MANUAL DO SOFTWARE

TEXT MINING SUITE® V.2.4.7

por

InText Mining Ltda.

www.intext.com.br

atendimento@intext.com.br

Versão do Manual: 10

ÍNDICE

1 Introdução

O software Text Mining Suíte é um conjunto de ferramentas para Descoberta de Conhecimento em Textos (Text Mining ou Mineração de Textos).

A principal técnica do software é a análise de conceitos (ou contextos ou temas) presentes nos textos. Conceitos representam objetos, eventos, situações ou idéias do mundo real. Eles são representados por palavras. Entretanto, a análise de palavras pode levar a erros no entendimento dos conceitos presentes numa coleção textual. Por exemplo, numa coleção de ocorrências policiais, alguns casos são registrados usando o termo "assassinato" e outros usando o termo "homicídio". Uma análise apenas de palavras poderia indicar, por exemplo, que 30% dos casos são sobre "assassinato" e 20% sobre "homicídio". Entretanto, a análise correta deveria ser que 50% dos casos envolvem o crime de homicídio ou assassinato.

Outro erro comum na análise simples de palavras é verificar se uma determinada palavra aparece num texto. Entretanto, o sentido da frase pode mudar se a palavra "não" estiver presente antes do verbo.

Assim, o software TMS proporciona ao usuário analisar os conceitos presentes nos textos ao invés de palavras.

Conceitos são característicos e dependentes da aplicação (ou área de domínio). Por exemplo, na área médica, conceitos podem ser sintomas de doenças ou características de pacientes. Em discursos políticos, conceitos podem ser ideologias. O usuário é quem deve definir que conceitos são interessantes para análise e como eles podem aparecer nos textos.

Análises de conceitos podem ser usadas em diferentes aplicações, por exemplo:

- análise de discursos;
- análise qualitativa de resultados de pesquisas (com questões abertas);
- análise de registros tais como chamados e manutenções técnicas, ocorrências policiais, casos jurídicos ou médicos.

Neste *manual*, somente serão explicadas algumas ferramentas para análise de textos. O símbolo ⓘ será usado para indicar uma informação importante que pode gerar problemas no uso do software. É algo com que o usuário deve ter atenção.

ⓘ Acompanham a este software os arquivos referentes às listas de stopwords (stw-espanhol.stw, stw-inglês.stw, stw-portugues.stw).

ⓘ Vídeos sobre o uso do software estão disponíveis no Youtube no canal

www.youtube.com/intextmining

2 Liberação do Software

Para ser utilizado, o software precisa ser liberado ou ativado. Isto se faz solicitando à InText Mining um arquivo de liberação. Junto da solicitação, deverá ser enviado o código numérico fornecido pelo software.

3 Diferenças para a Versão Anterior

Diferenças em relação a versões anteriores a 2.4.7:
Correção de bugs.

Novidades em relação a versões anteriores a 2.3.8:
Não há mais limitação no nome dos arquivos textos (os nomes podem ter até 255 caracteres). Entretanto, os sistemas operacionais podem não conseguir gerenciar arquivos com nomes muito longos ou cujo caminho de diretórios seja muito longo.
ⓘ ATENÇÃO: se você possui uma versão anterior, as listas de documentos antigas não poderão mais ser utilizadas a partir da versão 2.3.8. Você deverá criar novas listas de documentos. Da mesma forma, todos os resultados intermediários também deverão ser gerados novamente (resultados de classificação, clusterização, recuperação por similaridade, etc).

Em todas as ferramentas, nas caixas onde aparecem os nomes dos arquivos textos (lista de documentos), ao se dar 2 cliques no nome do documento ou texto, uma caixa é aberta mostrando o conteúdo do arquivo texto.

Novidades em relação a versões anteriores a 2.3.3:

Foi incorporado um novo Wizard (Assistente) para Comparação de Textos. Este assistente é útil para trabalhos de Inteligência Competitiva, podendo comparar descrições de produtos, sites de concorrentes, etc.

Novidades em relação a versões anteriores a 2.3.0:

a) Nos resultados da mineração feita pelo Wizard, agora é possível filtrar as associações apresentadas ou ordenar (por confiança ou alfabeticamente).

b) Bugs corrigidos:
- Criação de conceitos na ontologia: alguns conceitos novos eram perdidos.
- Salva automaticamente o projeto após cada passo (usuário não precisa salvar); se der algum problema (ex: travar o software em algum passo), o Wizard reinicia o projeto do passo em que parou.

4 Requisitos e Cuidados

Há as seguintes restrições ou limitações no software:

- somente podem ser analisados arquivos no **formato .TXT** (há uma ferramenta para transformar arquivos html em txt);
- os nomes de arquivos textos devem ter no máximo 255 **caracteres sem espaços e sinais especiais a não ser hífen ou sublinhado (underlined)**;
- os nomes de conceitos não devem ter mais de 15 caracteres, nem espaços ou caracteres especiais (somente sublinhado é permitido e não se deve usar o hífen);
- para os nomes de listas de documentos e listas de conceitos não há restrições;
- a versão de demonstração (Demo) só minera no máximo 100 textos.

Alguns cuidados que podem evitar problemas na execução:

- todos os arquivos textos devem estar no mesmo diretório;
- se for utilizado o Wizard e vários textos estiverem dentro de um arquivo, o Wizard poderá separar os textos, mas só poderá haver um **ÚNICO** arquivo texto no diretório;
- deixe o software e os arquivos com as listas de stopwords (arquivos .stw) no mesmo diretório;
- o software utiliza muita memória RAM, por isto, ao executar minerações demoradas (com grandes volumes de textos ou com textos muito grandes), procure retirar da memória RAM programas desnecessários;
- para minerações com grandes volumes de textos (mais de 100) ou com textos muito grandes (mais de 1 kbytes), utilize um computador com boa capacidade de processamento e memória;
- algumas ferramentas têm opções para utilizar o método "contextual": neste caso, é necessário fazer antes a classificação dos textos utilizando uma ontologia (os conceitos e suas regras devem ter sido definidos antes e deve ter sido criada uma lista de conceitos ou contextos);
- atenção para o ponto decimal: ao informar valores numéricos com casas decimais, o usuário deve ter atenção porque, dependendo do sistema operacional utilizado, o ponto decimal muda (pode ser . ou ,);
- em máquinas com pouca capacidade de processamento, algumas execuções podem demorar; procure não utilizar outro software ao mesmo tempo; se for utilizar algo em paralelo, o software TMS pode não mostrar as telas de acompanhamento do status do processamento (ou indicadores de status na linha abaixo da tela); entretanto, o software não deve travar; para certificar-se de que o software continua o processamento, verifique se o tamanho dos arquivos aumenta (se sim, o software está processando);
- em algumas ferramentas, o software não mostra uma tela indicando que terminou o processamento, mas apenas uma mensagem abaixo da tela; também podem ficar vazias algumas coisas na tela; verifique a mensagem na parte de baixo da tela;
- certifique-se de que o arquivo de stopwords foi baixado junto com o software (pelo menos um arquivo); junto com o software, são disponibilizadas 3 listas de stopwords (cada uma num arquivo diferente do tipo .stw).

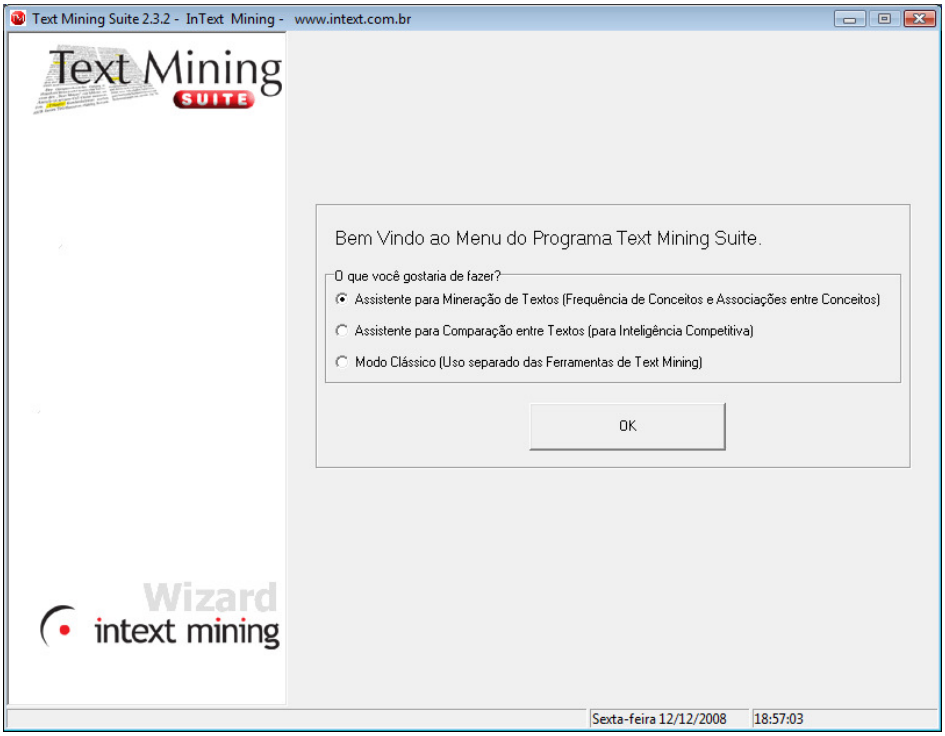
Sistema Operacional recomendado: Windows Vista

Foram relatados problemas com o Windows XP (ver na seção de Troubleshooting).

5 Tela Inicial do software

O software inicia com um menu com as seguintes opções:

a) Wizard (Assistente) para Mineração de Textos (Análise de Freqüência de Conceitos e Associações entre Conceitos)
b) Wizard (Assistente) para Comparação entre Textos (para Inteligência Competitiva)
c) Modo Clássico: permite utilizar todas as ferramentas da suíte.



A vantagem de utilizar um Wizard é que o usuário não precisa saber que ferramentas estão sendo utilizadas, nem como usar, nem em que ordem. Basta seguir os passos, respondendo às perguntas ou opções oferecidas.

O primeiro Wizard (Mineração de Textos tradicional) segue um processo padrão de Mineração dos Textos, com os seguintes passos:

- Preparação dos textos: escolha da linguagem, separação dos textos (se estiverem todos dentro de um mesmo documento), análise e representação interna dos textos;
- Criação da ontologia: definição de conceitos e das regras para identificação dos conceitos nos textos;
- Mineração sobre os conceitos (Concept-based Text Mining): descoberta dos conceitos nos textos, análise estatística dos conceitos (freqüência nos textos) e descoberta de associações entre conceitos (Data Mining).

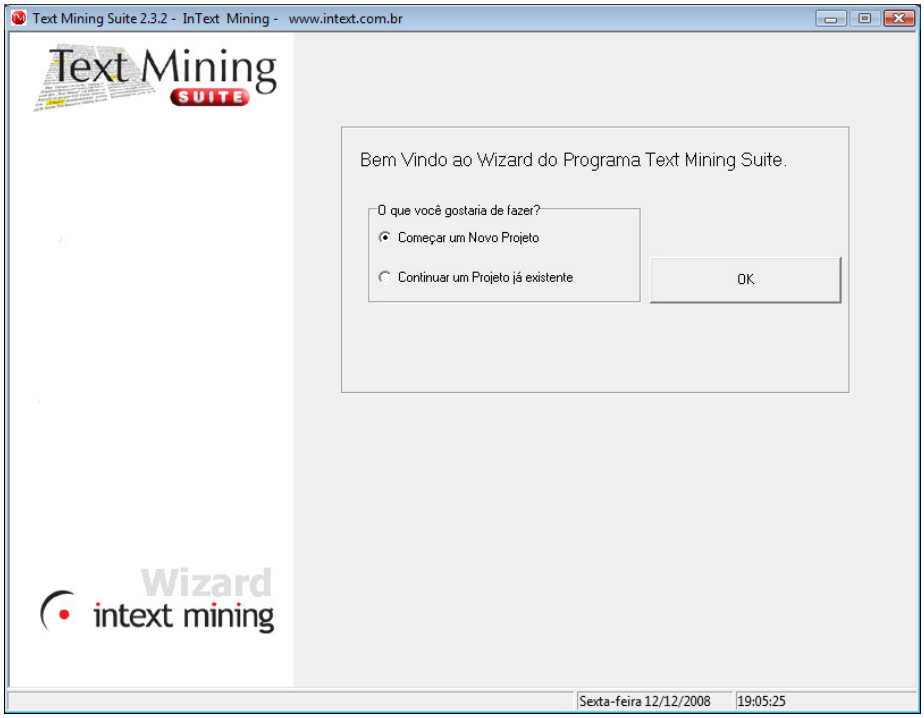
O segundo Wizard (Comparação entre Textos) permite comparar palavras que aparecem em diversos textos, com os seguintes passos:

- Preparação dos textos: escolha da linguagem, separação dos textos (se estiverem todos dentro de um mesmo documento), análise e representação interna dos textos;
- Análise de palavras que aparecem em mais de um texto:

- Análise de palavras exclusivas de cada texto (ao ser selecionado um texto, o software apresenta as palavras que só aparecem neste texto).

O Modo Clássico permite acessar outras ferramentas de Mineração que não estão no processo padrão de mineração, ou seja, que não são utilizadas pelo Wizard.

Os Assistentes iniciam perguntando se o usuário deseja iniciar um novo projeto ou continuar um projeto já existente. Na primeira opção, o software inicia pelo 1º passo. Se for escolhido continuar um projeto existente, o software solicita o nome do projeto (arquivo do tipo .ini) e vai direto ao passo em que o projeto foi interrompido (ou salvo).



6 Wizard (Assistente) para Mineração de Textos

Para utilização do Wizard, basta seguir os passos respondendo às perguntas ou escolhendo dentre as opções fornecidas. Após, as escolhas, basta clicar no botão "avançar" (canto inferior direito da tela). Se quiser refazer algum passo, clique no botão "voltar".

Se quiser sair do software, interrompendo a mineração, utilize os botões "sair":

- "Sair, salvando status do projeto": no próximo reinício do software (próxima vez que for executado), o Wizard continuará o projeto do mesmo passo onde estava (desde que solicitado "continuar um projeto já existente");
- "Sair, cancelando este passo": interrompe a execução do software e abandona (cancela) o projeto em andamento;
- "Voltar ao início": cancela o projeto em andamento e retorna para o menu inicial do software.

6.1 Passo 1 – Configurar Projeto

O título é obrigatório. Os campos "cliente" e "responsável" são opcionais.

Utilize o botão "alterar" para indicar a pasta onde os estão a serem minerados estão armazenados. O software identifica todos os arquivos texto (.txt) na pasta selecionada. Somente arquivos textos podem ser minerados pelo software.

ⓘ Somente deixe na pasta selecionada os arquivos textos que você deseja minerar.

6.2 Passo 2 – Preparar Textos

Este passo fará a preparação dos textos, gerando arquivos com representações internas dos textos (um arquivo do tipo lpl para cada arquivo texto).

- **Textos separados X Textos juntos num mesmo arquivo**

Se os textos estão cada um num arquivo txt separado, basta escolher a opção "vários arquivos".
Utilize a opção "todos num mesmo arquivo", se os textos a serem minerados estão juntos num mesmo arquivo txt. Neste caso, deve haver uma string (seqüência de caracteres) que separa os textos dentro do arquivo único. Esta string deverá ser informada no campo "string separador". Neste caso, o software vai separar os textos em arquivos individuais, colocando um nome padrão do tipo **"nome_projeto-XXXXX.txt",** onde XXXX é um número seqüencial (isto é feito automaticamente pelo módulo separador).

- **Escolha da língua em que estão os textos**

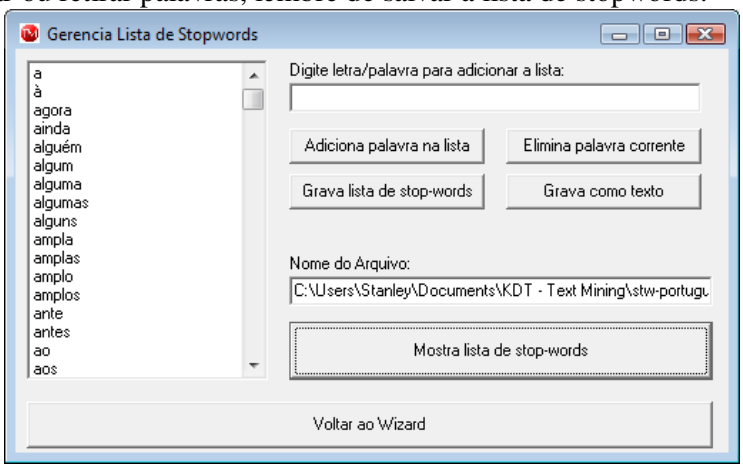
O software TMS aceita textos em português, espanhol ou inglês.
Neste passo, o software seleciona a lista de stopwords a serem utilizadas (arquivos do tipo stw).
ⓘ Para textos em Português, o software procurará pelo arquivo com nome "stw-portugues.stw"; para textos em Inglês, pelo arquivo "stw-ingles.stw" e para textos em Espanhol, pelo arquivo "stw-espanhol.stw". Certifique-se de que os arquivos .stw estejam no mesmo diretório do software.

Este passo é importante porque o software irá desconsiderar as stopwords na representação interna dos textos. Stopwords são palavras muito freqüentes e sem significado tais como artigos, preposições e algumas conjunções.

- **Gerência de Stopwords**

O usuário pode modificar a lista de stopwords, acrescentando ou retirando palavras.

ⓘ Se acrescentar ou retirar palavras, lembre de salvar a lista de stopwords.



6.3 Passo 3 – Ontologia

Uma ontologia é um conjunto de Conceitos (temas ou assuntos) e as regras que permitem identificar estes conceitos nos textos.

ⓘ Os nomes de conceitos não podem ter mais que 15 caracteres. Não devem ser usados caracteres especiais além de letras e números. Somente o caractere "_" (sublinhado ou underlined) é permitido (não utilizar hífen).

Opções:

- Usar uma Ontologia existente: permite reutilizar uma Ontologia que já foi criada anteriormente; neste caso, também é possível, após a importação da Ontologia, modificar os conceitos ou suas regras;
- Criar uma nova Ontologia: permite criar um novo conjunto de conceitos e regras;
- Ver uma Ontologia existente: permite apenas visualizar os conceitos e regras de uma ontologia já existente.

Para ver os conceitos existentes, utilize o combo de seleção de conceitos existentes.

No método **Determinístico**, o usuário deve especificar regras para identificação do conceito nas frases. Cada regra deve ser especificada em uma única linha e será analisada dentro de cada frase (cada frase do texto será analisada em relação a todas as regras definidas). Este método é melhor quando se precisa fazer uma análise de contexto mais detalhada. Por exemplo, se as palavras “bola” e “futebol” aparecerem na mesma frase, pode-se afirmar que o conceito “futebol” está presente.

As regras não podem ter mais que 30 caracteres e devem incluir as palavras positivas, aquelas que obrigatoriamente devem aparecer na frase (avaliação por conjunção = E lógico), e as negativas (opcionais), indicadas por um sinal de menos “-“ antes da palavra. As palavras negativas indicam que o conceito não existe. Isto é útil para casos onde o “não” aparece, o que inverte o significado.

Então, para que uma regra seja verdadeira, todas as palavras positivas devem aparecer em alguma frase e nesta mesma frase não deve aparecer nenhuma das palavras negativas definidas na mesma regra. A ordem das palavras não será avaliada, mas elas devem ser

fornecidas com no mínimo um espaço entre elas. Neste método, não é necessário fornecer pesos, pois se uma das regras for verdadeira, o conceito será identificado.

O caracter $ ao final de uma palavra é utilizado para indicar que outros caracteres podem aparecer ou não.

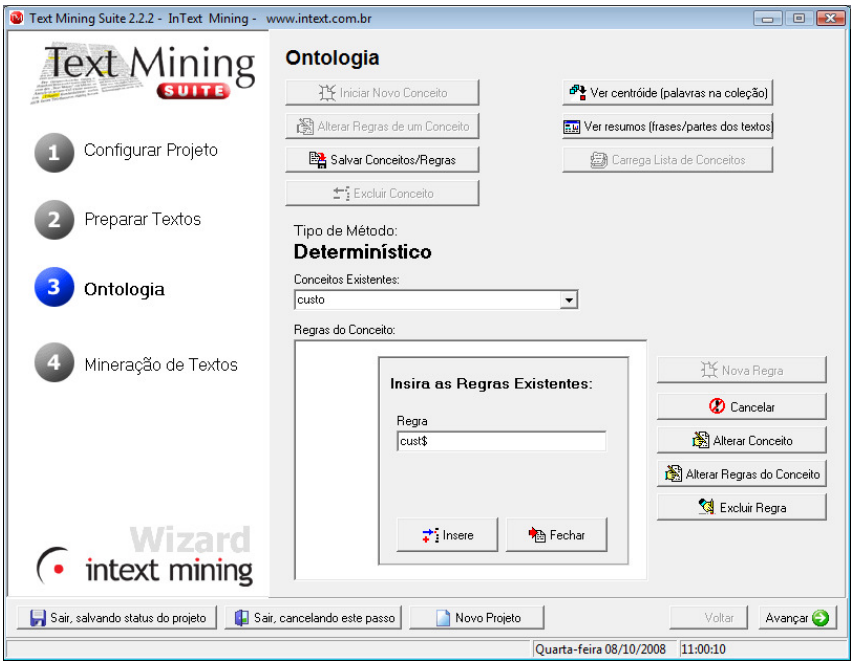
No processo de classificação, cada regra verdadeira soma 1 a um contador de presença do conceito no texto.

Exemplo de conceito e regras:

Regras para o Conceito Alcoolismo:
alcool$
álcool$
bebe imoderadamente
bebe muito –não
não pára beber

Explicação:
A 1ª regra indica que se uma palavra iniciada por "alcool" for encontrada na frase, o conceito alcoolismo estará presente no texto.
A 2ª regra é semelhante à 1ª mas utiliza uma variação com acento.
A 3ª regra indica que, se numa frase, aparecerem as palavras "bebe" e "imoderadamente", o conceito estará presente. Note-se que estas palavras não precisam estar necessariamente juntas (pode haver palavras entre elas) nem nesta ordem, mas devem estar na mesma frase.
A 4ª regra indica que as palavras "bebe" e "muito" devem aparecer na mesma frase (não importando a ordem ou se há outras palavras entre elas) para o conceito estar presente. Entretanto, o sinal – indica que a palavra "não" é negativa, ou seja, não pode aparecer na frase (se ela aparecer, o conceito não está presente).
A 5ª regra indica que as palavras devem aparecer na mesma frase (não importando a ordem ou se há palavras entre elas).

6.3.1 Iniciar novo conceito



Opções:

- Cancelar: cancela a operação em andamento
- Alterar conceito: altera nome
- Alterar regras do conceito: usuário deve selecionar uma regra e esta será exibida para ser modificada
- Excluir regra: exclui a regra selecionada

ⓘ Não esquecer de "salvar conceitos e regras" após a inserção de conceitos ou modificação de regras (seja um conceito novo ou já existente).

6.3.2 Alterar regras de um conceito

Usuário deve selecionar um conceito e o software mostrará todas as regras deste conceito. Após, selecionar uma regra e clicar em "alterar regras do conceito".

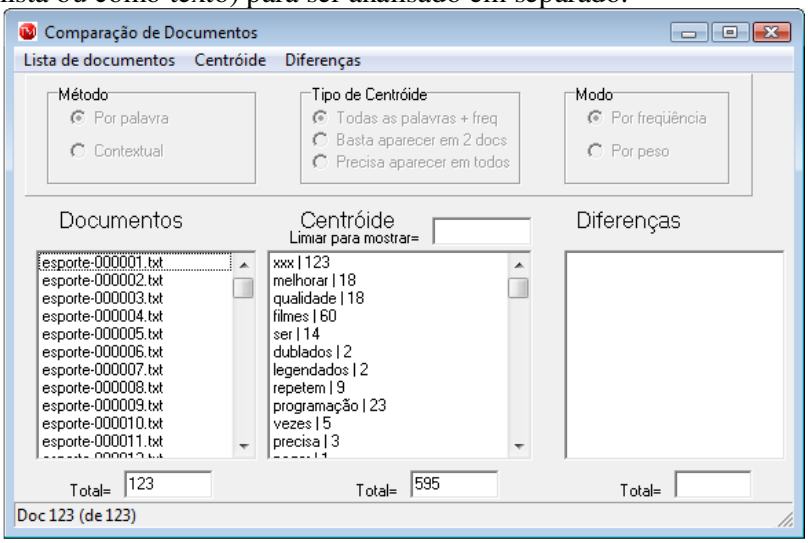
6.4 Ferramentas de Auxílio à Criação da Ontologia

Para criar a Ontologia, o usuário deve conhecer a forma de expressão dos conceitos, ou seja, como as pessoas estão falando dos conceitos (que palavras e expressões estão sendo utilizadas).

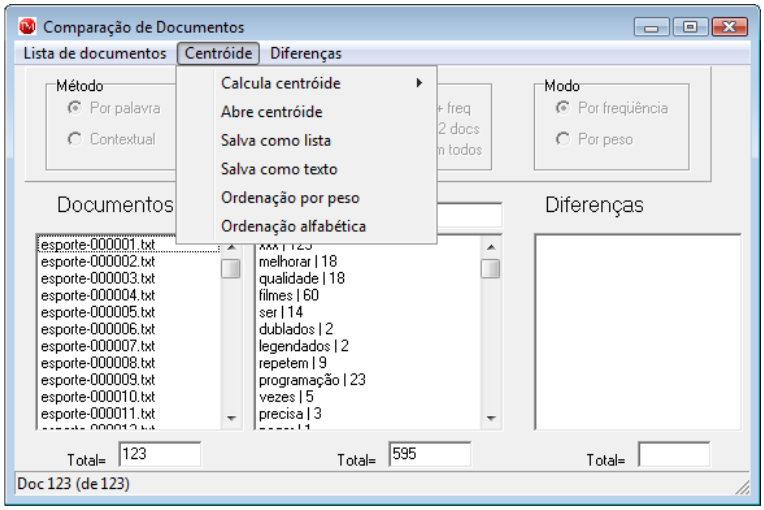
Abaixo, são explicadas duas ferramentas que auxiliam neste processo.

6.4.1 Encontrar Centróide (Análise Léxica)

O Centróide é a lista de todas as palavras que aparecem nos textos. Ao lado da palavra, o número indica a quantidade de textos onde esta palavra aparece na coleção a ser minerada. Pode-se ordenar esta lista por peso ou alfabeticamente. Também é possível salvar o centróide (salvar como lista ou como texto) para ser analisado em separado.



Clicar em “Centróide” para acessar opções de salvar ordenar e abrir centróide já salvo.



6.4.2 Extrair Resumos dos Textos (Frases contendo certas palavras)

Esta ferramenta comporta-se como descrito na seção 12.

6.5 Passo 4 – Mineração de Textos

Na passagem do passo 3 para o o 4º passo, acontecem 3 sub-processos:

1) Identificação de Conceitos nos Textos
Conforme as regras definidas na Ontologia, os conceitos são identificados nos textos.

2) Extrair centróide contextual
É feita a análise de distribuição (freqüência) dos conceitos nos textos. O resultado é ujma lista de conceitos com a sua freqüência absoluta (número de textos onde aparece o conceito) e a freqüência relativa (valor em percentual)

3) Fazer associações conceituais e gerar regras do tipo Se-Então
Este sub-processo identifica associações entre conceitos, mostrando a confiança (probabilidade condicional da associação) e o suporte (número de casos onde a associação aparece). As associações são apresentadas na forma de regras Se-Então.

Os resultados são apresentados em duas caixas:

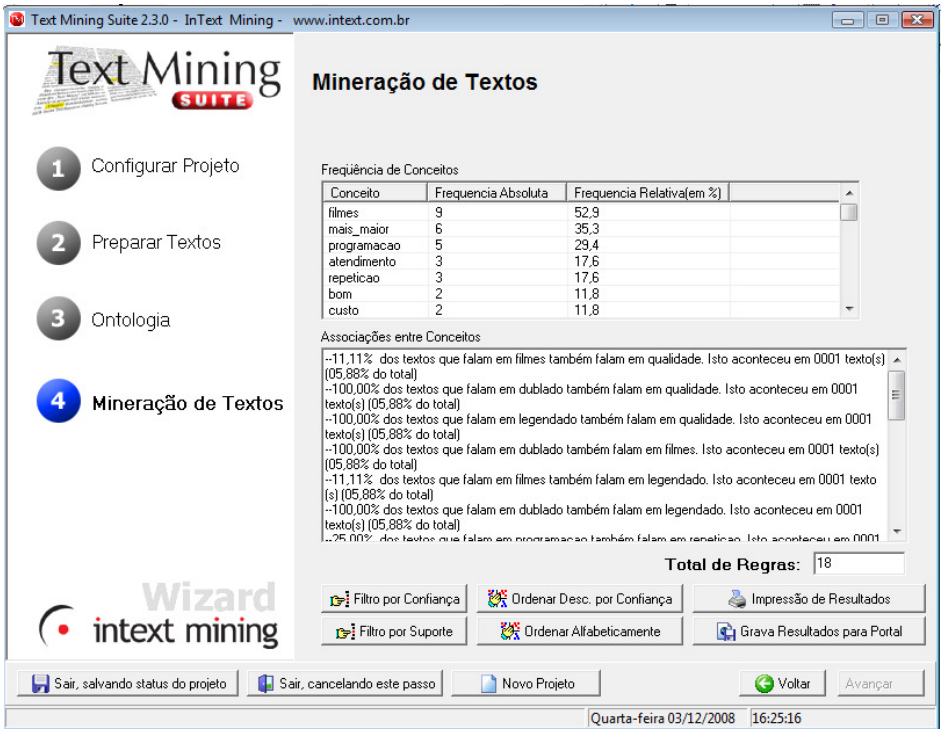
a) Freqüência de Conceitos:
a caixa mais acima apresenta os conceitos e a freqüência de cada um deles; a freqüência absoluta informa o número de textos em que o conceito aparece e a freqüência relativa

apresenta um valor percentual relativo à freqüência absoluta (% de textos onde o conceito aparece).
Para ordenar a lista de conceitos e freqüências, basta clicar no título da coluna desejada. Clicando-se na 1ª coluna ("Conceito"), a lista será ordenada alfabeticamente pelo nome do conceitos. Clicando-se na 2ª ou 3ª coluna, a lista será ordenada de foram decrescente pelo valor da freqüência (absoluta ou relativa).

b) Associações entre Conceitos:
a 2ª caixa de resultados apresenta as associações descobertas entre os conceitos. As associações aparecem no formato "X% os textos que falam em KL também falam em MN. Isto acontece em Y texto(s) (Z% do total)".
O valor X apresenta a confiança da regra (probabilidade condicional) entre os conceitos KL e MN. O número Y apresenta o valor absoluto do suporte e o número Z apresenta o suporte em valor percentual. Há uma caixa que informa o total de regras de Associação listadas nos resultados.

Há 5 botões que podem ajudar:

- Filtro por Confiança: o usuário deve informar um valor mínimo para a confiança; isto permite selecionar um grupo de regras de associação, eliminando regras com confiança menor que o valor informado;
- Filtro por Suporte: o usuário deve informar um valor mínimo (valor absoluto) para o suporte; isto permite selecionar um grupo de regras de associação, eliminando regras com suporte menor que o valor informado;
- Ordenar Desc. Por Confiança: ordena as regras apresentadas na caixa de resultados por valor de confiança (do maior para o menor);
- Ordenar Alfabeticamente: ordena as regras apresentadas na caixa de resultados de forma alfabética pelo nome do 1º conceito;
- Impressão de Resultados: serve para gerar um arquivo texto com todos os resultados (freqüências de conceitos e associações entre conceitos).



7 Assistente (Wizard) para Comparação de Textos

Este Assistente ajuda na comparação de diversos textos, encontrando palavras que aparecem em mais de um texto e também quais as palavras exclusivas de cada texto (ou seja, que palavras aparecem somente num texto).

Para utilização do Wizard, basta seguir os passos respondendo às perguntas ou escolhendo dentre as opções fornecidas. Após, as escolhas, basta clicar no botão "avançar" (canto inferior direito da tela). Se quiser refazer algum passo, clique no botão "voltar".

Se quiser sair do software, interrompendo a mineração, utilize os botões "sair":

- "Sair, salvando status do projeto": no próximo reinício do software (próxima vez que for executado), o Wizard continuará o projeto do mesmo passo onde estava (desde que solicitado "continuar um projeto já existente");
- "Sair, cancelando este passo": interrompe a execução do software e abandona (cancela) o projeto em andamento;
- "Voltar ao início": cancela o projeto em andamento e retorna para o menu inicial do software.

7.1 Passo 1 – Configurar Projeto

Este passo é feito como no Wizard anterior (seção 6.1).

7.2 Passo 2 – Preparar Textos

Este passo é feito como no Wizard anterior (seção 6.2).

7.3 Passo 3 – Análise dos Documentos (Comparação dos Textos)

Este passo pode ser demorado. Ao final da análise, aparece uma tela em branco como a abaixo.

Para ver os resultados, o usuário deve selecionar o tipo de resultado a ser apresentado.

a) Palavras que aparecem em mais de um texto

Selecione esta opção para ver as palavras que aparecem em mais de um texto do conjunto de documentos.

Clique na coluna "Palavra" para ordenar a lista alfabeticamente pelo nome da palavra.

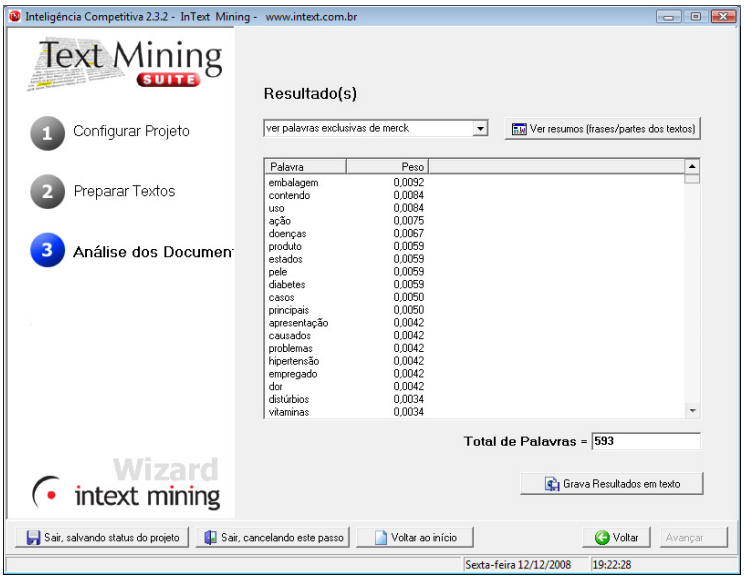
Clique na coluna "Peso" para ordenar a lista pelo peso das palavras de forma decrescente.

b) Palavras exclusivas de cada texto

Selecione a opção "ver palavras exclusivas de ..." para ver as palavras que aparecem somente no texto indicado.

Clique na coluna "Palavra" para ordenar a lista alfabeticamente pelo nome da palavra.

Clique na coluna "Peso" para ordenar a lista pelo peso das palavras de forma decrescente.



É possível gravar num arquivo texto a lista de palavras e pesos que está sendo apresentada na caixa. Para tanto, selecione o botão "Grava resultados em texto".

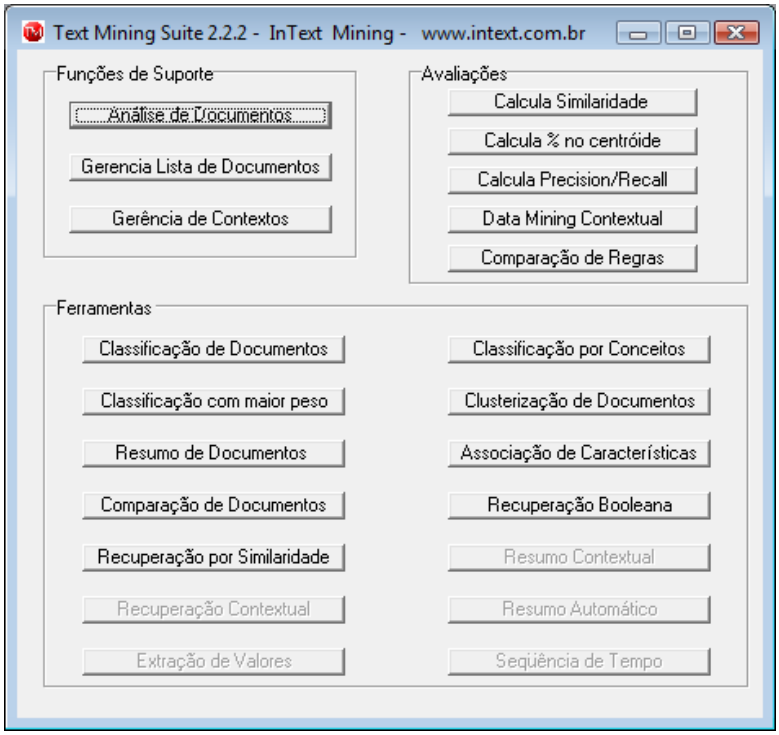
7.3.1 Botão "Ver resumos (frases/partes dos textos)

Esta ferramenta comporta-se como descrito na seção 12.

8 Modo Clássico

Este módulo está dividido em 3 conjuntos de ferramentas:

- as funções de suporte;
- as ferramentas de Text Mining (funções de análise);
- avaliações.



8.1 Visão geral do processo de descoberta

Primeiro, é necessário fazer o **pré-processamento** dos textos, para que sejam criadas representações internas, sobre as quais trabalharão as ferramentas.

Depois, deve-se **criar uma lista de documentos**. Podem ser criadas várias listas sobre a mesma coleção de textos.

Após estes 2 passos iniciais, as ferramentas de Mineração poderão ser utilizadas.

Todas as ferramentas de Mineração exigem uma lista de documentos como entrada e também permitem que os resultados (quando forem uma lista de documentos) possam ser gravados como "ldoc" (listas de documentos), para serem utilizados como entrada em outras ferramentas.

8.2 Análise de Conteúdos de Arquivos Textos

Em todas as ferramentas, nas caixas onde aparecem os nomes dos arquivos textos (lista de documentos), ao se dar 2 cliques no nome do documento ou texto, uma caixa é aberta mostrando o conteúdo do arquivo texto.

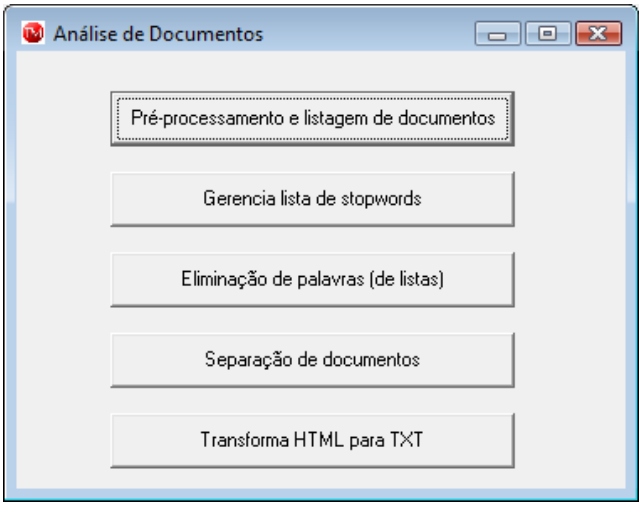
9 Pré-Processamento (Funções de Suporte)

Os documentos textuais (cada arquivo da coleção) devem ser preparados para as análises. Os textos serão analisados e representados num formato próprio para tratamento pelo software. Para cada arquivo texto, serão criados dois arquivos internos, com o mesmo nome mas com as extensões .DAT e .LPL. Este último formato (lista de palavras) é o formato padrão dos arquivos a serem usados pelo software. O formato .LPL é uma lista de palavras, cada uma com um valor numérico associado. No caso dos textos, este valor (ou peso) é a freqüência relativa da palavra no texto (número de aparições dividido pelo total de palavras no texto). Na representação interna dos textos, não serão incluídas as chamadas *stopwords*, que são palavras muito comuns ou genéricas, tais como preposições, artigos e conjunções. A lista de *stopwords* deverá ser definida pelo usuário. Junto com o software, há 3 listas pré-definidas, uma para português (stopword.stw), uma para inglês (stw-inglês.stw) e uma para espanhol (stw-espanhol.stw). Estas listas devem ser revisadas pelo usuário e podem conter qualquer tipo de palavra ou termo (mesmo numerais).

No caso de outros tipos de arquivos, o valor numérico associado à palavra indica o grau de importância da palavra ou do conceito. Por exemplo, se o arquivo LPL for o centróide (lista das palavras mais comuns) de uma coleção, o valor indica a média dos pesos da palavra nos textos onde aparece ou o número de textos onde ela aparece (depende do método usado para cálculo do centróide). No caso dos contextos ou conceitos, o valor numérico indica uma espécie de probabilidade de a palavra aparecer em textos que contenham o referido contexto ou conceito. Se o arquivo LPL contiver conceitos ao invés de palavras (por exemplo, no cálculo do centróide pelo método contextual ou depois da identificação de conceitos nos textos), então o valor numérico associado a cada conceito indica o grau de presença do conceito (no centróide ou no texto).

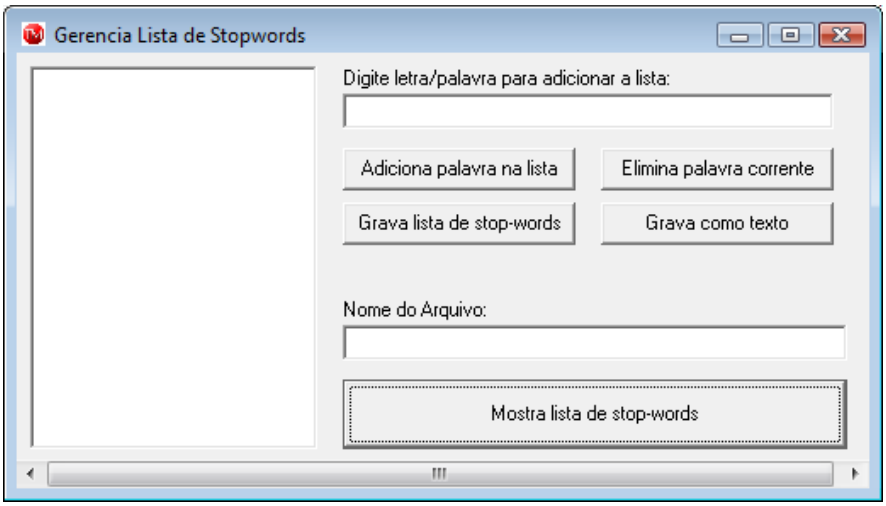
9.1 Preparando a coleção textual (Módulo Análise de Documentos)

Os documentos textuais devem ser analisados e representados num formato interno do software. Serão retiradas as stopwords. Há uma função extra para separar textos que estejam em um único arquivo.



9.1.1 Gerencia lista de stopwords

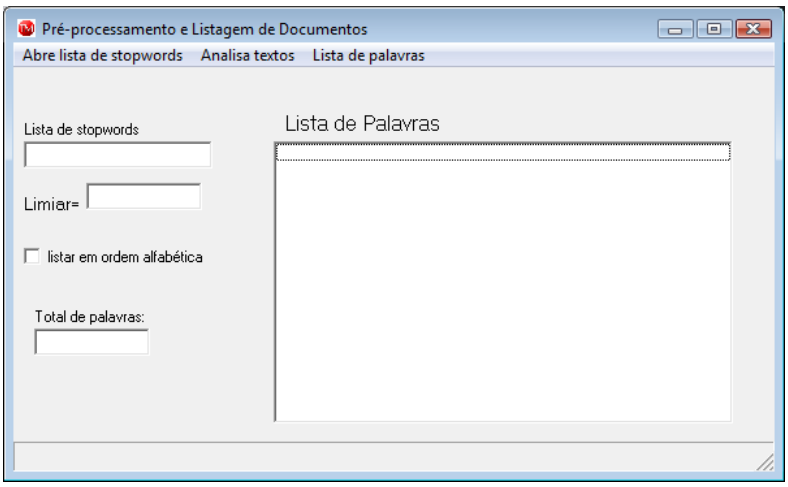
Se não há uma lista de *stopwords* já criada, deve-se criar uma. Para incluir uma palavra na lista, colocar a palavra no campo abaixo do botão "adiciona palavra na lista" e clicá-lo. Não esquecer de gravar a lista. ⓘ



9.1.2 Pré-processamento e listagem de documentos

Primeiro, selecionar uma lista de stopwords (já deve existir ou ter sido criada previamente na função anteriormente explicada). Clicar no botão "analisa textos" e selecionar os textos que serão analisados (utilizar multi-seleção). Para permitir a análise de mais de um texto ao mesmo tempo, deve-se selecionar os vários textos acionando a tecla CTRL junto com o botão do mouse ou clicar no 1º texto da lista e depois clicar no último segurando a tecla SHIFT.

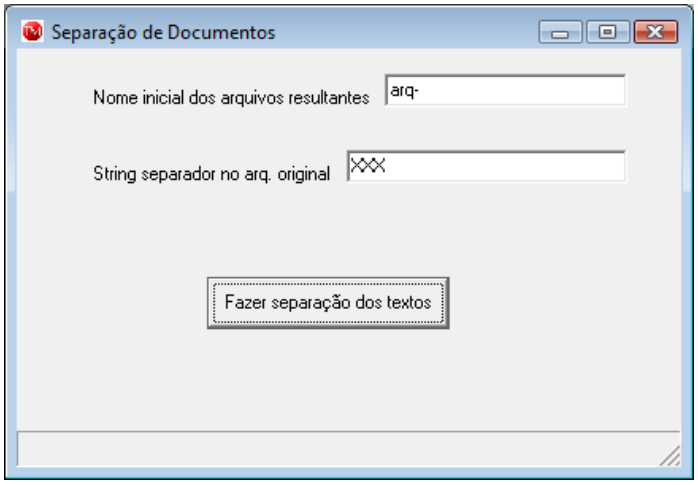
ⓘ Ao final do processo, a caixa "lista de palavras" continuará vazia. Uma mensagem será apresentada na parte de baixa da tela informando que o software terminou de analisar todos os documentos.



9.1.3 Função extra: separação de documentos

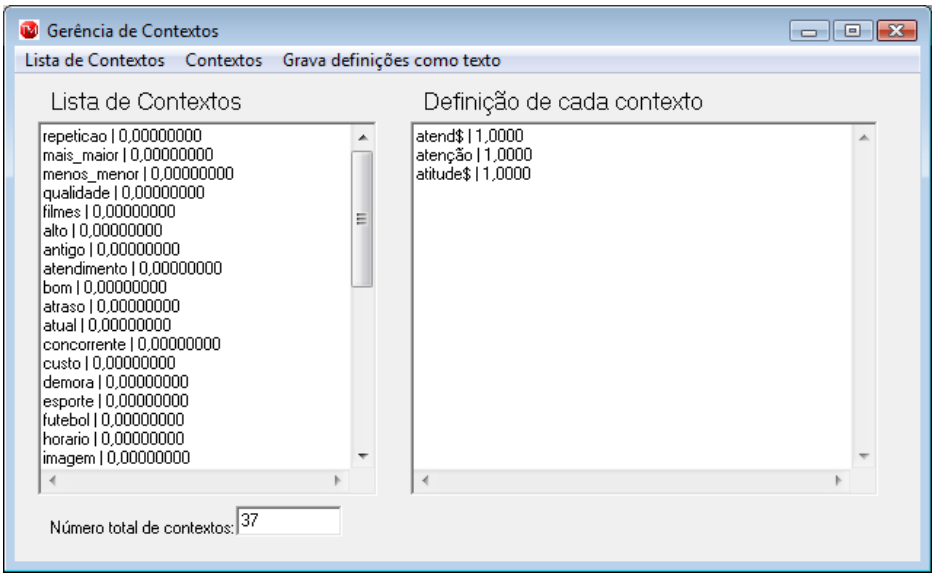
Se todos os textos estão em um único arquivo texto, há uma função para separá-los. Isto permitirá analisar cada texto como independente dos demais. O resultado desta função é que serão criados arquivos-texto independentes para cada texto. O nome a ser dado a cada novo arquivo gerado é uma combinação de um nome comum a todos, fornecido pelo usuário no primeiro campo, e um número inteiro a ser designado pelo software (como um contador). Para que esta função possa ser usada, deve existir entre os textos uma mesma seqüência de caracteres (string separador), a qual deve ser fornecida pelo usuário no segundo campo.

ⓘ O string separador considera maiúsculas e minúsculas.



9.2 Definindo os conceitos ou contextos (Módulo Gerência de Contextos)

Antes de fazer a classificação (identificação dos conceitos nos textos), o usuário deve definir que conceitos deverão ser analisados e como.



9.2.1 Criando e gravando a lista de conceitos

Na caixa à esquerda, o usuário deve fornecer a lista de conceitos (ou contextos) desejados (um em cada linha), cuidando para não deixar a última linha em branco. O nome do conceito não deve exceder a 15 caracteres e não pode ter caracteres brancos no meio nem hífen (somente o caractere sublinhado é permitido) ⓘ. Depois o usuário deve gravar a lista de contextos (pode dar o nome que desejar).

9.2.2 Definindo cada conceito

O próximo passo é definir cada conceito, ou seja, as regras para identificá-los nos textos. As regras para identificação dos conceitos serão definidas na caixa à direita, uma vez para cada conceito em separado. Há dois métodos de identificação que podem ser usados:

a) método "por todo texto" (probabilístico)

Neste método, o usuário especifica uma lista de palavras que, se encontradas no texto (em qualquer parte), darão um indício da presença do conceito naquele texto. O usuário deve fornecer, em cada linha, uma palavra e seu peso como a seguir

palavras | pesos

O peso deve indicar o grau de indício, ou seja, o quanto a palavra indica a presença do conceito. Este deve ser um valor relativo, ou seja, umas palavras são mais importantes que outras. Portanto, devem ser fornecidos valores entre 0 e 1.

Podem ser usadas também palavras negativas, isto é, que devem diminuir o indício de presença do conceito no texto. Isto é útil para análises de contexto. Por exemplo, se "bola" aparecer no texto, é possível que o conceito "futebol" esteja presente. Entretanto, se "sorvete" aparecer no texto, esta possibilidade deve diminuir. Para determinar palavras negativas, basta ao usuário utilizar pesos negativos.

No processo de classificação, os pesos de todas as palavras encontradas serão avaliados para dar o indício total da presença do conceito no texto, resultando num valor que é o grau de presença do conceito no texto.

b) método "por frase" (determinístico)

Neste segundo método, o usuário deve especificar regras para identificação do conceito nas frases. Cada regra deve ser especificada em uma única linha e será analisada dentro de cada frase e não mais no texto todo (cada frase do texto será analisada em relação a todas as regras definidas). Este método é melhor quando se precisa fazer uma análise de contexto mais detalhada. Por exemplo, se as palavras "bola" e "futebol" aparecerem na mesma frase, pode-se afirmar que o conceito "futebol" está presente.

As regras não podem ter mais que 30 caracteres e devem incluir as palavras positivas, aquelas que obrigatoriamente devem aparecer na frase (avaliação por conjunção = E lógico), e as negativas (se houver), indicadas por um sinal de menos "-" antes da palavra. Neste método, entretanto, as palavras negativas indicam que o conceito não existe. Isto é útil para casos onde o "não" aparece, o que inverte o significado.

Então, para que uma regra seja verdadeira, todas as palavras positivas devem aparecer em alguma frase e nesta mesma frase não deve aparecer nenhuma das palavras negativas definidas na mesma regra. A ordem das palavras não será avaliada, mas elas devem ser fornecidas com no mínimo um espaço entre elas. Neste método, não é necessário fornecer pesos, pois se uma das regras for verdadeira, o conceito será identificado.

No processo de classificação, cada regra verdadeira soma 1 a um contador de presença do conceito no texto.

Em ambos os métodos, o processo de classificação (explicado mais adiante) pode determinar um limiar para decidir se o conceito está ou não presente no texto. Este limiar pode ser maior que zero.

ⓘ Não esquecer de gravar cada conceito após a definição de suas regras. Utilizar o mesmo nome dado ao conceito na lista à esquerda.

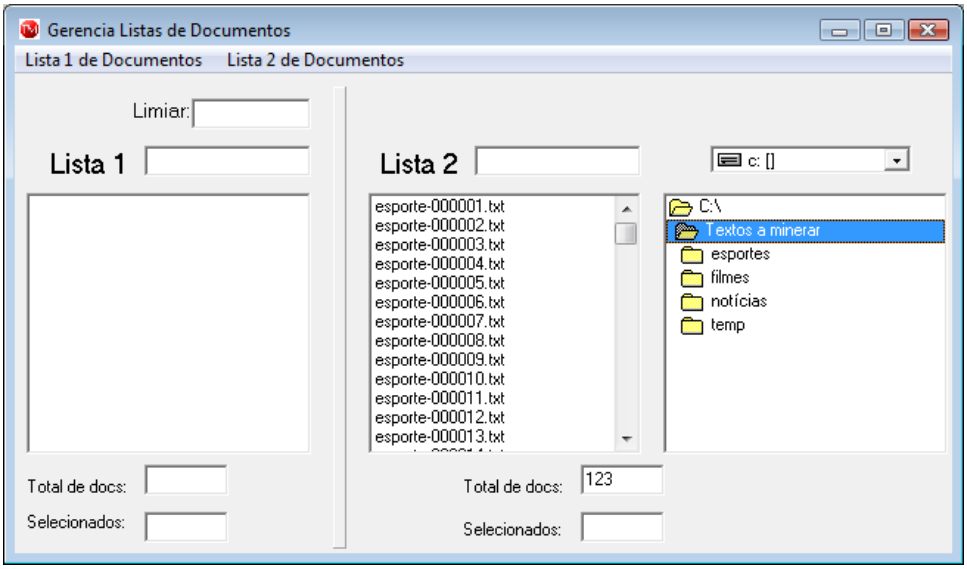
9.2.3 Gravando cada conceito em separado (mesmo nome no arquivo e na lista)

Após a definição das regras dos conceitos (em qualquer um dos dois métodos), é necessário gravá-lo (cada conceito em separado). O usuário deve fornecer como nome do arquivo de conceito exatamente o mesmo nome dado na lista de contexto na caixa à esquerda ⓘ. Os nomes de conceitos não devem ter mais de 15 caracteres, nem espaços ou caracteres especiais ⓘ.

9.3 Trabalhando com listas de documentos (Módulo Gerencia Lista de Documentos)

Os módulos de análise trabalham todos sobre listas de documentos. Isto permite ao usuário especificar somente uma vez que conjunto de documentos irá analisar. Podem ser criadas várias listas sobre a mesma coleção (formando subcoleções) e pode haver documentos comuns a mais de uma lista.

Neste módulo, há espaço para criação ou visualização de duas listas ao mesmo tempo.



9.3.1 Nova lista

A opção “nova lista” limpa a caixa correspondente.

9.3.2 Acrescenta documentos na lista

Esta opção permite acrescentar documentos numa lista em branco ou já listada na caixa. Para acrescentar, marcar o documento quando solicitado. Se o usuário quiser, poderá marcar mais de um usando a tecla CTRL simultaneamente. Não esquecer de gravar a lista depois. ⓘ

9.3.3 Abre lista de documentos

Esta opção limpa a caixa e permite visualizar os documentos de uma lista pré-existente.

9.3.4 Acrescenta lista inteira

Neste caso, os documentos listados na caixa não são apagados e os documentos de uma lista pré-existente são adicionados à caixa.

9.3.5 Elimina doc da lista

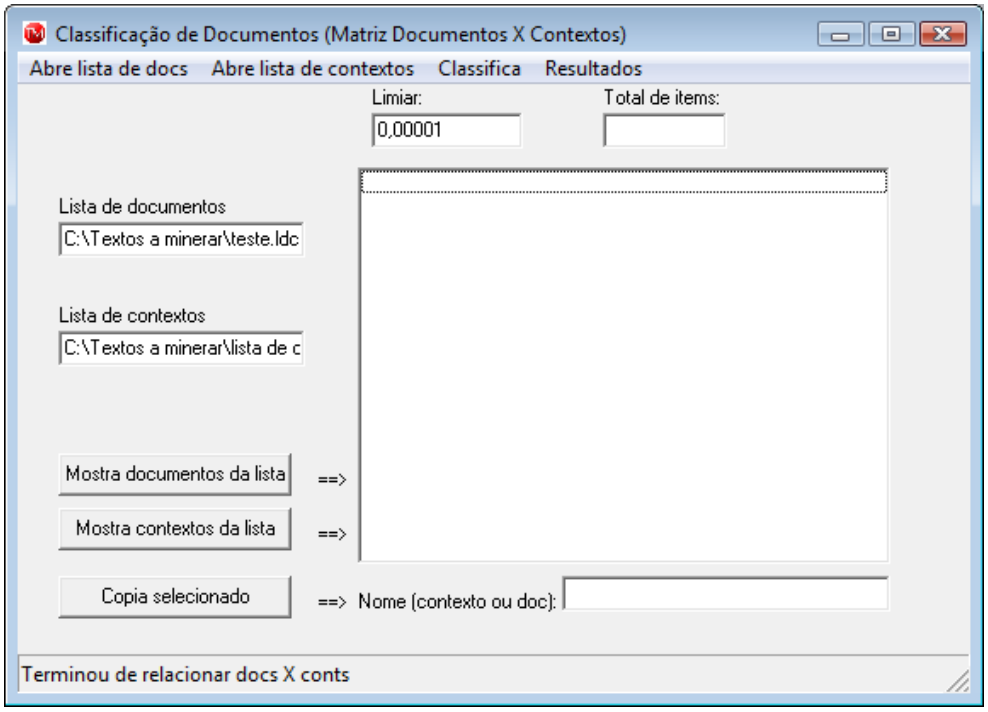
O usuário deve marcar um documento da lista e então acionar esta opção para eliminar o documento da lista.

9.3.6 Gravando listas

Ao final das operações anteriores, o usuário não deve esquecer de gravar a lista ⓘ.

10 Classificação de Documentos

Este módulo de análise permite a identificação dos conceitos previamente definidos nos textos de uma coleção (lista de documentos).



10.1 O processo de classificação

Primeiro o usuário deve selecionar uma lista de documentos e uma lista de contextos nas opções correspondentes. Estas listas já devem ter sido criadas previamente. Depois o usuário deve escolher o método de classificação a ser usado.

10.1.1 Método “todo texto” (probabilístico) X Método “por frase”(determinístico)

Como explicado anteriormente, o método “todo texto” avalia a presença de cada palavra no texto todo, somando o peso correspondente (ser for uma palavra negativa, o peso negativo será somado também). Ao final, a soma total dará o grau de presença do conceito no texto. Um limiar pode ser usado pelo usuário para separar os falsos casos (graus muito baixos não indicam a presença do conceito).

No método “por frase”, as regras definidas serão avaliadas em cada frase do texto. Cada regra verdadeira soma 1 no grau final. Também neste caso, um limiar pode ser usado para filtrar casos indesejados.

O processo de classificação inicia ao se clicar o botão “classifica” e pode demorar algumas horas dependendo do tamanho da coleção e do número e da complexidade dos conceitos definidos ⓘ.

O resultado do processo de classificação é como uma matriz relacionando documentos e conceitos e o grau deste relacionamento, variando de zero ao infinito e indicando o quanto um conceito está presente num documento.

10.1.2 Listando documentos por conceito/contexto – uso de limiar

Para visualizar os documentos classificados em um conceito/contexto, isto é, os documentos que possuem o conceito, colocar o nome do conceito no campo “nome (contexto ou doc)” e acionar o botão “mostra docs de um contexto”. Para agilizar esta tarefa, é possível listar os conceitos com o botão “mostra contextos da lista”, marcar um dos conceitos e clicar o botão “copia selecionado”.

10.1.3 Listando conceitos/contextos de um documento

Para mostrar os conceitos associados a um documento ou identificados nele, o procedimento é semelhante ao anterior: colocar o nome do documento no campo “nome” e acionar o botão “mostra contextos de um doc” ou então listar os documentos com o botão “mostra documentos da lista”, marcar um documento e clicar o botão “copia selecionado”.

10.1.4 Usando limiares

Nas funções anteriores (listar conceitos ou documentos), o usuário pode definir um limiar para apresentação. Este limiar será usado para filtrar documentos ou conceitos com grau acima ou igual. Na listagem de documentos de um conceito, serão mostrados os documentos que se relacionam com o conceito fornecido com grau acima ou igual ao limiar. Na listagem de conceitos de um documento, serão mostrados os conceitos identificados no documento com grau igual ou superior ao limiar.

Se nenhum limiar for especificado, será assumido o valor zero e serão mostrados todos os documentos ou contextos, mesmo com grau zero. Qualquer valor pode ser fornecido como limiar.

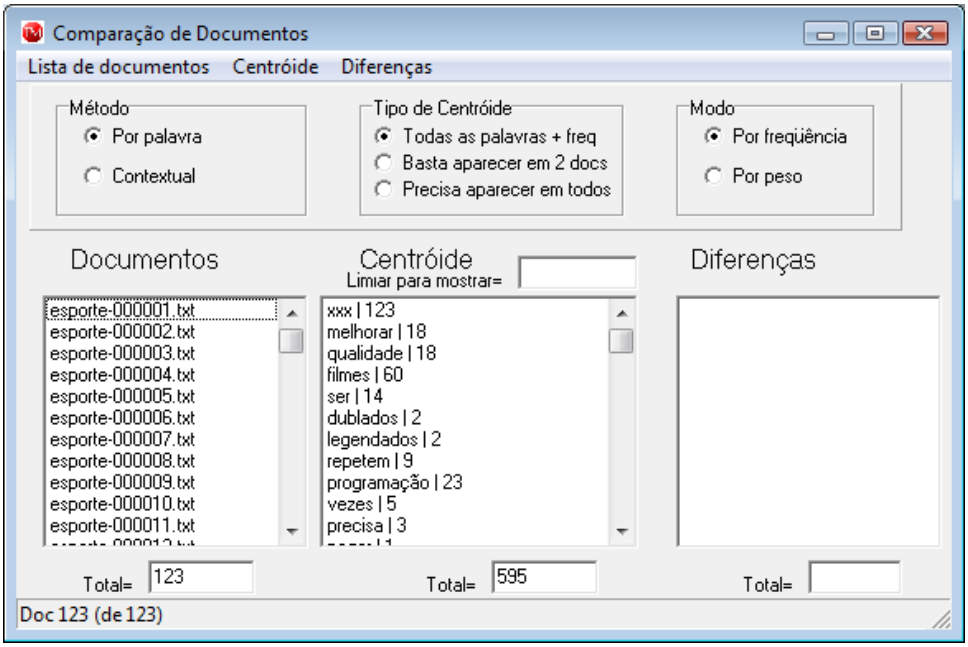
10.1.5 Gravando resultados para análise posterior (subcoleções)

Os documentos listados como resposta nas funções anteriores podem ser gravados como uma lista de documentos, incluindo o grau associado ao documento. Isto permite que sejam criadas subcoleções de acordo com classificações ou conceitos identificados. Estas subcoleções (novas listas de documentos) poderão ser usadas como entrada em qualquer outro módulo.

11 Módulo Comparação de Documentos

O primeiro tipo de análise possível é a identificação dos conceitos mais comuns numa coleção. Outro tipo de análise é a identificação de itens exclusivos (diferenças) a cada documento.

11.1 Identificação do Centróide de uma Coleção



11.1.1 Tipo de método

Na abordagem por conceitos, deve ser usado o método "contextual", isto é, serão analisados os conceitos identificados nos textos e não as palavras.

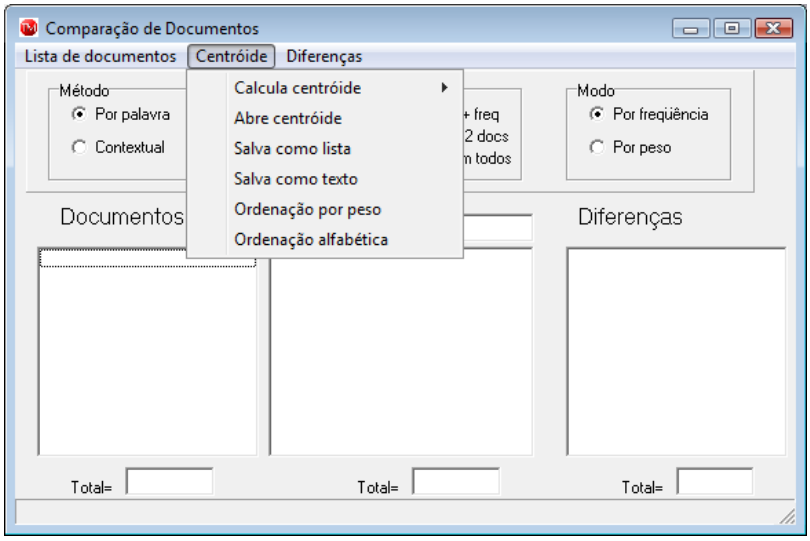
A opção "por palavra", indica que serão analisadas as palavras presentes nos textos.

11.1.2 Tipo de centróide

- "todas as palavras + freq": coloca no centróide todas as palavras presentes em todos os textos; ao lado da palavra, coloca o número de textos onde ela aparece;
- "basta aparecer em 2 docs": coloca no centróide somente palavras que apareçam em mais de um texto (dois ou mais); o peso é calculado pela freqüência relativa da palavra em cada texto (uma média);
- "precisa aparecer em todos docs": somente farão parte do centróide as palavras que aparecem em todos os textos da coleção.

11.1.3 Modo de cálculo

Deve ser selecionado o modo de cálculo "por freqüência" para a 1ª opção acima (tipo de centróide). Para as demais, utilizar o modo "por peso".



11.1.4 Calculando o centróide

O usuário deve selecionar uma lista de documentos e acionar a opção "calcula centróide (novo)".

11.1.5 Gravando centróide (mesmo formato de textos)

O centróide resultante pode ser gravado pelo usuário para análise posterior. Como o formato é o mesmo da representação interna dos textos (LPL = lista de palavras), o centróide pode ser visualizado no módulo "análise de documentos", função "pré-processamento e listagem de documentos", bastando selecionar um centróide ao invés de um documento.

11.1.6 Centróide de centróide (por classe)

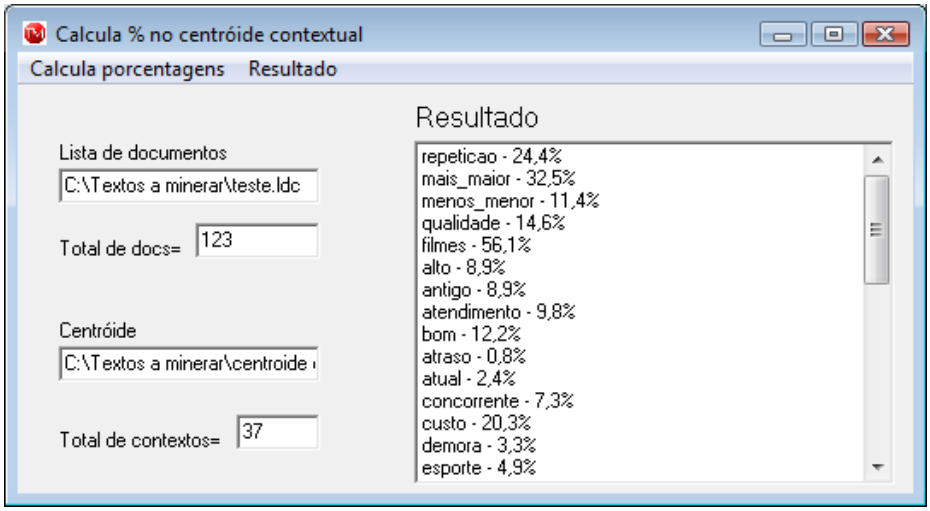
Por usar o mesmo formato interno de documentos, pode-se extrair um centróide de centróides, por exemplo, quando um centróide identifica uma subcoleção.

11.1.7 Centróide Percentual (Módulo Calcula % no centróide - Avaliações)

O resultado do cálculo do centróide indica o número absoluto de documentos onde cada conceito aparece. Para apresentar os resultados de forma percentual, pode-se usar o módulo "calcula % no centróide" entre as funções de avaliação na tela inicial.

Neste caso, ao acionar a opção "calcula porcentagens", será solicitado ao usuário para indicar a lista de documentos usada e o centróide resultante (arquivo gravado anteriormente, calculado pelo método "contextual").

O resultado é a lista de conceitos e o % de textos onde aparece cada um.

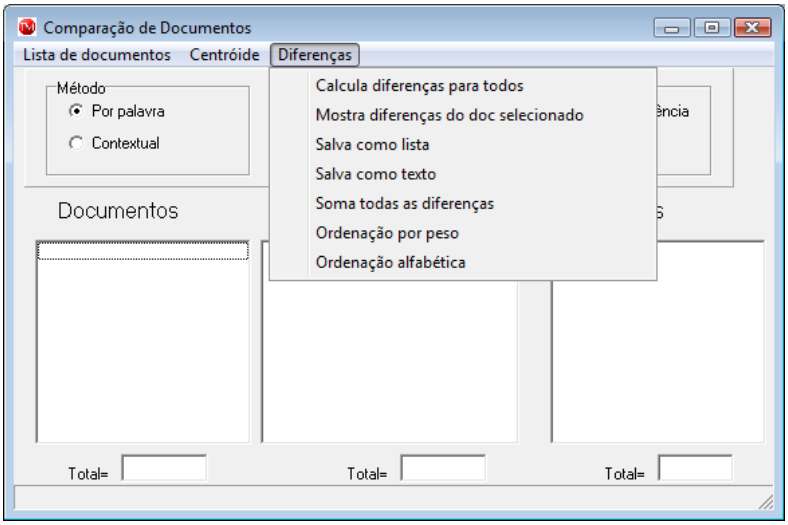


11.2 Análise de diferenças

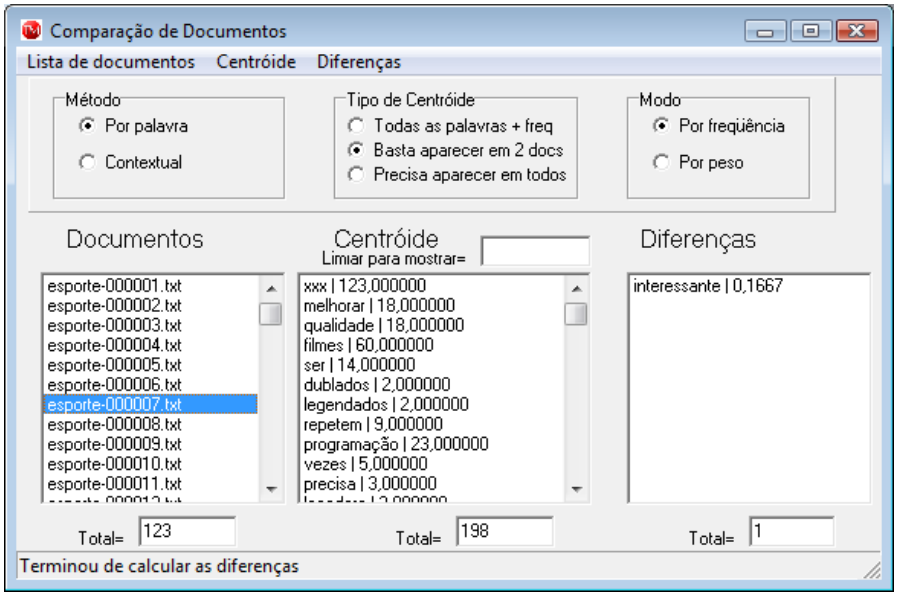
Após o cálculo do centróide, que resulta nos conceitos ou palavras comuns a uma coleção, pode-se calcular as diferenças, isto é, identificar os conceitos ou palavras que aparecem em somente um documento.

Na mesma tela, acionar a opção "calcula diferenças de todos docs". Será solicitado que o usuário selecione o arquivo centróide gravado (deve ter sido calculado antes). As diferenças serão armazenadas no formato DIF.

ⓘ As diferenças só podem ser calculadas se o centróide tiver sido gerado com o tipo de centróide "Basta aparecer em 2 docs".



Para mostrar as diferenças de cada documento, basta selecionar o documento desejado na lista que aparece na caixa à esquerda e acionar a opção "mostra diferenças do doc selecionado". Com as diferenças mostradas na caixa à direita, é possível gravá-las no formato LPL, como o centróide.

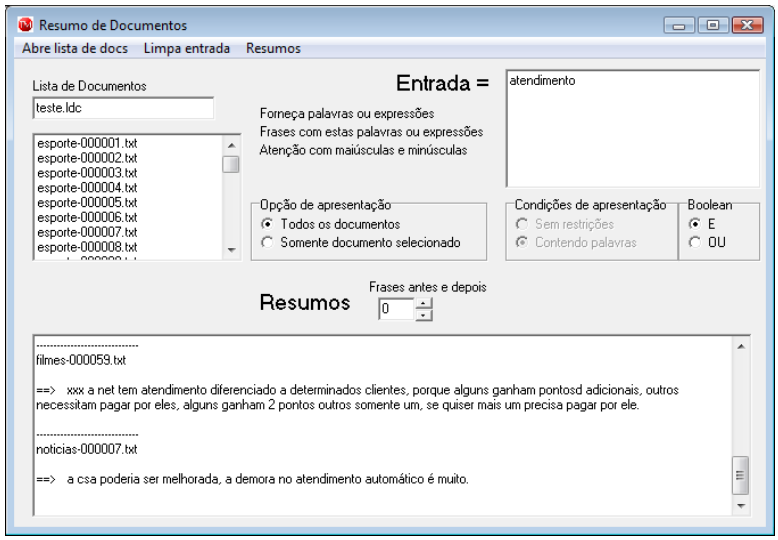


Assim como é possível calcular centróide de centróides, é possível identificar diferenças entre centróides. Isto será útil para encontrar características únicas por classe ou sub-coleção.

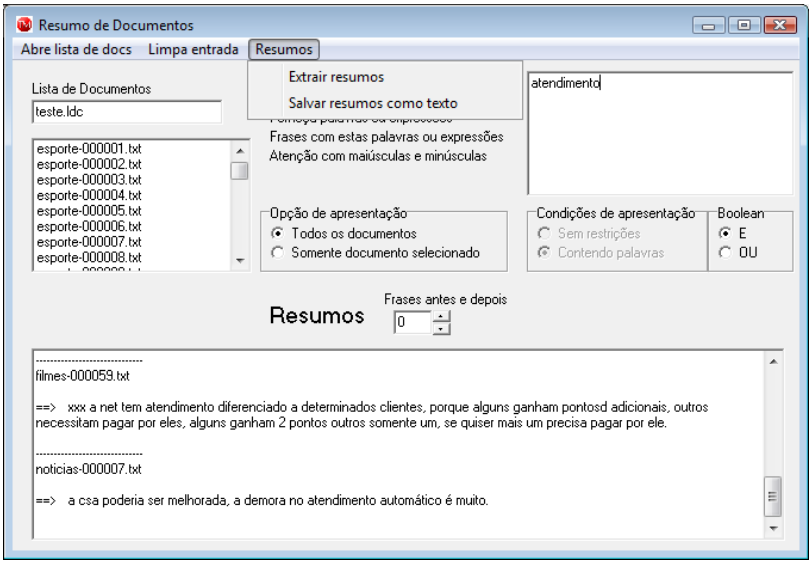
12 Módulo de Extração de Resumos (extração de frases contendo certas palavras)

Esta ferramenta extrai frases segundo condições estabelecidas pelo usuário. As condições indicam as palavras que devem aparecer nas frases, para que estas façam parte do resumo.
No resultado, serão apresentadas as frases que satisfazem as condições, sendo indicados também os nomes dos arquivos onde elas aparecem.
As palavras a serem analisadas nas frases devem ser fornecidas na caixa "Entrada".
É possível fornecer várias palavras (uma em cada linha na caixa de Entrada). Utilize a opção E ou OU (boolean) para encontrar frases com todas as palavras ou apenas uma delas.

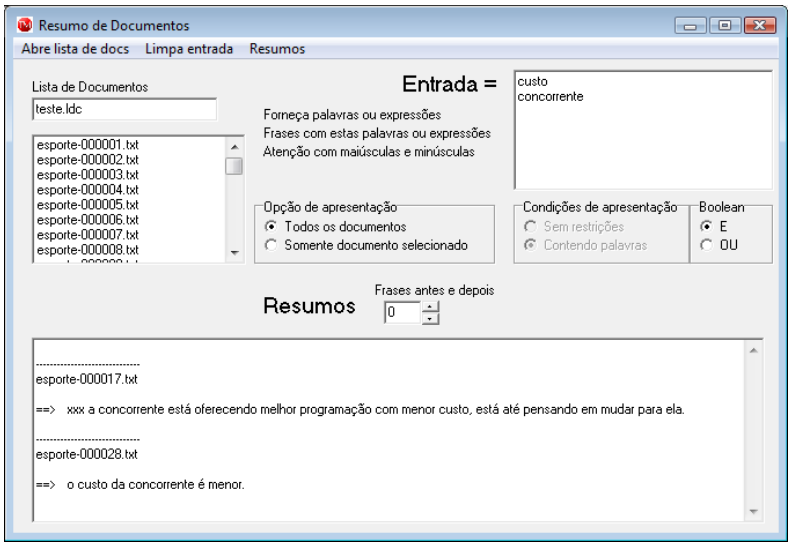
A opção de "frases antes/depois" permite escolher quantas frases (antes e depois) da frase encontrada serão apresentadas junto com a frase encontrada. Isto permite ver o contexto da frase.



Clique em "Resumos" para escolher as opções de extrair os resumos ou de salvar o resultado.



Se desejar, utilize mais de uma palavra na caixa de Entrada, mas cada uma numa linha diferente. Se quiser a palavra exata (e não como prefixo), coloque um espaço após a palavra. Não esqueça de escolher o operador booleano (E ou OU) para indicar se deseja frases onde as palavras aparecem todas (operador E) ou apenas uma delas (OU).



13 Análise e Mineração de Associações

Como explicado no início deste manual, um dos tipos de análise possível sobre os conceitos é a associação. O resultado é apresentado na forma de regras "X ➔ Y", onde X e Y são conceitos e a implicação significa que "quando X aparece em um documento, Y também aparece com um certo grau de confiança e suporte" (com explicado em 3.2).

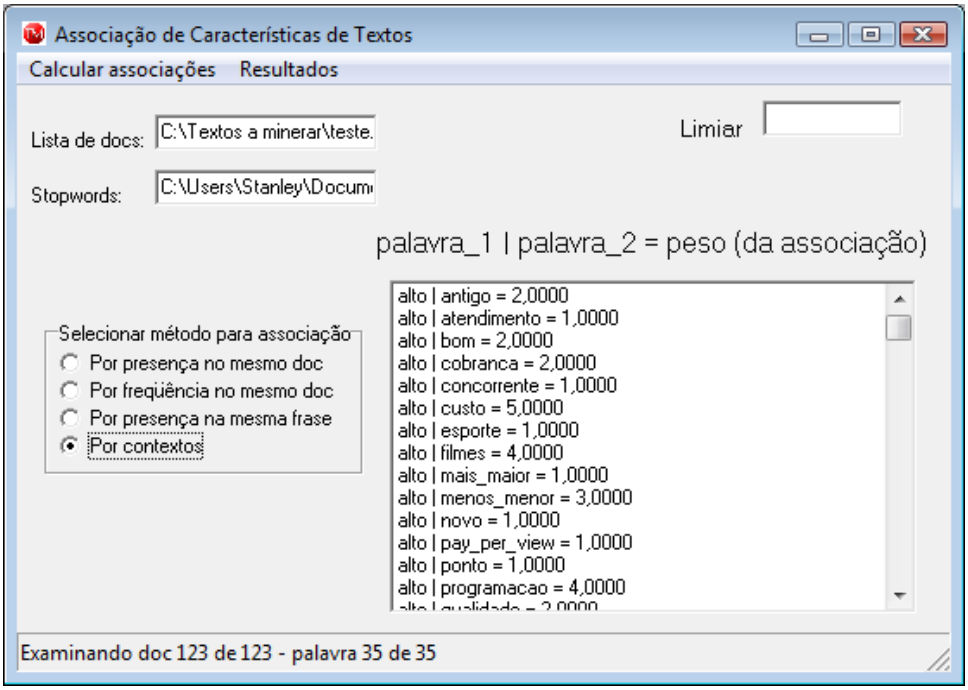
13.1 Calculando associações (Módulo Associação de termos)

As associações são calculadas no módulo "associação de termos" entre as ferramentas de Text Mining que aparecem na tela inicial (funções de análise).

Para tanto, o usuário deve selecionar o método contextual ("por contextos"). Após acionar a opção "calcular associação entre termos", será solicitado que o usuário selecione uma lista de stopwords (não será usada neste tipo de abordagem) e uma lista de documentos.

O resultado são regras do tipo "X | Y = Z", onde X e Y são palavras ou conceitos, e Z corresponde ao número de documentos onde X e Y aparecem juntos.

ⓘ O processo com qualquer um dos 3 primeiros métodos é bastante demorado.



13.2 Processo de Mineração (Módulo Data Mining Contextual - Avaliações)

O resultado da função anterior (não esquecer de gravar o resultado no passo anterior ⓘ), ainda não avalia os graus de suporte e confiança, é necessário utilizar o módulo "data mining contextual", presente entre as funções de avaliação da tela inicial.

Primeiro o usuário deve selecionar as listas de contextos e de documentos e selecionar o arquivo gravado com as associações (resultado da função anterior). Depois, basta acionar a opção "calcula confiança/suporte".

Os resultados poderão ser filtrados por valores de suporte ou de confiança, selecionando o tipo de limiar e fornecendo um valor no campo "limiar para mostrar".

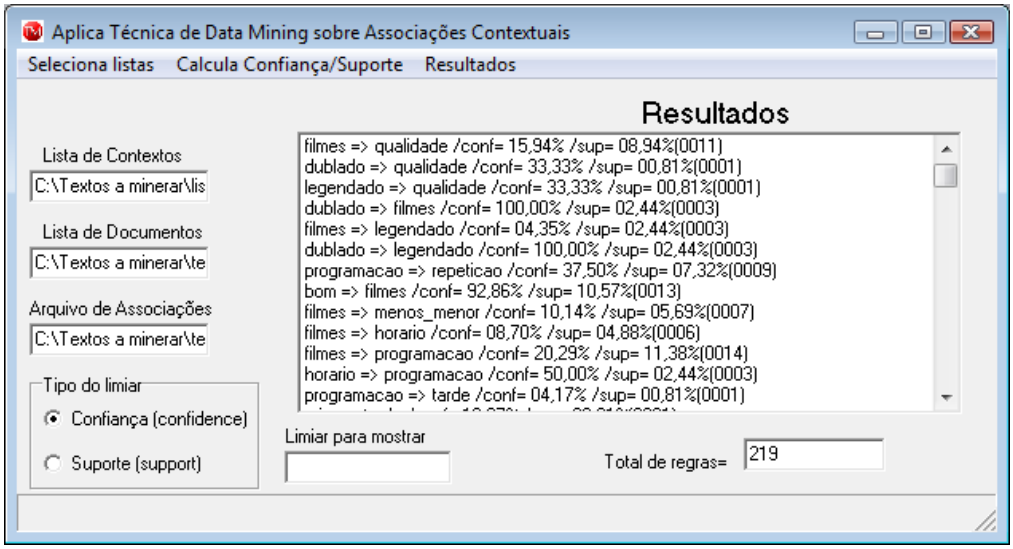
O resultado é composto por uma lista de regras de associação (do tipo Se-Então). A interpretação da regra deve ser feita da seguinte forma:

Tomando como exemplo a regra abaixo:

filmes => qualidade /conf= 15,94% /sup= 08,94%(0011)

Nota-se que a confiança (probabilidade condicional) é de 15,94% e o suporte (número de textos onde a regra aparece) é igual a 8,94% do total (11 textos em toda a coleção).

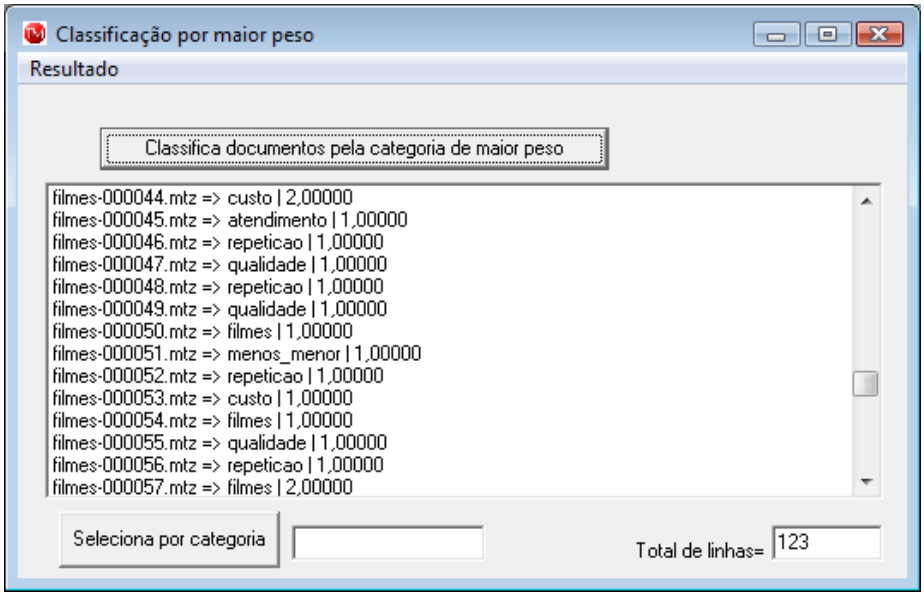
Interpretação: "15,94% dos textos que falam sobre filmes também falam em qualidade".



14 Classificação com Maior Peso

Esta ferramenta permite encontra a classe (conceito, tema ou assunto) com mais peso associada a um documento. Isto é útil porque o software pode encontrar vários conceitos em um documento. Entretanto, em alguns casos, é necessário ficar com o conceito de maior peso, como por exemplo, quando se quer saber o assunto principal de um documento.

Esta ferramenta mostra cada documento da coleção e o conceito de maior peso associado a cada documento.



15 Recuperação por Similaridade

Esta ferramenta permite encontrar textos similares a um texto escolhido ou mesmo em relação a uma lista de documentos.

Isto funciona como um método k-NN (dos Vizinhos Mais Próximos).

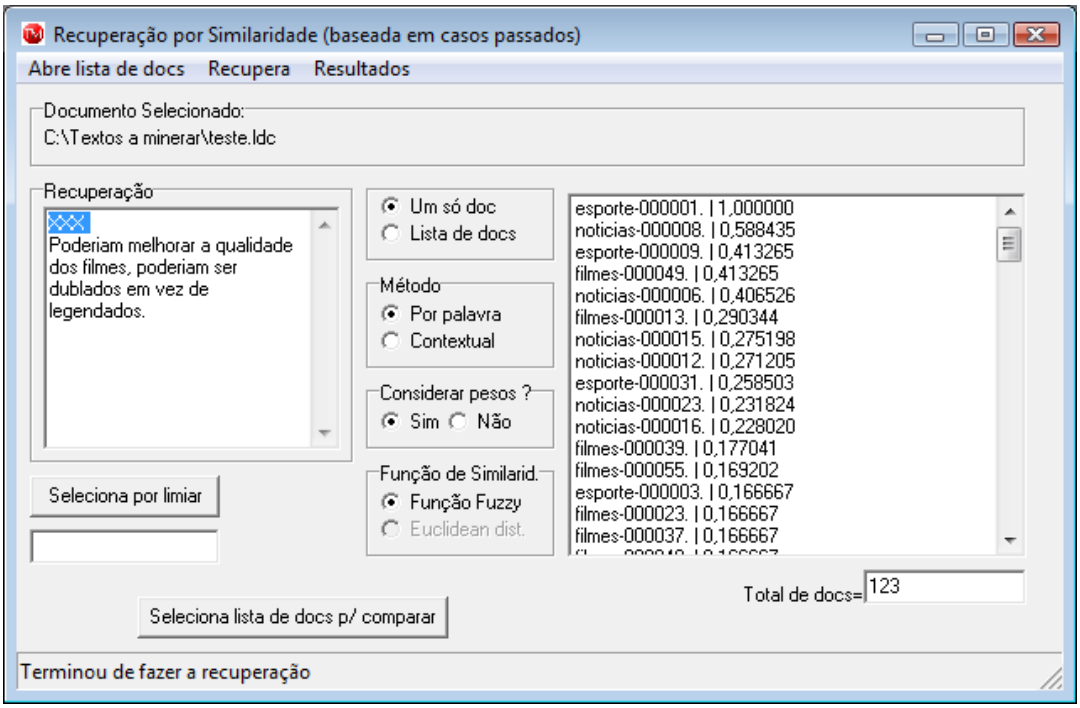
Utilize os métodos “um só doc”, “por palavra”, “sim (considerar pesos)” e “função Fuzzy”.

Primeiro, é necessário abrir uma lista de documentos (os que serão avaliados).

Ao clicar em “Recupera”, o software solicita o documento parâmetro a ser comparado com todos os outros.

Na caixa à esquerda, o software mostra um ranking dos documentos da coleção por similaridade ao documento parâmetro. Note que o próprio documento é similar a ele mesmo com grau igual a 1.

O valor de similaridade é um número entre 0 e 1 indicando o quanto os textos são similares. Textos iguais terão similaridade igual a 1 e textos que não compartilham nenhuma palavra em comum terão similaridade Zero.



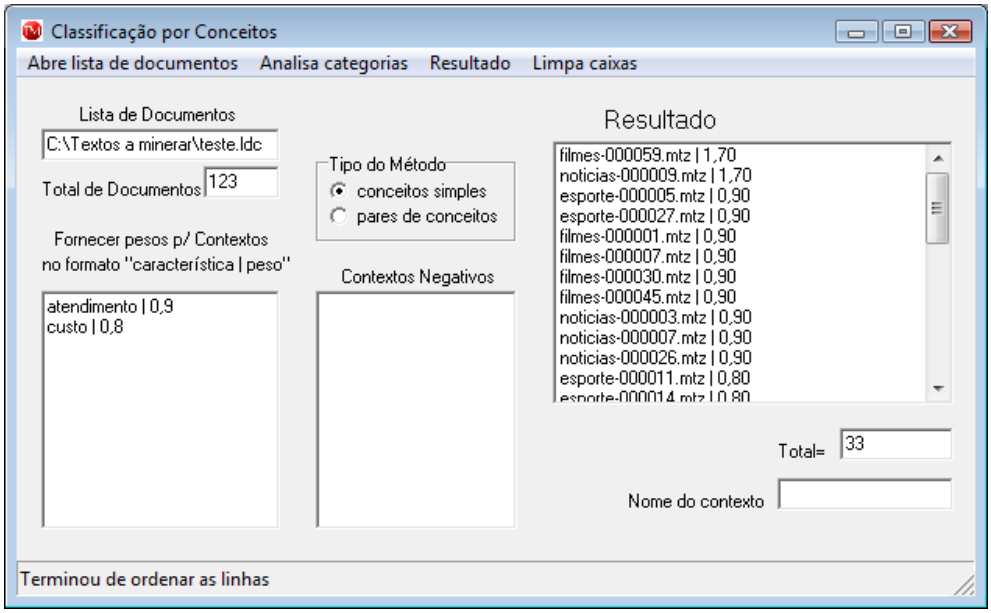
16 Classificação por Conceitos

Esta ferramenta permite encontrar textos segundo um perfil de conceitos presentes. Ela é útil principalmente para análise de currículos vitae de pessoas.

Após selecionar a lista de documentos a ser analisada, indique na caixa à esquerda os conceitos procurados e utilize um peso para cada conceito (separando com o caracter barra vertical |) .

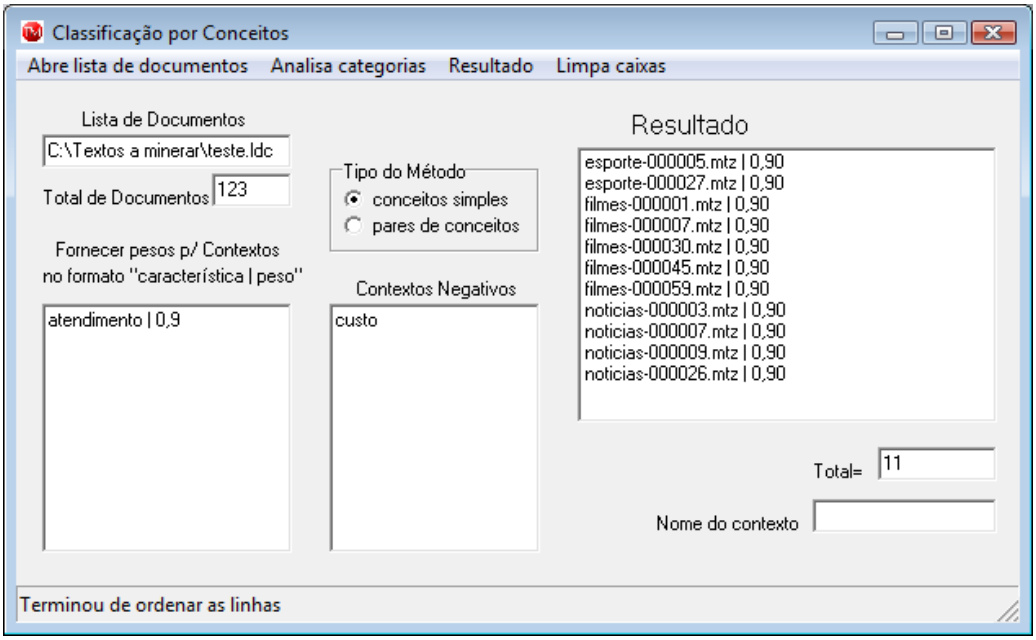
A ferramenta irá montar um ranking dos documentos que contenham estes conceitos. Lembrando que cada conceito presente num texto tem um peso associado, indicando o grau de importância do conceito no texto. Ficarão no topo do ranking os textos que tiverem os conceitos fornecidos na entrada e os que tiverem estes conceitos com mais peso de importância.

O ranking resultante ainda indica o grau com que cada texto satisfaz o perfil desejado.



Poderão ser utilizados conceitos negativos (informados na 2ª caixa). Isto significa que farão parte do resultado apenas os textos que satisfizerem o perfil de conceitos positivos e que não tiverem os conceitos informados como negativos.

No exemplo abaixo, os textos resultantes devem falar em “atendimento” e não devem ter o conceito “custo”.

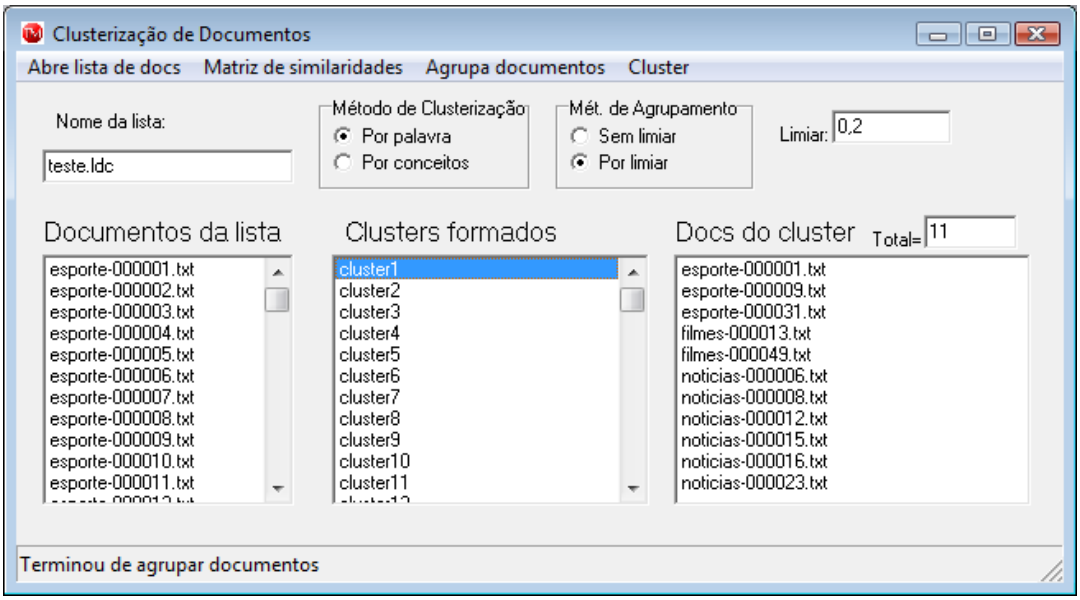

Classificação por Conceitos
Abre lista de documentos
Analisa categorias
Resultado
Limpa caixas
Lista de Documentos
C:\Textos a minerar\teste.ldc
Total de Documentos
123
Fornecer pesos p/ Contextos
no formato "característica | peso"
atendimento | 0,9
Tipo do Método
conceitos simples
pares de conceitos
Contextos Negativos
custo
Resultado
esporte-000005.mtz | 0,90
esporte-000027.mtz | 0,90
filmes-000001.mtz | 0,90
filmes-000007.mtz | 0,90
filmes-000030.mtz | 0,90
filmes-000045.mtz | 0,90
filmes-000059.mtz | 0,90
noticias-000003.mtz | 0,90
noticias-000007.mtz | 0,90
noticias-000009.mtz | 0,90
noticias-000026.mtz | 0,90
Total=
11
Nome do contexto
Terminou de ordenar as linhas

17 Clusterização de Documentos (Agrupamento)

Esta ferramenta agrupa os documentos por similaridade. Ela cria grupos contendo documentos similares e colocar documentos que não são similares em grupos diferentes. Os Grupos também são chamados Clusters.

Primeiro é preciso informar a lista de documentos. Depois criar a matriz de similaridades.

Para agrupar os documentos (opção "agrupa documentos"), é necessário informa um limiar. Este limiar é um valor mínimo de similaridade para que os elementos sejam agrupados, ou seja, só estarão no mesmo cluster documentos que tenham similaridade acima deste valor.



O algoritmo de clustering utilizado é o Star. Ele toma um elemento ainda não agrupado e cria um novo cluster. Então coloca neste cluster todos os elementos similares ao primeiro mas com grau de similaridade superior ao valor do limiar.

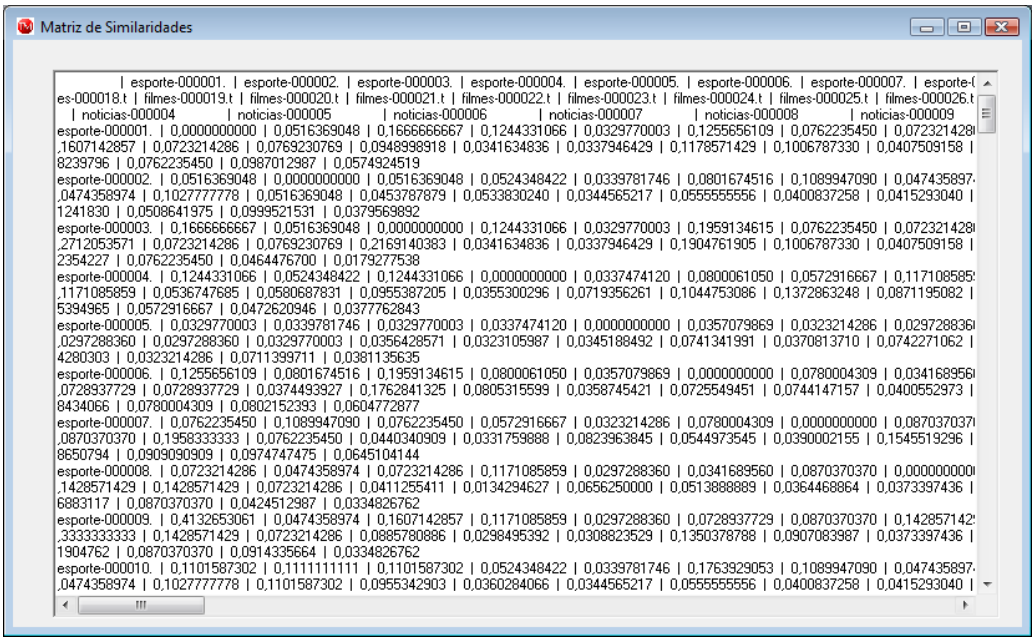
Pode-se utilizar o método "por palavra" (avalia a similaridade entre os textos pelas palavras em comum) ou o método "por conceitos" (avalia a similaridade entre os textos pelos conceitos em comum. No 2º caso, é necessário fazer antes a classificação dos textos (e a geração da Ontologia com conceitos e suas regras).

Para ver os documentos alocados em cada cluster, selecione um cluster na lista de "clusters formados" e selecione a opção "mostra docs de um cluster" no menu "Cluster".

ⓘ Como cada cluster resultante é formado por uma lista de documentos, você pode gravar cada lista em separado, representando cada cluster. Depois poderá utilizar a ferramenta de Comparação de textos para encontrar o centróide de cada cluster. Poderá também fazer o centróide dos centróides e encontrar as diferenças entre cada centróide; isto permitirá saber o que há de comum entre os elementos de cada cluster e também o que há de exclusivo (diferenças) em cada cluster.

Para determinar o valor do limiar, utilize a opção “mostra matriz” (no menu “Matriz de Similaridades”) para ver a matriz e os graus de similaridade calculados. A sugestão é utilizar um grau de similaridade médio (entre o maior e o menor graus apresentados na matriz).

Abaixo segue um exemplo de matriz de similaridades.

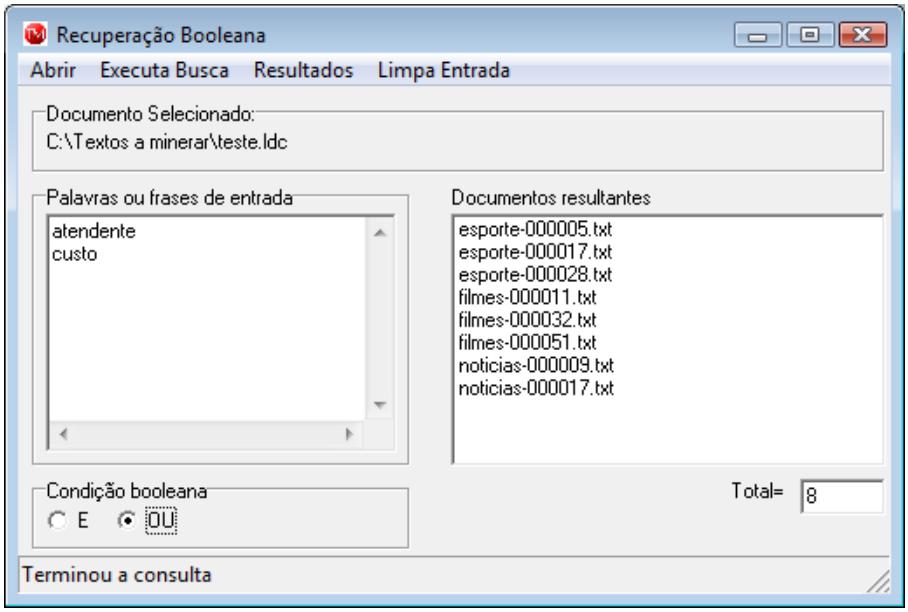


18 Recuperação Booleana

Esta ferramenta implementa a busca de textos por presença de palavras. Após escolher a lista de documentos, forneça palavras a serem procuradas (formando uma consulta, que pode ser gravada e posteriormente recuperada).

Ao escolher a opção "executa busca", a ferramenta procura textos que tenham as palavras fornecidas de acordo com a opção Booleana (E ou OU).

ⓘ Atenção: a busca é por palavras exatas, ou seja, atente para plural/singular e acentos.

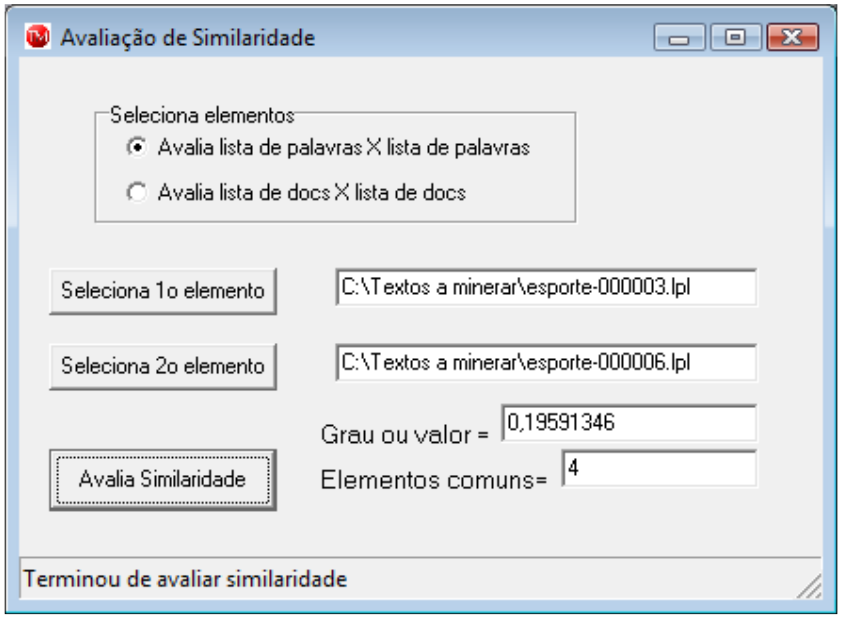


19 Similaridade entre Textos

Esta ferramenta avalia a similaridade entre 2 documentos ou entre duas listas de documentos.

O grau de similaridade resultante é um valor numérico relativo, entre 0 e 1. Se os documentos foram idênticos, o grau de similaridade será 1.

Também são informados os elementos (palavras) em comum.



20 TROUBLESHOOTING (Resolução de Problemas)

- **Em operações demoradas, o TMS parece travar.**

Não atualiza a mensagem embaixo da tela com o número de textos ou conceitos analisados e o no título da janela aparece o texto "não está respondendo".
➔ Na maioria das vezes, isto não é um problema, mas apenas uma incompatibilidade com o sistema operacional. O TMS deve estar executando ainda, porém algumas operações são demoradas mesmo (por exemplo, associações entre palavras).
Para ter certeza que o TMS está funcionando, verifique se o led (luz) do HD (disco rígido) está piscado. Se a luz não estiver piscando, é provável que o TMS tenha parado mesmo.
Além disto, verifique o tamanho dos arquivos no diretório de textos. Se alguns arquivos estiverem mudando de tamanho (aumentando), o TMS está executando perfeitamente. Os arquivos utilizados pelo TMS internamente são os com as seguintes extensões: .ass (para associações), .lpl (para preparação de textos e cálculo de centróide), .mtz (para classificação), .dif (para cálculo de diferenças entre documentos).

- **FILE NOT FOUND (Arquivo não encontrado).**

➔ Verifique se o arquivo de stopwords (.stw) está no mesmo diretório onde está o arquivo executável do TMS (Text Mining Suíte v.X.X.X.exe).
➔ O TMS pode não ter encontrado algum arquivo texto. Verifique as restrições quanto ao nome de arquivos texto (capítulo 4 deste manual). Procure colocar todos os arquivos texto num diretório, assim como os arquivos de resultados e listas de documentos.

- **Separação de Documentos**

se for utilizado o Wizard e vários textos estiverem dentro de um mesmo arquivo, o Wizard poderá separar os textos, mas só poderá haver um ÚNICO arquivo texto no diretório.

- **Mensagem do tipo "fazer preparação dos docs antes".**

➔ Pode ser que o TMS não tenha encontrado algum texto; pode ser causa do nome do arquivo texto (tamanho muito grande ou uso de caracteres especiais no nome). Verifique as restrições no capítulo 4 deste manual.
➔ Pode ser que os arquivos textos ainda não tenham sido preparados e o TMS não encontrou os arquivos com extensão .lpl (representação interna dos arquivos texto). Ver seção 8.1 deste manual.

- **Problemas com Windows XP**

No Windows XP, quando se está tentado minerar muitos arquivos (por exemplo, mais de mil textos), pode ocorrer problemas na preparação dos textos. Ao serem selecionados todos ou vários arquivos no diretório, o TMS não faz a análise de todos os documentos selecionados.
➔ Este é um problema do Sistema Operacional Windows XP (no Windows Vista, este problema não ocorre). A solução é fazer a preparação dos documentos por partes, ou seja, selecionando sub-grupos de textos para preparação.
Note: isto é feito somente na preparação; para mineração e uso da ferramenta que Gerencia Lista de Documentos, podem ser selecionados todos os arquivos texto e utilizados numa lista só.

- **Lista de Stopwords não encontrada**

Para textos em Português, o software procurará pelo arquivo com nome "stw-portugues.stw"; para textos em Inglês, pelo arquivo "stw-ingles.stw" e para textos em Espanhol, pelo arquivo

“stw-espanhol.stw”. Certifique-se de que os arquivos .stw estejam no mesmo diretório do software.