Distributed Information Systems
Fall Semester – 2020
CS-423

Time and Place
Lecture: Monday 13:15-15:00 Zoom
https://epfl.zoom.us/j/96988744528

Exercise: Monday 15:15-16:00 Discord

Karl Aberer
Distributed Information Systems Laboratory

Welcome to a Special Semester!

Everything is different

We will be working largely online

Everything might evolve and change throughout the semester

Program of today:

- Hour 1: Everything you need to know on the organization
- Hour 2 & 3: An overview of main concepts that will be covered during the semester

Goals of the Course

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

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

Understand which are **key problems** relevant for DIS?

- e.g. modeling, storage, indexing, retrieval, mining, recommending, integration, etc.

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

- e.g. vector space retrieval, association rule mining, schema mapping etc.

Assumption: basic knowledge in databases, e.g. from CS-422 Database Systems

Focus of the Course

Master important **Models and Algorithms** for representing and processing information:

Data Science

Conceptual foundations to practically use tools and platforms for Data Science

- Complementary to *Applied Data Analysis* by Bob West

Other Related Courses

In synergy with

- Applied Data Analysis

Complementary to

- Introduction to database systems
- Database systems

Some overlaps possible with

- Introduction to machine learning
- Machine learning
- Introduction to natural language processing
- Internet analytics

Which masters 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

Did you take Applied Data Analysis?

1. Yes
2. No

The Course - Lecture

Webinar https://epfl.zoom.us/j/96988744528

- Standard online ex cathedra lecture
- Use Chat tool to ask questions
  - Will be answered in public
- Alternatively QA tool
  - Will be answered privately by assistants
- Quizzes using Zoom (anonymous)

Video recording

https://tube.switch.ch/channels/45c71cb4

Materials

Web platform: Moodle

- General announcements will be published on Moodle
- Course notes and exercises will be published on the Web in advance: https://lsir.github.io/DIS/

Exercises

Weekly exercises

- 2-3 problems to solve

Most problems will be (simple) programming exercises

- Uses Python
- Focus on understanding the techniques (not programming skills etc)

Exercises and exam questions from previous years will be made available as well

Exercise Platform

We will be using Discord for communicating with assistants on exercises

https://discord.com/invite/rQ7cen3

- Dedicated per-topic channels
- Check whether question has already been answered
- Life answers during exercises session

“Continuous control”

Due to the current situation no graded continuous control

But

- Midterm programming exercise
- 2 Quizzes

Will allow to test your skills

Grading

Final Exam: 100%

- Questions similar to the question in exercises and quizzes
- will assume you attended the lecture
- will assume you did the exercises
- 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)

Are you on Campus today?

1. Yes
2. No

Are you planning to be on campus when it is your turn?

1. Yes
2. No

Have you already used Discord?

1. Yes
2. No

Schedule

Week	Date	Cont. Eval.	Area	Topic
1	14 September 2020		**Introduction**	Distributed Information Systems - An Overview
2	21 September 2020			*Holiday*
3	28 September 2020		**Information Retrieval**	Basic Text Retrieval Models
4	05 October 2020			Indexing and Relevance Feedback
5	12 October 2020	Prog. Midterm		Advanced Retrieval Methods
6	19 October 2020			Word Embeddings and Link-based Retrieval
7	26 October 2020		**Data Mining**	Frequent Itemset Mining
8	02 November 2020			Clustering and Classification
9	09 November 2020	Quiz		Classification Methodology
10	16 November			Document Classification and Recommender
11	23 November 2020			Social network mining
12	30 November 2020		**From Documents to Knowledge**	Semantic Web
13	07 December 2020	Quiz		Entity and Information Extraction
14	14 December 2020			Data Integration and Knowledge Graphs

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

– Chi Thang Duong	thang.duong@epfl.ch	BC 130
– Tugrulcan Elmas	tugrulcan.elmas@epfl.ch	INN 134
– Smeros Panayiotis	panayiotis.smeros@epfl.ch	BC 142
– Jeremie Rappaz	jeremie.rappaz@epfl.ch	INM 035

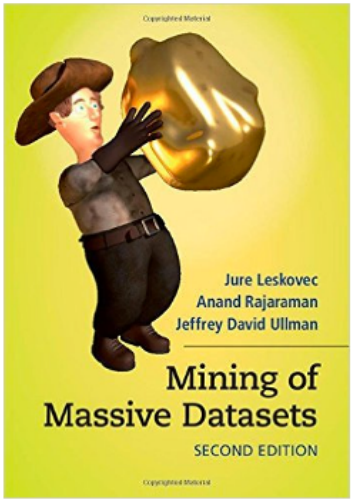
References

Parts of the course are based on the following text books

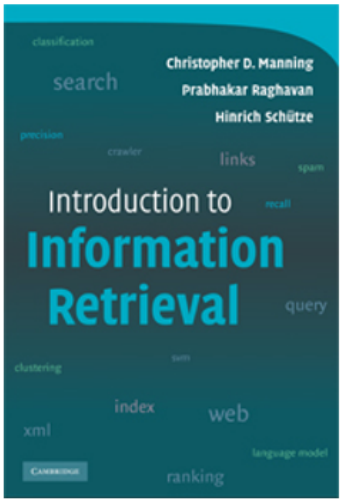
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval (Acm Press Series), Addison Wesley, 1999.
- Jiawei Han, Data Mining: concepts and techniques, Morgan Kaufman, 2000.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge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

Exam Date