IRWA Final Project (2025)

Summary

Based on what you learned from theoretical classes, the seminars, the lab exercises, and your own research, you are asked to build a RAG (Retrieval Augmented Generation) system based on a custom search engine implementing different indexing and ranking algorithms.

You are asked to implement four incremental steps that must be delivered on predefined dates:

Part	Topic	Delivery Date
Part 1	Text Processing and Exploratory Data Analysis	24/10/2025
Part 2	Indexing and Evaluation	31/10/2025
Part 3	Ranking and Filtering	20/11/2025
Part 4	RAG, User Interface, and Web Analytics	29/11/2025

The requirements definitions for each part will be published in the Aula Global in the corresponding week.

The programming language must be Python3.

Create a GitHub repository to upload and share your code.

Group

The project must be completed in groups of 3 people (deadline for group creation is 15/10/2025, end of the day).

Project Deliverables

You must deliver for each of the parts the following artifacts:

1. A **PDF report,** uploaded to Aula Global in a section that will be available for that purpose. Name the report file like **IRWA-2025-uXXXXXX-uXXXXXX-uXXXXXX-part-N.pdf** where uXXXXXX is each student's ID. The report must include: an explanation of the decisions you made for implementing the different algorithms, the assumptions made, and everything else you consider relevant to explain to the evaluators.
2. The **repository TAG** that names the part that you are delivering, i.e., **IRWA-2025-part-N**. The TAG creation date must be before the delivery deadline in order to be considered. You can find here a guide on how to work with Git. Section 2.6 addresses tagging.

The deadline time is at 23:00 on the requested date.
Delivery late: -20% of the grade per day.

The delivered code must be properly documented (especially the defined functions) and equipped with a proper README file in the project's root folder. The README should include step-by-step instructions about how to run the code and how to select the different functions, algorithms, and/or other options to run the ranking scores.

Remember to mention the GitHub URL and TAG in the report for each part.

Project Template

The project template can be found here: https://github.com/trokhymovych/irwa-search-engine

Project Evaluation

The team will be evaluated based on the code, corresponding results, the quality of the README.md explaining how to run the code, and the comprehensibility of the reports.
Your code will be evaluated mainly on the implemented algorithms and the level of reproducibility.

Each part is worth 1 point (4 total).

AI Usage Policy:

If AI tools are used, you **must** acknowledge this in your submission (e.g., "I used ChatGPT to review my draft explanation of the pre-processing decisions, and debug the stemming function"). You remain responsible for verifying the accuracy, originality, and appropriateness of all submitted work. Misuse of AI tools may be considered plagiarism or academic dishonesty.

Permitted Uses:

1. You may use AI to review drafts or check grammar and formatting.
2. You may use AI to brainstorm approaches to tasks (e.g., data pre-processing strategies) but not to generate final answers verbatim.
3. AI tools may assist in coding or debugging, but you must understand and be able to explain any code you submit.

Prohibited Uses

1. Copy-pasting AI-generated answers as your own without critical review or modification.
2. Using AI to bypass the thinking process required in reflective or analytical tasks.