Distributed Information Systems
Fall Semester – 2024
CS-423

Time and Place
Lecture: Thursday 10:15 – 12:00, CM2
https://epfl.zoom.us/j/63657026419

Exercise: Thursday 12:15-14:00, CM2

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/63657026419
  - 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/1652

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 of 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 — BC 108

Assistants
- Romanou Angelika — angelika.romanou@epfl.ch
- Negar Foroutan — negar.foroutan@epfl.ch
- Borges Ribeiro Beatriz Maria — beatriz.borges@epfl.ch
- Ismayilzada Mahammad — mahammad.ismayilzada@epfl.ch

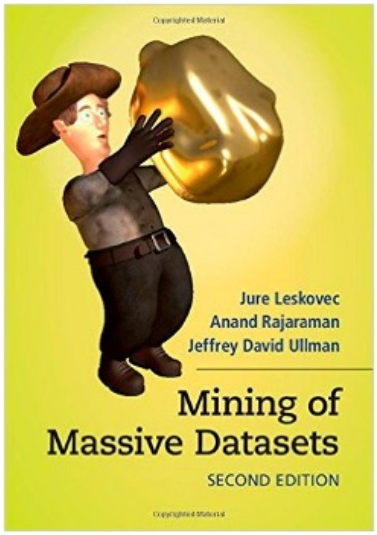
References

Parts of the course are based on the following text books

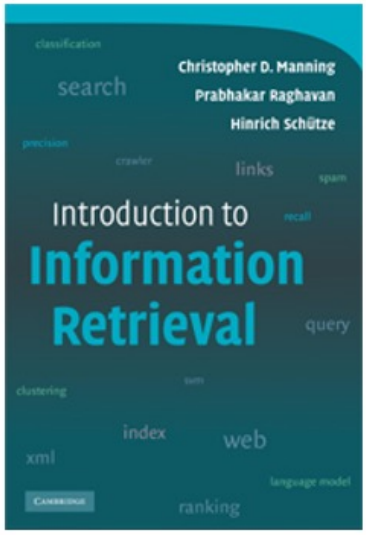
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval (Acm Press Series), Addison Wesley, 1999.
- Jiawei Han, Data Mining: concepts and techniques, Morgan Kaufman, 2000.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge 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://nlp.stanford.edu/IR-book/

Part 1 Information Retrieval

Part 2: Recommender Systems

Part 3: Information Extraction

Part 4: Graph Analytics

Part 5: Data Indexing and Mining