Information Retrieval

Projects

1 General Information

Within the project a student must implement a specialized search system, i.e., a search system with a certain topic. For example, wikipedia search, news search, code search, etc. The topic is chosen by a student.
To implement a search system, a student must implement the following steps:

1. Data acquisition (crawling and beyond)
2. Data processing and storage
3. Search functionality and features (see below)
4. Search interface
5. Qualitative and quantitative evaluation

The implemented search system should have at least one special feature. Some examples of features are:

- Entities
- Time
- Links

- Personalization
- Diversification
- Opinion
- Many other, so let's discuss

2 Outcomes and Reporting

Each project must result in the following:

- A working search system with at least one special feature
- Mid-term presentation (November 17, preliminary)
- Final presentation (December 15, preliminary)
- Final report (December 22)

2.1 Project report

The report contributes 40% to the final grade. It must be written in English and use LaTeXand the `sigconf` template from `https://www.acm.org/publications/proceedings-template`. The report must be not shorter than 4 and not longer than 8 pages. The report must contain the following information:

- Project description, where the search system, data and features are clearly outlined.
- Architecture/design of the system with explanation and justification of the choices you made when designing the system.
- Problems you encountered during implementation and corresponding solutions (if any).
- Reflection on the project.
- Summary of the work done so far and directions for future development.

- Two or three screenshots of your system (can go beyond 8 pages).

The report must be updated continuously during the project. In the end of each step (see Sections 1 and 2.3) an updated report must be submitted to `mailto:ir.spbau@gmail.com`. **IMPORTANT:** please mention the words "project report" and the name of the step in the title of the mail (e.g., "project report, data acquisition", "project report, search and features", etc.).

The final report containing all required information must be submitted no later than December 22.

2.2 Project presentations

The mid-term project presentation (November 17, preliminary) is a 7 minute presentation + 5 minutes for questions. In this presentation, a student describes the problem, chosen approach, current status, and what still needs to be done. This presentation will be graded by the teachers and students together. The grade will be indicative and will not contribute to the final grade.

The final project presentation (December 15, preliminary) will be 12 mins. In this presentation, a student briefly describes the problem and chosen approach, focusing on the achieved results and demonstration. This presentation will be graded by the teachers only. This grade will contribute 10% to the final grade.

2.3 Grading

	Not acceptable ($< 60\%$)	Acceptable (60%–70%)	Good (75%–85%)	Excellent ($\geq 90\%$)
Time constraints	• More than one week late	• One week late	• On time	• On time

Data acquisition (20%, 3 weeks, October 20)	• No data acquisition is implemented, only an existing dataset	• Basic crawling is implemented • Small amount of data (100–1k documents)	• Crawling with politness • Considerable amount of data (10k–100k documents) • One data source	• Crawling with politeness and one of the following: distributed crawling or updating repository • Large amount of data (more than 100k documents) • More than one data source
Data processing and storage (20%, 2 weeks, November 3)	• Data is stored in plain files	• One data storage method is implemented	• Data is pre-processed (cleaning, stop-word removal, stemming)	• Different data is stored using different methods (e.g., indexing for unstructured data and database for structured)
Search and features (20%, 3 weeks, November 24)	• Only Boolean search, no ranking • No features	• Search with ranking • One feature	• Two features	• More than two features

Interface (10%, 1 week, December 1)	• No interface	• Basic interface with a query box and a list of results	• Results are presented with snippets • The features are the part of the interface (the corresponding functionality and results are present) • The interface is intuitive and reflects the purpose of the system	• Query terms are highlighted • Additional visualizations
Evaluation (20%, 2 week, December 15)	• No evaluation	• One type of evaluation: qualitative (user study) or quantitative (offline evaluation) • Two assessors	• Both types of evaluation • Predefined evaluation tasks • Four assessors	• Quantitative (offline) evaluation of features • More than four assessors • Multiple assessors per task/relevance judgement

Presentation and report (10%)	• Some aspects of the project are not covered • No demo during the presentation • The report is in the wrong format	• All aspects of the project are covered, but with little details • The presentation includes demo • The report is in the right format, at least 4 pages • The report does not contain reflection and does not discuss future work	• All aspects of the project are properly covered (see requirements for the presentation and report) • The report covers all necessary details, clear, well-written, not more than 8 pages (excluding appendices and screenshots) • The report contains reflection and discusses future work	• The presentation is of high quality (clear, structured, good slides, well presented) • The report can be turned into a short paper for a conference in information retrieval and web search