340459 - PEDT - Text Data Processing and Mining

Coordinating unit: 340 - EPSEVG - Vilanova i la Geltrú School of Engineering
Teaching unit: 723 - CS - Department of Computer Science
Academic year: 2019
Degree: BACHELOR'S DEGREE IN INFORMATICS ENGINEERING (Syllabus 2018). (Teaching unit Optional)
ECTS credits: 6
Teaching languages: Catalan

Teaching staff
Coordinator: Neus Català Roig

Opening hours
Timetable: See the current office hours in the EPSEVG people list:
https://web3.epsevg.upc.edu/coneix-lepsevg/directori-epsevg

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. CEC07. 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 volumes.
I. CEC01. 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. CEC04. CEC04. 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

<table>
<thead>
<tr>
<th>Study load</th>
<th>Hours large group:</th>
<th>Hours small group:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total learning time: 60h</td>
<td>30h</td>
<td>50.00%</td>
</tr>
<tr>
<td></td>
<td>30h</td>
<td>50.00%</td>
</tr>
</tbody>
</table>
# 340459 - PEDT - Text Data Processing and Mining

## Content

| 1. Processes for obtaining text data | **Learning time**: 6h 30m  
Theory classes: 4h  
Laboratory classes: 2h 30m |
|--------------------------------------|---------------------------------------------------------------|

**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. Scarping or web crawling.

**Related activities:**
LABORATORY  
QUIZZES  
PROJECT

| 2. Preprocessing of text data | **Learning time**: 7h 30m  
Theory classes: 4h  
Laboratory classes: 3h 30m |
|--------------------------------|---------------------------------------------------------------|

**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

| 3. Natural Language Processing: major tasks and applications | **Learning time**: 13h  
Theory classes: 6h  
Laboratory classes: 7h |
|-------------------------------------------------------------|---------------------------------------------------------------|

**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
4. NLP tools and datasets for text mining | Learning time: 12h  
Theory classes: 6h  
Laboratory classes: 6h  

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  

5. Introduction to neural networks and Deep Learning applied to text mining | Learning time: 12h  
Theory classes: 6h  
Laboratory classes: 6h  

Description:  
Text vectorization: bag-of-words, tf-idf, word embeddings.  

Related activities:  
LABORATORY  
QUIZZES  
PROJECT  

6. Project presentations | Learning time: 9h  
Theory classes: 4h  
Laboratory classes: 5h  

Description:  
Presentations of student projects.  

Related activities:  
PROJECT  

Qualification 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.
340459 - PEDT - Text Data Processing and Mining

Regulations for carrying out activities

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:
