CS 516: Information Retrieval and Text Mining

Information Technology University (ITU)
Fall 2025
Course Instructor: Dr. Ahmad Mustafa

Homework Assignment 3

Due Date: 11.59 pm, 30th November, 2025

Design Problem

In this assignment, you will design and implement a complete information retrieval (IR) system that runs locally on your machine. You may use any retrieval strategy (or combination of strategies), such as Boolean retrieval, Vector Space Models (e.g., TF–IDF) etc.

Your goal is to design a coherent, justifiable, and well-evaluated retrieval system using the provided dataset of text documents (accessible here).

Creativity is encouraged, but your system must be reproducible and fully local (no cloud-hosted vector databases).

Requirements

You are free to design your IR system however you like, as long as it satisfies the following:

1. Local Implementation

Your system must run end-to-end on a local machine (Windows, Mac, or Linux).
You may use any mainstream programming language or libraries.

Cloud-hosted vector databases (e.g., Pinecone, Chroma Cloud, Weaviate Cloud, Elasticsearch clusters) are **not** allowed.

Local libraries (e.g., scikit-learn, gensim, rank-bm25, FAISS local install) are allowed.

2. Reproducible Pipeline

Your submission must include:

- source code
- a `README` with instructions to run your system

This ensures instructors can test your pipeline.

3. Technical Report

Write a technical report documenting your retrieval system. Your report should follow the template provided below.

Plagiarism & AI Use Policy

Your work on this assignment must be **your own**. You may discuss ideas with classmates, but **all code and writing must be written by you**.

You may use external resources or AI tools (e.g., ChatGPT, Copilot), **but you must clearly disclose every instance of AI use** with screenshots of prompts and responses. Failing to disclose AI assistance counts as plagiarism.

The following are not allowed:

- Copying code or text from other students
- Sharing your code with anyone
- Using online repositories or AI tools to generate solutions without disclosure

Violations may result in a **zero on the assignment** and potential academic misconduct action.

Technical Report Format

1. System Architecture

1.1 System Diagram

(Insert a block diagram showing all modules—data ingestion, preprocessing, indexing, retrieval, reranking, evaluation, etc.)

1.2 Figure Caption

Provide a short caption (2–3 sentences) describing the high-level architecture and the flow from raw documents to ranked results.

2. Description of the Retrieval System

Provide a detailed but clear description of your system design. This includes your data preprocessing steps (e.g., normalization, capitalization-handling, tokenization), indexing techniques (boolean, TF-IDF, BM25 etc.), scoring and ranking criteria. Justification should be provided where appropriate for any modifications added to the “standard”, bare-bones retrieval pipeline.

3. Evaluation

Provide a detailed description of how you evaluated your retrieval system. This may include both qualitative and quantitative approaches, as well as an appraisal of how efficient the retrieval system is in terms of its memory footprint, querying speed etc.

4. Discussion

Discuss the major findings from your results, any shortcomings you noticed, how you plan to improve the system etc.

5. References

Cite any research papers, textbooks, public code repositories, blogs or tutorials used to help you with this assignment. Use any consistent citation format.

6. Disclosure of AI Use

6.1 Summary of AI Usage

Document all AI tools used:

- ChatGPT
- GitHub Copilot
- Claude
- Others

6.2 Evidence of AI Assistance

For each instance of AI-generated content:

- Insert screenshots of prompts and responses
- Indicate where the AI output appears in your code or report
- Briefly justify any modifications you made to the AI output

Format example:

Figure A1 shows the prompt used to generate the initial BM25 scoring function. The final implementation was modified for efficiency (lines 30–42 of my code).

Submission Instructions

You must submit:

1. **A single PDF report** using this template.

2. **A clean, well-commented GitHub repository** containing:
    - All code
    - A README with reproducible instructions
    - Any configuration or environment files
3. **Screenshots documenting AI use**, included in the PDF.