COURSE DATA

DATA SUBJECT

Code: 36429
Name: Natural language processing
Cycle: Undergraduate Studies
ECTS Credits: 6
Academic year: 2026-27

STUDY (S)

Degree	Center	Acad. year	Period
1400 - Degree in Computer Engineering	Escola Tècnica Superior d'Enginyeria	4	Second quarter
1406 - Degree in Data Science	Escola Tècnica Superior d'Enginyeria	3	Second quarter

SUBJECT-MATTER

Degree	Subject-matter	Character
1400 - Degree in Computer Engineering	Optional subject	ELECTIVES
1406 - Degree in Data Science	Machine Learning and Data Mining	COMPULSORY

COORDINATION

PELLICER VALERO OSCAR JOSE

SUMMARY

Natural Language Processing (NLP) is a discipline at the intersection of artificial intelligence, computer science, and linguistics, aiming to equip machines with the ability to understand, interpret, and generate human language. This course offers a comprehensive journey through the field, starting from theoretical and practical fundamentals and advancing to the state-of-the-art techniques.

Key phases of an NLP project will be covered, including text acquisition, cleaning, and preprocessing. Fundamental text representation techniques will be studied, from classic models like Bag of Words and TF-IDF to semantic embeddings that capture contextual meaning.

The course also delves into the Deep Learning architectures that have revolutionized the field, with an analysis of the attention mechanism and the Transformer architecture. From there, it will explore the operation and application of Large Language Models, such as those from the GPT family. Advanced techniques like efficient fine-tuning (QLoRA), prompt engineering, and the construction of complex systems like Retrieval-Augmented Generation (RAG) will be examined.

Through theoretical and practical sessions, students will acquire the necessary skills to design and implement solutions for real-world problems such as text classification, information extraction, question-answering systems, and conversational agents.

PREVIOUS KNOWLEDGE

RELATIONSHIP TO OTHER SUBJECTS OF THE SAME DEGREE

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

OTHER REQUIREMENTS

For a proper follow-up of the course, the following knowledge is recommended:
- Intermediate level of programming in Python.
- Fundamentals of Machine Learning: Understanding of the basic concepts of supervised and unsupervised learning (classification, regression, clustering) and model evaluation metrics.
- Basic knowledge of linear algebra and probability: Familiarity with vectors, matrices, and probability concepts, necessary to understand how the models work.
- Previous experience with Python's Data Science ecosystem libraries, such as pandas, numpy, and scikit-learn.

COMPETENCES / LEARNING OUTCOMES

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

(CB5) Students must have developed the learning skills needed to undertake further study with a high degree of autonomy.

(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.

(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.

DESCRIPTION OF CONTENTS

1. Fundamentals and tools of Natural Language Processing

This introductory unit establishes the conceptual and practical foundations of the course. It will explore the definition and history of Natural Language Processing (NLP), analyzing the inherent complexity of human language and its components, from morphology to pragmatics. On a practical level, it will cover text handling in Python, including advanced use of strings and the power of regular expressions for pattern manipulation and searching. Finally, techniques for acquiring textual corpora, such as web scraping, will be introduced, and the ecosystem of libraries and frameworks to be used throughout the course will be presented, laying the groundwork for the complete lifecycle of an NLP project.

2. Text engineering and linguistic analysis

The objective of this unit is to learn how to transform raw, unstructured text into clean, enriched data ready to be processed by algorithms. Cleaning and normalization techniques, such as lowercasing and handling special characters, will be studied in depth. The fundamental task of tokenization will be addressed, both at the word and sub-word level, which is crucial for modern models. The content then delves into computational linguistic analysis, covering morphological analysis for reducing words to their stem or lemma (stemming and lemmatization), Part-of-Speech tagging to identify the function of each word, syntactic dependency parsing, and Named Entity Recognition (NER).

3. Vector representation of text

This unit focuses on the critical step of converting text into numerical representations that a machine can understand, exploring the evolution from lexical to semantic models. It will begin with classic sparse representations, such as the Bag-of-Words (BoW) model and its refinement using TF-IDF weighting, understanding their strengths and limitations. Next, a conceptual leap will be made towards dense representations with Word Embeddings. Foundational algorithms like Word2Vec, GloVe, and FastText will be analyzed in detail, which revolutionized the field by capturing semantic and contextual relationships in a continuous vector space, allowing for algebraic operations on word meanings.

4. Modeling and applications with text vectors

Once the text has been converted into meaningful vectors, this unit explores how to use these

representations to solve specific problems using Machine Learning algorithms. It will cover document classification, one of the most common NLP applications, with use cases like sentiment analysis. It will delve into unsupervised text mining through Topic Modeling, using techniques like LSI and LDA to discover latent themes in large document collections. Finally, the fundamentals of Information Retrieval systems will be studied, from classic ranking algorithms like BM25 to modern semantic search approaches based on embedding similarity.

5. Deep Learning architectures and contextual models

This unit marks the transition to the deep neural network architectures that define the current state of the art. It will present the historical context with Recurrent Neural Networks (RNN, LSTM) and Convolutional Neural Networks (CNN) for sequence processing. The main focus will be on the Transformer architecture, breaking down its most innovative component: the self-attention mechanism. It will be analyzed how this architecture allows for non-sequential text processing, capturing long-range dependencies. As a primary application of these ideas, the BERT (Bidirectional Encoder Representations from Transformers) model will be studied in depth.

6. Large Language Models (LLMs) and generative applications

Building on the Transformer architecture, this unit dives fully into the paradigm of Large Language Models (LLMs). The characteristic decoder-only architecture of the GPT family of models will be analyzed, and their lifecycle will be detailed: from massive-scale pre-training to the fine-tuning phases using instruction tuning and Reinforcement Learning from Human Feedback (RLHF). On a practical level, techniques for interacting with and adapting these models will be explored, such as prompt engineering, the creation of Retrieval-Augmented Generation (RAG) systems to base responses on external knowledge, and efficient fine-tuning methodologies like LoRA and QLoRA. Finally, this knowledge will be applied to the development of advanced generative applications such as machine translation, abstractive text summarization, and conversational agents.

7. Natural Language Processing laboratory

This practical block is designed to tangibly apply the theoretical concepts developed throughout the course. The planned practical sessions are:
Lab 0: Python Fundamentals for NLP (non-contact)
Lab 1: Text Manipulation with Regular Expressions
Lab 2: Textual Data Acquisition: Web Scraping
Lab 3: Linguistic Analysis and Preprocessing
Lab 4: Document Modeling and Classification
Lab 5: Applications with Large Language Models (LLMs)
Lab 6: Building Systems with LLMs

WORKLOAD

PRESENCIAL ACTIVITIES

Activity	Hours

Theory	30,00
Laboratory	20,00
Classroom practices	10,00
Total hours	**60,00**

NON PRESENCIAL ACTIVITIES

Activity	**Hours**
Attendance at other activities	0,00
Individual or group project	20,00
Independent study and work	15,00
Preparation of lessons	25,00
Preparation for assessment activities	10,00
Resolution of case studies	20,00
Total hours	**90,00**

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) (35%)

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

- SE2-2 (15%) Attendance and evaluation of lab sessions (non-recoverable activity)

- SE3 - Continuous evaluation of the student. (15%)

- SE3-1 (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 (35%), SE-3 (15%).

Particular considerations on the evaluation:

- To pass the course, it is necessary to obtain a minimum grade of 5 (out of 10) in each of the evaluation sections SE1-1, SE1-2 and SE2-1.

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

In any case, the evaluation system will be governed by the Regulations of Evaluation and Qualification of the University of Valencia for bachelor's and master's degrees. (https://webges.uv.

es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdictoSeleccionado=5639)

Copying or plagiarism of any activity that is part of the evaluation will result in the impossibility of passing the course, and the student will then be subject to the appropriate disciplinary procedures indicated in the ACTION PROTOCOL FOR FRAUDULENT PRACTICES AT THE UNIVERSITY OF VALENCIA (ACGUV 123/2020).

REFERENCES

Jurafsky, D., & Martin, J. H. (2025). Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

Lane, H., & Dyshel, M. (2025). Natural language processing in action. Simon and Schuster

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc."

Vaswani, A., et al. (2017). "Attention Is All You Need". Advances in Neural Information Processing Systems 30 (NIPS 2017)

Devlin, J., et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (pp. 4171-4186)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 36, 10088-10115