ESCUELA TÉCNICA SUPERIOR DE
INGENIERÍA

DTC-MBD-522 Unstructured Information

SEMESTER:	Spring
CREDITS:	6 ECTS (4 hrs. per week)
LANGUAGE:	Spanish/English
DEGREES:	Master in Big Data Technologies and Advanced Analytics

Course overview

This course examines how to extract information from the huge variety of documents of the Internet. To do so, the course is divided in two main blocks. The first one is centered in extracting information from a web page, filtering the unwanted labels and getting just the relevant information to be stored in a database. The second part if focused in the social media and the knowledge extraction from small comments, like tweets. The course will be complemented with practical exercises to understand the theoretical concepts.

Prerequisites

Basis of data filtering.

Basic knowledge of Programming in R and Python is required for the practice sessions.

Course contents

Theory:

1. Introduction

2. Text and voice mining

   2.1. Natural language processing

      2.1.1. Morphologycal analyzer

      2.1.2. Sintactycal analyzer

      2.1.3. Semantic analyzer

   2.2. Introduction and problems of knowledge extraction of unstructured information

   2.3. Web mining

      2.3.1. Data extraction and Web processing. Text Mining

         2.3.1.1. Concept Classifier

         2.3.1.2. Text Classifier

      2.3.2. Web content mining

         2.3.2.1. Extraction of information, summaries, Q/A systems.

         2.3.2.2. XML Mining)

      2.3.3. Mining the structure of the Web

      2.3.4. Web usage mining

         2.3.4.1. Navigation patterns

         2.3.4.2. Customized contents

   2.4. Voice mining

   2.5. Business Solutions

3. Social Networks analysis

   3.1. Opinion and sentiment analysis

   3.2. Classification and extraction (polarity of sentiment and degrees of positivity, detection of subjectivity and identification of opinion.

   3.3. Terminological analysis (presence of terms vs. frequency, parts of speech, syntax, negations, topics)

   3.4. Applications in different domains, impact and implications

   3.5. Business Solutions

Textbook

- **Bird S., Klein E., Loper, E.**, (2009). *Natural Language Processing with Python.* O'Reylly.
- **Manning, R., Schutze, H.** (1999). *Foundations of statistical and natural language processing.* MIT Press

Grading

The following conditions must be accomplished to pass the course:

- A minimum overall grade of at least 5 over 10.
- A minimum grade in all lab projects and in the final exam of 4 over 10.

The overall grade is obtained as follows:

- Final exam accounts for 35% of the final grade if the grade in this exam is at least 4. In other case, final exam accounts for 100 % of the overall grade.
- Mid-term exam accounts for 15%.
- Practical assignment accounts for 50% of the final grade.

This document is a brief outline of the course and does not replace the official program of study