Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Spatio-Temporal Trend Analysis of the Brazilian Elections based on Twitter Data

Bruno Justino Garcia Praciano

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. -Ing João Paulo Lustosa da Costa

Brasília
2014

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Spatio-Temporal Trend Analysis of the Brazilian Elections based on Twitter Data

Bruno Justino Garcia Praciano

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. -Ing João Paulo Lustosa da Costa (Orientador)
ENE/UnB

Prof. Dr. Rafael Timóteo de Sousa Júnior
ENE/UnB

Dr. -Ing Ricardo Kerhle Miranda
ENM/UnB

Prof. Dr. José Edil Guimarães de Medeiros
Coordenador do Curso de Engenharia da Computação

Brasília, 24 de dezembro de 2014

Dedicatória

Na *dedicatória* o autor presta homenagem a alguma pessoa (ou grupo de pessoas) que têm significado especial na vida pessoal ou profissional. Por exemplo (e citando o poeta): *Eu dedico essa música a primeira garota que tá sentada ali na fila. Brigado!*

Agradecimentos

Nos *agradecimentos*, o autor se dirige a pessoas ou instituições que contribuíram para elaboração do trabalho apresentado. Por exemplo: *Agradeço aos gigantes cujos ombros me permitiram enxergar mais longe. E a Google e Wikipédia.*

Resumo

O *resumo* é um texto inaugural para quem quer conhecer o trabalho, deve conter uma breve descrição de todo o trabalho (apenas um parágrafo). Portanto, só deve ser escrito após o texto estar pronto. Não é uma coletânea de frases recortadas do trabalho, mas uma apresentação concisa dos pontos relevantes, de modo que o leitor tenha uma ideia completa do que lhe espera. Uma sugestão é que seja composto por quatro pontos: 1) o que está sendo proposto, 2) qual o mérito da proposta, 3) como a proposta foi avaliada/validada, 4) quais as possibilidades para trabalhos futuros. É seguido de (geralmente) três palavras-chave que devem indicar claramente a que se refere o seu trabalho. Por exemplo: *Este trabalho apresenta informações úteis a produção de trabalhos científicos para descrever e exemplificar como utililzar a classe LaTeX do Departamento de Ciência da Computação da Universidade de Brasília para gerar documentos. A classe `UnB-CIC` define um padrão de formato para textos do CIC, facilitando a geração de textos e permitindo que os autores foquem apenas no conteúdo. O formato foi aprovado pelos professores do Departamento e utilizado para gerar este documento. Melhorias futuras incluem manutenção contínua da classe e aprimoramento do texto explicativo.*

Palavras-chave: Big Data, Aprendizado de Máquina Supervisionado, Análise de Sentimentos, Máquina de Vetor de Suporte

Abstract

O *abstract* é o resumo feito na língua Inglesa. Embora o conteúdo apresentado deva ser o mesmo, este texto não deve ser a tradução literal de cada palavra ou frase do resumo, muito menos feito em um tradutor automático. É uma língua diferente e o texto deveria ser escrito de acordo com suas nuances (aproveite para ler `http://dx.doi.org/10.6061%2Fclinics%2F2014(03)01`). Por exemplo: *This work presents useful information on how to create a scientific text to describe and provide examples of how to use the Computer Science Department's LaTeX class. The `UnB-CIC` class defines a standard format for texts, simplifying the process of generating CIC documents and enabling authors to focus only on content. The standard was approved by the Department's professors and used to create this document. Future work includes continued support for the class and improvements on the explanatory text.*

Keywords: Big Data, Supervised Machine Learning, Sentiment Analysis, Support Vector Machine

Contents

Chapter 1

Introduction

Este documento serve de exemplo da utilização da classe `UnB-CIC` para escrever um texto cujo objetivo é apresentar os resultados de um trabalho científico. A sequência de ideias apresentada deve fluir claramente, de modo que o leitor consiga compreender os principais conceitos e resultados apresentados, bem como encontrar informações sobre conceitos secundários.

1.1 Motivation

1.2 Problems

1.3 Objectives

1.4 Outline

1.5 Related work

1.6 Chapters description

Chapter 2

Machine Learning

2.1 Basic Concepts

2.1.1 Training and testing phases

2.1.2 Learning of paradigms

2.1.3 Performance measures

2.2 Support Vector Machine

2.3 Naive Bayes

2.4 Decision Trees

2.5 Logistic Regression

Chapter 3

Text Mining

3.1 Data Colection

3.1.1 Social Networks

3.2 Natural Language Processing

3.3 Pre-Processing

3.4 Tokenization

3.5 Bag of Words

3.6 Term Frequency

3.7 Term Frequency - Inverse Document Frequency (TF-IDF)

3.8 Stemming

3.9 Stop Words

3.10 Sentiment Analysis

Chapter 4

Methodology

4.1 Crawling and Tweet Extraction

4.2 Data Pre-Processing

4.3 Lexical Dictionary

4.4 Sentiment Classification

4.5 Data Visualization

Chapter 5

Results

5.1 Perfomance evaluation of trend analysis

5.2 N-Fold Cross Validation

5.3 Error evaluation of the sentiment analysis via SVM

5.4 Spatio Trend Analysis

5.5 Election Results

Chapter 6

Conclusion

Este documento serve de exemplo da utilização da classe `UnB-CIC` para escrever um texto cujo objetivo é apresentar os resultados de um trabalho científico. A sequência de ideias apresentada deve fluir claramente, de modo que o leitor consiga compreender os principais conceitos e resultados apresentados, bem como encontrar informações sobre conceitos secundários.

Referências

Appendix A

A.1 Appendix

Anexo I

Documentação Original UnB-CIC (parcial)

```
% -*- mode: LaTeX; coding: utf-8; -*-
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% File    : unb-cic.cls (LaTeX2e class file)
%% Authors : Flávio Maico Vaz da Costa
%%               (based on previous versions by José Carlos L. Ralha)
%% Version : 0.96
%% Updates : 0.5  [??/11/2004] - Initial release. don't remember the day.
%%         : 0.75 [04/04/2005] - Fixed font problems, UnB logo
%%                               resolution, keywords and palavras-chave
%%                               hyphenation and generation problems,
%%                               and a few other problems.
%%         : 0.8  [08/01/2006] - Corrigido o problema causado por
%%                               bancas com quatro membros. O quarto
%%                               membro agora é OPCIONAL.
%%                               Foi criado um novo comando chamado
%%                               bibliografia. Esse comando tem dois
%%                               argumentos onde o primeiro especifica
%%                               o nome do arquivo de referencias
%%                               bibliograficas e o segundo argumento
%%                               especifica o formato. Como efeito
%%                               colateral, as referências aparecem no
%%                               sumário.
%%         : 0.9 [02/03/2008]  - Reformulação total, com nova estrutura
%%                               de opções, comandos e ambientes, adequação
%%                               do logo da UnB às normas da universidade,
%%                               inúmeras melhorias tipográficas,
```

```
%%                              aprimoramento da integração com hyperref,
%%                              melhor tratamento de erros nos comandos,
%%                              documentação e limpeza do código da classe.
%%          : 0.91 [10/05/2008] - Suporte ao XeLaTeX, aprimorado suporte para
%%                              glossaries.sty, novos comandos \capa, \CDU
%%                              e \subtitle, ajustes de margem para opções
%%                              hyperref/impressao.
%%          : 0.92 [26/05/2008] - Melhora do ambiente {definition}, suporte
%%                              a hypcap, novos comandos \fontelogo e
%%                              \slashedzero, suporte [10pt, 11pt, 12pt].
%%                              Corrigido bug de seções de apêndice quando
%%                              usando \hypersetup{bookmarksnumbered=true}.
%%          : 0.93 [09/06/2008] - Correção na contagem de páginas, valores
%%                              load e config para opção hyperref, comandos
%%                              \ifhyperref e \SetTableFigures, melhor
%%                              formatação do quadrado CIP.
%%          : 0.94 [17/04/2014] - Inclusão da opção mpca.
%%          : 0.95 [06/06/2014] - Remoção da opção "mpca", inclusão das opções
%%                              "doutorado", "ppginf", e "ppca" para identificar
%%                              o programa de pós-graduação. Troca do teste
%%                              @mestrado por @posgraduacao.
%%          : 0.96 [24/06/2014] - Ajuste do nome do curso/nome do programa.
%%
```