UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Information Retrieval

2627-2-F1802Q127

Aims

Knowledge and Understanding
The objective of the course is to provide knowledge and understanding of the main concepts, models, and techniques at the basis of the design of Information Retrieval systems (also known as Search Engines)..
The course introduces Information Retrieval as the area concerned with the automatic access to textual information in response to users' information needs expressed through queries.
Students will acquire knowledge of the main stages of the pipelines at the basis of the functioning of an Information Retrieval System, including text preprocessing, text representation, indexing, query processing, retrieval, ranking, and evaluation.
They will understand the main retrieval models used to estimate the relevance of documents with respect to a user query, including exact-match, lexical, probabilistic, neural, and hybrid approaches.
The course will show how Information Retrieval is grounding modern technologies, including Web search engines, domain-specific search systems, Retrieval-Augmented Generation, and LLM-based conversational systems.
The laboratory activities will support the theoretical part of the course through the practical design, implementation, and evaluation of Information Retrieval applications using open-source software.
During the laboratory activities, students will have the opportunity to address a real-world search problem and to present their own solution. This activity will allow students to apply the knowledge acquired during the course and to challenge their ability to analyze a problem, make technical design choices, and develop an original solution.

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. 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,

and unambiguously to different types of audiences. These skills will be fostered during the course and evaluated as part of the final examination, specifically through the project presentation.

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 introduces the main concepts of Information Retrieval and the role of search systems in accessing textual information.
- The course presents techniques for text preprocessing, text representation, and indexing, including bag-of-words representations and inverted indexes.
- The main IR models for estimating the relevance of documents with respect to users' queries will be introduced, from Boolean retrieval and vector space models (with a particular focus on BM25) to probabilistic models and neural models.
- The course will also cover the standard methodology for evaluating search systems, including test collections, relevance judgments, ranking metrics, and statistical significance testing.
- The techniques at the basis of Web search engines will be presented, together with approaches for user interaction, personalization, domain-specific search, and learning to rank for web search.
- The final part of the course will introduce modern neural and LLM-based retrieval systems, including neural text representations, dense retrieval, sparse and hybrid neural retrieval, reranking, Retrieval-Augmented Generation, and search in agentic systems.

Detailed program

1. Introduction to Information Retrieval and search systems. Main concepts of Information Retrieval, search engines, digital libraries, enterprise search, modern AI-based search systems, and examples of IR applications.
2. Text preprocessing, representation, and indexing. Tokenization, normalization, stemming, lemmatization, stopwords, n-grams, metadata, document fields, bag-of-words representations, inverted indexes, posting lists, positional indexes, index compression, and indexing pipelines.
3. Queries, relevance, and exact-match retrieval. Query types, document units, information needs, user intent, relevance, relevance judgments, Boolean retrieval, Boolean operators, phrase queries, proximity search, and structured queries.
4. Lexical ranking models. Term frequency, inverse document frequency, TF-IDF, vector space model, cosine similarity, probabilistic retrieval, BM25, BM25 parameters, fielded BM25, and comparison between TF-IDF and BM25.
5. Evaluation of Information Retrieval systems. Test collections, topics, queries, qrels, relevance judgments, and standard IR evaluation measures.
6. User interaction, personalization, and Web search. Query reformulation, search sessions, click behavior, search logs, personalization, exploratory search, crawling, Web indexing, link analysis, PageRank, spam detection, freshness, and snippets.
7. Domain-specific search and Learning to Rank. Scientific search, biomedical search, legal search, enterprise search, e-commerce search, domain terminology, domain constraints, learning to rank (pointwise, pairwise, and listwise approaches), ranking features, LambdaMART, and click-based learning.
8. Neural representations and dense retrieval. Word embeddings, contextual embeddings, sentence embeddings, document embeddings, transformer-based representations, bi-encoders, and vector databases.

9. Sparse, dense, hybrid, and multi-stage neural retrieval. Learned sparse retrieval, neural lexical matching, SPLADE-style methods, hybrid sparse-dense retrieval, score fusion, candidate generation, neural reranking, cross-encoders, late interaction models, ColBERT-style ranking, and efficiency-effectiveness tradeoffs.
10. LLM-based, RAG, and agentic search systems. Query rewriting, query expansion, synthetic query generation, LLM-based reranking, result summarization, answer generation, RAG architecture, chunking, context assembly, grounding, citation generation, hallucination control, evaluation of RAG systems, tool use, agentic search, memory, Web agents, and open research challenges.

Prerequisites

Basic knowledge of statistics, linear algebra, and programming (e.g. Python) is required.

Teaching form

The course will be taught in English and includes both lectures and laboratory exercises.
The course includes 40 hours of lectures and 12 hours of laboratory exercises.
Most lectures last 2 hours, although some intensive lectures may last 3 hours depending on the topic.

Lectures are delivered through a combination of teaching formats.
* The first part of each lecture is mainly based on frontal teaching and introduces theoretical concepts, formal models, and technical methods.
* The second part includes interactive teaching activities, such as guided discussion, examples, short exercises, and analysis of practical cases, in order to encourage active student participation and support the application of the concepts introduced during the lecture.

During the laboratory sessions, students will use open-source software to design, implement, and evaluate Information Retrieval components.
Seminars led by internationally recognized experts will be organized.
Some lectures may be delivered remotely in an interactive lecture format.

Textbook and teaching resource

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.
- Bhaskar Mitra and Nick Craswell. An Introduction to Neural Information Retrieval. Foundations and Trends in Information Retrieval, Volume 13, Issue 1, 2018.

Semester

First semester.

Assessment method

- Individual written examination composed of exercises, open questions, and closed questions related to the course content.
- The assessment also includes a laboratory project, which may be developed individually or in groups of 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 project is to use open-source software to develop technological solutions to search problems addressed in the course.
- The project will focus on real application scenarios that require students to design, implement, and evaluate Information Retrieval components introduced during the course.

Office hours

To be agreed with the teacher.

Sustainable Development Goals