Agenda

1. Organization

2. Lab Project

Organization
Communication

- ❏ Slides, Announcements & Materials will be available at
  `temir.org/teaching/information-retrieval-ss24/`
  `information-retrieval-ss24.html`

- ❏ Communication channels are Discord and email
  - – Official announcements via Mail (check your student mails regularly!)
  - – Discord for Q&A and group communication

Organization
Lab Sessions

❏ Lab sessions throughout the semester

❏ Not every week will have a lab session taking place (see course page)

❏ We are available for questions outside the scheduled times via Discord

– Feel free to use the discussion board!

Organization
Lab Sessions

❑ Lab sessions throughout the semester

❑ Not every week will have a lab session taking place (see course page)

❑ We are available for questions outside the scheduled times via Discord

– Feel free to use the discussion board!

https://discord.gg/3f79K635XD

Lab Project
Overview

Goal

- ❑ Building and evaluating an information retrieval system for a specific domain
  - – Scientific requirements: related work search, hypothesis testing, writing
  - – Technical requirements: data handling, indexing, selection and implementation of suitable retrieval models, evaluation of search quality
  - – Submission of a written report and software on the TIRA platform

Lab Project
Overview

Goal

- ❑ Building and evaluating an information retrieval system for a specific domain
  - Scientific requirements: related work search, hypothesis testing, writing
  - Technical requirements: data handling, indexing, selection and implementation of suitable retrieval models, evaluation of search quality
  - Submission of a written report and software on the TIRA platform

Deliverables & Grading

- ❑ Lab is organized around 4 milestones throughout the semester
- ❑ At each milestone, we expect each group to hand in a deliverable
- ❑ First 3 milestones have to be passed; final milestone is graded (30% of module grade)

Lab Project
Task & Data

❑ Task: Scholarly Search

– Given a queried (scientific) topic, search for related papers

– Examples: Google Scholar, SemanticScholar

❑ Data: Union of the IR and ACL Anthology

– Two existing scholarly search engines as basis

  • Information retrieval papers: https://ir.webis.de/anthology/

  • Natural language processing papers: https://aclanthology.org/

– All 126 958 papers (title + abstract) from the IR and ACL Anthology

– Contains titles, abstracts, and some other metadata

❑ TIRA as an evaluation platform & leaderboard

Lab Project
Milestones

❏ 4 Milestones in total, each with an associated deliverable

❏ Mirror the development process of a search engine:

   – Milestone 1: Defining the Task ($\rightarrow$ Queries)

   – Milestone 2: Collecting Evaluation Data ($\rightarrow$ Relevance Judgments)

   – Milestone 3: Build a Prototype ($\rightarrow$ Initial System)

   – Milestone 4: Evaluate & Improve ($\rightarrow$ Refined System)

❏ Final results of the lab:

   – retrieval system for scholarly search (software submission)

   – scientific investigation of its properties (paper submission)

Lab Project

Create topics for the supplied search task and data collection.

- ❑ A topic is a description of a users' information need
  - – A **text** entered into the IR system as query
  - – A **description** of the underlying information need
  - – A **narrative** describing what is relevant to the query

ID	1
Text	retrieval system improving effectiveness
Description	What papers focus on improving the effectiveness of a retrieval system?
Narrative	Relevant papers include research on what makes a retrieval system effective and what improves the effectiveness of a retrieval system. Papers that focus on improving something else or improving the effectiveness of a system that is not a retrieval system are not relevant.

- ❑ **Due Date:** 22.04.2024
- ❑ **Deliverable:** Valid topic file (XML, see course page for example)

Lab Project
Milestone II: Relevance Assessments

Assess the relevance of documents retrieved for your topic

❑ Given your topics, we will supply a set of retrieved documents

- Retrieval is done by pooling several baseline systems
- For each topic and system, documents are retrieved and pooled

❑ Annotate these documents w.r.t. their relevance to the topic

Query ID	1
Text	retrieval system improving effectiveness
Description	What papers focus on improving the effectiveness of a retrieval system?
Narrative	Relevant papers include research on what makes a retrieval system [...]
Document ID	2005.ipm_journal-ir0anthology0volumeA41A1.7
Document	In this paper we will present a language-independent probabilistic model [...]
Relevancy	1

❑ Due Date: 06.05.2022
❑ Deliverable: Completed batch on the annotation platform

Lab Project
Milestone III: IR System

Build and evaluate your own IR system using
your topics and relevance assessments.

❏ Implement your IR system

- Training data will be supplied; compute resources available
- Final system should be deployed to the TIRA platform

❏ Evaluate your IR system

- The previously annotated topics are used for testing
- Testing is carried out using the TIRA platform

❏ Shortly reflect on the assignment in a written report

❏ Due Date: 10.06.2024
❏ Deliverable: TIRA submission

Lab Project

Derive and test a research question with your IR system.

❑ Come up with a hypothesis that can be tested with your system

❑ Conduct statistical analysis to address the hypothesis
 – Hyp. should separate two configurations (modified c., control c.)
 – Hyp. should motivate a distribution over both groups to test against

❑ Summarize your system and findings in a short report

"System effectiveness will be higher when boosting
the rank of documents under 1000 characters."

❑ Due Date: 01.07.2024
❑ Deliverable: Report (approx. 3 pages + references)

Questionnaire

Raise your hand if you...

- ❑ ... have used Python before?

- ❑ ... have preexisting knowledge in ML?

- ❑ ... have worked on data analysis before?

- ❑ ... have done data annotation before?

- ❑ ... have done scientific writing before?

- ❑ ... have worked in a cluster environment before?

- ❑ ... have used Docker before?

Project Groups

- ❑ You can work in groups of up to 4 people

- ❑ Already know whom you want to work with?
  - – Yes! → see below
  - – No! → raise your hand to be assigned to a group

- ❑ Send us a mail with names of all group members until next week!

- ❑ Each group will then receive:
  - – A unique group name
  - – A Discord channel
  - – A TIRA account

Formulating Topics

What makes a good topic?

Formulating Topics

What makes a good topic?

- ❑ Reflecting real information needs and congruent to the task
- ❑ Retrievability (supported by document collection)
- ❑ Helps in distinguishing systems
- ❑ Hard vs. Easy?

Formulating Topics

Reflecting real information needs

- ❏ Option 1: Browse through lecture slides and try to collect more information on some concept that interests you
- ❏ Option 2: Imagine you have to write exam questions. An exam question could form the basis of an information need.
- ❏ Option 3: Combine IR concepts to other concepts that are important to you / our society.

Formulating Topics

Retrievability: Test that relevant documents exist for your information need

- ❑ We provide a baseline retrieval system in a Google Colab for experimentation: colab.research.google.com/drive/1jXf8Yi1RgbJnvMCA6hWF9z7hgB9ln5AQ

- ❑ You can try out multiple query variants for your topic to adjust for the difficulty of the topic

    – E.g., „PageRank "vs. „measure importance of web pages "