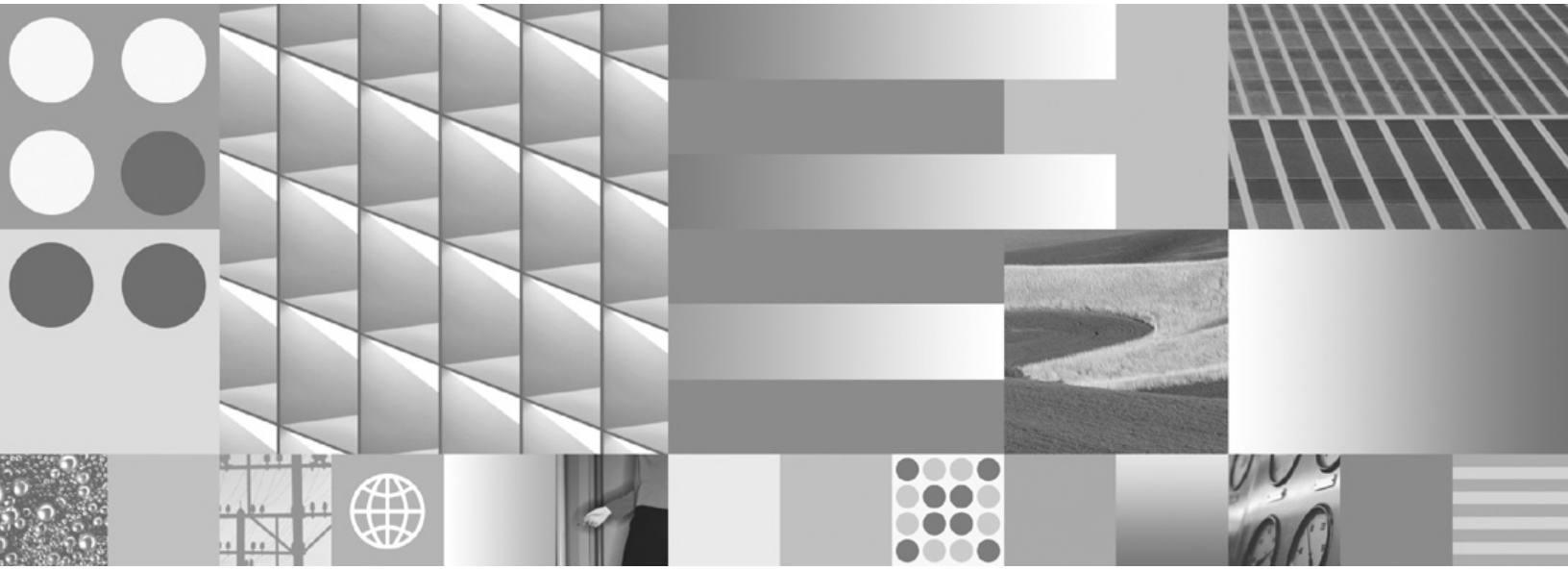
IBM OmniFind Enterprise Edition

Versão 8.4

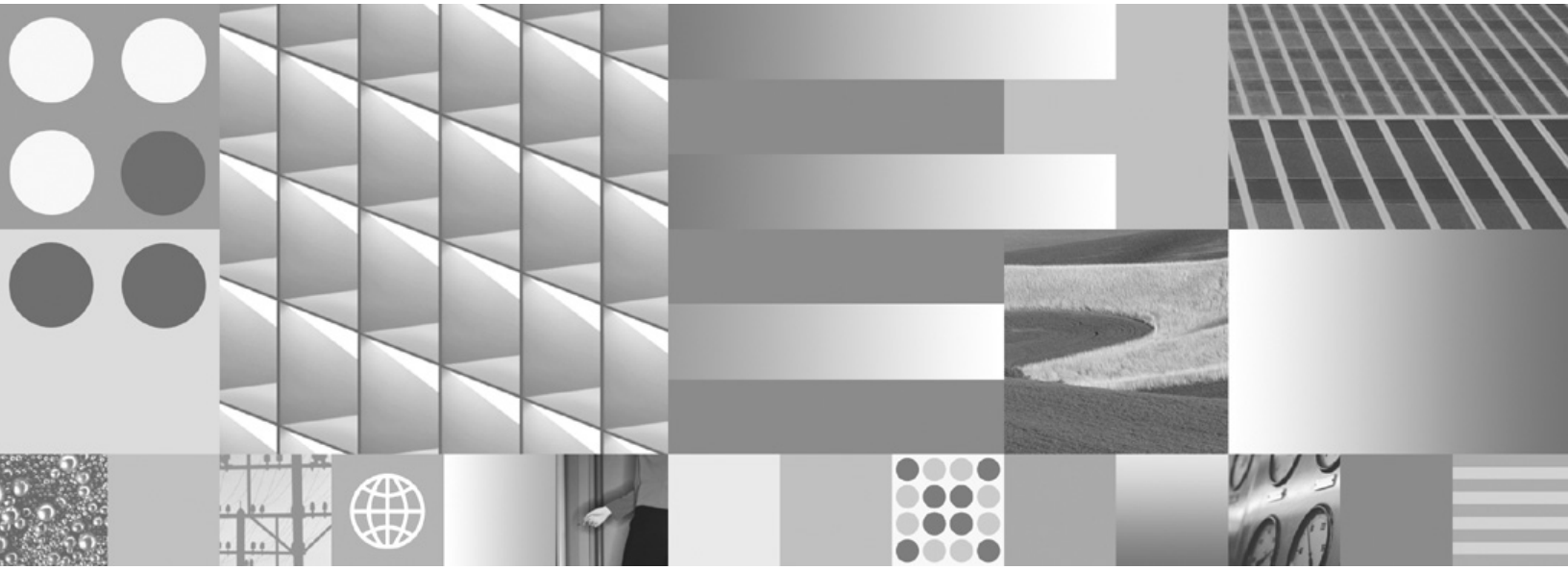


Integração da Análise de Texto

SC17-5463-01

IBM OmniFind Enterprise Edition

Versão 8.4



Integração da Análise de Texto

SC17-5463-01

Nota

Antes de utilizar as informações contidas nesta publicação, bem como o produto a que se referem, certifique-se de que lê as informações gerais incluídas na secção "Avisos e marcas comerciais" na página 119.

Segunda Edição (Novembro de 2006)

Pode encomendar publicações da IBM online ou através do representante IBM local:

- Para encomendar publicações online, avance para o IBM Publications Center em www.ibm.com/shop/publications/order.
- Para encontrar o representante IBM local, avance para o IBM Directory of Worldwide Contacts em www.ibm.com/planetwide.

Índice

Suporte linguístico para a procura semântica

O Enterprise Search oferece o suporte de procura linguístico para documentos de texto na maior parte dos idiomas indo-europeus e asiáticos, incluindo japonês.

Pode utilizar o suporte linguístico para melhorar a qualidade dos resultados da procura.

O processamento linguístico é executado em duas etapas: quando um documento é processado para ser adicionado ao índice e quando um utilizador introduz uma consulta de procura.

O Enterprise Search inclui apenas a funcionalidade linguística granular ou básica utilizada para determinar o idioma de um documento de entrada de dados e segmentar a sequência de entradas de documentos em palavras ou testemunhos.

Se o utilizador souber que as procuras serão restringidas principalmente a procuras por palavra-chave básicas ou procuras de de XML nativas que utilizam a estrutura de documentos, o processamento linguístico incluído no Enterprise Search abrange adequadamente as necessidades do utilizador.

A maior parte das informações nos documentos de texto é desestruturada, o que dificulta a utilização de forma eficiente porque não é fácil aceder ao significado das informações.

A procura de palavras-chave é simples, mas nem sempre é satisfatório se pretender procurar para além de meras palavras no documento, tal como ilustrado nos seguintes exemplos:

- Nos casos de colaboração, as informações não estão sempre explicitamente marcadas, por exemplo, um endereço ou um número de telefone numa mensagem de correio electrónico. Na realidade, o termo *número de telefone* pode até nem ser utilizado. Em vez disso, a mensagem de correio electrónico pode conter uma expressão tal como ″pode contactar-me pelo 219999999″. Frequentemente, o utilizador nem sempre sabe como as informações que pretende procurar existem no documento e, idealmente, pretenderia introduzir uma consulta como ″Número de telefone da Bárbara″ ao procurar o número de telefone de alguém que se chama Bárbara. No entanto, esta consulta não terá êxito, porque as palavras *número de telefone* não ocorrem no documento.
- Na inteligência competitiva, os documentos mencionam os concorrentes e os produtos que fornecem ou o sítio da Web do concorrente que mudou ao longo dos últimos três meses de venda de um conjunto de produtos para outro. Neste caso, o utilizador pode introduzir uma consulta como ″Produtos de Silva & C.ª″ ou ″Produtos de Silva & C.ª de Nov. de 2004 até Jan. de 2005″. Na primeira consulta, o termo *produtos* representa um produto ou um leque de produtos, mas a consulta não devolverá os produtos fornecidos pela empresa Silva & C.ª, uma vez que está a procurar o termo *produtos*. A amostra aplica-se à consulta que inclui um período de tempo específico. É quase impossível consultar um período de tempo utilizando a procura por palavra-chave.
- Na gestão de relações com clientes, os documentos podem mencionar problemas nos travões dos automóveis em oficinas de reparação na área do Porto. Os relatórios das oficinas de reparação descrevem situações tais como ″sapata ajustada devido a fuga no sistema hidráulico″. O utilizador que consulta mais

informações detalhadas pode introduzir uma consulta como ″oficinas de reparação de problemas nos travões a norte do Porto″. No entanto, esta consulta pode não devolver quaisquer relatórios que falam sobre ″sapata ajustada devido a fuga no sistema hidráulico″ porque os termos *problemas nos travões* ou *oficinas de reparação* não ocorrem nos relatórios. Além disso, estes relatórios podem mencionar apenas o nome da rua ou bairro da oficina de reparação, não o endereço completo incluindo o nome da cidade do Porto.

- Em investigação, os documentos descrevem um medicamento específico amplamente vendido através de várias marcas comerciais e a respectiva relação com pelo menos uma doença mencionada no mesmo parágrafo. O utilizador ocasional pode introduzir uma consulta utilizando um dos termos populares de um medicamento esperando um leque mais detalhado das várias doenças incluindo sintomas. No entanto, a consulta pode não devolver documentos satisfatórios porque o termo popular pode não ser utilizado nos documentos e, frequentemente, estes documentos nem sequer mencionam a palavra *doença*, apenas o nome da doença em si.

Nestes exemplos, a procura que o utilizador necessita de efectuar nas vastas colecções das origens de informações que existem hoje em dia apresenta novos desafios que requer uma análise sofisticada que ultrapasse o nível de segmentação e a análise com base nos dicionários que é oferecida no Enterprise Search. A maior parte das informações que são interessantes não estão explicitamente controladas nem marcadas de qualquer modo no documento original. Em vez disso, o conteúdo do documento tem de ser analisado para reconhecer e localizar conceitos de interesse, por exemplo, entidades nomeadas como pessoas, organizações, locais, instalações e produtos, e as possíveis relações entre estas entidades.

As informações que pretende identificar e extrair nos documentos de texto são específicos do utilizador e do domínio. Para ajudar a conceber os seus próprios algoritmos de análise, a IBM oferece o IBM Unstructured Information Management Architecture (UIMA), um contexto de software e arquitectura que ajuda a criar as funções de análise avançada para localizar as informações de interesse nas colecções de documentos no Enterprise Search.

Conceitos relacionados

“Integração da análise de texto personalizada” na página 3
Após ter criado a análise personalizada fora do Enterprise Search utilizando a Unstructured Information Management Architecture (UIMA), pode integrar a lógica da análise no Enterprise Search utilizando a consola de administração do Enterprise Search.

“Conceitos básicos utilizados no processamento de análise de texto” na página 4
Os conceitos básicos que são utilizados no processamento de análise de texto incluem anotadores, resultados de análise, estrutura funcional, tipo, tipo de sistema, estrutura de anotação e análise comum.

Integração da análise de texto personalizada

Após ter criado a análise personalizada fora do Enterprise Search utilizando a Unstructured Information Management Architecture (UIMA), pode integrar a lógica da análise no Enterprise Search utilizando a consola de administração do Enterprise Search.

A UIMA consiste numa plataforma aberta que identifica os componentes para cada função de análise conceptualmente distinta e garante que estes componentes possam ser facilmente reutilizados e combinados.

A análise linguística avançada pode incluir uma combinação de muitas e diferentes tarefas de análise. A análise começa com a detecção de idioma e segmentação e continua com o reconhecimento de parte do discurso, seguido pela análise gramatical aprofundada. As últimas tarefas incluem a identificação, por exemplo, da relação entre determinadas substâncias químicas e o aparecimento de sintomas específicos. Cada passo no processo de análise depende dos resultados no passo anterior.

A lógica da análise para cada passo encontra-se no *anotador (annotator)*. Os anotadores combinam para formar uma cadeia de processamento que itera através de cada documento na colecção para identificar novas informações e armazenar estas informações para o processamento na direcção do fluxo.

Os anotadores que são responsáveis pela identificação e representação do conteúdo da análise nos documentos de texto encontram-se num *motor de análise*, um conceito central em UIMA. Um motor de análise pode conter um único anotador ou pode ser um composto de muitos motores, cada um por sua vez contendo anotadores.

A UIMA fornece apenas os blocos de criação básicos para criar, testar e implementar os seus próprios motores de análise. Não fornece quaisquer funcionalidades de análise linguística sob a forma de motores de análise pré-configurados que pode implementar no ambiente de UIMA. No entanto, o processamento linguístico, que é aplicado no Enterprise Search está disponível como um conjunto de anotadores com o qual pode trabalhar na UIMA.

Para trabalhar com UIMA, tem de instalar o UIMA Software Development Kit. O kit de desenvolvimento está disponível no IBM developerWorks. Visite a zona do WebSphere Information Integrator para obter informações no sítio da Web http://www.ibm.com/developerworks/db2/zones/db2ii/. O UIMA Software Development Kit (SDK) inclui uma implementação Java do contexto UIMA para a entrada em vigor, descrição, composição e implementação de componentes UIMA.

O UIMA SDK fornece também um conjunto de ferramentas e utilitários para utilizar UIMA num ambiente de desenvolvimento baseado em Eclipse (suplementos Eclipse). Para obter informações sobre o Eclipse, consulte www.eclipse.org e a documentação de UIMA para obter instruções sobre como instalar o UIMA Software Development Kit no Ambiente de Desenvolvimento Interactivo do Eclipse.

Conceitos relacionados

"Suporte linguístico para a procura semântica" na página 1
O Enterprise Search oferece o suporte de procura linguístico para documentos de texto na maior parte dos idiomas indo-europeus e asiáticos, incluindo japonês.

"Conceitos básicos utilizados no processamento de análise de texto"
Os conceitos básicos que são utilizados no processamento de análise de texto incluem anotadores, resultados de análise, estrutura funcional, tipo, tipo de sistema, estrutura de anotação e análise comum.

Conceitos básicos utilizados no processamento de análise de texto

Os conceitos básicos que são utilizados no processamento de análise de texto incluem anotadores, resultados de análise, estrutura funcional, tipo, tipo de sistema, estrutura de anotação e análise comum.

Os *anotadores* contêm a lógica que analisa um documento e identifica os dados descritivos dos registos sobre o documento como um todo (referidos como metadados do documento) e componentes no documento. Estes dados descritivos são referidos como *resultados da análise*. Os resultados da análise anotam qualquer subcadeia contígua (também referida como grupo de recursos de rede) do documento de texto. Idealmente, os resultados da análise correspondem às informações que pretende procurar.

Uma *estrutura funcional* é a estrutura de dados subjacente que representa um resultado de análise. Uma estrutura funcional é uma estrutura atributo-valor. Cada estrutura funcional pertence a um *tipo (type)* e cada tipo tem um conjunto especificado de funcionalidades válidas ou atributos (propriedades), semelhante a uma classe de Java. As funcionalidades têm um tipo de intervalo a indicar o tipo de valor que a funcionalidade tem de ter, tal como, Cadeia.

Por exemplo, a expansão do texto ″José Mateus Bolota″ pode ser expandida por uma anotação do tipo `Pessoa` com as funcionalidades `nomePessoa`, `idade`, `nacionalidade` e `profissão`.

O *sistema de tipos* define os tipos de objectos (estruturas funcionais) que podem ser identificados num documento. O sistema de tipos define todas as estruturas funcionais possíveis nos termos de tipos e funcionalidades (atributos), semelhante a uma hierarquia de classes em Java. Pode definir qualquer número de tipos diferentes num sistema de tipos. Um sistema de tipos é específico de domínio e aplicação.

A maior parte dos anotadores de análise de texto produz os respectivos resultados de análise sob a forma de *anotações*. As anotações são um tipo especial de estrutura funcional designado para o processamento da análise linguística. Uma anotação expande ou abrange uma parte do texto de entrada e está definida nos termos do respectivo início e posições finais no texto de entrada.

Por exemplo, uma anotador que reconhece expressões monetárias cria para o texto ″100,55 Dólares norte-americanos″ uma anotação do tipo `monetaryExpression` que abrange o texto com a funcionalidade `currencySymbol` definida como ″$″.

Todos os anotadores no modelo UIMA armazenam os dados nas estruturas funcionais.

Todas as estruturas funcionais estão representadas numa estrutura de dados central denominada *estrutura de análise comum*. Toda a permuta de dados é processada utilizando a estrutura de análise comum.

A estrutura de análise comum contém os seguintes objectos:

- O documento de texto
- A descrição do sistema de tipos que indica os tipos, subtipos e respectivas funcionalidades
- Os resultados da análise que descrevem o documento ou regiões do documento
- Um repositório de índice que suporta o acesso e a iteração através dos resultados da análise

Conceitos relacionados

"Suporte linguístico para a procura semântica" na página 1
O Enterprise Search oferece o suporte de procura linguístico para documentos de texto na maior parte dos idiomas indo-europeus e asiáticos, incluindo japonês.

"Integração da análise de texto personalizada" na página 3
Após ter criado a análise personalizada fora do Enterprise Search utilizando a Unstructured Information Management Architecture (UIMA), pode integrar a lógica da análise no Enterprise Search utilizando a consola de administração do Enterprise Search.

Algoritmos de análise de texto

O UIMA Software Development Kit inclui APIs e ferramentas com as quais pode criar anotadores (algoritmos de análise incluindo o tipo de descrição do sistema) e incorporar estes anotadores nos motores de análise.

A documentação de UIMA inclui um guia de estilos de iniciação que ajuda a criar estes componentes. O kit de desenvolvimento de software inclui utilitários para testar e visualizar os resultados e um motor de procura semântica de pequena escala para indexar os resultados da análise. Pode também executar uma procura semântica mais avançada comparando com informações armazenadas no índice.

Uma vez que o UIMA Software Development Kit não fornece quaisquer anotadores pré-configurados e porque quaisquer anotadores personalizados que desenvolva utilizando UIMA e, em seguida, integre no Enterprise Search se fundamentam nos resultados dos anotadores base do Enterprise Search, pode utilizar o pacote anotador base no ambiente do UIMA. Consulte a documentação de UIMA para obter informações sobre como incluir a funcionalidade de detecção de idioma e de segmentação de testemunhos antes de executar os algoritmos de análise de texto personalizados no ambiente de UIMA.

Após desenvolver e testar os motores de análise utilizando o UIMA Software Development Kit, tem de criar um ficheiro PEAR (Processing Engine ARchive) para executar os algoritmos numa colecção de documentos do Enterprise Search. Este ficheiro de arquivo inclui todos os recursos requeridos para implementar a funcionalidade de análise personalizada como motores de análise no Enterprise Search. O modo como é descrito um arquivo na documentação de UIMA fornecida no Kit de Desenvolvimento de Software.

O arquivo criado para carregar no Enterprise Search tem apenas de conter a lógica de análise personalizada. Não pode conter qualquer dos anotadores base do Enterprise Search mesmo que a lógica de análise personalizada se fundamente nos

resultados do anotador base porque os anotadores base executam sempre antes de qualquer análise personalizada no Enterprise Search.

Para obter informações sobre como configurar e implementar uma solução de procura semântica no Enterprise Search, execute o guia de iniciação mencionado em http://www.ibm.com/developerworks/db2/zones/db2ii/. O guia de iniciação orienta-o nos passos necessários para a implementação de algoritmos de análise de texto personalizada no Enterprise Search e mostra-lhe como utilizar os resultados da análise nas consultas para melhorar os resultados da procura.

Tarefas relacionadas

"Utilizar os anotadores base do Enterprise Search em UIMA" na página 8
Pode utilizar os anotadores no pacote anotador base do Enterprise Search para desenvolver os novos anotadores no âmbito do UIMA Software Development Kit (SDK) e para mapear resultados da análise para tabelas JDBC.

Fluxo de trabalho para integração de análise personalizada

Crie e teste os algoritmos de análise de texto personalizados utilizando UIMA Software Development Kit e, em seguida, implemente e execute as colecções de documentos no Enterprise Search.

Para desenvolver algoritmos de análise e integrá-los no Enterprise Search:

1. Planear e estruturar:
   a. Determine quais as informações que pretende procurar. Quais são os documentos que pretender obter? Quais são os conceitos e relações necessários para cada tarefa de procura específica? Por exemplo, os nomes de produtos e empregados podem ser necessários para melhorar o objectivo geral das procuras num sítio da Web interno de uma empresa farmacêutica, enquanto que as pessoas na área da investigação e desenvolvimento necessitam de utilizar variantes de nomes de fármacos e consultar relações fármaco-causa-cura.
   b. Especifique o tipo de análise de texto de que necessita para obter as informações nos documentos que pretende procurar.
   c. Se a colecção contiver documentos XML, decida se pretende explorar a marcação XML na solução. No Enterprise Search, pode utilizar a marcação XML numa de duas formas:
      - Se for possível utilizar a marcação XML na análise personalizada (por exemplo, os documentos contiverem os elementos `<resumo>` ou `<tópico>` que podem ser úteis num anotador de categorização ou resumo), crie um ficheiro de mapeamento de elementos XML para a estrutura de análise comum.
      - Se pretender utilizar a marcação XML nas consultas conforme aparece no documento, tem de activar o mapeamento XML nativo.
   d. Determine quais as informações dos resultados da análise de texto armazenadas na estrutura de análise comum a que pretende ter acesso utilizando a procura semântica. Crie o ficheiro de mapeamento da estrutura de análise comum para o índice.
   e. Determine se pretende armazenar os resultados da análise numa base de dados relacional, por exemplo, para identificar tendências e associações utilizando a comunicação ou aplicações de exploração de dados. Crie o ficheiro de mapeamento da estrutura de análise comum para a base de dados.

f. Estruture a aplicação da procura semântica. Determine a utilização que o utilizador faz da procura de funções adicionais da procura semântica. Estruture a interface do utilizador.

2. Desenvolver: Actividades do UIMA Software Development Kit
   a. Defina os passos de análise individual.
   b. Descreva o sistema tipo dos mapeamentos e algoritmos de análise.
   c. Desenvolva os algoritmos de análise (anotadores) para cada passo da análise e incorpore os anotadores nos motores de análise utilizando o UIMA Software Development Kit. Crie qualquer análise personalizada utilizando a funcionalidade básica (identificação de idioma e segmentação) no pacote de anotadores base do Enterprise Search.
   d. Após testar os algoritmos de análise em UIMA, torne o motor de análise num ficheiro PEAR (Processing Engine Archive). O arquivo tem de conter apenas os algoritmos de análise e não a funcionalidade básica linguística do Enterprise Search.

      Quando concebe uma solução de análise de texto, pode incluir vários módulos de análise fornecidos em mais do que um ficheiro PEAR. A UIMA fornece um meio de intercalar dois ou mais ficheiros PEAR num único ficheiro PEAR que pode carregar e executar no Enterprise Search. A opção de intercalar ficheiros PEAR garante que não existem colisões de nomenclatura, que as funções de entrada e saída são intercaladas correctamente e que não existe substituição de parâmetros se os parâmetros intercalados nos descritores anotadores tiverem o mesmo nome. Consulte a documentação de UIMA para obter instruções sobre como intercalar ficheiros PEAR.

3. Implementar: Actividades do Enterprise Search
   a. Carregue o ficheiro de arquivo do motor de processamento (.pear) no Enterprise Search. Forneça um nome para o componente de análise de texto através do qual lhe possa fazer referência no Enterprise Search.
   b. Associe uma ou mais colecções de documentos com o componente de análise de texto.
   c. Se aplicável, para cada colecção, carregue e seleccione o mapeamento de elemento XML para a estrutura de análise comum que definiu para a análise personalizada.
   d. Se aplicável, para cada colecção, carregue e seleccione o mapeamento da estrutura de análise comum para a base de dados que definiu para a análise personalizada.
   e. Para cada colecção, carregue e seleccione o mapeamento da estrutura de análise comum para o índice que definiu para a procura semântica.
   f. Se necessário, configure a aplicação de procura semântica personalizada, por exemplo, implemente a interface do utilizador de procura baseada no navegador num servidor da aplicação.
   g. Pesquise, analise e indexe os documentos na colecção de procura semântica como faria para uma colecção baseada em palavras-chave.

Tarefas relacionadas

"Utilizar os anotadores base do Enterprise Search em UIMA" na página 8
Pode utilizar os anotadores no pacote anotador base do Enterprise Search para desenvolver os novos anotadores no âmbito do UIMA Software Development Kit (SDK) e para mapear resultados da análise para tabelas JDBC.

Utilizar os anotadores base do Enterprise Search em UIMA

Pode utilizar os anotadores no pacote anotador base do Enterprise Search para desenvolver os novos anotadores no âmbito do UIMA Software Development Kit (SDK) e para mapear resultados da análise para tabelas JDBC.

O conjunto dos anotadores base inclui:

- **Anotador do ID do idioma**

  Detecta o idioma de um documento. Para obter os parâmetros de configuração e capacidades, consulte o ficheiro descritor jlangid.xml.
- **anotador de procura do dicionário FROST**

  Fornece a segmentação e detecção de frases, com base nos dicionários do IBM LanguageWare. Para testemunhos, as informações linguísticas adicionais, por exemplo, o formulário base ou lema, são geradas. Para obter os parâmetros de configuração e capacidades, consulte o ficheiro descritor jfrost.xml.
- **Segmentador de espaços em branco**

  Pode executar a segmentação baseada em espaços em branco em todos os documentos de idiomas europeus ou outros scripts separados com espaços em branco. Para além disso, o anotador consegue efectuar a segmentação n-grama nos seguintes scripts de texto: árabe, han, hebraico, hiragana, katakana, laosiano, mongol, tailandês, YI e hangul. Esta lista inclui todos os scripts de texto asiático principais e significa que o anotador suporta japonês, chinês e coreano.

  Para obter os parâmetros de configuração e capacidades, consulte o ficheiro descritor jtok.xml.
- **Anotador de expressões globais**

  Detecta as entidades ou expansões de informações num documento de texto com base em expressões globais. Pode personalizar o anotador de expressões globais para detectar as entidades de texto de que necessita definindo as suas próprias regras. Um anotador de expressões globais de amostra que detecta números de telefone, URLs e endereços de correio electrónico nos documentos de texto está incluído no pacote anotador base.
- **Consumidor da estrutura de análise comum para a base de dados**

  O consumidor da estrutura de análise comum para a base de dados preenche uma base de dados relacional com resultados da análise de texto específicos.

O pacote anotador base do Enterprise Search é um ficheiro zipado que contém anotadores da análise de texto base com o anotador de expressões globais e o consumidor da estrutura de análise comum para a base de dados. O anotador de ID do idioma, o anotador de procura em dicionários FROST e o segmentador de espaços em branco são os anotadores da análise de texto base que executam sempre antes de qualquer análise de texto personalizada quando os documentos são analisados no Enterprise Search.

Uma vez que os anotadores da análise de texto base executam sempre antes de qualquer análise de texto personalizada no Enterprise Search e já que toda a análise de texto personalizada é baseada na saída de dados dos anotadores base, pode utilizar estes anotadores no ambiente de UIMA quando desenvolve e testa os anotadores personalizados.

O anotador de expressões globais e o consumidor da estrutura de análise comum para a base de dados são opções adicionais que pode seleccionar na consola de administração do Enterprise Search quando configura as opções de processamento de texto. Pode também utilizá-las em UIMA. Para a personalização avançada do

anotador de expressões globais, recomenda-se que utilize as ferramentas do UIMA SDK fornecidas para personalizar o anotador.

Para executar qualquer destes anotadores em UIMA, tem de ter o UIMA Software Development Kit (SDK) instalado. Está disponível no sítio da Web IBM developerWorks em http://www.ibm.com/developerworks/db2/zones/db2ii/.

Para instalar o pacote do anotador na instalação do UIMA SDK:

1. Localize o pacote anotador OF_base_annotators.zip na instalação do Enterprise Search (WebSphere Information Integrator OmniFind Edition) no directório *ES_INSTALL_ROOT*/packages/uima.
2. Copie o ficheiro zipado para o directório raiz da instalação UIMA SDK.
3. Extraia o ficheiro zipado para adicionar os ficheiros anotadores base do Enterprise Search à estrutura de directórios especificada da instalação do UIMA SDK. O ficheiro tt_core_typesystem.xml será sobreposto. Se pretender manter a versão antiga deste ficheiro, guarde-a antes de extrair o ficheiro zipado.
4. Para definir o caminho da classe, abra o script setUIMAClasspath no directório bin e adicione uma linha no final do script que inicie o script OFAnnotEnv.
5. Se pretender utilizar quaisquer tipos específicos do Enterprise Search ou personalizados em UIMA, consulte a documentação do UIMA SDK para obter informações sobre como os definir.

Após instalar o pacote anotador base, pode encontrar os ficheiros descritores do anotador no directório *UIMA_SDK_INSTALL*/docs/examples/descriptors/analysis_engine. O ficheiro of_tokenization.xml lista os anotadores da análise de texto base (o anotador de ID do idioma, o anotador de procura em dicionários FROST e o segmentador de espaços em branco) na sequência pela qual são utilizados no âmbito do Enterprise Search.

Os ficheiros descritores contêm os mesmos valores de configuração utilizados no Enterprise Search. Pode alterar os valores para fins de depuração no UIMA SDK. No entanto, não altera estes ficheiros descritores no sistema do Enterprise Search. Ao efectuar alterações nestes ficheiros poderá causar a instabilidade do sistema ou problemas no desempenho.

O pacote anotador base do Enterprise Search contém apenas os dicionários que são requeridos para processar documentos em inglês. Se pretender processar outros idiomas no ambiente de desenvolvimento, siga estes passos:

1. Localize os dicionários do Enterprise Search na respectiva instalação em *ES_INSTALL_ROOT*/configurations/parserservice/jediidata/frost/resources.
2. Copie o conteúdo do directório na instalação do UIMA SDK local em *UIMA_SDK_INSTALL*/data/frost/resources.

Para verificar se o pacote anotador foi instalado com êxito:

1. Abra o Visual Debugger (CVD) na Estrutura de Análise Comum (CAS, Common Analysis Structure) no seguinte directório: *UIMA_SDK_INSTALL*/bin/cvd[.bat/.sh].
2. Faça clique em **Executar (Run) → carregar TAE (load TAE)**.
3. Seleccione o ficheiro especificador do motor de análise de texto of_tokenization.xml no directório *UIMA_SDK_INSTALL*/docs/examples/descriptors/analysis_engine.
4. Carregue um documento amostra e execute o motor de análise de texto. Terá de visualizar as anotações do tipo `uima.tt.TokenAnnotation` no CVD.

Se executar qualquer dos anotadores de análise de texto base antes dos anotadores personalizados no ambiente de desenvolvimento e os anotadores personalizados utilizarem tipos definidos pela análise de texto base, incluem uma referência para o ficheiro `tt_core_typesystem` na secção do sistema tipo do especificador do anotador personalizado. O ficheiro `tt_core_typesystem` encontra-se no directório *UIMA_SDK_INSTALL*/docs/examples/descriptors/analysis_engine. Consulte o ficheiro jtok.xml no directório analysis_engine para obter uma exemplo de como incluir referências nos ficheiros descritores.

Tarefas relacionadas

"Visualizar o anotador base e os resultados da análise de texto personalizada" na página 13
Para visualizar os resultados da análise produzidos após a análise e por quaisquer anotadores no Enterprise Search, tem de actualizar as propriedades da colecção de documentos para produzir uma versão XML legível dos resultados da análise que são armazenados na estrutura de análise comum.

"Activar a procura semântica fácil utilizando o anotador de expressões globais" na página 85
Para activar a procura semântica fácil utilizando sinónimos, tem de adicionar o anotador de expressões globais, o ficheiro de mapeamento da estrutura de análise comum para o índice e o dicionário de sinónimos de amostra ao sistema Enterprise Search e associar estes recursos à colecção.

"Utilizar o consumidor de estrutura de análise comum para base de dados em UIMA"
Antes de poder utilizar o consumidor de estrutura de análise comum para base de dados em UIMA, tem de efectuar alterações no ficheiro descritor do consumidor e escrever o ficheiro de mapeamento da estrutura de análise comum para a base de dados.

"Utilizar o anotador de expressões globais em UIMA" na página 13
Utilize o anotador de expressões globais para detectar entidades ou unidades de informações num documento de texto. Pode personalizar o anotador para o domínio do assunto para cumprir as suas necessidades de procura.

Utilizar o consumidor de estrutura de análise comum para base de dados em UIMA

Antes de poder utilizar o consumidor de estrutura de análise comum para base de dados em UIMA, tem de efectuar alterações no ficheiro descritor do consumidor e escrever o ficheiro de mapeamento da estrutura de análise comum para a base de dados.

Antes de poder executar o consumidor de estrutura de análise comum para base de dados no ambiente de UIMA, é necessário efectuar os seguintes procedimentos:

1. Abra o ficheiro descritor XML cas2jdbc.xml em *UIMA_SDK_INSTALL*/docs/examples/descriptors/cas_consumer. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML à sua escolha.
2. Modifique o parâmetro **mappingFile** para incluir o caminho absoluto onde se localiza o ficheiro de mapeamento da estrutura de análise comum para a base de dados, por exemplo, D:\temp\MyMapping.xml
3. Modifique o parâmetro **docMetadata_Type** para especificar o tipo de UIMA a partir do qual os metadados para as funcionalidades foram obtidos, por exemplo, uima.tcas.DocumentAnnotation.
4. Modifique o parâmetro **docId_Feature** para incluir a funcionalidade ou caminho de funcionalidade no tipo de metadados a partir do qual o ID numérico de um documento (do tipo número inteiro) é obtido. Este processo é

requerido por todas as funcionalidades incorporadas que requerem o ID, tais como, `docId()`, `uniqueId()`, `objectId()` e `fsId()`.

5. Não defina o parâmetro **encryptionClass** uma vez que é utilizado apenas Enterprise Search para permitir que o consumidor de estrutura de análise comum para base de dados funcione com ficheiros de mapeamento codificado.
6. Guarde o ficheiro.
7. Copie os ficheiros de biblioteca EMF (common.jar, ecore.jar e ecore.xmi.jar) a partir do directório lib da instalação do Enterprise Search para o directório lib da instalação de UIMA. O ficheiro cc_cas2jdbc.jar já se encontra no directório lib da instalação de UIMA.
8. Crie o ficheiro de mapeamento da estrutura de análise comum para a base de dados que define quais os resultados da análise de texto a armazenar numa base de dados. Pode utilizar o ficheiro de mapeamento sampleMapping.xml em *UIMA_SDK_INSTALL*/docs/examples/descriptors/cas_consumer como uma amostra para criar o seu próprio ficheiro de mapeamento.

   Utilize o ficheiro de esquema XML denominado CasToJDBCMapping.xsd em *UIMA_SDK_INSTALL*/docs/examples/descriptors/cas_consumer para validar o ficheiro de mapeamento da estrutura de análise comum para a base de dados. Por motivos de desempenho, o consumidor de estrutura de análise comum para base de dados não valida o ficheiro de mapeamento, tem de ser o utilizador a fazê-lo.

O modo como executar o consumidor em UIMA encontra-se descrito na documentação de UIMA.

A seguinte amostra mostra como os parâmetros obrigatórios tem de estar definidos no descritor:

```
 ...
 <nameValuePair>
<name>mappingFile</name>
  <value>
<string>D:/temp/MyMapping.xml</string>
  </value>
 </nameValuePair>
<nameValuePair>
<name>docMetadata_Type</name>
  <value>
<string>uima.tcas.DocumentAnnotation</string>
  </value>
 </nameValuePair>
<nameValuePair>
<name>docId_Feature</name>
  <value>
<string>end</string>
  </value>
 </nameValuePair>

...
```

A tabela mostra os parâmetros de configuração pela ordem de apresentação no ficheiro descritor e indica quais os que são obrigatórios:

Tabela 1. Parâmetros de configuração no ficheiro descritor do consumidor de estrutura de análise comum para a base de dados

Parâmetro	Descrição	Obrigatório
mappingFile	O caminho absoluto para o ficheiro de mapeamento da estrutura de análise comum para a base de dados, por exemplo, D:/temp/sample.xml. Nos sistemas Windows, utilize "/" como separador de caminho.	verdadeiro
encryptionClass	Não defina este parâmetro, só é utilizado no Enterprise Search para permitir o consumidor de estrutura de análise comum para a base de dados para trabalhar com ficheiros de mapeamento codificados.	falso
docMetadata_Type	O tipo UIMA a partir do qual os metadados para funcionalidades incorporadas são obtidos.	verdadeiro
docId_Feature	A funcionalidade ou caminho da funcionalidade no tipo de metadados a partir do qual o ID numérico do documento é obtido. Tem de ser do tipo número inteiro (integer) e é necessário para todas as funcionalidades incorporadas que requerem o ID, tais como, `uniqeId()`, `objectId()` e `fsId()`.	verdadeiro
docUri_Feature	A funcionalidade ou caminho da funcionalidade no tipo de metadados de onde provém o URI do documento. Tem de ser do tipo cadeia.	falso
IsCompleted_Feature	A funcionalidade ou caminho da funcionalidade no tipo de metadados que sinaliza se o documento actual está dividido em várias estruturas de análise comum.	falso
chunkNumber_Feature	A funcionalidade ou caminho da funcionalidade no tipo de metadados que assinala o número subsequente da parte actual.	falso

Utilizar o anotador de expressões globais em UIMA

Utilize o anotador de expressões globais para detectar entidades ou unidades de informações num documento de texto. Pode personalizar o anotador para o domínio do assunto para cumprir as suas necessidades de procura.

Para executar o anotador de expressões globais de amostra que detecta os números de telefone, URLs e endereços de correio electrónico ou utilizar o anotador de amostra como base para criar a sua própria versão personalizada do anotador de expressões globais no ambiente de UIMA, é necessário:

1. O descritor do anotador de expressões globais no directório *UIMA_SDK_INSTALL*/docs/examples/descriptors/analysis_engine.
2. O conjunto de regras de amostra e a descrição do sistema tipo no directório *UIMA_SDK_INSTALL*/docs/examples/regex.
3. Um ficheiro de texto exemplo que o conjunto de regras de amostra pode ser aplicado, no directório *UIMA_SDK_INSTALL*/docs/data denominado of_sample_regex.txt.

O modo como executar o anotador em UIMA encontra-se descrito na documentação de UIMA.

Visualizar o anotador base e os resultados da análise de texto personalizada

Para visualizar os resultados da análise produzidos após a análise e por quaisquer anotadores no Enterprise Search, tem de actualizar as propriedades da colecção de documentos para produzir uma versão XML legível dos resultados da análise que são armazenados na estrutura de análise comum.

Acerca desta tarefa

Utilize a serialização XML dos resultados da análise do anotador armazenados na estrutura de análise comum para:

- Visualize os resultados após a análise, antes dos anotadores base serem processados.
- Visualize os resultados após a análise e segmentação (a executar os anotadores base do Enterprise Search). Este processo pode ajudá-lo a determinar as estruturas de dados de entrada para qualquer análise personalizada que pretenda desenvolver e que executará sempre após os anotadores base.
- Visualize e valide os resultados de uma análise personalizada executada numa pequena colecção de documentos no Enterprise Search com a finalidade de efectuar testes antes de decidir executar a análise numa colecção completa.

A serialização XML produz dois conjuntos de resultados:

- Os resultados após a análise. Estes incluem mapeamentos de campos e metadados de documentos.
- Os resultados após a análise e segmentação e, se estiver seleccionada, a análise de texto personalizada. Estes incluem todos os testemunhos e anotações produzidos.

Procedimento

Para produzir uma versão XML legível dos resultados da análise:

1. Abra o ficheiro collection.properties em *ES_NODE_ROOT*/master_config/<CollectionID>.parserdriver antes de começar a analisar os documentos na sua colecção.
2. Para visualizar os resultados após a análise, adicione a seguinte linha ao ficheiro collection.properties:
   `trevi.parser.dumpXCas=<o_directório_cópia_de_memória>`

   É necessário que já exista o directório da cópia de memória.

   a. Seleccione o tipo de saída que pretende. A saída inclui sempre a descrição do sistema tipo utilizada para analisar os resultados denominada OmniFindParserTypeSystem.xml. Adicione uma das seguintes linhas:

      - Para visualizar a saída dos últimos 25 ficheiros processados, adicione `trevi.parser.maxXCasFileCount=25`.

        O próprio utilizador pode determinar o número de ficheiros, mas recomenda-se que não defina este valor muito elevado.

        Lembre-se que a memória tampão de saída do ficheiro é constantemente substituída após ser alcançado o tamanho da memória tampão máximo. Este procedimento também implica que o documento com o número mais elevado não necessita de ser o último processado.

        A saída inclui os seguintes ficheiros: OmniFindParserXCasDump1.xml seguido de OmniFindParserXCasDump2.xml, etc., até serem listados 25 ficheiros.

      - Para visualizar a saída de documentos específicos, adicione o URI do documento `trevi.parser.xCasURI.1=ficheiro://home/test/ficheiro1.txt`.

        Pode adicionar qualquer número de documentos, no entanto, os documentos têm de estar numerados por ordem crescente começando em 1 sem intervalos entre os números. Por exemplo, o segundo documento seria `trevi.parser.xCasURI.2=ficheiro://home/test/ficheiro2.txt` e o terceiro `trevi.parser.xCasURI.3=ficheiro://home/test/ficheiro3.txt`

        A saída inclui os seguintes ficheiros: OmniFindParserXCasDumpURI_1.xml, OmniFindParserXCasDumpURI_2.xml e assim sucessivamente para todos os nomes de ficheiros que foram listados

3. Para visualizar os resultados após a segmentação, adicione a seguinte linha:
   `trevi.tokenizer.dumpXCas=<o_directório_cópia_de_memória>`

   Novamente, é necessário que já exista o directório da cópia de memória.

   a. Seleccione o tipo de saída que pretende. A saída criada também inclui sempre a descrição do sistema tipo utilizada para a segmentação e para os resultados da análise de texto, denominada OmniFindTypeSystem.xml. Adicione uma das seguintes linhas:

      - Para visualizar a saída dos últimos 25 ficheiros processados, adicione `trevi.tokenizer.maxXCasFileCount=25`.

        O próprio utilizador pode determinar o número de ficheiros, mas recomenda-se que não defina este valor muito elevado.

        Lembre-se que a memória tampão de saída do ficheiro é constantemente substituída após ser alcançado o tamanho da memória tampão máximo. Este procedimento também implica que o documento com o número mais elevado não necessita de ser o último processado.

        A saída inclui os seguintes ficheiros: OmniFindXCasDump1.xml, OmniFindXCasDump2.xml, etc., até serem listados 25 ficheiros.

- Para visualizar a saída de documentos específicos, adicione o URI do documento `trevi.tokenizer.xCasURI.1=ficheiro://home/test/ficheiro1.txt`.

  Pode adicionar qualquer número de documentos, no entanto, os documentos têm de estar numerados por ordem crescente começando em 1 sem intervalos entre os números. Por exemplo, o segundo documento seria `trevi.tokenizer.xCasURI.2=ficheiro://home/test/ficheiro2.txt` e o terceiro `trevi.tokenizer.xCasURI.3=ficheiro://home/test/ficheiro3.txt`

  A saída inclui os seguintes ficheiros: OmniFindXCasDumpURI_1.xml, OmniFindXCasDumpURI_2.xml e assim sucessivamente para todos os nomes de ficheiros que foram listados

No Enterprise Search, pode utilizar o Visualizador de Anotação XCAS (XCAS Annotation Viewer) para visualizar o conteúdo dos ficheiros XML. Inicie o Visualizador de Anotação XCAS executando o ficheiro de script `xcasAnnotationViewer` localizado no directório ES_INSTALL_ROOT/bin. Surge um pedido de informação a pedir:

- O directório da cópia de memória onde os resultados são colocados após a análise ou segmentação
- O ficheiro descritor, OmniFindParserTypeSystem.xml (para resultados do analisador) ou OmniFindTypeSystem.xml (para resultados da segmentação e da análise), como no directório da cópia da memória.

Ao seleccionar um documento da lista serão apresentados os resultados da análise para o documento. Ao clicar numa anotação evidenciada no documento são apresentados os detalhes da anotação.

Descrição do sistema tipo

O sistema tipo define os tipos de objectos e respectivas propriedades (ou funcionalidades) que podem ser instanciadas numa estrutura de análise comum.

Cada motor de análise tem as suas próprias descrições do sistema tipo que descrevem os requisitos de entrada e tipos de saída para os anotadores no motor de análise. As descrições do sistema tipo são específicas do domínio de aplicação.

Os sistemas tipo incluem as definições dos tipos, respectivas propriedades e hierarquia por herança simples dos tipos. Uma estrutura de análise comum tem de estar em conformidade com determinado sistema tipo.

Os tipos e funcionalidades que são definidos na descrição do sistema tipo têm também de ser utilizados em todos os ficheiros de mapeamento que estão associados à análise do documento, incluindo o ficheiro de mapeamento de elementos XML para a estrutura de análise comum, o ficheiro de mapeamento da estrutura de análise comum para o índice e o ficheiro de mapeamento da estrutura de análise comum para a base de dados.

A descrição do sistema tipo de um anotador pode fazer parte do descritor do anotador ou pode estar contido num ficheiro descritor de sistema tipo separado. Por vezes, faz parte do descritor de outro anotador contido no mesmo motor de análise.

Quando tiver concluído o desenvolvimento e testes do motor de análise no ambiente de UIMA, o ficheiro de arquivo (ficheiro .pear) que o utilizador criou e carregou para o Enterprise Search contém os ficheiros lógicos de análise bem como a descrição do sistema tipo.

Os anotadores base do Enterprise Search utilizam três descrições do sistema tipo; uma descrição do sistema tipo de núcleo que está sempre incluída e duas outras que pode activar opcionalmente para alterar o processamento da análise base da colecção de documentos para o modo de análise avançada. A necessidade de incluir uma ou ambas as descrições do sistema tipo expandidas depende dos resultados do processamento da análise de texto adicionais que pretender incluir durante o processamento da análise base.

Pode activar o modo de análise avançada incluindo um ou ambos os sistemas tipo de extensão. No modo de análise avançada, as funcionalidades de análise adicionais são disponibilizadas durante o processamento da análise base e são guardadas na estrutura de análise comum. Por exemplo, se requerer mais informações sobre um testemunho (mais informações sobre a funcionalidade), tais como, todos os lemas possíveis para o testemunho ou se o lema for uma palavra de paragem ou parte do discurso do lema, ou funcionalidades especiais para o processamento morfológico, também para japonês, necessita de activar o modo de análise avançada.

Tarefas relacionadas

"Mudar do modo de análise base para o modo de análise avançada"
Para alterar o processamento da colecção de documentos que é executado pelos anotadores base do Enterprise Search a partir do modo de análise base para o modo de análise avançada, tem de incluir as descrições do sistema tipo para o modo de análise avançada.

Referências relacionadas

"Tipos e funcionalidades definidos no Enterprise Search" na página 17
O sistema tipo definido no Enterprise Search abrange o processamento de metadados do documento e análise linguística básica.

Mudar do modo de análise base para o modo de análise avançada

Para alterar o processamento da colecção de documentos que é executado pelos anotadores base do Enterprise Search a partir do modo de análise base para o modo de análise avançada, tem de incluir as descrições do sistema tipo para o modo de análise avançada.

Restrições

Existem duas descrições do sistema tipo que pode seleccionar para activar o modo de análise avançada:

- A descrição `tt_extension_typesystem`, que inclui mais informações de funcionalidade de tipo lexical detalhadas sobre lemas.
- A descrição`dlt_extension_typesystem`, que inclui funcionalidades morfológicas adicionais e tipos lexicais especiais.

Procedimento

Para mudar o processamento de colecção base para o modo de análise avançada:

1. Abra o ficheiro tt_core_typesystem.xml no directório *ES_NODE_ROOT*/master_config/*IDColecção*.parserdriver/specifiers. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML à sua escolha.
2. Remova os controlos de comentário que rodeiam o elemento <import> na secção <imports> para incluir um ou ambos os ficheiros de descrição do sistema tipo.

   ```
   <imports>
   <!-- importa tt_extension_typsystem para a análise avançada -->
   <!-- <import location="tt_extension_typesystem.xml"/>-->
   <!-- importa o sistema tipo com a extensão dlt -->
   <!-- <import location="dlt_extension_typesystem.xml"/> -->
   </imports>
   ```
3. Abra os dois ficheiros descritores jfrost.xml e jfrost_ngram.xml e modifique o conteúdo do elemento <outputs> para incluir os tipos (num elemento <type>) e funcionalidades (num elemento <feature>) listados no elemento <description> na secção <capabilities> que pretende incluir durante a análise. Guarde as alterações.
4. Abra o ficheiro descritor jtok.xml e modifique o conteúdo do elemento <outputs> para incluir as funcionalidades (num elemento <feature>) listadas no elemento <description> na secção <capabilities> que pretende incluir durante a análise. Guarde as alterações.
5. Abra o ficheiro descritor es_tok_no_stw.xml e, também neste caso, modifique o conteúdo do elemento <outputs> para incluir as funcionalidades (num elemento <feature>) listadas no elemento <description> na secção <capabilities> que pretende incluir durante a análise. Guarde as alterações.
6. Quando mudar para o modo de análise avançada, tem de analisar novamente a colecção de documentos.

   Conceitos relacionados

   "Descrição do sistema tipo" na página 15
   O sistema tipo define os tipos de objectos e respectivas propriedades (ou funcionalidades) que podem ser instanciadas numa estrutura de análise comum.

   Referências relacionadas

   "Tipos e funcionalidades definidos no Enterprise Search"
   O sistema tipo definido no Enterprise Search abrange o processamento de metadados do documento e análise linguística básica.

Tipos e funcionalidades definidos no Enterprise Search

O sistema tipo definido no Enterprise Search abrange o processamento de metadados do documento e análise linguística básica.

Os tipos utilizados no Enterprise Search são definidos em três ficheiros de descrição do sistema tipo separados, começando pelo ficheiro de descrição do sistema tipo que contém os tipos núcleo sempre requeridos para toda a análise linguística básica e continua com as descrições do sistema tipo que definem as funcionalidades linguísticas avançadas que são, normalmente, apenas requeridas no modo de análise avançada.

A análise linguística básica sob a forma de reconhecimento e segmentação do idioma do documento é executada quando um documento é indexado, independentemente da análise personalizada estar ou não seleccionada. Durante a análise de documentos básica, a descrição tt_core_typesystem é utilizada e são

adicionadas as seguintes informações à estrutura de análise comum que pode utilizar na análise personalizada subsequente:

- Os metadados do documento do tipo `com.ibm.es.tt.DocumentMetaData`.
- As informações da estrutura do documento tais como anotações de frase e parágrafo do tipo `uima.tt.SentenceAnnotation` e `uima.tt.ParagraphAnnotation`.
- As anotações lexicais tais como testemunhos e compostos do tipo `uima.tt.TokenAnnotation`.

A descrição `tt_core_typesystem` é adequada para a maior parte do processamento da análise de texto.

Se pretender alterar o processamento de colecções para o modo de análise avançada, pode incluir os seguintes dois sistemas tipo. Os sistemas tipo incluem, principalmente, as funcionalidades que não são criadas durante o processamento linguístico básico.

- `tt_extension_typesystem` que inclui mais informações de funcionalidades sobre testemunho, lema, parágrafo e frase
- `dlt_core_typesystem` que contém alguns dos tipos de anotação expandida do IBM LanguageWare, por exemplo, URLs e endereços. Também inclui funcionalidades morfológicas que não são utilizadas frequentemente.

tt_core_typesystem

Os seguintes tipos e funcionalidades são definidos na descrição de `tt_core_typesystem`:

uima.tcas.DocumentAnnotation
A anotação do documento contém metadados do documento e tem a seguinte funcionalidade:

- `categories` com categorias de documentos adicionadas por um utilitário de categorização de texto. Cada categoria adicionada é do tipo `com.tt.CategoryConfidencePair`
- `languageCandidates` com os idiomas de documento detectados automaticamente durante a análise. Os idiomas são adicionados a uma lista do tipo `com.tt.LanguageConfidencePair`, com o idioma mais provável listado em primeiro lugar
- `id` com o ID de documento, tal como o URL

uima.tt.TTAnnotation
Este o tipo de raiz para anotações definidas em `tt_core_typesystem`. O respectivo supertipo é `uima.tcase.Annotation`. Tem os seguintes tipos:

uima.tt.DocStructureAnnotation
As anotações sobre a estrutura do documento. Tem os seguintes subtipos:

uima.tt.SentenceAnnotation
Frases

uima.tt.ParagraphAnnotation
Parágrafo do documento

uima.tt.LexicalAnnotation
As anotações lexicais tais como testemunhos e expressões de várias palavras. Tem os seguintes subtipos:

uima.tt.TokenLikeAnnotation
As anotações de testemunho único que podem ter as seguintes funcionalidades:

- `tokenProperties` com as propriedades do testemunho
- `lemma` com o lema ou raiz do termo
- `normalizedCoveredText` com a representação normalizada do texto abrangido

Este tipo de anotação tem os seguintes subtipos:

uima.tt.TokenAnnotation
Os testemunhos reais a serem distinguidos dos componentes comuns.

uima.tt.CompPartAnnotation
Os componentes compostos de um termo.

uima.tt.CompoundAnnotation
A anotação de um testemunho composto. Normalmente, o testemunho composto expande mais do que uma anotação do testemunho.

uima.tt.MultiTokenAnnotation
A anotação lexical consistindo em mais do que um testemunho. Este tipo de anotação tem os seguintes subtipos:

uima.tt.StopwordAnnotation
As anotações das palavras de paragem. As palavras de paragem podem também ser as palavras de vários termos.

uima.tt.SynonymAnnotation
A anotação de um termo para o qual existem sinónimos. Tem a funcionalidade `synonyms` que lista os sinónimos encontrados para o termo.

uima.tt.SpellCorrectionAnnotation
A anotação de um termo para o qual existem correcções de ortografia. Tem a funcionalidade `correctionTerms` que lista as correcções prováveis numa ordem começando pelas correcções mais prováveis.

uima.tt.MultiWordAnnotation
A anotação de um termo de várias palavras.

uima.CAS.TOP
A raiz do sistema tipo. Tem os seguintes subtipos:

uima.tt.KeyStringEntry
O tipo abstracto de estruturas de dados da Cadeia (String). Inclui a funcionalidade `key` que contém a chave de cadeia e o seguinte subtipo:

uima.tt.Lemma
Entradas de lemas do dicionário.

uima.tt.CategoryConfidencePair
O valor de fiabilidade para a categoria encontrada. Tem as seguintes funcionalidades:

- categoryString com o nome da categoria
- categoryConfidence com o valor de fiabilidade para a categoria
- mostSpecific com o sinalizador a indicar se esta categoria é a mais específica para o documento
- taxonomy com o nome da taxonomia de onde deriva a categoria

uima.tt.LanguageConfidencePair
O valor de fiabilidade para a categoria encontrada. Este tipo inclui as funcionalidades languageConfidence, language e languageID.

tt_extension_typesystem

A funcionalidade tt_extension_typesystem inclui as funcionalidades de análise de texto para um processamento mais avançado.

uima.tt.TokenLikeAnnotation
Este tipo de anotação em tt_extension_typesystem tem as seguintes funcionalidades:

- lemmaEntries lista todos os lemas possíveis para o testemunho. Os itens da lista são do tipo uima.tt.Lemma
- tokenNumber
- stopwordToken

uima.tt.Lemma
Esta anotação do tipo uima.tt.KeyStringEntry tem as seguintes funcionalidades:

- isStopword é verdadeiro (true) se o lema for uma palavra de paragem
- isDeterminer é verdadeiro (true) se o lema for um determinante
- partOfSpeech. Existem os seguintes códigos de descrição do número de parte do discurso:
  - 0: desconhecido
  - 1: pronome
  - 2: verbo
  - 3: substantivo
  - 4: adjectivo
  - 5: advérbio
  - 6: aposição
  - 7: interjeição
  - 8: conjunção

uima.tt.DocStructureAnnotation
As anotações sobre a estrutura do documento. Tem os seguintes subtipos:

uima.tt.SentenceAnnotation
Frase do documento. Tem a funcionalidade sentenceNumber.

uima.tt.ParagraphAnnotation
Parágrafo do documento. Tem a funcionalidade paragraphNumber.

dlt_extension_typesystem

A funcionalidade dlt_extension_typesystem inclui as funcionalidades adicionais utilizadas por IBM LanguageWare.

uima.tt.LexicalAnnotation
Esta anotação tem os seguintes subtipos:

uima.tt.TokenLikeAnnotation
Em `dlt_extension_typesystem`, esta anotação tem as seguintes funcionalidades:

- `synonymEntries`
- `frost_TokenType`
- `inflectedForms`
- `spellAid`
- `decomposition`

com.ibm.dlt.uimatypes.FilePath

com.ibm.dlt.uimatypes.Email

com.ibm.dlt.uimatypes.Number

com.ibm.dlt.uimatypes.URL

com.ibm.dlt.uimatypes.Date

com.ibm.dlt.uimatypes.Time

com.ibm.dlt.uimatypes.Tel

com.ibm.dlt.uimatypes.Currency

com.ibm.dlt.uimatypes.Acronym

uima.tt.TokenLikeAnnotation
Este tipo de anotação em `dlt_extension_typesystem` tem o seguinte tipo:

com.ibm.dlt.uimatypes.MWU
Este tipo é utilizado pelo IBM LanguageWare para anotar as expressões de várias palavras.

uima.tt.KeyStringEntry
As anotações de cadeia. Tem os seguintes subtipos:

uima.tt.Lemma
Tem as seguintes funcionalidades:

- `frost_Constraints` com os sinalizadores de restrição
- `frost_MorphBitMasks` contendo uma matriz de máscara de bits morfológica
- `frost_ExtendedPOS` com mais informações de parte do discurso, tal como, JPOS para japonês e CPOS para chinês
- `frost_JKom` contendo dados morfológicos em japonês
- `frost_JPStart` contendo dados de análise de início do japonês
- `morphID` contendo propriedades do lema

uima.tcas.Annotation
Tem o seguinte subtipo:

com.ibm.dlt.uimatypes.Decomp_Analysis
Análise estrutural completa de um composto. Tem as seguintes funcionalidades:

- `headComponentIndex` com o componente principal do composto
- `route` contendo uma lista de testemunhos que abrange um único encaminhamento de decomposição

Referências relacionadas

"Amostra da descrição do sistema tipo"
A descrição do sistema tipo descreve as estruturas funcionais (as estruturas de dados subjacentes que representam os resultados da análise) utilizadas na análise personalizada.

Amostra da descrição do sistema tipo

A descrição do sistema tipo descreve as estruturas funcionais (as estruturas de dados subjacentes que representam os resultados da análise) utilizadas na análise personalizada.

A descrição do sistema tipo tem de fazer parte do arquivo de motor de análise (ficheiro .pear) importado do ambiente UIMA para o Enterprise Search.

A seguinte descrição do sistema tipo de amostra descreve os relatórios policiais que contêm informações sobre suspeitos, local do crime, hora do crime e data:

A mesma descrição do sistema tipo de amostra é utilizada em todos os tópicos de análise de texto que debatem diferentes tipos de mapeamento que pode seleccionar com a análise personalizada.

```
<?xml version="1.0" encoding="UTF-8"?>
<typeSystemDescription>
   <name>Sistema Tipo dos Relatórios Policiais</name>
   <description>Descrição do sistema tipo para
        relatórios policiais</description>
   <version>1.0</version>
   <types>
     <typeDescription>
       <name>com.ibm.omnifind.types.RelatórioPolicial</name>
       <description>Anota um relatório policial</description>
       <supertypeName>uima.tcas.Annotation</supertypeName>
       <features>
         <featureDescription>
           <name>hora</name>
           <description>Hora do crime reportado
              </description>
           <rangeTypeName>com.ibm.omnifind.types.Hora</rangeTypeName>
         </featureDescription>
         <featureDescription>
           <name>data</name>
           <description>Quando ocorreu o crime</description>
           <rangeTypeName>com.ibm.omnifind.types.Data</rangeTypeName>
         </featureDescription>
         <featureDescription>
           <name>localização</name>
           <description>Onde ocorreu o crime</description>
           <rangeTypeName>com.ibm.omnifind.types.Cidade</rangeTypeName>
         </featureDescription>
         <featureDescription>
           <name>suspeitosConhecidos</name>
           <description>Contém anotações do tipo Suspeito</description>
           <rangeTypeName>uima.cas.FSArray</rangeTypeName>
         </featureDescription>
         <featureDescription>
           <name>descriçãoCrime</name>
           <description>Curta descrição do crime</description>
           <rangeTypeName>uima.cas.String</rangeTypeName>
         </featureDescription>
       </features>
     </typeDescription>
     <typeDescription>
       <name>com.ibm.omnifind.types.Cidade</name>
```

```
  <description>O nome de uma cidade</description>
  <supertypeName>uima.tcas.Annotation</supertypeName>
  <features>
    <featureDescription>
      <name>nomeCidade</name>
      <description>O nome da cidade</description>
      <rangeTypeName>uima.cas.String</rangeTypeName>
    </featureDescription>
    <featureDescription>
      <name>zonaCidade</name>
      <description>O nome da zona</description>
      <rangeTypeName>uima.cas.String</rangeTypeName>
    </featureDescription>
  </features>
</typeDescription>
<typeDescription>
  <name>com.ibm.omnifind.types.Pessoa</name>
  <description>Uma anotação de pessoa</description>
  <supertypeName>uima.tcas.Annotation</supertypeName>
  <features>
    <featureDescription>
      <name>função</name>
      <description>Por exemplo, suspeito ou testemunha</description>
      <rangeTypeName>uima.cas.String</rangeTypeName>
    </featureDescription>
    <featureDescription>
      <name>nomePróprio</name>
      <description>O nome próprio da pessoa</description>
      <rangeTypeName>uima.cas.String</rangeTypeName>
    </featureDescription>
    <featureDescription>
      <name>sobrenome</name>
      <description>O sobrenome da pessoa</description>
      <rangeTypeName>uima.cas.String</rangeTypeName>
    </featureDescription>
    <featureDescription>
      <name>título</name>
      <description>Por exemplo, Sr. ou Mna.</description>
      <rangeTypeName>uima.cas.String</rangeTypeName>
    </featureDescription>
    <featureDescription>
      <name>género</name>
      <description>Masculino ou feminino</description>
      <rangeTypeName>uima.cas.String</rangeTypeName>
    </featureDescription>
  </features>
</typeDescription>
<typeDescription>
  <name>com.ibm.omnifind.types.Suspeito</name>
  <description>Um suspeito encontrado</description>
  <supertypeName>com.ibm.omnifind.types.Pessoa</supertypeName>
  <features>
    <featureDescription>
      <name>descrição</name>
      <description>Descrição do suspeito,
      por exemplo, com barba e óculos escuros</description>
      <rangeTypeName>uima.cas.String</rangeTypeName>
    </featureDescription>
  </features>
</typeDescription>
<typeDescription>
  <name>com.ibm.omnifind.types.Data</name>
  <description>Uma data</description>
  <supertypeName>uima.tcas.Annotation</supertypeName>
  <features>
    <featureDescription>
      <name>ano</name>
```

```
            <description>O ano, por exemplo, 2005</description>
            <rangeTypeName>uima.cas.Integer</rangeTypeName>
          </featureDescription>
          <featureDescription>
            <name>mês</name>
            <description>O mês em dígitos, por exemplo, 7</description>
            <rangeTypeName>uima.cas.Integer</rangeTypeName>
          </featureDescription>
          <featureDescription>
            <name>dia</name>
            <description>O dia em dígitos</description>
            <rangeTypeName>uima.cas.Integer</rangeTypeName>
          </featureDescription>
          <featureDescription>
            <name>diaDaSemana</name>
            <description>O dia de semana, por exemplo, Segunda</description>
            <rangeTypeName>uima.cas.String</rangeTypeName>
          </featureDescription>
          <featureDescription>
            <name>trimestre</name>
            <description>O trimestre, por exemplo, Q1-2005</description>
            <rangeTypeName>uima.cas.String</rangeTypeName>
          </featureDescription>
          <featureDescription>
            <name>englDate</name>
            <description>Data como dd/mm/aaaa</description>
            <rangeTypeName>uima.cas.String</rangeTypeName>
          </featureDescription>
        </features>
      </typeDescription>
      <typeDescription>
        <name>com.ibm.omnifind.types.Hora</name>
        <description>Uma hora</description>
        <supertypeName>uima.tcas.Annotation</supertypeName>
        <features>
          <featureDescription>
            <name>horas</name>
            <description>Horas de 00-23</description>
            <rangeTypeName>uima.cas.Integer</rangeTypeName>
          </featureDescription>
          <featureDescription>
            <name>minutos</name>
            <description>Minutos na hora</description>
            <rangeTypeName>uima.cas.Integer</rangeTypeName>
          </featureDescription>
          <featureDescription>
            <name>horaDoDia</name>
            <description>Períodos horários, tal como, manhã, tarde</description>
            <rangeTypeName>uima.cas.String</rangeTypeName>
          </featureDescription>
        </features>
      </typeDescription>
    </types>
</typeSystemDescription>
```

Marcação XML na análise e procura

Pode mapear informações em estruturas XML num documento directamente para uma estrutura de análise comum sem escrever um anotador de UIMA.

Se os documentos na colecção estiverem em XML e pretender explorar a marcação XML durante a análise de texto ou procura semântica, tem as seguintes opções:

Procura XML nativa

Utilize esta opção se pretender utilizar todos os controlos e atributos XML à medida que aparecem no documento durante a procura semântica. Por exemplo, se tiver documentos de facturação que contenham um elemento `<addressee>`, a activação da procura XML nativa permite-lhe utilizar este controlo numa consulta de procura semântica para procurar determinado nome de cliente no âmbito deste elemento.

Com esta opção, a estrutura XML do documento é representada na estrutura de análise comum utilizando o tipo `com.ibm.es.tt.MarkupTag`. Para cada controlo XML, é criada uma anotação deste tipo. Esta anotação contém o nome do controlo, os respectivos atributos e o conteúdo do atributo. Estas informações são sempre indexadas e estão acessíveis para a procura semântica.

A procura XML nativa não requer um ficheiro de configuração de mapeamento. Pode activar a procura XML nativa a partir da consola de administração para o Enterprise Search.

Mapeamento de elementos XML para a estrutura de análise comum

Utilize esta opção nos seguintes casos:

- A semântica de determinados elementos XML é exacta e pode ser utilizada noutros passos de análise de texto. Estes passos de análise podem operar directamente nas anotações e funcionalidades criadas a partir de estruturas XML e estão protegidos de formatos potencialmente diferentes dos documentos originais. Por exemplo, o elemento `<addressee>` nos documentos de facturação contêm normalmente nomes de clientes. Ao utilizar o mapeamento de elementos XML para a estrutura de análise comum, o conteúdo deste elemento pode ser mapeado directamente para anotações do tipo `Cliente`. Um anotador pode, em seguida, inferir uma relação de Cliente-localizado-em, utilizando as informações em redor da anotação `Cliente`.
- Pretende limitar o âmbito do processamento de um anotador personalizado para áreas especificadas na entrada XML. Por exemplo, poderá pretender limitar a análise do conteúdo dos controlos `<comentárioTécnico>` apenas num anotador que detecte problemas em carros.
- Pretende restringir o processamento da análise de texto e subsequente procura em determinadas partes do documento XML e filtrar conteúdo irrelevante ou não textual.
- Pretende mapear controlos XML com nomes diferentes para uma expansão comum que seja utilizada na procura semântica. Por exemplo, o mapeamento de `<cabeçalhoPrincipal>` ou `<doc>` para o título.

Nestes casos, tem de criar um ficheiro de mapeamento de elementos XML para a estrutura de análise comum que define quais os elementos que mapeiam para determinadas estruturas funcionais. As estruturas funcionais que definir no ficheiro de mapeamento são criadas quando os documentos são analisados e são acedidos pelos anotadores personalizados.

Pode utilizar mais do que um ficheiro de mapeamento de elementos XML para a estrutura de análise comum para uma colecção de documentos. O ficheiro de mapeamento que se destina a um documento XML é determinado pelo elemento `<identificador>`. O elemento `<identificador>` no ficheiro de mapeamento tem de corresponder ao elemento raiz no documento XML. Por exemplo, se o elemento raiz do documento for `doc`, o valor do elemento `<identifier>` no ficheiro de mapeamento tem também de ser "doc".

Se não for encontrada uma correspondência, o programa procura um ficheiro de mapeamento com o elemento <identificador> definido como Predefinição (Default). Se não for encontrado um mapeamento predefinido, as secções textuais do documento (sem informações de controlo) são mapeadas para a anotação do documento na estrutura de análise comum.

Se pretender extrair informações apenas contidas nas partes relevantes de um documento, enquanto ignora as partes irrelevantes, especifique simplesmente quais os elementos XML nos documentos que contêm informações relevantes. Este processo é referido como extracção de conteúdo. Por exemplo, pode extrair a entrada especificada nos elementos de título e corpo, enquanto ignora a entrada em autor, data, ID e editor.

A extracção de conteúdo pode melhorar o processamento da análise para os seguintes tipos de documentos XML:

- Os documentos que contêm grandes quantidades de conteúdo não sujeito a análise, por exemplo, anexos binários. A utilização da extracção de conteúdo reduz significativamente o tamanho do documento, acelerando o processo e evitando os erros de análise que começam a partir de dados não adequados.
- Os documentos nos quais o texto está intercalado com texto irrelevante, por exemplo, documentos que contêm informações editoriais no âmbito dos controlos <nota>. Ao ignorar estas informações obtém melhores resultados quando analisa o conteúdo do documento.

A procura XML nativa e as opções de extracção do conteúdo no mapeamento de elementos XML para a estrutura de análise comum são opções mutuamente exclusivas, porque qualquer do conteúdo ou apenas o conteúdo especificado pode ser considerado. Se especificar a extracção do conteúdo, o mapeamento XML nativo é ignorado. Sem a extracção do conteúdo, pode ter mapeamento de elementos XML para a estrutura de análise comum e a procura XML nativa.

Todos os tipos e funcionalidades que utiliza no ficheiro de configuração têm de estar descritos na descrição do sistema tipo dos passos da análise personalizada. pode criar um descritor do sistema tipo no ambiente UIMA utilizando o suplemento Component Descriptor Editor Eclipse. Este suplemento permite-lhe criar um ficheiro descritor sem necessitar de ter conhecimentos sobre a sintaxe XML necessária.

Após criar e testar a análise personalizada, utilize o assistente de conversão UIMA PEAR (Processing Engine ARchive) para criar um arquivo que contenha os ficheiros de análise personalizada incluindo a descrição do sistema tipo. Em seguida, pode carregar o arquivo de análise personalizada e os ficheiros de mapeamento de de elementos XML para a estrutura de análise comum no Enterprise Search utilizando a respectiva consola de administração.

Tarefas relacionadas

"Criar um ficheiro de mapeamento de elementos XML para a estrutura de análise comum" na página 27
Num ficheiro de mapeamento de XML para a estrutura de análise comum, pode utilizar o intervalo completo de opções de configuração para o mapeamento de XML para tipos de dados UIMA.

Criar um ficheiro de mapeamento de elementos XML para a estrutura de análise comum

Num ficheiro de mapeamento de XML para a estrutura de análise comum, pode utilizar o intervalo completo de opções de configuração para o mapeamento de XML para tipos de dados UIMA.

Acerca desta tarefa

O ficheiro de mapeamento de XML para a estrutura de análise comum é mostrado no seguinte exemplo.

O relatório policial de amostra tem controlos XML para o tipo de crime, data do crime, localização do crime, agente principal, a esquadra de polícia onde está empregado o agente policial, a descrição do suspeito e outras informações. Estas indicações são seguidas por uma secção de corpo. Por exemplo:

```
<relatório>
 <doc>
  <tipoCrime>Roubo de carro</tipoCrime>
  <dataCrime>23/04/05 21:23</dataCrime>
  <localizaçãoCrime>R. Principal 27, Porto Covo, Sines, Setúbal</localizaçãoCrime>
  <posto agentePrincipal="Ten">Joaquim
   <apelido>Costa</apelido>
  </agentePrincipal>
  <esquadraPolícia>14ª Esquadra</esquadraPolícia>
  <descriçãoSuspeito>Masculino, cabelo escuro, óculos escuros,
    calças de ganga azuis com um casaco escuro, provavelmente
    preto</descriçãoSuspeito>
  <outrasinformações>Um Mercedes CLK foi roubado no dia 23/04/2005 de um parque de
   estacionamento em frente ao Restaurante Lagoa Azul na
   R. Principal 27, Porto Covo.(número de série: 32 2761 50871)</outrasinformações>
  <corpo>Um Mercedes CLK foi roubado no dia 23/04/2005 de um parque de
   estacionamento em frente ao Restaurante Lagoa Azul na R. Principal 27,
   Porto Covo.(número de série: 32 2761 50871)

É preto e tem pneus largos da Michelin.

Testemunhas na frente do restaurante viram dois homens com roupa escura
ir embora no carro a alta velocidade. O carro foi encontrado
abandonado na Av. Almirante em Grândola. O depósito de gasolina estava vazio.
Os bancos estavam muito manchados e o banco de trás foi vandalizado.
Nada foi roubado de dentro do carro....</corpo>
 </doc>
 <image>
  <--! imagem do local do crime como cadeia codificada em base64 -->
 </image>
</relatório>
```

Com base no relatório amostra, um ficheiro de mapeamento de XML para a estrutura de análise comum pode ter a seguinte estrutura. A amostra utiliza o sistema tipo definido para o cenário de um relatório policial.

```
<?xml version="1.0"?>
<xmlCasInitializerConfiguration
 xmlns="http://www.ibm.com/2005/uima/jedii_ci_xml">

<identifier>Default</identifier>
<description>Configuração da amostra</description>

<contentElements>
  <element>/relatório/doc</element>
</contentElements>

<elementToTypeMappings>
```

```
  <elementToTypeMapping>
    <element>//doc//agentePrincipal</element>
    <type>com.ibm.omnifind.types.Pessoa</type>
    <featureValueAssignment>
      <feature>função</feature>
      <basicValue default="Agente principal">
          </basicValue>
    </featureValueAssignment>
    <featureValueAssignment>
      <feature>género</feature>
      <basicValue default="masculino"
         useAttributeValue="sexo"/>
    </featureValueAssignment>
    <featureValueAssignment>
         <feature>sobrenome</feature>
      <values concatenate="true" delimiter=" ">
        <basicValue useAttributeValue="posto"
                default="Ten"/>
        <basicValue useElementContent="apelido"/>
      </values>
    </featureValueAssignment>
  </elementToTypeMapping>
  <elementToTypeMapping>
    <element>//doc</element>
       <type>com.ibm.omnifind.types.RelatórioPolicial</type>
    <featureValueAssignment>
      <feature>descriçãoCrime</feature>
      <basicValue useElementContent="abstract"
               trim="true">
          </basicValue>
    </featureValueAssignment>
  </elementToTypeMapping>
 </elementToTypeMappings>

</xmlCasInitializerConfiguration>
```

Restrições

O ficheiro de mapeamento está dividido em duas secções:

Elemento <contentElements>
: Utilize este elemento se pretender a extracção de conteúdo específico. O ficheiro de mapeamento de amostra extrai o conteúdo na secção `<doc>` de um documento e ignora outras secções no documento. No relatório policial de XML, a imagem pode ser demasiado grande e não muito útil para o processamento de texto. Ao especificar `<doc>` como elementos de conteúdo e não `<image>`, a imagem é filtrada antes de ser iniciado qualquer processamento de texto.

<elementToTypeMappings>
: Utilize este elemento para especificar quais os elementos de XML individuais (especificados num elemento `<elementToTypeMapping>`) no documento a mapear para determinadas estruturas funcionais na estrutura de análise comum.

  Se utilizar a opção de extracção do conteúdo, os elementos XML que são especificados na secção `<elementToTypeMappings>` têm de estar contidos no âmbito dos elementos XML especificados na secção `<contentElements>`.

Procedimento

Para criar um ficheiro de mapeamento de XML para a estrutura de análise comum:

1. Crie um ficheiro XML. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML para validar o XML. O esquema XSD para o ficheiro de mapeamento denomina-se XMLCasInitSchema.xsd e encontra-se na instalação do Enterprise Search em *ES_INSTALL_ROOT*/packages/uima/configuration_xsd/.
2. Inclua os mapeamentos num elemento `<xmlCasInitializerConfiguration xmlns="http://www.ibm.com/2005/uima/jedii_ci_xml">`. O espaço de nomes (especificado no atributo xmlns) tem de estar exactamente conforme mostrado.
3. Adicione um elemento `<contentElements>` se pretender extrair conteúdo específico das secções no documento e um elemento `<elementToTypeMappings>` que especifica quais os elementos XML individuais que pretende mapear para determinadas estruturas funcionais na área de análise comum.
4. Adicione um elemento `<identifier>` e um elemento `<description>`. O identificador determina qual o mapeamento a utilizar para determinados documentos XML. Um identificador tem de conter o elemento raiz do documento, tal como `doc`. Se o identificador estiver definido como Predefinição (Default), o elemento raiz do documento é irrelevante e o mapeamento é aplicado a qualquer documento XML.
5. Adicione um elemento `<contentElements>` se pretender extrair informações contidas apenas nas partes relevantes de um documento. Tem o seguinte elemento componente:
   - Um ou mais elementos `<element>` que contenham o caminho de um elemento XML no documento e segue a sintaxe XPath, por exemplo `<element>/doc/tipoCrime</element>`.
6. Adicione um elemento `<elementToTypeMappings>` se pretender especificar quais os elementos XML no documento que pretende mapear para determinadas estruturas funcionais na estrutura de análise comum. Tem os seguintes elementos componente:
   - Um ou mais elementos `<elementToTypeMapping>`. Este elemento tem de ter os seguintes elementos imbricados:
     - Um elemento `<element>` que é utilizado para especificar o caminho de um elemento XML e segue a sintaxe XPath: Uma barra inicial (/) significa que é fornecido um caminho completo. Por exemplo, `outrasinformações` sob o elemento raiz `doc`. Duas barras (//) significam qualquer subconjunto do caminho. Por exemplo, `dataNascimento` tem de ocorrer no âmbito de `agentePrincipal`, apesar de outros elementos poderem ocorrer entre estes dois.
     - Um elemento `<type>`, que especifica um tipo que é definido na descrição do sistema tipo. Tem de ser do tipo `Annotation`.
     - Zero ou mais elementos `<featureValueAssignment>`.
7. Num elemento `<featureValueAssignment>`, atribua um nome do tipo Cadeia (String) no elemento `<feature>` atribua um valor no elemento `<basicValue>`. Vários elementos `<basicValue>` podem ser adicionados entre um elemento `<values>`.

   O elemento `<basicValue>` pode ter atributos. Que incluem `useAttributeValue`, `useElementContent`, `default` e `trim`.

Utilize `useAttributeValue` se pretender utilizar o valor de um atributo como valor para uma funcionalidade. O seguinte exemplo

```
<elementToTypeMapping>
  <element>/doc//agentePrincipal</element>
  <type>com.ibm.omnifind.types.Pessoa</type>
  <featureValueAssignment>
```

```
    <feature>função</feature>
    <basicValue default="Agente principal"/>
  </featureValueAssignment>
  <featureValueAssignment>
     <feature>género</feature>
    <basicValue default="masculino" useAttributeValue="sexo"/>
  </featureValueAssignment>
</elementToTypeMapping>
```

resulta na seguinte saída:

- Para cada controlo XML <agentePrincipal>, que ocorre algures no âmbito de um controlo XML <doc> no documento, é criada uma estrutura funcional do tipo com.ibm.omnifind.types.Pessoa.
- Se o controlo <agentePrincipal> contiver um atributo sexo, a funcionalidade género da recém criada estrutura funcional é definida com o valor do atributo.

Utilize o atributo useElementContent para adicionar conteúdo como valor de uma funcionalidade. Por exemplo, no seguinte fragmento de mapeamento:

```
<elementToTypeMapping>
  <element>//doc</element>
      <type>com.ibm.omnifind.types.RelatórioPolicial</type>
  <featureValueAssignment>
     <feature>descriçãoCrime</feature>
    <basicValue useElementContent="abstract" trim="true"/>
  </featureValueAssignment>
</elementToTypeMapping>
```

O texto abrangido pelo elemento <outras informações> em <doc> torna-se o valor da estrutura funcional descriçãoCrime. Todos os espaços em branco iniciais e de seguimento são removidos.

Pode ser especificado mais do que um valor entre o elemento <values> para os seguintes casos:

- A funcionalidade a ser definida é do tipo StringArray.
- Muitas cadeias são concatenadas numa só cadeia utilizando o atributo delimitador e, deste modo, mapeie para uma funcionalidade do tipo Cadeia (String). Por exemplo, o título Sr. é uma constante, o nome próprio é o valor de um atributo e o apelido é abrangido por um elemento XML:

```
<elementToTypeMapping>
   <element>//doc//agentePrincipal</element>
  <type>com.ibm.omnifind.types.Pessoa</type>
  <featureValueAssignment>
        <feature>sobrenome</feature>
    <values concatenate="true" delimiter=" ">
      <basicValue default="Sr."/>
       <basicValue useAttributeValue="posto"
         default="Ten."/>
       <basicValue useElementContent="apelido"/>
    </values>
  </featureValueAssignment>
</elementToTypeMapping>
```

Os valores da funcionalidade da cadeia são extraídos do ficheiro de mapeamento tal como estão. Os valores retêm quaisquer espaços em branco iniciais ou de seguimento. No entanto, os nomes de tipos e funcionalidades são aparados de quaisquer espaços em branco. Por exemplo,
<type>**com.ibm.omnifind.types.Pessoa**</type> torna-se
<type>**com.ibm.omnifind.types.Pessoa**</type>.

Defina as condições nos atributos utilizando o elemento `<condition>`. Por exemplo, a estrutura funcional do tipo `com.ibm.omnifind.types.Pessoa` é criada apenas se `<descriçãoSuspeito>` ocorrer no documento com o atributo `armado` definido como `sim`:

```
<elementToTypeMapping>
  <element>//descriçãoSuspeito</element>
  <type>com.ibm.omnifind.types.Pessoa</type>
  <condition attribute="armado"  value="sim"/>
</elementToTypeMapping>
```

Com base no relatório policial de amostra e no ficheiro de mapeamento definido, são criadas as seguintes estruturas funcionais:

com.ibm.omnifind.types.RelatórioPolicial

- texto abrangido: ″Roubo de carro 23/04/05 21:23, R. Principal 27 - Porto Covo - Sines - Setúbal, Joaquim Costa 14ª Esquadra, Masculino, cabelo escuro, óculos escuros, calças de ganga azuis com um casaco escuro, provavelmente preto. Um Mercedes CLK foi... Nada foi roubado de dentro do carro.
- início = 2
- fim = 904
- suspeitosConhecidos = nenhum
- descriçãoCrime = ″Um Mercedes CLK foi roubado no dia 23/04/2005 de um parque de estacionamento em frente ao Restaurante Lagoa Azul na R. Principal 27, Porto Covo.(número de série: 32 2761 50871)″

com.ibm.omnifind.types.Pessoa

- texto abrangido = ″Joaquim Costa″
- início = 112
- fim = 127
- função = ″Agente principal″
- nomePróprio = nenhum
- sobrenome = ″Ten Costa″
- género = ″masculino″

Após criar o ficheiro de mapeamento, tem de carregá-lo no Enterprise Search e seleccionar o ficheiro de mapeamento de XML para a estrutura de análise comum com as outras selecções de análise personalizada utilizando a consola de administração do Enterprise Search.

Conceitos relacionados

“Marcação XML na análise e procura” na página 24
Pode mapear informações em estruturas XML num documento directamente para uma estrutura de análise comum sem escrever um anotador de UIMA.

Referências relacionadas

“Amostra da descrição do sistema tipo” na página 22
A descrição do sistema tipo descreve as estruturas funcionais (as estruturas de dados subjacentes que representam os resultados da análise) utilizadas na análise personalizada.

Resultados da análise de texto

Todos os resultados da análise de texto são armazenados na estrutura de análise comum.

Normalmente, os anotadores lêem e escrevem na estrutura de análise comum. Os consumidores da estrutura de análise comum (*consumidores da CAS*) só lêem a partir da estrutura de análise comum. Os consumidores da CAS efectuam o processamento final dos resultados de análise armazenados na estrutura de análise comum. O Enterprise Search contém dois consumidores da CAS:

- O consumidor que indexa o conteúdo da estrutura de análise comum num motor de procura. Este consumidor requer um ficheiro de mapeamento da estrutura de análise comum para o índice seleccionado com a análise de texto personalizada na consola de administração do Enterprise Search.
- O consumidor que preenche uma base de dados relacional com resultados da análise específicos. Este consumidor também requer um ficheiro de mapeamento da estrutura de análise comum para a base de dados seleccionado com as opções de análise de texto personalizada na consola de administração do Enterprise Search.

Se necessário, pode implementar consumidores da CAS personalizados no Enterprise Search. Consulte a documentação de UIMA para obter informações sobre como escrever um consumidor. Para obter informações sobre como carregar e utilizar o consumidor no Enterprise Search, consulte o sítio da Web IBM UIMA developerWorks em http://www.ibm.com/developerworks/db2/zones/db2ii/.

Conceitos relacionados

"Indexar o mapeamento para resultados da análise personalizada" na página 37
Após executar a análise personalizada numa colecção de documentos, pode utilizar o motor de procura no Enterprise Search para criar um índice a partir de informações armazenadas na estrutura de análise comum que é criada pelos algoritmos de análise personalizada.

"Mapeamento da base de dados para os resultados da análise seleccionada" na página 45
Depois de ter analisado os documentos no Enterprise Search, pode armazenar resultados da análise de texto seleccionado numa base de dados que suporte JDBC.

Caminhos de funcionalidade

Um caminho de funcionalidade fornece uma forma de aceder aos valores de funcionalidade nas estruturas de análise comum, semelhantes às instruções XPath utilizadas para aceder aos elementos de XML num documento XML.

Os caminhos de funcionalidade são úteis se pretender aceder a uma estrutura funcional que combine funcionalidades complexas, por exemplo, funcionalidades que sejam de valor de matriz ou apontem para outra estrutura funcional. Ao utilizar um caminho de funcionalidade, pode associar o valor de uma funcionalidade directamente a uma estrutura funcional e armazenar este valor no índice de procura semântica ou numa base de dados.

Por exemplo, tenha em consideração um anotador que identifique carros e as respectivas marcas. Cria anotações do tipo `carro` com o atributo `marca`. No entanto, `marca` não contém a empresa real (por exemplo, Chevrolet) mas contém uma estrutura funcional do tipo `Empresa`, que em si própria tem um atributo de valor em cadeia `nomeempresa`. Para permitir uma consulta semântica que combine nomes de carros e nome de empresas, um caminho de funcionalidade `marca/nomeempresa` é utilizado para anexar o valor de `nomeempresa` à expansão "carro" que é gerada para a anotação do carro. Deste modo, permite a consulta "Pretendo documentos que contenham carros feitos pela Chevrolet", utilizando '/carro[@marca="Chevrolet"]'.

Um caminho de funcionalidade é uma sequência de nomes de funcionalidade (f1/.../fn) com as seguintes propriedades:

- O valor de um caminho de funcionalidade pode ser Cadeia (String), Número Inteiro (Integer), Flutuante (Float) ou uma matriz de um destes tipos.
- Todas as funcionalidades no âmbito do caminho f1 a fn-1 têm de ter um tipo complexo, ou seja, do tipo `uima.cas.TOP`, `uima.cas.FSArray`, `uima.cas.FSList` ou de um dos respectivos subtipos.
- A última funcionalidade fn no caminho pode incluir um tipo complexo. Adicionalmente, pode incluir um (sub)tipo de `uima.cas.Float`, `uima.cas.Integer`, `uima.cas.String`, `uima.cas.FloatArray`, `uima.cas.IntegerArray`, `uima.cas.StringArray`, `uima.cas.FloatList`, `uima.cas.IntegerList` ou `uima.cas.StringList`.
- Opcionalmente, uma funcionalidade pode ter tipo. O nome do tipo totalmente qualificado tem de ser adicionado como prefixo ao nome da funcionalidade e ser separado por dois pontos. Por exemplo, `f1/com.ibm.es.AlgumTipo:f2/.../fn`.

Pode estreitar o âmbito do tipo de determinada funcionalidade. Por exemplo, tenha em consideração uma funcionalidade `InfoAdicional` do tipo `uima.cas.TOP`. Se tiver conhecimento que o valor da funcionalidade `InfoAdicional` for na realidade do tipo `InfoFuncionário` quer tem a funcionalidade `salário`, pode aceder a esta funcionalidade utilizando `InfoAdicional/InfoFuncionário:salário`. Tenha em atenção que neste exemplo, o caminho da funcionalidade `InfoAdicional/salário` resultaria num erro, uma vez que `salário` não foi definido para o tipo `uima.cas.TOP`.

As funcionalidades que são de valor de matriz ou de lista têm as seguintes propriedades adicionais:

- Utilize os parêntesis rectos ([<número>]) para seleccionar determinado elemento na matriz ou lista. Uma matriz inicia com (0). Por exemplo, para seleccionar o primeiro elemento na matriz das empresas, utilize empresas[0]. O marcador especial [last] pode ser utilizado para seleccionar a última entrada numa matriz, independentemente do respectivo tamanho, por exemplo, empresas[last].
- Utilize os parêntesis rectos vazios ([]) para denotar todos os elementos. Apenas é permitido um conjunto de parêntesis rectos ([]) num caminho de funcionalidade. Por exemplo, se existir uma matriz de suspeitos, o caminho da funcionalidade `suspeitosConhecidos[]/com.ibm.omnifind.types.Suspeito:apelido` recolhe todos os apelidos dos suspeitos numa matriz de Cadeia (String).
- Quando um caminho de funcionalidade que devolve uma matriz for utilizado durante a indexação, os elementos da matriz são concatenados (separados por espaços em branco) e escritos no índice como um único atributo ou campo de vários termos.
- O elemento seguinte no caminho da funcionalidade tem de ter um tipo. O nome do tipo é o tipo de elementos no âmbito da matriz. Por exemplo, tenha em consideração uma funcionalidade do tipo `Info`. Este tipo tem uma funcionalidade com o nome `empresas`, cujo âmbito é uma matriz FSArray. Os elementos da matriz são do tipo `Empresa`. `Empresa`, por sua vez, tem uma funcionalidade com o nome `lucro`. Para obter o lucro da terceira empresa, escreva (utilizando os nomes de tipo totalmente qualificados) `empresas[2]/Empresa:lucro`.

Funcionalidades incorporadas

As funcionalidades incorporadas são nomes de funcionalidade predefinidos com semântica especial. Podem ser utilizadas para aceder a informações que não estão contidas na estrutura funcional em si, por exemplo, o tipo de estrutura funcional

ou o texto abrangido de uma anotação. Podem ser utilizadas num caminho de funcionalidade que o último ou único elemento.

As seguintes funcionalidades incorporadas podem ser utilizadas em ambos os ficheiros de configuração do mapeamento:

- `fsId()` devolve o ID da estrutura funcional. O ID devolvido é um número inteiro (32 bits). Utilize esta funcionalidade incorporada para aceder a partes de um documentos que correspondam exactamente à consulta.
- `typeName()` devolve o tipo de objecto da estrutura de análise comum como uma cadeia. O tipo é o nome de tipo totalmente qualificado incluindo quaisquer prefixos de espaço de nomes, por exemplo, `uima.tcas.Annotation`. No contexto da base de dados, `typeName()` é especialmente útil se armazenar tipos e subtipos na mesma coluna e pretender conhecer um tipo real de uma anotação ou estrutura funcional. O seguinte exemplo armazena o tipo de pessoa, tal como, *suspeito* ou *testemunha*, na coluna da função.

```
<explicitMappingRule  applyToSubTypes="false">
   <type>com.ibm.omnifind.types.Pessoa</type>
   <table>sample.pessoa</table>
   <featureMappings>
     <featureMapping>
       <feature>typeName()</feature>
       <column>função</column>
     </featureMapping>
   </featureMappings>
 </explicitMappingRule>
```

- `coveredText()` devolve o texto que é expandido pelo objecto de análise comum. `coveredText()` está disponível apenas para anotações e respectivos subtipos. Não utilize esta funcionalidade incorporada em estruturas funcionais que não sejam sub-somadas pelo tipo de anotação. O seguinte exemplo armazena o nome de um suspeito na coluna nomeSuspeito.

```
<implicitMappingRule applyToSubTypes="false">
   <type>com.ibm.omnifind.types.Suspeito</type>
   <relation>sample.pessoa</relation>
   <featureMappings>
     <featureMapping>
       <feature>coveredText()</feature>
       <column>nomeSuspeito</column>
       <length>128</length>
     </featureMapping>
   </featureMappings>
 </implicitMappingRule>
```

- [] devolve um parâmetro identificador à entrada do contentor actual (matriz ou lista). A funcionalidade implica uma iteração, o que significa que uma entrada é efectuada na tabela de bases de dados ou índice para cada elemento na matriz ou lista. O seguinte exemplo é retirado de um ficheiro de mapeamento da estrutura de análise comum para a base de dados no qual a função incorporada [:index] também é permitida.

```
<implicitMappingRule applyToSubTypes="false">
   <type>uima.cas.FSArray</type>
   <table>sample.suspeitosConhecidos</table>
   <featureMappings>
     <featureMapping>
       <feature>uniqueId()</feature>
       <column>arrayId</column>
     </featureMapping>
     <featureMapping>
       <feature>[:index]</feature>
       <column>arrayIndex</column>
     </featureMapping>
     <featureMapping>
```

```
      <feature>[]/com.ibm.omnifind.types.Suspeito:uniqueId()</feature>
      <column>IDsuspeito</column>
    </featureMapping>
  </featureMappings>
</implicitMappingRule>
```

As seguintes funcionalidades incorporadas podem ser utilizadas apenas no ficheiro de mapeamento da estrutura de análise comum para a base de dados:

- `uniqueId()` devolve o ID exclusivo global da estrutura funcional. O ID exclusivo devolvido é uma cadeia de comprimento fixo (27 caracteres) e é uma concatenação do resultado do `fsId()`, `docId()`, `docTimestamp()` e do número da parte actual, uma vez que os documentos podem ser divididos em partes de várias estruturas de análise comum no Enterprise Search.

  A cadeia devolvida pode incluir quaisquer caracteres entre ″a-z″ e ″A-Z″, os números ″0-9″, ponto e vírgula (″;″) e dois pontos (″:″).

  O resultado de `uniqueId()` pode ser utilizado como chave principal para tabelas.
- `objectId()` devolve o ID da anotação ou estrutura funcional. `objectId()` é semelhante a `uniqueId()`, excepto que não contém o resultado de `docTimestamp()`. O ID devolvido é exclusivo apenas numa colecção em que os documentos são analisados uma vez. Se requerer exclusividade em todos os documentos e versões de documentos, tem de utilizar `uniqueId()`.

  A cadeia devolvida da funcionalidade incorporada `objectId()` tem um comprimento fixo de 16 caracteres e pode incluir quaisquer caracteres entre ″a-z″ e ″A-Z″, os números ″0-9″, ponto e vírgula (″;″) e dois pontos (″:″).

  Se o `uniqueId()` ou `objectId()` fizerem referência às estruturas funcionais que estão vazias, o valor predefinido estabelecido na definição da tabela de bases de dados é retirado, não são armazenados objectos vazios de um tipo referenciado.
- `docId()` devolve o ID de documento. O valor de retorno é do tipo de número inteiro (32 bits).

  O seguinte exemplo mostra estas funcionalidades incorporadas:

```
<explicitMappingRule applyToSubTypes="true">
       <type>com.ibm.omnifind.types.RelatórioPolicial</type>
  <table>sample.RelatórioPolicial</table>
  <featureMappings>
    <featureMapping>
      <feature>uniqueId()</feature>
      <column>IDRelatórioPolicial</column>
    </featureMapping>
    <featureMapping>
      <feature>docId()</feature>
      <column>IDdocRelatórioPolicial</column>
    </featureMapping>
  </featureMappings>
</explicitMappingRule>
```

- `docUri()` devolve o URI de documento.
- `docTimestamp()` devolve a hora (em milissegundos) de quando o documento foi processado. Esta funcionalidade incorporada é útil para rastrear versões de documentos, por exemplo, se pretender saber se a versão do documento que está a utilizar é a mais recente passada pela ferramenta de sequências de hiperligações.

```
<explicitMappingRule applyToSubTypes="false">
       <type>com.ibm.omnifind.types.RelatórioPolicial</type>
<relation>amostra.RelatórioPolicial</relation>
  <featureMappings>
    <featureMapping>
      <feature>uniqueId()</feature>
      <column>IDRelatórioPolicial</column>
```

```
      </featureMapping>
      <featureMapping>
        <feature>docTimestamp()</feature>
        <column>versãoRelatório</column>
      </featureMapping>
    </featureMappings>
  </explicitMappingRule>
```

- `parentId()` devolve o `fsId()` da estrutura funcional que inclui um mapeamento de contentor. `parentId()` é válido apenas no âmbito do contexto de um mapeamento de contentor.
- `uniqueParentId()` devolve o `uniqueId()` da anotação ou estrutura funcional incluída num mapeamento de contentor. Esta funcionalidade incorporada também é válida apenas no âmbito do contexto de mapeamento de contentor.
- `[:index]` devolve o índice da entrada do contentor actual (matriz ou lista).

Tarefas relacionadas

"Obter partes de um documento que correspondam a uma consulta de procura semântica" na página 56
Pode obter apenas as partes de um documento que correspondam exactamente à consulta através do mapeamento das estruturas funcionais relevantes para o índice e base de dados e especificando a expansão na consulta de procura semântica.

Filtros

Os filtros são utilizados para restringir regras de mapeamento nos ficheiros de mapeamento da estrutura de análise comum para o índice e nos ficheiros de mapeamento da estrutura de análise comum para a base de dados. Apenas quando o filtro está definido como verdadeiro (true) são adicionados os resultados da análise ao índice ou a uma tabela JDBC.

O elemento `<filter>` é opcional e utilizado para restringir os mapeamentos apenas para funcionalidades com determinado valor de atributo. É útil se pretender que um atributo funcione como um parâmetro para indicar o que será para indexar ou adicionar à base de dados. Por exemplo, as pessoas e organizações podem ser registadas numa anotação do tipo `EntityAnnotation`. A respectiva funcionalidade denominada `type` está definida como `pessoa` ou `organização`. Para extrair apenas as pessoas e não as organizações, pode adicionar o seguinte filtro à regra de mapeamento:

```
<filter syntax="FeatureValue">type = "pessoa"</filter>
```

Cada expressão de filtro toma a forma:

```
<FeaturePath> <Operador> <Literal>
```

em que:

- FeaturePath, corresponde a um caminho de funcionalidade na estrutura de análise comum.
- Operador, corresponde a =, !=, <, <=, > ou >=. Tenha em atenção que < (e apenas <) tem de ser expresso como `&lt;`.
- Literal, é um número inteiro, número de vírgula flutuante (não é suportada qualquer sintaxe expoente) ou literal de cadeia, colocado entre aspas, com aspas incorporadas e barras invertidas antecedidas por uma barra invertida.

`<FeaturePath>`, `<Operador>` e `<Literal>` têm de ser separados por um espaço em branco.

Os seguintes exemplos são filtros válidos:

- `<filter syntax="FeatureValue"> foo = "olá mundo" </filter>`

  A funcionalidade ″foo″ contém a cadeia ″olá mundo″.
- `<filter syntax="FeatureValue"> foo &lt; 42 </filter>`

  A funcionalidade ″foo″ contém um valor inteiro inferior a 42.
- `<filter syntax="FeatureValue"> marca/empresa = "Chevrolet" </filter>`

  O caminho de funcionalidade ″marca/empresa″ em que a funcionalidade ″marca″ contém uma estrutura funcional com uma funcionalidade ″empresa″ com o valor ″Chevrolet″.
- `<filter syntax="FeatureValue"> bar7 >= 0,5 </filter>`

  A funcionalidade ″bar7″ contém um valor flutuante superior ou igual a 0,5.

Indexar o mapeamento para resultados da análise personalizada

Após executar a análise personalizada numa colecção de documentos, pode utilizar o motor de procura no Enterprise Search para criar um índice a partir de informações armazenadas na estrutura de análise comum que é criada pelos algoritmos de análise personalizada.

O mapeamento dos resultados da análise para campos, expansões de texto e atributos no índice do Enterprise Search permitem utilizar estas informações nas consultas. A combinação da análise com o Enterprise Search, que tem capacidade para indexar as palavras e expansões de texto, permite a procura semântica.

Ao utilizar o ficheiro de mapeamento da estrutura de análise comum para o índice, pode determinar quais os resultados da análise na estrutura de análise comum que pretende indexar.

Pode utilizar diferentes estilos para mapear estruturas funcionais na estrutura de análise comum para o índice do Enterprise Search.

Anotação (Annotation)

Se indexar as estruturas funcionais na estrutura de análise comum utilizando o estilo de anotação, todas as anotações dos tipos especificados são armazenadas no índice como expansões passíveis de serem procuradas.

Por exemplo, se uma estrutura funcional que expande determinada área de texto for do tipo `pessoa` e estiver indexada utilizando o estilo de anotação, são possíveis as seguintes consultas:

Tabela 2. Consultas de amostra

Informações requeridas	Possível consulta
Pretendo todos os documentos que contêm pelo menos um nome de pessoa	<pessoa/>
Pretendo todos os documentos em que Chefe está contido no âmbito de uma anotação de pessoa	<pessoa>chefe</pessoa>
Pretendo todos os documentos em que Idioma é mencionado na mesma frase que um dos meus concorrentes	<frase><pessoa>Idioma</pessoa> <concorrente/></frase>

Os atributos das estruturas funcionais são também indexados como parte da expansão. Por exemplo, tenha em consideração um anotador que detecte carros e armazena a marca do carro como uma funcionalidade `marca` da anotação `carro`. Deste modo, permite o seguinte tipo de consulta: ″Pretendo documentos que mencionem carros da marca Chevrolet″.

Campo (Field)
Utilize este estilo se pretender disponibilizar o conteúdo das estruturas funcionais durante a procura utilizando as capacidades de procura do campo no Enterprise Search. Desta forma, o conteúdo de uma estrutura funcional pode ser apresentado nos resultados da procura ou utilizado na procura paramétrica.

Por exemplo, se mapear dosagens de medicamentos para um campo paramétrico, pode utilizar a seguinte consulta: ″Pretendo todos os documentos que referem determinado medicamento tomado com determinada dosagem acima de 100 miligramas″.

Quebra (Breaking)
Utilize este estilo se pretender que determinada funcionalidade seja interpretada como um delimitador de limpeza, por exemplo, secções ou parágrafos. O Enterprise Search detecta frases e parágrafos por predefinição. Utilize este estilo apenas se a análise personalizada detectar elementos estruturais adicionais num documento que pretende ter uma interpretação diferente.

Os resultados da análise podem ser utilizados também para afectar a classificação do documento no Enterprise Search, mesmo para simples consultas de palavras chave. Este processo é executado em dois passos:

1. Mapeie as estruturas funcionais para expansão ou campos passíveis de serem procurados, utilizando o estilo de mapeamento Anotação (Annotation) ou Campo (Field).
2. Defina uma classe hierárquica utilizando a consola de administração do Enterprise Search e mapeie a nome de campo ou expansão para esta classe hierárquica.

Se o utilizador introduzir um termo de procura que esteja contido no âmbito da estrutura funcional, o documento é classificado como superior. Por exemplo, tenha em consideração um anotador que detecte nomes de pessoas e empresas. Ao mapear estas estruturas de funcionalidade para expansões (tal como ″pessoa″ e ″empresa″) e, em seguida, mapear estas expansões para classes hierárquicas, o resultado da procura para ″intervalo″ classifica os documentos como superiores os que referem a empresa ″Intervalo″ em relação aos que meramente contêm o termo ″intervalo″.

Após escrever o ficheiro de mapeamento da estrutura de análise comum para o índice, pode carregá-lo para o Enterprise Search utilizando a consola de administração.

Tarefas relacionadas

“Criar o ficheiro de mapeamento da estrutura de análise comum para o índice” na página 39
Ao utilizar o ficheiro de mapeamento da estrutura de análise comum para o índice, pode determinar quais os resultados da análise na estrutura de análise comum que pretende indexar para activar a procura.

Criar o ficheiro de mapeamento da estrutura de análise comum para o índice

Ao utilizar o ficheiro de mapeamento da estrutura de análise comum para o índice, pode determinar quais os resultados da análise na estrutura de análise comum que pretende indexar para activar a procura.

Acerca desta tarefa

O ficheiro de mapeamento da estrutura de análise comum para o índice está no XML. O ficheiro de mapeamento da estrutura de análise comum para o índice de amostra baseia-se no sistema tipo definido para o cenário de um relatório policial.

```
<?xml version="1.0" encoding="UTF-8"?>
<indexBuildSpecification
xmlns="http://www.ibm.com/of/822/consumer/index/xml">
 <skipCondition>
   <type>com.ibm.uima.tt.DocumentAnnotation</type>
   <filter syntax="FeatureValue">toBeprocessed = 0</filter>
 </skipCondition>

 <indexBuildItem>
   <name>com.ibm.omnifind.types.Pessoa</name>
   <indexRule>
     <style name="Annotation">
       <attributemappings>
         <mapping>
           <feature>função</feature>
           <indexName>função</indexName>
         </mapping>
         <mapping>
           <feature>título</feature>
           <indexName>título</indexName>
         </mapping>
         <mapping>
           <feature>género</feature>
           <indexName>género</indexName>
         </mapping>
       </attributemappings>
     </style>
   </indexRule>
 </indexBuildItem>
 <indexBuildItem>
   <name>com.ibm.omnifind.types.Suspeito</name>
   <indexRule>
     <style name="Annotation"/>
     <style name="Field">
       <attribute name="parametric" value="false"/>
       <attribute name="fieldSearchable"
         value="true"/>
       <attribute name="returnable" value="true"/>
     </style>
   </indexRule>
 </indexBuildItem>
 <indexBuildItem>
   <name>com.ibm.omnifind.types.Cidade</name>
   <indexRule>
       <style name="Annotation">
       <attributemappings>
         <mapping>
           <feature>zonaCidade</feature>
           <indexName>zona</indexName>
         </mapping>
       </attributemappings>
     </style>
   </indexRule>
```

```
</indexBuildItem>
<indexBuildItem>
  <name>com.ibm.omnifind.types.Data</name>
  <indexRule>
    <style name="Field">
      <attribute name="fixedName" value="Data"/>
      <attribute name="fieldSearchable"
          value="true"/>
      <attribute name="returnable" value="true"/>
    </style>
    <style name="Field">
      <attribute name="fixedName" value="hora"/>
      <attribute name="valueFeature"  value="hora"/>
      <attribute name="parametric" value="true"/>
    </style>
  </indexRule>
  <filter syntax="FeatureValue">ano="2005"</filter>
</indexBuildItem>
<indexBuildItem>
  <name>com.ibm.omnifind.types.RelatórioPolicial</name>
  <indexRule>
    <style name="Annotation">
      <attribute name="fixedName"
         value="RelatórioPolicial"/>
      <attributemappings>
        <mapping>
          <feature>descriçãoCrime</feature>
          <indexName>descriçãoCrime</indexName>
        </mapping>
        <mapping>
          <feature>time/coveredText()</feature>
          <indexName>hora</indexName>
        </mapping>
        <mapping>
          <feature>date/englDate</feature>
          <indexName>data</indexName>
        </mapping>
        <mapping>
          <feature>location/coveredText()</feature>
          <indexName>localização</indexName>
        </mapping>
        <mapping>
          <feature>suspeitosConhecidos[]/com.ibm.omnifind.types.
          Suspeito:sobrenome</feature>
          <indexName>apelidosSuspeitos</indexName>
        </mapping>
      </attributemappings>
    </style>
  </indexRule>
 </indexBuildItem>
</indexBuildSpecification>
```

Restrições

O ficheiro de mapeamento da estrutura de análise comum para o índice tem de conter todos os resultados da análise que pretende procurar nas consultas.

Procedimento

Para criar o ficheiro de mapeamento da estrutura de análise comum para o índice:

1. Crie um ficheiro XML. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML à sua escolha. O esquema XSD para o

ficheiro de mapeamento denomina-se CasToIndexMapping.xsd e encontra-se na instalação do Enterprise Search em *ES_INSTALL_ROOT*/packages/uima/configuration_xsd/.

2. Inclua os mapeamentos num elemento `<indexBuildSpecification xmlns="http://www.ibm.com/of/822/consumer/index/xml">`. O espaço de nomes (especificado no atributo xmlns) tem de estar exactamente conforme mostrado.
3. Adicione um elemento `<skipCondition>` para proibir determinados documentos de serem indexados, com base em determinado valor da funcionalidade. Este elemento é opcional. No exemplo, não serão indexados os documentos que contêm uma estrutura de dados do tipo `com.ibm.uima.tt.DocumentAnnotation` com uma funcionalidade denominada `toBeProcessed` definida como zero.
4. Adicione um ou mais elementos `<indexBuildItem>`, que contenham o mapeamento de determinada estrutura funcional na estrutura de análise comum, a uma estrutura no índice.
5. Saia e valide o ficheiro XML.

Elemento `<indexBuildItem>`

O ficheiro de mapeamento da estrutura de análise comum para o índice contém um ou mais elementos `<indexBuildItem>`. Cada elemento descreve o mapeamento de determinada estrutura funcional na estrutura de análise comum para uma estrutura no índice (uma expansão ou campo).

O elemento `<name>` contém o tipo de estrutura funcional. Existem duas formas de especificar um tipo:

- O nome do tipo completo. Por exemplo, `com.ibm.omnifind.types.Suspeito`.
- Um carácter global. Por exemplo, `com.ibm.omnifind.types.*`. O carácter global pode ser adicionado apenas no final da especificação do tipo.

Utilize apenas os subtipos de `uima.tcas.Annotation` como itens de criação do índice. Se uma estrutura funcional for um subtipo de `uima.cas.TOP` (e não de `uima.tcas.Annotation`), pode aceder a esta estrutura funcional utilizando um caminho de funcionalidade que comece por uma anotação.

Se o tipo A for um subtipo do tipo B (na amostra, `com.ibm.omnifind.types.Suspeito` como um subtipo para `com.ibm.omnifind.types.Pessoa`) e existirem os elementos `<indexBuildItem>` Ia e Ib definidos para ambos os tipos, processa-se do seguinte modo:

- Cada regra de índice que é definida em Ib aplica-se às estruturas funcionais do tipo B e às estruturas funcionais do tipo A.
- Cada regra de índice definida em Ia aplica-se apenas às estruturas funcionais do tipo A

No exemplo, o elemento `<indexBuildItem>` definido para as anotações `com.ibm.omnifind.types.Pessoa` também se aplica às anotações `com.ibm.omnifind.types.Suspeito`. São criadas duas expansões para uma anotação do suspeito: uma com o nome Pessoa e a outra Suspeito.

O elemento `<filter>` é opcional e utilizado para restringir o mapeamento de `<indexBuildItem>` apenas para estruturas funcionais com determinado valor de atributo. É útil se pretender que um atributo funcione como um parâmetro para indicar o que será para indexar. Por exemplo, as pessoas e organizações podem ser registadas numa anotação do tipo `EntityAnnotation`. A respectiva funcionalidade

denominada type está definida como pessoa ou organização. Para extrair apenas as pessoas e não as organizações, pode adicionar o seguinte filtro:

```
<filter syntax="FeatureValue">type = "pessoa"</filter>
```

Para além disso, pode escolher indexar pessoas e organizações sob diferentes nomes de expansão, por exemplo, pessoa e organização. Para tal, defina dois elementos <indexBuildItem> do tipo EntityAnnotation e utilize dois filtros na funcionalidade type para accionar as pessoas ou organizações.

Elemento <indexRule>

Cada elemento <indexBuildItem> contém um elemento <indexRule>. Cada elemento <indexRule> contém todas as informações necessárias para mapear uma estrutura de funcionalidade na estrutura de análise comum para o índice como um estilo de campo (field), uma anotação (annotation) e um estilo de quebra (breaking). Os estilos de anotação e campo suportam um número de atributos. Não pode utilizar o estilo do termo, que é suportado no UIMA Software Development Kit no Enterprise Search (o estilo do termo é ignorado).

Para os estilos de anotação e campo, existem as seguintes alternativas quando especifica o nome da anotação ou do campo no índice:

- Utilize fixedName se pretender que cada estrutura funcional esteja acessível no índice com o mesmo nome. No seguinte exemplo, cada estrutura de funcionalidade do tipo com.ibm.omnifind.types.Pessoa será mapeada para uma expansão com o nome ″Pessoa″ no índice.

  ```
  <indexBuildItem>
     <name>com.ibm.omnifind.types.Pessoa</name>
     <indexRule>
       <style name="Annotation">
         <attribute name="fixedName" value="Pessoa" />
       </style>
     </indexRule>
  </indexBuildItem>
  ```

  Este processo permite consultas como ″Pretendo documentos em que Chefe está contido como um nome de pessoa″. A consulta é expressa da seguinte forma utilizando os fragmentos XML: @xmlf2::'<Pessoa>Chefe</Pessoa>'

- Utilize nameFeature se a anotação armazenar diferentes entidades a quem pretenda permitir o acesso utilizando diferentes expansões em função do valor de determinada funcionalidade da anotação. No seguinte exemplo, com.ibm.tt.EntityAnnotation é indexado como uma expansão de pessoa ou organização, em função do valor da funcionalidade denominada type. A funcionalidade pode também ser um caminho de funcionalidade.

  ```
  <indexBuildItem>
     <name>com.ibm.tt.EntityAnotation</name>
     <indexRule>
       <style name="Annotation">
         <attribute name="nameFeature" value="type" />
       </style>
     </indexRule>
  </indexBuildItem>
  ```

  Este processo permite consultas como ″Pretendo documentos sobre a organização QUEM″ (em oposição ao termo português ″quem″). A consulta é expressa da seguinte forma na sintaxe XPath limitada: @xmlp::'/organização[ftcontains="QUEM"]'

- Se nenhum dos atributos acima indicado for utilizado, é utilizado o nome abreviado do tipo de anotação no elemento <indexBuildItem>. Esta é a predefinição. Por exemplo:

```
<indexBuildItem>
   <name>com.ibm.uima.tutorial.NúmeroSala</name>
   <indexRule>
     <style name="Annotation" />
     <style name="Field" />
   </indexRule>
 </indexBuildItem>
```

Este elemento <indexBuildItem> resulta nas anotações e campos com o nome NúmeroSala preenchidos com o texto abrangido por com.ibm.uima.tutorial.NúmeroSala.

Elemento <style name="Annotation" />

A anotação (annotation) no elemento <style> especifica como pode aceder a informações de expansão no Enterprise Search. Para além de permitir a utilização dos atributos fixedName e nameFeature, este estilo também suporta o elemento <attributemappings>. No âmbito deste elemento, é possível mapear o valor de uma funcionalidade para um atributo da expansão resultante no índice, que pode utilizar subsequentemente numa expressão de procura.

Cada mapeamento é efectuado no âmbito de um elemento <mapping> separado. O elemento <feature> contém um caminho de funcionalidade e o elemento <indexName> contém o nome do atributo que é utilizado no índice para armazenar o valor de <feature>. Por exemplo,

```
<mapping>
  <feature>marca/nomeEmpresa</feature>
  <indexName>empresa</indexName>
</mapping>
```

Este elemento <mapping> armazena o valor da funcionalidade no caminho marca/nomeEmpresa directamente no atributo de índice empresa.

O mapeamento dos valores da funcionalidade para atributos de índice é especialmente útil se o sistema tipo utilizado durante a análise de texto for complexo, incluindo muitas estruturas funcionais imbricadas. Ao utilizar o elemento <mapping>, os atributos relevantes podem ser expostos, permitindo-lhe utilizá-los nas consultas sem o conhecimento detalhado da estrutura do sistema tipo original.

Elemento <style name="Field" />

O campo (field) no elemento <style> especifica como pode aceder a informações de campo no Enterprise Search. Para além dos atributos fixedName e nameFeature, pode definir os seguintes atributos.

parametric
: Se estiver definido como verdadeiro (true), o valor do campo pode ser procurado utilizando a procura paramétrica, por exemplo, #dosagem:>100.

fieldSearchable
: Se estiver definido como verdadeiro (true), o valor do campo pode ser utilizado na procura, por exemplo, marca:Bayer.

returnable
: Se estiver definido como verdadeiro (true), o campo e respectivos valores são devolvidos no resultado da procura.

As informações de campo são sempre passíveis de serem procuradas, isto é, as informações de campo estão acessíveis nas procuras normais de palavras-chave.

O atributo opcional `valueFeature` define qual o valor de funcionalidade a utilizar como valor de campo. Se a estrutura funcional for uma anotação e o atributo não estiver definido, o texto abrangido da anotação é utilizado como valor de campo. No exemplo,

```
<indexBuildItem>
   <name>com.ibm.omnifind.types.Data</name>
   <indexRule>
     <style name="Field">
       <attribute name="fixedName" value="data"/>
       <attribute name="fieldSearchable"
           value="true"/>
       <attribute name="returnable" value="true"/>
     </style>
     <style name="Field">
       <attribute name="fixedName" value="hora"/>
       <attribute name="valueFeature"  value="hora"/>
       <attribute name="parametric" value="true"/>
     </style>
   </indexRule>
   <filter syntax="FeatureValue">ano="2005"</filter>
</indexBuildItem>
```

são gerados dois campos para `com.ibm.omnifind.types.Data`. Um campo com o nome `data` contém o texto abrangido, por exemplo, 17:15. Outro campo contém o valor do atributo `hora`. Aqui pode efectuar uma consulta utilizando 'hora::<17'.

Elemento `<style name="Breaking" />`

O valor de quebra (Breaking) no elemento `<style>` não inclui quaisquer elementos.

Após criar o ficheiro XML, tem de carregá-lo no Enterprise Search e seleccionar o ficheiro de mapeamento da estrutura de análise comum para o índice com as outras selecções de análise personalizada utilizando a consola de administração do Enterprise Search.

Conceitos relacionados

"Indexar o mapeamento para resultados da análise personalizada" na página 37
Após executar a análise personalizada numa colecção de documentos, pode utilizar o motor de procura no Enterprise Search para criar um índice a partir de informações armazenadas na estrutura de análise comum que é criada pelos algoritmos de análise personalizada.

"Caminhos de funcionalidade" na página 32
Um caminho de funcionalidade fornece uma forma de aceder aos valores de funcionalidade nas estruturas de análise comum, semelhantes às instruções XPath utilizadas para aceder aos elementos de XML num documento XML.

Referências relacionadas

"Filtros" na página 36
Os filtros são utilizados para restringir regras de mapeamento nos ficheiros de mapeamento da estrutura de análise comum para o índice e nos ficheiros de mapeamento da estrutura de análise comum para a base de dados. Apenas quando o filtro está definido como verdadeiro (true) são adicionados os resultados da análise ao índice ou a uma tabela JDBC.

"Amostra da descrição do sistema tipo" na página 22
A descrição do sistema tipo descreve as estruturas funcionais (as estruturas de dados subjacentes que representam os resultados da análise) utilizadas na análise personalizada.

Mapeamento da base de dados para os resultados da análise seleccionada

Depois de ter analisado os documentos no Enterprise Search, pode armazenar resultados da análise de texto seleccionado numa base de dados que suporte JDBC.

Esta versão suporta apenas DB2 Universal Database, versão 8.2.2 (com.ibm.db2.jcc.DB2Driver versão 2.3) ou superior e Oracle 10g (oracle.jdbc.driver.OracleDriver versão 1.0).

Para DB2 Universal Database e Oracle, pode optar por inserir os resultados da análise directamente na base de dados ou por gerar os ficheiros equivalentes de carregamento específicos da base de dados e o script correspondente que executa os comandos de carregamento.

O mapeamento dos resultados da análise para tabelas na base de dados permite-lhe utilizar estas informações nos passos de processamento de informações empresariais subsequentes ou aceder directamente às partes relevantes de um documento que correspondam à consulta de procura semântica.

O ficheiro de mapeamento da estrutura de análise comum para a base de dados contém informações de configuração de ligação da base de dados e descreve quais os resultados da análise personalizada que deverão ser armazenados em determinadas tabelas e colunas. Os nomes de tabela e coluna no ficheiro de mapeamento têm de corresponder às tabelas e colunas que são criados na base de dados.

Depois de ter escrito o ficheiro de mapeamento da estrutura de análise comum para a base de dados, pode carregá-lo para o Enterprise Search utilizando a consola de administração.

Tarefas relacionadas

"Criar o ficheiro de mapeamento da estrutura de análise comum para a base de dados" na página 47
Para adicionar resultados de análise a uma base de dados, tem de criar ficheiro de mapeamento da estrutura de análise comum para a base de dados que contenha as informações de configuração da ligação da base de dados e uma descrição dos resultados da análise de texto personalizada que deverão ser armazenados e em que tabelas e colunas.

Armazenar resultados da análise numa base de dados

Para armazenar os resultados da análise seleccionada numa base de dados que suporte JDBC, tem de escrever o ficheiro de mapeamento da estrutura de análise comum para a base de dados que define quais os resultados da análise a armazenar numa base de dados e as bibliotecas do controlador JDBC necessárias têm de estar no caminho definido no ficheiro de mapeamento.

Para armazenar resultados da análise numa base de dados que suporte JDBC:

1. Decida quais os resultados da procura que pretende armazenar na base de dados. Crie uma base de dados que contenha as tabelas com todas as colunas necessárias dos tipos de dados apropriados.
2. Num editor de XML, escreva o ficheiro de mapeamento da estrutura de análise comum para a base de dados com os dados de configuração da base de dados e os resultados da análise que pretende armazenar. Para determinar quais os resultados da análise a incluir no ficheiro de mapeamento, tem de conhecer o sistema tipo subjacente que é utilizado quando os documentos são processados.
3. Coloque as bibliotecas do controlador JDBC num directório no nó indexador para que possam ser acedidas pelo sistema do Enterprise Search.
4. Carregue e seleccione o ficheiro de mapeamento utilizando a consola de administração do Enterprise Search.

Utilizar conjuntos de ficheiros de carregamento

Pode armazenar resultados da análise directamente numa base de dados que suporte JDBC ou pode configurar o processamento para utilizar os conjuntos de ficheiros de carregamento e carregar os dados uma base de dados numa fase posterior.

A utilização de conjuntos de ficheiros de carregamento tem as seguintes vantagens:

- No total, um conjunto de ficheiros de carregamento nunca pode ser maior do que o tamanho do ficheiro máximo suportado pelo sistema operativo
- Pode começar a carregar dados numa base de dados assim que um conjunto de ficheiros de carregamento fique cheio e não tem de parar e reiniciar o analisador de documentos para evitar conflitos do acesso ao ficheiro

A mudança de um conjunto de ficheiros de carregamento para o seguinte é efectuada ao nível de um documento, mesmo que o documento esteja dividido em partes nas várias estruturas de análise comum. Após um documento ter sido processado e se um ficheiro de carregamento no conjunto de ficheiros de carregamento actual exceder o limite definido, é utilizado um novo ficheiro de carregamento. Deste modo, garante a consistência do conjunto de ficheiros de carregamento. Após o conteúdo de um conjunto de ficheiros de carregamento ser carregado na base de dados, o modelo de dados permanece consistente, uma vez que todas as entradas na tabela principal contêm as entradas correspondentes na tabela de bases de dados.

Os ficheiros de carregamento e ficheiros de script são identificados pela extensão .cur do ficheiro. Quando um conjunto de ficheiros de carregamento for fechado, os ficheiros mudam de nome para ter a extensão .dat. Este processo indica que os ficheiros podem ser copiados ou movidos para um servidor de base de dados enquanto o analisador de documentos ainda está em execução.

Pode especificar o tamanho de um ficheiro de carregamento. Quando o limite do tamanho do ficheiro de carregamento for alcançado, é iniciado um novo conjunto de ficheiros de carregamento. O tamanho do ficheiro de carregamento é especificado no ficheiro de mapeamento da estrutura de análise comum para a base de dados na secção do elemento XML `<loadFile>`. O parâmetro loadFileSize é definido utilizando o elemento `<loadFileSize>` e é especificado em megabytes com 10 <= loadFileSize <= 10240 (10MB <= loadFileSize <= 10GB). O elemento `<loadFileSize>` é opcional. Se nenhum valor for definido, o valor predefinido é 1024MB (1GB).

Os ficheiros de carregamento isolados num conjunto são numerados utilizando um número de dez dígitos que identifica qual ficheiro pertence a que determinado conjunto de ficheiros de carregamento. Um conjunto de ficheiros de carregamento é fechado quando:

- Um ficheiro de carregamento no conjunto excede o limite de tamanho definido
- O processamento parou porque o analisador parou ou ocorreu um erro

Se o analisador for reiniciado, o processamento continua a partir do momento em que parou utilizando um novo conjunto de ficheiros de carregamento.

Criar o ficheiro de mapeamento da estrutura de análise comum para a base de dados

Para adicionar resultados de análise a uma base de dados, tem de criar ficheiro de mapeamento da estrutura de análise comum para a base de dados que contenha as informações de configuração da ligação da base de dados e uma descrição dos resultados da análise de texto personalizada que deverão ser armazenados e em que tabelas e colunas.

Acerca desta tarefa

O ficheiro de mapeamento da estrutura de análise comum para a base de dados está no XML. A amostra que se segue baseia-se no sistema tipo definido para o cenário de um relatório policial.

No exemplo, apenas os relatórios policiais e as cidades que aparecem nesses relatórios criminais são adicionados à base de dados. O exemplo mostra a utilização de funcionalidades incorporadas e o mapeamento do elemento `<constant>`.

```
<?xml version="1.0" encoding="UTF-8"?>
<cas2JdbcConfiguration xmlns="http://www.ibm.com/uima/consumer/jdbc/100/xml">
   <databaseConnection>
     <connectionUrl>db2://minhaMáquina:minhaPorta/minhaBaseDados</connectionUrl>
     <driver type="jdbc">com.ibm.db2.jcc.DB2Driver</driver>

     <driverLibraries>
       <driverLibrary>C:\db2\db2jcc.jar</driverLibrary>
       <driverLibrary>C:\db2\db2jcc_license_cu.jar</driverLibrary>
       <driverLibrary>C:\db2\db2jcc_license_cisuz.jar</driverLibrary>
     </driverLibraries>

     <authentication>
       <username>meuUtilizador</username>
       <password>minhaPalavra-passe</password>
     </authentication>

     <loadFile>
       <loadFileDirectory>/home/cas2jdbc/load/</loadFileDirectory>
<loadFileSize>1048</loadFileSize>
       <loadScript>/home/cas2jdbc/load/load.sh</loadScript>
     </loadFile>

   </databaseConnection>

   <cas2JdbcMappingSpec>
     <skipCondition>
       <name>com.ibm.uima.tt.DocumentAnnotation</name>
       <filter syntax="FeatureValue">toBeProcessed=0</filter>
     </skipCondition>

     <cas2JdbcMappings>
```

```
      <explicitMappings>
       <explicitMappingRule applyToSubtypes="false">
        <type>com.ibm.omnifind.types.RelatórioPolicial</type>
        <table>sample.RelatórioPolicial</table>
        <featureMappings>
         <featureMapping>
          <feature>uniqueId()</feature>
          <column>IDRelatórioPolicial</column>
         </featureMapping>
         <featureMapping>
          <feature>location/uniqueId()</feature>
          <column>IDlocalizaçãoCrime</column>
         </featureMapping>
        </featureMappings>
        <filter syntax="FeatureValue">location/coveredText()="Coimbra"</filter>
       </explicitMappingRule>
      </explicitMappings>

      <implicitMappings>
       <implicitMappingRule applyToSubtypes="false">
        <type>com.ibm.omnifind.types.Cidade</type>
         <table>sample.Cidade</table>
        <featureMappings>
         <featureMapping>
          <feature>uniqueId()</feature>
          <column>IDlocalizaçãoCrime</column>
         </featureMapping>
         <featureMapping>
          <feature>coveredText()</feature>
          <column>nomeCidade</column>
          <length>150</length>
         </featureMapping>
         <featureMapping>
          <constant>Portugal</constant>
          <column>país</column>
         </featureMapping>
        </featureMappings>
       </implicitMappingRule>
      </implicitMappings>

     </cas2JdbcMappings>
   </cas2JdbcMappingSpec>
</cas2JdbcConfiguration>
```

Procedimento

Para criar o ficheiro de mapeamento da estrutura de análise comum para a base de dados:

1. Crie um ficheiro XML. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML à sua escolha. O esquema XSD para o ficheiro de mapeamento denomina-se CasToJDBCMapping.xsd e encontra-se na instalação do Enterprise Search em *ES_INSTALL_ROOT*/packages/uima/configuration_xsd/.
2. Inclua os mapeamentos num elemento `<cas2JdbcConfiguration xmlns="http://www.ibm.com/uima/consumer/jdbc/100/xml">`. O espaço de nomes (especificado no atributo xmlns) tem de estar exactamente conforme mostrado.
3. Adicione um elemento `<databaseConnection>` que contenha todas as informações de configuração da ligação da base de dados e um elemento `<cas2JdbcMappingSpec>` que descreva as regras de mapeamento para os resultados da análise que são armazenados na base de dados ou ficheiros de carregamento.

4. Adicione os seguintes elementos componente ao elemento <databaseConnection>:
   - Obrigatório: Um elemento <connectionUrl>. Este elemento contém o URL de ligação da base de dados. Em funcionalidade da implementação do controlador JDBC, pode utilizar o acesso local ou remoto à base de dados.
   - Obrigatório: Um elemento <driver>. Este elemento contém o nome da classe do controlador JDBC, por exemplo, com.ibm.db2.jcc.DB2Driver para DB2 ou oracle.jdbc.driver.OracleDriver para Oracle.
   - Obrigatório: Um elemento <driverLibraries>. Este elemento contém as bibliotecas do controlador. Cada biblioteca está listada num elemento <driverLibrary>. As bibliotecas encontram-se no directório de instalação do DB2 ou Oracle. Para DB2, as bibliotecas são c:\dir_db2\db2jcc.jar, c:\dir_db2\db2jcc_license_cu.jar e c:\dir_db2\db2jcc_license_cisuz.jar. Para Oracle, a biblioteca a incluir é c:\dir_oracle\classes12.zip.

     Certifique-se de que as bibliotecas de controladores estão sempre ao mesmo nível de manutenção que o servidor de applets do DB2.
   - Obrigatório: Um elemento <authentication>. Este elemento contém o nome do utilizador e a palavra-passe para a base de dados.
   - Opcional: Um elemento <loadFile>. Este elemento contém os seguintes elementos componentes:
     - O directório do ficheiro de carregamento num elemento <loadFileDirectory>.
     - Opcional: O tamanho do ficheiro de carregamento num elemento <loadFileSize>. Os limites de tamanho do ficheiro de carregamento são 10 <= loadFileSize <= 10240 (10MB <= loadFileSize <= 10GB). Se nenhum valor for definido, a predefinição é 1024 MB (1GB).
     - O nome do script de carregamento num elemento <loadScript>.

     Se não especificar um elemento <loadFile>, todos os dados são armazenados directamente na base de dados utilizando JDBC.

     Tem também de adicionar todos os parâmetros de configuração da base de dados quando utiliza os ficheiros de carregamento específicos da base de dados e os scripts.
5. Adicione os seguintes elementos componente ao elemento <jdbcMappingSpec>:
   - Opcional: Um elemento <skipCondition>. Se não estiver definida qualquer condição para ignorar, todos os documentos são processados.

     ```
     <skipCondition>
       <name>com.ibm.uima.tt.DocumentAnnotation</name>
       <filter syntax="FeatureValue">toBeProcessed=0</filter>
     </skipCondition>
     ```

     No exemplo, os documentos que contêm uma anotação do tipo com.ibm.uima.tt.DocumentAnnotation com uma funcionalidade denominada toBeProcessed definida como zero não será considerada.
   - Um elemento <cas2JdbcMappings> que mostra quais os tipos e funcionalidades que são mapeados para determinadas tabelas de bases de dados e colunas. O elemento contém uma secção de mapeamentos explícitos e implícitos.
6. Adicione um elemento <explicitMappings>. Este elemento é obrigatório. Tem de ter um ou mais elementos <explicitMappingRule> que definam os mapeamentos explícitos e só pode ser definido para os tipos de anotação e respectivos subtipos. Se um mapeamento estiver definido na secção de

mapeamentos explícitos, todas as anotações que correspondem à definição de mapeamento serão armazenadas na base de dados.

7. Opcional: Adicione um elemento <implicitMappings>. Este elemento suporta todos os tipos de estrutura funcional. Se este elemento existir, tem de conter pelo menos um elemento <implicitMappingRule>. Os mapeamentos que estão definidos na secção de mapeamentos implícitos são adicionados à base de dados apenas se os tipos de anotação correspondente estiverem referenciados por outra anotação que possa corresponder a uma regra de mapeamento implícito ou explícito.

   O objectivo do mapeamento implícito consiste em permitir armazenar apenas os resultados da análise que aparecem em determinado contexto. Por exemplo, se o mapeamento para uma anotação do tipo com.ibm.omnifind.types.Cidade for implícito, apenas as cidades que são referidas pela definição do mapeamento com.ibm.omnifind.types.RelatórioPolicial na secção de mapeamento explícitos são armazenados na base de dados. O que significa que apenas as cidades mencionadas nos relatórios policiais são adicionadas à base de dados.

   Se existir uma regra de mapeamento explícito para a anotação Cidade, todas as cidades são adicionadas à base de dados. Em ambos os casos, se uma cidade for referida por vários relatórios policiais, é adicionada à base de dados apenas uma vez.

8. Os elementos <explicitMappingRule> e <implicitMappingRule> têm de conter o atributo applyToSubtypes, que, se estiver definido como verdadeiro (true), armazena não apenas a estrutura funcional e que está listada no elemento <type>, como também todas as estruturas funcionais daí derivadas. Adicione os seguintes elementos componente aos elementos <explicitMappingRule> e <implicitMappingRule>:
   - Um elemento <type> que contém o tipo de estrutura funcional.
   - Um elemento <table> que contém o esquema de base de dados e o nome da tabela. A sintaxe segue a regra esquema.nome_tabela ou apenas nome_tabela se nenhum esquema estiver definido.
   - Um elemento <featureMappings> com um ou mais elementos <featureMapping> ou um elemento <containerMapping>.
   - Opcional: Um elemento <filter> que contém uma condição que é avaliada de cada vez que a regra de mapeamento faz correspondência. Se a condição avaliar como verdadeiro (true), a anotação ou estrutura funcional é armazenada na base de dados. No exemplo, apenas os relatórios policiais que relatam crimes cometidos em Lisboa são armazenados na base de dados.

9. A estrutura componente do elemento <featureMapping> varia em função de estar ou não a mapear uma funcionalidade ou uma constante.

   Se estiver a mapear uma funcionalidade ou caminho de funcionalidade, os elementos componente incluem:
   - Um elemento <feature> com o nome da funcionalidade. A funcionalidade tem de ser definida para a estrutura funcional no elemento tipo (type). Pode também utilizar uma construção de caminho da funcionalidade ou qualquer das funcionalidades incorporadas definidas no sistema.
   - Opcional: Um elemento <length> com o comprimento que uma cadeia pode ter na coluna de base de dados especificada. As cadeias mais longas são truncadas.
   - Um elemento <column> com o nome da coluna na qual o valor da funcionalidade deve ser armazenado. As colunas da base de dados que não

são utilizadas em qualquer mapeamento da funcionalidade utilizam o valor predefinido (normalmente nulo) que está configurado na base de dados.

Certifique-se de que o valor do elemento funcionalidade (feature) está armazenado numa coluna do tipo apropriado. A seguinte tabela mostra quais os tipos de UIMA que correspondem a determinados tipos de bases de dados.

Tabela 3. Mapeamento entre tipos de UIMA e tipos de bases de dados

Tipo de UIMA ou funcionalidade incorporada	Tipo de dados DB2 recomendado	Tipo de dados Oracle recomendado
Float	REAL	FLOAT
String	VARCHAR	VARCHAR2
Integer	INTEGER	INTEGER
uniqueId(), uniqueParentId()	CHAR(27)	CHAR(27)
objectId(), parentId()	CHAR(16)	CHAR(16)
docTimestamp()	BIGINT	LONG
fsId()	INTEGER	INTEGER

Para obter uma constante, os elementos de mapeamento da funcionalidade do componente são os seguintes:

- Um elemento `<constant>` que contém o valor de uma constante.
- Um elemento `<column>` com o nome da coluna ao qual o valor da constante deve ser adicionado.

10. O elemento `<containerMapping>` contém o mapeamento para uma funcionalidade do tipo de contentor (matriz ou lista). Este elemento tem de ser utilizado apenas para os tipos de contentor. Tem os seguintes elementos componente:
    - Um elemento `<feature>` com o nome da funcionalidade. Pode também utilizar uma construção de caminho da funcionalidade ou qualquer das funcionalidades incorporadas definidas no sistema.
    - Um elemento `<table>` que contém o esquema de base de dados e o nome da tabela. A sintaxe segue a regra `esquema.nome_tabela` ou apenas `nome_tabela` se nenhum esquema estiver definido.
    - Um ou mais elementos `<featureMapping>` que contêm os nomes das estruturas funcionais e os nomes de colunas às quais as funcionalidades são adicionadas.
11. Guarde e valide o ficheiro XML utilizando o esquema fornecido.

Após criar o ficheiro XML, tem de carregá-lo no Enterprise Search e seleccionar o ficheiro de mapeamento da estrutura de análise comum para a base de dados com as outras selecções de análise personalizada utilizando a consola de administração do Enterprise Search.

Conceitos relacionados

"Mapeamento da base de dados para os resultados da análise seleccionada" na página 45
Depois de ter analisado os documentos no Enterprise Search, pode armazenar resultados da análise de texto seleccionado numa base de dados que suporte JDBC.

"Caminhos de funcionalidade" na página 32
Um caminho de funcionalidade fornece uma forma de aceder aos valores de

funcionalidade nas estruturas de análise comum, semelhantes às instruções XPath utilizadas para aceder aos elementos de XML num documento XML.

Referências relacionadas

"Filtros" na página 36
Os filtros são utilizados para restringir regras de mapeamento nos ficheiros de mapeamento da estrutura de análise comum para o índice e nos ficheiros de mapeamento da estrutura de análise comum para a base de dados. Apenas quando o filtro está definido como verdadeiro (true) são adicionados os resultados da análise ao índice ou a uma tabela JDBC.

"Funcionalidades incorporadas" na página 33
As funcionalidades incorporadas são nomes de funcionalidade predefinidos com semântica especial. Podem ser utilizadas para aceder a informações que não estão contidas na estrutura funcional em si, por exemplo, o tipo de estrutura funcional ou o texto abrangido de uma anotação. Podem ser utilizadas num caminho de funcionalidade que o último ou único elemento.

"Amostra da descrição do sistema tipo" na página 22
A descrição do sistema tipo descreve as estruturas funcionais (as estruturas de dados subjacentes que representam os resultados da análise) utilizadas na análise personalizada.

Mapeamento do tipo de contentor

O tipo de contentor é um dos tipos de matriz ou lista incorporados na estrutura de análise comum. O mapeamento do tipo de contentor é uma forma de mapeamento de valores de matriz ou lista para uma base de dados relacional.

Existem duas abordagens para processar tipos de contentor no ficheiro de mapeamento da estrutura de análise comum para a base de dados. Um método utiliza a construção da funcionalidade incorporada definida e uma tabela de ligação genérica que contém matrizes ou listas que são valores de uma regra de mapeamento da funcionalidade. À medida que diferentes matrizes ou listas são armazenadas na mesma tabela de ligação, a tabela não dá qualquer indicação sobre a relação das informações armazenadas.

No segundo método, a definição da tabela de ligação que é definida utilizando um elemento `<containerMapping>` denota explicitamente a relação entre as informações especificadas pretendidas.

Um exemplo do possível aspecto de um mapeamento de tabela de ligação genérica é apresentado de seguida. Existe uma relação n:m entre os relatórios policiais e as pessoas suspeitas, o que significa que um suspeito pode ser mencionado em mais do que relatório policial e um relatório policial pode mencionar mais do que um suspeito.

A tabela genérica `amostra.fsarray` no exemplo é a tabela de ligação entre os relatórios policiais e os suspeitos. Se existir outro tipo de mapeamento para além de `com.ibm.omnifind.types.RelatórioPolicial` com uma funcionalidade do tipo `com.ibm.omnifind.types.FSArray`, também é mapeado para esta tabela. Pode mesmo assim consultar a tabela para a relação entre um relatório policial e um suspeito correctamente, no entanto, não pode concluir, simplesmente olhando para a tabela, que contém a relação ou ligação entre os relatórios policiais e possíveis suspeitos.

```
<cas2JdbcMappings>
  <explicitMappings>
    <explicitMappingRule applyToSubtypes="false">
        <type>com.ibm.omnifind.types.RelatórioPolicial</type>
```

```
      <table>sample.RelatórioPolicial</table>
    <featureMappings>
      <featureMapping>
        <feature>uniqueId()</feature>
        <column>IDRelatórioPolicial</column>
      </featureMapping>
      <featureMapping>
        <feature>suspeitosConhecidos/uniqueId()</feature>
        <column>IDmatrizSuspeito</column>
      </featureMapping>
      <featureMapping>
        <feature>location/nomeCidade</feature>
        <column>cidade</column>
      </featureMapping>
    </featureMappings>
  </explicitMappingRule>
</explicitMappings>

<implicitMappings>
  <implicitMappingRule applyToSubtypes="false">
    <type>com.ibm.omnifind.types.Suspeito</type>
    <table>sample.suspeito</table>
    <featureMappings>
      <featureMapping>
        <feature>uniqueId()</feature>
        <column>IDsuspeito</column>
      </featureMapping>
      <featureMapping>
        <feature>sobrenome</feature>
        <column>apelido</column>
      </featureMapping>
      <featureMapping>
        <feature>descrição</feature>
        <column>descrição</column>
      </featureMapping>
    </featureMappings>
  </implicitMappingRule>

  <implicitMappingRule applyToSubtypes="false">
    <type>uima.cas.FSArray</type>
    <table>sample.fsarray</table>
    <featureMappings>
      <featureMapping>
        <feature>uniqueId()</feature>
        <column>arrayId</column>
      </featureMapping>
      <featureMapping>
        <feature>[:index]</feature>
        <column>arrayIndex</column>
      </featureMapping>
      <featureMapping>
        <feature>[]/uniqueId()</feature>
        <column>IDsuspeito</column>
      </featureMapping>
    </featureMappings>
  </implicitMappingRule>
</implicitMappings>

</cas2JdbcMappings>
```

A seguir são mostradas as tabelas de bases de dados baseadas nas regras genéricas de mapeamento acima indicadas.

Tabela 4. A tabela amostra.RelatórioPolicial

IDRelatórioPolicial	**IDmatrizSuspeito**	**cidade**
aaa...1	**bbb...1**	Sines
aaa...2	**bbb...2**	Lagos

Tabela 5. A tabela amostra.fsarray

arrayId	**arrayIndex**	**IDsuspeito**
bbb...1	1	*ccc...1*
bbb...1	2	*ccc...2*
bbb...2	1	*ccc...3*

Tabela 6. A tabela amostra.suspeito

IDsuspeito	**apelido**	**descrição**
ccc...1	Barreto	Compleição escura
ccc...2	Silva	Usa óculos
...	...	...

O exemplo mostra o mapeamento para as matrizes da estrutura funcional. Pode aplicar este tipo de mapeamento também a StringArray, IntegerArray, e FloatArray. Se incluir as regras de mapeamento para estas matrizes de valor simples, substitua `[]/uniqueId()` por `[]`.

Pode ser utilizada a mesma abordagem da tabela genérica para listas de estruturas funcionais, bem como para listas de tipos simples (StringList, IntegerList e FloatList).

Uma forma mais simples de processar relações consiste em utilizar um elemento de mapeamento de contentor explícito que define a iteração entre os elementos contidos nas matrizes ou listas.

Um exemplo de mapeamento que denota uma tabela de ligação explícita é apresentado de seguida. Novamente, existe uma relação n:m entre os relatórios policiais e as pessoas suspeitas. No entanto, desta vez, a tabela `amostra.relatórios_suspeitos` é a tabela de ligação entre os relatórios policiais e as pessoas suspeitas.

Nesta abordagem, não tem de se preocupar com o processamento de IDs de matriz nem com o mapeamento de entradas de cabeça ou cauda para os tipos de lista. A tabela de ligação contém uma relação explícita.

```
<cas2JdbcMappings>
  <explicitMappings>
    <explicitMappingRule applyToSubtypes="false">
        <type>com.ibm.omnifind.types.RelatórioPolicial</type>
        <table>sample.RelatórioPolicial</table>
      <featureMappings>
        <featureMapping>
          <feature>uniqueId()</feature>
          <column>IDRelatórioPolicial</column>
        </featureMapping>
        <featureMapping>
          <feature>location/nomeCidade</feature>
          <column>cidade</column>
```

```
        </featureMapping>
        <featureMapping>
          <feature>suspeitosConhecidos</feature>
          <containerMapping>
            <table>sample.relatórios_suspeitos</table>
            <featureMapping>
              <feature>com.ibm.omnifind.types.RelatórioPolicial
                /objectId()</feature>
              <column>IDRelatórioPolicial</column>
            </featureMapping>
            <featureMapping>
              <feature>suspeitosConhecidos/[]/objectId()</feature>
              <column>IDsuspeito</column>
            </featureMapping>
          </containerMapping>
        </featureMapping>
      </featureMappings>
    </explicitMappingRule>
  </explicitMappings>

  <implicitMappings>
    <implicitMappingRule applyToSubtypes="false">
      <type>com.ibm.omnifind.types.Suspeito</type>
      <table>sample.suspeito</table>
      <featureMappings>
        <featureMapping>
          <feature>objectId()</feature>
          <column>IDsuspeito</column>
        </featureMapping>
        <featureMapping>
          <feature>sobrenome</feature>
          <column>apelido</column>
        </featureMapping>
        <featureMapping>
          <feature>descrição</feature>
          <column>descrição</column>
        </featureMapping>
      </featureMappings>
    </implicitMappingRule>
  </implicitMappings>

</cas2JdbcMappings>
```

Um elemento `<containerMapping>` é utilizado para definir a iteração entre os elementos contidos na matriz. No exemplo, a tabela de ligação `amostra.relatórios_suspeitos` contém uma ligação para as colunas `IDRelatórioPolicial` e `IDsuspeito`. Não imbrique os elementos `<containerMapping>`.

A seguir são mostradas as tabelas de bases de dados baseadas nas regras explícitas de mapeamento da tabela de ligação.

Tabela 7. A tabela amostra.RelatórioPolicial

IDRelatórioPolicial	**cidade**
aaa...1	Sines
aaa...2	Lagos

Tabela 8. A tabela amostra.relatórios_suspeito

IDRelatórioPolicial	**IDsuspeito**
bbb...1	*ccc...1*

Tabela 8. A tabela amostra.relatórios_suspeito (continuação)

IDRelatórioPolicial	IDsuspeito
bbb...2	*ccc...2*
...	...

Tabela 9. A tabela amostra.suspeito

IDsuspeito	apelido	descrição
ccc...1	Barreto	Compleição escura
ccc...2	Silva	Usa óculos
...	...	...

Referências relacionadas

"Funcionalidades incorporadas" na página 33
As funcionalidades incorporadas são nomes de funcionalidade predefinidos com semântica especial. Podem ser utilizadas para aceder a informações que não estão contidas na estrutura funcional em si, por exemplo, o tipo de estrutura funcional ou o texto abrangido de uma anotação. Podem ser utilizadas num caminho de funcionalidade que o último ou único elemento.

Obter partes de um documento que correspondam a uma consulta de procura semântica

Pode obter apenas as partes de um documento que correspondam exactamente à consulta através do mapeamento das estruturas funcionais relevantes para o índice e base de dados e especificando a expansão na consulta de procura semântica.

Para aceder a todas as instâncias de um tipo de anotação específico no resultado da procura, por exemplo, para obter todas as pessoas, inclua um mapeamento de estilo de campo para o tipo de anotação e marque-o como passível de ser devolvido (returnable) no ficheiro de mapeamento da estrutura de análise comum para o índice. Por exemplo:

```
<indexBuildItem>
 <name>com.ibm.omnifind.types.Pessoa</name>
 <indexRule>
  <style name="Annotation"/>
  <style name="Field">
   <attribute name="returnable" value="true"/>
  </style>
 </indexRule>
</indexBuildItem>
```

Neste exemplo, as anotações do tipo `com.ibm.omnifind.types.Pessoa` são mapeadas para uma expansão com o nome Pessoa no índice do Enterprise Search, onde podem ser acedidos durante a procura semântica. Para além disso, o texto abrangido das anotações, por exemplo, os nomes completos das pessoas é armazenado como um campo passível de ser devolvido. Para obter estes valores de anotações, chame getFields("Pessoa") em cada objecto de resultado que seja devolvido a partir da consulta de procura (palavra-chave ou semântica). Este método devolve uma matriz Cadeia (String) com os valores das anotações, neste caso, os nomes da pessoas.

No entanto, esta abordagem devolve todas as instâncias de determinado tipo de anotação e não é adequada se limitar o processamento dos resultados a

documentos que correspondam exactamente à consulta. Por exemplo, um documento pode mencionar cinco pessoas. No entanto, na consulta de procura semântica '<sentence><pessoa/>IBM</sentence> o utilizador está interessado apenas na pessoa que é mencionada na mesma frase em que aparece o termo IBM. O utilizador não está interessado noutras pessoas.

Para aceder e processar estruturas funcionais que correspondam exactamente à consulta:

1. Mapeie os tipos de estrutura funcional relevantes para o índice do Enterprise Search utilizando o estilo de mapeamento da anotação. Por exemplo:

```
<indexBuildItem>
 <name>com.ibm.omnifind.types.Pessoa</name>
 <indexRule>
  <style name="Annotation"/>
 </indexRule>
</indexBuildItem>
```

2. Mapeie os tipos de estrutura funcional relevantes para as tabelas JDBC. Como parte do mapeamento, tem de incluir duas colunas para o URI do documento e para o ID da estrutura funcional. Apesar de poder mapear todos os tipos de estrutura funcional para a mesma tabela de bases de dados, deve mapear cada tipo para uma tabela diferente. Por exemplo:

```
<explicitMappingRule applyToSubtypes="false">
 <type>com.ibm.omnifind.types.Pessoa</type>
 <table>sample.pessoa</table>
 <featureMappings>
  <featureMapping>
   <feature>objectId()</feature>
   <column>primaryId</column>
  </featureMapping>
  <!-- Contém o texto abrangido da anotação-->
  <featureMapping>
   <feature>coveredText()</feature>
   <column>nomePessoa</column>
  </featureMapping>
  <!-- São incluídos aqui outros mapeamentos-->
  <!-- Para aceder às anotações de pessoa relevantes no resultado
  da consulta-->
  <featureMapping>
   <feature>docUri()</feature>
   <column>docUri</column>
  </featureMapping>
  <featureMapping>
   <feature>fsId()</feature>
   <column>annotationId</column>
  </featureMapping>
 </featureMappings>
</explicitMappingRule>
```

3. Pesquise, analise e indexe os documentos.
4. Obtenha os IDs das instâncias que correspondem à consulta. Na API de procura e de índice (SIAPI, Search and Index API), estas instâncias são referidas como elementos de destino. Um elemento de destino especifica a expansão de entrada para ser devolvida. É definida da seguinte forma:
   - Nos fragmentos de XML, o elemento de destino é identificado por um sinal de cardinal (#) adicionado como prefixo. O sinal de cardinal só é permitido uma vez e pode aparecer em qualquer ponto na consulta do fragmento de XML. Por exemplo: `$xmlf2::'<sentence><#pessoa/>IBM</sentence>'`
   - Em XPath por predefinição, o elemento de destino é o último campo na expressão XPath.

- Aceda a estas instâncias utilizando o método Result.getProperty(″ElementoDestino″). A propriedade devolvida é uma concatenação de cadeia de todos os IDs de ocorrências que são separados por espaços. Cada ocorrência na propriedade pode ser traduzida num valor inteiro.

5. A SIAPI não devolve as estruturas funcionais em si, apenas os respectivos IDs de ocorrências. Estes IDs correspondem ao valor fsId() que é armazenado na tabela de bases de dados. Para obter estas instâncias e respectivas informações associadas, é necessário que a aplicação:
   a. Seleccione a tabela de bases de dados correcta, em função do nome de expansão do elemento de destino. No exemplo, a aplicação contém um mapeamento de pessoa para a tabela amostra.Pessoa. Estas informações são deduzidas a partir do ficheiro de mapeamento da estrutura de análise comum para o índice, que entrega o nome da expansão, e do ficheiro de mapeamento da estrutura de análise comum para a base de dados, que entrega o nome da tabela.
   b. Para cada objecto do resultado no resultado da procura:
      1) Analise a cadeia que é devolvida por Result.getProperty (″ElementoDestino″) para localizar os IDs das ocorrências.
      2) Emita uma instrução SELECT para a tabela utilizando o URI do resultado (acessível utilizando Result.getDocumentId()) como valor na coluna docUri e os IDs das ocorrências como valor na coluna annotationId. Os nomes da coluna dependem do ficheiro de mapeamento. Os nomes da coluna são retirados do exemplo anterior.

      As linhas devolvidas contêm as informações que são armazenadas para a estrutura da funcionalidade, por exemplo, o texto abrangido ou atributos específicos da estrutura da funcionalidade, tais como, ″apelido″ ou ″local de nascimento″.

Certifique-se de que as actualizações efectuadas à base de dados sºão sincronizadas com as actualizações do índice no Enterprise Search. Se a base de dados contiver informações desactualizadas (por exemplo, porque utilizou ficheiros de carregamento da base de dados e não actualizou a base de dados, mas renovou ou reorganizou o índice), alguns IDs das ocorrências podem não ser encontrados na base de dados. O Enterprise Search mantém um registo apenas da última versão do documento no respectivo índice. Deste modo, os IDs da ocorrência são válidos apenas para o último documento.

Se armazenar várias versões do mesmo documento na mesma tabela de bases de dados, podem existir várias linhas que correspondem aos mesmos IDs das ocorrências, cada um para versões diferentes do documento. Neste caso, tem de definir uma coluna de versão do documento e preenchê-la utilizando a lógica da aplicação ou funcionalidades incorporadas como docTimestamp(). Deste modo, pode filtrar o resultado para obter apenas a versão do documento mais recente.

Conceitos relacionados

“Termo de consulta de procura semântica” na página 59
O termo de consulta de procura semântica é comunicado como um termo opaco.

Tarefas relacionadas

“Criar o ficheiro de mapeamento da estrutura de análise comum para o índice” na página 39
Ao utilizar o ficheiro de mapeamento da estrutura de análise comum para o

índice, pode determinar quais os resultados da análise na estrutura de análise comum que pretende indexar para activar a procura.

"Criar o ficheiro de mapeamento da estrutura de análise comum para a base de dados" na página 47
Para adicionar resultados de análise a uma base de dados, tem de criar ficheiro de mapeamento da estrutura de análise comum para a base de dados que contenha as informações de configuração da ligação da base de dados e uma descrição dos resultados da análise de texto personalizada que deverão ser armazenados e em que tabelas e colunas.

Aplicações de procura semântica

Quatro tipos de informações de documentos são armazenados no índice do Enterprise Search que pode consultar nas aplicações de procura utilizando a interface de API de procura e de índice (SIAPI, Search and Index API).

Os quatro tipos diferentes de informações incluem:

- Palavras de texto que são encontradas num documento, por exemplo, uma expressão como *software de computador*.
- Nomes de expansão, por exemplo, um documento XML que inclui <autor>José</autor>, fornece a expansão <autor>.
- Nomes de atributo, por exemplo, um documento XML que inclui <autor países DeOrigem=POR>José</autor>, fornece o atributo ″países DeOrigem″.
- Valores de atributo, por exemplo, POR é o valor do atributo ″países DeOrigem″.

O idioma de consulta SIAPI inclui o termo da consulta de procura semântica. O termo especifica um padrão de arbusto. Um arbusto é uma pequena árvore com folhas. Cada folha representa os quatro tipos de informações (palavras de texto, nomes de expansão, etc). Os nós internos da árvore especificam como a respectiva ocorrência num documento se relacionam entre si. Existem cinco tipos de nós internos que especificam as relações:

- and
- or
- not
- in_the_span_of
- attribute_in_the_span_of

Diz-se que um documento satisfaz determinado termo de procura semântica se incluir ocorrências das folhas e as restrições especificadas pelos nós internos (as relações definidas) forem respeitadas.

O termo de consulta de procura semântica ajuda a obter documentos de melhor qualidade. Não só é possível efectuar a procura utilizando combinações booleanas da palavra e anotações, como também obter documentos em que, por exemplo, *José* aparece no autor com o nome de expansão ou em que os termos *ibm* e *procura* aparecem na mesma frase.

Termo de consulta de procura semântica

O termo de consulta de procura semântica é comunicado como um termo opaco.

Há duas formas de sintaxe para expressar um termo opaco na API de procura e de índice (SIAPI, Search and Index API):

- Fragmentos XML

- XPath limitado

O termo de consulta de fragmento XML parece um fragmento bem balanceado de um documento XML. Um termo de consulta de fragmento XML tem como prefixo o sinal de termo opaco `@xmlf2::`, seguido pela expressão de fragmento XML entre plicas ('...').

No entanto, os termos de consulta XPath limitada têm como prefixo `@xmlxp::`, seguido pela consulta XPath entre plicas ('...').

Tal como com termos de consulta geral na interface de API de procura e de índice (SIAPI, Search and Index API), cada termo pode ter um modificador de aspecto:

Sinal de adição (+)
O termo tem de aparecer.

Prefixo =
O termo tem de corresponder exactamente.

Prefixo til (~)
Tem em consideração sinónimos do termo de consulta.

Sufixo til (~)
Tem em consideração palavras com o mesmo lema do termo de consulta.

Sinal de Cardinal (#)
O termo é evidenciado.

Os seguintes exemplos mostram consultas de fragmento XML:

@xmlf2::'<Cidade>Sines</Cidade>'
Encontra documentos que incluem a expansão (anotação) ″Cidade″ contendo a cadeia ″Sines″.

@xmlf2::'<Pessoa género=″feminino″/>'
Encontra documentos em que uma pessoa do género feminino está anotada.

@xmlf2::'<Pessoa><.or><@género>feminino</@género> <@título>Sra</@título><@título>Mna</@título><⁄.or></Pessoa>'
Encontra documentos que especificam uma pessoa como sendo uma mulher, pelo género e pelo título.

@xmlf2::'<Pessoa género=″masculino″ função=″suspeito″/> <RelatórioPolicial><@descriçãoCrime><.or>assalto roubo</.or>-acidente </@descriçãoCrime></RelatórioPolicial> <Cidade>Sines<.or> <@zona>Porto Covo</@zona><@zona>Grândola</@zona></.or></Cidade>'
Encontra documentos que especificam pessoas do género masculino consideradas como suspeitos e uma anotação RelatórioPolicial que é atribuída pela cadeia *assalto* ou *roubo* no atributo ″descriçãoCrime″, mas não a cadeia *acidente*. Os documentos têm também de conter uma anotação ″cidade″ que abrange a palavra do texto *Sines*, uma anotação atribuída com a zona *Porto Covo* ou *Grândola*.

As consultas XPath correspondentes têm a seguinte estrutura:

@xmlxp::'//Cidade ftcontains (″Sines″)'
Encontra documentos que incluem a expansão (anotação) ″Cidade″ contendo a cadeia *Sines*.

@xmlxp::'//RelatórioPolicial[Cidade ftcontains(″Sines″)]'

Encontra documentos que incluem a expansão (anotação) ″Cidade″ na expansão ″RelatórioPolicial″ contendo a cadeia *Sines*.

@xmlxp::'//Pessoa[@género=″feminino″ ou @título ftcontains(″Mna″) ou @título ftcontains(″Sra″)]'

Encontra documentos em que uma pessoa do género feminino está anotada. No atributo género, o valor tem de corresponder exactamente, ao passo que para o atributo título, *Mna* e *Sra* não é necessário corresponder exactamente ao valor do atributo.

Suporte de sinónimos em aplicações de procura

Pode expandir os resultados da procura procurando documentos que contêm sinónimos dos termos da consulta.

Normalmente, os sinónimos incluem termos de várias palavras, tais como, nomes de produtos como *WebSphere Information Integrator OmniFind*. Os termos de várias palavras contidos no dicionário de sinónimos estão correctamente identificados nas consultas do utilizador e não têm de aparecer entre aspas.

A API de Procura e de Índice (SIAPI, Search and Index API) para o Enterprise Search suporta várias formas dos utilizadores procurarem sinónimos dos termos da consulta:

- A sintaxe da consulta SIAPI suporta o operador til (~) para expansão de sinónimos. Se o utilizador adicionar como prefixo este operador a um termo de consulta, a expansão de sinónimos é efectuada para a palavra. Por exemplo, a consulta `~WAS` devolve documentos que se referem a WebSphere Application Server e a quaisquer outros sinónimos que existem para esta abreviatura.
- A expansão de sinónimos pode ser activada utilizando a interface de expansão de sinónimos SIAPI a partir de uma aplicação de procura. Os termos da consulta podem ser expandidos automaticamente para incluir sinónimos ou a aplicação de procura pode incluir as opções que permitem ao utilizador especificar se os sinónimos dos termos da consulta deverão ser devolvidos nos resultados da procura.

  Durante a expansão automática de sinónimos, a procura de sinónimos é efectuada em todas as palavras da consulta. Os resultados da procura incluem documentos que contêm termos da consulta ou sinónimos dos termos da consulta. A SIAPI também suporta a conversão de uma lista de expansões de sinónimos para a consulta submetida.

Não utilize o suporte de sinónimos para texto que seja processado utilizando a segmentação n-grama.

Criar um ficheiro XML para sinónimos

Para expandir consultas no Enterprise Search de modo a incluir sinónimos dos termos da consulta, tem de especificar quais as palavras que se qualificam como sinónimos entre si num ficheiro XML. Este ficheiro XML é utilizado para criar um ficheiro de dicionário binário que carrega para o Enterprise Search e atribui às colecções apropriadas.

Acerca desta tarefa

O ficheiro XML que lista os sinónimos tem de estar em conformidade com um esquema específico. Trata-se de um exemplo de ficheiro XML para sinónimos:

```
<?xml version="1.0" encoding="UTF-8"?>
<synonymgroups xmlns="http://www.ibm.com/of/822/synonym/xml">
  <synonymgroup>
    <synonym>Think Pad</synonym>
    <synonym>Notebook</synonym>
    <synonym>Notebooks</synonym>
  </synonymgroup>
  <synonymgroup>
```

```
    <synonym>WebSphere Application Server</synonym>
    <synonym>WAS</synonym>
  </synonymgroup>
</synonymgroups>
```

Restrições

Tem de agrupar palavras que são sinónimos entre si (os elementos <synonym>) num elemento <synonymgroup>. Um sinónimo pode incluir espaços em branco, mas não pode incluir caracteres de pontuação, tais como, uma vírgula (,) ou barra vertical (|), uma vez que estes caracteres podem interferir com a sintaxe de consulta do Enterprise Search.

Tem enumerar todas as inflexões possíveis dos termos que adiciona como sinónimos, tais como as formas do singular e plural de uma palavra. Não é necessário enumerar as normalizações do termo, tais como, a remoção de acentos ou tremas (o Enterprise Search processa a normalização automaticamente), nem incluir variantes de minúsculas e minúsculas do termo. Por exemplo, se pretender incluir o termo météo como sinónimo, não necessita de incluir também o termo METEO.

Procedimento

Para criar uma lista de sinónimos para o Enterprise Search:

1. Crie um ficheiro XML. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML à sua escolha. O esquema XSD para o ficheiro XML denomina-se synonyms.xsd e encontra-se na instalação do Enterprise Search em *ES_INSTALL_ROOT*/packages/uima/configuration_xsd/.
2. Adicione um elemento <synonymgroup> e, em seguida, insira um elemento <synonym> para cada palavra que deva ser tratada como um sinónimo de outras palavras no grupo de sinónimos.

   Certifique-se de que inclui os mapeamentos num elemento <synonymgroups xmlns="http://www.ibm.com/of/822/synonym/xml">. O espaço de nomes (especificado no atributo xmlns) necessita de estar exactamente conforme mostrado.
3. Repita o passo precedente até especificar todos os sinónimos que pretende utilizar para procurar documentos numa colecção do Enterprise Search.
4. Guarde e saia do ficheiro XML.

Após criar o ficheiro XML, tem de convertê-lo para um dicionário de sinónimos de forma a que possa ser adicionado ao sistema do Enterprise Search.

Criar um dicionário de sinónimos

Após criar ou actualizar uma lista de sinónimos num ficheiro XML, tem de converter o ficheiro XML num dicionário de sinónimos binário.

Acerca desta tarefa

Para criar um dicionário de sinónimos, utilize a ferramenta de linha de comandos denominada essyndictbuilder, que é fornecida com o WebSphere II OmniFind Edition. A ferramenta está no directório *ES_INSTALL_ROOT*/bin.

A entrada na ferramenta é o ficheiro XML que lista os sinónimos e a saída da ferramenta é um dicionário de sinónimos. O dicionário tem de ter o sufixo `.dic`. Por exemplo, `c:\meusdicionários\produtos.dic`.

A localização predefinida para ambos os ficheiros é o directório onde o script é invocado. Se existir um dicionário com o mesmo nome, o script produz um erro.

O tamanho máximo de um `.dic` no Enterprise Search é de 8 MB.

Procedimento

Para criar um dicionário de sinónimos para o Enterprise Search:

1. No servidor de índices, inicie sessão como administrador do Enterprise Search. Este ID de utilizador foi especificado quando o WebSphere II OmniFind Edition foi instalado.
2. Introduza o seguinte comando, em que *ficheiro_XML* é o caminho totalmente qualificado para o ficheiro XML que contém a lista de sinónimos e *ficheiro_DIC* é o caminho totalmente qualificado para o dicionário de sinónimos.

   AIX, Linux ou Solaris: `essyndictbuilder.sh` *ficheiro_XML ficheiro_DIC*
   Windows: `essyndictbuilder.bat` *ficheiro_XML ficheiro_DIC*

Após criar um dicionário de sinónimos, utilize a consola de administração do Enterprise Search para adicionar o dicionário ao sistema do Enterprise Search e associá-lo a uma ou mais colecções.

Apenas o ficheiro .dic gerado é carregado no sistema do Enterprise Search. Certifique-se de que o ficheiro XML de origem é mantido num ambiente de acesso controlado e de que efectua regularmente cópia de segurança do ficheiro. O utilizador necessita deste ficheiro XML para actualizar o dicionário de sinónimos.

Dicionários de palavras de paragem personalizados

Pode definir o vocabulário específico da empresa que é removido de uma consulta para aumentar a relevância da procura.

Existem dois tipos de suporte de palavras de paragem no Enterprise Search:

- O reconhecimento da palavra de paragem específico de idioma que remove todas as palavras utilizadas frequentemente como *um* e *o* de uma consulta de várias palavras. O dicionário de palavras de paragem que existe para cada idioma não pode ser modificado pelos utilizadores. Este reconhecimento de palavras de paragem é executado automaticamente em todas as consultas para melhorar a relevância da procura.
- O reconhecimento de palavras de paragem personalizado ou definido pelo utilizador que remove vocabulário específico da empresa das consultas. Este dicionário de palavras de paragem, definido pelo administrador, pode conter apenas vocabulário especial. O dicionário de palavras de paragem definido pelo utilizador não substitui os dicionários de palavras de paragem específicos de idioma do Enterprise Search que contêm palavras comuns. Os dicionários de palavras de paragem definidos pelo utilizador são independentes do idioma.

Normalmente, as palavras de paragem definidas pelo utilizador incluem termos de várias palavras, tais como, nomes de produtos como *WebSphere Information Integrator OmniFind*. Os termos de várias palavras contidos no dicionário de palavras de paragem estão correctamente identificados nas consultas do utilizador e não têm de aparecer entre aspas.

Os termos compostos de idiomas germânicos estão também correctamente identificados nas consultas. Um termo composto é a combinação de duas ou mais palavras que é utilizado como uma única palavra. Os compostos lexicais como *Reisebüro* (agência de viagens) não são considerados compostos.

Os termos compostos numa consulta são divididos em termos individuais que constituem o composto. Se qualquer dos termos individuais que constitua o composto estiver no dicionário de palavras de paragem, o termo composto não é removido da consulta.

Por exemplo, o termo de consulta `Versicherungspolice` (apólice de seguro) devolve documentos que contêm os termos compostos *Lebensversicherungspolice* (apólice de seguro de vida) e *Haftpflichtversicherungspolice* (apólice de seguro contra terceiros). Mesmo que a palavra *Police* esteja listada no dicionário de palavras de paragem, o termo de consulta composto `Versicherungspolice` não é removido da consulta.

Tem de listar vocabulário específico da empresa num ficheiro XML que tem de converter para um dicionário de palavras de paragem de forma a que possa ser adicionado ao sistema do Enterprise Search.

Pode seleccionar qual o dicionário de palavras de paragem a utilizar na consola de administração do Enterprise Search. Pode seleccionar um dicionário de palavras de paragem para cada colecção. Um dicionário de palavras de paragem pode ser partilhado por várias colecções.

Criar um ficheiro XML para palavras de paragem

Para remover vocabulário específico da empresa a partir de consultas, tem de especificar quais as palavras que se qualificam como palavras de paragem num ficheiro XML.

Acerca desta tarefa

O ficheiro XML que lista as palavras de paragem tem de estar em conformidade com um esquema específico referido no documento XML. Este é um exemplo de um ficheiro XML para palavras de paragem:

```
<?xml version="1.0" encoding="UTF-8"?>
<stopWords xmlns="http://www.ibm.com/of/83/stopwordbuilder/xml">
   <stopWord>OmniFind Edition</stopWord>
   <stopWord>WAS</stopWord>
   <stopWord>...</stopWord>
</stopWords>
```

Restrições

Uma palavra de paragem pode incluir espaços em branco, mas não pode incluir caracteres de pontuação, tais como, uma vírgula (,) ou barra vertical (|), uma vez que estes caracteres podem interferir com a sintaxe de consulta do Enterprise Search.

Não é necessário enumerar as normalizações do termo, tais como, a remoção de acentos ou tremas (o Enterprise Search processa a normalização automaticamente). Por exemplo, se pretender incluir o termo météo como palavra de paragem, não necessita de incluir também o termo METEO.

Procedimento

Para criar uma lista de palavras de paragem para o Enterprise Search:

1. Crie um ficheiro XML. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML que pode validar o XML. O esquema XSD para o ficheiro XML denomina-se stopWords.xsd e encontra-se na instalação do Enterprise Search em *ES_INSTALL_ROOT*/packages/uima/configuration_xsd/.
2. Adicione um elemento `<stopWord>` para cada palavra que deva ser tratada como uma palavra de paragem.

   Certifique-se de que inclui os mapeamentos num elemento `<stopWords xmlns="http://www.ibm.com/of/83/stopwordbuilder/xml">`. O espaço de nomes (especificado no atributo xmlns) necessita de estar exactamente conforme mostrado.
3. Repita o passo precedente até especificar todas as palavras de paragem que pretende remover das consultas quando os utilizadores procuram colecções do Enterprise Search.
4. Guarde e saia do ficheiro XML.

Após criar o ficheiro XML, tem de convertê-lo para um dicionário de palavras de paragem de forma a que possa ser adicionado ao sistema do Enterprise Search.

Criar um dicionário de palavras de paragem

Após criar ou actualizar uma lista de palavras de paragem num ficheiro XML, tem de converter o ficheiro XML num dicionário de palavras de paragem.

Acerca desta tarefa

Para criar um dicionário de palavras de paragem, utilize a ferramenta de linha de comandos denominada esstopworddictbuilder, que é fornecida com o WebSphere II OmniFind Edition. A ferramenta está no directório *ES_INSTALL_ROOT*/bin.

A entrada na ferramenta é o ficheiro XML que lista as palavras de paragem e a saída da ferramenta é um dicionário de palavras de paragem. O dicionário tem de ter o sufixo `.dic`. Por exemplo, `c:\meusdicionários\palavrasparagemproduto.dic`.

A localização predefinida para ambos os ficheiros é o directório onde o script é invocado. Se existir um dicionário com o mesmo nome, o script produz um erro.

O tamanho máximo de um `.dic` no Enterprise Search é de 8 MB.

Procedimento

Para criar um dicionário de palavras de paragem para o Enterprise Search:

1. No servidor de índices, inicie sessão como administrador do Enterprise Search. Este ID de utilizador foi especificado quando o WebSphere II OmniFind Edition foi instalado.
2. Introduza o seguinte comando, em que *ficheiro_XML* é o caminho totalmente qualificado para o ficheiro XML que contém a lista de palavras de paragem e *ficheiro_DIC* é o caminho totalmente qualificado para o dicionário de palavras de paragem.

   AIX, Linux ou Solaris: `esstopworddictbuilder.sh` *ficheiro_XML ficheiro_DIC*
   Windows: `esstopworddictbuilder.bat` *ficheiro_XML ficheiro_DIC*

Após criar um dicionário de palavras de paragem, utilize a consola de administração do Enterprise Search para adicionar o dicionário ao sistema do Enterprise Search e associá-lo a uma ou mais colecções.

Apenas o ficheiro .dic gerado é carregado no sistema do Enterprise Search. Certifique-se de que o ficheiro XML de origem é mantido num ambiente de acesso controlado e de que efectua regularmente cópia de segurança do ficheiro. O utilizador necessita deste ficheiro XML para actualizar o dicionário de palavras de paragem.

Dicionários de palavras hierárquicas personalizados

Pode definir termos específicos ou termos de várias palavras que aumentam ou diminuem o valor da classificação do documento no qual aparece o termo.

Cada termo no dicionário hierárquico está associado a um factor hierárquico que pode ter um intervalo entre -10 e +10. Aos termos que pretende ver em particular nos documentos resultantes são alocados um factor hierárquico mais elevado, enquanto que àqueles que não pretende que sejam apresentados ou em combinação com termos hierárquicos mais elevados é concedido um valor mais baixo. Os valores -1, 0 e 1 não tem efeito hierárquico.

Se um termo de consulta que está listado no dicionário hierárquico com determinado factor hierárquico aparece num documento obtido, o valor da classificação do documento é aumentado ou diminuído em função do valor hierárquico. O valor hierárquico atribuído a um termo é relativo uma vez que também é afectado por outros factores. Deste modo, se o termo X aumentar por B1 e o termo Y por B2, e B1 > B2, nesse caso, hierarquia(X) >= hierarquia(Y).

Normalmente, uma palavra hierárquica inclui termos de várias palavras, tais como, nomes de produtos como *WebSphere Information Integrator OmniFind*. Os termos de várias palavras contidos no dicionário de palavras hierárquicas estão correctamente identificados nas consultas do utilizador e não têm de aparecer entre aspas.

Os dicionários de palavras hierárquicas são independentes do idioma.

Os termos compostos de idiomas germânicos estão também correctamente identificados nas consultas. Um termo composto é a combinação de duas ou mais palavras que é utilizado como uma única palavra. Os compostos lexicais como *Reisebüro* (agência de viagens) não são considerados compostos.

Os termos compostos numa consulta são divididos em termos individuais que constituem o composto. Se existirem valores hierárquicos em função dos termos individuais de um composto, os documentos obtidos são classificados, apesar do valor atribuído ser inferior àquele que se aplica se o termo aparecer por si próprio no documento (e não como parte de um composto. Este procedimento alarga o âmbito de procura que é útil nos casos em que apenas poucos documentos são encontrados com o composto completo.

Por exemplo, o termo de consulta `Versicherungspolice` (apólice de seguro) devolve documentos que contêm os termos compostos *Lebensversicherungspolice* (apólice de seguro de vida) e *Haftpflichtversicherungspolice* (apólice de seguro contra terceiros). Se a palavra *Police* (apólice) existir no dicionário de palavras hierárquicas, ao documento que contém o termo de consulta composto `Versicherungspolice` é atribuído um valor hierárquico.

Tem de listar termos com o respectivo valor hierárquico num ficheiro XML que tem de converter para um dicionário de palavras hierárquicas de forma a que possa ser adicionado ao sistema do Enterprise Search.

Pode seleccionar qual o dicionário de palavras hierárquicas a utilizar na consola de administração do Enterprise Search. Pode ser seleccionado um dicionário de palavras hierárquicas para cada colecção. Um dicionário de palavras hierárquicas pode ser partilhado por várias colecções.

Criar um ficheiro XML para palavras hierárquicas

Para aumentar ou diminuir a importância de determinados documentos resultantes, tem de especificar quais as palavras que podem influenciar a classificação de documentos num ficheiro XML.

Acerca desta tarefa

O ficheiro XML que lista as palavras hierárquicas tem de estar em conformidade com um esquema específico referido no ficheiro XML. Este é um exemplo de um ficheiro XML para palavras hierárquicas:

```
<?xml version="1.0" encoding="UTF-8"?>
<boostTerms xmlns="http://www.ibm.com/of/83/boostbuilder/xml">
    <!-- agrupa termos hierárquicos por valor hierárquico-->
   <boostTermList boost="5">
    <!-- cada termo pode especificar a expansão de sinónimos separadamente-->
     <term useVariants="true">OmniFind Edition</term>
     <term useVariants="false">Edition</term>
   <term>OmniFind</term>
  </boostTermList>
  <boostTermList boost="8">
       <term useVariants="true">WAS</term>
       <term>term9</term>
   </boostTermList>
</boostTerms>
```

Restrições

Pode agrupar termos que partilham o mesmo valor hierárquico num elemento `<boostTermList>`, mas um valor hierárquico pode ocorrer várias vezes, por exemplo, se pretender ordenar palavras hierárquicas alfabeticamente no ficheiro XML.

Uma palavra hierárquica pode incluir espaços em branco, mas não pode incluir caracteres de pontuação, tais como, uma vírgula (,) ou barra vertical (|), uma vez que estes caracteres podem interferir com a sintaxe de consulta do Enterprise Search.

Os termos hierárquicos podem ter variantes, tais como, acrónimos ou abreviaturas. Pode enumerar todas as variantes no dicionário de palavras hierárquicas; no entanto, se planear utilizar um dicionário de sinónimos bem como um dicionário de palavras hierárquicas e já tiver adicionado termos e as respectivas variantes ao dicionário de sinónimos, já não tem de adicionar estas variantes à lista de palavras hierárquicas. Em vez disso, pode simplesmente definir o atributo `useVariants` como verdadeiro (true) para a variante que adicionar ao dicionário de palavras hierárquicas. Todas as variantes deste termo listado no dicionário de sinónimos que ocorre em qualquer dos documentos obtidos têm influência sobre o valor de classificação para estes documentos.

Não é necessário enumerar as normalizações do termo, tais como, a remoção de acentos ou tremas (o Enterprise Search processa a normalização automaticamente). Por exemplo, se pretender incluir o termo météo como palavra hierárquica, não necessita de incluir também o termo METEO.

Procedimento

Para criar uma lista de palavras hierárquicas para o Enterprise Search:

1. Crie um ficheiro XML. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML à sua escolha. O esquema XSD para o ficheiro XML denomina-se boostTerms.xsd e encontra-se na instalação do Enterprise Search em *ES_INSTALL_ROOT*/packages/uima/configuration_xsd/.
2. Inclua os mapeamentos num elemento `<boostTerms xmlns="http://www.ibm.com/of/83/boostbuilder/xml">`. O espaço de nomes (especificado no atributo xmlns) necessita de estar exactamente conforme mostrado.
3. Adicione um elemento `<boostTermList>` para agrupar todos os termos que partilham o valor hierárquico especificado.

   Os valores hierárquicos pode ter um intervalo entre -10 e 10. Por exemplo, `<boostTermList boost="-5">` ou `<boostTermList boost="5">`.

   A importância dos documentos que contêm os termos especificados serão aumentados ou reduzidos de acordo com o valor hierárquico especificado.
4. Adicione um elemento `<term>` para cada termo que utiliza o valor hierárquico especificado.

   Se pretender incluir variantes de uma palavra hierárquica que estão listadas no dicionário de sinónimos, defina o atributo `useVariants` no elemento `<term>` como verdadeiro (true). A predefinição é falso (false). Se não forem encontradas variantes no dicionário de sinónimos, nenhuma mensagem de erro é produzida.
5. Repita os passos precedentes até especificar todos os termos que deverão ser utilizados como palavras hierárquicas quando os utilizadores procuram colecções do Enterprise Search.
6. Guarde e saia do ficheiro XML.

Após criar o ficheiro XML, tem de convertê-lo para um dicionário de palavras hierárquicas de forma a que possa ser adicionado ao sistema do Enterprise Search.

Criar um dicionário de palavras hierárquicas

Após criar ou actualizar uma lista de palavras hierárquicas num ficheiro XML, tem de converter o ficheiro XML num dicionário de palavras hierárquicas.

Acerca desta tarefa

Para criar um dicionário de palavras hierárquicas, utilize a ferramenta de linha de comandos denominada esboostworddictbuilder, que é fornecida com o WebSphere II OmniFind Edition. A ferramenta está no directório ES_INSTALL_ROOT/bin.

A entrada na ferramenta é o ficheiro XML que lista as palavras hierárquicas e a saída da ferramenta é um dicionário de palavras hierárquicas. O dicionário tem de ter o sufixo `.dic`. Por exemplo, `c:\meusdicionários\palavrashierárquicasproduto.dic`.

A localização predefinida para ambos os ficheiros é o directório onde o script é invocado. Se existir um dicionário com o mesmo nome, o script produz um erro.

O tamanho máximo de um `.dic` no Enterprise Search é de 8 MB.

Procedimento

Para criar um dicionário de palavras hierárquicas para o Enterprise Search:

1. No servidor de índices, inicie sessão como administrador do Enterprise Search. Este ID de utilizador foi especificado quando o WebSphere II OmniFind Edition foi instalado.
2. Introduza o seguinte comando, em que *ficheiro_XML* é o caminho totalmente qualificado para o ficheiro XML que contém a lista de palavras hierárquicas e *ficheiro_DIC* é o caminho totalmente qualificado para o dicionário de palavras hierárquicas. Se pretender utilizar também um dicionário de sinónimos, adicione o caminho totalmente qualificado para o dicionário de sinónimos após o nome do dicionário hierárquico. A atribuição de nome a um dicionário de sinónimos é opcional.

   UNIX: `esboostworddictbuilder.sh` *`fich_XML fich_DIC ficheiro_SYNDIC`*
   Windows: `esboostworddictbuilder.bat` *`fich_XML fich_DIC fich_SYNDIC`*

Após criar um dicionário de palavras hierárquicas, utilize a consola de administração do Enterprise Search para adicionar o dicionário ao sistema do Enterprise Search e associá-lo a uma ou mais colecções.

Apenas o ficheiro .dic gerado é carregado no sistema do Enterprise Search. Certifique-se de que o ficheiro XML de origem é mantido num ambiente de acesso controlado, com a estratégia de cópia de segurança apropriada na posição correcta. O utilizador necessita deste ficheiro XML para actualizar o dicionário de palavras hierárquicas.

Tarefas relacionadas

"Criar um dicionário de sinónimos" na página 64
Após criar ou actualizar uma lista de sinónimos num ficheiro XML, tem de converter o ficheiro XML num dicionário de sinónimos binário.

Análise de texto incluída no Enterprise Search

A análise de texto incluída no Enterprise Search inclui a detecção e segmentação do idioma do documento.

Quando um documento for processado, o Enterprise Search determina o idioma que o documento e interrompe a sequência de texto de entrada em unidades ou testemunhos distintos.

Durante a procura, o utilizador ou uma aplicação, tem de seleccionar manualmente o idioma da consulta. A cadeia de consulta é segmentada, analisada e procurada no índice.

A análise do documento e da cadeia de consulta pode ser dividida:

- Suporte não baseado em dicionários básicos. Inclui a segmentação n-grama e de espaços em branco. O suporte não baseado em dicionários básicos também contém segmentação de frases.
- Suporte linguístico baseado em dicionários. Inclui a segmentação e formação de lemas da palavra e frase.

  O processamento linguístico envolve a análise lexical, que é o processo de criação de representações alternativas do texto de entrada que associa todos os dados disponíveis do dicionário aos testemunhos que são reconhecidos no texto de entrada. A qualidade da procura é grandemente melhorada utilizando o processamento do idioma avançado.

  Conceitos relacionados

  "Identificação do idioma"
  Antes que possa ocorrer a segmentação da palavra e da frase, a normalização de caracteres ou a formação de lemas, o Enterprise Search tem de determinar o idioma do documento origem.

  "Suporte linguístico para segmentação não baseada em dicionários" na página 77
  Para documentos em idiomas que não são suportados pela tecnologia de análise lexical, o Enterprise Search fornece suporte básico sob a forma de segmentação de espaços em branco baseada em Unicode e segmentação n-grama.

Identificação do idioma

Antes que possa ocorrer a segmentação da palavra e da frase, a normalização de caracteres ou a formação de lemas, o Enterprise Search tem de determinar o idioma do documento origem.

O Enterprise Search consegue detectar automaticamente os seguintes idiomas:

Tabela 10. Idiomas suportados para identificação automática do idioma

Africanês	Árabe	Balinês
Basco	Catalão	Chinês (Tradicional e Simplificado)
Checo	Dinamarquês	Neerlandês
Inglês	Finlandês	Francês

Tabela 10. Idiomas suportados para identificação automática do idioma (continuação)

Alemão	Grego	Hebraico
Islandês	Irlandês (Gaélico)	Italiano
Japonês	Coreano	Malaio
Norueguês (Bokmål)	Polaco	Português
Romeno	Russo	Espanhol
Sueco	Tagalog	Tailandês
Turco	Vietnamês	

Os processos linguísticos no Enterprise Search detectam o idioma de um documento origem durante a indexação, não durante o processamento de consultas.

No Enterprise Search, pode especificar para detectar o idioma de um documento automaticamente ou seleccionar um idioma a utilizar.

Se seleccionar a detecção automática do idioma e o analisador não puder determinar o idioma de um documento, o analisador utiliza o idioma especificado quando cria a ferramenta de sequências de hiperligações na consola de administração do Enterprise Search.

Se não seleccionar a detecção automática do idioma, o idioma especificado é sempre utilizado. Especifique o idioma do documento editando as propriedades da ferramenta de sequências de hiperligações na consola de administração do Enterprise Search. O idioma predefinido é o inglês.

Os documentos para os quais não existem dicionários específicos de idioma são processados utilizando uma tecnologia independente do idioma, tal como, a segmentação de espaços em branco e segmentação n-grama.

A tecnologia de detecção de idioma do Enterprise Search adequa-se aos documentos monolingues. Se um documento for multilingue, é efectuada uma tentativa para determinar o idioma predominante que é utilizado no documento. No entanto, os resultados da análise nem sempre são satisfatórios.

O idioma de um documento pode ser utilizado para restringir os resultados da procura apenas aos documentos com determinado idioma. Por exemplo, se procurar documentos sobre Jacques Chirac numa colecção de documentos multilingue, pode limitar os resultados da procura para incluir apenas os documentos escritos em francês. Ao definir o idioma dos documentos resultantes é uma opção de procura avançada que pode seleccionar na consola de administração do Enterprise Search.

Conceitos relacionados

"Análise de texto incluída no Enterprise Search" na página 75
A análise de texto incluída no Enterprise Search inclui a detecção e segmentação do idioma do documento.

"Suporte linguístico para segmentação não baseada em dicionários" na página 77
Para documentos em idiomas que não são suportados pela tecnologia de análise lexical, o Enterprise Search fornece suporte básico sob a forma de segmentação de espaços em branco baseada em Unicode e segmentação n-grama.

Suporte linguístico para segmentação não baseada em dicionários

Para documentos em idiomas que não são suportados pela tecnologia de análise lexical, o Enterprise Search fornece suporte básico sob a forma de segmentação de espaços em branco baseada em Unicode e segmentação n-grama.

Segmentação de espaços em branco baseada em Unicode
Este método do processamento linguístico utiliza o espaço em branco entre palavras como delimitador de palavra.

Segmentação n-grama
Este método do processamento linguístico trata as sequências de sobreposição dos caracteres *n* como uma única palavra. Este método simples de segmentação é suficiente para muitas tarefas de obtenção.

Estes métodos são independentes de qualquer dicionário de idioma e não inclui a tecnologia de processamento linguístico sofisticado, tal como a redução da forma base.

A segmentação n-grama é utilizada para idiomas como o tailandês, que não utiliza espaços em branco como delimitadores. O mesmo método aplica-se ao hebraico e árabe. Apesar destes dois idiomas utilizarem delimitadores de espaços em branco, a segmentação n-grama devolve melhores resultados do que a forma básica de segmentação de espaços em branco baseada em Unicode.

Quando criar a colecção, pode também opcionalmente seleccionar para segmentar documentos em chinês e japonês utilizando a segmentação n-grama.

Para remover quaisquer caracteres de espaço em branco, por exemplo, novos caracteres de linha ou tabulação, durante a segmentação n-grama, tem de activar as definições de parâmetros no ficheiro denominado collection.properties em *ES_NODE_ROOT*/master_config/<IDColecção>.parserdriver antes de começar a analisar os documentos. Os parâmetros requeridos para remover os caracteres de parâmetro incluem:

- **removeCjNewLineChars**: Se estiver definido como *verdadeiro (true)*, este parâmetro remove qualquer sequência de caracteres de tabulação e de nova linha que ocorre entre os caracteres de idioma chinês ou japonês. A predefinição é `removeCjNewlineChars=false`.
- **removeCjNewLineCharsMode**: Se estiver definido como *tudo (all)*, este parâmetro remove os caracteres de espaço em branco independentemente do contexto do carácter. Por exemplo, os caracteres de espaço em branco são também removidos do texto inglês. Se pretender trabalhar com esta opção, tem de adicionar o parâmetro ao ficheiro de propriedades. Apenas `removeCjNewlineCharsMode=all` é válido, todos os outros valores são ignorados.

Conceitos relacionados

"Análise de texto incluída no Enterprise Search" na página 75
A análise de texto incluída no Enterprise Search inclui a detecção e segmentação do idioma do documento.

"Identificação do idioma" na página 75
Antes que possa ocorrer a segmentação da palavra e da frase, a normalização de caracteres ou a formação de lemas, o Enterprise Search tem de determinar o idioma do documento origem.

Segmentar caracteres numéricos como testemunho n-grama

Para segmentar caracteres numéricos para além dos caracteres de duplo byte como testemunhos n-grama, tem de activar uma definição de parâmetro no ficheiro descritor do segmentador n-grama e de espaço em branco.

Acerca desta tarefa

O processamento predefinido de caracteres numéricos no segmentador n-grama e de espaço em branco consiste em processar todos os caracteres numéricos como testemunhos segmentados de espaço em branco. Para segmentar caracteres numéricos como testemunhos n-grama, tem de alterar a definição do modo n-grama no ficheiro descritor do anotador. Não pode alterar esta definição utilizando a consola de administração do Enterprise Search.

Procedimento

A predefinição do modo n-grama é denominada normal e processa os caracteres numéricos e caracteres SBCS como caracteres segmentados por espaço em branco. Para activar o modo n-grama numérico:

1. Pare o analisador para a sua colecção.
2. Pare o tempo de execução para a sua colecção.
3. Abra o ficheiro descritor do anotador denominado jtok.xml no directório *ES_NODE_ROOT*/master_config/<CollectionID>.parserdriver/specifiers. O item CollectionID é o ID que foi especificado para a colecção (ou que foi atribuído pelo sistema) quando a colecção foi criada.
4. Altere a definição do parâmetro **NgramMode** de normal para numérico.
5. Reinicie o analisador para a sua colecção.
6. Reiniciar o tempo de execução.

Suporte linguístico para segmentação baseada em dicionários

Se o idioma de um documento for correctamente detectado e estiverem disponíveis os dicionários específicos de idioma, aplica-se o processamento linguístico apropriado.

A segmentação é o processo pelo qual o texto de entrada está dividido em unidades lexicais distintas. Este processo inclui algumas das seguintes actividades de processamento linguístico:

Segmentação de palavras
A segmentação de palavras é utilizada para idiomas que não utilizam espaços em branco (ou delimitadores) entre palavras, tais como, delimitadores e chinês.

Formação de lemas
A formação de lemas é uma forma de processamento linguístico que determina o lema para cada forma de palavra que ocorra no texto. O *lema* de uma palavra abrange a respectiva forma base mais as formas flexivas que partilham a mesma parte do discurso. Por exemplo, o lema para *ir* abrange *ir*, *vai*, *foi*, *ido* e *indo*. Os lemas para formas do singular e plural do grupo dos substantivos (tais como, *caracol* e *caracóis*). Os lemas para as formas de comparativo e superlativo do grupo dos adjectivos (tais como, *bom*, *melhor* e *óptimo*). Os lemas para casos diferentes do mesmo pronome do grupo de pronomes (tais como *I*, *eu*, *me* e *meu*).

A formação de lemas requer um dicionário para a indexação e procura.

O Enterprise Search indexa os lemas e as palavras flexionadas e forma lemas de todas as palavras flexionadas numa consulta. A formação de lemas melhora a qualidade da procura localizando documentos que contêm variantes de uma palavra flexionada na consulta. Por exemplo, os documentos que contêm a palavra *cães* são encontrados quando uma consulta inclui a palavra *cão*.

Divisão das contracções

A qualidade da procura é melhorada identificando as contracções e dividindo-as nas respectivas partes componentes. Por exemplo:

daquele é dividido em *de* + *aquele*
Horse's é dividido em *Horse* + *'s*

Identificação clítica

Os clíticos são uma forma especial de contracção e a qualidade da procura é melhorada pela determinação das respectivas partes componentes. Um *clítico* é um elemento que se comporta como um afixo e uma palavra. No entanto, os clíticos são difíceis de identificar uma vez que também fazem parte da formação da palavra. Ao contrário de outros fenómenos morfológicos (estrutura da palavra), os clíticos ocorrem numa estrutura sintáctica e a respectiva anexação às palavras não faz parte das regras de formação das palavras. Por exemplo:

reparti-lo-emos tem os componentes *repartir* + *lo* + *emos*
l'avenue tem os componentes *le* + *avenue*
dell'arte tem os componentes *dello* + *arte*.

Reconhecimento de caracteres não alfabéticos

Os processos linguísticos reconhecem caracteres não alfabéticos. Em função da lógica interna dependente do idioma, alguns caracteres não alfabéticos são devolvidos como unidades lexicais separadas de diferentes tipos e alguns são agrupados.

Por exemplo, os apóstrofos são considerados parte das palavras, no caso de clíticos, e são considerados pontos finais, no caso de abreviaturas desconhecidas. Os URLs, endereços de correio electrónico e datas são divididos em vários tokens.

Reconhecimentos de abreviaturas

Os processos linguísticos reconhecem as abreviaturas que existem no dicionário como uma unidade lexical. Se a abreviatura não se encontrar no dicionário é reconhecida como um item lexical, mas a abreviatura não terá quaisquer informações de dicionário associadas.

O reconhecimento correcto das abreviaturas é vital para o reconhecimento de frases. Por exemplo, o ponto no final de uma abreviatura não é necessariamente o final de uma frase.

Reconhecimento do marcador de fim de frase

Os processos linguísticos identificam correctamente os marcadores de fim de frase para a segmentação de frases.

O suporte linguístico baseado em dicionários está disponível para os seguintes idiomas:

Tabela 11. Idiomas suportados

Árabe	Italiano

Tabela 11. Idiomas suportados (continuação)

Chinês (Simplificado e Tradicional)	Japonês
Checo	Coreano
Dinamarquês	Norueguês (Bokmål)
Neerlandês	Polaco
Inglês	Português (Nacional e Brasileiro)
Finlandês	Russo
Francês (Nacional e Canadiano)	Espanhol
Alemão (Nacional e Suíço)	Sueco
Grego	

Conceitos relacionados

"Segmentação de palavras em japonês"
Se o documento de texto ou a cadeia de consulta for reconhecido como sendo japonês, o Enterprise Search efectua a segmentação de palavras relevante utilizando a tecnologia de análise morfológica que está optimizada para o idioma japonês.

"Variantes ortográficas em japonês"
O japonês utiliza muitas variantes ortográficas. As variantes de Katakana são as mais importantes uma vez que Katakana é frequentemente utilizado para escrever e pronunciar palavras estrangeiras. Muitas variantes de Katakana são geralmente utilizadas em japonês.

Segmentação de palavras em japonês

Se o documento de texto ou a cadeia de consulta for reconhecido como sendo japonês, o Enterprise Search efectua a segmentação de palavras relevante utilizando a tecnologia de análise morfológica que está optimizada para o idioma japonês.

Um exemplo desta optimização é a decomposição de palavras. O japonês utiliza um grande número de palavras compostas. Estas palavras são decompostas em testemunhos de tamanho optimizado para alcançar os melhores resultados da procura. Os palavras flexionadas e as preposições também são decompostas para melhorar o desempenho da procura.

Conceitos relacionados

"Suporte linguístico para segmentação baseada em dicionários" na página 78
Se o idioma de um documento for correctamente detectado e estiverem disponíveis os dicionários específicos de idioma, aplica-se o processamento linguístico apropriado.

"Variantes ortográficas em japonês"
O japonês utiliza muitas variantes ortográficas. As variantes de Katakana são as mais importantes uma vez que Katakana é frequentemente utilizado para escrever e pronunciar palavras estrangeiras. Muitas variantes de Katakana são geralmente utilizadas em japonês.

Variantes ortográficas em japonês

O japonês utiliza muitas variantes ortográficas. As variantes de Katakana são as mais importantes uma vez que Katakana é frequentemente utilizado para escrever e pronunciar palavras estrangeiras. Muitas variantes de Katakana são geralmente utilizadas em japonês.

O Enterprise Search utiliza um dicionário de variantes para mapear variantes de Katakana típicas com as respectivas formas base (semelhante a um lema) de forma a que todos os documentos, incluindo os que possuem variantes ortográficas da palavra Katakana na cadeia da consulta, sejam encontrados.

O Enterprise Search também suporta variantes Okurigana típicas, que são finalizações de palavras Kanji escritas em Hiragana.

Conceitos relacionados

"Suporte linguístico para segmentação baseada em dicionários" na página 78
Se o idioma de um documento for correctamente detectado e estiverem disponíveis os dicionários específicos de idioma, aplica-se o processamento linguístico apropriado.

"Segmentação de palavras em japonês" na página 80
Se o documento de texto ou a cadeia de consulta for reconhecido como sendo japonês, o Enterprise Search efectua a segmentação de palavras relevante utilizando a tecnologia de análise morfológica que está optimizada para o idioma japonês.

Remoção de palavras de paragem

No Enterprise Search, todas as palavras de paragem, por exemplo, as palavras comuns, tais como *um* e *o*, são removidas de várias consultas de palavras para aumentar o desempenho da procura.

O reconhecimento de palavras de paragem em japonês baseia-se em informações gramaticais, por exemplo, o Enterprise Search reconhece se a palavra é um substantivo ou um verbo. Para outros idiomas, o Enterprise Search utiliza listas especiais.

Não são removidas palavras de paragem durante o processamento das consultas se:

- Todas as palavras numa consulta forem palavras de paragem. Se todos os termos de consulta forem removidos durante o processamento das palavras de paragem, nesse caso, o conjunto de resultados está vazio. Para se certificar de que os resultados da procura são devolvidos, a remoção de palavras de paragem é desactivada quando todos os termos da consulta forem palavras de paragem. Por exemplo, se a palavra *carro* for uma palavra de paragem e procurar *carro*, os resultados da procura contêm documentos que correspondem à palavra *carro*. Se procurar *carro buick*, os resultados da procura contêm documentos que correspondem à palavra *buick*.
- A palavra numa consulta é precedida pelo sinal de adição (+).
- A palavra faz parte de uma correspondência exacta.
- A palavra está dentro de uma expressão, por exemplo, "Adoro o meu carro".

Conceitos relacionados

"Normalização de caracteres" na página 82
A normalização de caracteres é um processo que pode melhorar a recuperação. O melhoramento da recuperação pela normalização de caracteres significa que são obtidos mais documentos mesmo que os documentos não correspondam exactamente à consulta.

Normalização de caracteres

A normalização de caracteres é um processo que pode melhorar a recuperação. O melhoramento da recuperação pela normalização de caracteres significa que são obtidos mais documentos mesmo que os documentos não correspondam exactamente à consulta.

O Enterprise Search utiliza a normalização da compatibilidade Unicode que inclui a normalização de caracteres com metade da largura para caracteres de largura total asiáticos.

O Enterprise Search também remove os pontos centrados Katakana, que são utilizados como delimitadores de palavras compostas em japonês.

Outras formas de normalização de caracteres incluem:

Normalização de maiúsculas e minúsculas
Por exemplo, localizando documentos com *EUA* quando procura *eua*.

Expansão de tremas
Por exemplo, localizando documentos que contêm *schoen* quando procura *schön*.

Remoção de acentos
Por exemplo, localizando documentos que contêm *é* quando procura *e*.

Remoção de outros diacríticos
Por exemplo, localizando documentos que contêm *ç* quando procura *c*.

Expansão de ligaturas
Por exemplo, localizando documentos que contêm *Æ* quando procura *ae*.

Todas as normalizações funcionam em ambos os sentidos. Pode localizar documentos que contêm *eua* quando procura *EUA*, documentos que contêm palavras com *e* quando procura *é*, etc. Estas normalizações podem também ser combinadas. Por exemplo, pode localizar documentos que contêm *météo* quando procura *METEO*.

As normalizações são baseadas em propriedades de caracteres Unicode e não são dependentes do idioma. Por exemplo, o Enterprise Search suporta a remoção de diacríticos para hebraico e a expansão de ligaturas para árabe.

Conceitos relacionados

"Remoção de palavras de paragem" na página 81
No Enterprise Search, todas as palavras de paragem, por exemplo, as palavras comuns, tais como *um* e *o*, são removidas de várias consultas de palavras para aumentar o desempenho da procura.

Anotador de expressões globais

O anotador de expressões globais permite-lhe executar a análise de texto personalizada sem ter necessidade de implementar o seu próprio motor de análise de texto. Com base num conjunto de regras (expressões globais) que pode definir por si próprio, o anotador de expressões globais detecta as estruturas de informações em documentos de texto e cria anotações das informações detectadas na estrutura de análise comum.

O anotador de expressões globais detecta entidades ou unidades de informação em documentos de texto, por exemplo, números de telefone, códigos de produtos, números do edifício e das salas ou endereços, com base em expressões globais. Se uma das expressões globais corresponder a partes do texto do documento, o anotador de expressões globais cria as anotações correspondentes que incluem a parte das informações correspondente. Estas anotações são armazenadas na estrutura de análise comum e podem posteriormente ser procuradas mapeando estes resultados de análise para o índice do Enterprise Search, utilizando um ficheiro de mapeamento da estrutura de análise comum para o índice. Em alternativa, um ficheiro de mapeamento da estrutura de análise comum para a base de dados pode ser criado para armazenar as anotações numa base de dados que suporte JDBC.

O conjunto de regras (expressões globais) que o utilizador define está armazenado num ficheiro de configuração XML (também referido como ficheiro de conjunto de regras). O anotador de expressões globais contém a lógica de análise que processa essas expressões globais. Suporta a sintaxe de expressões globais em Java 1.4.

A descrição do sistema tipo do anotador de expressões globais tem de definir os tipos e funcionalidades de anotações que são utilizadas e criadas pelo anotador de expressões globais. Em função da complexidade da área de aplicação do anotador de expressões globais (por exemplo, se forem requeridos mais tipos do que os definidos no anotador de expressões globais fornecido), as funções de entrada e saída adicionais têm de ser definidas no descritor do anotador de expressões globais. Os tipos utilizados no descritor têm de corresponder aos tipos na descrição do sistema tipo do anotador.

O anotador de expressões globais está incluído no Enterprise Search como um ficheiro PEAR (Processing Engine ARchive) passível de ser implementado que é configurado com regras de amostra para detectar números de telefone, URLs e endereços de correio electrónico.

Conceitos relacionados

"Ficheiro do conjunto de regras" na página 86
No anotador de expressões globais, o ficheiro do conjunto de regras XML define as regras, na forma de expressões globais, que são utilizadas para analisar o documento de texto.

Tarefas relacionadas

"Definir regras de expressão global" na página 87
O conjunto de regras define as expressões globais que são comparadas com o texto no documento e as acções que o anotador de expressões globais tem de executar se existir correspondência de padrões.

Referências relacionadas

"Descritor do anotador" na página 92
O descritor XML do anotador de expressões globais contém informações descritivas sobre o anotador de expressões globais que é necessário para executar o anotador.

"Registar" na página 95
Todas as mensagens de registo do anotador de expressões globais são escritas no ficheiro de registo da colecção actual.

Procura semântica fácil utilizando o anotador de expressões globais

O Enterprise Search inclui o motor de análise de expressões globais pré-configurado com um conjunto de regras que permite detectar números de telefone, URLs e endereços de correio electrónico nos documentos de texto.

Pode utilizar esta configuração de amostra do motor de análise de expressões globais para permitir ao Enterprise Search localizar números de telefone reais em documentos sem procurar a palavra-chave *número de telefone* nos documentos. Para consultar as construções detectadas pelo anotador de expressões globais, também é fornecido um ficheiro de mapeamento da estrutura de análise comum para o índice de amostra. Além disso, um simples método é demonstrado através do qual pode emitir consultas semânticas poderosas através de simples palavras-chave. Este método utiliza o suporte de sinónimos do Enterprise Search para expandir automaticamente simples consultas de palavras-chave nas consultas semânticas. É fornecido um dicionário de sinónimos de amostra que ilustra este mecanismo. Pode encontrar todos os ficheiros de que necessita para utilizar o anotador de expressões globais com a configuração de amostra em ES_INSTALL_ROOT/packages/uima/regex.

Para muitos cenários de aplicação, pode ser suficiente modificar apenas ligeiramente as regras de expressão global que são fornecidas com a configuração de amostra de forma a ajustar o anotador de expressões globais para corresponder às suas necessidades.

No entanto, para personalizar totalmente o anotador, recomenda-se a utilização do UIMA SDK. Com esta finalidade, o anotador de expressões globais é também incluído no pacote anotador base do Enterprise Search localizado em ES_INSTALL_ROOT/packages/uima/.

Tarefas relacionadas

"Activar a procura semântica fácil utilizando o anotador de expressões globais" na página 85
Para activar a procura semântica fácil utilizando sinónimos, tem de adicionar o anotador de expressões globais, o ficheiro de mapeamento da estrutura de análise comum para o índice e o dicionário de sinónimos de amostra ao sistema Enterprise Search e associar estes recursos à colecção.

"Personalizar o anotador de expressões globais" na página 91
Pode personalizar a configuração de amostra do anotador de expressões globais para detectar novas entidades (por exemplo, números de série dos produtos) ou adaptar as regras de expressão global às entidades existentes (por exemplo, para detectar números de telefone específicos da empresa) efectuando pequenas alterações ao conjunto de regras de amostra e ficheiros do sistema tipo.

"Visualizar o anotador base e os resultados da análise de texto personalizada" na página 13
Para visualizar os resultados da análise produzidos após a análise e por quaisquer anotadores no Enterprise Search, tem de actualizar as propriedades

da colecção de documentos para produzir uma versão XML legível dos resultados da análise que são armazenados na estrutura de análise comum.

Activar a procura semântica fácil utilizando o anotador de expressões globais

Para activar a procura semântica fácil utilizando sinónimos, tem de adicionar o anotador de expressões globais, o ficheiro de mapeamento da estrutura de análise comum para o índice e o dicionário de sinónimos de amostra ao sistema Enterprise Search e associar estes recursos à colecção.

Consequentemente, o anotador de expressões globais processará os documentos durante a fase de análise, o indexador adicionará os resultados da análise personalizada ao índice e o serviço de procura pode utilizar o dicionário de sinónimos semânticos fornecido para procurar os resultados de análise personalizada através de simples palavras-passe que são automaticamente expandidas em consultas semânticas.

Procedimento

Para activar a procura semântica fácil:

1. Adicione o motor de análise de texto personalizada de expressões globais denominado of_regex.pear em *ES_INSTALL_ROOT*/packages/uima/regex ao sistema Enterprise Search utilizando a consola de administração do Enterprise Search.
2. Associe o motor de análise de texto de expressões globais à sua colecção.
3. Adicione o ficheiro de mapeamento da estrutura de análise comum para o índice denominado of_sample_regex_cas2index.xml no directório *ES_INSTALL_ROOT*/packages/uima/regex. Deste modo, mapeia os resultados da análise personalizada (anotações) que o anotador de expressões globais produz nas expansões passíveis de serem procuradas no índice do Enterprise Search. Em seguida, pode utilizar o fragmento XML ou as consultas XPath para procurar estas expansões.
4. Pesquise, analise e indexe a sua colecção. Nesta altura, após a indexação ter terminado, pode introduzir uma consulta de procura de XML utilizando uma expressão de fragmento XML, por exemplo, `@xmlf2::'<#phonenumber>'`, utilizando a aplicação de procura. No entanto, a finalidade de activar a procura semântica por sinónimos consiste em permitir utilizar consultas como `Número de telefone da Bárbara` e fazer com que o sistema converta a consulta para Bárbara @xmlf2::'<#phonenumber>'.
5. Adicione o dicionário de sinónimos binário de amostra fornecido denominado of_sample_synonym_dic.dic no directório *ES_INSTALL_ROOT*/packages/uima/regex ao sistema Enterprise Search utilizando a consola de administração. Pode efectuar alterações no dicionário de amostra XML da origem ou utilizá-lo como base para criar o seu próprio dicionário e, em seguida, convertê-lo num novo ficheiro de dicionário utilizando a ferramenta essyndictbuilder. O dicionário de sinónimos de amostra XML denomina-se of_sample_synonym_dic.xml, também em *ES_INSTALL_ROOT*/packages/uima/regex.
6. Associe o dicionário de sinónimos à sua colecção e inicie (ou reinicie) o serviço de procura para a colecção.
7. Na aplicação da procura, seleccione a opção para procurar automaticamente sinónimos utilizando a expansão semântica. Após activar esta opção, a aplicação da procura reescreve as consultas de palavras-chave básicas em

consultas de fragmentos XML e inclui expressões que localizam as expansões passíveis de serem procuradas que identificam os números de telefone, endereços de correio electrónico e URLs.

8. Na aplicação da procura, introduza uma consulta a pedir um número de telefone, por exemplo, `número de telefone da Bárbara`. A consulta procura documentos que contêm três palavras-chave *Bárbara*, *telefone* e *número*, bem como documentos que contenham a palavra-chave *Bárbara* e expansões de números e caracteres nos documentos que correspondam a expressões globais definidas para um número de telefone. As palavras-chave e os números de telefone que são encontrados são realçados nos resultados da procura.

Pode ver quais as palavras-chave que se traduzem em consultas semânticas no dicionário de sinónimos de amostra fornecidos.

```
<?xml version="1.0" encoding="UTF-8"?>
<synonymgroups xmlns="http://www.ibm.com/of/822/synonym/xml">
  <synonymgroup>
<synonym>número de telefone</synonym>
<synonym>número de tlf</synonym>
<synonym>nº de telefone</synonym>
<synonym>nº de tlf</synonym>
<synonym>@xmlf2::'&lt;#phonenumber/&gt;'</synonym>
  </synonymgroup>
  <synonymgroup>
<synonym>número de fax</synonym>
<synonym>número fax</synonym>
<synonym>nº de fax</synonym>
<synonym>nº fax</synonym>
<synonym>@xmlf2::'&lt;#phonenumber/&gt;'</synonym>
  </synonymgroup>
  <synonymgroup>
<synonym>endereço de correio electrónico</synonym>
<synonym>endereço de email</synonym>
<synonym>@xmlf2::'&lt;#email/&gt;'</synonym>
  </synonymgroup>
  <synonymgroup>
<synonym>URL</synonym>
<synonym>unified resource locator</synonym>
<synonym>endereço Web</synonym>
<synonym>@xmlf2::'&lt;#url/&gt;'</synonym>
  </synonymgroup>
</synonymgroups>
```

Conceitos relacionados

"Procura semântica fácil utilizando o anotador de expressões globais" na página 84

O Enterprise Search inclui o motor de análise de expressões globais pré-configurado com um conjunto de regras que permite detectar números de telefone, URLs e endereços de correio electrónico nos documentos de texto.

Ficheiro do conjunto de regras

No anotador de expressões globais, o ficheiro do conjunto de regras XML define as regras, na forma de expressões globais, que são utilizadas para analisar o documento de texto.

As regras especificam, por ordem sequencial, onde no documento de texto o anotador tem de procurar e, em seguida, qual a acção a executar se uma correspondência for encontrada.

Quando o anotador de expressões globais for chamado, o ficheiro do conjunto de regras XML que contém os padrões de expressão global é compilado e comparado com as partes do texto do documento. Se for encontrada uma correspondência parcial ou total, a anotação que está associada à regra específica é criada e armazenada na estrutura de análise comum.

Os tipos utilizados nas regras têm de estar definidos na descrição do sistema tipo do anotador de expressões globais.

O anotador de expressões globais processa uma regra de cada vez, começando pela primeira regra no ficheiro do conjunto de regras XML. Para cada regra, a expressão global compilada correspondente é comparada com as anotações criadas num passo anterior, por exemplo, as anotações criadas pelos anotadores que processaram o documento antes do anotador de expressões globais. As anotações que correspondem às regras têm de ser do mesmo tipo que os tipos de funções de entrada especificados no descritor do anotador de expressões globais.

Se for encontrada uma correspondência, o tipo de anotação criado na regra que é accionada tem também de ser especificado como um tipo de função de saída válido no descritor do anotador de expressões globais. As novas anotações que são criadas por uma regra anterior podem ser utilizadas como anotações de entrada para as regras que são accionadas posteriormente no conjunto de regras XML.

Conceitos relacionados

"Anotador de expressões globais" na página 83
O anotador de expressões globais permite-lhe executar a análise de texto personalizada sem ter necessidade de implementar o seu próprio motor de análise de texto. Com base num conjunto de regras (expressões globais) que pode definir por si próprio, o anotador de expressões globais detecta as estruturas de informações em documentos de texto e cria anotações das informações detectadas na estrutura de análise comum.

Tarefas relacionadas

"Definir regras de expressão global"
O conjunto de regras define as expressões globais que são comparadas com o texto no documento e as acções que o anotador de expressões globais tem de executar se existir correspondência de padrões.

Referências relacionadas

"Descritor do anotador" na página 92
O descritor XML do anotador de expressões globais contém informações descritivas sobre o anotador de expressões globais que é necessário para executar o anotador.

"Registar" na página 95
Todas as mensagens de registo do anotador de expressões globais são escritas no ficheiro de registo da colecção actual.

Definir regras de expressão global

O conjunto de regras define as expressões globais que são comparadas com o texto no documento e as acções que o anotador de expressões globais tem de executar se existir correspondência de padrões.

Acerca desta tarefa

O ficheiro do conjunto de regras XML tem de seguir a sintaxe de regras delineada no seguinte exemplo. Este é um ficheiro do conjunto de regras para o anotador de expressões globais de amostra que reconhece números de telefone, URLs e endereços de correio electrónico.

O elemento de nível superior é um elemento <ruleSet> que consiste em um ou mais elementos <rule>. Cada elemento <rule> por sua vez define uma expressão global de Java consistindo de um atributo regEx bem como dos atributos matchStrategy e matchType. A acção é definida no elemento <createAnnotation> que especifica o ID de anotação e o tipo de anotação.

```
<?xml version="1.0" encoding="UTF-8"?>
<ruleSet xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="ruleSet.xsd">
<!-- Phone Number -->
<!-- Esta regra corresponde a diferentes formas de escrever números de telefone,
por exemplo, 01234-12345, 01234 / 122-32, (001234)12345,
+49 (0) 123412345, (123) 123 1234,
1-800-IBM-4YOU -->
 <rule regEx="(?x)(\s|\b)(
0{1,2}[1-9]{1}[0-9]{1,5}\x20?[-/\\]\x20?[1-9]{1}([0-9]{1,8}-?)
  {1,3}[0-9]{1,}
|\(0[1-9]{1}[0-9]{1,3}\)\x20?[1-9]{1}[0-9]{2,8}
|\(00[1-9]{1}[0-9]{1,8}\)\x20?[1-9]{1}[0-9]{2,10}
|\((0\x20?[1-9]{1}[0-9]{1,3}|00\x20?[1-9]{1}[0-9]{1,8})\)\x20?[1-9]
{1}[0-9] {1,3}(\x20[0-9]{2,4}){1,5}
|(0\x20?[1-9]{1}[0-9]{1,3}|00\x20?[1-9]{1}[0-9]{1,8})\x20?[/\\]\x20?
[1-9]{1} [0-9]{1,3}(\x20[0-9]{2,4}){1,5}
|\(?\+[1-9]{1}[0-9]{0,3}\)?([-\x20]|\x20?\(0\))[-\x20]?[1-9]
   {1}[0-9]{1,10}
|\(?\+[1-9]{1}[0-9]{0,3}\)?([-\x20]|\x20?\(0\))[-\x20]?[1-9]
{1}[0-9]{1,3}[-\x20]([0-9]{2,5}[-\x20]?){1,4}
|(1-)?[0-9]{3}-[0-9]{3}-[0-9]{4}
|\([1-9]{1}[0-9]{2}\)\x20[0-9]{3}[-\x20][0-9]{4}
|1-(800|888|877|866)-([A-Z0-9]{7}|[A-Z0-9]{3}-[A-Z0-9]
   {4}|[A-Z0-9]{4}-[A-Z0-9]{3})
)(?!(\d|\x20\d|-\d))(\s|\b)"
  matchStrategy="matchAll" matchType="uima.tcas.DocumentAnnotation">
  <createAnnotation id="phonenumber" type="com.ibm.es.uima.PhoneNumber">
   <begin group="0"/>
   <end group="0"/>
  </createAnnotation>
 </rule>
<!-- potential Phone Number -->
<!-- Esta regra corresponde a números que se assemelham a números de
telefone mas podem também ser outra coisa. Por exemplo, 0123 1234 123,
+123456789, 123 123 1234 -->
 <rule regEx="(?x)(\s|\b)(
  0[1-9]{1}[0-9]{1,3}\x20[1-9]{1}[0-9]*\x20?([0-9]{2,}\x20?)+
  |00\x20?[1-9]{1}[0-9]{0,3}\x20[1-9]{1}[0-9]{1,3}\x20?[1-9]
{1}([0-9]{2,}\x20?)+
  |\+[1-9]{1}[0-9]{0,3}[1-9]{1}[0-9]{6,}
  |[1-9]{1}[0-9]{2}\x20[0-9]{3}\x20[0-9]{4}
  )(?!(\d|\x20\d|-\d))(\s|\b)"
  matchStrategy="matchAll" matchType="uima.tcas.DocumentAnnotation">
  <createAnnotation id="potential_phonenumber"
     type="com.ibm.es.uima.PotentialPhoneNumber">
   <begin group="0"/>
   <end group="0"/>
  </createAnnotation>
 </rule>
<!-- URL Annotation -->
<!-- Esta regra corresponde a URLs, por exemplo, http://www.ibm.com -->
 <rule regEx="(?x)(\s|\b)(
    http://[\w\-]+([\.][\w\-]+)+([/][\w\~\(\)\-\?=%\u0026\#]*)*
    |www.[\w\-]+([\.][\w\-]+)+([/][\w\~\(\)\-\?=%\u0026\#]*)*
```

```
   )(\s|\b)"
  matchStrategy="matchAll" matchType="uima.tcas.DocumentAnnotation">
  <createAnnotation id="url" type="com.ibm.es.uima.URL">
   <begin group="0"/>
   <end group="0"/>
  </createAnnotation>
 </rule>
<!-- Email Annotation -->
<!-- Esta regra corresponde a endereços de correio electrónico, por
exemplo, nome@domínio.com -->
 <rule regEx="(?x)(\s|\b)(
    [a-zA-Z0-9][\w\.-]*[a-zA-Z0-9]@[a-zA-Z0-9]([\.-]?\w)*\.[a-zA-Z]
    {2,3})(\s|\b)"
  matchStrategy="matchAll" matchType="uima.tcas.DocumentAnnotation">
  <createAnnotation id="email" type="com.ibm.es.uima.Email">
   <begin group="0"/>
   <end group="0"/>
  </createAnnotation>
 </rule>
</ruleSet>
```

Procedimento

Para criar o conjunto de regras XML para o anotador de expressões globais que define as expressões globais personalizadas:

1. Crie um ficheiro XML. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML à sua escolha. O esquema XSD para o ficheiro do conjunto de regras XML é denominado ruleSet.xsd, que pode localizar na instalação do Enterprise Search no directório *ES_INSTALL_ROOT*/packages/uima/regex/.
2. Inclua os mapeamentos num elemento `<ruleSet xmlns="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="ruleSet.xsd">`. O espaço de nomes é especificado no atributo xmlns e tem de estar exactamente conforme mostrado.
3. Adicione um elemento `<rule>` que contenha um atributo `regEx` com o padrão de expressão global, um atributo `matchStrategy` e um atributo `matchType`.

   O anotador suporta totalmente a sintaxe de expressão global do Java 1.4. Para uma introdução às expressões globais e para visualizar a sintaxe completa, consulte a documentação de Java em http://java.sun.com/j2se/1.4.2/docs/api/java/util/regex/Pattern.html.

   `matchStrategy` especifica como procurar, por exemplo, se todas as correspondências têm de ser encontradas no documento ou se a correspondência de texto tem de ser exacta. Estão disponíveis três diferentes estratégias de correspondência:

   - `matchFirst` pára na primeira sequência de texto que corresponder ao padrão global
   - `matchAll` localiza todas as sequências num documento que corresponde ao padrão global
   - `matchComplete` apenas as sequências de texto que são uma correspondência exacta. Por exemplo, se tivermos um padrão "foo", apenas o termo "foo" seria correspondido, "barrafoo" não resultaria numa correspondência.

   `matchType` determina o tipo de anotação que é comparado à regra. Deste modo, pode restringir a expressão global a corresponder, por exemplo, no âmbito de uma anotação do testemunho existente. Assim, evita a correspondência de demasiado conteúdo no âmbito de uma regra. Os tipos possíveis são os tipos de anotação de entrada permitidos para o anotador (definidos no descritor do anotador), tais como, uima.tt.DocumentAnnotation,

uima.tt.ParagraphAnnotation e tipos definidos pelo utilizador, tais como, foo.bar.MyAnnotation. Por vezes, o tipo de saída de uma regra é utilizado como tipo de entrada de uma regra subsequente. `matchType` permite-lhe restringir o âmbito da procura de determinadas regras.

4. Adicione um elemento `<createAnnotation>` que define a acção que o anotador de expressões globais deverá executar se for encontrada uma correspondência.

   Cada elemento `createAnnotation` tem dois atributos:

   - `id` identifica de forma exclusiva a anotação e é utilizado para fazer referência à anotação
   - `type` especifica o tipo de anotação que é criada

5. Adicione os seguintes elementos componentes que definem a posição de correspondência para o elemento `<createAnnotation>`:

   - Obrigatório: `<begin>` especifica onde a correspondência começa. Este elemento tem dois atributos:
     - Obrigatório: `group` identifica o grupo de captura Java. Pode ser atribuído um valor entre 0 (correspondência de sequência de texto completa) e 9 (vários grupos de captura)
     - Opcional: `location` indica uma posição dentro do grupo de correspondência (respeitante ao posicionamento dos parênteses), quer sejam `start` (abrir parêntese curvo) ou `end` (fechar parêntese curvo).
   - Obrigatório: `<end>` especifica onde a correspondência termina. Este elemento tem dois atributos:
     - Obrigatório: `group` identifica o grupo de captura. Pode ser atribuído um valor entre 0 (correspondência de sequência de texto completa) e 9 (grupos de correspondência subsequentes e cada vez mais pequenos)
     - Opcional: `location` indica uma posição dentro do grupo de correspondência (respeitante ao posicionamento dos parênteses), quer sejam `start` (abrir parêntese curvo) ou `end` (fechar parêntese curvo).
   - Opcional: `<setFeature>` cria uma funcionalidade e atribui esta última à anotação. Este elemento tem dois atributos:
     - `name` é o nome da funcionalidade conforme definida na descrição do sistema tipo
     - `type` especifica o tipo de valor da funcionalidade, Cadeia (String), Número Inteiro (Integer), Float e Referência (Reference). O tipo tem de ser o mesmo que o tipo de intervalo definido para a funcionalidade na descrição do sistema tipo do anotador.

       As funcionalidades do tipo `Reference` são utilizadas para criar uma ligação entre duas anotações para modelar uma relação semântica. O conteúdo do elemento `<setFeature>` tem de estar definido para o `id` do elemento `<createAnnotation>` ao qual pretende estabelecer ligação.

Conceitos relacionados

"Ficheiro do conjunto de regras" na página 86
No anotador de expressões globais, o ficheiro do conjunto de regras XML define as regras, na forma de expressões globais, que são utilizadas para analisar o documento de texto.

Personalizar o anotador de expressões globais

Pode personalizar a configuração de amostra do anotador de expressões globais para detectar novas entidades (por exemplo, números de série dos produtos) ou adaptar as regras de expressão global às entidades existentes (por exemplo, para detectar números de telefone específicos da empresa) efectuando pequenas alterações ao conjunto de regras de amostra e ficheiros do sistema tipo.

O ficheiro do conjunto de regras modificado e a descrição do sistema tipo têm de ser adicionados ao ficheiro de arquivo do motor de processamento (ficheiro PEAR) de expressão global. Após actualizar o ficheiro PEAR, pode adicionar novamente o motor de análise de texto de expressão global ao sistema Enterprise Search.

Para uma personalização mais elaborada do anotador de expressões globais, recomenda-se vivamente que utilize as ferramentas do UIMA SDK. Estas ferramentas ajudam a criar ou actualizar a descrição do sistema tipo e os ficheiros descritores, para combinar possivelmente o anotador com outros para formar um motor de análise agregada e para criar um novo arquivo de motor de processamento (ficheiro PEAR) que inclui todos os recursos necessários para utilizar o anotador no Enterprise Search. Para obter informações sobre as ferramentas que estão disponíveis de forma a ajudá-lo nestas tarefas, consulte a documentação do UIMA SDK.

Procedimento

Para adaptar o anotador de expressões globais adicionando novas regras e entidades ou para alterar as regras existentes, pode actualizar o ficheiro PEAR do anotador de expressões globais de amostra fornecido da seguinte forma:

1. Crie um novo directório denominado xml no sistema.
2. Copie o ficheiro de regras de amostra of_sample_regex_rules.xml no directório *ES_INSTALL_ROOT*/packages/uima/regex/ para o directório xml e modifique o ficheiro para incluir as regras de correspondência padrão personalizadas. Para evitar erros de sintaxe XML, utilize um editor de XML ou ferramenta de criação de XML à sua escolha.
3. Copie o ficheiro de descrição do sistema tipo correspondente of_sample_typesystem.xml do directório ES_INSTALL_ROOT/packages/uima/regex/ para o directório xml e modifique o ficheiro para incluir as definições para os tipos que as novas regras requerem.
4. Se adicionar apenas algumas regras novas ou alterar regras existentes, não é necessário que altere o descritor do anotador. Se planear efectuar outras alterações ou se utilizar passos de análise personalizada adicionais, verifique se o descritor do anotador tem ser modificado.
5. Utilize um utilitário de arquivo à sua escolha para actualizar uma cópia do ficheiro PEAR do anotador de expressões globais para incluir os seus dois ficheiros actualizados. Por exemplo, copie o ficheiro of_regex.pear a partir de ES_INSTALL_ROOT/packages/uima/regex/ para o directório ascendente do directório xml que criou. Em seguida, utilize a ferramenta da linha de comandos jar do Java (por exemplo, parte do IBM Java SDK) para emitir os seguintes comandos a partir do directório ascendente:

   ```
   "jar -uf of_regex.pear -C xml/ of_sample_regex_rules.xml"
   "jar -uf of_regex.pear -C xml/ of_sample_regex_typesystem.xml"
   ```

6. Utilize a consola de administração do Enterprise Search para adicionar o anotador de expressões globais como um motor de análise de texto personalizada ao sistema do Enterprise Search e associe-o a uma colecção de documentos de teste.
7. Verifique os resultados da análise produzidos pelo anotador de expressões globais actualizando as propriedades da colecção de documentos para produzir uma saída XML legível dos resultados da análise que estão armazenados na estrutura de análise comum utilizando a funcionalidade de cópia de memória (dump) XCAS.
8. Processe os documentos de teste e utilize o Visualizador de Anotação XCAS (XCAS Annotation Viewer) para visualizar o conteúdo dos ficheiros XML.
9. Se estiver satisfeito com as anotações que são criadas pelo anotador com base nas expressões globais personalizadas, edite as propriedades da colecção de documentos novamente para desactivar o analisador de produzir uma saída de XML legível dos resultados da análise. Se forem necessárias mais modificações no ficheiro do conjunto de regras, tem de repetir os passo que permitem actualizar o ficheiro PEAR.
10. Crie o ficheiro de mapeamento da estrutura de análise comum para o índice para indexar os resultados da análise ou o ficheiro de mapeamento da estrutura de análise comum para a base de dados se pretender adicionar os resultados a uma base de dados. Pode utilizar o ficheiro de mapeamento da estrutura de análise comum para o índice fornecido como ponto de partida para criar o ficheiro de mapeamento da estrutura de análise comum para o índice.
11. Utilize a consola de administração do Enterprise Search para adicionar os ficheiros de mapeamento e associá-los a toda a colecção de documentos.
12. Procure as anotações utilizando o fragmento XML ou as consultas XPath ou, em alternativa, utilizando a expansão semântica durante a procura de sinónimos.

Conceitos relacionados

"Procura semântica fácil utilizando o anotador de expressões globais" na página 84
O Enterprise Search inclui o motor de análise de expressões globais pré-configurado com um conjunto de regras que permite detectar números de telefone, URLs e endereços de correio electrónico nos documentos de texto.

Tarefas relacionadas

"Visualizar o anotador base e os resultados da análise de texto personalizada" na página 13
Para visualizar os resultados da análise produzidos após a análise e por quaisquer anotadores no Enterprise Search, tem de actualizar as propriedades da colecção de documentos para produzir uma versão XML legível dos resultados da análise que são armazenados na estrutura de análise comum.

Descritor do anotador

O descritor XML do anotador de expressões globais contém informações descritivas sobre o anotador de expressões globais que é necessário para executar o anotador.

Se só estiver a utilizar o anotador de expressões globais e mais nenhum passo de análise personalizada, apenas é requerido que altere o descritor se:

- Pretender alterar o nome do ficheiro do ficheiro de conjunto de regras (no elemento `<externalResourceDependencies>`).

- Pretender utilizar mais do que um ficheiro de conjunto de regras .
- Pretender alterar o nome do ficheiro descrição do sistema tipo.

Se estiver a utilizar passos de análise personalizada adicionais, é requerido que altere o descritor se:

- Pretender que a análise personalizada utilize as anotação criados pelo anotador de expressões globais. Neste caso, tem de actualizar as funções de saída no descritor do anotador.
- Tiver definido regras de expressão global que têm de corresponder aos tipos de anotações criadas nos passos de análise personalizada anteriores. Neste caso, tem de actualizar as funções de entrada no descritor do anotador.

Utilize as ferramentas do UIMA SDK para criar ou actualizar o descritor do anotador e recriar o arquivo de motor de processamento (ficheiro .pear) que inclui todos os recursos necessários para utilizar o anotador no Enterprise Search. Consulte a documentação de UIMA para obter informações sobre as ferramentas que estão disponíveis para ajudar o utilizador nestas tarefas.

Parâmetros de configuração

O anotador de expressões globais apenas tem um parâmetro de configuração denominado `String2NumberImpl` que tem de ser definido com o nome da classe que implementa a interface `com.ibm.uima.an_regex.String2Number`. O anotador de expressões globais tem de ser fornecido com uma implementação desta classe, caso contrário ocorrerá uma excepção. Se pretender personalizar o anotador de expressões globais para corresponder às suas necessidades, pode fornecer a sua própria implementação da interface `String2Number` passando o nome da classe no ficheiro descritor XML.

A interface `String2Number` declara dois métodos, `toInt(String)` e `toFloat(String)`, que transformam uma representação de cadeia de um número inteiro ou valor float num valor correspondente respectivo. Estes dois métodos são utilizados para transformar um número que contém caracteres separadores num valor de Número Inteiro (Integer) ou Float válido de Java.

A implementação predefinida de `com.ibm.uima.an_regex.String2Number_impl` considera um ponto final (.) como um separador decimal e uma vírgula (,) como um separador de milhares. Por exemplo, se 1.999,00 for encontrado num documento de texto, `toInt` converte-o em 1999. `toFloat` devolve 1999,00.

Amostra

A secção do parâmetro de configuração do descritor é da seguinte forma:

```
<configurationParameters>
 <configurationParameter>
<name>String2NumberImpl</name>
<description>Implementação da interface
com.ibm.uima.an_regex.String2Number</description>
<type>String</type>
<multiValued>false</multiValued>
<mandatory>true</mandatory>
 </configurationParameter>

 <configurationParameterSettings>
  <nameValuePair>
<name>String2NumberImpl</name>
   <value>
```

```
<string>com.ibm.uima.an_regex.impl.String2Number_impl</string>
   </value>
  </nameValuePair>
 </configurationParameterSettings>
</configurationParameters>
```

Funções

As funções de entrada e saída do anotador de expressões globais e os idiomas suportados por estas são definidas na secção de funções do descritor do anotador.

As funções de entrada (tipos de entrada) no ficheiro descritor têm de cumprir com os tipos de correspondência utilizados no ficheiro de conjunto de regras. Se as regras apenas utilizarem o tipo `uima.tt.DocumentAnnotation`, não tem de declarar quaisquer funções de entrada uma vez que este tipo está sempre definido. Todos os outros tipos têm de ser definidos.

Os tipos de anotação criados pelo anotador de expressões globais são especificados na secção de funções da saída. Estes tipos têm de corresponder aos tipos de saída declarados no ficheiro de conjunto de regras.

Uma vez que o anotador de expressões globais é independente do idioma, especifique `x-unspecified`, que representa qualquer idioma.

Descrição do sistema tipo

A secção da descrição do sistema tipo no descritor XML do anotador de expressões globais define o sistema tipo utilizado pelo anotador. Os tipos utilizados no ficheiro XML do conjunto de regras e mencionados nas secções de funções de entrada e saída no descritor do anotador têm de corresponder aos tipos definidos na descrição do sistema tipo.

Amostra

A secção da descrição do sistema tipo do descritor importa o ficheiro XML do descritor do sistema tipo:

```
<typeSystemDescription>
 <imports>
  <import location="./xml/of_sample_regex_typesystem.xml"/>
 </imports>
</typeSystemDescription>
```

Recursos externos

A secção de recursos externos do descritor contém os ficheiros e classes requeridos pelo anotador.

O anotador de expressões globais requer o ficheiro do conjunto de regras. O ficheiro do conjunto de regras é disponibilizado para o anotador de expressões globais através da interface `com.ibm.uima.an_regex.FileResource`, que é implementada pela classe `com.ibm.uima.an_regex.impl.FileResource_impl`. Para passar as regras personalizadas para o anotador de expressões globais, tem de fornecer o nome do ficheiro do conjunto de regras e adicionar a localização do ficheiro ao caminho da classe. A chave que o anotador de expressões globais utiliza para aceder ao ficheiro do conjunto de regras denomina-se `RuleSetDefinition`. Não altere esta chave, caso contrário o anotador de expressões globais não encontrará o conjunto de regras e o anotador não conseguirá inicializar.

Amostra

A secção de recursos externos do descritor é da seguinte forma:

```
<externalResourceDependencies>
 <externalResourceDependency>
<key>RuleSetDefinition</key>
<description>Definição do conjunto de regras</description>
<interfaceName>com.ibm.uima.an_regex.FileResource</interfaceName>
<optional>false</optional>
 </externalResourceDependency>
</externalResourceDependencies>
<resourceManagerConfiguration>
 <externalResources>
  <externalResource>
<name>of_samples_regex_rules</name>
<description>Ficheiro da definição do conjunto de regras para números
das salas</description>
   <fileResourceSpecifier>
<fileUrl>file:of_samples_regex_rules.xml</fileUrl>
   </fileResourceSpecifier>
   <implementationName>
    com.ibm.uima.an_regex.impl.FileResource_impl</implementationName>
  </externalResource>
 </externalResources>
 <externalResourceBindings>
  <externalResourceBinding>
<key>RuleSetDefinition</key>
<resourceName>of_samples_regex_rules</resourceName>
  </externalResourceBinding>
 </externalResourceBindings>
</resourceManagerConfiguration>
```

Conceitos relacionados

"Anotador de expressões globais" na página 83
O anotador de expressões globais permite-lhe executar a análise de texto personalizada sem ter necessidade de implementar o seu próprio motor de análise de texto. Com base num conjunto de regras (expressões globais) que pode definir por si próprio, o anotador de expressões globais detecta as estruturas de informações em documentos de texto e cria anotações das informações detectadas na estrutura de análise comum.

"Ficheiro do conjunto de regras" na página 86
No anotador de expressões globais, o ficheiro do conjunto de regras XML define as regras, na forma de expressões globais, que são utilizadas para analisar o documento de texto.

Referências relacionadas

"Registar"
Todas as mensagens de registo do anotador de expressões globais são escritas no ficheiro de registo da colecção actual.

Registar

Todas as mensagens de registo do anotador de expressões globais são escritas no ficheiro de registo da colecção actual.

Os ficheiros de registo da colecção estão localizados em ES_NODE_ROOT/logs/ e têm nomes com o formato *<id_colecção>_<data_actual>*.log. É possível visualizar os ficheiros de registo utilizando os scripts esviewlogs.sh/.bat.

Existem sete níveis de registo possíveis:

- Erro (Error)

- Aviso (Warning)
- Informações (Info)
- Configuração (Config)
- Optimizado (Fine)
- Detalhado (Finer)
- Pormenorizado (Finest)

Não pode alterar o mapeamento para mensagens de Erro e Aviso. Por predefinição, apenas mensagens Informações (Info), Aviso (Warning) e Erro (Error) são escritas no ficheiro de registo. Estes são os níveis de registo padrão utilizados pelo Enterprise Search. Os outros níveis de registo podem ser mapeados para obter informações mais detalhadas.

Para receber mensagens de registo a partir do anotador de expressões globais, o nível de registo tem de estar definido, pelo menos, como Configuração (Config). Neste nível, o anotador regista definições de configuração, tal como, o ficheiro do conjunto de regras que é utilizado e o nome da classe de implementação para a interface `com.ibm.uima.an_regex.String2Number`.

Se definir o nível de registo para Detalhado (Finer), por exemplo, o anotador regista quais as anotações que não foi possível criar. Este nível pode ajudar a determinar porque nem todas as anotações que está a esperar foram criadas. Por exemplo, poderia existir um erro numa das expressões globais ou um grupo de captura opcional pode não ter encontrado correspondência para qualquer texto no documento. Do mesmo modo, se especificar que uma funcionalidade deve ser definida com a sequência de texto que corresponde a um grupo de captura e não existir sequência de texto correspondente, a funcionalidade é definida como nula.

Para obter informações muito detalhadas, defina o nível de registo como Pormenorizado (Finest). Neste nível, o anotador regista o padrão de expressão global actual, a parte do texto de documento que está a ser analisada e quaisquer anotações e funcionalidades que tenham sido criadas. Ao utilizar o registo muito detalhado, especialmente os níveis de registo Detalhado (Finer) e Pormenorizado (Finest), tem um impacto negativo no desempenho global do anotador.

Se requerer o mapeamento do nível de registo detalhado, modifique o ficheiro de configuração denominado tokenizer.properties em *ES_NODE_ROOT*/master_config/parserservice/ alterando a definição da configuração `trevi.tokenizer.jedii.InformationalLevelMapping=Info` em `trevi.tokenizer.jedii.InformationalLevelMapping=Finest`, por exemplo.

Para activar as alterações do nível do registo, tem de parar todos os processos do analisador utilizando a consola de administração. Em seguida, tem de parar e reiniciar de seguida a sessão de inicialização do analisador a partir da linha de comandos chamando:

```
>esadmin session parserservice stop
>esdamin session parserservice start
```

Depois disso, a análise pode ser novamente iniciada e deve agora ter o novo nível de registo. Tem de repetir estes passos sempre que alterar o nível do registo.

Conceitos relacionados

"Anotador de expressões globais" na página 83
O anotador de expressões globais permite-lhe executar a análise de texto personalizada sem ter necessidade de implementar o seu próprio motor de

análise de texto. Com base num conjunto de regras (expressões globais) que pode definir por si próprio, o anotador de expressões globais detecta as estruturas de informações em documentos de texto e cria anotações das informações detectadas na estrutura de análise comum.

"Ficheiro do conjunto de regras" na página 86
No anotador de expressões globais, o ficheiro do conjunto de regras XML define as regras, na forma de expressões globais, que são utilizadas para analisar o documento de texto.

Referências relacionadas

"Descritor do anotador" na página 92
O descritor XML do anotador de expressões globais contém informações descritivas sobre o anotador de expressões globais que é necessário para executar o anotador.

Documentação de Enterprise Search

Pode ler a documentação do OmniFind Enterprise Edition em formato PDF ou HTML.

O programa de instalação do OmniFind Enterprise Edition instala automaticamente o centro de informações do IBM Content Discovery, que inclui as versões HTML da documentação dos produtos OmniFind Enterprise Edition, Versão 8.4 e WebSphere Information Integrator Content Edition, Versão 8.3. Para instalação em vários servidores, o Information Center é instalado em todos os servidores de procura. Se não instalar o centro de informações, quando clicar na ajuda, o centro de informações abre num sítio da Web da IBM.

Para consultar versões instaladas dos documentos em PDF, avance para ES_INSTALL_ROOT/docs/*locale*/pdf. Por exemplo, para encontrar documentos em inglês, avance para ES_INSTALL_ROOT/docs/en_US/pdf.

Para aceder às versões em PDF da documentação em todos os idiomas disponíveis, consulte o sítio da Web OmniFind Enterprise Edition, Version 8.4 documentation.

Pode também aceder a descarregamentos de produtos, pacotes de correcções, notas técnicas e ao centro de informações a partir do sítio da Web OmniFind Enterprise Edition Support.

A seguinte tabela mostra a documentação disponível, nomes de ficheiros e localizações.

Tabela 12. Documentação para Enterprise Search

Título	Nome do ficheiro	Localização
Information Center		http://publib.boulder.ibm.com/infocenter/discover/v8r4/
Manual de Instalação para Enterprise Search	iiysi.pdf	ES_INSTALL_ROOT/docs/*locale*/pdf/
Manual de Iniciação Rápida (Este documento também está disponível em cópia impressa em inglês, francês e japonês.)	QuickStartGuide_*locale de duas letras*.pdf	ES_INSTALL_ROOT/docs/*locale*/pdf/
Requisitos de Instalação para Enterprise Search	iiysr.txt ou iiysr.htm	ES_INSTALL_ROOT/docs/*locale*/ (pode também aceder a este ficheiro a partir do painel de lançamento da instalação)
Administração de Enterprise Search	iiysa.pdf	ES_INSTALL_ROOT/docs/*locale*/pdf/
Manual de Programação e Referência de API para Enterprise Search	iiysp.pdf	ES_INSTALL_ROOT/docs/en_US/pdf/
Manual de Correcção de Problemas e Referência de Mensagens	iiysm.pdf	ES_INSTALL_ROOT/docs/*locale*/pdf/

Tabela 12. Documentação para Enterprise Search (continuação)

Título	Nome do ficheiro	Localização
Integração de Análise de Texto	iiyst.pdf	ES_INSTALL_ROOT/docs/*locale*/pdf/
Suplemento para o Google Desktop Search	iiysg.pdf	ES_INSTALL_ROOT/docs/*locale*/pdf/
Notas de Edição	iiysn.pdf	Disponível no sítio da Web OmniFind Enterprise Edition, Version 8.4 documentation (pode também aceder a este ficheiro a partir do painel de lançamento da instalação)

Acessibilidade do WebSphere Information Integrator OmniFind Edition

As interfaces do utilizador do IBM WebSphere Information Integrator OmniFind Edition e a documentação estão acessíveis.

Programa de instalação

Pode utilizar teclas de atalho para navegar e avançar no programa de instalação do WebSphere Information Integrator OmniFind Edition. A tabela seguinte descreve algumas teclas de atalho.

Tabela 13. Teclas de atalho para o programa de instalação

Acção	**Atalho**
Destacar um selector de opção	Tecla de seta
Seleccionar um selector de opção	Tecla Tab
Destacar um selector de acção	Tecla Tab
Seleccionar um selector de acção	Tecla Enter
Ir para janela seguinte ou anterior, ou cancelar	Destacar um selector de acção premindo a tecla Tab e premir Enter
Tornar inactiva a janela activa	Ctrl + Alt + Esc

Consola de administração de Enterprise Search e centro de informações

A consola de administração e o centro de informações são interfaces baseadas no navegador que pode visualizar no Microsoft Internet Explorer ou Mozilla FireFox. Consulte a ajuda online para o Internet Explorer ou FireFox para obter uma lista de teclas de atalhos e outras funcionalidades de acessibilidade para o navegador.

Documentação PDF

Pode ver toda a documentação de Enterprise Search em PDF. Os documentos PDF são acessíveis através do Adobe Acrobat Version 6.0. Os documentos em PDF são estruturados e devem ser passíveis de serem lidos pela maioria dos leitores de ecrã.

Glossário de termos para Enterprise Search

Este glossário define termos utilizados nas interfaces e na documentação de Enterprise Search.

lista de controlo de acessos (ACL)
Uma lista que composta por um ou vários IDs de utilizador ou grupos de utilizadores e respectivos privilégios associados. As listas de controlo de acessos são utilizadas para controlar o acesso dos utilizadores a itens e objectos.

função administrativa
Uma classificação de um utilizador que determina as funções que o utilizador pode executar na consola de administração de Enterprise Search. A função também determina quais as colecções que o utilizador pode gerir.

motor de análise
Ver motor de análise de texto.

resultados da análise
As informações que são produzidas por anotadores. Os resultados de análise são escritos numa estrutura de dados denominada estrutura de análise comum. Os resultados de análise produzidos pelos motores de análise de texto personalizados (anotadores) podem ser disponibilizados para procura através da respectiva inclusão no índice de Enterprise Search.

anotação
Informações sobre um grupo de recursos de rede de texto. Por exemplo, uma anotação pode indicar que um grupo de recursos de rede de texto representa um nome de empresa. Em Unstructured Information Management Architecture (UIMA), uma anotação é um tipo especial de estrutura funcional.

anotador
Um componente de software que executa tarefas de análise linguística específicas e produz e regista anotações. Um anotador é o componente lógico de análise num motor de análise.

procura booleana
Uma procura na qual um ou mais termos da procura são combinados utilizando operadores como AND, NOT e OR.

classe hierárquica
Uma especificação que pode influenciar a classificação relativa de um documento nos resultados da procura.

palavra hierárquica
Uma palavra que pode influenciar a classificação relativa de um documento nos resultados de procura. Durante o processamento da consulta, a importância de um documento que contenha uma palavra hierárquica pode ser aumentada ou diminuída, dependendo de uma classificação predefinida para a palavra.

categoria
Um grupo de documentos que têm propriedades semelhantes.

árvore de categorias
Uma hierarquia de categorias que é apresentada na consola de administra de Enterprise Search.

certificado
Um documento digital que associa uma chave pública à identidade do proprietário do certificado, permitindo dessa forma a autenticação do proprietário do certificado. Um certificado é emitido por uma entidade certificadora.

entidade certificadora
Uma organização que emite certificados e autentica entidades (individuais ou organizações) que estejam envolvidas em transacções electrónicas. As entidades certificadoras garantem que as duas partes envolvidas na troca de informações são realmente quem dizem ser.

normalização de caracteres
Um processo no qual as formas variantes de um carácter, tais como maiúsculas e marcas diacríticas, são reduzidas a um formato comum.

clítico Uma palavra que funciona sintacticamente em separado, mas que está ligada foneticamente a outra palavra. Um clítico pode ser escrito como ligado ou separado da palavra à qual está associado. Exemplos comuns de clíticos incluem a parte final de uma contracção em inglês (*would**n't*** ou *you**'re***).

colecção
Um conjunto de dados e opções para pesquisar, analisar, indexar e procurar as origens de dados.

estrutura de análise comum (CAS)
Uma estrutura que armazena o conteúdo e os metadados de um documento, e todos os resultados de análise produzidos por um motor de análise de texto. Durante a análise de documentos, todas as trocas de dados são processadas utilizando a estrutura de análise comum.

consumidor de estrutura de análise comum (consumidor CAS)
Um consumidor de estrutura de análise comum efectua o processamento final dos resultados de análise armazenados na estrutura de análise comum. Por exemplo, um consumidor indexa o conteúdo da estrutura de análise comum num motor de procura ou preenche uma base de dados relacional com resultados de análise específicos.

nível de comunicação comum (CCL)
A infra-estrutura de comunicações que une os diversos componentes (controlador, analisador, ferramenta de sequências de hiperligações, servidor de índices) do WebSphere Information Integrator OmniFind Edition.

extracção de conceitos
Função de análise de texto que identifica itens de vocabulário significativos (tais como pessoas, locais ou produtos) em documentos de texto e produz uma lista desses itens. Ver, também, extracção de temas.

espaço de sequência de hiperligações
Um conjunto de origens que correspondem a padrões especificados (tais como Uniform Resource Locators (URLs), nomes de bases de dados, caminhos de sistemas de ficheiros, nomes de domínios, e endereços de IP) que uma ferramenta de sequências de hiperligações lê para obter itens para indexação.

ferramenta de sequências de hiperligações
Um programa de software que obtém documentos de origens de dados e recolhe informações que podem ser utilizadas para criar índices remissivos de procura.

credencial
Informações detalhadas, que são adquiridas durante a autenticação, que descrevem atributos da identidade relacionados com segurança, do utilizador e de quaisquer associações de grupo. As credenciais podem ser utilizadas para executarem um grande número de serviços, tais como autorização, examinação e delegação.

motor de análise de texto personalizada
Um motor de análise de texto que é criado utilizando o kit de desenvolvimento de software (SKD) Unstructured Information Management Architecture (UIMA) e pode ser adicionado ao conjunto de motores de análise de texto de Enterprise Search padrão (também conhecidos como anotadores base de Enterprise Search). Ver, também, motor de análise de texto.

origem de dados
Qualquer repositório de dados do qual seja possível obter dados, tal como a Web, bases de dados relacionais e não relacionais e sistemas de gestão de conteúdos.

tipo de origem de dados
Um agrupamento de origens de dados de acordo com o protocolo que está a ser utilizado para aceder aos dados.

arquivo de dados
Uma estrutura de dados em que os documentos são mantidos na sua forma analisada. O analisador escreve ao arquivo de dados. O arquivo de dados é utilizado para criar o índice, bem como para gerar resumos de procura. Não confundir com arquivo de dados não processados.

criação de índice delta
O processo de adicionar informações novas a um índice remissivo existente num sistema de Enterprise Search. Contraste com criação de índice principal.

desenfileirar
Para remover os itens de uma fila.

diacríticos
Uma marca que é adicionada à letra para alterar a pronúncia de uma palavra ou para distinguir entre palavras semelhantes, tal como uma marca de acento ou o trema alemão.

descobridor
Uma função de uma ferramenta de sequências de hiperligações que determina quais as origens de dados que estão disponíveis para a ferramenta de sequências de hiperligações obter informações.

nome distinto
O nome que identifica de modo único uma entrada num dicionário. Um nome distinto é constituído por atributo:pares de valores, separado por vírgulas. Para além disso, também é um conjunto de pares de valores de nome (tais como NP=nome da pessoa e P=País ou região) que identificam de modo único uma entidade num certificado digital.

Document Object Model
Um sistema no qual um documento estruturado, tal como um ficheiro XML, é visualizado como uma árvore de objectos que podem ser acedidos e actualizados de forma programada.

arquivo Domino Document Manager
Uma base de dados Domino Document Manager que é utilizada para organizar documentos. Os arquivos contêm bases de dados Domino.

Biblioteca Domino Document Manager
Uma base de dados Domino Document Manager que é o ponto de entrada para o Domino Document Manager.

Domino Internet Inter-ORB Protocol (DIIOP)
Uma tarefa de servidor executada num servidor e funciona com o Solicitador de Pedido Objecto Domino para permitir a comunicação entre applets Java que são criados com as classes de Java do Notes e o servidor Domino. Os utilizadores de navegadores e servidores Domino utilizam o DIIOP para comunicarem e trocarem dados de objectos.

classificação dinâmica
Um tipo de classificação no qual os termos na consulta são analisados em relação aos documentos que estão a ser procurados para determinar a classificação de resultados. Ver, também, classificação baseada em texto. Contraste com classificação estática.

resumo dinâmico
Um tipo de resumo no qual os termos da procura são destacados e os resultados da procura contêm frases que representam da melhor forma os conceitos do documento que o utilizador procura. Contraste comresumo estático.

colocar em fila
Colocar itens numa fila.

administrador de Enterprise Search
Uma função administrativa que permite a um utilizador gerir todo o sistema de Enterprise Search.

anotadores base de Enterprise Search
Um conjunto de motores de análise de texto personalizada utilizado em Enterprise Search para o processamento de análise de documentos predefinido.

carácter de alteração de controlo
Um carácter que suprime ou selecciona um significado especial para um ou mais caracteres que venham depois.

origem de dados externa
Uma origem de dados para federação que não é pesquisada, analisada ou indexada pelo WebSphere Information Integrator OmniFind Edition. As procuras de origens de dados externas são delegadas à interface de programação da aplicação para consultas das mesmas origens de dados.

caminho funcional
Caminho utilizado para aceder ao valor de uma funcionalidade numa estrutura funcional de Unstructured Information Management Architecture (UIMA).

estrutura funcional
A estrutura de dados subjacente que representa o resultado da análise de texto. Uma estrutura funcional é uma estrutura atributo-valor. Cada

estrutura funcional é de um certo tipo, e cada tipo tem um conjunto especial de funções válidas ou atributos, muito semelhante a uma classe de Java.

procura federada
Uma capacidade de procura que permite procuras através de vários serviços de procura e devolve uma lista consolidada de resultados da procura.

federação
O processo de combinar sistemas de atribuição de nomes de forma a que o sistema agregado possa processar nomes compostos de qualquer sistema de atribuição de nomes.

campo A parte identificável mais pequena de um registo.

procura por campo
Uma consulta restrita a um determinado campo.

procura de texto livre
Uma procura na qual o termo da procura é exprimido como texto de forma livre.

índice remissivo de texto completo
Uma estrutura de dados que consulta itens de dados para permitir que a procura encontre rapidamente documentos que contenham os termos da consulta.

procura aproximada
Uma procura que devolve palavras com ortografia semelhante à do termo da consulta.

procura híbrida
Uma combinação de procura booleana e procura de texto livre.

gestão de identidade
A capacidade de validar as credenciais actuais de um utilizador com controlos de acesso nativos. Se uma origem de dados estiver protegida por um produto que suporte a autenticação de início de sessão único (SSO, single sign-on), e se a ferramenta de sequências de hiperligações estiver configurada para utilizar a segurança de SSO, os mecanismos de SSO serão utilizados para autenticar o utilizador. Caso contrário, as credenciais do utilizador são codificadas num arquivo protegido que possa ser actualizado quando os controlos de acesso nativos forem alterados.

índice remissivo
Consultar índice remissivo de texto completo.

fila de índice remissivo
Uma lista de pedidos para processamento da criação de índices principais e diferenciais.

extracção de informações
Um tipo de extracção de conceitos que reconhece automaticamente itens de vocabulário relevantes, tais como nomes, termos e expressões, em documentos de texto.

endereço de IP
O endereço de 32-bit exclusivo que identifica um computador central na rede.

Java Database Connectivity (JDBC)
Uma norma industrial para conectividade independente da base de dados

entre a plataforma Java e um vasto leque de bases de dados. A interface JDBC fornece uma API ao nível de chamada para acesso de base de dados com base em SQL.

JavaScript
Uma linguagem da Web utilizada em navegadores e servidores da Web.

JavaServer Pages (JSP)
Uma tecnologia de escrita de servidor que permite que código Java seja incorporado dinamicamente dentro de páginas da Web (ficheiros HTML) e que seja executado quando a página é apresentada, de forma a devolver conteúdo dinâmico a um cliente.

Java virtual machine (JVM)
Uma implementação de software que executa código Java compilado (applets e aplicações).

Katakana
Um conjunto de caracteres constituído por símbolos utilizados em um dos dois alfabetos fonéticos japoneses comuns, que é utilizado principalmente para escrever palavras estrangeiras de forma fonética.

ficheiro de arquivo de chaves
Um ficheiro de base de dados que contém chaves públicas que são armazenadas como certificados do assinante e chaves privadas que são armazenadas em certificados pessoais.

identificação do idioma
Uma função de Enterprise Search que determina o idioma de um documento.

lema A forma base de uma palavra. O lemas são significativos em idiomas altamente flexionados, como o checo.

formação de lemas
O processo de consulta do lema para uma determinada palavra num dicionário. A formação de lemas é diferente de stemming na medida em que o stemming é algorítmico e, em geral, não funciona com um dicionário que elabore uma lista de palavras de um idioma.

afinidade lexical
Relação entre palavras de procura num documento que estão próximas umas das outras ao nível do significado. A afinidade lexical é utilizada para calcular a pertinência de um resultado.

biblioteca
Um objecto de sistema que serve como um directório para outros objectos. Ver, também, Biblioteca do Domino Document Manager.

ditongo
Dois ou mais caracteres ligados de forma a aparecerem como um único carácter, tal como a junção de a e e, que forma o ditongo æ.

Lightweight Directory Access Protocol (LDAP)
Um protocolo aberto que utiliza TCP/IP para fornecer acesso a directórios que suportem um modelo X.500 e que não está sujeito aos requisitos de recurso do Directory Access Protocol X.500 mais complexos.

procura linguística
Um tipo de procura que procura, obtém e indexa um documento com

termos que reduzidos à forma base (por exemplo, para que *caracteres* seja indexado como *carácter*) ou expandido com a forma base (tal como com palavras compostas).

análise de ligações
Um método baseado na análise de hiperligações entre documentos e utilizado para determinar quais as páginas na colecção que são importantes para os utilizadores.

federador local
Um federador cliente que é associado através de um conjunto de objectos passíveis de serem procurados.

local do Lotus QuickPlace
Um local de reunião da Web que é fornecido pelo Lotus QuickPlace que permite que participantes dispersos geograficamente colaborem em projectos e comuniquem online numa área de trabalho estruturada e protegida.

sala do Lotus QuickPlace
Uma área particionada de um local do Lotus QuickPlace que é restrita a membros autorizados que partilham um interesse comum e necessitam de trabalhar colectivamente.

criação de índice principal
O processo de criação do índice completo num sistema de Enterprise Search. Contraste com criação de índice delta.

carácter de máscara
Um carácter que é utilizado para representar caracteres opcionais antes, no meio e depois de um termo da procura. Os caracteres de máscara são utilizados normalmente para encontrar variações de um termo num índice remissivo. Ver, também, carácter global.

tipo MIME
Um padrão da Internet para identificar o tipo de objecto que está a ser transferido através da Internet.

supervisor
Um utilizador de Enterprise Search que tem autoridade para observar processos de nível de colecção.

consulta de idioma natural
Tipo de procura que analisa expressões escritas (tais como ″Quem é o director do departamento financeiro?″) em vez de simples conjuntos de palavras-chave.

carácter de mudança de linha
Um carácter de controlo faz com que a posição de impressão ou de apresentação se mova uma linha para baixo. Alguns sistemas necessitam de mais de um carácter.

segmentação n-grama
Um método de análise que considera sequências sobrepostas de um determinado número de caracteres como uma única palavra, em vez de utilizar um espaço em branco para delimitar palavras como na segmentação de espaço em branco baseada em Unicode.

directiva de não seguimento
Uma directiva numa página da Web que dá instruções a robôs (tal como a ferramenta de sequências de hiperligações da Web) para não seguirem ligações encontradas nessas páginas.

directiva sem índice remissivo
Uma directiva numa página da Web que dá instruções a robôs (tal como a ferramenta de sequências de hiperligações da Web) para não incluírem o conteúdo dessas páginas no índice remissivo.

Chamada de procedimento remoto do Notes (NRPC)
Mecanismo de comunicação do Lotus Notes utilizado para todas as comunicações de Notes para Notes.

operador
Um utilizador de Enterprise Search que tem a autoridade para observar, iniciar e parar processos de nível de colecção.

procura paramétrica
Um tipo de procura que procura objectos que contenham um atributo ou valor numérico, tal como datas, números inteiros ou outros tipos de dados numéricos dentro de um intervalo especificado.

analisador
Um programa que interpreta documentos adicionados ao arquivo de dados de Enterprise Search. O analisador extrai informações dos documentos e prepara-os para indexar, procurar e obter.

controlador de analisador
Serviço de Enterprise Search que alimenta o serviço analisador com documentos. Existe um controlador de analisador por cada colecção. O serviço de controlador de analisador de uma colecção corresponde ao analisador da colecção na consola de administração de Enterprise Search.

serviço analisador
O serviço de Enterprise Search que processa a totalidade da análise de documentos e do processamento da análise de texto em colecções de documentos. Existe sempre um serviço analisador em execução, no mínimo.

local Uma localização virtual visível no portal onde indivíduos e grupos se encontram para colaborarem. Num portal, cada utilizador tem um local pessoal para trabalho privado, e os indivíduos e grupos têm acesso a uma variedade de locais partilhados, que tanto podem ser locais públicos como restritos. Ver, também, local do Lotus QuickPlace.

classificação popular
Um tipo de classificação baseado na popularidade do documento que é adicionado a uma classificação existente de um documento.

Portal Document Manager (PDM)
Permite aos utilizadores ter um repositório de documentos central para trabalho de equipa. Os administradores têm a capacidade de gerir os documentos de forma eficiente e podem controlar o modo como os utilizadores interagem com as informações.

arquivo de motor de processamento
Um ficheiro de arquivo .pear zip que inclui um motor de análise Unstructured Information Management Architecture (UIMA) e todos os recursos requeridos para o utilizar em análises personalizada em Enterprise Search.

procura de proximidade
Um tipo de procura que procura determinadas palavras na mesma frase, parágrafo ou documento.

servidor proxy
Um servidor que faz de intermediário para pedidos HTTP da Web que sejam alojados por uma aplicação ou um servidor da Web. Um servidor proxy faz de substituto para os servidores de conteúdo na empresa.

ligação rápida
Associação entre um Uniform Resource Identifier (URI) e palavras-chave ou expressões.

classificação
O processo de atribuir um valor inteiro a cada documento nos resultados da procura de uma consulta. A ordem dos documentos nos resultados da procura é baseada na pertinência em relação à consulta. Uma classificação mais alta significa uma correspondência mais aproximada. Ver, também, classificação dinâmica e classificação estática.

arquivo de dados não processados
Estrutura de dados onde os documentos pesquisados são armazenados antes de serem enviados ao analisador. As ferramentas de sequências de hiperligações escrevem ao arquivo de dados não processados, e o analisador lê a partir do arquivo de dados não processados. Uma vez analisados os documentos, estes são removidos do arquivo de dados não processados. Não confundir com arquivo de dados.

anotador de expressões globais
O anotador de expressões globais detecta entidades ou unidades de informação num documento de texto, como, por exemplo, números de telefone, números de produtos, nomes de funcionários ou endereços, com base em expressões globais que descrevem os padrões exactos procurados no documento de texto. Se uma das expressões globais corresponder a partes do texto do documento, o anotador de expressões globais cria as anotações correspondentes que incluem a correspondência, ou parte da mesma. Estas expressões anotadas são posteriormente armazenadas, ou no índice de Enterprise Search, através de um ficheiro de correlação de índice, ou numa base de dados compatível com JDBC, através de um ficheiro de correlação de base de dados.

federador remoto
Um federador de servidor que associa um conjunto de objectos passíveis de serem procurados.

Robots Exclusion Protocol
Um protocolo que permite aos administradores de sítios da Web indicarem aos robôs visitantes quais as partes do sítio que não devem ser visitadas pelo robô.

sala Um programa que permite aos utilizadores criarem documentos para outros lerem, responderem a comentários de outros utilizadores e reverem o estado do projecto e as datas de conclusão. Os utilizadores também pode conversar com outros utilizadores que se encontrem na mesma sala. Ver, também, Sala do Lotus QuickPlace.

categoria baseada em regras
Categorias criadas por regras que especificam quais os documentos que estão associados a quais categorias. Por exemplo, pode definir regras para associação de documentos que contenham ou excluam determinadas palavras, ou que correspondam a um padrão de Uniform Resource Identifier (URI), com categorias específicas.

âmbito
Grupo de Uniform Resource Identifiers (URIs) relacionados utilizado para definir o âmbito de um pedido de procura.

aplicação de procura
Um programa que processa consultas, procura o índice remissivo, devolve os resultados da procura e obtém os documentos origem num sistema de Enterprise Search.

memória cache de procura
Uma memória tampão que mantém os dados e os resultados de pedidos de procura anteriores.

motor de procura
Um programa que aceita um pedido de procura e devolve uma lista de documentos ao utilizador.

ficheiros de índices de procura
O conjunto de ficheiros no qual um índice remissivo é armazenado no motor de procura.

resultados da procura
Uma lista de documentos que correspondem ao pedido da procura.

Secure Sockets Layer (SSL)
Um protocolo de segurança que fornece privacidade de comunicações.

token de segurança
Informações sobre a identidade e a segurança que são utilizadas para autorizar o acesso a documentos numa colecção. Diferentes tipos de origem de dados suportam diferentes tipos de tokens de segurança. Os exemplos incluem funções de utilizador, IDs de utilizador, IDs de grupo e outras informações que podem ser utilizadas para controlar o acesso a conteúdos.

página de lista de seeds
No WebSphere Portal, uma página XML que contém ligações para as páginas disponíveis num portal. As ferramentas de sequências de hiperligações podem utilizar a lista de seeds para identificar os documentos a pesquisar. A página da lista de seeds contém também metadados armazenados em conjunto com os documentos pesquisados no índice de Enterprise Search.

Uniform Resource Locator (URL) inicial
O ponto de partida para uma sequência de hiperligações.

segmentação
A divisão de texto em unidades lexicais distintas. O processamento não baseado em dicionários inclui segmentação n-grama e espaço em branco, ao passo que o suporte baseado em dicionários inclui segmentação de palavras, frases e parágrafos, e formação de lemas.

procura semântica
A procura semântica melhora o paradigma de procura por palavra-chave através da incorporação de mais conhecimento linguístico e do domínio da solução de procura. A tecnologia que inclui e aplica este conhecimento é referida como análise de texto.

servlet
Um programa de Java que é executado num servidor da Web e estende a funcionalidade do servidor gerando conteúdos dinâmicos como resposta aos pedidos de clientes da Web. Os servlets são utilizados frequentemente para ligarem bases de dados à Web.

shingle
Uma cadeia de tokens (palavras) consecutivos retirados de uma frase. Por exemplo, de ″Esta é uma frase muito curta.″, os shingles de 3 palavras (ou trigramas) são:

Esta é uma
é uma frase
uma frase muito
frase muito curta

Os shingles podem ser utilizados em linguística estatística. Por exemplo, se dois textos diferentes tiverem muitos shingles em comum, os textos estão provavelmente relacionados de alguma forma.

página de erros esporádicos
Uma página especial que explica o problema detalhadamente se um servidor de HTTP não conseguir devolver a página que o cliente pediu, e, configura o servidor de HTTP para devolver estas páginas em vez de uma resposta que consista apenas num cabeçalho com um código de retorno a indicar o problema.

classificação estática
Um tipo de classificação no qual os factores sobre os documentos que estão a ser classificados, tais como a data, o número de ligações que apontam para o documento e etc., aumentam a classificação. Contraste comclassificação dinâmica.

resumo estático
Um tipo de resumo no qual os resultados da procura contêm um resumo específico, armazenado do documento. Contraste comresumo dinâmico.

stemming
Consultar stemming de palavras.

palavra de paragem
Uma palavra que é utilizada frequentemente, tal como *o*, *um* ou *e*, que é ignorada pela aplicação de procura.

remover palavras de paragem
O processo de remover palavras de paragem da consulta para ignorar palavras comuns e devolver resultados mais relevantes.

resumo
O processo de incluir instruções em resultados da procura para descrever brevemente o conteúdo de um documento. Ver, também, resumo dinâmico e resumo estático.

dicionário de sinónimos
Um dicionário que permite ao utilizador procurar sinónimos dos termos da consulta quando procuram uma colecção.

taxonomia
Uma classificação de objectos em grupos baseada em semelhanças. No Enterprise Search, uma taxonomia organiza os dados em categorias e subcategorias. Ver, também, árvore de categorias.

análise de texto
O processo de extrair semântica e outras informações do texto para melhorar a possibilidade de obtenção de dados numa colecção.

motor de análise de texto
Um componente de software que é responsável por encontrar e representar conteúdos semânticos e de contexto em textos.

classificação baseada em texto
O processo de atribuir um número inteiro a um documento que signifique a importância do documento em relação aos termos numa consulta. Um valor inteiro mais elevado significa uma correspondência mais aproximada com a consulta. Ver, também, classificação dinâmica.

segmentação de texto
Ver segmentação.

extracção de temas
Um tipo de extracção de conceitos que reconhece automaticamente itens de vocabulário relevantes em documentos de texto para extrair o tema ou tópico de um documento.Ver, também, extracção de conceitos.

token As unidades textuais básicas que são indexadas por Enterprise Search. Os tokens podem ser as palavras num idioma ou outras unidades de texto adequadas para indexação.

tokenização
Ver segmentação.

segmentador
Um programa de segmentação que digitaliza texto e determina se e quando uma série de caracteres pode ser reconhecida como um token.

carácter de seguimento
Um carácter que tem a última posição numa palavra.

sistema de tipos
O sistema de tipos define os tipos de objectos (estruturas funcionais) que podem ser identificados num documento por um motor de análise de texto. O sistema de tipos define todas as estruturas funcionais possíveis em termos de tipos e funções. Pode definir qualquer número de tipos diferentes num sistema de tipos. Um sistema de tipos é específico de domínio e aplicação.

segmentação de espaços em branco baseada em Unicode
Um método de segmentação que utiliza propriedades de carácter Unicode para distinguir entre token e caracteres separadores.

Uniform Resource Identifier (URI)
Uma cadeia de caracteres compacta que identifica um recurso abstracto ou físico.

Uniform Resource Locator (URL)
Uma sequência de caracteres que representa recursos de de informação num computador ou numa rede como a Internet. Esta sequência de caracteres inclui o nome abreviado do protocolo que é utilizado para aceder ao recurso de informação e às informações utilizadas pelo protocolo para localizar o recurso de informação.

Universal Resource Name (URN)
Um elemento de protocolo da Internet que consiste numa cadeia curta de caracteres em conformidade com uma determinada sintaxe. A cadeia inclui um nome que pode ser utilizado para fazer referência a um recurso.

Unstructured Information Management Architecture (UIMA)
Uma arquitectura da IBM que define um contexto para implementar sistemas para a análise de dados não estruturados.

agente do utilizador
Uma aplicação que procura a Web e deixa informações próprias nos sítios que visita. No Enterprise Search, a ferramenta de sequências de hiperligações da Web é um agente do utilizador.

ferramenta de sequências de hiperligações da Web
Uma classe de software robô que explora a Web obtendo um documento da Web e seguindo as ligações dentro desse documento.

procura de termo ponderado
Uma consulta em que é dada mais importância a determinados termos.

carácter global
Um carácter que é utilizado para representar caracteres opcionais antes, no meio ou depois de um termo da procura.

stemming de palavras
Um processo de normalização linguística no qual as formas variantes de uma palavra são reduzidas a um formato comum. Por exemplo, palavras como *programação*, *programado*, e *programável* são reduzidas a *programa*.

XML Path Language (XPath)
Uma linguagem que identifica exclusivamente ou dirige-se a partes de um documento de XML de origem. O XPath também fornece opções básicas para manipular cadeias, números e operadores booleanos.

Aceder a informações sobre o Content Management and Discovery

Encontram-se disponíveis informações sobre produtos do IBM Content Management and Discovery por telefone ou na Web.

Os números de telefone aqui disponibilizados são válidos nos E.U.A.:

- Para encomendar produtos ou para obter informações gerais: 1-800-IBM-CALL (1-800-426-2255)
- Para encomendar publicações: 1-800-879-2755

É possível encontrar informações sobre os produtos do IBM Content Management and Discovery na Web através de http://www.ibm.com/software/sw-bycategory/subcategory/SWB40.html. Este sítio contém ligações que podem ajudá-lo a:

- Obter informações sobre os produtos
- Adquirir os produtos
- Participar em testes experimentais e beta para os produtos
- Obter suporte sobre produtos

Para aceder à documentação do produto:

1. Visite a página Web em http://www.ibm.com/software/sw-bycategory/subcategory/SWB40.html.
2. Seleccione um produto sobre o qual pretenda obter mais informações, por exemplo, WebSphere Information Integrator OmniFind Edition. Este sítio contém ligações para:
   - Documentação do produto, incluindo notas de edição e centros de informações online
   - Requisitos do sistema
   - Descarregamentos do produto
   - Pacotes de correcções
   - Novidades sobre o produto
   - Materiais de suporte do produto, tais como, documentação técnica e IBM Redbooks
   - Newsgroups e grupos de utilizadores
   - Instruções para encomendar livros
3. Faça clique na ligação Suporte (Support), no lado esquerdo da página.
4. Na secção Informações (Learn), seleccione o tipo de documentação que pretende visualizar. Se estiver disponível um centro de informações para o produto seleccionado, pode seleccionar a ligação correspondente ao centro de informações.

Fornecer comentários sobre a documentação

Agradecemos o envio de quaisquer comentários sobre estas informações ou sobre outra documentação da IBM.

O fornecimento de informações ajuda a IBM a prestar informações de qualidade. Agradecemos o envio de quaisquer comentários sobre estas informações ou sobre

outra documentação que o IBM Software Development envia com os respectivos produtos. Pode utilizar qualquer dos seguintes métodos para fornecer comentários:

1. Envie os comentários utilizando o formulário de comentários para leitores online que se encontra na página Web www.ibm.com/software/awdtools/rcf/.
2. Envie os comentários por mensagem de correio electrónico para comments@us.ibm.com. Inclua o nome do produto, a respectiva versão e o nome e part number das informações (se aplicável). Se estiver a comentar sobre texto específico, inclua a localização do texto (por exemplo, um título, um número de tabela ou número de página).

Contactar a IBM

Para contactar a assistência a clientes da IBM nos E.U.A. ou Canadá, ligue 1-800-IBM-SERV (1-800-426-7378).

Para obter informações sobre as opções de assistência disponíveis, telefone para um dos seguintes números:

- Nos E.U.A.: 1-888-426-4343
- No Canadá: 1-800-465-9600

Para localizar um escritório da IBM no seu país ou região, consulte o sítio da Web IBM Directory of Worldwide Contacts em www.ibm.com/planetwide.

Avisos e marcas comerciais

Avisos

Estas informações foram desenvolvidas para produtos e serviços disponibilizados nos E.U.A. Os produtos, serviços ou funções descritos neste documento poderão não ser disponibilizados pela IBM noutros países. Consulte o seu representante IBM para obter informações sobre os produtos e serviços actualmente disponíveis na sua área. Quaisquer referências, nesta publicação, a produtos, programas ou serviços IBM não significam que apenas esses produtos, programas ou serviços IBM possam ser utilizados. Qualquer outro produto, programa ou serviço, funcionalmente equivalente, poderá ser utilizado em substituição daqueles, desde que não infrinja nenhum direito de propriedade intelectual da IBM. No entanto, é da inteira responsabilidade do utilizador avaliar e verificar o funcionamento de qualquer produto, programa ou serviço não IBM.

Neste documento, podem ser feitas referências a patentes ou a pedidos de patente pendentes. O facto de este documento lhe ser fornecido não lhe confere quaisquer direitos sobre essas patentes. Para solicitar pedidos de informação sobre licenças, tais pedidos deverão ser endereçados, por escrito, a: IBM Director of Licensing; IBM Corporation; North Castle Drive; Armonk, NY 10504-1785 U.S.A.

Para endereçar os seus pedidos de informação sobre licenças relacionados com informações de conjunto de caracteres de duplo byte (DBCS, Double Byte Character Set), contacte o Departamento de Propriedade Intelectual do seu país ou envie-os, por escrito, para: IBM World Trade Asia Corporation Licensing 2-31 Roppongi 3-chome, Minato-ku Tokyo 106-0032, Japan

O parágrafo seguinte não se aplica ao Reino Unido nem a nenhum outro país onde estas cláusulas sejam inconsistentes com a lei local: A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO “TAL COMO ESTÁ (AS IS)”, SEM GARANTIA DE QUALQUER ESPÉCIE, EXPLÍCITA OU IMPLÍCITA, INCLUINDO, MAS NÃO SE LIMITANDO ÀS GARANTIAS IMPLÍCITAS DE NÃO INFRACÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO FIM. Alguns Estados não permitem a exclusão de garantias, explícitas ou implícitas, em determinadas transacções; esta declaração pode, portanto, não se aplicar ao seu caso.

Esta publicação pode conter imprecisões técnicas ou erros de tipografia. A IBM permite-se fazer alterações periódicas às informações aqui contidas; essas alterações serão incluídas nas posteriores edições desta publicação. A IBM pode introduzir melhorias e/ou alterações ao(s) produto(s) e/ou programa(s) descrito(s) nesta publicação em qualquer momento, sem aviso prévio.

Quaisquer referências, nesta publicação, a sítios da Web não IBM são fornecidas apenas para conveniência e não constituem, em caso algum, aprovação desses sítios da Web. Os materiais existentes nesses sítios da Web não fazem parte dos materiais destinados a este produto IBM e a utilização desses sítios da Web será da exclusiva responsabilidade do utilizador.

A IBM pode usar ou distribuir quaisquer informações que lhe forneça, da forma que julgue apropriada, sem incorrer em nenhuma obrigação para consigo.

Os Licenciados deste programa que pretendam obter informações sobre o mesmo com o objectivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) a utilização recíproca das informações que tenham sido trocadas, deverão contactar a IBM através do seguinte endereço:

IBM Corporation J46A/G4
555 Bailey Avenue
San Jose, CA 95141-1003 U.S.A.

Tais informações poderão estar disponíveis, sujeitas aos termos e às condições adequados, incluindo, em alguns casos, o pagamento de um encargo.

O programa licenciado descrito neste documento e todo o material licenciado disponível para o programa são fornecidos pela IBM nos termos das Condições Gerais IBM (IBM Customer Agreement), do Acordo de Licença Internacional para Programas IBM (IPLA, IBM International Program License Agreement) ou de qualquer acordo equivalente entre ambas as partes.

Quaisquer dados de desempenho aqui contidos foram determinados num ambiente controlado. Assim sendo, os resultados obtidos noutros ambientes operativos podem variar significativamente. Algumas medições podem ter sido efectuadas em sistemas ao nível do desenvolvimento, pelo que não existem garantias de que estas medições sejam iguais nos sistemas disponíveis habitualmente. Para além disso, algumas medições podem ter sido calculadas por extrapolação. Os resultados reais podem variar. Os utilizadores deste documento devem verificar os dados aplicáveis ao seu ambiente específico.

As informações relativas a produtos não IBM foram obtidas junto dos fornecedores desses produtos, dos seus anúncios publicados ou de outras fontes de divulgação ao público. A IBM não testou esses produtos e não pode confirmar a exactidão do desempenho, da compatibilidade ou de quaisquer outras afirmações relacionadas com produtos não IBM. Todas as questões sobre as capacidades dos produtos não IBM deverão ser endereçadas aos fornecedores desses produtos.

Todas as afirmações relativas às directivas ou tendências futuras da IBM estão sujeitas a alterações ou descontinuação sem aviso prévio, representando apenas metas e objectivos.

Estas informações contêm exemplos de dados e relatórios utilizados em operações comerciais diárias. Para ilustrá-los o melhor possível, os exemplos incluem nomes de indivíduos, firmas, marcas e produtos. Todos estes nomes são fictícios e qualquer semelhança com nomes e moradas reais é mera coincidência.

LICENÇA DE COPYRIGHT:

Esta publicação contém programas de aplicação exemplo na linguagem de origem, que ilustra técnicas de programação em várias plataformas operativas. Pode copiar, modificar e distribuir estes programas exemplo de qualquer forma, sem encargos para com a IBM, com a finalidade de desenvolver, utilizar, comercializar ou distribuir programas de aplicação conformes à interface de programação de aplicações e destinados à plataforma operativa para a qual os programas exemplo são escritos. Estes exemplos não foram testados exaustivamente sob todas as condições. Deste modo, a IBM não garante nem se responsabiliza pela fiabilidade, assistência ou funcionamento implícito destes programas. Pode copiar, modificar e distribuir estes programas exemplo de qualquer forma, sem encargos para com a

IBM, com a finalidade de desenvolver, utilizar, comercializar ou distribuir programas de aplicação conformes à interfaces de programação de aplicações da IBM.

Cada cópia, ou qualquer parte destes programas exemplo, ou qualquer trabalho derivado dos mesmos, tem de incluir um aviso de direitos de autor, do seguinte modo:

Outside In (®) Viewer Technology, © 1992-2006 Stellent, Chicago, IL., Inc. Todos os Direitos Reservados.

IBM XSLT Processor Materiais Licenciados - Propriedade da IBM ©Copyright IBM Corp., 1999-2006. Todos os Direitos Reservados.

Marcas comerciais

Este tópico lista as marcas comerciais da IBM e determinadas marcas comerciais não IBM.

Consulte o sítio da Web http://www.ibm.com/legal/copytrade.shtml para obter informações sobre as marcas comerciais da IBM.

Os seguintes termos são marcas comerciais ou marcas comerciais registadas de outras empresas:

Java e todas as marcas comerciais baseadas em Java e logotipos são marcas comerciais ou marcas comerciais registadas da Sun Microsystems, Inc. nos Estados Unidos e/ou noutros países.

Microsoft, Windows, Windows NT e o logotipo do Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou noutros países.

Intel, Intel Inside (logotipos), MMX e Pentium são marcas comerciais da Intel Corporation nos Estados Unidos e/ou noutros países.

UNIX é uma marca comercial registada de The Open Group nos Estados Unidos e noutros países.

Linux é uma marca comercial da Linus Torvalds nos Estados Unidos e/ou noutros países.

Outros nomes de empresas, produtos ou serviços podem ser marcas comerciais ou marcas de serviços de outras empresas.

Índice Remissivo

U

V

SC17-5463-01