Pontifícia Universidade Católica
do Rio de Janeiro

Fábio de Azevedo Soares

Categorização Automática de Textos Baseada em Mineração de Textos

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo Programa de Pós-graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica da PUC-Rio.

Orientadora: Prof. Marley M. B. R. Vellasco

Co-Orientador: Prof. Emmanuel P. L. Passos

Rio de Janeiro

Junho de 2013

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Fábio de Azevedo Soares

Categorização Automática de Textos Baseada em Mineração de Textos

Tese apresentada como requisito parcial para obtenção do grau Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Profa. Marley Maria Bernardes Rebuzzi Vellasco
Orientadora
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Emmanuel Piseces Lopes Passos
Co-Orientador
Aposentado do IME

Profa. Karla Tereza Figueiredo Leite
UEZO

Prof. Rubens Nascimento Melo
Departamento de Informática -PUC-Rio

Prof. Ronaldo Ribeiro Goldschmidt
UFRRJ

Prof. Douglas Mota Dias
Departamento de Engenharia Elétrica - PUC-Rio

Prof. Cláudio Márcio do Nascimento Abreu Pereira
Comissão Nacional de Energia Nuclear

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico - PUC-Rio

Rio de Janeiro, 10 de Junho de 2013

Fábio de Azevedo Soares

Graduou-se Bacharel em Ciência da Computação em 2006. Mestre pela PUC-Rio em 2008. Atua como Desenvolvedor de Softwares, principalmente no desenvolvimento de Sistemas de Apoio à Decisão. Leciona para o ensino universitário. Tem interesse na pesquisa de novos algoritmos, principalmente, na área de Mineração de Textos e Aprendizado de Máquina.

Ficha Catalográfica

Soares, Fábio de Azevedo

Categorização automática de textos baseada em mineração de textos / Fábio de Azevedo Soares ; orientadores: Marley M. B. R. Vellasco, Emmanuel P. L. Passos. – 2013.

158 f. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2013.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Mineração de textos. 3. Categorização. 4. Framework. 5. Português brasileiro. 6. Automática. I. Vellasco, Marley M. B. R. II. Passos, Emmanuel P. L. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Agradecimentos

A Deus por ser e pela oportunidade de todo dia começar de novo.

A Gabriela Soares por dar um novo sentido a minha vida, ser a minha fonte de amor e esperança, ensinar mais do que posso aprender e pelo sorriso de cada dia.

Ao professor Emmanuel Passos pelo apoio, dedicação, paciência, amizade e inspiração de vida.

À professora Marley Vellasco pela confiança depositada, oportunidade de realizar este trabalho e exemplo de mestre.

Aos meus pais pelo carinho, preocupação e expectativa de conclusão deste trabalho.

Ao meu amigo Thiago Mendonça por incentivar e compreender mais do que precisava.

Ao CNPq e à CAPES pelo apoio financeiro.

À PUC-Rio e à Vice Reitoria Acadêmica (VRAc) pela bolsa de isenção que me foi concedida.

Resumo

Soares, Fábio de Azevedo; Vellasco, Marley Maria Bernardes Rebuzzi (Orientadora); Passos, Emmanuel Piseces Lopes Passos (Co-Orientador). **Categorização Automática de Textos Baseada em Mineração de Textos.** Rio de Janeiro, 2013. 158p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A Categorização de Documentos, uma das tarefas desempenhadas em Mineração de Textos, pode ser descrita como a obtenção de uma função que seja capaz de atribuir a um documento uma categoria a que ele pertença. O principal objetivo de se construir uma taxonomia de documentos é tornar mais fácil a obtenção de informação relevante. Porém, a implementação e a execução de um processo de Categorização de Documentos não é uma tarefa trivial: as ferramentas de Mineração de Textos estão em processo de amadurecimento e ainda, demandam elevado conhecimento técnico para a sua utilização. Além disso, exercendo grande importância em um processo de Mineração de Textos, a linguagem em que os documentos se encontram escritas deve ser tratada com as particularidades do idioma. Contudo há grande carência de ferramentas que forneçam tratamento adequado ao Português do Brasil. Dessa forma, os objetivos principais deste trabalho são pesquisar, propor, implementar e avaliar um *framework* de Mineração de Textos para a Categorização Automática de Documentos, capaz de auxiliar a execução do processo de descoberta de conhecimento e que ofereça processamento linguístico para o Português do Brasil.

Palavras-chave

Mineração de Textos; Categorização; Framework; Português brasileiro; Automática.

Abstract

Soares, Fábio de Azevedo; Vellasco, Marley Maria Bernardes Rebuzzi (Advisor); Passos, Emmanuel Piseces Lopes Passos (Co-Advisor). **Automatic Text Categorization Based on Text Mining.** Rio de Janeiro, 2013. 158p. Ph.D. Thesis - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Text Categorization, one of the tasks performed in Text Mining, can be described as the achievement of a function that is able to assign a document to the category, previously defined, to which it belongs. The main goal of building a taxonomy of documents is to make easier obtaining relevant information. However, the implementation and execution of Text Categorization is not a trivial task: Text Mining tools are under development and still require high technical expertise to be handled, also having great significance in a Text Mining process, the language of the documents should be treated with the peculiarities of each idiom. Yet there is great need for tools that provide proper handling to Portuguese of Brazil. Thus, the main aims of this work are to research, propose, implement and evaluate a Text Mining Framework for Automatic Text Categorization, capable of assisting the execution of knowledge discovery process and provides language processing for Brazilian Portuguese.

Keywords

Text Mining; Text Categorization; Framework; Brazilian Portuguese; Automatic.

Sumário

1 Introdução 16

1.1. Motivação 16

1.2. Objetivos do Trabalho 20

1.3. Trabalhos relacionados 21

1.4. Organização da Tese 28

2 Mineração de Textos: Fundamentos 30

2.1. Definição 30

2.2. Principais Elementos 34

2.3. Documentos textuais são estruturados 36

2.4. Características representativas de um documento 37

2.5. Abordagens ao processo de Mineração de Textos 40

2.5.1. Análise Estatística 40
2.5.2. Análise Semântica 41
2.6. Áreas correlatas a Mineração de Textos 42

2.6.1. Ciência Cognitiva 42
2.6.2. Processamento de Linguagem Natural 43
2.6.3. Aprendizado de Máquina 44
2.6.4. Estatística 45
2.6.5. Recuperação de Informação 46
2.6.6. Mineração de Dados 47

3 Metodologia de Mineração de Textos 49

3.1. Coleta de Dados 50

3.2. Pré-Processamento 51

3.2.1. Tokenização 52
3.2.2. Remoção de *stopwords* 54
3.2.3. Processamento de Linguagem Natural 55
3.3. Indexação 62

3.3.1. Indexação Textual 63
3.3.2. Indexação Temática 64
3.4. Mineração 65

3.5. Análise 66

3.5.1. Precisão 68

3.5.2. Abrangência 68
3.5.3. Medida-F 69
3.5.4. Precisão x Abrangência 69

4 Recuperação de Informação 71

4.1. Introdução 71

4.2. Histórico da área de Recuperação de Informação 73

4.2.1. 1ª Fase – Décadas de 50 e 60 73
4.2.2. 2ª Fase – Décadas de 70 e 80 73
4.2.3. 3ª Fase – Década de 90 em diante 74
4.3. Recuperação de Informação Clássica 75

4.3.1. Modelos de Representação de Documentos 77

5 Categorização de Textos 86

5.1. Introdução 86

5.2. Histórico da área de Categorização de Textos 87

5.2.1. 1ª Fase - Até o final década de 80 87
5.2.2. 2ª Fase - Década de 90 em diante 87
5.3. Definição 88

5.4. Tipos de Classificadores 89

5.5. Modelagem da categorização 90

5.6. Tipos de categorização 91

5.7. Aplicações de Categorização de Textos 92

5.7.1. Organização de documentos 92
5.7.2. Filtragem de Documentos 92
5.7.3. Desambiguação Lexical de Sentido 93
5.8. Aprendizagem de Máquina em CT 94

5.8.1. Aprendizagem Supervisionada 94
5.8.2. Treinamento e Teste 95
5.8.3. *k*-Nearest Neighbors 96
5.8.4. SVM 97
5.8.5. Combinação de Classificadores 100
5.9. Ferramentas de Mineração de Textos 103

5.9.1. Weka 103
5.9.2. Text Mine 104
5.9.3. TMSK 105
5.9.4. RIKTEXT 106
5.9.5. STATISCA Text Miner 106

6 Framework proposto 109

6.1. Definição 109

6.2. Ambiente de desenvolvimento 110

6.3. Objetivos 110

6.4. Coleta 111

6.5. Pré-Processamento 111

6.5.1. Tokenização 111
6.5.2. Análise/Remoção de *stopwords* 112
6.5.3. Processamento de Linguagem Natural 114
6.5.4. Redução de características 121
6.5.5. Indexação 122
6.5.6. Classificadores implementados 122
6.5.7. Técnicas de combinação de classificadores 123
6.6. Corpus 123

6.7. Assistência Inteligente 126

7 Estudos de Caso 129

7.1. Coleta 129

7.2. Treinamento 129

7.3. Resultados 129

7.3.1. Tokenização 129
7.3.1. Remoção de *stopwords* 132
7.3.2. PLN - Identificação de classes gramaticais 134
7.3.3. PLN - Lematização 135
7.3.4. Thesaurus 137
7.3.5. Seleção de características 138
7.3.6. Mineração 141

8 Conclusões e Trabalhos Futuros 147

8.1. Conclusões 147

8.2. Trabalhos Futuros 148

Referências Bibliográficas 150

Lista de Figuras

Figura 1 - Processo de obtenção de conhecimento 17
Figura 2 – Integridade semântica de um SGBD 32
Figura 3 – Sites brasileiros quanto à frequência de modificação do conteúdo 35
Figura 4 - Coleções de documentos com elementos em comum 35
Figura 5 – Algumas estruturas sintáticas de um trecho de texto 36
Figura 6 – Documentos em formatos fracamente estruturado e semiestruturado (respectivamente) 37
Figura 7 – Modelos de representação baseados em palavras e termos 40
Figura 8 – Multidisciplinaridade da Mineração de Textos 42
Figura 9 – Modelo simples de aprendizagem de máquina 45
Figura 10 – Linhas cronológica das etapas de um processo de Mineração de Textos (por Aranha) 49
Figura 11 – Processo de representação estruturada de um texto 52
Figura 12 - Metodologia de identificação de tokens proposta por KONCHADY 54
Figura 13 - Processo de tokenização seguido por remoção de stopwords 55
Figura 14 - Reconhecimento de anáfora com informações do contexto 57
Figura 15 – Erros de um processo de stemming: overstemming e understemming 58
Figura 16 – Derivações de um mesmo radical identificadas pelo algoritmo de Porter 59
Figura 17 - Representação de um índice invertido 64
Figura 18 - Estrutura básica de um Dicionário Thesaurus 65
Figura 20 – Gráfico de compensação entre precisão e abrangência 70
Figura 21 - Sistema Clássico de Recuperação de Informação 75
Figura 22 – Etapas possíveis no processo de Indexação de documentos textuais 77
Figura 23 - Representação vetorial do documento Di no espaço n-dimensional (*n =2*) 79
Figura 24 – Algoritmo KNN - Seleção baseada nos *k* (*= 3)* vizinhos 96

Figura 25 – Máquina de Vetores de Suporte 98
Figura 26 – Abordagens SVM para problemas não binários 99
Figura 27 – Parâmetros de configuração do filtro *StringToWordVector* do software Weka 104
Figura 28 – Interface de coleta do software STATISCA 108
Figura 29 – Classes gramaticais segundo a NGB 115
Figura 30 - Estrutura do dicionário Thesaurus utilizado no Sistema de MT 120
Figura 31 - Exemplo de documento do corpus CETENFolha 125
Figura 32 - Exemplo de modelagem de execução para o *k*-NN 127
Figura 33 - Exemplo de documento do corpus CETENFolha 131
Figura 34 - Exemplo de documento do corpus CETENFolha 131
Figura 35 - Representação de documentos na forma de *bag of words* 132
Figura 36 - Resultado do processo de remoção de *stopwords* baseado em listas 133
Figura 37 - Resultado do processo de remoção de *stopwords* do domínio 134
Figura 38 - Identificação de classes gramaticais 135
Figura 40 - Fluxograma da lematização não verbal 137
Figura 41 - Substituição de termos por consulta ao Thesaurus 138
Figura 42 - Seleção de características 140
Figura 43 - Subconjuntos disponíveis 143
Figura 44 - Diagrama de estados 143
Figura 45 - Solução de melhor desempenho 145
Figura 46 - Desempenho lematização 146

Lista de Tabelas

Tabela 1 - Resumo comparativo dos trabalhos relacionados quanto ao corpus, modelo de representação dos documentos e atribuição de pesos utilizados 27
Tabela 2 - Resumo comparativo dos trabalhos relacionados quanto a técnica de PLN, tarefa de MT e modelos de classificadores utilizados 28
Tabela 3 - As duas abordagens para a Análise de Textos e suas principais Áreas de Conhecimento 41
Tabela 4 - As principais características de cada uma das abordagens para a Análise de Textos 41
Tabela 5 - Marcação de *tags* para Reconhecimento de Entidades Nomeadas 60
Tabela 6 - Exemplo de classificações distintas de uma mesma entidade 61
Tabela 7 – Visualização das regras para concessão de empréstimos em uma tabela 67
Tabela 8 - Comparação entre Recuperação de Dados x Recuperação de Informação 72
Tabela 9 - Resultados conflitantes de SVMs binárias 99
Tabela 10 - Lista de cem *stopwords* utilizadas na etapa de Pré-processamento 113
Tabela 11 - Exemplos ambíguos de identificação de classes gramaticais 116
Tabela 12 - Modelagem dos dados baseada em *sliding window* 117
Tabela 13 - Informações adicionais sobre o CETENFolha 124
Tabela 14 - Planejamento de ações 127
Tabela 15 - Planejamento de ações: Tokenização 130
Tabela 16 - Planejamento de ações: Remoção de *stopwords* 132
Tabela 17 - Planejamento de ações: PLN - Identificação de classes gramaticais 134
Tabela 18- Planejamento de ações: PLN - Lematização 135
Tabela 19- Planejamento de ações: PLN - Thesaurus 137
Tabela 20- Planejamento de ações: Seleção de características 139
Tabela 21- Planejamento de ações: Mineração 141

Tabela 22 - Configurações de execução do algoritmo SVM 141
Tabela 23 - Configurações de execução do algoritmo KNN 141
Tabela 24 - Configuração do melhor resultado obtido 144
Tabela 25 - Relação categoria x classificador 145

Lista de Equações

Equação 1 - Teorema de Bayes 46
Equação 2 - Fórmula da métrica de desempenho “Precisão” 68
Equação 3 - Fórmula da métrica de desempenho “Abrangência” 68
Equação 4 - Fórmula da métrica de desempenho "Medida-F" 69
Equação 5 - Cálculo da medida TF em um documento 80
Equação 6 - Cálculo da medida TF-IDF em um documento 81
Equação 7 - Cálculo do escore de relevância de um termo 81
Equação 8 - Cálculo do Coeficiente de Correlação 82
Equação 9 - Cálculo do Ganho de Informação 84
Equação 10 - Cálculo de similaridade entre documentos por meio do cosseno 85

Lista de Siglas

CT	Categorização de Textos
ETL	Extract Transform Load
IA	Inteligência Artificial
KDD	Knowledge Discovery in Databases
KDT	Knowledge Discovery in Texts
k-NN	k-Nearest Neighbors
NILC	Núcleo Interinstitucional de Linguística Computacional
PLN	Processamento de Linguagem Natural
RN	Redes Neurais
SGBD	Sistema Gerenciador de Bancos de Dados
SVM	Support Vector Machine
VISL	Visual Interactive Syntax Learning