CS 429 Information Retrieval

Assignment 3: Relevance Feedback

Points: 100

Submission deadline: Friday, 2 August, 11:59 PM

Late Submission: Monday, 5 August, 11:59 PM with 10% penalty

Note: This is individual work

In this assignment, you will run an experiment to study the effects of relevance feedback on the recall, precision and Mean Average Precision (MAP) values of an IR system. The IR system will use a vector space model with cosine similarity (tf-idf weighting). You will run the study on the TIME dataset, provided along with the assignment.

Part A: Cosine Similarity and Rocchio's algorithm [40pts]

- You will implement a cosine similarity measure with *tf-idf* weighting. Your index should contain the information that you will need to calculate the cosine similarity measure such as *tf* and *idf* values. You may reuse code from the previous assignments as needed
- Implement the Rocchio algorithm for query refinement. Your system should display results and then prompt the user for providing positive and negative feedback. **Use α= 1, β= 0.75, and γ= 0.15** as parameters for the Rocchio's algorithm.

Part B: Experimental study [35pts]

- Run your system for at least 3 queries **from the test bed**. Pick queries that have 5 or more relevant documents (see TIME.REL file). For each query, you will perform a series of 5 relevance feedback and plot the change in precision, recall and MAP
- You will prepare a report on the experimental study where you will provide at least the following details for each of the queries:
  - Query text and ID (provided in the testbed)
  - Precision, recall and MAP values of the query
  - IDs of documents which are Positive and Negative feedback provided for each query *during each iteration* of the Rocchio algorithm
  - For each *iteration* of the Rocchio algorithm, provide the terms of the new query and their weights
- Your report will have 3 plots (precision vs Rocchio iteration, recall vs Rocchio iteration, and MAP measure vs Rocchio iteration) that depict the progressive change in the performance values over the iterations of the Rocchio algorithm, for the three queries.
- Also discuss any **query drift** that you may observe in your results.
- Note: The queries provided in the testbed have varying number of relevant documents (see TIME.REL file). This can be a problem when calculating the performance values, if *k* is kept constant during the retrieval. For the experimental study, assume that the number of relevant

documents is provided to the system along with the query. In other words*, the value of k will change with the query*.

Part C: Pseudo Relevance Feedback [25pts]

One of the challenges of user feedback is that the user may not be willing to provide feedback. In such cases, pseudo relevance feedback can be used. You will compare the performance of your user feedback-based system from **Part A** against a pseudo feedback mechanism, where the top 3 results of the system are considered to be relevant. Run this system with the **same queries** from your experimental study in **Part B**. Compare its recall, precision and MAP values against the system using user feedback and add to the report from **Part B**.

Other instructions:

- Comment your code appropriately
- You may reuse the code from earlier assignments

Attachments:

1. Collection of documents **(time.zip)**
2. Skeleton code **(index.py)** – implement the functions in the code
    a Use additional functions as needed
3. Sample output (**sample_output.txt**)

Submission:

Submit the following files on Blackboard as a .zip file.

1. index.py
2. Report.pdf: with performance analysis and other discussions.
3. Output.txt: containing the output generated by your code for the three queries. This is for testing your code.