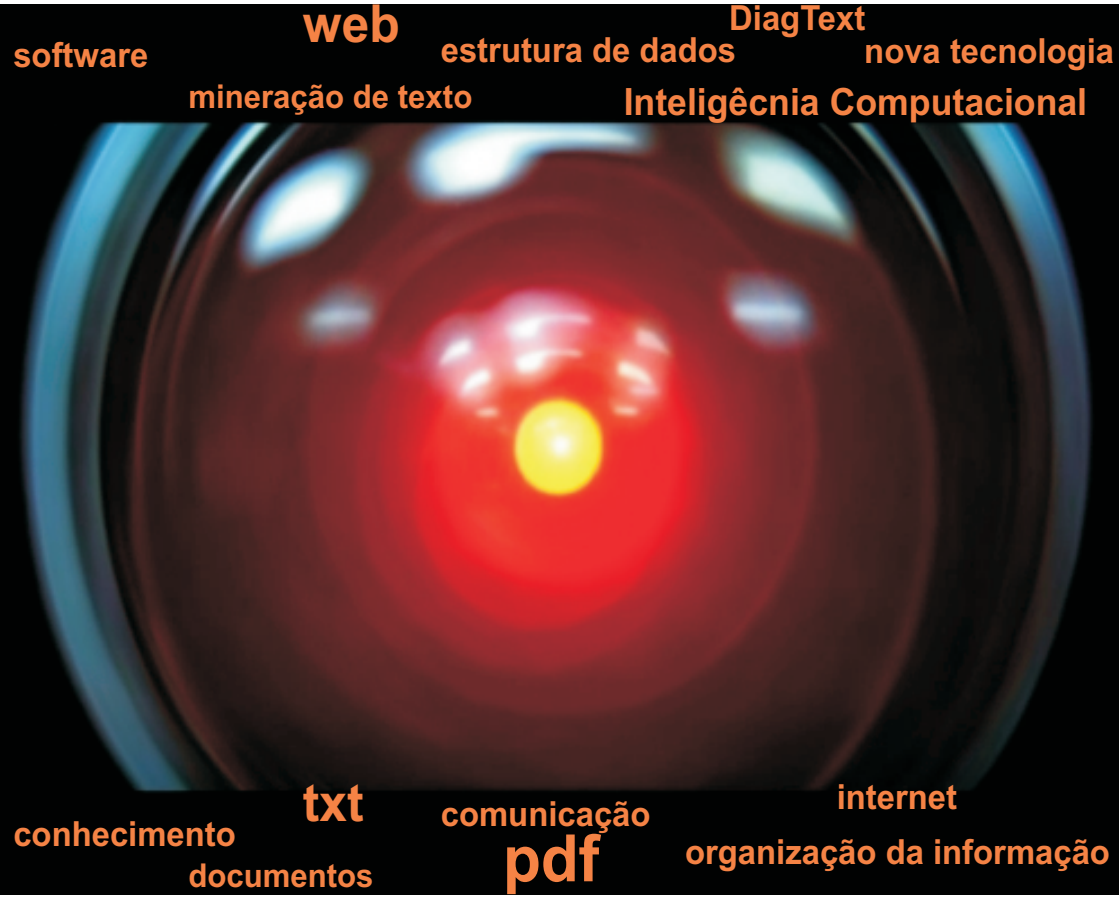
Documentos

Dezembro, 2008 84

ISSN 1677-9274

DiagText: Manual do Usuário



Embrapa

ISSN 1677-9274
Dezembro, 2008

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 84

DiagText: Manual do Usuário

Silvia Maria Fonseca Silveira Massruhá
Helano Póvoas de Lima
André Della Libera Zanchetta
Raphael Fuini Ricciotti

Embrapa Informática Agropecuária
Campinas, SP
2008

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Av. André Tosello, 209
Cidade Universitária "Zeferino Vaz" – Barão Geraldo
Caixa Postal 6041
13083-970 – Campinas, SP
Telefone (19) 3211-5700 – Fax (19) 3211-5754
URL: http://www.cnptia.embrapa.br
e-mail: sac@cnptia.embrapa.br

Supervisão editorial: *Suzilei Almeida Carneiro*
Normalização bibliográfica: *Marcia Izabel Fugisawa Souza*
Revisão de texto: *Adriana Farah Gonzalez*
Editoração eletrônica: *Área de Comunicação e Negócios (ACN)*

1ª. edição on-line - 2008

DiagText : manual do usuário / Silvia Maria Fonseca Silveira

Massruhá ... [et al.]. – Campinas : Embrapa Informática Agropecuária, 2008.

26 p. : il. – (Documentos / Embrapa Informática Agropecuária ; 84)

ISSN 1677-9274

1. Mineração de dados. 2. DiagText. I. Massruhá, Silvia Maria Fonseca Silveira. II. Série.

CDD – 21st Ed.
006.3

Autores

Silvia Maria Fonseca Silveira Massruhá
Doutora em Computação Aplicada,
Pesquisadora da Embrapa Informática Agropecuária
Av. André Tosello, 209, Barão Geraldo
Caixa Postal 6041 - 13083-970 - Campinas, SP
Telefone: 19-3211-5814
e-mail: silvia@cnptia.embrapa.br

Helano Póvoas de Lima
Graduado em Ciências da Computação,
Analista da Embrapa Informática Agropecuária
Av. André Tosello, 209, Barão Geraldo
Caixa Postal 6041 - 13083-970 - Campinas, SP
Telefone: 19-3211-5816
e-mail: helano@cnptia.embrapa.br

André Della Libera Zanchetta
Graduando em Ciências da Computação,
Estagiário da Embrapa Informática Agropecuária
Av. André Tosello, 209, Barão Geraldo
Caixa Postal 6041 - 13083-970 - Campinas, SP
Telefone: 19-3211-5737
e-mail: andrecomacento@gmail.com

Raphael Fuini Ricciotti
Graduando em Engenharia da Computação,
Estagiário da Embrapa Informática Agropecuária
Av. André Tosello, 209, Barão Geraldo
Caixa Postal 6041 - 13083-970 - Campinas, SP
Telefone: 19-3211-5737
e-mail: r_ricciotti@yahoo.com.br

Apresentação

À Embrapa é apresentada, diariamente, uma vasta demanda de informação tanto bibliográfica, de natureza técnico-científica, quanto referencial. Por outro lado, as mudanças ocorridas na sociedade devido aos impactos das novas tecnologias de informação e de comunicação exigem da Embrapa novos procedimentos para organização da informação disponível que resultem em efetiva transferência de tecnologia.

Este tutorial visa apresentar uma ferramenta, denominada Diagtext, que foi desenvolvida no Laboratório de Inteligência Computacional (LabIC) da Embrapa Informática Agropecuária com o objetivo inicial de auxiliar o processo de extração de informação de documentos textuais e identificar agrupamentos que podem dar subsídios ao especialista para decidir como melhor categorizar/classificar e utilizar o recurso. Para isso, foram utilizadas técnicas de mineração de textos.

O DiagText pode ser utilizado em diversas áreas de conhecimento, uma vez que foi desenvolvido de forma a ser independente do idioma no qual os textos se encontram e do assunto que tratam.

Este manual visa viabilizar o uso do DiagText para extração de informação a partir de dados não estruturados tanto em formato (.txt) quanto em formato (.pdf).

Eduardo Delgado Assad
Chefe-Geral

Sumário

DiagText: Manual do Usuário

Silvia Maria Fonseca Silveira Massruhá
Helano Póvoas de Lima
André Della Libera Zanchetta
Raphael Fuini Ricciotti

Introdução

DiagText é uma ferramenta que teve como objetivo inicial auxiliar o processo de extração de informações de documentos textuais que descrevem doenças de culturas agrícolas para formação de uma árvore de decisão baseada nos sintomas das doenças avaliadas. Para isso, foram utilizadas técnicas de mineração de textos (Rezende, 2002; Weiss et al., 2005).

Apesar do objetivo inicial ser a avaliação de doenças agrícolas e seus sintomas, o DiagText pode ser utilizado para diversos fins, em diversas áreas, uma vez que foi desenvolvido de forma a ser independente do idioma no qual os textos se encontram e do assunto que tratam.

Existem duas formas principais de abordagem dos dados textuais (Rezende, 2002). A análise semântica que é baseada na funcionalidade e no significado dos termos ou palavras nos textos e a análise estatística baseada na frequência dos termos ou palavras. Esses tipos de abordagem podem ser utilizados sozinhos ou em conjunto para analisar dados. A abordagem estatística foi adotada nesta primeira versão do DiagText.

O DiagText foi baseado no *software* Eurekha!, desenvolvido no Instituto de Informática da Universidade Federal do Rio Grande do Sul (Wives, 2004).

Instalação

Passo 1: O DiagText foi desenvolvido sob a plataforma Java verão 1.6. Portanto, para executá-lo, é preciso instalar e configurar a máquina virtual Java (JRE) da mesma versão ou posterior. Para tanto, acesse a página de *downloads* da Sun (http://java.sun.com/javase/downloads/index.jsp), faça o *download* da versão mais recente do *JRE* (Java Runtime Environment) e prossiga com a instalação.

Passo 2: faça o *download* do arquivo a partir do CD contendo o *software.* Pode estar no formato de compressão Zip (.zip) ou Rar(.rar).

Passo 3: faça a descompactação do conteúdo (sem alterar o mesmo ou sua estrutura de pastas) em um diretório que não possua 'espaços' no caminho total.

Exemplos:

Windows:
c:/Arquivos de Programas/DiagText/ [inválido]
c:/usuario/DiagText/ [válido]
c:/windows/temp/Diag Text [inválido]

Linux:
/home/Desktop/diagText [válido]
/tmp/Pasta do usuario [inválido]

Utilização

Iniciando

No Windows

Dentro da pasta na qual o arquivo foi descompactado, clique sobre o ícone verde de nome "DiagText".

No Linux

Via linha de comando, dentro do diretório no qual o arquivo foi descompactado, execute o comando:

```
"sh diagText.sh"
```

Certifique-se de que o arquivo `"diagText.sh"` tem permissão de execução. Caso não possua permissão, execute o comando abaixo:

```
"chmod ugo+x diagText.sh"
```

Visão Geral

Na Fig. 1 está apresentada a tela principal do DiagText que contempla 3 partes principais.

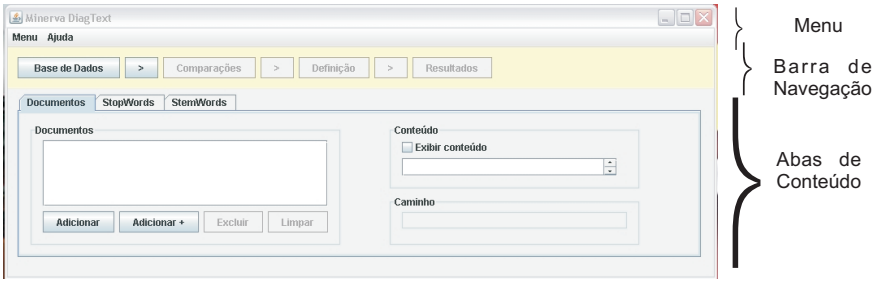


Fig. 1. Tela inicial do DiagText: seleção dos documentos.

Menu: Contém opções básicas da ferramenta em geral. Informações sobre o *software* em Ajuda > Sobre e saída em Menu > Sair.

Barra de Navegação: Contém as fases do processo de mineração seguido pelo DiagText. Para mudar de etapa (seja indo para a próxima ou voltando para alguma anterior), deve-se clicar sobre o nome da etapa desejada (quando se pretende voltar) ou sobre a tecla de seta ('>') para ir para a próxima.

Abas de Conteúdo: Contém as abas referentes à fase atual do processo de mineração em que o usuário se encontra. Para cada fase, um novo conjunto de abas contendo informações relevantes e opções de escolha é apresentado ao usuário para que este faça as definições desejadas.

Etapas da mineração de textos

O processo de extração de conhecimento a partir de dados não estruturados utilizando o DiagText pode ser dividido em 4 etapas principais: Preparação da Base de Dados (Base de Dados), Análise de similaridade dos documentos (Comparações), Definição de parâmetros para agrupamentos (Definição) e Apresentação de resultados (Resultados).

Base de Dados (preparação dos documentos)

Nessa primeira etapa são definidas as informações básicas de entrada da

mineração a ser executada: quais documentos serão avaliados, quais palavras devem ser desconsideradas (*StopWords*) e quais palavras podem ser associadas e consideradas sinônimos (*StemWords*). Cada grupo desses tem sua própria aba para definições.

Documentos

O DiagText suporta dois formatos de texto: arquivos de texto puro (que contenham a extensão '.txt') e arquivos PDF - Portable Document Format (extensão .pdf).

A codificação de caracteres para arquivos de texto puro é a codificação padrão do sistema operacional no qual a ferramenta se encontra em execução. Por exemplo: sendo o sistema operacional *Windows* 2000, a codificação de caracteres utilizada por padrão é ISO-8859-1, mas se o sistema for Ubuntu Linux, a codificação será UTF-8.

Opções (Fig. 1.):

- Adicionar: adiciona um único arquivo suportado à lista de arquivos a serem avaliados.
- Adicionar+: adiciona todos os arquivos suportados de um diretório à lista de arquivos a serem avaliados.
- Excluir: remove um arquivo da lista de arquivos a serem avaliados.
- Limpar: remove todos os arquivos da lista de arquivos a serem avaliados.

StopWords

Em muitos casos, algumas palavras são irrelevantes no contexto de associação entre arquivos (caso mais comum: artigos) e desconsiderá-las tende a melhorar os resultados que venham a ser obtidos. Esse conjunto de palavras que ocorrem com muita frequência é denominado *StopWords*.

Como muitas vezes o mesmo conjunto de palavras é irrelevante em contextos de documentos distintos, é oferecido ao usuário a opção de manipular categorias de *StopWords,* que nada mais são do que listas de *StopWords*

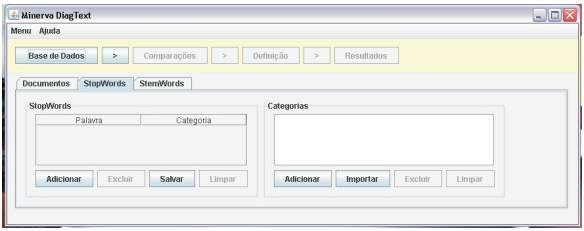


Fig. 2. Tela de definição de *StopWords.*

salvas para utilização e eliminação posterior.
Opções de *StopWords* singulares (Fig. 2):

- Adicionar: adiciona uma palavra na lista de *StopWords* consideradas.
- Excluir: remove a palavra selecionada da lista de *StopWords* consideradas.
- Salvar: salva a atual lista de *StopWords* considerada em uma categoria, cujo nome é definido pelo usuário.
- Limpar: remove todas as *StopWords* da lista de *StopWords* considerada.

Opções de categorias de *StopWords:*

- Adicionar: adiciona uma categoria de *StopWords* na lista de consideradas. Para isso é aberta uma janela contendo a lista de categorias salvas no sistema.
- Importar: adiciona todas as *StopWords* de um arquivo qualquer do sistema na lista de *StopWords* consideradas. Veja abaixo a respeito dos arquivos de *Stopwords.*
- Excluir: remove a categoria de *StopWords* selecionada do projeto, removendo também todas as *StopWords* associadas àquela categoria
- Limpar: remove todas as categorias do projeto, (e consequentemente não considerando mais nenhuma *StopWord* associada a qualquer categoria).

Observação: Os documentos lidos pelo DiagText contendo as *StopWords* devem seguir o seguinte padrão: serem arquivos de texto simples, estarem codificados em ISO 8859-1 e possuírem uma palavra por linha, sem espaços.

Exemplo de conteúdo de documento aceito:

Uma
Um
De
a
Ele
Ela
para

StemWords

Algumas palavras, a título de classificação de textos, deveriam ser associadas entre si para que tivessem o mesmo significado. Esse conjunto de palavras é denominado *StemWords.* Por exemplo: suponha que uma

característica importante num texto é a descrição de um veículo qualquer com cor vermelha. Para uma máquina, as expressões 'carro vermelho', 'carros vermelhos' e 'motos avermelhadas' não têm nada em comum (palavras distintas), mas para os humanos todos eles têm o mesmo significado de "um veículo com cor vermelha". O processo de eliminação de prefixos e sufixos das palavras é denominado *stemming.* No exemplo acima, por meio do processo de *stemming,* pode-se associar “vermelho”, “vermelhos”, “avermelhadas” com o significado "vermelho" e, pode-se também associar "carro", "carros" e "motos" para que todas tenham o significado de "veículo" e possam ser aproximados no cálculo de similaridades entre documentos.

Assim como no caso das *StopWords,* as *StemWords* também podem ser manipuladas por meio de categorias (a título de conveniência ao usuário).

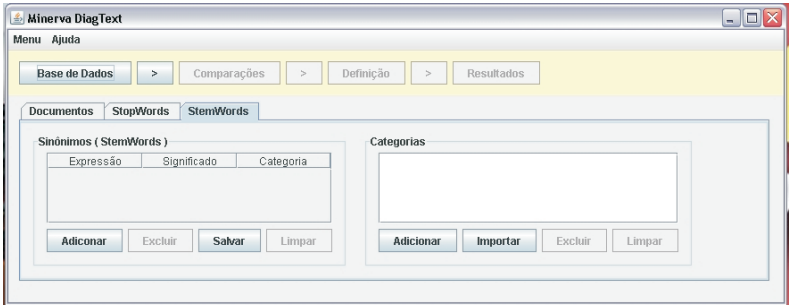


Fig. 3. Tela de definição de *StemWords.*

Opções de *StemWords* singulares (Fig. 3):

- Adicionar : adiciona uma palavra (associada a seu significado) na lista de *StemWords* consideradas.
- Apagar: remove uma palavra da lista de *StemWords* consideradas.
- Salvar: salva a atual lista de *StemWords* consideradas em uma categoria, cujo nome é definido pelo usuário.
- Limpar: remove todas as *StemWords* da lista de *StemWords* consideradas.

Opções de categorias de *StemWords*

- Adicionar: adiciona uma categoria de *StemWords* na lista de consideradas. Para isso é aberta uma janela contendo a lista de categorias salvas no sistema.
- Importar: adiciona todas as *StemWords* de um arquivo qualquer do sistema na lista de *StemWords* consideradas. Veja a observação abaixo a respeito dos arquivos de *StemWords*

- Excluir : remove a categoria de *StemWords* selecionada do projeto, removendo também todas as *StemWords* e associações relacionadas àquela categoria
- Limpar: remove todas as categorias do projeto, (e consequentemente não considerando mais nenhuma *StemWord* associada a qualquer categoria)

Observação: Os documentos lidos pelo DiagText contendo as *StemWords* têm uma estrutura baseada na estrutura dos arquivos de saída da ferramenta PreText, desenvolvida na USP - São Carlos (Matsubara et al., 2003) e, em geral, adota a mesma extensão (.all):

- A mesma estrutura é adotada justamente para que arquivos de *stemização* gerados pelo PreText possam ser importados para o DiagText.

- Basicamente, o documento deve ser um arquivo de texto simples, com codificação *Unicode,* e deve ser dividido em várias partes, cada parte associada a um significado. Cada palavra associada àquele significado deve ser colocada após a linha que contenha o significado e com 6 caracteres de espaço entre o começo da linha e a palavra.

- Exemplo de conteúdo de documento aceito:

```
amarelo:
      amarelado
      amarela
moto:
      motos
cachorro:
      cão
      cadela
```

- Significando que sempre toda palavra "amarelado" e "amarela" serão consideradas como "amarelo", e que toda palavra "motos" será considerada "moto".

Comparações (análise de similaridades)

Nessa segunda etapa são exibidas informações referentes às semelhanças entre os documentos de entrada (considerando as *StopWords* e *StemWords).* São apresentadas a matriz de similaridade e a lista de termos para redefinição de peso.

Similaridades

Entre a transição da etapa de definição da base de dados e a de análise de similaridades são feitos cálculos para definição do nível de semelhança entre entre documentos. Desses cálculos é gerada uma matriz com linhas e colunas dos documentos utilizados, associando cada par de documentos ao valor (em uma escala de 0 a 1) do nível de semelhança entre cada um deles.

Por exemplo: para que possamos saber o nível de semelhança entre os documentos "docX" e "docY", procuramos na tabela a célula, cuja linha seja a linha de "docX" e a coluna de "docY" (ou o inverso).

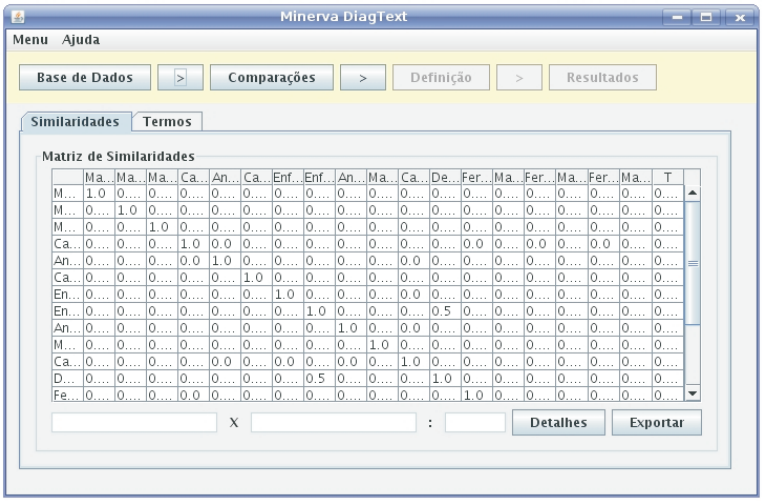


Fig. 4. Matriz de similaridades dos documentos.

- Detalhes: exibe uma janela com informações referentes à(s) semelhança(s) entre os dois documentos da célula selecionada da matriz (informações tais como os termos ou palavras comuns a ambos e suas frequências relativas).
- Exportar HTML: salva a matriz gerada em um arquivo HTML simples, visando facilitar a leitura dos dados e transferência de informação.

Termos

Algumas vezes, pode ser interessante alterar o peso de um determinado termo nos documentos de uma coleção, forçando uma maior (ou menor) consideração de sua presença. Na janela abaixo, podemos alterar o "peso"

de cada termo encontrado, de tal forma que seja multiplicado o número de aparições daquela palavra em cada documento pelo seu peso definido.

Exemplo: suponha que o peso da palavra "doença" seja alterado de 1 para 3. Assim, se um documento apresenta a palavra "doença" repetida 7 vezes, então esse documento passará a ter a mesma palavra repetida 21 vezes, para qualquer critério de comparação.

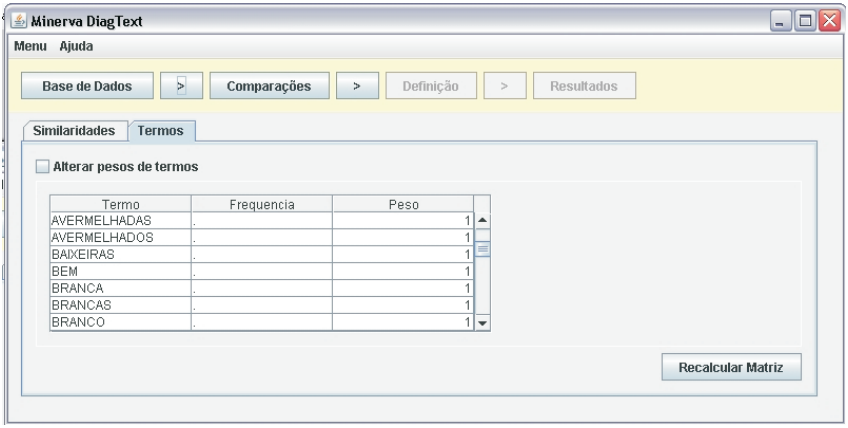


Fig. 5. Tela de alteração de peso dos termos.

Opções (Fig. 5):

- Alterar peso dos termos: Inicia processo de mudança de peso dos termos. Termina apenas quando se desmarca essa opção ou quando se recalcula a matriz de similaridades, clicando em Recalcular Matriz.
- Recalcular Matriz : Após a redefinição dos valores de peso dos termos, a matriz de similaridades dos termos precisa ser recalculada para futuros fins de comparação. Esse botão encerra o processo de redefinição, atualizando a devida matriz.

Definição (parâmetros para agrupamentos)

Nessa terceira etapa são definidas as informações referentes à forma de agrupamento a ser utilizada na divisão dos documentos, como o algoritmo, se for divisão simples ou hierárquica e o nível de semelhança mínimo.

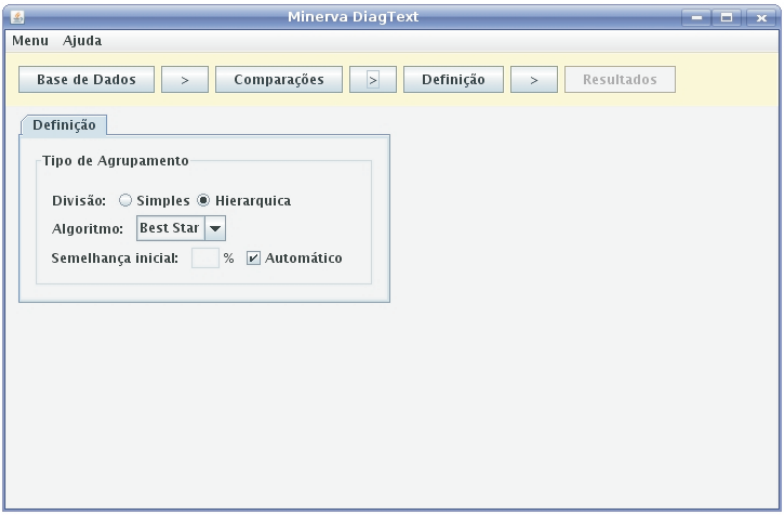


Fig. 6. Definição do tipo de agrupamento.

Opções (Fig. 6):

- Divisão: é definido o nível de divisões a serem realizadas. Se for definido como 'Simples', é realizada uma única divisão em grupos. Se for definido como 'Hierárquica', cada grupo gerado inicialmente será dividido em grupos internos menores, conforme o grau de semelhança entre seus elementos.
- Algoritmo: lista os possíveis algoritmos de agrupamentos. Até a versão atual, são suportados 2 algoritmos: Best Star e Cliques. Esses algoritmos pertencem à classe de métodos grafo-teoréticos que são baseados em teoremas e axiomas conhecidos da teoria de grafos. Maiores detalhes podem ser vistos em Wives (1999).
- Semelhança inicial: utilizada pelo algoritmo Best Star. Define o grau mínimo de semelhança entre os elementos que formarão os primeiros grupos do agrupamento, em uma escala de 0 a 100. Pode ser definido como 'Automático', para que o próprio programa se encarregue de definir o valor que julga mais apropriado.

Resultados (avaliação e visualização dos resultados)

Nessa quarta e última etapa são mostrados os resultados referentes à divisão realizada. São exibidos de duas maneiras: por meio de uma árvore de documentos (Fig. 7) e de uma árvore hiperbólica (Fig. 8).

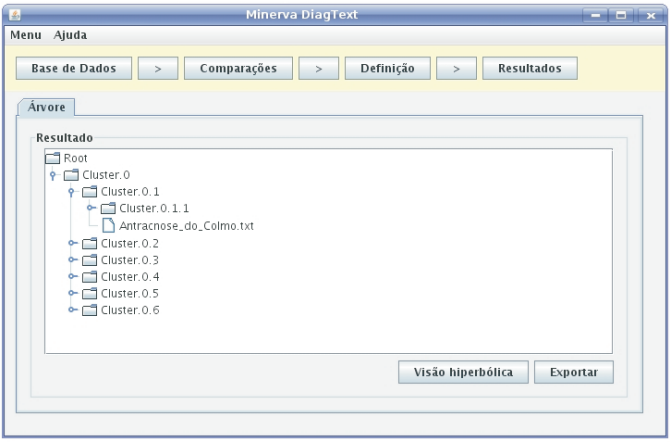


Fig. 7. Apresentação dos agrupamentos gerados.

Opções (Fig. 7):

- Visão hiperbólica (Fig. 8) : abre o navegador *web* padrão definido no sistema operacional exibindo a representação dos agrupamentos em uma árvore hiperbólica.
- Exportar : Salva os arquivos referentes à leitura da árvore hiperbólica no computador de uma forma independente do programa para efeito de intercâmbio de dados.

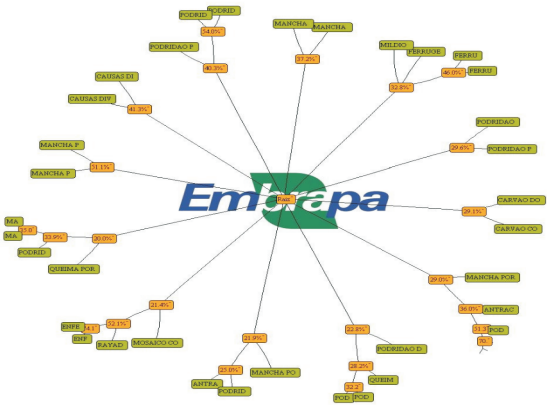


Fig. 8. Apresentação dos agrupamentos em uma árvore hiperbólica.

Exemplo de uso

O usuário André tem aproximadamente 19 documentos de texto referentes a doenças de milho e deseja saber quais doenças são semelhantes em termos de sintomas.

André então inicia o DiagText. Como todos os documentos de texto se encontram no diretório "c:\Doenças\Milho\", na tela de entrada, na aba de Documentos, o usuário clica em "Adicionar+" e navega até o "c:\Doenças\Milho\" (entrando na pasta 'Milho') e clicando em "Abrir" na nova janela que surgiu.

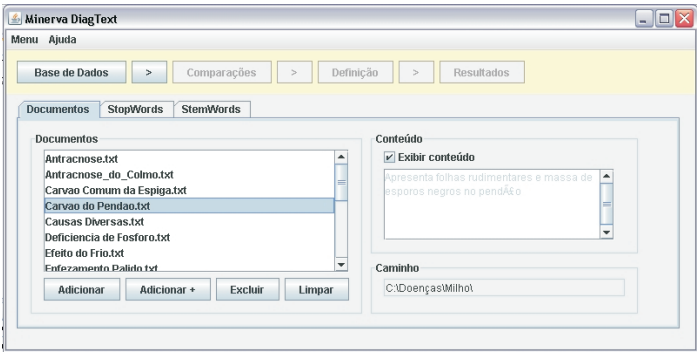


Fig. 9. Exemplo de uso: Base de Dados -> Documentos

Observando a lista de documentos avaliados, André percebe que havia colocado um documento de texto sobre futebol por engano naquela mesma pasta em algum momento anterior. Para não envolvê-lo no processamento, decidiu removê-lo, selecionando-o na lista e clicando em Excluir.

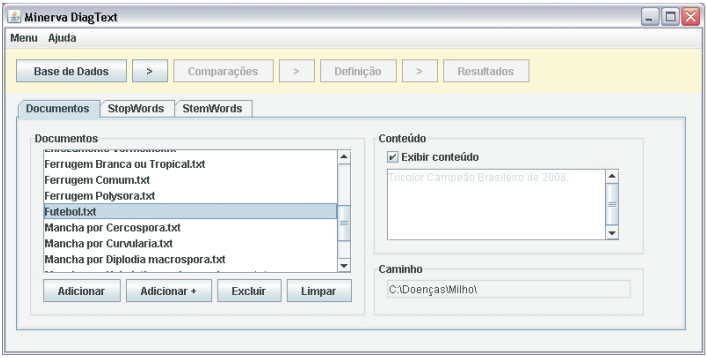


Fig. 10. Exemplo de uso: Base de Dados -> Documentos - Excluir

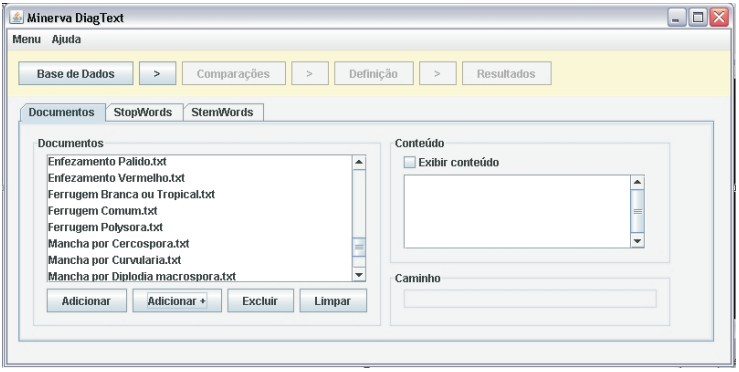


Fig. 11. Exemplo de uso: Base de Dados → Documentos

Com a lista de documentos agora correta, André vai para a próxima etapa, clicando no botão seta ">" ao lado do botão "Base de Dados" na barra de navegação (Figura 11). Em "Comparações" (Fig. 12), como o usuário não tem interesse em vasculhar a matriz de semelhança em busca dos níveis numéricos de semelhança entre cada documento, André passa direto para a próxima etapa clicando no último botão de seta ">" clicável, ao lado do botão "Comparações" na barra de navegação.

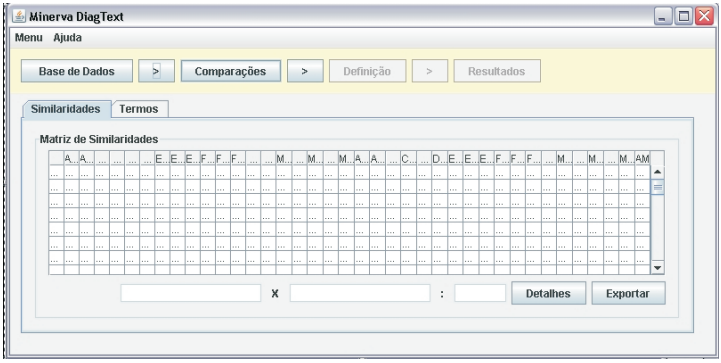


Fig. 12. Exemplo de uso: Comparações → Similaridades

Sem muito conhecimento dos algoritmos (Figu. 13), André deixa as opções apresentadas na aba de Definição como as padrões e vai para a próxima etapa (Fig.14). A ele é apresentado uma estrutura de pastas contendo apenas uma pasta fechada. Clicando na pasta da Fig. 14, o usuário observa as pastas internas e os documentos internos de cada uma.

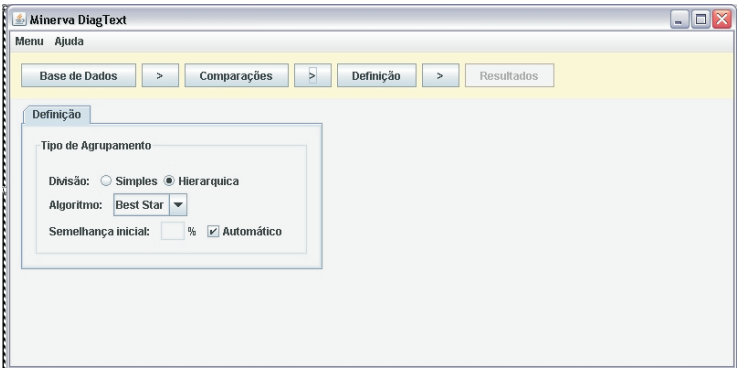


Fig. 13. Exemplo de uso: Definição

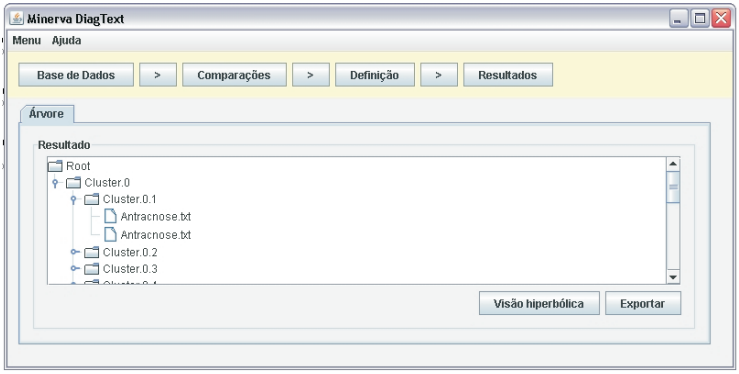


Fig. 14. Exemplo de uso: Resultados

No entanto, ele considera essa visão pouco prática e pouco informativa. Quando clica em "Visão Hiperbólica", no canto inferior direito da janela (Fig. 14), seu navegador padrão é aberto e, como se encontra devidamente configurado para execução de *applets java*, é exibido a mesma árvore vista na estrutura de pastas, mas de uma forma mais dinâmica e navegável (Fig. 15). Passando o mouse sobre os grupos (representados pelos quadrados de cor laranja), André visualiza quais palavras são comuns a todos os documentos (representados pelos quadrados de cor verde) associados àquele grupo.

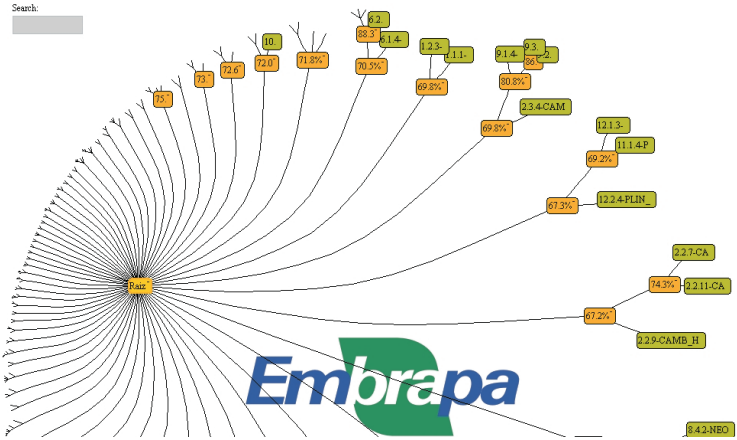


Fig. 15. Exemplo de uso: Visão Hiperbólica

Ele percebe, então, que existem muitos artigos em comum a vários documentos. Como esses artigos são irrelevantes para o que André está procurando considerar, o mesmo decide ignorar todos os artigos em sua consideração.

Volta, então, para a fase de Base de Dados clicando no botão de mesmo texto na barra de navegação. Lá, dentro da aba de *StopWords,* clica em Adicionar, dentro do bloco de Categorias. Na nova janela que surgiu, ele seleciona a categoria "milho" e clica em "Adicionar" e depois sai da janela, clicando em "Sair" (Fig. 16). Feito isso, verifica os novos resultados, percorrendo novamente cada etapa passada até chegar em "Resultados".

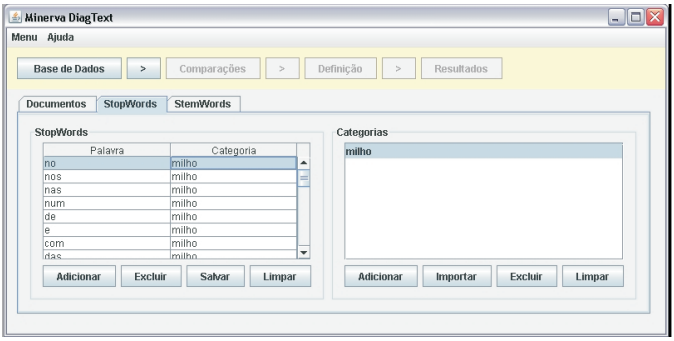


Fig. 16. Exemplo de uso: Base de Dados → *StopWords*

Verificou então que a árvore ficou mais fácil de ser lida, mas ainda apresenta muitas palavras irrelevantes. Palavras como “coloração" e "planta" estão presentes em praticamente todos os documentos e não adicionam nenhuma informação útil ao documento para critérios de agrupamento. Outras palavras são relevantes e de igual significado para André, mas se encontram apresentadas em termos distintos (tais como "vermelho", "avermelhado" e "avermelhada" representam para ele o mesmo significado).

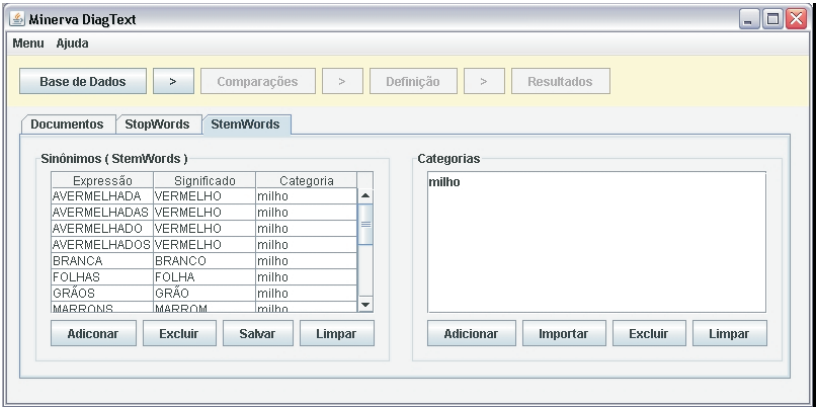


Fig. 17. Exemplo de uso: Base de Dados → *StemWords*

O usuário volta então para a fase de "Base de Dados" e insere as palavras "coloração" e "planta" na lista de *StopWords,* clicando no botão "Adicionar", na área de *StopWords* dentro da aba de *StopWords* e preenchendo o campo que surgiu com as palavras "coloração" e "planta".

Adiciona também a associação da palavra "avermelhada" com o significado de "vermelho", abrindo a aba de *StemWords* e, dentro da área de Sinônimos *(StemWords),* clicando em "Adicionar" e preenchendo o campo "Expressão" com "Avermelhada" e "Significado" com "Vermelho".

Procede novamente até a fase de resultados e reavalia a árvore gerada, repetindo o processo até gerar uma árvore que lhe agrade.

Agradecimentos

Agradecimentos especiais ao Dr. Leandro Wives, do Instituto de Informática da Universidade Federal do Rio Grande do Sul, que nos concedeu gentilmente o código fonte do Eurekha!. Também agradecemos ao Conselho Nacional de Pesquisa - CNPQ pela bolsa de iniciação científica (PBIC)

concedida à Embrapa no âmbito do qual foi desenvolvido este trabalho.

Referências Bibliográficas

MATSUBARA, E. T.; MARTINS, C. A.; MONARD, M. C. *Pretext:* uma ferramenta para pré-processamentos de textos utilizando a abordagem bag-of-words. São Carlos, SP: ICMC-USP, 2003. 57 p. (Relatório Técnico ICMC-USP, 209).

REZENDE, S. O. *Sistemas inteligentes:* fundamentos e aplicações. Barueri: Editora Manole, 2002. 525 p.

WEISS, S. M.; INDURKYA, N.; ZHANG, T.; DAMERAU, F.J. *Text mining*: predictive methods for analyzing unstructured information. New York: Springer, 2005. 236 p.

WIVES, L. K. *Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de "clustering".* 1999. 102 f. Dissertação (Mestrado) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

WIVES, L. K. *Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos*. 2004. 136 f. Tese (Doutorado) -- Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Embrapa
Informática Agropecuária

Ministério da
Agricultura, Pecuária
e Abastecimento

BRASIL
UM PAÍS DE TODOS
GOVERNO FEDERAL