CSE 4392/5369 Special Topic: Information Retrieval

Homework 11 - Spring 2026

Due Date: April 20, 2026, 11:59 PM

Question 1 - 100%

In this assignment, you will build a mini information retrieval system over a document collection. The goal is to understand the basic IR pipeline, including inverted index construction, Boolean retrieval, and ranked retrieval with tf-idf.

- Dataset: Twenty Newsgroups Dataset

1. Implement the preprocessing and indexing pipeline. Your system should read all documents, tokenize the text, convert tokens to lowercase, remove punctuation, and build an inverted index. Each term in the vocabulary should map to a sorted postings list of document IDs in which the term appears.
2. Implement two retrieval modes on the same document collection:
   - Boolean retrieval for two-term AND queries (e.g., `information AND retrieval`);
   - Ranked retrieval for free-text queries using tf-idf scoring.

   For ranked retrieval, compute a relevance score for each document based on the sum of tf-idf weights of the query terms appearing in the document, and return the top 5 documents ranked by score in descending order.
3. Write a short report that includes:
   - a description of your implementation, along with a brief discussion of your findings and an explanation of the observed results;
   - one example of a Boolean query and its returned results;
   - one example of a ranked query and its top returned documents;
   - brief answers to the following questions:
     (a) What is the difference between Boolean retrieval and ranked retrieval?
     (b) In what kind of search scenario is tf-idf ranking more useful than Boolean retrieval?

Submission Format: Submit one zip file via Canvas containing only the `.pdf` version of your homework (typed submissions are preferred; scanned images must be readable), the corresponding source files, and a `README` file describing how to run the code. The zip file must be named `lastname_studentID_hw11.zip`.