Academic Year/course: 2022/23

32138 - Natural Language Procesing

Syllabus Information

Academic Course: 2022/23

Academic Center: 803 - Masters Centre of the Department of Translation and Language Sciences

Study: 8037 - Theoretical and Applied Linguistics - MA

Subject: 32138 - Natural Language Procesing

Credits: 5.0

Course: 1

Teaching languages:

Theory: Group 1: English

Teachers: Thomas Simon Brochhagen

Teaching Period: Second Quarter

Presentation

This class gives an introduction to central aspects of natural language processing. It puts an emphasis on hands-on experience with the acquisition, manipulation, curation, and processing of linguistic data. It covers both symbolic and statistical methods, from a theoretical and practical angle.

The main goal of this class is for students to acquaint themselves with state of the art techniques used in industry and academia to structure language data and extract information from it; as well as to empower them to apply this knowledge to new problems outside the scope of the class.

Associated skills

- Programming (python)
- Linguistic data acquisition, manipulation, curation, and processing
- Machine learning
- Quantitative reasoning applied to language sciences

Learning outcomes

The main goal is for students to acquaint themselves with state of the art techniques used in industry and academia to structure language data and extract information from it. This goal can be further subdivided into two. First, from a theoretical perspective, the aim is for students to understand how linguistic data can be processed and analyzed with different computational methods; and to recognize what the advantages and drawbacks of different choices to do so are. Second, on the practical side, the aim is to enable students to be able to process natural language data on their own, and to be able to build on the knowledge acquired in this class to tackle problems not covered in class.

Sustainable Development Goals

SDG 5 Gender Equality, with an emphasis on empowering women with technology

SDG 9 Industry, innovation and infrastructure, with emphasis on scalability and data quality for low-resource languages

SDG 4 Quality education, through technical empowerment

Prerequisites

Knowledge of *python 3*. See the "Recommendations" section of the specialization in Computational Linguistics of the Master's for more information: https://www.upf.edu/web/masterlinguistica/linguistica-computacional.

Contents

Section 1: Handling text

Section 2: Language models

Section 3: Tagging

Section 4: Parsing

Section 5: Information extraction

Section 6: Other topics of interest to students (e.g., human in the loop & annotation)

Teaching Methods

The course is largely based on a flipped classroom format. Students are expected to prepare weekly readings and to lead a portion of a weekly session (see "Evaluation"), the remaining time is devoted to theoretical discussions and practical applications of the concepts introduced.

Evaluation

- 20% participation in class discussions/presentations
- 80% exercises (exercise 1: 25%; exercise 2: 25%; exercise 3: 30%)

Bibliography and information resources

Core Resources:

- Jurafsky, Daniel & Martin, James H. (2021), Speech and Language Processing. 3d edition (in progress). Available at https://web.stanford.edu/~jurafsky/slp3/
- Jurafsky, Daniel & Manning, Christopher D. (2019), Natural Language Processing lectures, You Tube: https://www.youtube.com/watch?v=8rXD5-xhemo&list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z&index=1
- Vasiliev, Yuli (2020), Natural Language Processing with Python and SpaCy: A Practical Introduction

Other recommended readings and resources:

- Franke, Michael (2021), An Introduction to Data Analysis (https://michael-franke.github.io/intro-data-analysis/index.html)
- Winter, Bodo (2019): Statistics for Linguists: An Introduction using R
- Thomas, Joy A. & Thomas M. Cover (1991): Elements of Information Theory
- Monarch, R. (2021): Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI