Course guide
340459 - PEDT - Text Data Processing and Mining

Last modified: 17/05/2023

Unit in charge: Vilanova i la Geltrú School of Engineering
Teaching unit: 723 - CS - Department of Computer Science.

Degree: BACHELOR'S DEGREE IN INFORMATICS ENGINEERING (Syllabus 2018). (Optional subject).

Academic year: 2023 **ECTS Credits:** 6.0 **Languages:** Catalan

LECTURER

Coordinating lecturer: Neus Català Roig

Others:

PRIOR SKILLS

- To know and use comfortably basic concepts of linear algebra, discrete mathematics, probability and statistics.
- To know basic concepts of programming in Python.
- To know basic concepts studied in the subjects of MIDA and REIN.

REQUIREMENTS

To have passed ESTA, AMEP and DABD courses or at least being enrolled.
It is recommended to have taken MIDA and REIN.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
I_CECO7. CEC07. Ability to learn and develop techniques of computing learning and design and implement applications and systems which use them, including those dedicated to automatic information and knowledge extraction from large data volums.
I_CECO1. CEC01. Ability to have a thorough understanding of the fundamental principles and models of computation, ability to apply the principles to interpret, select, evaluate, model, and create new concepts, theories, applications and advance the technological development related to computing.
I_CECO4. CECO4. Ability to learn basics, paradigms and techniques of intelligent systems and analyze, design and build systems, services and computing applications that use these techniques in any scope.
I_CEIS6. CEIS6. Ability to design appropriate solutions in one or more application domains using software engineering methods that integrate ethical, social, legal and economic aspects.
I_CEIS4. CEIS4. Ability to identify and analyze problems and design, develop, deploy, test and document software solutions based on an adequate knowledge of theories, models and techniques.

TEACHING METHODOLOGY

The methodological approach consists of:
- 2 hours per week of lecture classes in which the teacher presents subject matter to students (theory lectures and problem-solving sessions),
- 2 hours per week in the computer classroom, in which students will do the work specified in the script with the guidance of the teacher.

LEARNING OBJECTIVES OF THE SUBJECT

Textual data is found everywhere, for example in books, articles, laws, financial analysis, medical records, social networks, etc. It is estimated that they represent between 80% and 90% of the data stored. In order to extract, summarize and analyze information from large volumes of textual data, specific methods are required. The field known as Text Mining uses computational techniques to extract information from textual data automatically.

The course covers the basic components of Natural Language Processing (NLP) and how they are used in Text Mining tasks. Applications such as Document Classification, Sentiment Analysis (or Opinion Mining) and Information Extraction are also studied.

STUDY LOAD

Type	Hours	Percentage
Hours small group	30,0	50.00
Hours large group	30,0	50.00

Total learning time: 60 h

CONTENTS

1. Processes for obtaining text data

Description:
To obtain or build a corpus. Creation of a corpus with data extracted from different sources: emails, Wikipedia articles, financial reports, literary works or websites of interest. Scrapping o web crawling.

Related activities:
LABORATORY
QUIZZES
PROJECT

Full-or-part-time: 6h 30m
Theory classes: 4h
Laboratory classes: 2h 30m

2. Preprocessing of text data

Description:
Simple syntactic processing: text clean-up, normalization and tokenization.
Advanced linguistic processing: Word Sense Disambiguation (WSD) and Part-of-Speech (PoS) tagging.

Related activities:
LABORATORY
QUIZZES
PROJECT

Full-or-part-time: 7h 30m
Theory classes: 4h
Laboratory classes: 3h 30m

3. Natural Language Processing: major tasks and applications

Description:
Introduction to NLP. Main tasks: Part-of-speech tagging, syntactic analysis and semantic interpretation.
Some applications (demos): document classification, document clustering, sentiment analysis, information extraction, automatic summarization, machine translation.

Related activities:
LABORATORY
QUIZZES
PROJECT

Full-or-part-time: 13h
Theory classes: 6h
Laboratory classes: 7h

4. NLP tools and datasets for text mining

Description:
NLP tools in Python: Scikit-Learn, Natural Language Toolkit (NLTK), Gensim, spaCy, NetworkX.
Datasets for text mining accessible on-line.

Related activities:
LABORATORY
QUIZZES
PROJECT

Full-or-part-time: 12h
Theory classes: 6h
Laboratory classes: 6h

5. Introduction to neural networks and Deep Learning applied to text mining

Description:
Text vectorization: bag-of-words, tf-idf, word embeddings.
Neural networks. Applications of Deep Learning in Text Mining.

Related activities:
LABORATORY
QUIZZES
PROJECT

Full-or-part-time: 12h
Theory classes: 6h
Laboratory classes: 6h

6. Project presentations

Description:
Presentations of student projects.

Related activities:
PROJECT

Full-or-part-time: 9h
Theory classes: 4h
Laboratory classes: 5h

GRADING SYSTEM

- Evaluation of the activities carried out during the laboratory sessions: 60%
- Realization and public presentation of a work of analysis on one of the topics studied in the course.: 30%
- Quizzes: 10%

Since 100% of the subject is evaluated through practical work, there is no overall final control and no reevaluation control in the form of a written examination.

EXAMINATION RULES.

Reports of laboratory sessions will be delivered online within a time limit for each session.
Quizzes are carried out in-person and are individual.

BIBLIOGRAPHY

Basic:
- Ignatow, Gabe; Mihalcea, Rada F. An Introduction to text mining : research design, data collection, and analysis. Thousand Oaks, California: SAGE Publications, Inc, 2018. ISBN 9781506337005.
- Jurafsky, Dan; Martin, James H. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition [on line]. 3rd ed. Upper Saddle River: Els autors, 2019 [Consultation: 28/04/2022]. Available on: https://web.stanford.edu/~jurafsky/slp3/.