UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Natural Language Processing

2627-2-FDS02Q011

Aims

The aim of the course is to provide an introduction to the fundamental concepts of Natural Language Processing (NLP) as well as an overview of the main tools used in the field. In addition, some applications of NLP will be presented, e.g. machine translation and hate speech detection.

The specific objectives with respect to the Dublin Descriptors (DdD) are the following:
1. Knowledge and understanding

- Become familiar with the fundamental concepts of Natural Language Processing (NLP).
- Understand the main statistical and neural techniques for natural language processing.
- Become familiar with the most recent tools and models in the NLP field, including Word2Vec, FastText, GloVe, BERT, GPT.

2. Applied knowledge and understanding

- Apply NLP algorithms and models to real-world problems such as machine translation and hate speech detection.
- Use open-source tools to implement NLP systems.
- Design and develop an NLP application starting from a specific domain

3. Autonomy of judgment

- Be able to choose the most appropriate methods and models to solve specific tasks related to natural language analysis.
- Analyze the effectiveness of linguistic models

4. Communication skills

- Present an NLP project clearly and with arguments, also through oral presentation.

- Communicate the results of the analysis to a technical and non-technical audience.
- Collaborate in a group for the realization of a complex project, sharing knowledge and strategies.

5. Ability to learn

- Develop independent study skills through the completion of optional assignments deriving from laboratory activities focused on the use of NLP models and tools.

Contents

The course content includes fundamental principles of Natural Language Processing (NLP) and offers an overview of the key tools utilized in this field. The course will cover a range of topics, ranging from statistical techniques to recent advancements in neural approaches. Moreover, the course incorporates practical demonstrations of different NLP applications, including machine translation, and hate speech detection.

Detailed program

Course introduction
- Rationalist and Empiricist Approaches to Language
- The Ambiguity of Language: Why NLP Is Difficult

Linguistic Essentials
- Lexical resources
- Zipf's laws
- Collocations
- Concordances
- Syntax

Frequentist Representation of Text (TF, TF-IDF, etc..) and Word Embeddings
- Word2Vec
- FastText
- Glove

Visualization of embeddings:
- Principal Components Analysis
- T-distributed stochastic neighbor embedding
- Uniform Manifold Approximation and Projection

Sequence-to-Sequence (RNN, LSTM)

Transformers and Large Language Models
- Attention Mechanisms: Self and Multi Head Attention

Contextualized Language Models:
- ELMO
- BERT
- GPT
- LLAMA

Prompting and Instruct Tuning

Language Model Evaluation Metrics

Interpretability and Explainability of Language Models

Prerequisites

Basic knowledge of statistics and programming languages.

Teaching form

The course will be taught in English, and it will consist of both lectures introducing the main topics and tutorial sessions where open-source tools will be explained.
Seminars held by experts at national and international levels may be part of the course.

24 lectures of 2 hours delivered in person.

Textbook and teaching resource

Daniel Jurafsky and James Martin, "Speech and Language Processing, 2nd Edition", Prentice Hall, 2008.

Emily M. Bender, "Linguistic Fundamentals for Natural Language Processing", Synthesis lectures on human language technologies, Morgan&Claypool Publishers, 2013.

Yoav Goldberg, "Neural Network Methods for Natural Language Processing", Synthesis lectures on human language technologies, Morgan&Claypool Publishers, 2017.

Mohammad Taher Pilehvar and Jose Camacho-collados, "Embeddings in Natural Language Processing", Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2021.

Semester

Second Semester

Assessment method

Project
• The project consists in the development of a natural language processing tool based on methods and models presented during the course.
• Each group/individual must identify a domain of interest and dataset for which it intends to address specific NLP tasks.
• The project must be presented orally
• The project is evaluated in the range [0-24] according to the following criteria:

1. Definition of the problem and objectives: max 2 points
2. Justification of the methodological choices: max 3 points
3. Implementation and correctness of the code: max 4 points
4. Analysis of the experimental results: max 6 points

5. Completeness and clarity of the final report: max 6 points
6. Conclusions and final considerations: max 3 points

Oral Exam
• The oral exam can have an outcome between [-8; +8]
• It consists of 4 questions about topics addressed during the course: -2 will be given for an incorrect answer or no answer, +2 for a correct answer.

There are no mid-term tests.

Office hours

To be agreed with the teacher

Sustainable Development Goals