Project Specification

The project involves conducting an IR experiment using advanced methods (e.g., Transformers). The experiment must be in the scope of information retrieval, containing evaluations and some advanced concepts covered in class. The project consists of four parts, graded separately:

- **Design Document** - Sketching your project idea.
- **Project** - Your implementation of an advanced IR project, done in groups.
- **Presentation** - presentation of your work in class, done with your group.
- **Report** - a description of your system, your experiments and findings in the form of a scientific paper.

The project will be conducted in groups of 4 people.

Programming Language: Python

Important Dates

- Group Formation until: 21.10.2025 23:59
- Design Document due: 02.12.2025 23:59 (early submission: 11.11.2025 23:59)
- Project Submission due: 13.01.2026 23:59
- Project Presentation due: 13.01.2026 23:59
- Project PDF-Report due: 13.01.2026 23:59
- Project Presentations: January lecture units (13.01., 20.01., 27.01.) - slots need to be booked in TC

Design Document (15p)

1 document per group
The document should present the general idea in a **maximum of three** A4 pages (including drawing/diagram, written content, and title page) and must include:

- Title
- Group Number

- Members + tentative roles (= responsibilities)
- Abstract
- Visual Depiction - see an example in Fig. 1

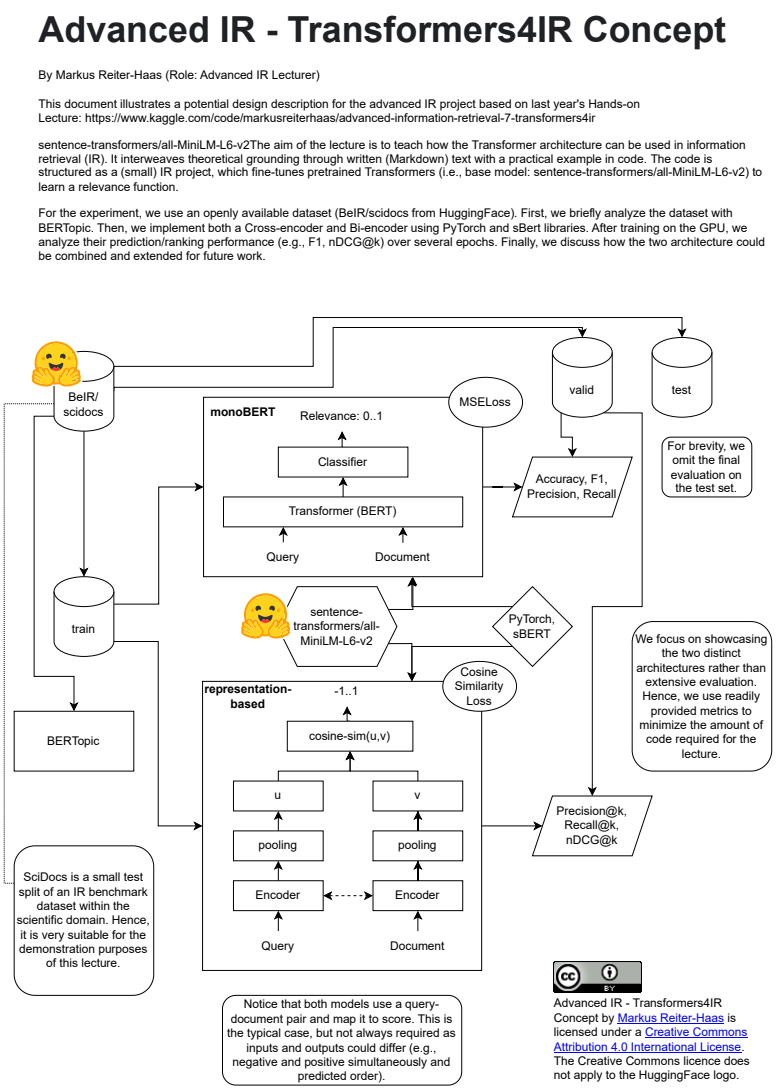

Figure 1: Example Sketch from WS22

The design document must clarify:

- **Idea** (goal, e.g., problem to solve formulated as research question)
- **Main Task** (focus of your work)

- **Dataset + Processing** (resources used)
- **Methods/Models** (how you intend to solve the problem)
- **Evaluation** (how results are measured)

Submission as PDF in TeachCenter.

Project (30p)

The project should be developed with your group and must be an advanced IR system. Submit the code in advance and provide a link to your code repository in TC.

Repository

The repository should be public (e.g., GitHub, Zenodo): **Be proud of your work.** If not possible, contact us at least 2 weeks in advance to share the code privately.

Important points that need to be considered:

- Include all your code parts for reproducibility
  - Data loading, preprocessing, training, evaluation, and plot generation
  - Code must be clean, well-structured, and readable, with meaningful naming conventions and comments explaining non-obvious logic
  - Do not include any personal information or API keys!
- Include all your generated plots (either separate or within Jupyter notebooks)
  - Ensure that all plots are clearly readable, with appropriate titles, axis labels, and legends where necessary
- Include a `README.md` file
  - At the beginning, include a brief description of the project.
  - Describe the project structure and experimental setup
  - Explain how we could reproduce your results
  - Briefly describe the dataset (e.g., source and location)
- *Optionally:* the dataset itself
- Ensure your folder structure is organized and files are clearly named
- Include a `requirements.txt` listing all dependencies needed to run your code
- If external code or data is used, provide proper citations or licenses

Top projects will get the possibility to be highlighted on the lab website:
https://socialcomplab.github.io/advancedIR-2023-showcase/
The affected groups will be informed after the presentation and asked for their permission.

Presentation (15p)

Submit slides as PDF in TC. **Each team member** must be present and actively take part in a **speaking role** during the presentation. The presentation should include the **5 slides specified below**.
You can include additional slides for plots, figures, or other additional material that supports your presentation. **Do not compromise the quality** of your plots, figures, or images so that they fit on as few slides as possible.
The duration of the presentation will be **9 minutes** and an additional **3 minutes** for questions.
Practice your presentation in advance and make sure to time yourself. The time limit is strict, if you exceed it, your presentation will be cut off.

Presentation Content

- Title slide
- Introduction (RQ/motivation)
- Data + Methods (+ optional theoretical background)
- Results (+ analysis/interpretation)
- Conclusion (incl. limitations/biases)

Title page must include:

- Name of the project (defined by group)
- Group number
- Group members + team roles
- Link to repository

Report (20p)

Write a report on your experiments in the format of a scientific paper **(maximum 4 pages of text, 12pt font, excluding the title page)**.

Important: The page limit applies only to the text. Do not compromise the quality of plots, figures, or images to fit everything into 4 pages. Your report may exceed 4 pages overall, as long as the total text content does not exceed 4 pages.

For better structure and readability, you may place all plots and figures in an appendix at the end, keeping the first four pages exclusively for text.

Submit as PDF in TC. The structure should be:

- **Title page** with group number and names
- **Introduction**: Motivation and research question

- **Related Work**: Brief literature or project discussion
- **Experiments and Results**: Conducted experiments and findings
- **Conclusion**: Key takeaways and future work

Tips

An important aspect of an IR experiment is the analysis and evaluation of IR systems/models. For instance, an image generator is not IR-related and its evaluation may be unsuitable for this course.

Organization:

- Start project work before handing in the design document. Test if design components (e.g., dataset) are feasible.
- Use feedback from the design document to improve the project.
- Seek help early if needed (e.g., in TeachCenter).

Programming:

- Jupyter is recommended, but for long tasks, scripts may be better.
- Use Numpy or PyTorch instead of plain Python.
- For large datasets, create a smaller subset and report statistics in the presentation.
- Attend lectures, especially on Transformers.

Presentation:

- Stick to time. Rehearse beforehand.
- Engage with the audience and observe time signals.
- Nervousness is normal; preparation is key.

Grading Considerations:

- A clean setup is crucial, though code style is not graded.
- The setup must avoid issues like data leakage.
- Results require analysis/interpretation beyond mere presentation.
- Negative results are valid but must be carefully analyzed and validated. For example, a high-performing random model indicates a methodological issue, not a valid negative outcome.