

# Information Retrieval

Exercise – Winter term 2025

`klara.gutekunst@uni-kassel.de`

# Agenda

1. Organization
2. Lab Projects
3. Cranfield Paradigm
4. Formulating Topics
5. Tutorial
6. Next Steps

# Organization

## Communication

- ❑ Slides, Announcements & Materials on the course website [[temir.org](https://temir.org)]
- ❑ Questions and communication via Discord and email [[discord.gg](https://discord.gg/temir)]  
(please use your full name as the server nickname)



# Organization

## Lab / Exercise

- ❑ Joint collaborative exercises with other universities:  
Jena, Kassel, Radboud, ...
- ❑ Lab project as group work (3 people per group)  
(i.e., about 1–2 groups this semester)
- ❑ Lab/exercise sessions throughout the semester
  - **Tuesday, 14:00 – 15:30, Room 1114**
  - Alternating weekly between lab and theory
  - Lab
    - Each group shortly presents their current progress
    - Discuss open questions and next steps
  - Exercise
    - Exercises for lecture contents
- ❑ We are also available for questions outside the scheduled times via [Discord](#)
  - Feel free to use the [discussion board](#) to discuss across universities!

# Lab Project

## Overview

- ❑ Goal
  - Build & evaluate an information retrieval system for web search
  - Search for related work, learn data handling, indexing, select & implement suitable retrieval models, evaluate search effectiveness
  - Identify (and solve) retrieval errors
  - Submit written report and software on the [TIRA](#) platform
  - Participate in international research

# Lab Project

## Overview

### □ Goal

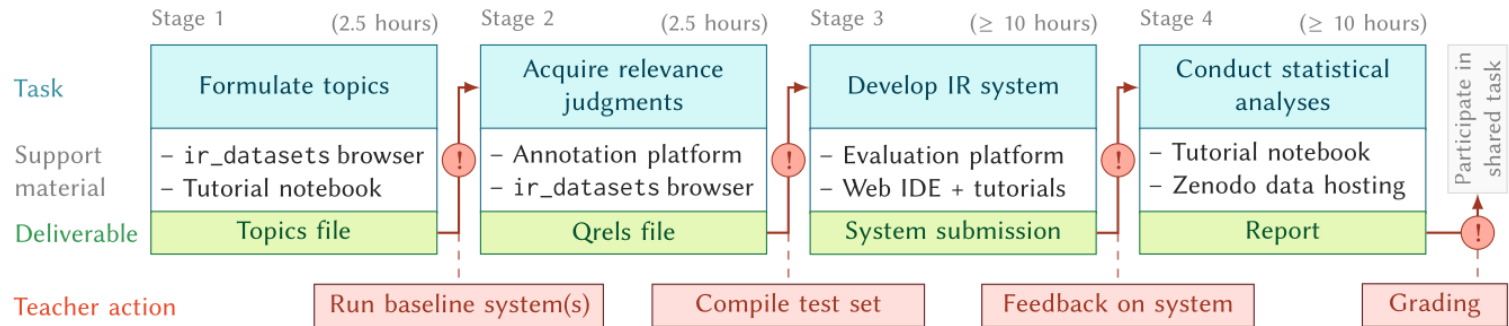
- Build & evaluate an information retrieval system for web search
- Search for related work, learn data handling, indexing, select & implement suitable retrieval models, evaluate search effectiveness
- Identify (and solve) retrieval errors
- Submit written report and software on the [TIRA](#) platform
- Participate in international research

### □ Resources

- Test collection: Open Web Index [[website](#)]
- [TIRA](#) as an evaluation platform & leaderboard [[website](#)]
- `ir_datasets` browser [[website](#)]
- Tutorials [[list](#)]

# Lab Project

## Overview



Overview of the course stages [SIGIR'24]

- ❑ Lab organized in 4 stages throughout the semester
- ❑ Each stage is finished by a deliverable  
→ Groups hand in their intermediate work
- ❑ Deliverables are exchanged with other universities

# Lab Project

## Task & Data

- ❑ Task: Web search
  - Aka: What you search everyday...
  - Very broad set of topics
- ❑ Data: Open Web Index [[website](#)]
  - OpenWebSearch.eu project [[website](#)]
  - > 1195 TB crawled data, > 34 TB indexed data [[website](#)]
  - Temporal slices (single days) of crawled, preprocessed and indexed data
- ❑ Further evaluation data will be collected in the lab itself



# OpenWebSearch.eu

Search for information ...

... in newspapers



Image: Generated by Firefly Image 4

... online



Image: Flaticon.com

Problem: [[Open Web Index website](#), [Michael Granitzer Keynote 2025 \(video\)](#)]

- ❑ Market **oligopoly**
- ❑ Current existing web indices: Google (USA), Bing (USA), Yandex (Russia), Baidu (China)
- ❑ 90% of queries are done via Google
- ❑ Indices are proprietary and not openly accessible
- ❑ Search engines can send requests to indices via API only
- ❑ Ranking methods via their APIs are not disclosed

Goal: Create free, open and unbiased **Web Index** [[video](#)]

# Lab Project

## Stage 1: Formulate topics

Create topics for the supplied search task and data collection.

- ❑ A topic is a description of a user's information need
  - A *text* that could be entered into the search system as a query
  - A precise *description* of the underlying information need
  - A *narrative* describing what is relevant to the topic (and what not)
- ❑ **Due Date:** Thursday, 06.11.2025, 23:59
- ❑ **Deliverable:** valid topics submitted [online](#)

---

|             |  |
|-------------|--|
| Topic 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. |

---

# Lab Project

## Stage 2: Acquire relevance judgments

Judge the relevance of documents retrieved for your topic.

- ❑ Given your topics, we will retrieve documents
  - Top-10 for each topic and baseline system
  - Some manually added documents
- ❑ Annotate these documents w.r.t. relevance to the topic (read the *narrative*)
- ❑ **Due Date:** tba
- ❑ **Deliverable:** Relevance judgments finished on [Doccano](#)

---

|                 |  |
|-----------------|--|
| Topic ID        | 1  |
| Query           | retrieval system improving effectiveness                                 |
| Topic desc.     | What papers focus on improving the effectiveness of a retrieval system?  |
| Topic narrative | Relevant papers include research on what makes a retrieval system [...]  |
| Document ID     | 2005.ipm_journal-ir0anthology0volumeA41A1.7                              |
| Document text   | In this paper we will present a language-independent probabilistic [...] |
| Relevance       | 1 (relevant)   |

---

# Lab Project

## Stage 3: Develop information retrieval 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 deployed to the **TIRA** platform
- ❑ Evaluate your IR system
  - Topics from **Open Web Index** and **TREC** are used for validation
  - Tests carried out on the **TIRA** platform
- ❑ Summarize your system and findings in a preliminary report
- ❑ **Due Date:** tba
- ❑ **Deliverable:** Preliminary report (max. 2 pages), **TIRA** submission

# Lab Project

## Stage 4: Conduct statistical analyses

Derive and test hypotheses with your previously constructed IR system.

- ❑ Look at types of retrieval errors
- ❑ Observe your system's behavior compared to baselines
- ❑ Formulate 2 hypotheses
- ❑ Hypotheses must be testable (more on that later)
- ❑ **Due Date:** tba
- ❑ **Deliverable:** Report (max. 2 pages)

# Lab Project

## Optional: Participate in a shared task

Polish your system and findings to compete in an international shared task.

- ❑ Shared task initiatives: TREC, CLEF, NTCIR
- ❑ Task might differ slightly
- ❑ **Due Date:** Shared task's deadline
- ❑ **Deliverable:** Shared task submission & paper

# Questionnaire

Raise your hand if you...

- ☐ ... have used Python before?
- ☐ ... have previous knowledge in ML?
- ☐ ... have worked on data analysis before?
- ☐ ... have done data annotation before?
- ☐ ... have done scientific writing before?
- ☐ ... have done statistical analyses before?
- ☐ ... have used Docker before?
- ☐ ... would like to participate in an international shared task?

# Cranfield Paradigm

- ❑ Retrieval systems evaluated using a set of:
  - Information needs (topics)
  - Documents
  - Relevance judgments
- ❑ Assumption: Each document's relevance can be assessed independently
- ❑ Applied in the evaluation of most shared tasks in IR

Do you see any problems with the Cranfield paradigm?



# Formulating Topics

What makes a good topic?

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

# Good Topics

---

|             |  |
|-------------|--|
| ID          | 794  |
| Text        | pet therapy  |
| Description | How are pets or animals used in therapy for humans and what are the benefits?  |
| Narrative   | Relevant documents must include details of how pet or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it. |

---

Why is this a good topic?

# Bad Topics

---

|             |  |
|-------------|--|
| ID          | 123  |
| Text        | pet therapy  |
| Description | Which physical therapies are there for pets and when is therapy necessary? |
| Narrative   | Relevant documents list pros/cons of therapies for pets.                   |

---

Why is this a bad topic?

# Tutorial

- ❑ Prepared Jupyter notebook [[notebook](#)] [[codespace](#)]
- ❑ Run on your own computer, on GitHub Codespaces, or on Google Colab  
[\[colab.research.google.com\]](https://colab.research.google.com)

Let's get started!

# Next Steps

## Project Groups

- ❑ Form groups of 3 people
  
- ❑ Each group will receive:
  - A unique group name (be creative!)
  - A Discord channel
  - A **TIRA** account
  
- ❑ **Email me** the names of all group members and Discord handles until Thursday, 31.10.2025, 23:59

# Next Steps

## Assignment

- ❑ Form project groups (email with names and Discord handles)
- ❑ Read about [Open Web Search](#), [Open Web Index](#) and [TREC](#)
- ❑ Complete the notebook about topics
- ❑ Write & submit your own topics **TODO** [\[form\]](#)
  - One topic per team member
  - **Deliverable:** valid topics submitted online
  - **Due Date:** Thursday, 06.11.2025, 23:59