las novamente. A sincronização de versões de biblioteca consiste em publicar novamente e/ou atualizar bibliotecas locais.

Sempre que você ativar uma sessão de ambiente de trabalho interativa ou fechar uma , será solicitado que você sincronize novamente quaisquer bibliotecas que precisem de atualização ou nova publicação. Além disso, é possível identificar facilmente o estado de sincronização de sua biblioteca local pelo ícone que aparece ao lado do nome da biblioteca na visualização em árvore ou visualizando a caixa de diálogo Propriedades da Biblioteca. Também é possível escolher fazer isso a qualquer momento por meio de seleções de menu. A tabela a seguir descreve os cinco estados possíveis e seus ícones associados.

Tabela 39. Estados de sincronização de biblioteca local

Ícone	Descrição do status da biblioteca local
	Não publicado - A biblioteca local nunca foi publicada.
	Sincronizada - As versões de bibliotecas públicas e locais são idênticas. Isso também se aplica à *Biblioteca Local*, que não pode ser publicada porque deve conter somente recursos específicos da sessão .
	Desatualizada - A versão da biblioteca pública é mais recente que a da versão local. É possível atualizar sua versão local com mudanças.
	Mais nova — A versão da biblioteca local é mais recente que a da versão pública. É possível publicar novamente sua versão local para a versão pública.
	Fora de sincronização - As bibliotecas local e pública contêm mudanças que as outras não contêm. Você deve decidir se irá atualizar ou publicar sua biblioteca local. Se atualizar, você perderá todas as mudanças feitas desde a última vez em que atualizou ou publicou. Se escolher publicar, você sobrescreverá as mudanças na versão pública.

Nota: Se você sempre atualizar suas bibliotecas ao ativar uma sessão de ambiente de trabalho interativa ou publicar ao fechar uma , é menos provável que você tenha bibliotecas fora de sincronização.

É possível publicar uma biblioteca novamente sempre que você achar que as mudanças na biblioteca beneficiariam outros fluxos que possam conter essa biblioteca. Então, se suas mudanças beneficiarem outros fluxos, é possível atualizar as versões locais nesses fluxos. Dessa forma, é possível criar fluxos

para cada contexto ou domínio que se aplique aos seus dados criando novas bibliotecas e/ou incluindo inúmeras bibliotecas públicas em seus recursos.

Se uma versão pública de uma biblioteca for compartilhada, há mais chances de surgirem diferenças entre as versões pública e local. Sempre que você ativa ou fecha e publica a partir de uma sessão de ambiente de trabalho interativa ou abre ou fecha um modelo a partir do Editor de Template , uma mensagem é exibida para permitir que você publique e/ou atualize quaisquer bibliotecas cujas versões não estão sincronizadas com aquelas na caixa de diálogo Gerenciar Bibliotecas. Se a versão da biblioteca pública for mais recente que a da versão local, uma caixa de diálogo perguntando se você gostaria de atualizar será aberta. É possível escolher se você deseja manter a versão local no estado em que ela se encontra em vez de atualizá-la com a versão pública ou mesclar atualizações na biblioteca local.

Publicando bibliotecas

Se você nunca tiver publicado uma biblioteca específica, a publicação envolverá a criação de uma cópia pública de sua biblioteca local no banco de dados. Se você estiver publicando uma biblioteca novamente, o conteúdo da biblioteca local substituirá o conteúdo da versão pública existente. Após você publicar novamente, é possível atualizar essa biblioteca em outras sessões de fluxo quaisquer para que suas versões locais fiquem sincronizadas com a versão pública. Mesmo que seja possível publicar uma biblioteca, uma versão local é sempre armazenada na sessão .

Importante! Se você fizer mudanças em sua biblioteca local e, nesse meio tempo, a versão pública da biblioteca também mudar, sua biblioteca será considerada fora de sincronização. É recomendado começar atualizando a versão local com as mudanças públicas, fazer todas as mudanças desejadas e, depois, publicar sua versão local novamente para que deixar as duas idênticas. Se fizer mudanças e publicar primeiro, você sobrescreverá quaisquer mudanças na versão pública.

Para Publicar Bibliotecas Locais no Banco de Dados

1. A partir dos menus, escolha **Recursos> Publicar Bibliotecas**. A caixa de diálogo Publicar Bibliotecas é aberta com todas as bibliotecas precisando de publicação selecionadas por padrão.
2. Marque a caixa de seleção à esquerda de cada biblioteca que deseja publicar ou publicar novamente.
3. Clique em **Publicar** para publicar as bibliotecas no banco de dados Gerenciar Bibliotecas.

Atualizando bibliotecas

Sempre que você ativa ou fecha uma sessão de ambiente de trabalho interativa , é possível atualizar ou publicar quaisquer bibliotecas que não estejam mais sincronizadas com as versões públicas. Se a versão da biblioteca pública for mais recente que a versão local, uma caixa de diálogo perguntando se você gostaria de atualizar a biblioteca será aberta. É possível escolher se você deseja manter a versão local em vez de atualizá-la com a versão pública ou substituir a versão local pela pública. Se uma versão pública de uma biblioteca for mais recente que sua versão local, é possível atualizar a versão local para sincronizar seu conteúdo com o da versão pública. Atualização significa incorporar as mudanças localizadas na versão pública à versão local.

Nota: Se você sempre atualizar suas bibliotecas ao ativar uma sessão de ambiente de trabalho interativa ou publicar ao fechar uma , é menos provável que você tenha bibliotecas fora de sincronização. Consulte o tópico “Compartilhando bibliotecas” na página 180 para obter mais informações.

Para atualizar bibliotecas locais

1. A partir dos menus, escolha **Recursos> Bibliotecas de Atualização**. A caixa de diálogo Atualizar Bibliotecas é aberta, com todas as bibliotecas precisando de atualização selecionadas por padrão.
2. Marque a caixa de seleção à esquerda de cada biblioteca que deseja publicar ou publicar novamente.
3. Clique em **Atualizar** para atualizar a biblioteca local.

Resolvendo conflitos

Conflitos de biblioteca local versus pública

Sempre que você ativa uma sessão de fluxo , o IBM SPSS Modeler Text Analytics executa uma comparação das bibliotecas locais e daquelas listadas na caixa de diálogo Gerenciar Bibliotecas. Se alguma biblioteca local em sua sessão não estiver sincronizada com as versões publicadas, a caixa de diálogo Aviso de Sincronização de Biblioteca será aberta. É possível escolher entre as seguintes opções para selecionar as versões de biblioteca que você deseja usar aqui:

- **Todas as bibliotecas locais para o arquivo.** Esta opção mantém todas as suas bibliotecas locais como elas são. Sempre é possível publicá-las novamente ou atualizá-las depois.
- **Todas as bibliotecas publicadas nesta máquina.** Esta opção substituirá as bibliotecas locais mostradas com as versões localizadas no banco de dados.
- **Todas as bibliotecas mais recentes.** Esta opção substituirá quaisquer bibliotecas locais mais antigas pelas versões públicas mais recentes do banco de dados.
- **Outro.** Esta opção permite selecionar manualmente as versões desejadas escolhendo-as na tabela.

Conflitos de Termo Forçado

Sempre que você incluir uma biblioteca pública ou atualiza uma biblioteca local, conflitos e entradas duplicadas podem ser descobertos entre os termos e os tipos na biblioteca e os termos e os tipos nas outras bibliotecas em seus recursos. Se isso acontecer, será solicitado que você verifique as resoluções de conflito propostas ou mude-as antes de concluir a operação na caixa de diálogo Editar Termos Forçados. Consulte o tópico “Forçando termos” na página 188 para obter informações adicionais.

A caixa de diálogo Editar Termos Forçados contém cada par de termos ou tipos conflitantes. As cores do plano de fundo alternativas são usadas para distinguir visualmente cada par de conflitos. Essas cores podem ser mudadas na caixa de diálogo Opções. Consulte o tópico “Opções: guia Exibir” na página 74 para obter mais informações. A caixa de diálogo Editar Termos Forçados contém duas guias:

- **Duplicatas.** Esta guia contém os termos duplicados localizados nas bibliotecas. Se um ícone de tacha aparecer após um termo, isso significa que essa ocorrência do termo foi forçada. Se um ícone de X preto aparecer, isso significa que essa ocorrência do termo será ignorada durante a extração porque ela foi forçada em outras partes.
- **Definido pelo Usuário.** Esta guia contém uma lista de quaisquer termos que foram forçados manualmente na área de janela de dicionário de tipo, e não por meio de conflitos.

Nota: A caixa de diálogo Editar Termos Forçados é aberta após você incluir ou atualizar uma biblioteca. Se cancelar essa caixa de diálogo, você não estará cancelando a atualização ou adição da biblioteca.

Para Resolver Conflitos

1. Na caixa de diálogo Editar Termos Forçados, selecione o botão de opções na coluna Uso para o termo que deseja forçar.
2. Quando tiver concluído, clique em **OK** para aplicar os termos forçados e feche a caixa de diálogo. Se clicar em **Cancelar**, você cancelará as mudanças feitas nessa caixa de diálogo.

Capítulo 16. Sobre dicionários de biblioteca

Os recursos usados para extrair dados de texto são armazenados em forma de modelos e bibliotecas. Uma biblioteca pode ser composta por três dicionários.

- O **dicionário de tipo** contém uma coleção de termos agrupados sob um rótulo ou nome de tipo. Quando o mecanismo de extração lê seus dados de texto, ele compara as palavras localizadas no texto com os termos definidos em seus dicionários de tipo. Durante a extração, as formas flexionadas do dos termos e sinônimos de um tipo são agrupadas sob um termo de resposta chamado conceito. Conceitos extraídos são designados ao dicionário de tipo no qual aparecem como termos. É possível gerenciar seus dicionários de tipo nas áreas de janela central e superior esquerda do editor - a árvore de biblioteca e a área de janela de termo. Consulte o tópico “Dicionários de tipo” na página 183 para obter mais informações.
- O **dicionário de substituição** contém uma coleção de palavras definidas como sinônimos ou elementos opcionais usada para agrupar termos semelhantes sob um termo de resposta chamado conceito nos resultados da extração final. É possível gerenciar seus dicionários de substituição na área de janela inferior esquerda do editor usando a guia Sinônimos e a guia Opcional. Consulte o tópico “Dicionários de substituição/sinônimo” na página 190 para obter mais informações.
- O **dicionário de exclusão** contém uma coleção de termos e tipos que serão removidos dos resultados da extração final. É possível gerenciar seus dicionários de exclusão na área de janela à direita do editor. Veja o tópico “Dicionários de exclusão” na página 193 para obter mais informações.

Consulte o tópico Capítulo 15, “Trabalhando com bibliotecas”, na página 175 para obter informações adicionais.

Dicionários de tipo

Um *dicionário de tipo* é composto por um nome de tipo, ou rótulo, e uma lista de termos. Os dicionários do tipo são gerenciados nas panes superior esquerda e central da guia Recursos da Biblioteca no editor. Você pode acessar esta visualização com **Visualizar> Editor de Recursos** nos menus, se você estiver em uma sessão interativa de ambiente de trabalho. Caso contrário, é possível editar dicionários para um modelo específico no Editor de Template.

Quando o mecanismo de extração lê seus dados de texto, ele compara as palavras localizadas no texto com os termos definidos em seus dicionários de tipos. Termos são palavras ou frases nos dicionários de tipos em seus recursos linguísticos.

Quando uma palavra corresponde a um termo, ela é designada ao nome do tipo para esse termo. Quando os recursos são lidos durante a extração, os termos que foram localizados no texto então passam por diversos passos de processamento antes de se tornarem conceitos na área de janela Resultados da Extração. Se for determinado pelo mecanismo de extração que vários termos pertencentes ao mesmo dicionário de tipo são sinônimos, eles serão agrupados sob o termo que ocorre com mais frequência e chamados de *conceito* na área de janela Resultados da Extração. Por exemplo, os termos `question` e `query` poderem aparecer sob o nome do conceito `question` no final.

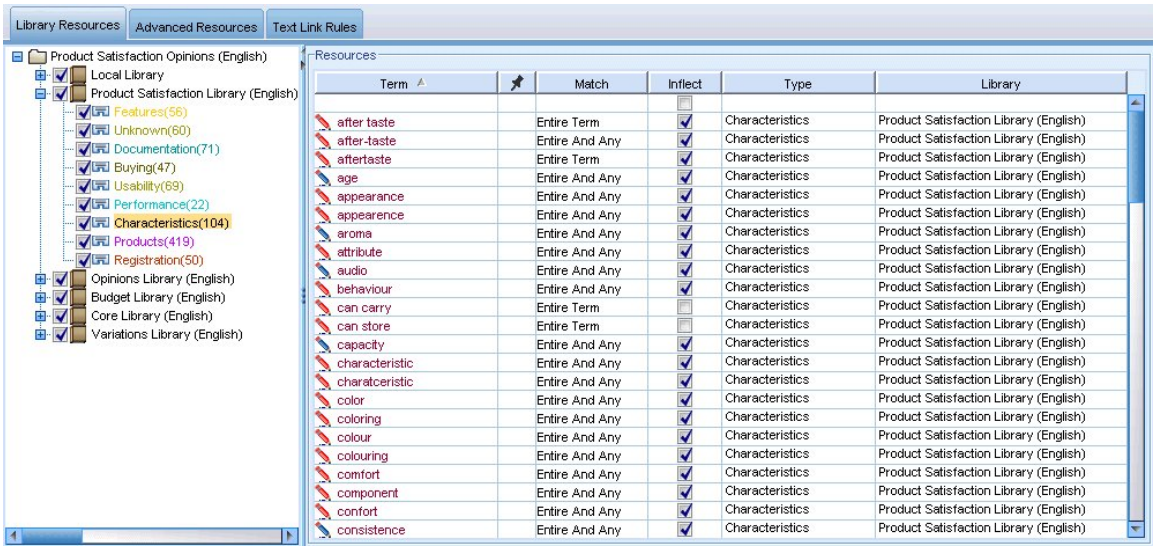


Figura 42. Árvore de biblioteca e área de janela de termo

A lista de dicionários de tipo é mostrada na área de janela em árvore de biblioteca à esquerda. O conteúdo de cada dicionário de tipo aparece na área de janela central. Dicionários de tipo consistem em mais do que apenas uma lista de termos. A maneira como as palavras e as frases em seus dados de texto correspondem aos termos definidos nos dicionários de tipo é determinada pela opção de correspondência definida. Uma **opção de correspondência** especifica como um termo é ancorado com relação a uma palavra ou frase candidata nos dados de texto. Consulte o tópico "incluindo termos" na página 186 para obter mais informações.

Além disso, é possível ampliar os termos em seu dicionário de tipo especificando se você deseja gerar e incluir automaticamente formas flexionadas dos termos no dicionário. Ao gerar formas flexionadas, você inclui automaticamente as formas no plural dos termos no singular, as formas no singular dos termos no plural e adjetivos no dicionário de tipo. Consulte o tópico "incluindo termos" na página 186 para obter informações adicionais.

Nota: Para a maioria das línguas, conceitos que não são encontrados em nenhum dicionário do tipo mas são extraídos do texto são digitados automaticamente como <Unknown>

Usando um asterisco em termos

Usar um asterisco (*) em termos é especialmente útil se você estiver lidando com uma linguagem aglutinativa que cria novas palavras compondo outras palavras juntas sem espaços interditados. Por exemplo, a palavra alemã *Übernachtungspreis*, que é composta por: *Übernachtung* + *s* + *Preis*.

Como exemplo, se você pesquisar em termos para `preis*` no tipo `Budget`, ele corresponderá a conceitos extraídos como `preiserhöhung`. Da mesma forma, `*preis` corresponderá a `Übernachtung` e `*preis*` corresponderá a `Übernachtungspreiserhöhung`.

Tipos integrados

IBM SPSS Modeler Text Analytics é entregue com um conjunto de recursos linguísticos em forma de bibliotecas enviadas e recursos compilados. As bibliotecas embarcadas contêm um conjunto de dicionários do tipo integrado, como `<Location>`, `<Organization>`, `<Person>`e `<Product>`.

Esses dicionários do tipo são usados pelo mecanismo de extração para atribuir tipos aos conceitos que ele extrai como atribuído o tipo `<Location>` ao conceito `paris`. Embora um grande número de termos tenha sido definido nos dicionários de tipo integrado, eles não cobrem cada possibilidade. Portanto, é possível incluí-los ou criar os seus. Para obter uma descrição do conteúdo de um determinado dicionário de tipo enviado, leia a anotação na caixa de diálogo Propriedades de Tipo. Selecione o tipo na árvore e escolha **Editar> Propriedades** a partir do menu de contexto.

Nota:

Além das bibliotecas embarcadas, os recursos compilados (também utilizados pelo motor de extração) contêm um grande número de definições complementares aos dicionários do tipo embutido, mas seu conteúdo não é visível no produto. Porém, é possível forçar um termo que foi tipificado pelos dicionários compilados em qualquer outro dicionário. Veja "Forçando termos" na página 188 para obter mais informações.

Criando tipos

É possível criar dicionários de tipos para ajudar a agrupar termos semelhantes. Quando os termos que aparecem nesse dicionário são descobertos durante o processo de extração, eles são designados a esse nome de tipo e extraídos sob um nome de conceito. Sempre que você cria uma biblioteca, uma biblioteca de tipos vazia é sempre incluída para que seja possível começar a inserir termos imediatamente.

Se estiver analisando texto sobre comida e desejar agrupar termos relacionados a vegetais, você pode criar seu próprio dicionário do tipo <Vegetables>. Você poderia então incluir termos como `carrot`, `broccoli` e `spinach` se achar que eles são termos importantes que aparecerão no texto. Depois, durante a extração, se algum desses termos for localizado, eles serão extraídos como conceitos e designados ao tipo <Vegetables>.

Você não precisa definir cada forma de uma palavra ou expressão, pois é possível escolher gerar as formas flexionadas dos termos. Quando você escolhe essa opção, o mecanismo de extração reconhece automaticamente as formas no singular e no plural dos termos entre as formas pertencentes a esse tipo. Essa opção é especialmente útil quando seu tipo contém principalmente nomes, já que é improvável que você queira formas flexionadas de verbos ou adjetivos.

A caixa de diálogo Propriedades de Tipo contém os seguintes campos.

Nome. O nome que você dá ao dicionário de tipo sendo criado. É recomendado não usar espaços nos nomes de tipo, principalmente se dois ou mais nomes de tipo começarem com a mesma palavra.

Nota: Existem algumas restrições sobre nomes de tipos e o uso de símbolos. Por exemplo, não use símbolos como "@" ou "!" dentro de um nome.

partida padrão. O atributo de correspondência padrão instrui o mecanismo de extração a fazer a correspondência desse termo com os dados de texto. Sempre que você inclui um termo nesse dicionário de tipo, esse é o atributo de correspondência automaticamente designado a ele. Sempre é possível mudar a opção de correspondência manualmente na lista de termos. As opções incluem: **Termo Inteiro**, **Início**, **Término**, **Qualquer um**, **Início ou Término**, **Inteiro e Início**, **Inteiro e Término**, **Inteiro e (Início ou Término)** e **Inteiro (sem compostos)**. Veja o tópico "incluindo termos" na página 186 para obter mais informações.

Adicionar a. Este campo indica a biblioteca na qual você criará seu novo dicionário de tipos.

Gerar formas inflected por padrão. Esta opção diz ao mecanismo de extração para usar morfologia gramatical para capturar e agrupar formas semelhantes dos termos que você inclui nesse dicionário, como formas no singular ou no plural do termo. Essa opção é especialmente útil quando seu tipo contém principalmente nomes. Quando você seleciona essa opção, todos os novos termos incluídos nesse tipo têm automaticamente essa opção, embora seja possível mudá-la manualmente na lista.

cor da fonte. Este campo permite distinguir os resultados deste tipo dos outros na interface. Se você selecionar **Usar cor pai**, a cor do tipo padrão também será usada para esse tipo de dicionário. Essa cor padrão é configurada na caixa de diálogo de opções. Consulte o tópico "Opções: guia Exibir" na página 74 para obter mais informações. Se você selecionar **Custom**, selecione uma cor na lista suspensa.

Anotação. Este campo é opcional e pode ser usado para quaisquer comentários ou descrições.

Para Criar um Dicionário de Tipo

1. Selecione a biblioteca na qual gostaria de criar um novo dicionário de tipo.
2. A partir dos menus, escolha **Ferramentas> Novo tipo**. A caixa de diálogo Propriedades de Tipo é aberta.
3. Insira o nome de seu dicionário de tipo na caixa de texto **Nome** e escolha as opções desejadas.

4. Clique em **OK** para criar o dicionário de tipo. O novo tipo fica visível na área de janela em árvore de biblioteca e aparece na área de janela central. É possível começar a incluir termos imediatamente. Para obter mais informações, consulte “incluindo termos” na página 186.

Nota: Estas instruções mostram como fazer alterações dentro da visualização Editor de Recursos ou o Editor de Template . Lembre-se de que também é possível fazer esse tipo de ajuste diretamente a partir da área de janela Resultados da Extração , área de janela Dados, área de janela Categorias ou caixa de diálogo Definições de Cluster nas outras visualizações. Consulte o tópico “refinando resultados da extração” na página 86 para obter mais informações.

incluindo termos

A área de janela em árvore da biblioteca exibe bibliotecas e pode ser expandida para mostrar os dicionários de tipo que elas contêm. Na área de janela central, uma lista de termos exibe os termos na biblioteca selecionada ou no dicionário de tipo, dependendo da seleção na árvore.

No Editor de Recursos, é possível incluir termos em um dicionário de tipo diretamente na área de janela de termo ou por meio da caixa de diálogo Incluir Novos Termos. Os termos que você inclui podem ser palavras simples ou palavras compostas. Você sempre encontrará uma linha em branco na parte superior da lista para permitir a inclusão de um novo termo.

Nota: Estas instruções mostram como fazer alterações dentro da visualização Editor de Recursos ou o Editor de Template . Lembre-se de que também é possível fazer esse tipo de ajuste diretamente a partir da área de janela Resultados da Extração , área de janela Dados, área de janela Categorias ou caixa de diálogo Definições de Cluster nas outras visualizações. Consulte o tópico “refinando resultados da extração” na página 86 para obter mais informações.

Coluna de Termos

Nesta coluna, insira palavras simples ou compostas na célula. A cor em que o termo aparece depende da cor do tipo em que o termo está armazenado ou forçado. É possível mudar as cores de tipo na caixa de diálogo Propriedades de Tipo. Consulte o tópico “Criando tipos” na página 185 para obter informações adicionais.

Coluna Forçar

Nesta coluna, ao se colocar um ícone de tacha nesta célula, o mecanismo de extração sabe ignorar quaisquer outras ocorrências desse mesmo termo em outras bibliotecas. Consulte o tópico “Forçando termos” na página 188 para obter informações adicionais.

Coluna de Correspondência

Nesta coluna, selecione uma opção de correspondência para instruir o mecanismo de extração a fazer a correspondência deste termo com os dados de texto. Consulte a tabela para obter exemplos. É possível mudar o valor padrão editando as propriedades de tipo. Consulte o tópico “Criando tipos” na página 185 para obter informações adicionais. A partir dos menus, escolha **Editar> Alterar Match**. A seguir estão opções de correspondência básicas, já que suas combinações também são possíveis:

- **Início**. Se o termo no dicionário corresponder à primeira palavra em um conceito extraído do texto, esse tipo será designado. Por exemplo, se você inserir `apple`, `apple tart` será uma correspondência.
- **end**. Se o termo no dicionário corresponder à última palavra em um conceito extraído do texto, esse tipo será designado. Por exemplo, se você inserir `apple`, `cider apple` será uma correspondência.
- **Qualquer um**. Se o termo no dicionário corresponder a qualquer palavra de um conceito extraído do texto, esse tipo será designado. Por exemplo, se você inserir `apple`, a opção **Qualquer** irá digitar `apple tart`, `cider applee cider apple tart` da mesma forma.
- **Termo Inteiro**. Se o conceito inteiro extraído do texto corresponder ao termo exato no dicionário, esse tipo será designado. A inclusão de um termo como **Termo Inteiro**, **Inteiro e Início**, **Inteiro e Término**, **Inteiro e Qualquer um** ou **Inteiro (sem compostos)** forçará a extração de um termo.

Além disso, uma vez que o tipo <Person> extrai apenas dois nomes de peças, como *edith piaf* ou *mohandas gandhi*, você pode querer adicionar explicitamente os primeiros nomes a este dicionário do tipo se você estiver tentando extrair um primeiro nome quando nenhum sobrenome for mencionado. Por exemplo, se você quiser pegar todas as instâncias de *edith* como um nome, você deve adicionar `edith` ao tipo <Person> usando **Termo de pneu** ou **Entire e Iniciar**.

- **Inteiro (sem compostos)**. Se o conceito inteiro extraído do texto corresponder ao termo exato no dicionário, esse tipo será designado e a extração será interrompida para que não faça a correspondência do termo com um composto mais longo. Por exemplo, se você inserir `apple`, a opção **Inteiro (sem compostos)** tipificará `apple` e não extrairá o composto `apple sauce`, a menos que isso seja forçado em algum outro lugar.

Na tabela a seguir, suponha que o termo `apple` esteja em um dicionário de tipo. Dependendo da opção de correspondência, essa tabela mostra quais conceitos seriam extraídos e tipificados se fossem localizados no texto.

Tabela 40. Exemplos Correspondentes

Opções de correspondência para **o termo:**  **`apple`**	**Conceitos Extraídos**			
	`apple`	**`apple` *tart***	***ripe* `apple`**	***homemade*** **`apple` *tart***
Termo Inteiro	✔			
Início		✔		
Término			✔	
Início ou Término		✔	✔	
Inteiro e Início	✔	✔		
Inteiro e Término	✔		✔	
Inteiro e (Início ou Término)	✔	✔	✔	
qualquer um		✔	✔	✔
Inteiro e Qualquer um	✔	✔	✔	✔
Inteiro (sem compostos)	✔	*nunca extraído*	*nunca extraído*	*nunca extraído*

Coluna Flexionada

Nesta coluna, selecione se o mecanismo de extração deve gerar formas flexionadas para este termo durante a extração para que todas sejam agrupadas juntas. O valor padrão para esta coluna é definido em Propriedades de Tipo, mas é possível mudar essa opção caso por caso diretamente na coluna. A partir dos menus, escolha **Editar> Mudar Inflection**.

Coluna Tipo

Nesta coluna, selecione um dicionário de tipo na lista suspensa. A lista de tipos é filtrada de acordo com sua seleção na área de janela em árvore de biblioteca. O primeiro tipo na lista é sempre o tipo padrão

selecionado na área de janela em árvore de biblioteca. A partir dos menus, escolha **Editar> Tipo de Mudança**.

Coluna de Biblioteca

Nesta coluna, a biblioteca na qual seu termo está armazenado aparece. É possível arrastar e soltar um termo em outro tipo na área de janela em árvore de biblioteca para mudar sua biblioteca.

Para Incluir um Único Termo em um Dicionário de Tipo

1. Na área de janela em árvore de biblioteca, selecione o dicionário de tipo no qual deseja incluir o termo.
2. Na lista de termos na área de janela central, digite seu termo na primeira célula vazia disponível e configure quaisquer opções desejadas para este termo.

Para Incluir Vários Termos em um Dicionário de Tipo

1. Na área de janela em árvore de biblioteca, selecione o dicionário de tipo no qual deseja incluir os termos.
2. A partir dos menus, escolha **Ferramentas> Novos Termos**. A caixa de diálogo Incluir Novos Termos é aberta.
3. Insira os termos que deseja incluir no dicionário de tipo selecionado digitando-os ou copiando e colando um conjunto de termos. Se você inserir diversos termos, deve-se separá-los usando o delimitador definido no diálogo Opções ou incluir cada termo em uma nova linha. Veja o tópico "Configurando opções" na página 73 para obter mais informações.
4. Clique em **OK** para incluir os termos no dicionário. A opção de correspondência é configurada automaticamente para a opção padrão para essa biblioteca de tipos. A caixa de diálogo fecha e os novos termos aparecem no dicionário.

Forçando termos

Se desejar que um termo seja designado a um tipo específico, é possível incluí-lo no dicionário de tipo correspondente. Entretanto, se houver diversos termos com o mesmo nome, o mecanismo de extração deverá saber qual tipo deve ser usado. Portanto, será solicitado que você selecione qual tipo deve ser usado. Isso é chamado de *forçamento* de termo em um tipo. Esta opção é mais útil ao anular a atribuição do tipo a partir de um dicionário compilado (interno, não editável). Em geral, é recomendado evitar completamente termos duplicados.

O forçamento não *removerá* as outras ocorrências desse termo; em vez disso, elas serão ignoradas pelo mecanismo de extração. Depois, é possível mudar qual ocorrência deve ser usada forçando ou removendo o forçamento de um termo. Você também pode precisar forçar um termo em um dicionário de tipo ao incluir uma biblioteca pública ou atualizar uma.

É possível ver quais termos são forçados ou ignorados na coluna Forçar, a segunda na área de janela de termo. Se um ícone de tacha aparecer, isso significa que essa ocorrência do termo foi forçada. Se um ícone de X preto aparecer, isso significa que essa ocorrência do termo será ignorada durante a extração porque ela foi forçada em outras partes. Além disso, quando você força um termo, ele aparece colorido para o tipo no qual foi forçado. Isto significa que se você forçou um termo que está em ambos Type 1 e Type 2 em Type 1, qualquer hora que você vir este termo na janela, ele aparecerá na cor da fonte definida para Type 1.

É possível dar um clique duplo no ícone para mudar o status. Se o termo aparecer em algum outro lugar, uma caixa de diálogo Resolver Conflitos será aberta para permitir que você selecione qual ocorrência deve ser usada.

Renomeando tipos

É possível renomear um dicionário de tipo ou mudar outras configurações do dicionário editando as propriedades de tipo.

Importante: É recomendado não usar espaços nos nomes de tipo, principalmente se dois ou mais nomes de tipo começarem com a mesma palavra. Também recomendamos não renomear os tipos nas bibliotecas Núcleo ou Pareceres ou mudar seus atributos de correspondência padrão.

Para Renomear um Tipo

1. Na área de janela em árvore de biblioteca, selecione o dicionário de tipo que deseja renomear.
2. Clique com o botão direito do seu mouse e escolha **Propriedade de Tipo** no menu de contexto. A caixa de diálogo Propriedades de Tipo é aberta.
3. Insira o novo nome para seu dicionário de tipo na caixa de texto Nome.
4. Clique em **OK** para aceitar o novo nome. O novo nome do tipo fica visível na área de janela em árvore de biblioteca.

Movendo tipos

É possível arrastar um dicionário de tipo para outra localização dentro de uma biblioteca ou para outra biblioteca na árvore.

Para Reordenar um Tipo dentro de uma Biblioteca

1. Na área de janela em árvore de biblioteca, selecione o dicionário de tipo que deseja mover.
2. A partir dos menus, escolha **Editar> Mover para cima** para mover o dicionário do tipo para cima de uma posição no painel da árvore da biblioteca ou **Editar> Mover para baixo** para motá-lo para baixo uma posição.

Para Mover um Tipo para outra Biblioteca

1. Na área de janela em árvore de biblioteca, selecione o dicionário de tipo que deseja mover.
2. Clique com o botão direito do seu mouse e escolha **Propriedade de Tipo** no menu de contexto. A caixa de diálogo Propriedades de Tipo é aberta. (Também é possível arrastar e soltar o tipo em outra biblioteca).
3. Na caixa de listagem Incluir em, selecione a biblioteca para a qual deseja mover o dicionário de tipo.
4. Clique em **OK**. A caixa de diálogo é fechada e agora o tipo está na biblioteca que você selecionou.

Desativando e excluindo tipos

Se desejar remover provisoriamente um dicionário de tipo, é possível desativá-lo cancelando a seleção da caixa à esquerda do nome do dicionário na área de janela em árvore de biblioteca. Isso sinaliza que você deseja manter o dicionário em sua biblioteca, mas quer que seu conteúdo seja ignorado durante a verificação de conflito e durante o processo de extração.

Também é possível excluir permanentemente os dicionários de tipo de uma biblioteca.

Para Desativar um Dicionário de Tipo

1. Na área de janela em árvore de biblioteca, selecione o dicionário de tipo que deseja desativar.
2. Clique na barra de espaço. A caixa de seleção à esquerda do nome do tipo é limpa.

Para Excluir um Dicionário de Tipo

1. Na área de janela em árvore de biblioteca, selecione o dicionário de tipo que deseja excluir.
2. A partir dos menus, escolha **Editar> Excluir** para excluir o dicionário do tipo.

Dicionários de substituição/sinônimo

Um *dicionário de substituição* é uma coleção de termos que ajudam a agrupar termos similares sob um termo de resposta. Dicionários de substituição são gerenciados na área de janela inferior da guia Recursos de Biblioteca. Você pode acessar esta visualização com **Visualizar> Editor de Recursos** nos menus, se você estiver em uma sessão interativa de ambiente de trabalho. Caso contrário, é possível editar dicionários para um modelo específico no Editor de Template.

É possível definir duas formas de substituições nesse dicionário: *sinônimos* e *elementos opcionais*. É possível clicar nas guias nessa área de janela para se alternar entre elas.

Após executar uma extração em seus dados de texto, talvez você encontre vários conceitos que são sinônimos ou formas flexionadas de outros conceitos. Com a identificação de elementos opcionais e sinônimos, é possível forçar o mecanismo de extração a mapeá-los para um único termo de resposta.

A substituição usando sinônimos e elementos opcionais reduz o número de conceitos na área de janela Resultados da Extração combinando-os em conceitos representantes mais significativos com uma maior frequência de Doc. contagens.

Sinônimos

Os sinônimos associam duas ou mais palavras que têm o mesmo significado. Também é possível usar sinônimos para agrupar termos com suas abreviações ou para agrupar palavras comumente digitadas incorretamente com a ortografia correta. É possível definir esses sinônimos na guia Sinônimos.

Uma definição de sinônimo é composta por duas partes. A primeira é um termo **Resposta**, que é o termo sob o qual você deseja que o mecanismo de extração agrupe todos os termos sinônimos. A menos que esse termo de resposta seja usado como um sinônimo de outro termo de resposta ou seja excluído, provavelmente ele se tornará o conceito que aparece na área de janela Resultados da Extração. A segunda é a lista de sinônimos que serão agrupados sob o termo de resposta.

Por exemplo, se você deseja que `automobile` seja substituído por `vehicle`, então `automobile` é o sinônimo e `vehicle` é o termo de destino.

Você pode inserir quaisquer palavras na coluna **Synonym** , mas se a palavra não for encontrada durante a extração e o termo tivesse uma opção de correspondência com `Entire`, então nenhuma substituição pode ocorrer. No entanto, o termo de resposta não precisa ser extraído para os sinônimos serem agrupados sob esse termo.

Elementos Opcionais

Elementos opcionais identificam palavras opcionais em um termo composto que pode ser ignorado durante a extração para manter juntos os termos semelhantes, mesmo que eles apareçam um pouco diferente no texto. Elementos opcionais são palavras simples que, se fossem removidas de uma palavra composta, poderiam criar uma correspondência com outro termo. Essas palavras simples podem aparecer em qualquer lugar dentro da composta -- no início, no meio ou no fim. É possível definir elementos opcionais na guia Opcional.

Por exemplo, para agrupar os termos `ibm` e `ibm corp` juntos, você deve declarar `corp` para ser tratado como um elemento opcional neste caso. Em outro exemplo, se você designar o termo `access` para ser um elemento opcional e durante a extração ambos `internet access speed` e `internet speed` forem encontrados, eles serão agrupados sob o termo que ocorre com mais frequência.

Definindo sinônimos

Na guia Sinônimos, você pode inserir uma definição de sinônimo na linha vazia na parte superior da tabela. Comece por definir o termo de destino e seus sinônimos. Também é possível selecionar a biblioteca na qual você gostaria de armazenar essa definição. Durante a extração, todas as ocorrências de sinônimos serão agrupadas sob o termo de destino na extração final. Consulte o tópico “incluindo termos” na página 186 para obter informações adicionais.

Por exemplo, se os seus dados de texto incluem muitas informações de telecomunicações, você pode ter esses termos: `cellular phone`, `wireless phone`e `mobile phone`. Neste exemplo, talvez você queira definir `cellular` e `mobile` como sinônimos de `wireless`. Se você definir esses sinônimos, cada ocorrência extraída de `cellular phone` e `mobile phone` será tratada como o mesmo termo que `wireless phone` e aparecerá junto na lista de termos.

Quando estiver construindo seu dicionário de tipos, você pode inserir um termo e pensar em três ou quatro sinônimos para ele. Nesse caso, você poderia inserir todos os termos e, então, seu termos de destino no dicionários de substituições e, em seguida, arrastar os sinônimos.

A substituição de sinônimo também é aplicada em formas flexionadas (como plural) do sinônimo. Dependendo do contexto, talvez você queira impor restrições em como os termos são substituídos. Certos caracteres podem ser usados para impor limites no modo como o processamento de sinônimo deve acontecer:

- **marca de exclamação (!).** Quando a marca de exclamação precede diretamente o sinônimo `!synonym`, isto indica que nenhuma forma inflecionada do sinônimo será substituída pelo termo de destino. No entanto, um ponto de exclamação precedendo diretamente o termo de resposta `!target-term` significa que você não quer que nenhuma parte do termo de resposta composto ou variantes recebam substituições adicionais.
- **Asterisco (*).** Um asterisco colocado diretamente após um sinônimo, tal como `synonym*`, significa que você deseja que esta palavra seja substituída pelo termo de destino. Por exemplo, se você definiu `manage*` como sinônimo e `management` como resposta, então `associate managers` será substituído pelo termo de resposta `associate management`. Também é possível incluir um espaço e um asterisco após a palavra (`synonym *`), como `internet *`. Se você definisse a resposta como `internet` e os sinônimos como `internet * *` e `web *`, então `internet access card` e `web portal` seriam substituídos por `internet`. Não é possível iniciar uma palavra ou uma sequência de caracteres com o curinga asterisco nesse dicionário.
- **Caret (^).** Um careta e um espaço que precede o sinônimo, como `^ synonym`, significa que o agrupamento sinônimo aplica-se apenas quando o termo inicia com o sinônimo. Por exemplo, se você definir `^ wage` como o sinônimo e `income` como a resposta e ambos os termos forem extraídos, eles serão agrupados juntos sob o termo `income`. No entanto, se `minimum wage` e `income` forem extraídos, eles não serão agrupados, uma vez que `minimum wage` não começa com `wage`. Um espaço deve ser colocado entre esse símbolo e o sinônimo.
- **Sinal de dólar ($).** Um espaço e um sinal de dólar seguindo o sinônimo, como `synonym $`, significa que o agrupamento de sinônimo se aplica apenas quando o termo termina com o sinônimo. Por exemplo, se você definir `cash $` como o sinônimo e `money` como a resposta e ambos os termos forem extraídos, eles serão agrupados juntos sob o termo `money`. No entanto, se `cash cow` e `money` forem extraídos, eles não serão agrupados, uma vez que `cash cow` não termina com `cash`. Um espaço deve ser colocado entre esse símbolo e o sinônimo.
- **Careta (^) e cifrão de dólar ($).** Se o sinal de careta e dólar forem usados juntos, como `^ synonym $`, um termo corresponde ao sinônimo apenas se for uma correspondência exata. Isso significa que nenhuma palavra pode aparecer antes ou depois do sinônimo no termo extraído para que o agrupamento de sinônimos aconteça. Por exemplo, você pode querer definir `^ van $` como o sinônimo e `truck` como a resposta para que somente `van` seja agrupado com `truck`, enquanto `marie van guerin` permanecerá inalterado. Além disso, sempre que você definir um sinônimo usando o acento circunflexo e o símbolo de dólar e essa palavra aparecer em qualquer lugar no texto de origem, o sinônimo será extraído automaticamente.

Para Incluir uma Entrada de Sinônimo

1. Com a área de janela de substituição exibida, clique na guia **Sinônimos** no canto inferior esquerdo.
2. Na linha vazia na parte superior da tabela, insira seu termo de resposta na coluna Resposta. O termo de destino inserido aparece colorido. Essa cor representa o tipo no qual o termo aparece ou é forçado, se for o caso. Se o termo aparecer em preto, isso significa que ele não aparecerá em nenhum dicionário de tipo.

3. Clique na segunda célula à direita do alvo e digite o conjunto de sinônimos. Separe cada entrada usando o delimitador global, conforme definido na caixa de diálogo Opções. Veja o tópico "Configurando opções" na página 73 para obter mais informações. Os termos inseridos aparecem coloridos. Essa cor representa o tipo no qual o termo aparece. Se o termo aparecer em preto, isso significa que ele não aparecerá em nenhum dicionário de tipo.
4. Clique na última célula para selecionar a biblioteca na qual deseja-se armazenar esta definição de sinônimo.

Nota: Estas instruções mostram como fazer alterações dentro da visualização Editor de Recursos ou o Editor de Template . Lembre-se de que também é possível fazer esse tipo de ajuste diretamente a partir da área de janela Resultados da Extração , área de janela Dados, área de janela Categorias ou caixa de diálogo Definições de Cluster nas outras visualizações. Consulte o tópico "refinando resultados da extração" na página 86 para obter mais informações.

Definindo elementos opcionais

Na guia Opcional, é possível definir elementos opcionais para a biblioteca desejada. Essas entradas são agrupadas juntas para cada biblioteca. Assim que uma biblioteca é incluída na área de janela em árvore de biblioteca, uma linha de elemento opcional vazia é incluída na guia Opcional.

Todas as entradas são automaticamente transformadas em palavras em letras minúsculas. O mecanismo de extração fará a correspondência de todas as entradas com as palavras em letras minúsculas e maiúsculas no texto.

Nota: Os termos são delimitados usando o delimitador definido no diálogo Opções. Veja o tópico "Configurando opções" na página 73 para obter mais informações. Se o elemento opcional sendo inserido incluir o mesmo delimitador como parte do termo, uma barra invertida deverá precedê-lo.

Para Incluir uma Entrada

1. Com a área de janela de substituição exibida, clique na guia Opcional no canto inferior esquerdo do editor.
2. Clique na célula na coluna Elementos Opcionais para a biblioteca na qual deseja incluir essa entrada.
3. Insira o elemento opcional. Separe cada entrada usando o delimitador global, conforme definido na caixa de diálogo Opções. Veja o tópico "Configurando opções" na página 73 para obter mais informações.

Desativando e excluindo substituições

É possível remover provisoriamente uma entrada desativando-a em seu dicionário. Quando você desativa uma entrada, ela é ignorada durante a extração.

Também é possível excluir quaisquer entradas obsoletas em seu dicionário de substituição.

Para Desativar uma Entrada

1. Em seu dicionário, selecione a entrada que deseja desativar.
2. Clique na barra de espaço. A caixa de seleção à esquerda da entrada é limpa.

Nota: Também é possível cancelar a seleção da caixa de seleção à esquerda da entrada para desativá-la.

Para Excluir uma Entrada de Sinônimo

1. Em seu dicionário, selecione a entrada que deseja excluir.
2. A partir dos menus, escolha **Editar> Excluir** ou pressione a tecla **Delete** em seu teclado. A entrada não está mais no dicionário.

Para Excluir uma Entrada de Elemento Opcional

1. Em seu dicionário, dê um clique duplo na entrada que deseja excluir.
2. Exclua o termo manualmente.

3. Pressione Enter para aplicar a mudança.

Dicionários de exclusão

Um *dicionário de exclusão* é uma lista de palavras, frases ou sequências de caracteres parciais. Quaisquer termos correspondentes a ou contendo uma entrada no dicionário de exclusão serão ignorados ou excluídos da extração. Dicionários de exclusão são gerenciados na área de janela direita do editor. Normalmente os termos que você inclui nessa lista são frases ou palavras preenchidas que são usadas no texto para continuidade, mas que, de fato, não incluem nada de importante no texto e podem tumultuar os resultados da extração. Ao incluir esses termos no dicionário de exclusão, é possível se certificar de que eles nunca serão extraídos.

Dicionários de exclusão são gerenciados na área de janela superior direita da guia Recursos de Biblioteca no editor. Você pode acessar esta visualização com **Visualizar> Editor de Recursos** nos menus, se você estiver em uma sessão interativa de ambiente de trabalho. Caso contrário, é possível editar dicionários para um modelo específico no Editor de Template.

No dicionário de exclusão, é possível inserir uma palavra, uma frase ou uma sequência parcial na linha vazia na parte superior da tabela. É possível incluir sequências de caracteres em seu dicionário de exclusão como uma ou mais palavras, ou até mesmo palavras parciais, usando o asterisco como um curinga. As entradas declaradas no dicionário de exclusão serão usadas para barrar a extração de conceitos. Se uma entrada também for declarada em algum outro lugar na interface, como em um dicionário de tipo, ela será mostrada com um tachado nos outros dicionários, indicando que está atualmente excluída. Essa sequência de caracteres não precisa aparecer nos dados de texto ou ser declarada como parte de qualquer dicionário de tipo a ser aplicado.

Nota: Se você adicionar um conceito ao dicionário de exclusão que também atua como destino em uma entrada de sinônimo, então o alvo e todos os seus sinônimos também serão excluídos. Consulte o tópico “Definindo sinônimos” na página 190 para obter informações adicionais.

Usando Curingas (*)

pode usar o curinga asterisco para denotar que deseja que a entrada de exclusão seja tratada como uma sequência parcial. Quaisquer termos localizados pelo mecanismo de extração contendo uma palavra que comece ou termine com uma sequência de caracteres inserida no dicionário de exclusão serão excluídos da extração final. No entanto, há dois casos em que o uso do curinga não é permitido:

- Caractere hífen (-) precedido por um curinga asterisco, como *-
- Apóstrofo (') precedido por um curinga asterisco, como *'s

Tabela 41. Exemplos de entradas de exclusão

Entrada	Exemplo	Resultados
palavra	*Próximo*	Nenhum conceito (ou seus termos) será extraído se contiver a palavra *próximo*.
frase	*por exemplo*	Nenhum conceito (ou seus termos) será extraído se contiver a frase *por exemplo*.
parcial	*copyright*	Excluirá quaisquer conceitos (ou seus termos) correspondentes a ou contendo as variações da palavra *copyright*, como `copyrighted`, `copyrighting`, `copyrights` ou `copyright 2010`.
parcial	*ware	Excluirá quaisquer conceitos (ou seus termos) correspondentes a ou contendo as variações da palavra *ware*, como `freeware`, `shareware`, `software`, `hardware`, beware ou `silverware`.

Para Incluir Entradas

- Na linha vazia na parte superior da tabela, insira um termo. O termo inserido aparece colorido. Essa cor representa o tipo no qual o termo aparece. Se o termo aparecer em preto, isso significa que ele não aparecerá em nenhum dicionário de tipo.

Para Desativar Entradas

É possível remover provisoriamente uma entrada desativando-a no dicionário de exclusão. Quando você desativa uma entrada, ela é ignorada durante a extração.

1. Em seu dicionário de exclusão, selecione a entrada que deseja desativar.
2. Clique na barra de espaço. A caixa de seleção à esquerda da entrada é limpa.

Nota: Você também pode desmarcar a caixa de seleção à esquerda da entrada para desativá-lo.

Para Excluir Entradas

É possível excluir quaisquer entradas desnecessárias em seu dicionário de exclusão.

1. Em seu dicionário de exclusão, selecione a entrada que deseja excluir.
2. A partir dos menus, escolha **Editar> Excluir**. A entrada não está mais no dicionário.

Capítulo 17. Sobre recursos avançados

Além do tipo, dos dicionários de exclusão e substituição, também é possível trabalhar com uma variedade de configurações de recursos avançados, como configurações de Agrupamento Difuso ou definições de tipo não linguístico. É possível trabalhar com esses recursos na guia Recursos Avançados na visualização Editor de Template ou Editor de Recursos.

Ao ir para a guia Recursos Avançado, é possível editar as informações a seguir:

- **Idioma de destino para recursos.** Usado para selecionar o idioma para o qual os recursos serão criados ou ajustados. Veja o tópico "Idioma de destino para recursos" na página 197 para obter mais informações.
- **Agrupamento Difuso (Exceções).** Usado para excluir pares palavras a partir do algoritmo de agrupamento difuso (correção de erro de ortografia). Veja o tópico "Agrupamento difuso" na página 197 para obter mais informações.
- **Entidades não Linguísticas.** Usadas para ativar e desativar as entidades não linguísticas que podem ser extraídas, bem como as expressões regulares e as regras de normalização, aplicadas durante a sua extração. Veja o tópico "Entidades não linguísticas" na página 198 para obter mais informações.
- **Tratamento de Idioma.** Usado para declarar as formas especiais de estruturar sentenças (padrões de extração e definições forçadas) e usar abreviações para o idioma selecionado. Veja o tópico "Manipulação de idioma" na página 202 para obter mais informações.

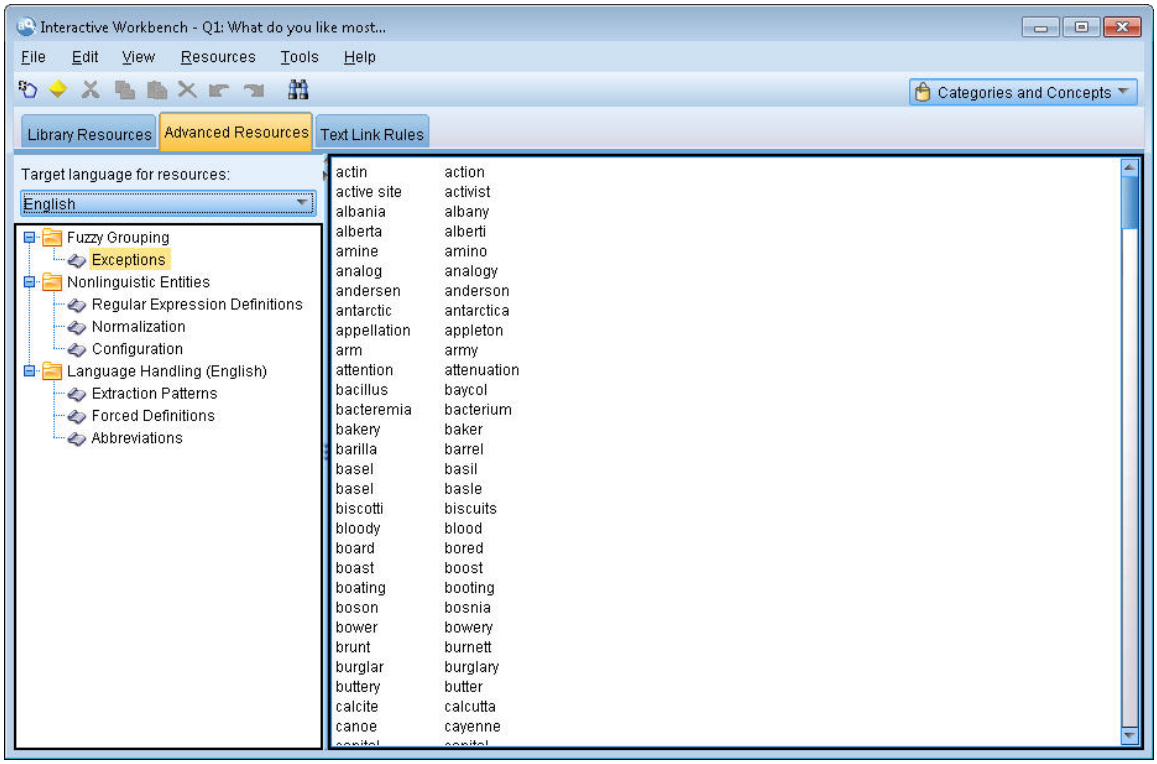


Figura 43. Editor de modelo de mineração de texto - guia recursos avançados

Nota: Você pode usar a barra de ferramentas Localizar / Substituir para encontrar informações rapidamente ou para fazer alterações uniformes em uma seção. Para obter mais informações, consulte "Substituição" na página 196.

Para Editar Recursos Avançados

1. Localize e selecione a seção de recursos que deseja editar. O conteúdo aparece na área de janela à direita.
2. Use os botões de menu ou da barra de ferramentas para recortar, copiar ou colar o conteúdo, se necessário.
3. Edite o(s) arquivo(s) que você deseja alterar usando as regras de formatação nesta seção. Suas mudanças serão salvas assim que forem feitas. Utilize as setas de desfazer ou refazer na barra de ferramentas para reverter para as mudanças anteriores.

Descoberta

Em alguns casos, talvez seja necessário localizar informações rapidamente em uma determinada seção. Por exemplo, se executar uma análise de ligação de texto, você pode ter centenas de macros e definições de padrão. Usando a variável Localizar, é possível localizar uma regra específica rapidamente. Para procurar informações em uma seção, é possível usar a barra de ferramentas Localizar.

Para Usar a Variável Localizar

1. Localize e selecione a recurso do recurso que deseja procurar. O conteúdo aparece na área de janela direita do editor.
2. A partir dos menus, escolha **Editar> Localizar**. A barra de ferramentas Localizar aparece no lado superior direito da caixa de diálogo Editar Recursos Avançados.
3. Insira a sequência de palavras que deseja procurar na caixa de texto. É possível usar os botões da barra de ferramentas para controlar a distinção entre maiúsculas e minúsculas, correspondência parcial e direção da procura.
4. Clique em **Localizar** para iniciar a procura. Se uma correspondência for localizada, o texto será destacado na janela.
5. Clique em **Localizar** novamente para procurar a próxima correspondência.

Nota: Ao trabalhar na aba Regras de Link de Texto, a opção Localizar só está disponível quando você visualizar o código-fonte.

Substituição

Em alguns casos, talvez seja necessário fazer atualizações mais abrangentes para seus recursos avançados. A variável Substituir pode ajudá-lo a fazer atualizações uniformes em seu conteúdo.

Para Usar a Variável Substituir

1. Localize e selecione a seção de recurso em que deseja procurar e substituir. O conteúdo aparece na área de janela direita do editor.
2. A partir dos menus, escolha **Editar> Substituto**. A caixa de diálogo Substituir é aberta.
3. Na caixa de texto **Localizar o que**, insira a sequência de palavras que deseja procurar.
4. Na caixa de texto **Substituir por**, insira a sequência que deseja usar no lugar do texto localizado.
5. Selecione **Apenas correspondência de palavras inteiras** se desejar localizar somente palavras completas.
6. Selecione **Correspondência de maiúsculas e minúsculas** se desejar localizar ou substituir somente palavras que correspondam exatamente às maiúsculas e minúsculas.
7. Clique em **Localizar Próximo** para localizar uma correspondência. Se uma correspondência for localizada, o texto será destacado na janela. Se não desejar substituir essa correspondência, clique em **Localizar Próximo** novamente até localizar uma correspondência que deseje substituir.
8. Clique em **Substituir** para substituir a correspondência selecionada.

9. Clique em **Substituir** para substituir todas as correspondências na seção. Uma mensagem será aberta com o número de substituições feitas.
10. Quando tiver concluído suas substituições, clique em **Fechar**. A caixa de diálogo é fechada.

Nota: Se você fez um erro de substituição, você pode desfazer a substituição fechando a caixa de diálogo e escolhendo **Editar> Undo** dos menus. Deve-se executar essa ação uma vez para cada mudança que você deseja desfazer.

Idioma de destino para recursos

Recursos são criados para um determinado idioma do texto. O idioma para o qual esses recursos são ajustados é definido na guia Recursos Avançados. É possível alternar para outro idioma se for necessário selecionando esse idioma na caixa de combinação **Idioma de destino para recursos**. Além disso, o idioma listado aqui aparecerá como aquele para qualquer pacote de análise de texto que você criar com esses recursos.

Importante: Raramente você precisará mudar o idioma em seus recursos. Isso pode causar problemas quando seus recursos não corresponderem mais ao idioma de extração. Embora raramente empregado, você pode mudar um idioma se planejou usar a opção de idioma ALL durante a extração porque esperava ter um texto em mais de um idioma. Ao mudar o idioma, você pode acessar, por exemplo, os recursos de tratamento de idioma para padrões de extração, abreviações e forçar definições para o idioma secundário no qual está interessado. No entanto, antes de publicar ou salvar as mudanças de recurso feitas ou executar outra extração, lembre-se configurar o idioma de volta para o idioma principal que está interessado em extrair.

Agrupamento difuso

No nó Mineração de Texto e em Configurações de Extração, se você selecionar **Acomodar ortografia para o limite mínimo de caractere raiz de**, você ativou o algoritmo de agrupamento difuso.

Agrupamentos difusos ajudam a agrupar palavras mal escritas com frequência ou palavras com ortografia semelhante removendo todas as vogais provisoriamente (exceto a primeira) e consoantes duplas ou triplas das palavras extraídas e comparando-as para ver se elas são iguais. Durante o processo de extração, a variável de agrupamento difuso é aplicada aos termos extraídos e os resultados são comparados para determinar se alguma correspondência foi localizada. Se sim, os termos originais serão agrupados na lista de extração final. Eles serão agrupados sob o termo que ocorre com mais frequência nos dados.

Nota: Se os dois termos sendo comparados são atribuídos a tipos diferentes, excluindo o tipo <Unknown> , então a técnica de agrupamento fuzzy não será aplicada a este par. Em outras palavras, os termos devem pertencer ao mesmo tipo ou ao tipo <Unknown> a fim de que a técnica seja aplicada.

Se você ativou essa variável e descobriu que duas palavras com ortografia semelhante foram agrupadas incorretamente, você pode querer excluí-las do agrupamento difuso. Isso pode ser feito inserindo os pares correspondidos incorretamente na seção Exceções na guia Recursos Avançados. Consulte o tópico Capítulo 17, “Sobre recursos avançados”, na página 195 para obter informações adicionais.

O exemplo a seguir demonstra como um agrupamento difuso é executado. Se o agrupamento difuso estiver ativado, estas palavras parecerão as mesmas e serão correspondidas da seguinte maneira:

```
color  -> colr                mountain -> montn
colour -> colr                montana  -> montn

modeling  -> modlng           furniture -> furntr
modelling -> modlng           furnature -> furntr
```

No exemplo anterior, provavelmente você gostaria de excluir `mountain` e `montana` para não serem agrupadas. Portanto, é possível inseri-las na seção Exceções da seguinte forma:

```
mountain      montana
```

Importante: Em alguns casos, as exceções de agrupamento difuso não impedem que duas palavras sejam emparelhadas porque certas regras de sinônimo estão sendo aplicadas. Nesse caso, você pode querer tentar inserir sinônimos usando o curinga de marca de exclamação (!) para proibir as palavras de se tornarem sinônimos na saída. Para obter mais informações, consulte "Definindo sinônimos" na página 190.

Regras de formatação para exceções de agrupamento difuso

- Defina somente um par de exceções por linha.
- Use palavras simples ou compostas.
- Use somente caracteres minúsculos para as palavras. Palavras maiúsculas serão ignoradas.
- Use um caractere TAB para separar cada palavra em um par.

Entidades não linguísticas

Ao trabalhar com determinados tipos de dados, talvez você esteja interessado em extrair datas, números de seguridade social, porcentagens ou outras entidades não linguísticas. Essas entidades são declaradas explicitamente no arquivo de configuração no qual é possível ativar ou desativar as entidades. Consulte o tópico "Configuração" na página 201 para obter informações adicionais. Para otimizar a saída a partir do mecanismo de extração, a entrada do processamento não linguístico é normalizada para agrupar entidades semelhantes de acordo com os formatos predefinidos. Veja o tópico "Normalização" na página 201 para obter mais informações.

Nota: Você pode ligar e desligar a extração de entidade não linguística nas configurações de extração.

Entidades Não Linguísticas Disponíveis

As entidades não linguísticas na tabela a seguir podem ser extraídas. O nome do tipo está entre parênteses.

Tabela 42. Entidades não linguísticas que podem ser extraídas

Endereços	(<Address>)
Aminoácidos	(<Aminoacid>)
Moedas	(<Currency>)
Datas	(<Date>)
Atraso	(<Delay>)
Dígitos	(<Digit>)
Endereços de E-mail	(<email>)
Endereços HTTP/URL	(<url>)
Endereço IP	(<IP>
Organizações	(<Organization>)
Porcentagens	(<Percent>)
Produtos	(<Product>)
Proteínas	(<Gene>)
Números de telefone	(<PhoneNumber>)
Tempos	(<Time>)
Número de seguridade social dos EUA	(<SocialSecurityNumber>)

Tabela 42. Entidades não linguísticas que podem ser extraídas (continuação)	
Pesos e medidas	(<Weights-Measures>)

Limpando Texto para Processamento

Antes da ocorrência da extração de entidades não linguísticas, o texto de entrada é limpo. Durante esse passo, as mudanças provisórias a seguir são feitas para que entidades não linguísticas possam ser identificadas e extraídas como tais:

- Qualquer sequência de dois ou mais espaços é substituída por um único espaço.
- Tabulações são substituídas por espaço.
- Sequências de caracteres ou caracteres únicos de fim de linha são substituídos por um espaço, enquanto sequências múltiplas de fim de linha são marcadas como o fim de um parágrafo. O fim de linha pode ser denotado por retornos de linha (CR) e feed de linha (LF) ou ambos juntos.
- Tags HTML e XML são removidas provisoriamente e ignoradas.

Definições de expressão regular

Durante a extração de entidades não linguísticas, talvez você queira editar ou incluir em definições de expressão regular que são usadas para identificar expressões regulares. Isso é feito na seção **Definições de Expressão Regular** na guia Recursos Avançados. Veja o tópico Capítulo 17, “Sobre recursos avançados”, na página 195 para obter mais informações.

O arquivo é divido em seções distintas. A primeira seção é chamada `[macros]`. Além dessa seção, pode existir uma seção adicional para cada entidade não linguística. É possível incluir seções nesse arquivo. Dentro de cada seção, regras são numeradas (*regexp1*, *regexp2*, e assim por diante). Essas regras devem ser numeradas sequencialmente de 1–*n*. Qualquer divisão na numeração fará com que o processamento desse arquivo seja completamente suspenso.

Em certos casos, uma entidade é dependente de idioma. Uma entidade é considerada dependente de idioma se usar um valor diferente de 0 para o parâmetro de idioma no arquivo de configuração. Consulte o tópico “Configuração” na página 201 para obter informações adicionais. Quando uma entidade é dependente de idioma, o idioma deve ser usado para prefixar o nome da seção, como `[english/PhoneNumber]`. Essa seção conteria regras que se aplicam somente aos números de telefone em inglês quando a entidade `PhoneNumber` recebesse um valor de 2 para o idioma.

Importante! Se você fizer mudanças nesse arquivo ou em qualquer outro no editor e o mecanismo de extração não funcionar mais como desejado, use a opção **Reconfigurar para Original** na barra de ferramentas para reconfigurar o arquivo para o conteúdo original fornecido. Esse arquivo requer certo nível de familiaridade com expressões regulares. Se você requerer assistência adicional nessa área, entre em contato com IBM Corp. para obter ajuda.

Caracteres Especiais. [] {} () \ * +? | ^ $

Todos os caracteres combinam com eles mesmos, exceto para os seguintes caracteres especiais, que são usados para um propósito específico em expressões: `.[{()\*+?|^$` Para usar esses caracteres como tal, eles devem ser precedidos por uma barra inverter (\) na definição.

Por exemplo, se você estiver tentando extrair endereços da web, o caractere de parada total é muito importante para a entidade, portanto, deve-se incluir uma barra invertida nele da seguinte forma:

```
www\.[a-z]+\.[a-z]+
```

Repetições Operadores e Quantificadores? + * {}

Para permitir que as definições sejam mais flexíveis, é possível usar diversos curingas que são padrão para expressões regulares. They are `* ? +`

- *Asterisco* * indica que há *zero ou mais* das sequências de caracteres precedentes. Por exemplo, ab*c corresponde a "*ac*", "*abc*", "*abbbc*", e assim por diante.

- *Sinal de mais +* indica que há *uma ou mais* das sequências de caracteres precedentes. Por exemplo, `ab+c` corresponde a "*abc*", "*abbc*", "*abbbc*", mas não a "*ac*".
- *Ponto de interrogação ?* indica que há *zero ou uma* das sequências de caracteres precedentes. Por exemplo, `modell?ing` corresponde a "*modeling*" e a "*modeling*".
- *Limitar repetição com colchetes {}* indica os limites da repetição. Por exemplo,

  `[0-9]{n}` corresponde a um dígito repetido exatamente *n* vezes. Por exemplo, `[0-9]{4}` corresponderá a "*1998*", mas não a "*33*" ou "*19983*".

  `[0-9]{n,}` corresponde a um dígito repetido *n ou mais* vezes. Por exemplo, `[0-9]{3,}` corresponderá a "*199*" ou "*1998*", mas não a "*19*".

  `[0-9]{n,m}` corresponde a um dígito repetido entre *n e m vezes inclusivas*. Por exemplo, `[0-9]{3,5}` corresponderá "*199*", "*1998*" ou "*19983*", mas não "*19*" nem "*199835*".

Espaços e Hifens Opcionais

Em alguns casos, você deseja incluir um espaço opcional em uma definição. Por exemplo, se você quisesse extrair moedas como "*pesos uruguaios*", "*peso uruguaio*", "*pesos do Uruguai*", "*peso do Uruguai*", "*pesos*" ou "*peso*", seria necessário lidar com o fato de que pode haver duas palavras separadas por um espaço. Nesse caso, essa definição deve ser escrita como `(uruguayan |uruguay )?pesos?`. Uma vez que *uruguayan* ou *uruguay* são seguidos por um espaço quando usado com *pesos*/*peso*, o espaço opcional deve ser definido dentro da sequência opcional `(uruguayan |uruguay )`. Se o espaço não estivesse na sequência opcional como `(uruguayan|uruguay)? pesos?`, ele não corresponderia em "*pesos*" ou "*peso*", uma vez que o espaço seria necessário.

Se você estiver procurando uma série de coisas incluindo caracteres hífen (-) em uma lista, o hífen deverá ser definido por último. Por exemplo, se você estiver procurando uma vírgula (,) ou um hífen (-), use `[,-]` e nunca `[-,]`.

Ordem das Sequências de Caracteres em Listas e Macros

É sempre necessário definir a sequência mais longa antes de uma mais curta ou a mais longa nunca será lida, já que a correspondência ocorrerá na mais curta. Por exemplo, se você estivesse procurando as sequências de caracteres "*bilhão*" ou "*bil*", então "*bilhão*" deveria ser definido antes de "*bil*". Portanto, por exemplo `(billion|bill)` e não `(bill|billion)`. Isso também se aplica a macros, já que macros são listas de sequências de caracteres.

Ordem de Regras na Seção de Definição

Defina uma regra por linha. Dentro de cada seção, regras são numeradas (*regexp1*, *regexp2*, e assim por diante). Essas regras devem ser numeradas sequencialmente de 1–*n*. Qualquer divisão na numeração fará com que o processamento desse arquivo seja completamente suspenso. Para desativar uma entrada, coloque um símbolo # no início de cada linha usada para definir a expressão regular. Para ativar uma entrada, remova o caractere # antes dessa linha.

Em cada seção, as regras mais específicas devem ser definidas antes das mais gerais para assegurar o processamento adequado. Por exemplo, se você estivesse procurando uma data no formato "*mês ano*" e no formato "*mês*", então a regra "*mês ano*" deveria ser definida antes da regra "*mês*". Veja aqui como isso deveria ser definido:

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

e não

```
#@# January
regexp1=$(MONTH)
```

```
     #@# January 1932
     regexp2=$(MONTH),? [0-9]{4}
```

Usando Macros em Regras

Sempre que uma sequência específica é usada em várias regras, é possível usar uma macro. Então, se você precisar mudar a definição dessa sequência, será necessário mudá-la somente uma vez, e não em todas as regras que se referem a ela. Por exemplo, supondo que você tivesse a seguinte macro:

```
     MONTH=((january|february|march|april|june|july|august|september|october|
     november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)
(\.)?)
```

Sempre que você se referir ao nome da macro, ele deverá ser colocado entre `$()`, como: `regexp1=$(MONTH)`

Todas as macros devem ser definidas na seção `[macros]`.

Normalização

Durante a extração de entidades não linguísticas, as entidades encontradas são normalizadas para grupos como entidades de acordo com formatos predefinidos. Por exemplo, símbolos monetários e seus equivalentes em palavras são tratados como a mesma coisa. As entradas de normalização são armazenadas na seção **Normalização** na guia Recursos Avançados. Veja o tópico Capítulo 17, “Sobre recursos avançados”, na página 195 para obter mais informações. O arquivo é divido em seções distintas.

Importante! Esse arquivo é somente para usuários avançados. É muito improvável que você precise mudar esse arquivo. Se você requerer assistência adicional nessa área, entre em contato com IBM Corp. para obter ajuda.

Regras de formatação para normalização

- Inclua somente uma entrada de normalização por linha.
- Respeite rigorosamente as seções no arquivo. Nenhuma nova seção pode ser incluída.
- Para desativar uma entrada, coloque um símbolo # no início dessa linha. Para ativar uma entrada, remova o caractere # antes dessa linha.

Datas em inglês na normalização

Por padrão, datas em um modelo em inglês são reconhecidas no formato de data estilo americano, ou seja, mês, dia, ano. Se você precisar mudar para o formato dia, mês, ano, desative a linha "format:US" (incluindo # no início da linha) e ative "format:UK" (removendo # dessa linha).

Configuração

É possível ativar e desativar os tipos de entidade não linguística que você deseja extrair no arquivo de configuração da entidade não linguística. Desativando as entidades não necessárias, é possível diminuir o tempo de processamento necessário. Isso é feito na seção **Configuração** na guia Recursos Avançados. Veja o tópico Capítulo 17, “Sobre recursos avançados”, na página 195 para obter mais informações. Se a extração não linguística estiver ativada, o mecanismo de extração lerá esse arquivo de configuração durante o processo de extração para determinar quais tipos de entidade não linguística devem ser extraídos.

A sintaxe para esse arquivo é a seguinte:

```
     #name<TAB>Language<TAB>Code
```

Tabela 43. Sintaxe para arquivo de configuração	
Rótulo da coluna	**Descrição**
#name	As palavras pelas quais entidades não linguísticas serão referenciadas nos outros dois arquivos necessários para a extração da entidade não linguística. Os nomes usados aqui fazem distinção entre maiúsculas e minúsculas.
Language	O idioma dos documentos . É melhor selecionar o idioma específico; no entanto, existe uma opção **Qualquer um**. As possíveis opções são: 0 = Qualquer um que é usado sempre que uma regexp não é específica de um idioma e poderia ser usada em diversos modelos com idiomas diferentes, por exemplo, um IP/URL/ endereço de email; 1 = Francês; 2 = Inglês; 4 = Alemão; 5 = Espanhol; 6 = Holandês; 8 = Português; 10 = Italiano.
Code	Código da parte do discurso. A maioria das entidades usa um valor de “s”, exceto em alguns casos. Os valores possíveis são: s = palavra comum; a = adjetivo; n = nome. Se ativado, as entidades não linguísticas serão extraídas primeiro e os padrões de extração serão aplicados para identificar sua função em um contexto mais amplo. Por exemplo, as porcentagens recebem um valor de “a.” Suponha que 30% sejam extraídos como uma entidade não linguística. Isso seria identificado como um adjetivo. Então, se seu texto contivesse um "aumento de salário de 30%", a entidade não linguística de “30%” se ajustaria ao padrão de parte do discurso “ann” (adjetivo nome nome).

Ordem de Definição de Entidades

A ordem em que as entidades são declaradas nesse arquivo é importante e afeta a forma como elas são extraídas. Elas são aplicadas na lista ordenada. A mudança na ordem mudará os resultados. As entidades não linguísticas mais específicas devem ser definidas antes das mais gerais.

Por exemplo, a entidade não linguística "Aminoacid" é definida por:

```
regexp1=($(AA)-?$(NUM))
```

em que $(AA) corresponde a “(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)”, que são sequências de três letras específicas correspondentes a determinados aminoácidos.

Por outro lado, a entidade não linguística "Gene" é mais geral e é definida por:

```
regexp1=p[0-9]{2,3}
regexp2=[a-z]{2,4}-?[0-9]{1,3}-?[r]
regexp3=[a-z]{2,4}-?[0-9]{1,3}-?p?
```

Se "Gene" for definido antes de "Aminoacid" na seção Configuração, então "Aminoacid" nunca será correspondido, uma vez que regexp3 de "Gene" sempre corresponderá primeiro.

Regras de Formatação para Configuração

- Use um caractere TAB para separar cada entrada em uma coluna.
- Não exclua nenhuma linha.
- Respeite a sintaxe mostrada na tabela precedente.
- Para desativar uma entrada, coloque um símbolo # no início dessa linha. Para ativar uma entidade, remova o caractere # antes dessa linha.

Manipulação de idioma

Cada idioma usado hoje tem maneiras especiais de expressar ideias, estruturar sentenças e usar abreviações. Na seção Tratamento de Idioma, é possível editar padrões de extração, forçar definições para esses padrões e declarar abreviações para o idioma que você selecionou na lista suspensa Idioma.

- Padrões de extração
- Definições forçadas
- Abreviações

Padrões de extração

Durante a extração de informações de seus documentos , o mecanismo de extração aplica um conjunto de padrões de extração de parte do discurso a uma "pilha" de palavras no texto para identificar termos candidatos (palavras e frases) para extração. É possível incluir ou modificar os padrões de extração.

Partes do discurso incluem elementos gramaticais, como nomes, adjetivos, passados particípios, determinadores, preposições, coordenadores, nomes, rubricas e partículas. Uma série desses elementos compõe um padrão de extração de parte do discurso. Nos produtos de mineração de texto do IBM Corp., cada parte do discurso é representada por um único caractere para facilitar a definição de seus padrões. Por exemplo, um adjetivo é representado pela letra minúsera *a*. O conjunto de códigos suportados aparece por padrão na parte superior de cada seção de padrões de extração padrão juntamente com um conjunto de padrões e exemplos de cada padrão para ajudá-lo a entender cada código que é usado.

Regras de formatação para padrões de extração

- Um padrão por linha.
- Use # no início de uma linha para desativar um padrão.

A ordem que você lista os padrões de extração é muito importante porque uma determinada sequência de palavras é lida somente uma vez pelo mecanismo de extração e é designada aos primeiros padrões de extração para os quais o mecanismo localiza uma correspondência.

Partes apoiadas de códigos de fala

A seguir, uma tabela de todas as partes suportadas de códigos de fala definidas no dicionário compilado em inglês.

Todas as partes de fala que são usadas em um determinado modelo estão listadas na parte superior de **Recursos Avançados** > **padrões de extração**.

A principal diferença entre o modelo de recursos básicos e o modelo de opinião é que quando determinadores mínimos ("d") e preposições ("c") são usados em básico, seus equivalentes estendidas ("e" e "r") são usados em opiniões. "0" e "1" têm um uso limitado em todos os modelos de opinião. Veja **Recursos Avançados** > **Manipulação de Idiomas (Inglês)** > **Definições forçadas e padrões de Extração**.

Outros modelos em inglês podem usar algumas partes de discurso não listadas no dicionário (por exemplo, "w" e "W", no template de Inteligência de Mercado). Mas, nesse caso, essas partes de discurso são atribuídas a palavras específicas sob **Recursos Avançados** > **Definições forçados**.

Tabela 44. Partes apoiadas de códigos de fala

Código	**Significado**	**Exemplo**
a	adjetivo	abdominal, azul ...
A	não usado	não usado
b	advérbio	freqüentemente, muitas vezes, muito, ...
B	não usado	não usado
c	Preposição	"de"
C	código interno para palavras digitada	
d	determinador	"o"

Tabela 44. Partes apoiadas de códigos de fala (continuação)

Código	Significado	Exemplo
D	não usado	não usado
e	determinador estendido	o, um, meu, seu ...
E	não usado	não usado
f	nome	João, Maria ...
F	não usado	não usado
g	não usado	não usado
G	adjetivo de nacionalidade	francês, americano ...
h	não usado	não usado
H	não usado	não usado
i	iniciais todas as letras únicas seguidas de "."	"a.", "w." e algumas letras simples como "w" (usado para extrair nomes de pessoa como John W. Doe)
I	não usado	não usado
j	não usado	não usado
J	não usado	não usado
k	não usado	não usado
K	não usado	não usado
l	não usado	não usado
L	não usado	não usado
m	substantivo ou desconhecido	cão, ibm
M	não usado	não usado
n	substantivo	cachorro
N	alguns nouns adequados	ibm
o	coordenação	"e", "&"
O	não usado	não usado
p	paral do passado	abandonado, acessorizado ...
P	não usado	não usado
q	não usado	não usado
Q	qualificador	caro, pequeno, bom, ...
r	preposição estendida	de, entre, contra, de ...
R	não usado	não usado
s	palavra vazia	qualquer palavra que não queremos extrair
S	não usado	não usado
t	título	mrs., Sra., capitão, brig., ...
T	não usado	não usado

Tabela 44. Partes apoiadas de códigos de fala (continuação)		
Código	**Significado**	**Exemplo**
u	desconhecida por definição, não no dicionário	
U	não usado	não usado
v	verbo	comer, comer, comer, comer, ...
V	verbo infinitivo	comer, ...
w	não usado	não usado
W	não usado	não usado
x	auxiliar	be
X	não usado	não usado
y	partícula	von, di, de, ... (usado para extrair nomes de pessoa: John von Doe)
Y	não usado	não usado
z	não usado	não usado
Z	não usado	não usado
0	adverb de opinião	Só em Opinião. Veja **Recursos Avançados** > **Manejo de Idiomas (Inglês)** > **definições forçadas**.
1	"a" em opiniões	Veja **Recursos Avançados** > **Manejo de Idiomas (Inglês)** > **Definições forçados**
2	não usado	não usado
3	não usado	não usado
4	não usado	não usado
5	não usado	não usado
6	não usado	não usado
7	não usado	não usado
8	não usado	não usado
9	não usado	não usado

Definições Forçadas

Durante a extração de informações de seus documentos , o mecanismo de extração varre o texto e identifica a parte do discurso para cada palavra que encontra. Em alguns casos, uma palavra pode se ajustar a vários papeis diferentes, dependendo do contexto. Se desejar forçar uma palavra para assumir um determinado papel na parte do discurso ou excluir a palavra completamente do processamento, é possível fazer isso na seção **Definição Forçada** da guia Recursos Avançados. Consulte o tópico Capítulo 17, “Sobre recursos avançados”, na página 195 para obter mais informações.

Para forçar um papel de parte do discurso para que certa palavra, deve-se incluir uma linha nesta seção usando a sintaxe a seguir:

```
term:code
```

Tabela 45. Descrição da sintaxe	
Entrada	**Descrição**
`term`	O nome de um termo.
`code`	Um código de um único caractere representando o papel de parte do discurso. É possível listar até seis códigos de parte do discurso diferentes por unitermo. Além disso, é possível impedir que uma palavra seja extraída de palavras/frases compostas usando o código de letra minúscula `s`, como `additional:s`.

Regras de formatação para definições forçadas

- Uma linha por palavra.
- Termos não podem conter dois pontos.
- Use a letra minúscula `s` como um código de parte do discurso para impedir que uma palavra seja completamente excluída.
- Use até seis códigos de parte do discurso por linha. Os códigos de parte do discurso suportados são mostrados na seção Padrões de Extração. Consulte o tópico "Padrões de extração" na página 203 para obter informações adicionais.
- Use o caractere asterisco (*) como um curinga no final de uma sequência de caracteres para correspondências parciais. Por exemplo, se você inserir `add*:s`, palavras como `add`, `additional`, `additionally`, `addendum` e `additive` nunca serão extraídas como um termo ou como uma parte de um termo de palavra composta. No entanto, se uma correspondência de palavra for declarada explicitamente como um termo em um dicionário compilado ou em definições forçadas, ela ainda será extraída. Por exemplo, se você inserir `add*:s` e `addendum:n`, `addendum` ainda será extraída se for localizada no texto.

Abreviações

Quando o mecanismo de extração está processando texto, geralmente ele considera qualquer ponto final que localiza como uma indicação de que uma sentença terminou. Normalmente isso está correto; no entanto, esse tratamento de caracteres de ponto final não se aplica quando há abreviações contidas no texto.

Se você extrair termos do seu texto e achar que certas abreviações foram tratadas incorretamente, será necessário declarar explicitamente essa abreviação nessa seção.

Nota: Se a abreviação já aparecer em uma definição de sinônimo ou estiver definida como um termo em um dicionário de tipo, não haverá necessidade de incluir a entrada da abreviação aqui.

Regras de formatação para abreviações

- Defina uma abreviação por linha.

Capítulo 18. Sobre regras de ligação de texto

A análise de ligação de texto (TLA) é uma tecnologia de correspondência de padrões usada para extrair relacionamentos localizados em seu texto usando um conjunto de regras. Quando a análise de ligação de texto é ativada para extração, os dados de texto são comparados com essas regras. Quando uma correspondência é localizada, o padrão de análise de ligação de texto é extraído e apresentado. Essas regras são definidas na guia Regras de Ligação de Texto.

Por exemplo, a extração de conceitos representando ideias simples sobre uma organização pode não ser interessante o suficiente para você, mas usando a TLA, você também poderia saber mais sobre as ligações entre diferentes organizações ou as pessoas associadas à organização. A TLA também pode ser usada para extrair pareceres sobre tópicos, como, por exemplo, como as pessoas se sentem em relação a um determinado produto ou experiência.

Para se beneficiar da TLA, você deve ter recursos que contenham regras de ligação de texto (TLA). Quando você seleciona um modelo, é possível ver quais modelos têm regras de TLA, dependendo se eles possuem ou não um ícone na coluna TLA.

Os padrões de análise de ligação de texto são localizados nos dados de texto durante a fase de correspondência de padrões do processo de extração. Durante essa fase, as regras são comparadas com os dados de texto e, quando uma correspondência é localizada, essa informação é extraída como um padrão. Há momentos em que você pode querer obter mais da análise de ligação de texto ou mudar como algo é correspondido. Nesses casos, é possível refinar as regras para adaptá-las a suas necessidades específicas. Isso é feito na guia Regras de Ligação de Texto.

Nota: A partir da versão 18.2, as Regras de Redesignação de Tipo (TRRs) estão disponíveis. As TRRs transformam uma sequência de tipos, macros e/ou tokens em um novo conceito com um tipo específico. Elas podem ser usadas em modelos de Opiniões para capturar opiniões com uma mudança em polaridade. Para obter mais informações, consulte “Regras de redesignação de tipo” na página 151.

Onde trabalhar nas Regras de Ligação de Texto

É possível editar e criar regras diretamente na guia Regras de Ligação de Texto na visualização Editor de Template ou Editor de Recursos. Para ajudá-lo a ver como as regras podem corresponder ao texto, é possível executar uma simulação nessa guia. Durante a simulação, uma extração é executada somente nos dados de simulação de amostra e as regras de ligação de texto são aplicadas para ver se algum padrão corresponde. Quaisquer regras correspondentes ao texto são mostradas na área de janela de simulação. Com base nas correspondências, é possível escolher editar regras e macros para mudar como o texto é correspondido.

Ao contrário de outros recursos avançados, as regras de TLA são específicas de bibliotecas; portanto, só é possível usar as regras de TLA de uma biblioteca por vez. De dentro do Editor de Template ou Editor de Recursos, acesse a guia **Regras de Ligação de Texto**. Nessa guia, é possível especificar a biblioteca em seu modelo que contém as regras de TLA que você deseja usar ou editar. Por esse motivo, é altamente recomendado armazenar todas as suas regras em uma biblioteca, a menos que haja um motivo muito específico que não seja desejado.

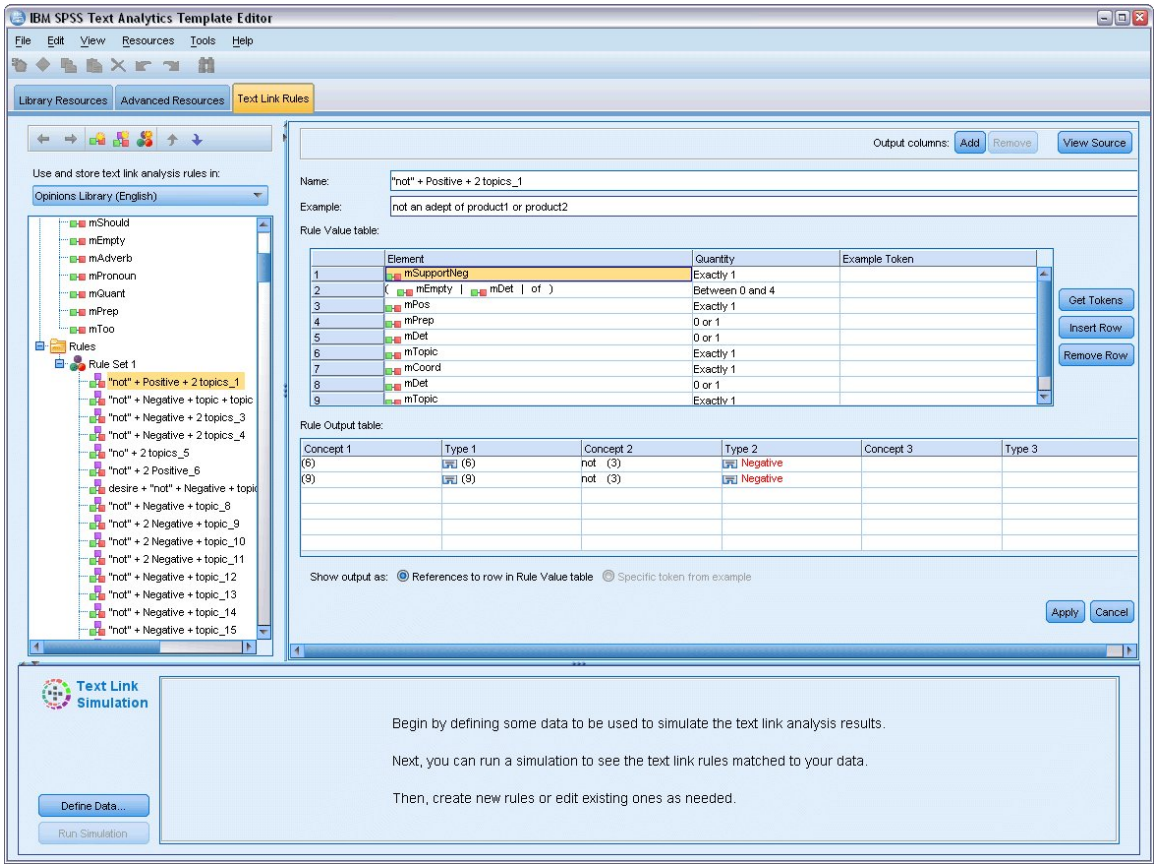


Figura 44. Guia Regras de Ligação de Texto

Por onde começar

Há inúmeras maneiras de se começar a trabalhar no editor da guia Regra de Ligação de Texto:

- Comece simulando resultados com alguns textos de amostra e edite ou crie regras de correspondência com base em como o conjunto atual de regras extrai padrões dos dados de simulação.
- Crie uma nova regra do zero ou edite uma existente.
- Trabalhe diretamente na visualização de origem.

Quando editar ou criar regras

Embora geralmente as regras de análise de ligação de texto entregues com cada modelo sejam adequadas para a extração de vários relacionamentos simples e complexos do seu texto, haverá momentos em que você desejará fazer algumas mudanças nessas regras ou criar as suas próprias regras. Por exemplo:

- Para capturar uma ideia ou uma relação que não está sendo extraída com as regras existentes criando uma nova regra ou macro.
- Para mudar o comportamento padrão de um tipo que você incluiu nos recursos. Geralmente isso requer a edição de uma macro, como `mTopic` ou `mNonLingEntities`. Consulte o tópico "Macros especiais: mTopic, mNonLingEntities, SEP" na página 214 para obter mais informações.
- Para incluir novos tipos em macros e regras de análise de ligação de texto existentes. Por exemplo, se você achar que o tipo `<Organization>` é muito amplo, você poderia criar novos tipos para organizações em vários setores de negócios diferentes como `<Pharmaceuticals>`, `<Car Manufacturing>`, `<Finance>`e etc. Nesse caso, deve-se editar as regras de análise de ligação de texto e/ou criar uma macro para levar esses novos tipos em conta e processá-los de acordo.

- Para incluir tipos em uma regra de análise de ligação de texto existente. Por exemplo, digamos que você tem uma regra que captura o seguinte texto `john doe called jane doe` mas você quer essa regra que capta as comunicações telefônicas para também capturar trocas de e-mails. Você poderia adicionar o tipo de entidade não linguística para e-mail à regra para que ele também capturasse textos como: `johndoe@ibm.com emailed janedoe@ibm.com`.
- Para modificar um pouco uma regra existente em vez de criar uma nova. Por exemplo, digamos que você tenha uma regra que corresponda ao texto a seguir `xyz is very good` mas deseja que essa regra também capture `xyz is very, very good`.

Simulando resultados da Análise de Ligação de Texto

Para ajudá-lo a definir novas regras de ligação de texto ou ajudá-lo a entender como certas sentenças são correspondidas durante a análise de ligação de texto, geralmente é útil pegar uma parte de amostra de texto e executar uma simulação. Durante a simulação, uma extração é executada somente nos dados de simulação de amostra usando o conjunto atual de recursos linguísticos e as atuais configurações de extração. O objetivo é obter os resultados simulados e usá-los para melhorar suas regras, criar outras novas ou entender melhor como ocorre a correspondência. Para cada parte do texto (sentença, palavra ou cláusula, dependendo do contexto), uma saída de simulação exibe a coleção de tokens e quaisquer regras de TLA que descobriram um padrão nesse texto. Um **token** é definido como qualquer palavra ou frase identificada durante o processo de extração.

Ao contrário de outros recursos avançados, as regras de TLA são específicas de bibliotecas; portanto, só é possível usar as regras de TLA de uma biblioteca por vez. De dentro do Editor de Template ou Editor de Recursos, acesse a guia **Regras de Ligação de Texto**. Nessa guia, é possível especificar a biblioteca em seu modelo que contém as regras de TLA que você deseja usar ou editar. Por esse motivo, é altamente recomendado armazenar todas as suas regras em uma biblioteca, a menos que haja um motivo muito específico que não seja desejado.

Importante! É altamente recomendado que, se usar um arquivo de dados, você assegure que o texto que ele contém seja pequeno para minimizar o tempo de processamento. O objetivo da simulação é ver como uma parte do texto é interpretada e entender como as regras correspondem a esse texto. Essas informações ajudarão você a gravar ou editar suas regras. Use o nó de análise de ligação de texto ou execute um fluxo com uma sessão interativa com a extração de TLA ativada para obter resultados para um conjunto de dados mais completo. Essa simulação é apenas para propósitos de teste e autoria de regra.

Definindo dados para simulação

Para ajudá-lo a ver como as regras podem corresponder ao texto, é possível executar uma simulação usando dados de amostra. O primeiro passo é definir os dados.

Definindo dados

1. Clique em **Definir Dados** na área de janela de simulação na parte inferior da guia **Regras de Ligação de Texto**. Alternativamente, se nenhum dado tiver sido previamente definido, escolha **Ferramentas** > **Executar Simulação** a partir dos menus. O assistente Dados de Simulação é aberto.
2. Especifique o tipo de dados selecionando um dos seguintes:
   - **Colar ou inserir texto diretamente** Uma caixa de texto é fornecida para você colar algum texto da área de transferência ou para inserir manualmente o texto desejado para ser processado. É possível inserir uma sentença por linha ou usar pontuação para dividir a sentença, como pontos finais ou vírgulas. Após você ter inserido seu texto, é possível iniciar a simulação clicando em **Executar Simulação**.
   - **Especificar uma origem de dados de arquivo** Esta opção indica que você deseja processar um arquivo contendo texto. Clique em **Avançar** para continuar no passo do assistente no qual é possível definir o arquivo a ser processado. Após o arquivo ter sido selecionado, é possível iniciar a simulação clicando em **Executar Simulação**. Os tipos de arquivo a seguir são suportados: `.txt` e `.text`. O arquivo de dados escolhido é lido 'no estado em que se encontra' durante a simulação. O arquivo

inteiro é tratado da mesma maneira que se você tivesse conectado um nó Lista de Arquivos a um nó Mineração de Texto.

Importante: É altamente recomendado que, se usar um arquivo de dados, você assegure que o texto que ele contém seja pequeno para minimizar o tempo de processamento. O objetivo da simulação é ver como uma parte do texto é interpretada e entender como as regras correspondem a esse texto. Essas informações ajudarão você a gravar ou editar suas regras. Use o nó de análise de ligação de texto ou execute um fluxo com uma sessão interativa com a extração de TLA ativada para obter resultados para um conjunto de dados mais completo. Essa simulação é apenas para propósitos de teste e autoria de regra.

3. Para iniciar o processo de simulação, clique em **Executar Simulação**. Um diálogo de progresso aparece. Se você estiver em uma sessão interativa, as configurações de extração usadas durante a simulação são aquelas selecionadas atualmente na sessão interativa (veja **Ferramentas** > **Configurações de Extração** na visualização de Conceitos e Categorias). Se você estiver no Editor de Template, as configurações de extração usadas durante a simulação são as configurações de extração padrão, que são as mesmas que aquelas mostradas na guia Especialista de um nó Análise de Ligação de Texto. Para obter informações adicionais, consulte "Entendendo resultados da simulação" na página 210.

Entendendo resultados da simulação

Para ajudá-lo a ver como as regras podem corresponder ao texto, é possível executar uma simulação usando dados de amostra e revisar os resultados. A partir daí, é possível mudar seu conjunto de regras para melhor ajustar seus dados. Quando a extração e o processo de simulação são concluídos, são apresentados os resultados da simulação.

Para cada "sentença" identificada durante a extração, são apresentadas várias informações, incluindo a "sentença" exata, o detalhamento dos tokens localizados nessa sentença de texto de entrada e, por fim, quaisquer regras correspondentes ao texto nessa sentença. Com "**sentença**", queremos dizer uma palavra, uma sentença ou uma cláusula, dependendo de como o extrator dividiu o texto em partes legíveis.

Um **token** é definido como qualquer palavra ou frase de palavras identificadas durante o processo de extração. Por exemplo, na sentença *Meu tio vive em Nova Iorque*, os tokens a seguir podem ser localizados durante a extração: *meu*, *tio*, *vive*, *em* e *nova iorque*. Adicionalmente, o *tio* poderia ser extraído como um conceito e digitado como <Unknown>, e *nova york* também poderia ser extraído como um conceito e digitado como <Location>. Todos os conceitos são tokens, mas nem todos os tokens são conceitos. Tokens também podem ser macros, sequências de caracteres literais e diferenças de palavras. Somente palavras ou frases de palavras tipificadas podem ser conceitos.

Ao trabalhar na sessão interativa ou no editor de recurso, você está trabalhando em nível de conceito. As regras de TLA são mais granulares, e os tokens individuais em uma sentença podem ser usados na definição de uma regra, mesmo que eles nunca sejam extraídos e digitados. Poder usar tokens que não são conceitos oferece às regras muito mais flexibilidade na captura de relacionamentos complexos em seu texto.

Se você tiver mais de uma sentença em seus dados de simulação, é possível voltar e avançar nos resultados clicando em **Próximo** e **Anterior**.

Nos casos em que uma sentença não corresponde a nenhuma regra TLA na biblioteca selecionada (consulte o nome da biblioteca acima da árvore nessa guia), os resultados são considerados não correspondentes e os botões **Próximo Não Correspondente** e **Anterior Não Correspondente** são ativados para permitir que você saiba que há um texto para o qual nenhuma regra localizou uma correspondência e para permitir que você navegue nessas instâncias rapidamente.

Após criar novas regras, editar suas regras ou mudar seus recursos ou configurações de extração, talvez você queira executar uma simulação novamente. Para executar uma simulação novamente, clique em **Executar Simulação** na área de janela de simulação e os mesmos dados de entrada serão usados novamente.

Os campos e as tabelas a seguir são mostrados nos resultados da simulação:

Texto de Entrada. A 'sentença' real identificada pelo processo de extração para os dados de simulação que você definiu no assistente. Com sentença, queremos dizer uma palavra, uma sentença ou uma cláusula, dependendo de como o extrator dividiu o texto em partes legíveis.

Visualização do Sistema. Uma coleção de tokens que o processo de extração identificou.

- **Token de Texto de Entrada.** Cada token localizado no texto de entrada. Tokens foram definidos anteriormente neste tópico.
- **Tipificado como.** Se um token foi identificado como um conceito e digitado, então o nome do tipo associado (tais como `<Unknown>`, `<Person>`, `<Location>`) é mostrado nesta coluna.
- **Macro Correspondente.** Se um token correspondeu a uma macro existente, o nome da macro associada será exibido nessa coluna.

Regras Correspondentes ao Texto de Entrada. Esta tabela mostra quaisquer regras de TLA que foram correspondidas com relação ao texto de entrada. Para cada regra correspondida, você verá o nome da regra na coluna **Saída de Regra** e os valores de saída associados para essa regra (pares de Conceito + Tipo). É possível dar um clique duplo no nome da regra correspondida para abri-la na área de janela do editor acima da área de janela de simulação.

Botão **Gerar Regra**. Se você clicar nesse botão na área de janela de simulação, uma nova regra será aberta no editor de regras acima da área de janela de simulação. Ele usará o texto de entrada como exemplo. Da mesma forma, qualquer token que foi tipificado e correspondido a uma macro durante a simulação é inserido automaticamente na coluna Elementos na tabela **Valores de Regra**. Se um token foi tipificado *e* correspondido a uma macro, o valor da macro será aquele usado na regra para simplificá-la. Por exemplo, a frase "*Eu gosto de pizza*" poderia ser digitada durante a simulação como `<Unknown>` e correspondia a macro `mTopic` se você estivesse usando os recursos básicos do Inglês. Nesse caso, `mTopic` será usado como o elemento na regra gerada. Consulte o tópico “Trabalhando com regras de ligação de texto” na página 215 para obter mais informações.

Navegando em regras e macros na árvore

Quando a análise de ligação de texto é executada durante a extração, as regras de ligação de texto armazenadas na biblioteca selecionada na guia **Regras de Ligação de Texto** são usadas.

Ao contrário de outros recursos avançados, as regras de TLA são específicas de bibliotecas; portanto, só é possível usar as regras de TLA de uma biblioteca por vez. De dentro do Editor de Template ou Editor de Recursos, acesse a guia **Regras de Ligação de Texto**. Nessa guia, é possível especificar a biblioteca em seu modelo que contém as regras de TLA que você deseja usar ou editar. Por esse motivo, é altamente recomendado armazenar todas as suas regras em uma biblioteca, a menos que haja um motivo forte ou específico que não seja desejado.

É possível especificar em qual biblioteca você deseja trabalhar na guia Regras de Ligação de Texto selecionando essa biblioteca na lista suspensa **Usar e armazenar regras de análise de ligação de texto em:** nesta guia. Quando a análise de ligação de texto é executada durante a extração, as regras de ligação de texto armazenadas na biblioteca selecionada na guia **Regras de Ligação de Texto** são usadas. Portanto, se você definiu regras de ligação de texto (regras de TLA) em mais de uma biblioteca, somente a primeira biblioteca em que as regras de TLA estão localizadas será usada para análise de ligação de texto. Por esse motivo, é altamente recomendado armazenar todas as suas regras em uma biblioteca, a menos que haja um motivo muito específico que não seja desejado.

Quando você seleciona uma macro ou uma regra na árvore, seu conteúdo é exibido na área de janela do editor à direita. Se você clicar com o botão direito em qualquer item na árvore, um menu de contexto será aberto para mostrar quais outras tarefas são possíveis, como:

- Criar uma nova macro na árvore e abri-la no editor à direita.
- Criar uma nova regra na árvore e abri-la no editor à direita.
- Criar um novo conjunto de regras na árvore.
- Cortar, copiar e colar itens para simplificar a edição.
- Excluir macros, regras e conjuntos de regras para removê-los dos recursos.

- Desativar macros, regras e conjuntos de regras para indicar que eles devem ser ignorados durante o processamento.
- Mover regras para cima ou para baixo para afetar a ordem de processamento.

Avisos na árvore

Avisos são exibidos com um triângulo amarelo na árvore e estão lá para informar que pode haver um problema. Passe o ponteiro do mouse sobre a macro ou regra com falha para exibir uma explicação pop-up. Na maioria dos casos, você verá algo como: **Aviso: Nenhum exemplo fornecido; Insira um exemplo**, portanto, é necessário inserir um exemplo.

Se você não tiver um exemplo, ou se o exemplo não corresponder à regra, não será possível usar a variável Obter Tokens; portanto, é recomendado inserir apenas um exemplo por regra.

Quando a regra estiver destacada em amarelo, isso significa que um tipo ou uma macro é desconhecido no editor de TLA. A mensagem será semelhante à: **Aviso: Tipo ou macro desconhecido**. Isso é para informá-lo de que um item que seria definido por `$something` na visualização de origem, por exemplo, `$myType`, não é um tipo anterior em sua biblioteca, nem uma macro.

Para atualizar o verificador de sintaxe, é preciso alternar para outra regra ou macro; não é necessário recompilar nada. Portanto, por exemplo, se a regra A exibir um aviso porque o exemplo está omisso, é necessário incluir um exemplo, clicar em uma regra superior ou inferior e retornar à regra A para verificar se agora ela está correta.

Trabalhando com macros

Macros pode simplificar a aparência das regras de análise de link de texto permitindo que você agrupe tipos, outras macros e strings literais (word) juntamente com um operador OR (`|`). A vantagem em usar macros é que não só você pode reutilizar macros em várias regras de análise de link de texto para simplificá-las, mas também possibilita fazer atualizações em uma macro em vez de ter que fazer atualizações ao longo de todas as suas regras de análise de link de texto. A maioria das regras de TLA divididas contém macros predefinidas. As macros aparecem na parte superior da árvore na área de janela esquerda da guia Regras de Ligação de Texto.

Os campos e as tabelas a seguir são mostrados nos resultados da simulação:

Nome. Um nome exclusivo identificando essa macro. Recomendamos prefixar os nomes de macros com uma letra m minúscula para ajudá-lo a identificar macros rapidamente em suas regras. Quando você se refere a macros manualmente em suas regras (fazendo uma edição sequencial ou na visualização de origem), é preciso usar o caractere prefixo $ para que o processo de extração saiba procurar esse nome especial. No entanto, se você arrastar e soltar o nome da macro e incluí-lo nos menus de contexto, o produto o reconhecerá automaticamente como uma macro e nenhum $ será incluído.

Tabela **Valor da Macro**.

- Um número de linhas representando todos os valores possíveis que essa macro pode representar. Esses valores fazem distinção entre maiúsculas e minúsculas.
- Esses valores podem incluir um ou uma combinação de tipos, sequências de caracteres literais, diferenças de palavras ou macros. Consulte o tópico “Elementos suportados para regras e macros” na página 221 para obter informações adicionais.
- Para inserir um valor para um elemento em uma macro, dê um clique duplo na linha na qual deseja trabalhar. Uma caixa de texto editável aparece, na qual é possível inserir uma referência de tipo, uma referência de macro, uma sequência de caracteres literal ou uma diferença de palavra. Como alternativa, clique com o botão direito na célula para exibir um menu contextual oferecendo listas de macros comuns, nomes de tipos e nomes de tipos não linguísticos. Para fazer referência a um tipo ou a uma macro, deve-se preceder o nome do tipo ou macro com um caractere ‘$’, como `$mTopic` para a macro `mTopic`. Durante a combinação de argumentos, deve-se usar parênteses `( )` para agrupar os argumentos e o caractere `|` para indicar um booleano OR.
- É possível incluir ou remover linhas na tabela Valor da Macro usando os botões à sua direita.

- Insira cada elemento em sua própria linha. Por exemplo, se você quisesse criar uma macro representando uma de 3 sequências de caracteres literais, como `am OR was OR is`, seria possível inserir cada sequência de caracteres literal em uma linha separada na visualização e sua tabela Macro conteria 3 linhas.

Criando e editando macros

É possível criar novas macros ou editar as existentes. Siga as diretrizes e as descrições para o editor de macro. Consulte o tópico "Trabalhando com macros" na página 212 para obter informações adicionais.

Criando Novas Macros

1. A partir dos menus, escolha **Ferramentas> Novo Macro**. Alternativamente, clique no ícone Nova Macro na barra de ferramentas da árvore para abrir uma nova macro no editor.
2. Insira um nome exclusivo e defina os elementos de valor da macro.
3. Clique em **Aplicar** ao concluir a verificação de erros.

Editando Macros

1. Clique no nome da macro na árvore. A macro é aberta na área de janela do editor à direita.
2. Faça as mudanças.
3. Clique em **Aplicar** ao concluir a verificação de erros.

Desativando e excluindo macros

Desativando macros

Se desejar que uma macro seja ignorada durante o processamento, é possível desativá-la. Isso pode causar avisos ou erros em quaisquer regras que ainda façam referência a essa macro desativada. Tome cuidado ao excluir e desativar macros.

1. Clique no nome da macro na árvore. A macro é aberta na área de janela do editor à direita.
2. Clique com o botão direito no nome.
3. Nos menus de contexto, escolha **Desativar**. O ícone da macro fica cinza e a macro se torna não editável.

Excluindo macros

Se desejar se livrar de uma macro, é possível excluí-la. Isso pode causar erros em quaisquer regras que ainda façam referência a essa macro. Tome cuidado ao excluir e desativar macros.

1. Clique no nome da macro na árvore. A macro é aberta na área de janela do editor à direita.
2. Clique com o botão direito no nome.
3. Nos menus de contexto, escolha **Excluir**. A macro desaparece da lista.

Verificando erros, salvando e cancelando

Aplicando mudanças de macro

Se você clicar fora do editor de macro ou em **Aplicar**, a macro será automaticamente varrida em busca de erros. Se um erro for localizado, você precisará corrigi-lo antes de mover para outra parte do aplicativo.

No entanto, se erros menos sérios forem detectados, somente um aviso será fornecido. Por exemplo, se sua macro contiver definições incompletas ou não referenciadas a tipos ou outras macros, uma mensagem de aviso será exibida. Após um clique em **Aplicar**, quaisquer avisos não corrigidos farão com que um ícone de aviso apareça à esquerda do nome da macro na Árvore de Rules e Macro na área de janela esquerda.

A aplicação de uma macro não significa que sua macro esteja permanentemente salva. A aplicação fará com que o processo de validação verifique erros e avisos.

Salvando recursos dentro de uma sessão de ambiente de trabalho interativa

1. Para salvar as mudanças feitas em seus recursos durante uma sessão de ambiente de trabalho interativa para que seja possível obtê-las na próxima vez que você executar seu fluxo, deve-se:
   - Atualizar seu nó de modelagem para garantir que você obterá os mesmos recursos na próxima vez que executar seu fluxo. Consulte o tópico "Atualizando nós de modelagem e salvando" na página 75 para obter informações adicionais. Em seguida, salve seu fluxo. Para salvar seu fluxo, faça isso na janela principal do IBM SPSS Modeler após atualizar o nó de modelagem.
2. Para salvar as mudanças feitas em seus recursos durante uma sessão de ambiente de trabalho interativa para que seja possível usá-las em outros fluxos, é possível:

- Atualizar o modelo que você usou ou criar um novo. Consulte o tópico "Criando e atualizando modelos" na página 162 para obter informações adicionais. Isso não salvará as mudanças no nó atual (consulte o passo anterior)
- Ou atualizar a TAP usada. Consulte o tópico "Atualizando Pacotes de Análise de Texto" na página 133 para obter informações adicionais.

Salvando Recursos dentro do Editor de Template

1. Primeiro, publique a biblioteca. Consulte o tópico "Publicando bibliotecas" na página 181 para obter informações adicionais.
2. Em seguida, salve o modelo por meio do **Arquivo> Salvar Modelo de Recursos** nos menus.

Cancelando mudanças de macro

1. Se quiser descartar as mudanças, clique em **Cancelar**.

Macros especiais: mTopic, mNonLingEntities, SEP

O modelo Opiniões (e modelos semelhantes), bem como os modelos Recursos Básicos são fornecidos com duas macros especiais chamadas `mTopic` e `mNonLingEntities`.

mTopic

Por padrão, a macro `mTopic` agrupa todos os tipos embarcados no template que provavelmente estarão conectados com uma opinião, como os seguintes tipos de biblioteca *Core* : `<Person>`, `<Organization>`, `<Location>`, e assim por diante, desde que o tipo não seja um tipo de opinião (por exemplo, `<Negative>` ou `<Positive>`) ou um tipo definido como uma entidade não linguística nos Recursos Avançados.

Sempre que você cria um novo tipo em um modelo Pareceres (ou semelhante), o produto assume que, a menos que esse tipo seja especificado em outra macro ou nas entidades não linguísticas da guia Recursos Avançados, ele será tratado da mesma maneira que os outros tipos definidos na macro `mTopic`.

Digamos que você tenha criado novos tipos nos recursos a partir de um modelo de Opinião: `<Vegetables>` e `<Fruit>`. Sem precisar fazer qualquer mudança, seus novos tipos serão tratados como tipos `mTopic`, assim, é possível descobrir automaticamente os pareceres positivo, negativo, neutro e contextual sobre seus novos tipos. Durante a extração, por exemplo, a sentença "*Eu gosto de brócolis, mas odeio toranja*" produziria os dois padrões de saída a seguir:

```
broccoli <Vegetables> + like <Positive>
```

```
grapefruit <Fruit> + dislike <Negative>
```

No entanto, se desejar processar esses tipos diferentemente dos outros em mTopic, é possível incluir o nome do tipo em uma macro existente, como `mPos`, que agrupa todos os tipos de pareceres positivos, ou criar uma nova macro que possa servir de referência posteriormente em uma ou mais regras.

Importante! Se você criar um novo tipo, como `<Vegetables>`, esse novo tipo será incluído como um tipo em mTopic, no entanto, esse nome de tipo não será explicitamente visível na definição de macro

mNonLingEntities

Similarmente, se você incluir novas entidades não linguísticas na seção **Entidades Não Linguísticas** da guia Recursos Avançados, elas serão processadas automaticamente como `mNonLingEntities`, a menos

que seja especificado de outra forma. Consulte o tópico “Entidades não linguísticas” na página 198 para obter mais informações.

SET

Também é possível usar a macro predefinida SEP, que corresponde ao separador global definido na máquina local, geralmente uma vírgula (,).

Trabalhando com regras de ligação de texto

Uma regra de análise de ligação de texto é um query booleano usado para executar uma correspondência em uma sentença. As regras de análise de ligação de texto contêm um ou mais dos argumentos a seguir: tipos, macros, sequências de caracteres literais ou diferenças de palavras. Deve-se ter pelo menos uma regra de análise de ligação de texto para se extrair resultados de TLA.

As áreas e os campos a seguir são exibidos na guia Regras de Ligação de Texto, Editor de Regra:

Nome campo. O nome exclusivo da regra de ligação de texto.

Exemplo campo. Opcionalmente, é possível incluir uma sentença ou uma sequência de palavras de exemplo que seriam capturadas por essa regra. É recomendado usar exemplos. Nesse editor, você poderá gerar tokens a partir desse texto de exemplo para ver como ele corresponde à regra e como ele será emitido.Um **token** é definido como qualquer palavra ou frase de palavras identificadas durante o processo de extração. Por exemplo, na sentença *Meu tio vive em Nova Iorque*, os tokens a seguir podem ser localizados durante a extração: *meu*, *tio*, *vive*, *em* e *nova iorque*. Adicionalmente, o *tio* poderia ser extraído como um conceito e digitado como <Unknown>, e *nova york* também poderia ser extraído como um conceito e digitado como <Location>. Todos os conceitos são tokens, mas nem todos os tokens são conceitos. Tokens também podem ser macros, sequências de caracteres literais e diferenças de palavras. Somente palavras ou frases de palavras tipificadas podem ser conceitos.

Tabela de Valor de Regra. Essa tabela contém os elementos da regra que são usados para fazer a correspondência de uma regra com uma sentença. É possível incluir ou remover linhas na tabela usando os botões à sua direita. A tabela consiste em 3 colunas:

- **Elemento** coluna. Insira valores como um ou uma combinação de tipos, strings literais, lacunas de palavras (<Any Token>) ou macros. Consulte o tópico “Elementos suportados para regras e macros” na página 221 para obter mais informações. Dê um clique duplo na célula do elemento para inserir informações diretamente. Como alternativa, clique com o botão direito na célula para exibir um menu contextual oferecendo listas de macros comuns, nomes de tipos e nomes de tipos não linguísticos. Lembre-se de que se você inserir informações na célula digitando-as, preceda a macro ou digite um nome com um caractere ‘$’, como $mTopic para a macro mTopic. A ordem em que você cria suas linhas de elemento é crítica para a maneira como a regra corresponderá ao texto. Durante a combinação de argumentos, deve-se usar parênteses ( ) para agrupar os argumentos e o caractere | para indicar um booleano OR. Lembre-se de que os valores fazem distinção entre maiúsculas e minúsculas.
- **Quantidade** coluna. Indica os números mínimo e máximo de vezes que o elemento deve ser localizado para ocorrer uma correspondência. Por exemplo, se quisesse definir uma diferença, ou uma série de palavras, entre outros dois elementos contendo de 0 a 3 palavras, você poderia escolher **Entre 0 e 3** na lista ou inserir os números diretamente na caixa de diálogo. O padrão é ‘**Exatamente 1**’. Em alguns casos, você vai querer um elemento opcional. Se esse for o caso, você terá uma quantidade mínima de 0 e uma quantidade máxima maior que 0 (ou seja 0 ou 1, entre 0 e 2). Observe que o primeiro elemento em uma regra não pode ser opcional, o que significa que ele não pode ter uma quantidade de 0.
- Coluna **Token de Exemplo**. Se você clicar em **Obter Tokens**, o programa dividirá o texto **Exemplo** em tokens e usará esses tokens para preencher essa coluna com aqueles que correspondem aos elementos definidos. Também é possível ver esses tokens na tabela de saída, se escolher isso.

Tabela de saída de regra Cada linha nesta tabela define como a saída de padrão TLA aparecerá nos resultados. A saída de regra pode produzir padrões de até seis pares de colunas Conceito/Tipo, cada um representando um *slot*. Por exemplo, o padrão de tipo <Location> + <Positive> é um padrão de dois slots que significa que ele é composto de 2 pares de coluna Concept / Type.

Nota: Termos na coluna **Elemento** da **Tabela de Valor da Regra**, ou em qualquer uma das colunas **Concept** da **Tabela de saída de regra** não pode iniciar com nenhum dos seguintes caracteres: `, #,%, ^, *, _,-,:, <, >, /, \, ou ".

Assim como o idioma nos lá liberdade para expressar as mesmas ideias básicas de várias formas diferentes, você pode ter diversas regras definidas para capturar a mesma ideia básica. Por exemplo, o texto *"Paris é um lugar que amo"* e o texto *"Eu realmente gosto muito de Paris e Florence"* representam a mesma ideia básica -- que você gosta de Paris -- mas são expressos de formas diferentes e exigem duas regras diferentes para serem capturados. No entanto, é mais fácil trabalhar com os resultados padrão se ideias semelhantes forem agrupadas. Por esta razão, enquanto você pode ter 2 regras diferentes para capturar estas 2 frases, você poderia definir a mesma saída para ambas as regras, como o padrão do tipo `<Location> + <Positive>` para que ele represente ambos os textos. E dessa maneira, é possível ver nem sempre que a saída imita a estrutura ou a ordem das palavras localizadas no texto original. Além disso, tal padrão de tipo poderia corresponder a outras frases e poderia produzir padrões de conceito como: `paris + like` e `tokyo + like`.

Para ajudá-lo a definir a saída rapidamente com menos erros, é possível usar o menu de contexto para escolher o elemento que você deseja ver na saída. Como alternativa, também é possível arrastar e soltar elementos da tabela Valor de Regra na saída. Por exemplo, se você tiver uma regra que contém uma referência à macro `mTopic` na linha 2 da tabela Valor de Regra, e desejar que o valor esteja em sua saída, é possível simplificar a ação de arrastar/soltar o elemento para `mTopic` para o primeiro par de colunas na tabela Saída de Regra. Isso preencherá automaticamente o Conceito e o Tipo do par selecionado. Ou, se você desejar que a saída comece com o tipo definido pelo terceiro elemento (linha 3) da tabela de valor de regra, arraste esse tipo da tabela Valor de Regra para a célula **Tipo 1** na tabela de saída. A tabela será atualizada para mostrar a referência de linha entre parênteses (3).

Como alternativa, é possível inserir essas referências manualmente na tabela dando um clique duplo na célula em cada coluna **Conceito** que você deseja emitir e inserindo o símbolo $ seguido pelo número da linha, como $2, para se referir ao elemento definido na linha 2 da tabela Valor de Regra. Quando você insere as informações manualmente, também é necessário definir a coluna **Tipo** e inserir o símbolo *#* seguido pelo número da linha, como *#*2, para se referir ao elemento definido na linha 2 da tabela Valor de Regra.

Além disso, você pode até combinar métodos. Digamos que você teve o tipo `<Positive>` na linha 4 da sua tabela de Valor de Regra. Você poderia arrastá-lo para a coluna `Type 2` e, em seguida, clicar duas vezes na célula na coluna `Concept 2` e, em seguida, digitar manualmente a palavra '*not*' na frente dele. A coluna de saída então leria `not (4)` na tabela, ou se você estava no modo de edição ou modo de origem `not $4`. Então seria possível clicar com o botão direito na coluna Tipo 1 e selecionar, por exemplo, a macro chamada `mTopic`. Então essa saída poderia resultar em um padrão de conceito tal como: `car + bad`.

A maioria das regras só tem uma linha de saída, mas haverá vezes em que mais de uma saída será possível e desejado. Nesse caso, defina uma saída por linha na tabela Saída de Regra.

Importante: Lembre-se de que outras operações de tratamento linguístico são executadas durante a extração de padrões de TLA. Portanto, quando a saída ler `t$3\t#3`, isso significa que o padrão finalmente exibirá o conceito final para o terceiro elemento e o tipo final para o terceiro elemento após todo o processamento linguístico ser aplicado (sinônimos ou outros agrupamentos).

- **Mostrar saída como**. Por padrão, a opção **Referências à linha na tabela Valor de Regra** é selecionada e a saída é mostrada usando as referências numéricas às linhas definidas na guia Valor de Regra. Se você tiver clicado anteriormente em Obter Tokens e tiver tokens na coluna Tokens de Exemplo na tabela Valor de Regra, será possível escolher visualizar a saída para esses tokens específicos escolhendo a opção.

Nota: Se não houver pares de saída de conceito / tipo suficientes mostrados na tabela de saída, você poderá adicionar outro par clicando no botão Add na barra de ferramentas do editor. Se três pares forem mostrados atualmente e você clicar em Incluir, mais duas colunas (Conceito 4 e Tipo 4) serão incluídas na tabela. Isso significa que agora você verá quatro pares na tabela de saída para todas as regras. Também é possível remover pares não usados, contanto que nenhuma outra regra no conjunto de regras nessa biblioteca use o par.

Regra de exemplo

Vamos supor que seus recursos contenham a regra de análise de ligação de texto a seguir e você tenha ativado a extração dos resultados de TLA:

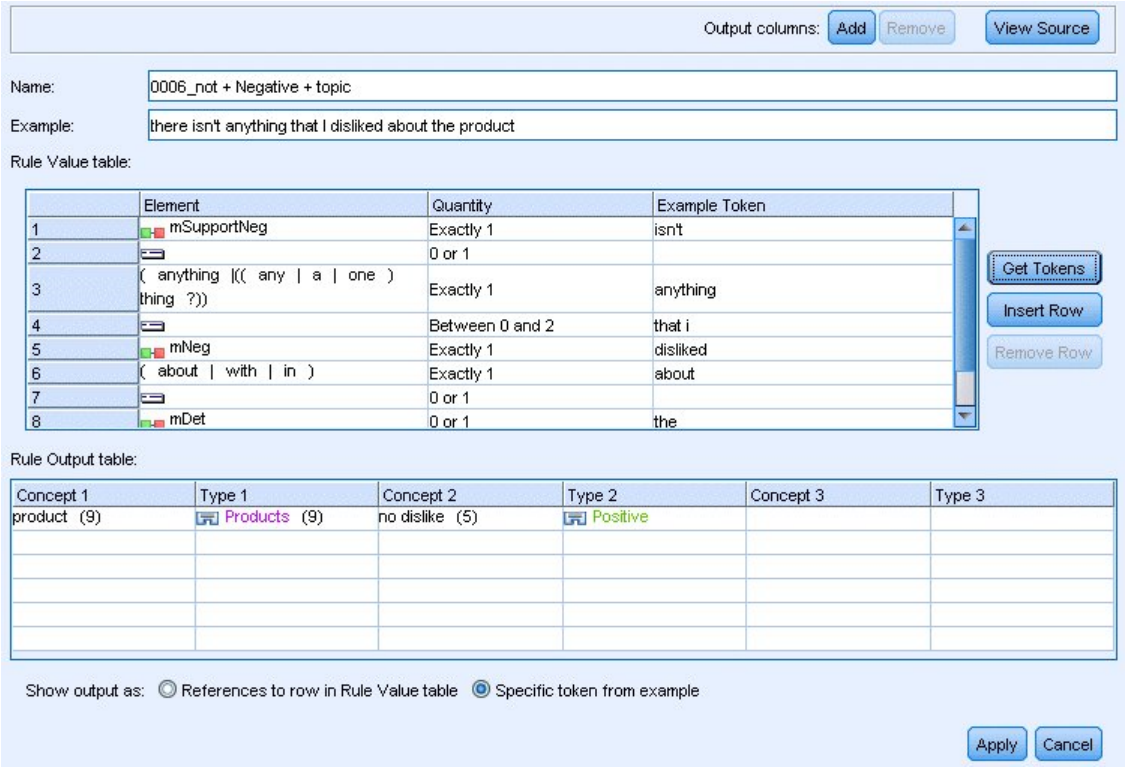


Figura 45. Guia Regras de Ligação de Texto: Editor de Regras

Sempre que você extrai, o mecanismo de extração lê cada sentença e tenta corresponder à seguinte sequência:

Tabela 46. Exemplo de sequência de extração

Elemento (linha)	**Descrição dos argumentos**
1	O conceito de um dos tipos representados pelas macros `mPos` ou `mNeg` ou a partir do tipo `<Uncertain>`.
2	Um conceito cujo tipo é um dos representados pela macro `mTopic`.
3	Uma das palavras representadas pela macro `mBe`.
4	Um elemento opcional, 0 ou 1 palavras, também referido como uma lacuna de palavras ou `<Any Token>`
5	Um conceito cujo tipo é um dos representados pela macro `mTopic`.

A tabela de saída mostra que tudo o que é desejado a partir desta regra é um padrão onde qualquer conceito ou tipo correspondente à macro `mTopic` que foi definido na linha 5 na **Regra de Valor da Regra** + qualquer conceito ou tipo correspondente ao `mPos`, `mNeg`, ou `<Uncertain>` como foi definido na linha 1 na tabela **Regra de Valor**. Isso pode ser `sausage + like` ou `<Unknown> + <Positive>`.

Criando e editando regras

É possível criar novas regras ou editar as existentes. Siga as diretrizes e as descrições para o editor de regra. Consulte o tópico "Trabalhando com regras de ligação de texto" na página 215 para obter informações adicionais.

Criando Novas Regras

1. A partir dos menus, escolha **Ferramentas> Nova Regra**. Alternativamente, clique no ícone Nova Regra na barra de ferramentas da árvore para abrir uma nova regra no editor.
2. Insira um nome exclusivo e defina os elementos de valor da regra.
3. Clique em **Aplicar** ao concluir a verificação de erros.

Editando regras

1. Clique no nome da regra na árvore. A regra é aberta na área de janela do editor à direita.
2. Faça as mudanças.
3. Clique em **Aplicar** ao concluir a verificação de erros.

Desativado e excluindo regras

Desativando regras

Se desejar que uma regra seja ignorada durante o processamento, é possível desativá-la. Tome cuidado ao excluir e desativar regras.

1. Clique no nome da regra na árvore. A regra é aberta na área de janela do editor à direita.
2. Clique com o botão direito no nome.
3. Nos menus de contexto, escolha **Desativar**. O ícone de regra fica cinza e a regra se torna não editável.

Excluindo regras

Se desejar se livrar de uma regra, é possível excluí-la. Tome cuidado ao excluir e desativar regras.

1. Clique no nome da regra na árvore. A regra é aberta na área de janela do editor à direita.
2. Clique com o botão direito no nome.
3. Nos menus de contexto, escolha **Excluir**. A regra desaparece da lista.

Verificando erros, salvando e cancelando

Aplicando mudanças de regra

Se você clicar fora do editor de regra ou em **Aplicar**, a regra será automaticamente varrida em busca de erros. Se um erro for localizado, você precisará corrigi-lo antes de mover para outra parte do aplicativo.

No entanto, se erros menos sérios forem detectados, somente um aviso será fornecido. Por exemplo, se sua regra contiver definições incompletas ou não referenciadas a tipos ou macros, uma mensagem de aviso será exibida. Após um clique em **Aplicar**, quaisquer avisos não corrigidos farão com que um ícone de aviso apareça à esquerda do nome da regra na árvore na área de janela esquerda.

A aplicação de uma regra não significa que sua regra esteja permanentemente salva. A aplicação fará com que o processo de validação verifique erros e avisos.

Salvando recursos dentro de uma sessão de ambiente de trabalho interativa

1. Para salvar as mudanças feitas em seus recursos durante uma sessão de ambiente de trabalho interativa para que seja possível obtê-las na próxima vez que você executar seu fluxo, deve-se:
   - Atualizar seu nó de modelagem para garantir que você obterá os mesmos recursos na próxima vez que executar seu fluxo. Consulte o tópico "Atualizando nós de modelagem e salvando" na página 75 para obter informações adicionais. Em seguida, salve seu fluxo. Para salvar seu fluxo, faça isso na janela principal do IBM SPSS Modeler após atualizar o nó de modelagem.

2. Para salvar as mudanças feitas em seus recursos durante uma sessão de ambiente de trabalho interativa para que seja possível usá-las em outros fluxos, é possível:
   - Atualizar o modelo que você usou ou criar um novo. Consulte o tópico “Criando e atualizando modelos” na página 162 para obter informações adicionais. Isso não salvará as mudanças no nó atual (consulte o passo anterior)
   - Ou atualizar a TAP usada. Consulte o tópico “Atualizando Pacotes de Análise de Texto” na página 133 para obter informações adicionais.

Salvando Recursos dentro do Editor de Template

1. Primeiro, publique a biblioteca. Consulte o tópico “Publicando bibliotecas” na página 181 para obter informações adicionais.
2. Em seguida, salve o modelo por meio do **Arquivo> Salvar Modelo de Recursos** nos menus.

Cancelando mudanças de regra

1. Se quiser descartar as mudanças, clique em **Cancelar** na área de janela do editor.

Ordem de processamento para regras

Quando a análise de ligação de texto for executada durante a extração, uma "sentença" (cláusula, palavra, frase) será correspondida com relação a cada regra sucessivamente, até que uma correspondência seja localizada ou até que todas as regras tenham sido esgotadas. O ranqueamento na árvore dita a ordem em que as regras são tentadas. A melhor prática diz que é necessário ordenar suas regras da mais específica para a mais genérica. As mais específicas devem estar na parte superior da árvore. Para mudar a ordem de uma regra ou conjunto de regras específicos, selecione **Mover para Cima** ou **Mover para Baixo**, a partir do menu de contexto Árvore de Regras e Macro ou das setas para cima e para baixo na barra de ferramentas.

Se você estiver *na visualização de origem*, não será possível mudar a ordem das regras movendo-as no editor. A regra que estiver mais no alto aparecerá na visualização de origem assim que for processada. É altamente recomendado reordenar somente regras na árvore para evitar problemas de copiar/colar.

Importante! Nas versões anteriores do IBM SPSS Modeler Text Analytics, era necessário ter um ID de regra numérico exclusivo. A partir da versão 18.5.0, só é possível indicar a ordem de processamento movendo uma regra para cima ou para baixo na árvore ou por seu ranqueamento na visualização de origem.

Por exemplo, suponha que seu texto contenha as duas sentenças a seguir:

Eu amo anchovas

Eu amo anchovas e pimentas verdes

Além disso, suponha que existam duas regras de análise de ligação de texto com os valores a seguir:

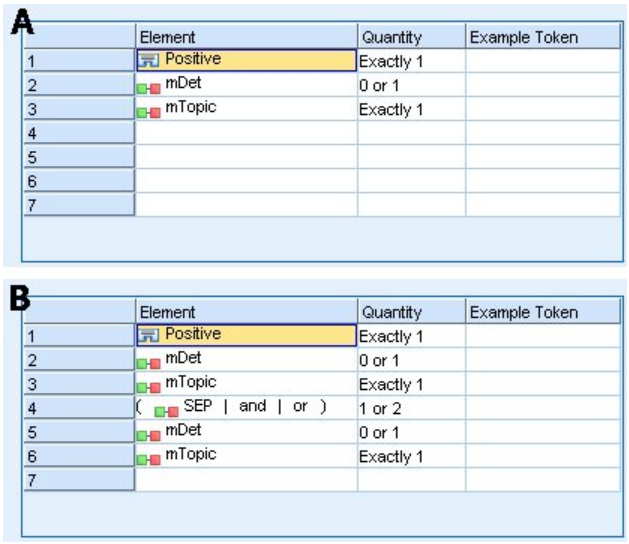


Figura 46. 2 Regras de Exemplo

Na visualização de origem, os valores de regra podem ser semelhantes aos seguintes:

A: `value = $Positive $mDet? $mTopic`

B: `value = $Positive $mDet? $mTopic ($SEP|and|or){1,2} $mDet? $mTopic`

Se a regra **A** estiver mais alta na árvore (mais perto da parte superior) do que a regra **B**, então, a regra **A** será processada primeiro e a frase *Eu amo anchovas e pimentas verdes* será primeiramente correspondido por `$Positive $mDet? $mTopic`, e produzirá uma saída padrão incompleta (`anchovies + like`) desde que foi correspondido por uma regra que não estava procurando 2 `$mTopic` correspondências.

Portanto, para capturar a verdadeira essência do texto, a regra mais específica, nesse caso, a **B**, deverá ser colocada mais no alto da árvore do que a mais genérica, nesse caso, a regra **A**.

Trabalhando com conjuntos de regras (Passagem Múltipla)

Um conjunto de regras é uma maneira útil de agrupar um conjunto relacionado de regras na Árvore de Regras e Macro para executar o processamento de passagem múltipla. Um conjunto de regras não tem uma definição própria além de um nome, e é usado para organizar suas regras em grupos significativos. Em alguns contextos, o texto é muito rico e variado para ser processado em uma única passagem. Por exemplo, durante o trabalho com dados de inteligência de segurança, o texto pode conter links entre indivíduos que são descobertos por meio de métodos de contato (*x chamou y*), por meio de relacionamentos familiares (*y cunhado de x*), por meio de troca de dinheiro (*x emprestou $100 para y*), e assim por diante. Nesse caso, é útil criar conjuntos especializados de regras de análise de ligação de texto, cada um focado em certo tipo de relacionamento, como para descobrir contatos, outro para descobrir membros da família, e assim por diante.

Para criar um conjunto de regras, selecione “Criar Conjunto de Regras” a partir do menu de contexto Árvore de Regras e Macro ou a partir da barra de ferramentas. É possível criar novas regras diretamente sob um nó Conjunto de Regras na árvore ou mover regras existentes para um Conjunto de Regras.

Quando você executa uma extração usando recursos nos quais as regras são agrupadas em conjuntos de regras, o mecanismo de extração é forçado a fazer passagens múltiplas por meio do texto para corresponder a diferentes tipos de padrões em casa passagem. Dessa maneira, uma "sentença" pode corresponder a uma regra em cada conjunto de regras, enquanto que, sem um conjunto de regras, ela só pode corresponder a uma única regra.

Nota: É possível incluir até 512 regras por conjunto de regras.

Criando novos conjuntos de regras

1. A partir dos menus, escolha **Ferramentas> Novo Conjunto de Regras**. Alternativamente, clique no ícone Novo Conjunto de Regras na barra de ferramentas da árvore. Um conjunto de regras aparece na árvore de regra.
2. Inclua novas regras nesse conjunto de regras ou mova regras existentes para o conjunto.

Desativando conjuntos de regras

1. Clique com o botão direito no nome do conjunto de regras na árvore.
2. Nos menus de contexto, escolha **Desativar**. O ícone do conjunto de regras fica cinza e todas as regras contidas nesse conjunto de regras também são desativadas e ignoradas durante o processamento.

Excluindo conjuntos de regras

1. Clique com o botão direito no nome do conjunto de regras na árvore.
2. Nos menus de contexto, escolha **Excluir**. O conjunto de regras e todas as regras que ele contém são excluídos dos recursos.

Elementos suportados para regras e macros

Os argumentos a seguir são aceitos para os parâmetros de valor nas macros e regras de análise de ligação de texto:

Macros

É possível usar uma macro diretamente em uma regra de análise de ligação de texto ou dentro de outra macro. Se você estiver inserindo o nome da macro manualmente ou de dentro de uma visualização de origem (ao contrário de selecionar o nome da macro de um menu de contexto), certifique-se de prefixar o nome com um caractere de símbolo de dólar ($), como `$mTopic`. O nome da macro faz distinção entre maiúsculas e minúsculas. É possível escolher a partir de qualquer macro definida na guia Regras de Ligação de Texto atual ao selecionar macros por meio de menus de contexto.

Tipos

É possível usar um tipo diretamente em uma macro ou regra de análise de ligação de texto. Se você estiver inserindo o nome do tipo manualmente ou na visualização de origem (ao contrário de selecionar o tipo de um menu de contexto), certifique-se de prefixar o nome do tipo com um caractere de símbolo de dólar ($), como `$Person`. O nome do tipo faz distinção entre maiúsculas e minúsculas. Se você usar os menus de contexto, é possível escolher entre qualquer um dos tipos no atual conjunto de recursos sendo usado.

Se fizer referência a um tipo não reconhecido, você receberá uma mensagem de aviso e a regra terá um ícone de aviso na Árvore de Regras e Macro até você corrigi-la.

Sequências de Caracteres Literais

Para incluir informações que nunca foram extraídas, é possível definir uma sequência de caracteres literal pela qual o mecanismo de extração procurará. Todas as frases ou palavras extraídas foram designadas a um tipo e, por esse motivo, elas não podem ser usadas em sequências de caracteres literais. Se você usar uma palavra que foi extraída, ela será ignorada, mesmo se seu tipo for `<Unknown>`.

Uma sequência de caracteres literal pode ter uma ou mais palavras. As regras a seguir se aplicam durante a definição de uma lista de sequências de caracteres literais:

- Coloque a lista de sequências de caracteres entre parênteses, como (his). Se houver uma opção de sequências de caracteres literais, cada sequência de caracteres deverá ser separada pelo operador OR, como `(a|an|the)` or `(his|hers|its)`.
- Use palavras simples ou compostas.
- Separe cada palavra na lista usando o caractere `|`, que é como um booleano OR.
- Insira as formas no singular e no plural se desejar ambas as correspondências. A flexão não é gerada automaticamente.
- Use somente letras minúsculas.

- Para reutilizar sequências de caracteres literais, defina-as como uma macro e use essa macro em suas outras macros e regras de análise de ligação de texto.
- Se uma sequência de caracteres contiver pontos (parada total) ou hifens, você deverá incluí-los. Por exemplo, para a correspondência de a.k.a no texto, insira os pontos junto com as letras a.k.a como a sequência de caracteres literal.

Operador de Exclusão

Use ! como operador de exclusão para impedir qualquer expressão da negação de ocupar um slot particular. Só é possível incluir um operador de exclusão manualmente por meio da edição de célula sequencial (dê um clique duplo na tabela Valor de Regra ou na tabela Valor de Macro) ou na visualização de origem. Por exemplo, se você adicionar $mTopic @{0,2} !($Positive) $Budget à sua regra de análise de link de texto, você está procurando por texto que contenha (1) um termo atribuído a qualquer um dos tipos na macro mTopic , (2) uma lacuna de palavra de zero a duas palavras longa, (3) nenhuma instâncias de um termo atribuído ao tipo <Positive> e (4) um termo atribuído ao tipo <Budget> . Isso pode capturar "*carros apresentam aumento de preço*", mas irá ignorar "*loja oferece descontos incríveis*".

Para usar esse operador, deve-se inserir o ponto de exclamação e os parênteses manualmente na célula do elemento dando um clique duplo na célula.

Word Lacunas (< Qualquer Token>)

Uma lacuna de palavras, também referida como <Any Token>, define uma gama numérica de tokens que podem estar presentes entre dois elementos. Diferenças de palavras são muito úteis quando correspondem a frases bastante similares que podem ser um pouco diferentes devido à presença de determinadores adicionais, fases preposicionais, adjetivos ou outras palavras.

Tabela 47. Exemplo dos elementos na tabela Valor de Regra sem uma diferença de palavra

#	Elemento
1	Unknown
2	mBeHave
3	Positive

Nota: Na visão de origem este valor é definido como: $Unknown $mBeHave $Positive

Esse valor corresponderá a sentenças como "*a equipe do hotel foi gentil*", em que *equipe do hotel* pertence ao tipo <Unknown>, *foi* está sob a macro mBeHave e *gentil* é <Positive>. Mas ele não corresponderá a "*a equipe do hotel foi muito gentil*".

Tabela 48. Exemplo dos elementos em uma tabela de Valor de Regra com uma lacuna de Token de < Qualquer Token>

#	Elemento
1	Unknown
2	mBeHave
3	
4	Positive

Nota: Na visão de origem este valor é definido como: `$Unknown $mBeHave @{0,1} $Positive`

Se você incluir uma diferença de palavras em seu valor de regra, ele corresponderá a "*a equipe do hotel foi gentil*" e a "*a equipe do hotel foi muito gentil*".

Na visualização de origem ou com a edição sequencial, a sintaxe para uma diferença de palavras é `@{#,#}`, em que `@` significa uma diferença de palavras e `{#,#}` define as palavras mínima e máxima aceitas entre o elemento precedente e o elemento seguinte. Por exemplo, `@{1,3}` significa que uma correspondência pode ser feita entre os dois elementos definidos, caso haja pelo menos uma palavra presente, mas não mais que três entre esses dois elementos. `@{0,3}` significa que uma correspondência pode ser feita entre os dois elementos definidos caso haja 0, 1, 2 ou 3 palavras presentes, mas não mais que três palavras.

Visualizando e trabalhando no modo de origem

Para cada regra e macro, o editor de TLA gera o código de origem subjacente que é usado pelo extrator para corresponder à e produzir a saída de TLA. Se você preferir trabalhar com o código em si, é possível visualizar esse código de origem e editá-lo diretamente clicando no botão "Visualizar Origem" na parte superior do Editor. A Visualização de Origem irá para e destacará a regra ou macro atualmente selecionada. Entretanto, é recomendado usar as áreas de janela do editor para reduzir a chance de erros.

Quando tiver concluído a visualização ou edição da origem, clique em **Origem de Saída**. Se você gerar uma sintaxe inválida para uma regra, será necessário corrigi-la antes de sair da visualização de origem.

Importante: Se você editar na visualização de origem, é altamente recomendado editar regras e macros uma por vez. Após editar uma macro, valide os resultados fazendo extração. Se você estiver satisfeito com o resultado, é recomendado salvar o modelo antes de fazer outra mudança. Se não estiver satisfeito com o resultado ou ocorrer um erro, reverta para seus recursos salvos.

Macros na visualização de origem

```
[macro]
name = macro_name
value = ([type_name|macro_name|literal_string|word_gap])
```

Tabela 49. Entradas de macro	
`[macro]`	Cada macro deve começar com a linha marcada [macro] para denotar o início de uma macro.
`name`	O nome da definição de macro. Cada nome deve ser exclusivo.
`value`	Uma combinação de um ou mais tipos, sequências de caracteres literais, diferenças de palavra ou macros. Consulte o tópico "Elementos suportados para regras e macros" na página 221 para obter mais informações. Durante a combinação de argumentos, deve-se usar parênteses ( ) para agrupar os argumentos e o caractere | para indicar um booleano OR.

Além das diretrizes e da sintaxe cobertas na seção sobre Macros, a visualização de origem tem algumas diretrizes adicionais não requeridas durante o trabalho na visualização do editor. Macros também devem respeitar o seguinte durante o trabalho no modo de origem:

- Cada macro deve começar com a linha marcada `[macro]` para denotar o início de uma macro.
- Para desativar um elemento, coloque um indicador de comentário (#) antes de cada linha.

Exemplo. Este exemplo define uma macro chamada mTopic. O valor para mTopic é a presença de um termo correspondente a *um* dos tipos a seguir: <Product>, <Person>, <Location>, <Organization>, <Budget> ou <Unknown>.

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

Regras na visualização de origem

```
[pattern(ID)]
name = pattern_name
value = [$type_name|macro_name|word_gaps|literal_strings]
output = $digit[\t]#digit[\t]$digit[\t]#digit[\t]$digit[\t]#digit[\t]
```

Tabela 50. Entradas de regra	
[pattern (<ID>)]	Indica o início de uma regra de análise de ligação de texto e fornece um ID numérico exclusivo usado para determinar a ordem de processamento.
name	Fornece um nome exclusivo para essa regra de análise de ligação de texto.
value	Fornece a sintaxe e os argumentos para corresponderem ao texto. Consulte o tópico "Elementos suportados para regras e macros" na página 221 para obter informações adicionais.
output	O formato de saída para os padrões correspondentes resultantes descobertos no texto. A saída nem sempre lembra o ranqueamento original exato dos elementos no texto de origem. Além disso, é possível ter diversas linhas de saída para uma determinada regra de análise de ligação de texto colocada cada saída em uma linha separada. Sintaxe para saída: • Separe a saída com o código da guia \t, como $1\t#1\t$3\t#3 • $ e um número requer o termo localizado correspondente ao argumento definido no parâmetro de valor nesse ranqueamento. Portanto, $1 significa o termo correspondente ao primeiro argumento definido para o valor. • # e um número chamadas para o nome do tipo do elemento nessa posição. Se um item for uma lista de sequências de caracteres literais, o tipo <Unknown> será designado. • Um valor de Null\tNull não criará nenhuma saída.

Além das diretrizes e da sintaxe cobertas na seção sobre Regras, a visualização de origem tem algumas diretrizes adicionais não requeridas durante o trabalho na visualização do editor. Regras também devem respeitar o seguinte durante o trabalho no modo de origem:

- Sempre que dois ou mais elementos são definidos, eles devem ser fechados entre parênteses se são ou não opcionais (por exemplo, ($Negative|$Positive) *ou* ($mCoord|$SEP)?). $SEP representa uma vírgula.
- O primeiro elemento em uma regra de análise de ligação de texto não pode ser um elemento opcional. Por exemplo, você não pode começar com value = $mTopic? ou value = @{0,1}.
- É possível associar uma quantidade (ou contagem de instâncias) a um token. Isso é útil durante a gravação de apenas uma regra que abrange todos os casos, em vez de gravar uma regra separada para cada caso. Por exemplo, você pode usar a sequência de caracteres literal ($SEP|and) se estiver tentando fazer a correspondência com , (vírgula) ou and. Se você estender isto incluindo uma quantidade para que a sequência de caracteres literal se torne ($SEP|and){1,2}, agora corresponderá qualquer uma das instâncias a seguir: "," "and" ", and".

- Os espaços não são suportados entre o nome macro e os caracteres $ e ? na regra de análise de link de texto `value`.
- Os espaços não são suportados na regra de análise de link de texto `output`.
- Para desativar um elemento, coloque um indicador de comentário (#) antes de cada linha.

Exemplo. Vamos supor que seus recursos contenham a regra de análise de ligação de texto TLA a seguir e você tenha ativado a extração dos resultados de TLA:

```
## Jean Doe was the former HR director of IBM in France
[pattern(201)]
name= 1_201
value = $Person ($SEP|$mDet|$mSupport|as|then){1,2} @{0,1} $Function
 (of|with|for|in|to|at) @{0,1} $Organization @{0,2} $Location
output = $1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

Sempre que você extrai, o mecanismo de extração lê cada sentença e tenta corresponder à seguinte sequência:

Tabela 51. Exemplo de sequência de extração

Posição	Descrição dos argumentos
1	O nome de uma pessoa ($Person),
2	Um ou dois do seguinte: vírgula ($SEP), determinador ($mDet), verbo auxiliar ($mSupport), sequências de caracteres “then” ou “as”,
3	0 ou 1 palavra (@{0,1})
4	Uma função ($Function)
5	Uma das sequências de caracteres a seguir: “of”, “with”, “for”, “in”, “to” ou “at”,
6	0 ou 1 palavra (@{0,1})
7	O nome de uma organização ($Organization)
8	0, 1 ou 2 palavras (@{0,2})
9	O nome de uma localização ($Location)

Esta regra de análise de ligação de texto de amostra corresponderia a sentenças ou frases como:

Jean Doe, diretor de RH da IBM na França

Jean Doe foi o primeiro diretor de RH da IBM na França

A IBM nomeou Jean Doe como diretor de RD da IBM na França

Essa regra de análise de ligação de texto de amostra produziria a seguinte saída:

```
jean doe <Person> hr director <Function> ibm <Organization> france <Location>
```

em que:

- `jean doe` é o termo correspondente a $1 (o primeiro elemento na regra de análise de link de texto) e <Person> é o tipo para `jean doe` (#1),
- `hr director` é o termo correspondente a $4 (o elemento 4th na regra de análise de link de texto) e <Function> é o tipo para `hr director` (#4),
- `ibm` é o termo correspondente a $7 (o elemento 7th na regra de análise de link de texto) e <Organization> é o tipo para `ibm`. (#7),
- `france` é o termo correspondente a $9 (o elemento 9th na regra de análise de link de texto) e <Location> é o tipo para `france` (#9)

Conjuntos de regras na visualização de origem

```
[set(<ID>)]
```

Onde `[set (<ID>)]` indica o início de um conjunto de regras e fornece um uso de ID numérico exclusivo para determinar a ordem de processamento dos conjuntos.

Exemplo. A sentença a seguir contém informações sobre indivíduos, sua função dentro de uma empresa e também as atividades fusão/aquisição dessa empresa.

```
Org1 Inc has entered into a definitive merger agreement with Org2 Ltd, said
John Doe, CEO of Org2 Ltd.
```

Você poderia gravar uma regra com várias saídas para manipular todas as saídas possíveis, como:

```
## Org1 Inc entered into a definitive merger agreement with Org2 Ltd, said
John Doe, CEO of Org2 Ltd.

[pattern(020)]
name=020
value = $Organization @{0,4} $ActionNouns @{0,6}  $mOrg @{1,2}
 $Person @{0,2} $Function @{0,1} $Organization
output = $1\t#1\t$3\t#3\t$5\t#5
output = $7\t#7\t$9\t#9\t$11\t#11
```

o que produziria os dois padrões de saída a seguir:

- `org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd<Organization>`
- `john doe <Person> + ceo <Function> + org2 ltd<Organization>`

Importante! Lembre-se de que outras operações de tratamento linguístico são executadas durante a extração de padrões de TLA. Neste caso, `merger` é agrupado sob `merges with` durante a fase de agrupamento de sinônimos do processo de extração. E já que `merges with` pertence ao tipo `<ActiveVerb>` , este nome de tipo é o que aparece na saída de padrão TLA final. Portanto, quando a saída ler `t$3\t#3`, isso significa que o padrão finalmente exibirá o conceito final para o terceiro elemento e o tipo final para o terceiro elemento após todo o processamento linguístico ser aplicado (sinônimos ou outros agrupamentos).

Em vez de gravar regras complexas como a anterior, talvez seja mais fácil gerenciar e trabalhar com duas regras. A primeira é especializada na descoberta de fusões/aquisições entre empresas:

```
[set(1)]
## Org1 Inc has entered into a definitive merger agreement with Org2 Ltd
[pattern(44)]
name=firm + action + firm_0044
value=$mOrg @{0,20} $ActionNouns @{0,6} $mOrg
output(1)=$1\t#1\t$3\t#3\t$5\t#5
```

que produziria `org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd <Organization>`

A segunda é especializada em indivíduos/funções/empresas:

```
[set(2)]
## said John Doe, CEO of Org2 Ltd
[pattern(52)]
name=individual + role + firm_0007
value=$Person @{0,3} $mFunction (at|of)? ($mOrg|$Media|$Unknown)
output(1)=$1\t#1\t$3\tFunction\t$5\t#5
```

que produziria `john doe <Person> + ceo <Function> + org2 ltd <Organization>`

Avisos

Estas informações foram desenvolvidas para produtos e serviços oferecidos em todo o mundo.

É possível que a IBM não ofereça os produtos, serviços ou recursos discutidos nesta publicação em outros países. Consulte seu representante IBM local para obter informações sobre os produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser utilizados. Qualquer produto, programa ou serviço funcionalmente equivalente que não infrinja nenhum direito de propriedade intelectual da IBM pode ser usado em substituição. Entretanto, a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade do Cliente.

A IBM pode ter patentes ou solicitações de patentes pendentes relativas a assuntos tratados nesta publicação. O fornecimento desta publicação não lhe garante direito algum sobre tais patentes. Pedidos de licença devem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
IBM Corporation
Botafogo
Rio de Janeiro, RJ
Brasil

Para pedidos de licença relacionados a informações de Conjunto de Caracteres de Byte Duplo (DBCS), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie pedidos de licença, por escrito, para:

Intellectual Property Licensing
IBM World Trade Asia Corporation Licensing
2-31 Roppongi 3-chome
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS NÃO SE LIMITANDO ÀS GARANTIAS IMPLÍCITAS DE MERCADO OU DE ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias expressas ou implícitas em certas transações; portanto, essa disposição pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar os produtos e/ou programas descritos nesta publicação, sem aviso prévio.

Referências nestas informações a Web sites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a esses websites. Os materiais contidos nesses websites não fazem parte dos materiais desse produto IBM e a utilização desses websites é de inteira responsabilidade do Cliente.

A IBM pode utilizar ou distribuir as informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o Cliente.

Licenciados deste programa que desejam obter informações sobre este assunto com objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) a utilização mútua das informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
IBM Corporation
Botafogo

Rio de Janeiro, RJ
Brasil

Tais informações podem estar disponíveis, sujeitas a termos e condições apropriadas, incluindo em alguns casos o pagamento de uma taxa.

O programa licenciado descrito nesta publicação e todo o material licenciado disponível são fornecidos pela IBM sob os termos do Contrato com o Cliente IBM, do Contrato Internacional de Licença do Programa IBM ou de qualquer outro contrato equivalente.

Os exemplos de clientes e dados de desempenho citados são apresentados com propósitos meramente ilustrativos. Os resultados reais de desempenho podem variar, dependendo das configurações e condições operacionais específicas.

As informações relativas a produtos não IBM foram obtidas junto aos fornecedores dos respectivos produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou estes produtos e não pode confirmar a precisão de seu desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Dúvidas sobre os recursos de produtos não IBM devem ser encaminhadas diretamente a seus fornecedores.

As declarações relacionadas aos objetivos e intenções futuras da IBM estão sujeitas a alterações ou cancelamento sem aviso prévio e representam apenas metas e objetivos.

Estas informações contêm exemplos de dados e relatórios utilizados nas operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de indivíduos, empresas, marcas e produtos. Todos estes nomes são fictícios e qualquer semelhança com nomes e endereços utilizados por uma empresa real é mera coincidência.

Marcas comerciais

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em várias jurisdições no mundo todo. Outros nomes de empresas, produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. Uma lista atual de marcas registradas da IBM está disponível na web em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml.

Intel, o logotipo Intel, Intel Inside, o logotipo Intel Inside, Intel Centrino, o logotipo do Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou suas subsidiárias nos Estados Unidos e em outros países.

Linux é uma marca registrada da Linus Torvalds nos Estados Unidos e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou em outros países.

UNIX é uma marca registrada do The Open Group nos Estados Unidos e/ou em outros países.

Java e todas as marcas comerciais e logotipos baseados em Java são marcas comerciais ou marcas registradas da Oracle e/ou de suas afiliadas.

Outros nomes de produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas.

Índice remissivo

Caracteres Especiais

A

B

C

D

E

F

G

H

I

L

M

N

O

P

R

S

T

U

V