UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Text Mining and Search

2324-2-FDS01Q013

Aims

The aim of the course is to provide an introduction to the fundamental concepts related to Text Representation and Text Mining techniques; moreover, in the course some Text Mining applications will be presented: Text Classification and Clustering, Topic Modelling, and Text Summarization. An introduction to Search Engines and Recommender Systems will be provided.

Contents

- This course will first provide the definition of Text Mining and will point out the basic differences between Data Mining and Text Mining.
- The issues of text pre-processing and analysis, and of text indexing and representation will be addressed.
- The course will then introduce some tasks involved by Text Mining, which include Text Clustering, Classification, Topic Modeling, and Text Summarization.
- Then the course will introduce the previously mentioned tasks. Some open source software for Text Mining will be introduced and practiced.

Detailed program

1. Definition of Text Mining and basic differences between Data Mining and Text Mining
2. Introduction to some tasks related to Text Mining
3. Text pre-processing, indexing and formal representation (BoW, Word Embedding, Introduction to Contextual Word Embedding techniques)
4. Text Classification and Clustering
5. Topic Modelling

6. Text Summarization
7. Introduction to Text Based Search Engines and to Recommender Systems
8. Open Source software for Text Mining and Search

Prerequisites

Basic knowledge of statistics and of programming languages.

Teaching form

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

Textbook and teaching resource

- Berry, M. W., & Kogan, J. (Eds.). (2010). Text mining: applications and theory. John Wiley & Sons.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. Fundamentals of artificial intelligence, 603-649.

Other specific books and articles on text mining that are accessible online will be recommended during the course.

Semester

First semester

Assessment method

Written exam, definition of a **laboratory project** (project work) 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. The **evaluation** of the written examination will be in thirtieths. 0 to 4 points will be added to this evaluation.

Office hours

To be agreed with the teachers

Sustainable Development Goals