VNIVERSITAT DE VALÈNCIA

Course Guide
36429 Natural language processing

COURSE DATA

Data Subject	
Code	36429
Name	Natural language processing
Cycle	Grade
ECTS Credits	6.0
Academic year	2020 - 2021

Study (s)

Degree	Center	Acad. year	Period
1400 - Degree in Computer Engineering	School of Engineering	4	Second term
1406 - Degree in Data Science	School of Engineering	3	Second term

Subject-matter

Degree	Subject-matter	Character
1400 - Degree in Computer Engineering	16 - Optional subject	Optional
1406 - Degree in Data Science	9 - Machine Learning and Data Mining	Obligatory

Coordination

Name	Department
VILA FRANCES, JOAN	242 - Electronic Engineering

SUMMARY

Most of the current data available for analysis consists of unstructured information in the form of natural language texts. Among this information we find web pages (Wikipedia, digital newspapers, blogs) or social networks (Facebook, Twitter). Being able to analyze these texts, using natural language processing (NLP) algorithms, is very useful for organizations to take better decisions.

The automatic learning algorithms are not capable of understanding free text or characters, so the NLP performs all the necessary processing to convert this data in text form into a format understandable by machines (numbers) and thus be able to perform any subsequent analysis. Among the most common applications of the NLP are text classification, information search and extraction, automatic translation or question-answering systems, among others.

All the steps of the NLP, from the capture of the text in any format to the manipulation and analysis of it to obtain the relevant information, are explained in the compulsory subject 36429, Natural Language Processing that is taught in the second term of the third course.

Theory classes will be given in Spanish and practical and laboratory classes will be given according to the teaching guide available in the web page of the degree.

PREVIOUS KNOWLEDGE

Relationship to other subjects of the same degree

There are no specified enrollment restrictions with other subjects of the curriculum.

Other requirements

It is recommended to have passed the subjects Linear Models (second year) and Machine Learning (first term of third year).

COMPETENCES (RD 1393/2007) // LEARNING OUTCOMES (RD 822/2021)

1400 - Degree in Computer Engineering

- C1 - Ability to know the fundamentals, paradigms and techniques in the field of intelligent systems, and to analyse, design and build computer systems, services and applications that use these techniques in any field of application.
- C2 - Ability to acquire, obtain, formalise and represent human knowledge in a computable form for solving problems through a computer system in any field, particularly in those related to aspects of computing, perception and action in intelligent environments.
- C3 - Ability to recognise and develop computational learning techniques and to design and implement applications and systems that use them, including those for the automatic retrieval of information and knowledge from large volumes of data.

1406 - Degree in Data Science

- Students must have developed the learning skills needed to undertake further study with a high degree of autonomy.
- (CG06) Ability to access and manage information in different formats for subsequent analysis in order to obtain knowledge from data.
- (CT04) To be responsible for ones own professional development and specialisation, applying the acquired knowledge in the identification of career opportunities and sources of employment.
- (CT05) Ability to evaluate the advantages and disadvantages of different methodological and / or technological alternatives in different fields of application.

- (CE03) Ability to solve classification, modelling, segmentation and prediction problems from a set of data.
- (CE07) Ability to model dependency between a response variable and several explanatory variables, in complex data sets, using machine learning techniques, interpreting the results obtained.

LEARNING OUTCOMES (RD 1393/2007) // NO CONTENT (RD 822/2021)

Knowing how to segment text into simple elements (CG06)

Know the natural language processing techniques (CG06, CT04, CE03, CE07).

To know and implement the most extended applications of natural language processing (CB05, CT05, CE03, CE07).

As a consequence of the acquired learning results, the students will acquire the following skills:

- Be able to load and process text in Python.

- Knowing how to use regular expressions on text.

- Know and know how to use the most important NLP libraries for Python.

- Be able to convert text into numerical vectors for later treatment with machine learning algorithms.

- Be able to design a text classification application.

- Be able to analyze large volumes of text to extract its most representative topics and perform information searches.

DESCRIPTION OF CONTENTS

1. Introduction to Natural Language Processing

1.1. What is NLP
1.2. The importance of text
1.3. Historical approaches to NLP
1.4. Workflow in NLP applications

2. Using text in Python

2.1. Text strings in Python
2.2. Regular expressions
2.3. Loading text
2.4. NLP libraries in Python
2.5. Capturing Web content (Web scraping)

3. Pre-processing of text

3.1. Cleaning and standardization of the text
3.2. Word division (tokens)
3.3. Morphological analysis (lemmas)
3.4. Semantic analysis (Part of Speech)
3.5. Grammar analysis (dependency analysis)

4. Feature extraction

4.1. Vector space models
4.2. Bag of Words model
4.3. Model TF-IDF
4.4. Word embeddings
4.5. Document vectors

5. Applications of the NLP

5.1. Classification
5.2. Information extraction
5.3. Name Entity Recognition
5.4. Topic Modeling
5.5. Sequential models

6. Natural Language Processing laboratory

In this block, a series of practical exercises will be carried out to apply the concepts of NLP in the computer classroom.
Practice 0. Using text in Python
Practice 1. Regular expressions
Practice 2. Web scraping
Practice 3. Using NLP libraries in Python
Practice 4. Pre-processing text in Python
Practice 5. Feature Extraction in text
Practice 6. Topic Modeling
Practice 7. Practical application: sentiment analysis

WORKLOAD

ACTIVITY	Hours	% To be attended
Theory classes	30,00	100
Laboratory practices	20,00	100
Classroom practices	10,00	100
Development of group work	10,00	0
Development of individual work	10,00	0
Study and independent work	15,00	0
Readings supplementary material	5,00	0
Preparation of evaluation activities	10,00	0
Preparing lectures	10,00	0
Preparation of practical classes and problem	10,00	0
Resolution of case studies	15,00	0
Resolution of online questionnaires	5,00	0
TOTAL	**150,00**	

TEACHING METHODOLOGY

Lessons will combine theoretical and practical content:

MD1 - Theoretical activities. Expository development of the subject with the participation of the students in the resolution of specific questions. Carrying out of individual evaluation questionnaires.

In the theoretical activities during presential lessons, the different aspects of the subject will be developed providing a global and integrating vision: lessons will foment, at any moment, the participation of the students (CB05, CT05).

MD2 - Practical activities. Learning through problem solving, exercises and case studies through which skills are acquired on the different theoretical aspects of the subject. (CB05, CG06, CE03, CE07)

The theoretical activities are complemented by computer practices with the aim of putting the basic concepts into use and extending them with the knowledge and experience acquired during the performance of the proposed work.

MD4 -Laboratory and/or computer classroom work. Learning through guided activities developed individually or in small groups and carried out in laboratories and/or computer classrooms. (CB05, CG06, CT04, CE03, CE07)

In addition to the classroom activities, students will be required to perform personal tasks (outside the classroom) on: issues and problems, as well as class and exam preparation (study). These tasks will mainly be done individually, in order to promote autonomous work, but additionally, tasks will be included, especially the preparation and resolution of laboratory practices, which require the participation of small groups of students (2-3) to promote the capacity of integration in work groups.

The University of Valencia's e-learning platform (Aula Virtual) will be used as a support for communication with students. Through it, students will have access to the teaching material used in class and the scripts of the laboratory practices, as well as the problems and exercises to be solved.

EVALUATION

The evaluation of the knowledge and competences achieved by the students will be done continuously throughout the course, and will consist of the following blocks:

- SE1 - Objective test, consisting of an exam with both theoretical and practical questions and problems (competence assessment CB05, CT05, CE03, CE07) (50%) (Note: All percentages refer to the final mark)

- SE1-1 (40%) Theory-Problem Test

- SE1-2 (10%) Laboratory test

- SE2 - Evaluation of laboratory practices, from the elaboration of works/memories and/or oral presentations (competence evaluation CB05, CG06, CT04, CE03, CE07) (30%)

- SE2-1 (20%) Implementation of a mini-project consisting of the development of a complete NLP application for text classification

- SE2-2 (10%) Attendance and evaluation of lab sessions (non-recoverable activity)

- SE3 - Continuous evaluation of the student. (20%)

- SE3-1 (5%) Regular attendance at planned face-to-face activities (competency assessment CB05) (Non-recoverable activity)

- SE3-2 (15%) Resolution of proposed questions and problems (competency assessment CB05, CG06, CE03) (Non- recoverable activity)

The final mark of the course will be calculated as the weighted average of each of all the previous sections, according to the following criteria: SE-1 (50%), SE-2 (30%), SE-3 (20%).

Particular considerations on the evaluation:

- It is necessary to obtain a minimum grade of 4 (out of 10) in each of the evaluation sections SE1-1, SE1-2 and SE2-1.

- The activities SE2-2, SE3-1 and SE3-2 are not recoverable.

In any case, the evaluation system will be ruled by the provisions of the University of Valencia's Evaluation and Qualification Regulations for Degrees and Masters:

https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdicto Seleccionado=5639

REFERENCES

Basic

- Sohom Ghosh, Dwight Gunning. Natural Language Processing Fundamentals. Packt Publishing, 2019.
- Akshay Kulkarni, Adarsha Shivananda. Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python. Apress, 2019 (disponible e-libro)
- Dipanjan Sarkar. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. Apress 2019 (disponible e-libro)
- Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python. OReally Media, 2009.

Additional

- Jacob Eisentein. Natural Language Processing. 2018 (disponible bajo licencia CC-BY-NC-ND)
- Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. Practical Natural Language Processing. OReally Media, 2020

ADDENDUM COVID-19

This addendum will only be activated if the health situation requires so and with the prior agreement of the Governing Council

The teaching methodology of the course will follow the Teaching Model approved by the Data Science Academic Committee (https://go.uv.es/cienciadatos/ModelDocentGCD2Q). In the event that the facilities are closed for health reasons that affect all or part of the course sessions, these will be replaced by non-presential sessions following the established timetable. If the closure affects a presential assessment test for the subject, it will be replaced by a test of a similar nature that will be carried out in virtual mode

through the computer tools supported by the University of Valencia. The percentages of each assessment test will remain unchanged, as established by this guide.