Course Introduction

[DAT640] Information Retrieval and Text Mining

Krisztian Balog
University of Stavanger

August 20, 2024

Instructors

Krisztian Balog
Lecturer

Petra Galuscakova
Lecturer

Weronika Lajewska
Teaching assistant

About me

- Professor at the University of Stavanger, Norway [0.7fte] (2016–)
- Staff Research Scientist at Google DeepMind [0.5fte] (2021–)
- Research on the use and development of information retrieval, natural language processing, and machine learning techniques for intelligent information access tasks
  - Conversational information access
  - Transparent and explainable search and recommender systems
  - User simulation
  - Personal knowledge graphs
- Teaching this course since 2019

What about you?

What is this course about?

- Techniques and methods for processing, mining, and searching in (massive) text collections:
  - **Previously:** ranking, classification, clustering
  - **This year:** ranking, ~~classification~~, ~~clustering,~~ LLMs, text generation
- **Information retrieval** (search engines): Analysis, organization, storage, and retrieval of information
- **Text mining** (text analytics): Deriving high-quality information from textual data by analyzing trends and patterns

Information Access Tasks

- *Pull mode*: user takes the initiative and uses a search engine to find information
- *Push mode*: the system takes the initiative and recommends relevant information to the user
- Search and recommendation are "two sides of the same coin" and involve:
  - Modeling a user's information need and preferences
  - Matching an information object with a user's interest
  - Ranking items accurately
  - Learning from user feedback
  - Evaluating a ranked list to assess its utility to a user
- *Mixed initiative*: conversational assistants facilitate both search and recommendation via natural language interactions

Prerequisites/requirements

- No formal prerequisites, **but** you are expected to know
  - Algorithms and data structures
  - Databases (basic concepts)
  - A bit of statistics
  - Python

Course format, expectations

Format

- **Lectures**
  - Tuesdays (weeks 34-46) and Wednesdays (weeks 34-38)
  - In-person, no recordings/streaming, slides are shared
- **Individual assignments** (weeks 34-39)
  - 10 assignments in total, automatically graded
  - Friday labs are dedicated to working on the assignments and getting help from TAs
- **Group project** (weeks 39-46)
  - Work in groups of 2-3 on a project (TBD)
  - Wednesdays are Fridays are dedicated to group project work; lecturer(s) available to give feedback
  - Grading based on functionality demonstrated in person (time slots to be booked in advance)
- Schedule is subject to changes

w	Tue	Wed	Fri
34	L1	L2	
35	L3	L4	
36	L5	L6	
37			
38	L7	L8	
39	L9		
40	L10		
41	L11		
42	L12		
43	L13		
44	L14		
45	L15		
46	L16		

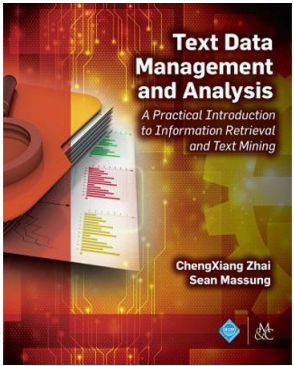
Course content

- Part I: Classic IR (L1-L4)
  - Text representation, similarity, preprocessing
  - Search engine architecture and basic retrieval models
  - Advanced retrieval models
  - Retrieval evaluation
- Part II: Neural IR (L5-L7)
  - Neural networks for language
  - Word embeddings and dense retrieval
  - Transformer-based models, LLMs
  - Transformers in IR
- Part III: Advanced topics (L8-L16)
  - Conversational information access
  - Knowledge bases and entity retrieval
  - Entity linking
  - Text generation
  - Evaluation and benchmarking
  - User simulation

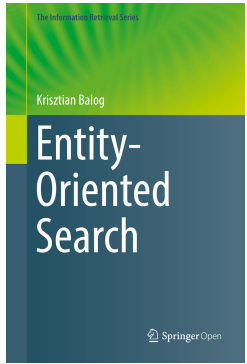
Resources

- **Everything is on Canvas**
  - Course schedule and curriculum
  - Lecture slides
  - Exercises and solutions
  - Assignments
  - etc.
- **Bring your own device (laptop)!**
  - Python 3.9+ (ideally, Anaconda distribution)

Textbooks



Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining (Zhai and Massung), ACM and Morgan & Claypool Publishers, 2016.



Entity-Oriented Search (Balog), Springer, 2018.

Grading

- **Project work** (40%)
  - 30% from individual assignments
  - 70% from group project work
  - No re-sit option!
- **Written exam** (60%)
  - Digital exam (Inspera)
  - Open book
  - Mixture of exercises, multiple choice, and essay questions
- Both need to be a pass to get a (passing) grade for the course!

	weight	mark
Project work	40%	A-F
Written exam	60%	A-F

Lectures

- Mixes theory, discussion, and hands-on exercises
- Slides will be made available on Canvas beforehand

Question/Discussion

When you see this box, it means you need to think about answering the question (and interact).

Exercise

When you see this box, you'll need to go to the referred exercise on Canvas.
Solutions will be published on Canvas after the lecture.

Individual assignments

- 10 assignments in total, all released at once (before Friday)
- Assignments are worth 40 points in total, max. 30 may be collected
- Points vary based on difficulty
- Single submission, graded automatically
- Deadlines are strict, no extensions, no exceptions!
- Assignments account for **30% of the project work final grade**
- Fridays labs are dedicated to working on assignments—this is the time and place to get help
- Delivery deadlines on Mondays at 08:00 (as requested by students)
- Three delivery cutoffs: Sep 2 (A1-A3), Sep 16 (A4-A6), and Sep 30 (A7-A10)

Assignments workflow

- For each assignment we provide
  - a task description
  - a requirements file containing list of python libraries
  - a skeleton code (Python file) that needs to be completed
  - a set of public tests you can use to verify your solution
- Assignments are graded automatically, using a combination of public and hidden tests
  - **public tests** are released with the assignment; if your solution passes these, it is likely correct
  - **hidden tests** will be used to test your solution after submission; these typically contain larger inputs/datasets, corner cases, or other inputs in order to test that you fully understood the methods and/or followed the instructions
- Upload completed Python file to Canvas

Group project

- Work in groups of 2 or 3 (TBD)
- Project: developing a conversational music recommender system
- Full freedom in choosing the programming language and tools used to implement the required functionality
- New functionality requirements are released each week
- Groups can get weekly feedback during the Wed and Fri slots
- Grading is based primarily on the working functionality demonstrated in person to the lecturers during dedicated timeslots
- Accounts for **70% of the project work final grade**

Contact

- Fridays labs are for working on the assignments. This is **the** time to get help!
  - Contacts for assignment-related issues: `weronika.lajewska@uis.no`
- If you need to talk to the lecturer, make an appointment via email (`krisztian.balog@uis.no`). No drop-ins unannounced!

Some simple rules

- We're back to regular teaching, which means physical lectures and in-person communication. (We're slow on email.)
- Deadlines are strict and immutable, no extensions (don't even ask)!
  - Application for exceptions has to be submitted 72 hours prior the deadline **in writing** with necessary **proof attached** (medical certificate, etc.)
- Discussions while working on assignments is OK. Copying someone else's solution is considered cheating and will get both of you into **serious** trouble (cases will be reported to the Exam Office)

Question

What is the value of in-person classes in 2024?

FAQ

- **Is it obligatory to attend the lectures?**
  No. However, it is highly recommended in order to get the full learning experience (especially that there is no recording/streaming). Also, note that everything that is covered in the lectures may be asked back at the exam.
- **Is it obligatory to attend the labs?**
  No. However, this is the time and place to work and get help on the assignments.

Question

Do you have any questions?

Exercise

Exercise

E0-1 Python basics