UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Natural Language Processing

2324-2-FDS01Q011

Aims

The course aims to introduce the foundational elements and the most recent advanced computational models related to natural language processing. At the end of the training activity, the student will have acquired knowledge and skills related to algorithms, tools and models for processing and analyzing natural language, in order to exploit the most recent state-of-the-art processing systems.

Contents

Foundations of natural language representation
Semantics of words
Large Language Models
NLP applications

Detailed program

1. Fundamentals
    - Rationalist and Empiricist Approaches to Language
    - The Ambiguity of Language: Why NLP Is Difficult
    - Linguistic Essentials
        - Words, Tokens, Lemmas, Stems
        - Parts of Speech and Morphology
        - Phrase Structure
    - Dirty Hands-on Text

        - 1.4.1 Lexical resources
        - 1.4.2 Word counts
        - 1.4.3 Zipf’s laws
        - 1.4.4 Collocations
        - 1.4.5 Concordances

2. Vector Semantics

    - Frequentist Representation of Text (TF, TF-IDF, etc..)
    - Word Embeddings
        - Word2Vec
        - FastText
        - Glove
    - Visualization of Embeddings:
        - Principal Components Analysis
        - T-distributed stochastic neighbor embedding
        - Uniform Manifold Approximation and Projection

3. Transformers and Large Language Models

    - Attention Mechanisms: Self and Multi Head Attention
    - Positional Embeddings
    - Transformers as Language Models
    - Pretraining Large Language Models
    - Prompting and Instruct Tuning
    - Interpretability and Explainability of Language Models

4. NLP Applications

    - Text and Token Classification
    - Chatbots and Dialog Systems
    - Word Sense Disambiguation
    - Topic Modeling
    - Machine Translation

Prerequisites

Useful, but not required: machine learning, python programming

Teaching form

Lectures and classroom exercises.
The course will be given in English.

Textbook and teaching resource

Cristopher MANNING and Hinrich SCHÜTZE. Foundations of Statistical Natural Language Processing. MIT Press.
Dan JURAFSKY and James H. MARTIN. Speech and Language Processing. Prentice Hall.

Semester

Second semester

Assessment method

Project and Oral Exam. Intermediate tests are absent.
The project will consist in the development of a natural language processing tool based on methods and models presented during the course. The project is evaluated in the range of 0-24 points.
The oral exam consists of 4 questions about theory addressed during the course and listed in the detailed program. For each question, an evaluation of -2 will be given, for an incorrect answer or no answer, and +2 points for a correct answer.

Office hours

By appointment to be agreed via email with the teacher.

Sustainable Development Goals