UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Text Mining and Natural Language Processing

2324-2-E311PV011

Aims

The aim of the course is to provide an introduction to the fundamental concepts related to the Linguistic aspects of human languages, and Natural Language Processing (NLP) techniques; moreover, in the course, some NLP applications will be presented, i.e. Information Retrieval and Machine Translation.

After successfully completing the course, students will be able to:

-describe basic linguistic aspects of human languages.
-explain the common computational vector space models for words applied in language technology.
-describe the challenges related to word vector models.
-know the main neural language models and apply them for different applications.

Contents

This course will first provide the notions of the morphological and syntactic structure of human languages, useful in creating more linguistically aware NLP systems.

The course will then introduce some NLP tasks and text representation techniques. Starting from statistical methods to modern neural approaches, an overview of fundamental techniques will be presented and practiced, such as the n-gram model, Word2Vec, the encoder-decoder paradigm, and neural language models. Open-source software for NLP will be introduced and used throughout the lab sessions.

Detailed program

Introduction to levels of linguistics analysis and typological differences

Morphology/morphophonology

Morphosyntax/syntax

Parts of speech

Heads, arguments, adjuncts

Argument types and grammatical functions

Mismatches between syntactic position and semantic roles

Resources

Introduction to some NLP tasks

Data Pre-Processing (eg. tokenization, NER, etc.)

Text representation (eg. tf-idf)

Statistical LM (eg. n-gram model)

Dense vector representation (eg. Word2Vec, FastText, etc.)

Deep Neural Approaches for NLP (eg. Encoder-Decoder, Neural Language Model)

Applications of NLP:

Information Retrieval

Machine Translation

Prerequisites

Basic knowledge of statistics, programming languages, and machine learning.

Teaching form

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

Textbook and teaching resource

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

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

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

Semester

Second Semester

Assessment method

Written and optional oral individual examination, definition of a laboratory project that can be developed also by groups of students (up to three students).

The written examination is aimed at assessing the level of understanding of the basic aspects taught during the course; it is constituted by a set of open questions.

The goal of the group project is the usage of open-source software that will be employed to develop technological solutions to the problems addressed in the course. In particular, real application areas will be considered, which require the definition of systems presented during the course.

Office hours

To be agreed with the teachers

Sustainable Development Goals