UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Information Retrieval

2122-2-F1801Q110

Aims

This course aims at introducing the basic concepts, the formal models and the main techniques to define and design Information Retrieval Systems (also called Search Engines, and in particular Web Search Engines when working on the Web to the aim of retrieving Web pages) and Information Filtering (IF) systems. In this context, the main problem is the assessment of the relevance of documents with respect to the information needs formulated in a user's query. Students will acquire the capability of understanding and defining algorithms for documents indexing and retrieval, and to use open source software to implement ad hoc search engines. They will also develop a search engine application by using open source software.

Contents

This course aims at introducing the basic concepts, the formal models and the main techniques to define and design Information Retrieval Systems (also called Search Engines, and in particular Web Search Engines when working on the Web to the aim of retrieving Web pages) and Information Filtering (IF) systems. In particular, various techniques for the analysis and the indexing of texts will be presented, also including a basic introduction to multimedia documents indexing. Moreover, the issue of estimating the relevance of documents to a query will be addressed: several models finalised at the assessment of a numeric estimate of relevance (degree or probability) of a document to a query will be explained. The main approaches to personalized search will be presented. The course will also introduce additional applications related to text analysis and mining, such as crawling, analysis and retrieval of user generated content on Social Media (e.g. Twitter, Facebook, etc.).

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

4. Information Retrieval models: basic models (Boolean model, Vector Space model, probabilistic models). Advanced models (e.g. 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:

• Personalized Search

• Social Search

• Domain Specific Search

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 it will be constituted of both lectures introducing the main topics and of sessions in a laboratory where the usage of an open source software for the implementation of search engines will be explained and experienced. Seminars taught by international experts will be organised.

Textbook and teaching resource

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval,

Cambridge University Press, 2008.

John Scott, Social Network Analysis (Third Ed.), SAGE, 2013.

Semester

First semester

Assessment method

Individual written examination constituted by both exercises and open questions related to the course content; oral examination on the results of the written examination and on possible 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.