

# **Distributed Information Systems**

## **Fall Semester – 2024**

### **CS-423**

#### **Time and Place**

**Lecture: Thursday 10:15 – 12:00, CM2**

<https://epfl.zoom.us/j/66237387610>

**Exercise: Thursday 12:15-14:00, CM1105**

**Karl Aberer**

Distributed Information Systems Laboratory

# Goals of the Course

Understand what is a "**Distributed Information System**"?

- e.g. Web Search Engines, Online Social Networks, etc.

Know which are **key tasks** relevant for DIS?

- e.g. retrieval, mining, recommending, information extraction, data integration etc.

Master **common methods** used to solve these problems

- e.g. vector space model, graph mining, word embeddings etc.

Pre-existing knowledge not required

Knowledge in databases and machine learning helpful

## **Disclaimer on GPT (LLMs)**

Many of these task can now be done with LLMs!

Why care about other methods?

- ▶ Understand the principles
- ▶ Alternative methods remain useful

## Related Courses

- Applied Data Analysis
- Introduction to database systems
- Database systems
- Introduction to machine learning
- Machine learning
- Introduction to natural language processing
- Internet analytics
- ...

## **Which master's program are you from?**

1. Computer Science
2. Communications
3. Data Science
4. Cybersecurity
5. Digital Humanities
6. Life Science
7. Electrical Engineering
8. Environmental Science
9. Others

# The Course - Lecture

## Standard online ex cathedra lecture

- Lecture streamed via Zoom
  - <https://epfl.zoom.us/j/66462767931>
  - Zoom QA tool to ask questions
  - Will be answered privately by assistants, or by the lecturer, depending on the questions
- Zoom Quizzes (anonymous)
- Zoom Chat to collect feedbacks

## Video recordings

- Check on Moodle

## **Are you planning to join the lecture live or virtually?**

1. I join live today, and plan to continue live
2. I join live today, but plan to join virtually
3. I join virtually today, and plan to continue virtually
4. I join virtually today, but plan to join live

# Materials

## Web platform: Moodle

- General announcements will be published on Moodle
- Course notes and project-related information will be published on the Web:  
<https://lsir.github.io/DIS/>
- Exercises and exam questions from previous years will be made available as well



## **Projects - Key element of the course**

### **Projects**

1. Information retrieval
2. Information extraction

**Done in groups of 2 or 3**

**Expected median workload: 30 hours**

**More details in exercise session**

# **Project Evaluation**

**Results - Metrics [Comparison with baselines]**

**Code**

- Working code
- Code quality and documentation

**2-page Report [Moodle submission]**

- Originality of approach
- Interpretation of results
- Presentation

## Exercise Platform

Ed Forum to ask questions offline:

<https://edstem.org/eu/courses/831/discussion/>

Both among students and with assistants

## Grading

Projects: 60%

- Each project contributes 30%

Final Exam: 40%

- Program problem similar to the projects
- will assume you attended the lecture
- will assume you did the projects
- examples from earlier years (exercises, exams) provided for preparation

## Exam Support

### **Your computer will be admitted to the exam**

- You will have Internet access
- But: communication not allowed (messaging, social platform etc.)
- You can use your notes (paper or electronically, all lecture materials)

# Schedule

Week	Date	Area	Topic	Project
1	Thursday, 12 September 2024	Introduction		
2	Thursday, 19 September 2024	Information Retrieval	Basic Information Retrieval	Information Retrieval
3	Thursday, 26 September 2024	Embedding techniques		
4	Thursday, 3 October 2024	Embedding techniques		
5	Thursday, 10 October 2023	Recommender Systems		
6	Thursday, 17 October 2023	Web Mining	Document Classification	
7	Thursday, 31 Oktober 2024	Link Ranking		
8	Thursday, 7 November 2024	Graph Mining		Information Extraction
9	Thursday, 14 November 2024	Information Extraction	Named Entity Recognition	
10	Thursday, 21 November 2024	Knowledge Representation		
11	Thursday, 28 November 2024	Information Extraction		
12	Thursday, 5 December 2024	Knowledge Inferences		
13	Thursday, 12 December 2024	Data Indexing and mining	Indexing for Information retrieval	
14	Thursday, 19 December 2024	Association Rule Mining		

# Lecturer

**Karl Aberer**

Head of LSIR

EPFL - I&C - LSIR  
BC108  
station 14  
CH-1015 Lausanne  
+41 21 693.46.73  
karl-aberer@epfl.ch



# Organizational Info

## Moodle

- <http://moodle.epfl.ch/course/view.php?id=4051>

## Lecturers

- Prof. Karl Aberer [karl.aberer@epfl.ch](mailto:karl.aberer@epfl.ch) BC 108

## Assistants

- Romanou Angelika [angelika.romanou@epfl.ch](mailto:angelika.romanou@epfl.ch)
- Negar Foroutan [negar.foroutan@epfl.ch](mailto:negar.foroutan@epfl.ch)
- Borges Ribeiro Beatriz Maria [beatriz.borges@epfl.ch](mailto:beatriz.borges@epfl.ch)
- Ismayilzada Mahammad [mahammad.ismayilzada@epfl.ch](mailto:mahammad.ismayilzada@epfl.ch)



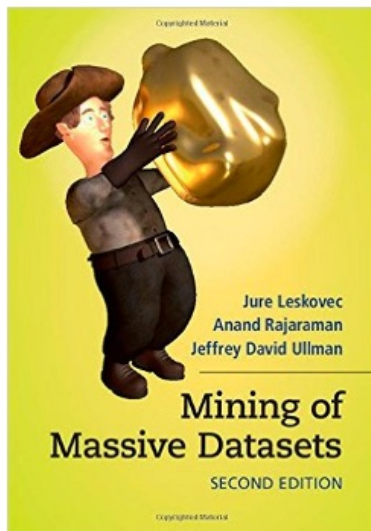
# References

Parts of the course are based on the following text books

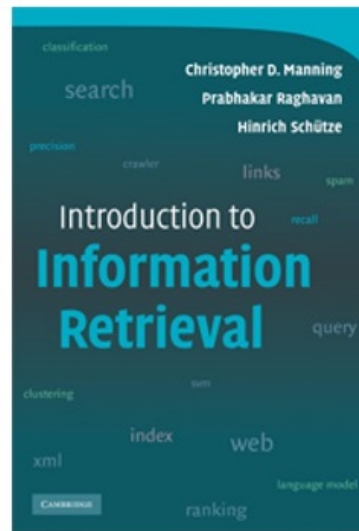
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval (Acm Press Series), Addison Wesley, 1999.
- Jiawei Han, Data Mining: concepts and techniques, Morgan Kaufman, 2000.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- J Leskovec, A Rajaraman, JD Ullman, Mining of Massive Datasets, 2014.

Further references to the literature will be given during the lecture

## Free books



[mmds.org](http://mmds.org)



<http://nlp.stanford.edu/IR-book/>

# Part 1 Information Retrieval

## Part 1: Information Retrieval

### 1.1 Introduction to Information Retrieval

### 1.2 Basic Information Retrieval

#### 1.2.1 Text-Based Information Retrieval

#### 1.2.2 Boolean Retrieval

#### 1.2.3 Vector Space Retrieval

#### 1.2.4 Probabilistic Information Retrieval

#### 1.2.5 Evaluating Information Retrieval

#### 1.2.6 Query Expansion

##### 1.2.6.1 User Relevance Feedback

##### 1.2.6.2 Global Query Expansion

### 1.3 Embedding Techniques

#### 1.3.1 Latent Semantic Indexing

#### 1.3.2 Latent Dirichlet Allocation

#### 1.3.3 Word Embeddings – skipgram, CBOW

#### 1.3.4 Fasttext

#### 1.3.5 Glove

# Part 2: Recommender Systems

## Part 2: Recommender Systems

- 2.1 Collaborative Filtering
- 2.2 Content-based Recommendation
- 2.3 Matrix Factorization
- 2.4 SLIM, Sparse Linear Methods
- 2.5 Evaluation of Recommender Systems

# Part 3: Information Extraction

Part 3: Information Extraction
3.1 Named Entity Recognition
3.1.1 Keyphrase extraction
3.1.2 Named entity recognition (NER)
3.1.3 Entity Disambiguation
3.2 Document Classification
3.2.1 kNN
3.2.2 Naïve Bayes Classifier
3.2.3 Classification Using Word Embeddings
3.2.4 Transformer Models
3.3 Knowledge Representation
3.3.1 Knowledge Representation
3.3.2 Semi-structured data
3.3.3 The Semantic Web
3.3.4 RDF - Resource Description Framework
3.3.5 Semantic Web Resources
3.4 Information Extraction
3.4.1 Information extraction (IE)
3.4.1.1 Hand-written patterns
3.4.1.2 Supervised machine learning
3.4.1.3 Bootstrapping
3.4.1.4 Distant supervision
3.4.1.5 Matrix Factorization
3.4.2 Taxonomy Induction

# Part 4: Graph Analytics

Part 4: Graph analytics
4.1 Link-Based Ranking
4.1.1 PageRank
4.1.2 Hyperlink-Induced Topic Search (HITS)
4.2 Mining Social Graphs
4.2.1 Louvain Modularity Algorithm
4.2.2 Girvan-Newman Algorithm
4.2 Knowledge Inference
4.2.1 Label Propagation
4.2.2 Link Prediction

# Part 5: Data Indexing and Mining

## Part 5: Data Indexing and Mining

### 5.1 Indexing for Information Retrieval

#### 5.1.1 Inverted Index

#### 5.1.2 Web-scale Indexing: Map-Reduce

#### 5.1.3 Link Indexing

#### 5.1.4 Distributed Retrieval

##### 5.1.4.1 Fagin's algorithm

##### 5.1.4.2 Threshold algorithm

### 5.2 Introduction to Data Mining

#### 5.2.1 Association Rule Mining

##### 5.2.1.1 Association Rules

##### 5.2.1.2 Scoring Function

##### 5.2.1.3 Apriori Algorithm

##### 5.2.1.4 FP Growth