UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Text Mining and Search

2627-2-FDS02Q013

Aims

Knowledge and Understanding

The aim of the course is to introduce the fundamental concepts of Natural Language Processing (NLP) and Text Mining techniques, providing a solid foundation for the automatic analysis of textual data. After an overview of basic methodologies, the course will cover specific tasks such as document classification and clustering, topic modeling, automatic text summarization, and Information Retrieval (IR).

Applying Knowledge and Understanding

Throughout lectures and lab activities, students are encouraged and assessed on their ability to apply the knowledge acquired to the topics covered in the course.

Making Judgements

The course aims to foster independent judgement and critical analysis skills in relation to the main challenges of natural language processing and Text Mining, as well as the related key tasks. These competencies will be further developed through in-class discussions and lab work.

Communication Skills

Development of the ability to clearly, consciously, and unambiguously communicate technical content, ideas, problems, and corresponding solutions to different types of audiences. These skills will be promoted during the course and assessed as part of the final examination.

Learning Skills

The course is designed to provide both theoretical knowledge and practical skills, offering a solid starting point for further individual study of the principles of text representation, analysis, and retrieval.

Contents

- The course will provide an introductory definition of Text Mining and Natural Language Processing (NLP), highlighting the main differences between Data Mining and Text Mining.
- Key text pre-processing techniques will be presented, along with issues related to text indexing and formal representation.
- Fundamental Text Mining applications will then be introduced, including document classification and clustering, topic modeling, automatic text summarization, and textual Information Retrieval (IR).
- The course will also present selected open-source tools useful for developing Text Mining applications.

Detailed program

1. Definition of Natural Language Processing (NLP), Text Mining, and main differences between Text Mining and Data Mining.
2. Brief introduction to selected Text Mining applications.
3. Text pre-processing techniques, indexing, and formal text representation (Bag-of-Words, Word Embedding, introduction to Contextualized Word Embedding techniques).
4. Text classification and clustering.
5. Topic modeling.
6. Automatic text summarization.
7. Introduction to text search engines.
8. Open-source tools for Text Mining and online Information Retrieval (IR).

Prerequisites

Basic knowledge of statistics and programming (preferably in Python).

Teaching form

- The course consists of 46 hours, of which 9 are laboratory sessions.
- The course is taught in English.
- Lessons are 2 hours long (with one session lasting 3 hours) and are conducted both in a lecture format (mainly at the beginning of the lesson) and in an interactive mode (during the lesson) to actively engage students.
- During laboratory activities, the use of open-source software is explained and practiced.
- Seminars by national and international experts may be scheduled.
- Some lessons may be conducted remotely in a lecture-interactive format.

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 and completion of a **laboratory project**, which **must be carried out in groups** of up to three students.

- The written exam aims to assess the understanding of the fundamental concepts taught in the course and consists of a series of open-ended questions.
- The goal of the group project, through the use of open source software, is to develop technological solutions to problems addressed during the lessons. In particular, real-world application areas are considered, requiring the design of systems whose foundations were presented during the course. The project is presented in person by the students to assess the skills they have actually acquired, both technical and critical/judgmental, while also developing their communication abilities.

The written exam will be **graded** out of 30. Upon obtaining a passing grade in the written exam (at least 18/30), 0 to 4 additional points will be added based on the project evaluation. The project will be evaluated based on its completeness with respect to the specifications, the correctness of the algorithmic solutions used, the clarity of presentation, and the quality of the material submitted.

No midterm exams are scheduled.

Office hours

To be agreed with the teachers.

Sustainable Development Goals

GOOD HEALTH AND WELL-BEING | QUALITY EDUCATION | GENDER EQUALITY | AFFORDABLE AND CLEAN ENERGY | INDUSTRY, INNOVATION AND INFRASTRUCTURE | PEACE, JUSTICE AND STRONG INSTITUTIONS