Text Mining

Academic Year: (2024 / 2025) **Review date: 30-04-2024**

Department assigned to the subject: Statistics Department

Coordinating teacher: UCAR MARQUES, IÑAKI

Type: Compulsory ECTS Credits : 3.0

Year : 1 Semester : 2

REQUIREMENTS (SUBJECTS THAT ARE ASSUMED TO BE KNOWN)

Data Programming (19138)

OBJECTIVES

- Knowledge of text mining structures and procedures.
- Ability to use basic methods for extracting information from textual data.
- Ability to apply processing techniques to prepare documents for statistical modeling.
- Ability to evaluate and use basic predictive models of textual information.

DESCRIPTION OF CONTENTS: PROGRAMME

1. Theoretical introduction to Natural Language Processing
1.1. Brief history of computational linguistics and main developments
1.2. What is Natural Language Processing and its role in Artificial Intelligence
1.3. Structure of a basic NLP pipeline
1.4. Most common tasks and applications in the industry
1.5. Current importance in the digital society, main initiatives

2. Practical introduction to automatic language analysis with R
2.1. Source text import, dataset design and creation of data structures
2.2. Text cleaning, removal of stopwords and symbols, missing values and duplicates
2.3. Splitting and tokenization processes
2.4. Basic analysis: word count, n-gram extraction, frequency tables
2.5. Intermediate analysis: distinctiveness analysis, tf-idf, bag of words

3. Introduction to sentiment analysis
3.1. What is automatic sentiment analysis in a text: opinion, emotion and intention of the speaker
3.2. Real-world cases of sentiment analysis in the industry and limitations
3.3. Practical training on automatic sentiment analysis: use of lexicons and dictionaries, automatic sentiment mapping, segmentation, word clouds
3.4. Creation of sentiment analysis graphs and reports

4. Introduction to topic modeling
4.1. What is topic modeling, main uses in the industry
4.2. Classifying text into categories: supervised and unsupervised methods
4.3. Practical training in topic modelling: word and topic association, natural group identification and characterization, common terms and overlapping
4.4. Creation of topic modeling graphs and reports for identification of representative ideas

5. Language models
5.1. What are pre-trained language models and their impact on NLP and Machine Learning development
5.2. Uses and implications in the industry and current status, main initiatives
5.3. Practical training on the use and evaluation of basic predictive models with text data

LEARNING ACTIVITIES AND METHODOLOGY

Training Activities:

- Theoretical-practical classes
- Tutorials
- Individual student work
- Partial and final examinations

Teaching Methods:

- Presentations in the professor's lecture room with computer and audiovisual support, in which the main concepts of the subject are developed and a bibliography is provided to complement the students' learning.
- Resolution of practical cases, problems, etc. raised by the professor, either individually or in a group.
- Presentation and discussion in class, under the moderation of the professor, of topics related to the content of the subject, as well as practical case studies.
- Developing pieces of work and reports, individually or in group.

ASSESSMENT SYSTEM

% end-of-term-examination:	40
% of continuous assessment (assigments, laboratory, practicals…):	60

- Participation in the class (10%)
- Individual or group work done during the course (50%)
- Final exam (40%)

In the extraordinary call, the evaluation system will be as follows:
1) Exam: 100%

BASIC BIBLIOGRAPHY

- Gabe Ignatow and Rada F. Mihalcea An Introduction to Text Mining: Research Design, Data Collection, and Analysis., SAGE Publications, 2017

- Silge, J., & Robinson, D. Text mining with R: A tidy approach, O'Reilly Media, 2017

ADDITIONAL BIBLIOGRAPHY

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed.), Stanford University, 2021

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed.), PEARSON, Prentice Hall, 2021

- Kumar, A., & Paul, A. Mastering text mining with r: Master text-taming techniques and build effective text-processing applications with R, Packt Publishing Limited, 2016

- Kwartler, T. Text mining in practice with R, Winley, 2017

- Marchette, D. J. Text data mining using R, Chapman & Hall Crc, 2018

- Ted Kwartler Text Mining in Practice with R, Wiley, 2017

BASIC ELECTRONIC RESOURCES

- Dan Jurafsky and James H. Martin . Speech and Language Processing (3rd ed.): http://https://web.stanford.edu/~jurafsky/slp3

- Julia Silge and David Robinson . Tex Mining with R: http://https://www.tidytextmining.com/