



BoardWords

Manual do Usuário
Versão 1.0.0

Sumário

1	Introdução	1
2	Funcionalidades	2
2.1	Textos	3
2.2	Stopwords	3
2.3	Processar	3
2.4	Palavras	4
2.5	Calendário	4
2.6	Painel	5
2.7	Salvar	8
2.8	Editor	8
2.9	Ajuda	8
3	Funcionamento	9

BoarWords é uma aplicação da área de Computação Gráfica, que agrega técnicas de Visualização de Informações e Mineração de Texto conjuntamente. Foi desenvolvido com o objetivo de auxiliar na análise de dados de texto gerados a partir de um fórum de discussões em um AVA (Ambiente Virtual de Aprendizagem). Dessa forma, a aplicação possui interesses específicos, mas acreditamos que isso não seja um fator limitante para sua generalização e utilização em outros ambientes de análise.

O nome da aplicação é uma adaptação do termo “*board of words*” (trad. tábua de palavras), que faz referência direta ao aspecto visual das principais funcionalidades da aplicação. Esse projeto é uma extensão do projeto desenvolvido por Pacheco Jr., intitulado “Processo de *Visual Analytics* para a Análise Qualitativa de Conteúdo em Fóruns de Discussão” [1], e foi concebido baseado nos bons resultados obtidos no mesmo. Através dele, foi possível extrair novas informações sobre o conjunto de dados analisados por Rinaldi [2], contribuindo para uma compreensão mais ampla do escopo analisado.

A principal contribuição desse projeto é a inserção de novas funcionalidades de Visualização, das quais o atributo temporal é o mais relevante, justificando o termo ‘Conteúdo-Temporal’ usado no título do trabalho. Através disso, será possível realizar uma análise mais detalhada dos dados textuais, disponibilizando ao usuário ou pesquisador um conjunto maior de informações a seu respeito.

Funcionalidades

As principais funcionalidades do BoardWords se referem ao esquema natural do processo de Mineração de Dados definido por Fayyad [3], conforme ilustrado na Figura 2.1.

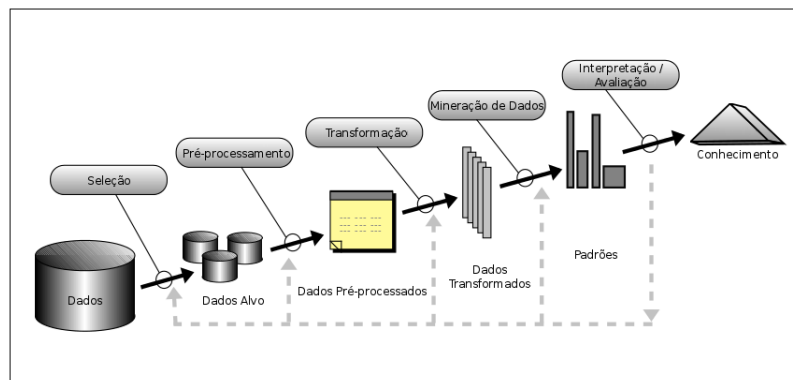


Figura 2.1: etapas de KDD segundo Fayyad; adaptado de [3].

No menu principal (Figura 2.2) estão localizadas as funções relacionadas às etapas de Seleção, Pré-processamento, Transformação e Mineração de Dados descritas por Fayyad, representadas pelos botões *Textos*, *Stopwords*, *Processar* e *Painel*. As demais funcionalidades servem como apoio à esse processo, oferecendo recursos de interatividade para trabalhar com os dados.

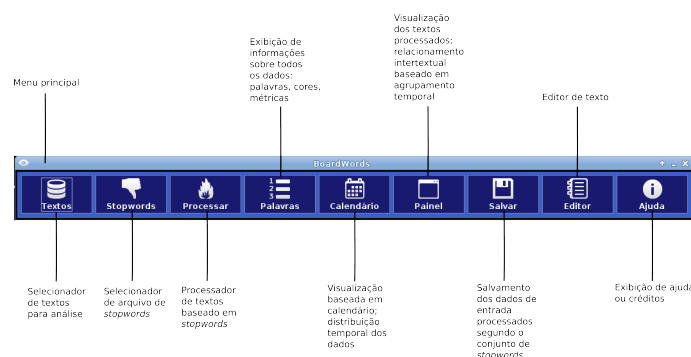


Figura 2.2: menu principal.

Observação: para o correto funcionamento do programa, todos os arquivos utilizados devem necessariamente ser do formato *txt* e possuir codificação de caracteres ISO 8859-1. Essa aplicação também depende da instalação e configuração correta do *software* JRE, na versão 6 ou superior.

2.1 Textos

Textos é o meio de seleção do diretório contendo os dados ou arquivos de texto a serem analisados. Os arquivos devem ser nomeados de acordo com o esquema *mmdaaaaa_hh-mm.txt*, que representam respectivamente o mês, dia, ano, hora e minuto de criação dos textos. Pelo fato dos dados analisados nesse projeto serem originários de um fórum de discussões em um AVA específico, o conteúdo dos textos segue o padrão abaixo:

```
De: <remetente>
Para: <destinatário>
dd/mm/aaaa hh:mm:ss
<texto>
```

O padrão utilizado nos arquivos de texto não é necessário para se utilizar a aplicação, pois os seus dados não são aproveitados, já que os atributos temporais são derivados apenas do nome dos arquivos, que obrigatoriamente devem seguir o modelo já citado. Dessa forma, a aplicação pode ser aplicada mais facilmente em outros ambientes, aumentando seu campo de atuação. Entretanto, levando em conta o objetivo inicial proposto para esse trabalho, os arquivos de texto devem possuir no mínimo três linhas de dados (relativas ao cabeçalho dos arquivos analisados) que, por serem desprezadas, podem possuir qualquer conteúdo.

2.2 Stopwords

Stopwords é o meio de seleção do arquivo que contém as *stopwords* ou palavras a serem filtradas dos dados de entrada, selecionados pelo botão *Textos*. O arquivo deve ser do tipo *txt* e seu conteúdo deve possuir o formato abaixo, onde cada *< palavra_n >* representa a *stopword* da *n*-ésima linha do arquivo:

```
<palavra_1>
<palavra_2>
...
```

2.3 Processar

Processar realiza o processamento dos dados de entrada baseado no arquivo de *stopwords*, separando as palavras ou em *startwords* ou em *stopwords* e removendo o conjunto de caracteres apresentados na Figura 2.3, considerados nulos nessa análise textual, assim como espaços em branco.



Figura 2.3: conjunto de caracteres considerados nulos.

2.4 Palavras

Palavras é o meio de exibição das *startwords* com suas respectivas cores de representação e dados das métricas aplicadas. A função é ilustrada pela Figura 2.4, e o conjunto de suas funcionalidades é apresentado na Tabela 2.1.

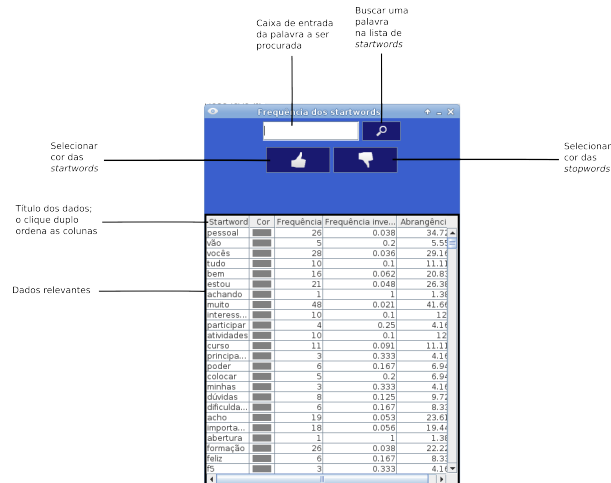


Figura 2.4: janela de exibição das frequências das palavras.

Ícone	Texto	Ação	Função
		Clique único	Realiza buscas por <i>startwords</i> na lista de palavras.
		Clique único	Seleciona a cor das <i>startwords</i> .
		Clique único	Seleciona a cor das <i>stopwords</i> .
	Startword	Clique único	Exibi uma palavra. Se o título for clicado, ordena os valores em ordem crescente/descrescente.
	Cor	Clique único	Exibi a cor de representação de uma palavra. Se o título for clicado, ordena os valores em ordem crescente/descrescente.
	Frequência	Clique único	Exibi a quantidade absoluta (em todos os textos) de ocorrências de uma palavra. Se o título for clicado, ordena os valores em ordem crescente/descrescente.
	Frequência inversa	Clique único	Exibi o inverso da quantidade absoluta de ocorrências de uma palavra, isto é, $ft_i = \frac{1}{ft_a}$, onde ft_a é a frequência total absoluta e ft_i é a frequência total inversa dessa palavra em todos os textos. Se o título for clicado, ordena os valores em ordem crescente/descrescente.
	Abrangência (%)	Clique único	Exibi o percentual de textos em que uma palavra é citada. Se o título for clicado, ordena os valores em ordem crescente/descrescente.

Tabela 2.1: subfuncionalidades de *Palavras*.

2.5 Calendário

Calendário é o meio de exibição da distribuição da data de criação dos dados sobre a forma de um calendário, ilustrado pela Figura 2.5. O conjunto de suas funcionalidades é apresentado na Tabela 2.2.



Figura 2.5: janela de exibição da distribuição dos textos.

Ícone	Texto	Ação	Função
		Clique único	Seleciona a cor dos dias válidos no calendário, onde será aplicado o percentual relativo à quantidade de textos presentes em determinado dia, representado por pontos ou retângulos.
		Clique único	Seleciona a cor dos dias inválidos (por exemplo, dia 32).
		Clique único	Seleciona a cor do painel de fundo da visualização.
		Clique único	Seleciona a cor de seleção dos dias.
		Clique único	Exporta os dados contidos nos dias selecionados para uma nova janela de Visualização.
		Clique único	Diminui a escala da visualização.
		Clique único	Aumenta a escala da visualização.
		Rolagem do <i>scroll</i> do mouse para baixo	Diminui a escala da visualização.
		Rolagem do <i>scroll</i> do mouse para cima	Aumenta a escala da visualização.
		Clique único sobre um dia válido	Seleciona/deseleciona o dia.
		Clique duplo sobre um dia válido	Seleciona/deseleciona o mês referente ao dia.
		Clique duplo sobre o painel de fundo	Deseleciona os dias selecionados.
		Posicionamento do cursor sobre um dia válido	Exibe as informações sobre o dia, no formato <i>dd/mm/aaaa</i> <i>< dia_da_semana ></i> <i>n texto(s)</i> , onde <i>dd/mm/aaaa</i> é o formato de data, <i>< dia_da_semana ></i> corresponde ao dia da semana daquele dia (segunda, por exemplo), e <i>n</i> é a quantidade de textos criados naquele dia.

Tabela 2.2: subfuncionalidades de *Calendário*.

2.6 Painel

Painel é o meio de criação de um ambiente de Visualização contendo todos os dados processados. O conjunto de suas funcionalidades é apresentado na Tabela 2.3, na Tabela 2.4 e na Tabela 2.5.


Ícone	Texto	Ação	Função
		Clique único	Processa e exibe a Visualização baseado nas configurações de entrada.
		Clique único	Configura os parâmetros da Visualização. Através dessa função, é exibida uma janela de configuração da Visualização a ser realizada, que tem como princípio o conceito de agrupamento de dados ou textos. O usuário pode escolher entre agrupar os textos por períodos de dias, meses ou anos, com valores entre [1-999], denominado <i>agrupamento de datas</i> , e períodos de minutos e horas, com valores variáveis [1-1440] e [1-24], respectivamente, denominado <i>agrupamento de horários</i> .
		Clique único	Importa um arquivo de configurações de Visualização, no formato: <pre>grouping:<boolean> legend:<boolean> frequencies:<boolean> grouping_dates:<day month year> value_grouping_dates:<integer> grouping_schedules:<minute hour> value_grouping_schedules:<integer></pre>
		Clique único	Exporta as configurações de Visualização atuais para um arquivo.
		Clique único	Seleciona a cor dos identificadores de horário de criação dos textos e dos agrupamentos.
		Clique único	Seleciona a cor de fundo do painel de visualização.
		Clique único	Seleciona a cor de seleção dos textos e palavras da Visualização.
		Clique único	Seleciona palavras no texto, baseado em um arquivo de entrada com aspecto similar ao arquivo de <i>stopwords</i> , onde cada linha possui uma palavra a ser destacada.
		Clique único	Seleciona a cor da fonte dos textos onde são exibidas as palavras selecionadas.
		Clique único	Exibe as palavras selecionadas sobre os arquivos de texto originais.
		Clique único	Seleciona os textos, baseado em um arquivo de entrada com aspecto similar ao arquivo de <i>stopwords</i> , onde cada linha possui uma palavra a ser destacada.
		Clique único	Exporta os textos selecionados para uma nova janela de Visualização.
		Clique único	Diminui a escala da visualização.
		Clique único	Aumenta a escala da visualização.
	Rolagem do <i>scroll</i> do mouse para baixo		Diminui a escala da visualização.
	Rolagem do <i>scroll</i> do mouse para cima		Aumenta a escala da visualização.
	Clique duplo na área de exibição dos textos originais		Deseleciona todas as palavras destacadas.
	Clique duplo sobre o painel de Visualização		Deseleciona os textos ou palavras selecionados.
	Posicionamento do cursor sobre o identificador do horário de criação dos textos		Exibe a data de criação do texto.

Tabela 2.3: subfuncionalidades de *Painel*.

Ícone	Texto	Ação	Função
		Clique duplo sobre as palavras do texto	Exibe o texto original em que está contida a palavra, sendo possível visualizar os demais textos analisados pelos botões de próximo texto e texto anterior na janela de exibição dos textos originais.
		Posicionamento do cursor sobre as palavras do texto	Exibe a data e hora de criação do texto no menu da janela (lado direito) e as informações do ponto; se <i>Frequência</i> não estiver selecionado, exibe apenas a palavra representada pelo ponto, senão, exibe as seguintes informações: <p>p:<palavra> fa:<fp_a> fr:<fp_r> fi:<fp_i></p>

Tabela 2.4: subfuncionalidades de *Painel*.

Ícone	Texto	Ação	Função
	Dia	Clique único	Agrupar os textos por grupos cronológicos de dias.
	Mês	Clique único	Agrupar os textos por grupos cronológicos de meses.
	Ano	Clique único	Agrupar os textos por grupos cronológicos de anos.
	Minuto	Clique único	Agrupar os textos por subgrupos cronológicos de minutos.
	Hora	Clique único	Agrupar os textos por subgrupos cronológicos de horas.
	Agrupamento	Clique único	Insere/remove o agrupamento de dados para realizar a Visualização.
	Legenda	Clique único	Exibe/omite os identificadores dos agrupamentos e horários de criação dos textos.
	Frequência	Clique único	Exibe/omite as frequências do cálculo de relevância das palavras no texto em que se encontram. As métricas calculadas são frequência absoluta (fp_a – frequência parcial absoluta), frequência relativa (fp_r – frequência parcial relativa) e frequência inversa de documentos (fp_i – frequência parcial inversa), definidas a seguir [4]: <ul style="list-style-type: none"> • fp_a: quantidade de vezes em que a palavra é citada no texto. • fp_r: razão entre a fp_a e a quantidade de palavras relacionadas no texto processado, dada pela função $fp_r = \frac{fp_a}{n}$, onde n é a quantidade total de palavras do texto processado. Através desse cálculo é possível ter a percepção do percentual de ocorrência que da palavra sobre o texto. • fp_i: razão entre a fp_a e o número de documentos em que a palavra ocorre ou a_a (abrangência absoluta), de onde é derivado o cálculo da abrangência das palavras em relação a todos os textos. A função que define fp_i é dada por $fp_i = \frac{fp_a}{a_a}$. <p>Através desse cálculo é possível destacar termos ou palavras que aparecem em menor escala nos textos e diminuir a importância dos termos mais frequentes, que aparecem mais comumente nos textos.</p>

Tabela 2.5: subfuncionalidades das configurações de Visualização de *Painel*.

2.7 Salvar

Salvar é o meio de armazenamento dos arquivos de entrada processados com base no arquivo de filtro contendo as *stopwords*. Quando os arquivos de texto são processados, são removidas as três primeiras linhas relativas ao cabeçalho do texto e todas as palavras que são *stopwords*. Os textos são comprimidos em uma única linha, onde as palavras, transformadas em minúsculas, são separadas por espaçamento único.

2.8 Editor

Editor é o meio de criação de um ambiente de edição de textos, que pode ser utilizado para várias finalidades, inclusive para o usuário registrar informações extraídas dos dados analisados. O conjunto de suas funcionalidades é apresentado na Tabela 2.6.






Ícone	Texto	Ação	Função
		Clique único	Seleciona um arquivo de texto para edição.
		Clique único	Cria um novo arquivo de texto.
		Clique único	Salva o arquivo de texto atual, sobrescrevendo-o se este já existir.
		Clique único	Salva o arquivo de texto atual com nome e localização variáveis.
		Clique único	Destaca uma palavra no texto.
		Clique duplo na área de exibição dos textos	Deseleciona todas as palavras destacadas.

Tabela 2.6: subfuncionalidades de *Editor*.

2.9 Ajuda

Ajuda é o meio de exibição do manual da aplicação e dos créditos de desenvolvimento.

Funcionamento

O BoardWords se baseia no modelo de Mineração de Dados definido por Fayyad (ver Figura 2.1). Dessa forma, as etapas necessárias para analisar dados através da aplicação são:

1. Seleção do diretório contendo os dados ou arquivos de entrada, através do botão *Textos* (menu principal);
2. Seleção do arquivo de *stopwords*, através do botão *Stopwords* (menu principal);
3. Processamento dos dados de entrada baseado no arquivo de *stopwords*, através do botão *Processar* (menu principal);

O usuário também pode utilizar as funções de *Editor* e *Ajuda*, independentes de outros procedimentos. Realizados os procedimentos necessários, todas as funcionalidades da aplicação tornam-se acessíveis ao usuário. Através do botão *Palavras* (menu principal), podem ser visualizadas informações sobre todos os dados de entrada, como *startwords*, frequências das palavras e abrangência do termo em todos os textos.

No botão *Calendário* (menu principal), pode ser realizada uma análise mais superficial ou geral do conjunto de dados de entrada, sendo possível identificar e selecionar regiões de maior interesse para uma análise mais detalhada, como picos de atividade, destacados pela maior intensidade de coloração.

O *Painel* é a principal funcionalidade da aplicação, pois através dele são realizadas as análises mais detalhadas do conjunto de dados. Os relacionamentos intertextuais são obtidos pelo princípio de agrupamento temporal, no qual o usuário pode organizar os textos por períodos de tempo, que podem ser por datas (dias, meses ou anos) ou por horários (minutos ou horas). No caso do agrupamento de horários, a entrada de valores divisores de 6 (1 minuto ou 6 horas, por exemplo) implica na coloração dos identificadores de agrupamento baseado na transição simbólica da luminosidade do sol sobre o dia, representada por 4 faixas de coloração. Esse artifício permite identificar mais rapidamente períodos relacionados pela proximidade.

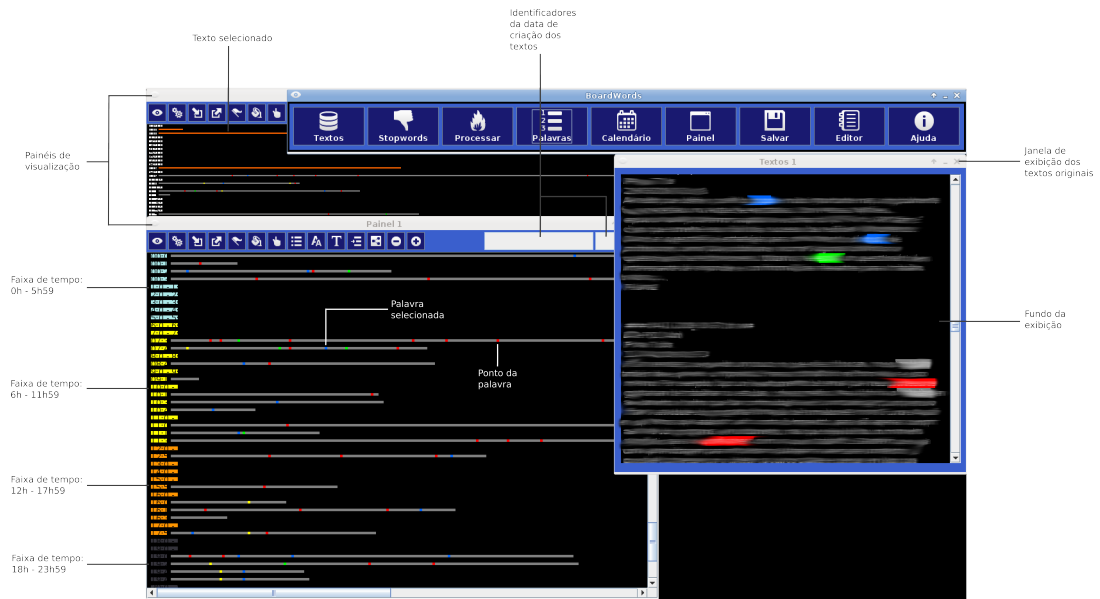


Figura 3.1: ambiente de análise múltipla.

Na Figura 3.1, é exibido um exemplo de análise múltipla da aplicação. Através da janela de exibição dos textos originais, o usuário pode relacionar a visualização com os textos originais (sem processamento), destacando as palavras que achar conveniente para a análise. Nesse exemplo, os textos originais aparecem borrados para manter o sigilo dos dados analisados.

Referências Bibliográficas

- [1] Pacheco Jr., J. C. Processo de visual analytics para a análise qualitativa de conteúdo em fóruns de discussão. *PIBIC*, *id.* 16856, Presidente Prudente, Nov. 2011.
- [2] RINALDI, R. P. *Desenvolvimento profissional de formadores em exercício: contribuições de um programa online*. 2009. Tese (Doutorado em Física) - Universidade Federal de São Carlos, Centro de Educação e Ciências Humanas, Curso de doutorado em Educação, São Carlos, 2009.
- [3] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*, Rhode Island, July 1996.
- [4] MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. *Technical Report*, *Instituto de Informática, Universidade Federal de Goiás*, Goiás, Brasil, , n. INF05/07, 2007.

PROJETO DE PESQUISA

Programa Institucional de Bolsas de Iniciação Científica – PIBIC 2012/2013

Orientador: Prof. Dr. Milton Hirokazu Shimabukuro (DMC).

Aluno: João Vítor Antunes Ribeiro (Bacharelado em Ciência da Computação).

Colaboração: Profa. Dra. Renata Portela Rinaldi (Departamento de Educação).

Unidade: Faculdade de Ciências e Tecnologia – Campus de Presidente Prudente
Departamento de Matemática e Computação - DMC.

Área: Ciência da Computação.

Título: Visualização Interativa de Dados para Suporte à Atividade de Análise Qualitativa ‘Conteúdo-Temporal’ de Fóruns de Discussão.

Agradecimentos: Alisson Fernando Coelho do Carmo
Leonardo Tadashi Nozawa

Copyright (C) 2012, 2013 BoardWords

This file is part of BoardWords.

BoardWords is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

BoardWords is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with BoardWords. If not, see <<http://www.gnu.org/licenses/>>.