NOVA IMS Information Management School

Text Mining

Roberto Henriques
Filipa Peleja

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Introduction

- Data mining aims to extract and identify useful information from large databases
- Web mining aims to understand customer behaviour, web site's effectiveness, quantify a marketing campaign success, etc.
- Text mining looks at documents, e-communications, e-commerce activities etc.

"Text mining or text analytics are broad umbrella terms describing a range of technologies for analyzing and processing semistructured and unstructured text data." **(Delen et al. 2012)[1]**

[1] Delen et al., Practical text mining and statistics for non-structured text data applications, pp.31, 2012.

Objectives

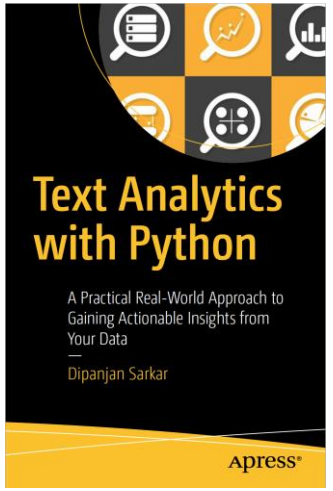
- By the end of the course, the student should:
  - Understand different types of text mining techniques.
  - Understand how to process semi-structured and unstructured text.
  - Understand text classification, text summarization and text similarity concepts and techniques
  - Understand semantic and sentiment analysis concepts and techniques
  - Demonstrate capacity to perform a practical work that requires the application of text mining techniques.
  - Be proficient with text mining and sentiment analysis python libraries

Course Unit Content

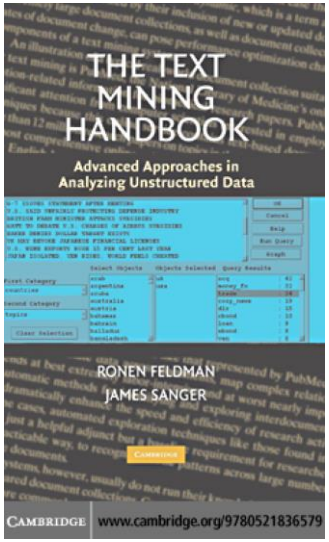
- CUC1. Introduction to Text Mining
- CUC2. Natural Language Basics
- CUC3. Processing and understanding text
- CUC4. Text Classification
- CUC5. Text Summarization
- CUC6. Text Similarity and Clustering
- CUC7. Semantic and Sentiment Analysis

Bibliography

- Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data, By Dipanjan Sarkar



- The Text Mining Handbook, By Ronen Feldman and James Sanger



Assessment methods

- 1st call
  - Exam (50%)
  - Project 1 (25%)
  - Project 2 (25%)
- 2nd call
  - Exam (50%)
  - Project 1 (25%)
  - Project 2 (25%)
- *Minimum grade of 8.0 (in 20) for the exam*
- *Minimum grade of 5.0 (in 20) for the projects*

NOVA IMS Information Management School

Program

Week	Date	Topics
1	16-Feb-18	Course presentation; Introduction to Text Mining
2	23-Mar-18	Natural Language Basics
3	2-Mar-18	Processing and Understanding Text; Text Tokenization and Text normalization
4	9-Mar-18	Understanding Text Syntax and Structure
5	16-Mar-18	Text Classification; Text Normalization
6	23-Mar-18	Feature Extraction; Classification Algorithms
7	6-Apr-18	Text Summarization; Text Normalization; Feature Extraction; Topic Modelling
8	13-Apr-18	Text Similarity and Clustering: Text Similarity; Term Similarity
9	20-Apr-18	Text Similarity and Clustering: Document Similarity
10	27-May-18	Semantic and Sentiment Analysis: Semantic Analysis; Word Sense Disambiguation
11	4-May-18	Semantic and Sentiment Analysis: Named Entity Recognition
12	11-May-18	Sentiment Analysis
13	18-May-18	Sentiment Analysis/Project support
14	25-May-18	Project support