UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Information Retrieval

2526-2-F1801Q110

Aims

Knowledge and Understanding
The objective of the course is to provide an introduction to the fundamental concepts, formal models, and techniques for implementing systems for the automatic retrieval of textual documents in digital form (“Information Retrieval” systems, aka Search Engines, or Web Search Engines when the documents to be retrieved consist of Web pages). After an introduction to the automatic analysis of texts and their representation (basic NLP techniques), the problem of information retrieval will be defined, which presupposes the estimate of the relevance of documents to the user's information needs expressed in a query. At the end of the course, the student will be able to understand the basic techniques related to the analysis and representation of texts in natural language, as well as the operation of a search engine. The student will also be able to use “open source” software to define Information Retrieval applications. The lab will be aimed at the implementation of an application.
Applying Knowledge and Understanding
Throughout lectures and lab activities, students are encouraged and evaluated on their ability to apply the knowledge acquired to the topics covered in the course.
Autonomy of judgment
The course aims to foster independent judgement and critical analysis skills in relation to the main challenges of natural language processing and search, 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 communicate technical content, ideas, problems, and corresponding solutions clearly, consciously, and unambiguously to different types of audiences. These skills will be fostered during the course and evaluated 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 first introduce the problem of text representation for automatic text processing, and the techniques used in many NLP applications for text pre-processing and formal text representation (including neural techniques). The main techniques for the design and implementation of search engines will then be presented, with mention of techniques for defining recommender systems.
The main IR models for estimating the relevance of a document with respect to the user's information needs expressed in a query will be then introduced, from the vector space model to the more recent neural models. The techniques at the basis of Web Searche Engines will be presented.
Some techniques for personalizing search will be presented as well; also the technique of Retrieval Augmented Generation will be explained.
The course will also introduce the issue of evaluating the effectiveness of search engines.

Detailed program

1. Introduction to Text Processing and Natural Language Processing (NLP)

2. Introduction to Text Mining and to some tasks related to NLP

3. Text pre-processing, indexing and formal representation of texts ((bag of words, word embeddings, statistical language models, neural language models - large language models)

4. Information Retrieval models: basic models (Boolean model, Vector Space model, probabilistic models). Advanced models (neural models). Introduction to multimedia information retrieval.

5. Web Search Engines: crawling, link analysis and other factors for estimating relevance of Web pages.

6. The evaluation of Search Engines.

7. Advanced topics

8. Introduction to open source software for the development of search engines.

Prerequisites

Basic knowledge of statistics and of linear algebra.

Teaching form

The course will be taught in English and includes both lectures (40 hours) and laboratory exercises (12 hours). Each lecture lasts 2 hours and is delivered using both a traditional format (especially at the beginning of the lecture) and an interactive format (during the lecture) to encourage active student participation.
During the laboratory sessions, the use of open-source software will be explained and tested. Seminars led by internationally recognized experts will be organized. Some lectures may be delivered in a remote, lecture-interactive format.

Textbook and teaching resource

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

Semester

First semester

Assessment method

Individual written examination constituted by both exercises and open questions related to the course content. Definition of a laboratory project that can be also developed by groups of students (up to three students).
The written examination is aimed at assessing the level of understanding of the basic theoretical and technical aspects taught during the course.
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 teacher.

Sustainable Development Goals