Summary
This course introduces the foundations of information retrieval, data mining and knowledge bases, which constitute the foundations of today's Web-based distributed information systems.

Content

Information Retrieval
1. Information Retrieval - Introduction
2. Text-Based Information Retrieval (Boolean, Vector space, probabilistic)
3. Inverted Files
4. Distributed Retrieval
5. Query Expansion
6. Embedding models (LSI, word2vec)
7. Link-Based Ranking

Mining Unstructured Data
1. Document Classification (knn, Naive Bayes, Fasttext, Transformer models)
2. Recommender Systems (collaborative filtering, matrix factorization)
3. Mining Social Graphs (modularity clustering, Girvan-Newman)

Knowledge Bases
1. Semantic Web
2. Keyphrase extraction
3. Named entity recognition
4. Information extraction
5. Taxonomy Induction
6. Entity Disambiguation
7. Label Propagation
8. Link Prediction

Learning Prerequisites
Recommended courses
Introductory courses to databases and machine learning are helpful, but not required. Programming skills in Python are helpful, but not required.
Learning Outcomes

By the end of the course, the student must be able to:

- Characterize the main tasks performed by information systems, namely data, information and knowledge management
- Apply collaborative information management models, like crowd-sourcing, recommender systems, social networks
- Apply knowledge models, their representation through Web standards and algorithms for storing and processing semi-structured data
- Apply fundamental models and techniques of text retrieval and their use in Web search engines
- Apply main categories of data mining techniques, local rules, predictive and descriptive models, and master representative algorithms for each of the categories

Teaching methods

Ex cathedra + programming projects (Python)

Assessment methods

50% Continuous evaluations of projects with bonus system during the semester
50% Final written exam (180 min) during exam session

Resources

Moodle Link

- https://go.epfl.ch/CS-423