ISSN 2236-8884

Identificação de Palavras Compostas como Auxílio à Mineração de Textos: Desenvolvimento do Compostas_v1

Identification of Compound Nouns in Text Mining: Development of Compostas_v1 Software

Yasmmin Martins[1], Breno F. Terra Azevedo[1-2], Helvia Pereira Pinto Bastos[1-2]

[1] Núcleo de Informática na Educação (NIE) – Instituto Federal Fluminense
Campos dos Goytacazes, RJ, *campus* Campos-Centro.
[2] Doutorado em Informática na Educação – Universidade Federal do Rio Grande do Sul
nim_asay@hotmail.com; bterra@iff.edu.br; helviabastos@yahoo.com.br

Abstract: *This paper describes Compostas_v1, a program developed with the objective of processing compound words in Portuguese. The aim of the project is to solve problems found in two other Text Mining programs. Thus, the article provides an overview of that field of research, and discusses other issues related to the processing of natural language.*

Key words: *Text Mining. Natural Language Processing. Compound Words.*

Resumo: O artigo descreve o programa Compostas_v1, desenvolvido, particularmente, com o objetivo de processar palavras compostas em Língua Portuguesa de forma a resolver lacunas encontradas em dois programas de Mineração de Texto. Assim, o trabalho apresenta noções gerais sobre esse campo de estudo, e discute algumas questões relacionadas ao processamento da linguagem natural.

Palavras Chaves: Mineração de Textos. Processamento da Linguagem Natural. Palavras Compostas.

Introdução

Este trabalho apresenta o *software* de mineração de textos Compostas_v1, como parte da pesquisa que está sendo realizada no projeto de iniciação científica "Mineração de Textos Eletrônicos"

ISSN 2236-8884

desenvolvida no Núcleo de Informática na Educação (NIE) do Instituto Federal Fluminense, *campus* Centro. Esse projeto visa ao desenvolvimento de funcionalidades que possam resolver algumas limitações encontradas nos programas Sobek (LORENZATTI, 2007) e Eurekha (WIVES, 1999).

A Mineração de Texto (*Text Mining*) se insere na área de conhecimento denominada Descoberta de Conhecimento em Texto (*Knowledge Discovery in Text – KDT*), campo de estudo e de aplicação crescente, particularmente devido à necessidade de processar e / ou filtrar o grande volume de informações encontrado na Internet.

As questões relacionadas à Mineração de Texto são discutidas na Seção 2 deste artigo, incluindo sua conceituação, aplicações e etapas de realização. A Seção 3 apresenta as implicações de natureza linguística envolvidas no processamento automático de textos que motivaram a criação do Compostas_v1 e, por fim, a Seção 4 fornece maior detalhamento do desenvolvimento desse *software*.

Mineração de Textos

A área de Mineração de Textos (MT) ou Descoberta de Conhecimento em Textos tem relação com outra grande área chamada Mineração de Dados (*Data Mining*). Enquanto esta se preocupa em descobrir padrões e informações existentes em banco de dados, a MT é um conjunto de métodos usados para explorar, organizar, achar e descobrir informações em uma base de textos. Hearst (2003) define a MT como sendo "a descoberta por computador de informação nova, previamente desconhecida, pela extração automática de informação de diferentes fontes escritas".

Os diferentes campos que realizam o processamento da linguagem apresentam interfaces de limites indefinidos. Exemplificando, a MT possui técnicas que auxiliam a análise de textos em campos da Linguística Computacional como a Linguística de Corpus e o Processamento da Linguagem Natural (OTHERO; MENUZZI, 2005, p.22). Por sua vez, MT envolve conhecimentos e se utiliza de outras áreas em seu processo, tais como Informática, Estatística, Linguística e Ciência Cognitiva. A área de Linguagem é usada para reconhecer as palavras, os textos, as classes, e identificar o sentido e a sintaxe; a de Informática é utilizada para fazer o processamento automático dos textos, os cálculos presentes ao longo do processo de mineração, assim como as demais desempenham sua função em algumas fases do processo.

Segundo Feldman e Sanger (2007), a mineração de textos pode ser definida como um pro-

ISSN 2236-8884

cesso intensivo de conhecimento no qual um usuário interage com uma grande quantidade de documentos utilizando ferramentas para análise dos mesmos. O objetivo é extrair informações úteis a partir de coleções de documentos. Essas informações são identificadas em padrões interessantes nos dados textuais não estruturados.

Os sistemas de mineração de textos baseiam-se em rotinas de pré-processamento, algoritmos para descoberta de padrões, e elementos para apresentação dos resultados. As etapas que compõem a arquitetura de um sistema para mineração de textos (Figura 1) são: operações de pré-processamento, geração de documentos processados, mineração, apresentação dos resultados. O usuário do sistema interage com a etapa de pré-processamento, com o núcleo de mineração e com a apresentação dos resultados.

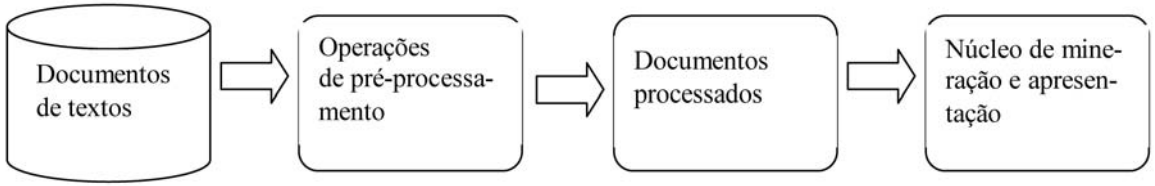

Figura 1: Arquitetura de um sistema para mineração de textos

Essas etapas do processo de mineração de texto envolvem problemas de natureza relacionados à linguagem, como discutido a seguir.

Questões linguísticas no processamento textual

O projeto de pesquisa em que se insere este trabalho consiste na descoberta de conhecimento e informações relevantes em fontes de dados não estruturados ou semiestruturados como *chats* e fóruns de discussão educacionais, ofertados em plataformas de gestão de cursos a distância, como o Moodle[1], por exemplo. Dois programas estão sendo aprimorados, como explicado a seguir.

[1] http://moodle.org

ISSN 2236-8884

O programa Sobek (LORENZATTI, 2007) permite construir grafos a partir de informações estatísticas obtidas do texto. Os grafos, por sua vez, apresentam informações referentes ao número absoluto e relativo de ocorrência dos termos (vértices) e relacionamentos (arestas) em determinado documento. O grafo obtido representa uma *rede dos conceitos* que foram trabalhados dentro do texto. Já o *software* Eurekha (Wives, 1999) verifica a frequência de ocorrência de termos (*lexicometria*) e agrupa os textos com conteúdos semelhantes (*clustering*). O objetivo inicial do projeto é aperfeiçoar esses programas buscando corrigir algumas falhas e adicionar funções inexistentes como identificação de palavras compostas e do significado correto dos termos numa determinada frase, conforme seu contexto estrutural e lexical.

É importante ressaltar que, neste estudo, considera-se como "palavra composta", todas as palavras cujos termos que as precedem ou sucedem funcionam como modificadores, complementando ou alterando seu sentido. Assim, para os autores deste trabalho, palavras compostas são todas aquelas que, juntas, possuem um sentido diferente das mesmas em separado. Essa é uma questão relevante para a compreensão de um texto, já que palavras isoladas podem ter sentidos diferentes quando acompanhadas por outras. Por exemplo, em "Rio de Janeiro", cada termo isoladamente tem seu próprio significados, mas juntos formam o nome da cidade, capital, município e estado. Da mesma forma, verifica-se que as palavras podem assumir sentidos diferentes de acordo com o contexto e estrutura da frase. Um exemplo desse caso é a palavra "manga", que pode ser (i) a parte que protege a lâmpada num lustre, (ii) a parte de uma camisa, (iii) a fruta, e (iv) como sinônimo do verbo "caçoar".

A chamada ambiguidade lexical é, geralmente, resolvida em programas denominados *parsers* ou *analisadores sintáticos*. Segundo Othero e Menuzzi (2005, p. 123), os parsers são programas de computador que realizam "a interpretação automática (ou semiautomática) de sentenças" pela classificação morfossintática de palavras e expressões dentro das frases.

O software Compostas_v1

Com dito acima, o *software* Compostas_v1 foi construído para suprir uma das necessidades encontradas nos *softwares* de mineração (Sobek e Eurekha) utilizados no projeto de iniciação científica "Mineração de Textos Eletrônicos". O código-fonte do *software* foi feito na linguagem de programação Java.

ISSN 2236-8884

O programa possui uma interface gráfica que faz com que o usuário o entenda melhor e possa usá-lo com maior facilidade. Assim, o usuário tanto pode identificar as palavras compostas, quanto visualizar a lista de palavras produzida por ele sem as *stopwords* (palavras que não são relevantes para a compreensão do texto e que, na etapa de pré-processamento, são eliminadas para que seja armazenada a menor e a mais completa quantidade de informações possível a serem mineradas).

Além de poder escrever as frases que desejar, o usuário poderá carregar outro arquivo em que queira fazer esse procedimento, por exemplo, um texto de 15 ou 20 linhas. Além disso, o programa também possui a função de apagar as listas de palavras geradas, e o texto que ele havia digitado.

O programa foi desenvolvido em vários passos, listados a seguir:

• Implementação de componentes gráficos para facilitar a compreensão e ter uma melhor interação com o usuário (esclarecendo, inclusive, o modo de uso do *software*);

• Adição de funcionalidades aos botões criados na fase anterior. Os botões que exercem as funções mais importantes são o "adicionar texto" e o "carregar"; os demais só possuem a função de limpar o conteúdo do campo de texto e das listas de palavras criadas.

O botão "adicionar texto" faz com que sejam feitas as seguintes atividades:

• Leitura de dois arquivos (por exemplo, do disco local e / ou de um *pen drive*), contendo palavras compostas e *stopwords*. Esses arquivos (todos com extensão ".txt"), contém (i) as palavras compostas a serem encontradas nos textos carregados pelos usuários, e (ii) palavras irrelevantes para a compreensão do texto.

• Retirada de todos os sinais de pontuação (vírgula, ponto de exclamação, ponto final, ponto de interrogação, etc.), por serem desnecessários para a identificação das palavras compostas.

• Retirada de todos os espaços e colocação de todo o texto em letra minúscula. Isso é necessário porque o programa faz a identificação de acordo com as letras encontradas e deve englobar tanto as maiúsculas quanto as minúsculas. Como no arquivo de comparação elas estão sem espaço e em minúsculas, o texto do usuário sofre essas alterações.

• Verificação se o texto do usuário contém as palavras do arquivo de termos compostos (as que estiverem no texto são guardadas numa lista de palavras compostas).

• Apresentação da lista de palavras compostas para o usuário.

• Divisão do texto em uma lista de palavras, separando-as por ocorrência de espaços. Até

ISSN 2236-8884

então, o programa faz os procedimentos considerando o texto todo, sem divisão do texto em palavras.

• Colocação de cada palavra dessa nova lista em letras minúsculas.

• Comparação de cada palavra dessa nova lista com as palavras que estão no arquivo *stopwords* – caso haja alguma, ela será retirada do texto.

• Apresentação do texto do usuário sem as *stopwords*.

O botão "Carregar arquivo", quando clicado, faz com que o programa:

• Gere uma janela que servirá para o usuário escolher o caminho de algum arquivo que ele queira usar para identificar as palavras compostas e retirar as *stopwords*.

• Leia o endereço do arquivo, extraia as linhas do texto e coloque cada linha em uma posição de uma lista.

• Reúna todas as linhas com a finalidade de obter um único texto. A partir dessa etapa, todos os outros procedimentos são iguais aos feitos pelo botão Adicionar.

O *software* faz a identificação de palavras compostas, processando-as e retirando-as de acordo com alguns critérios preestabelecidos, como os nomes compostos referentes às cidades e os separados por hífen. O *software* aceita, no momento, poucos casos de palavras compostas, mas em sua construção, essas palavras serão divididas em módulos que irão organizá-las de acordo com a situação que faz com que elas sejam compostas numa frase, por exemplo: (i) "Rio de Janeiro": palavra composta referente a um nome próprio; (ii) "Guarda-chuva": palavra composta separada por hífen; (iii) "Armário de aço": palavras ligadas pela preposição "de";. (iv) "Casa amarela": palavra composta formada por um substantivo modificado por um adjetivo.

Como o *software* ainda está em desenvolvimento, suas funções são limitadas, assim como as comparações por ele feitas. A partir da base já construída, pretende-se aperfeiçoar os módulos para processar, mais eficazmente, palavras compostas.

ISSN 2236-8884

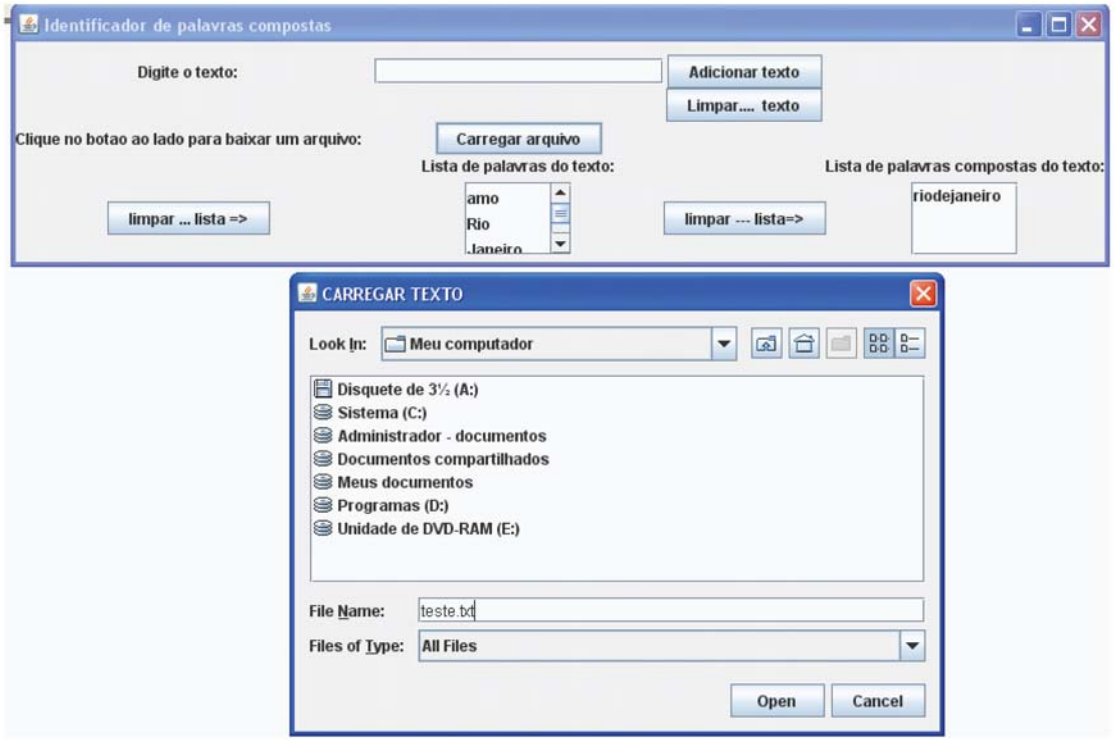

Figura 2: Ilustração do software em pleno funcionamento

Considerações Finais

Este artigo apresentou o programa Compostas_v1, um *software* de mineração de textos para o projeto "Mineração de Textos Eletrônicos" desenvolvido pelos autores no Núcleo de Informática na Educação do IF Fluminense (*campus* Campos-Centro).

A área de Descoberta de Conhecimento em Textos constitui um campo de estudo promissor e de larga aplicação em estudos que envolvam, por exemplo, a verificação de estados afetivos, da participação, e da opinião do usuário em ambientes digitais. Nesse sentido, o artigo buscou apresentar

ISSN 2236-8884

uma visão geral do que seja Mineração de Textos e de algumas questões de natureza linguística envolvidas no processamento automático de textos.

Pretende-se que, quando o Compostas_v1 for completado, ele possa ser integrado como uma ferramenta complementar dos programas Sobek e Eurekha de forma a atender aos objetivos dos pesquisadores. A ideia é que esses programas sejam implementados na plataforma Moodle como ferramentas auxiliares dos tutores na Educação a Distância. Dessa forma, os professores não precisarão analisar, de forma detalhada, os escritos de seus alunos, podendo fazer o diagnóstico de sua participação, desempenho ou eventuais dificuldades.

Referências

FELDMAN,, R.; SANGER, J. ***The Text Mining Handbook:*** Advanced Approaches in Analyzing Unstructured Data. Cambridge, MA: Cambridge University Press, 2007.

HEARST, M. ***What is Text Mining?***. 2003. Disponível em:<http://people.ischool.berkeley.edu/~hearst/text-mining.html>. Acesso em: jun. 2010.

LORENZATTI, A. ***SOBEK:*** uma Ferramenta de Mineração de Textos. Caxias do Sul/RS: UCS, 2007. Trabalho de Conclusão de Curso (Graduação).

OTHERO, G. A.; MENUZZI, S. M. Lingüística Computacional: teoria e prática. São Paulo: Parábola, 2005.

WIVES, L. K. ***Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de Clustering***. Dissertação (Mestrado) - UFRGS. Instituto de Informática. PPGC, Porto Alegre, 1999.