Natural Language Processing

Academic Year: (2022 / 2023)   **Review date: 20-05-2022**

Department assigned to the subject:

Coordinating teacher: ARENAS GARCIA, JERONIMO

Type: Electives ECTS Credits : 3.0

Year : 1 Semester : 2

REQUIREMENTS (SUBJECTS THAT ARE ASSUMED TO BE KNOWN)

* It is recommended to have passed the Machine Learning subject
* The Deep Learning subject also provides competences of interest, although it is not essential to have taken it. During the first sessions of the course, the necessary concepts for neural-based word and document embeddings will be reviewed.

OBJECTIVES

* Familiarize students with some commonly used methods for natural language processing, both for preprocessing unstructured text, and for building models based on machine learning
* Know various approaches for calculating semantic similarity between documents and their use to build and analyze semantic graphs
* Presentation of some tools for the interactive visualization of machine learning models and natural language processing based on graphs and interactive dashboards
* Familiarize students with some relevant applications of natural language processing
* Encourage maturity in the knowledge of these technologies, and the autonomy to deepen the concepts explained in class, by working on a final group project

DESCRIPTION OF CONTENTS: PROGRAMME

1. NLP Introduction
2. Word and document representation
  2.1. One-hot encoding
  2.2. Word Embeddings. Word2Vec. GloVe
  2.3. Other Embedding representations
3. Text preprocessing
  3.1. Corpus Acquisition & Document Parsing
  3.2. NLP pipelines
  3.3. Text homogeneization and cleaning. WordNet
  3.4. Named Entity Recognition
  3.5. Spacy and Spark NLP
4. Topic Modeling
  4.1. Latent Semantic Indexing
  4.2. Latent Dirichlet Allocation
  4.3. Visualization. Enriched BI Dashboards
5. Transformers
  5.1. Introduction to Transformers. Hugging Face
  5.2. Text Classification: Sentiment Analysis
  5.3. Zero-shot classification
  5.4. Text Generation
  5.5. Neural Machine Translation
  5.6. Question & Answering
6. Semantic Analysis
  6.1. Semantic Similarity Metrics
  6.2. Semantic Graphs
  6.3. Graph Analysis
  6.4. Graph Visualization
  6.5. Semantic Information Retrieval

LEARNING ACTIVITIES AND METHODOLOGY

The following learning activities and methodologies are employed:

- Combined master and lab clases: Master classes provide an overview of the main theoretical & mathematical concepts of natural language processing along with the analytic tools. In these classes, lab examples will be introduced as part of the theoretical expositions: all the formative sessions (lab availability provided) will take place in the lab to imbricate practical examples within the explanations to add dynamism to the class. This is also beneficial to solve different background issues.

- Final Project: Students will work on a project in which they will program a complete modular system of one of the tools explained in class. The students will be provided with some guidelines and some preparatory sessions by using problem-based learning.

Teachers are available during 2 hours per week for office hours.

ASSESSMENT SYSTEM

First call:
- Class Participation & Assignments (30%)
- Comprehension tests (30%)
- Final project (40%)

Second call:
- Final Exam (60%)
- Final project (40%)

% end-of-term-examination:	0
% of continuous assessment (assigments, laboratory, practicals…):	100

BASIC BIBLIOGRAPHY

 - Aston Zhang, Zachary C. Lipton, Mu Li, Alexander J. Smola Dive into Deep Learning, https://d2l.ai, 2020

 - Christopher D. Manning, Hinrich Schütze Foundations of Statistical Natural Language Processing, MIT Press, 1999

 - Dan Jurafsky and James H. Martin Speech and Language Processing, Prentice Hall, 2018

 - Li Deng (Editor), Yang Liu (Editor) Deep Learning in Natural Language Processing, Springer, 2018