EPFL

CS-423 **Distributed information systems**

Aberer Karl

Cursus	Sem.	Type
Biocomputing minor	H	Opt.
Civil & Environmental Engineering		Opt.
Communication systems minor	H	Opt.
Computer science	MA1, MA3	Opt.
Cybersecurity	MA1, MA3	Opt.
Data Science	MA1, MA3	Opt.
Digital Humanities	MA1, MA3	Opt.
Electrical and Electronical Engineering	MA1, MA3	Opt.
Energy Management and Sustainability	MA1, MA3	Opt.
Environmental Sciences and Engineering	MA1, MA3	Opt.
Learning Sciences		Obl.
SC master EPFL	MA1, MA3	Obl.

Language of teaching	English
Credits	4
Session	Winter
Semester	Fall
Exam	Written
Workload	120h
Weeks	14
Hours	**3 weekly**
Courses	2 weekly
Exercises	1 weekly
Number of positions	

Summary

This course introduces the key concepts and algorithms from the areas 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. Association Rule Mining
2. Document Classification (knn, Naive Bayes, Fasttext, Transformer models)
3. Recommender Systems (collaborative filtering, matrix factorization)
4. 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
9. Data Integration

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 exercises (Python)

Assessment methods

25% Continuous evaluations with bonus system during the semester
75% Final written exam (180 min) during exam session

Supervision

Office hours	Yes
Assistants	Yes
Forum	Yes