Comp 479/6791 # Project 2 **Fall 2023**

Objectives: Implement a naive indexer. Implement single term query processing. Implement and compare lossy dictionary compression.

Due date: October 9, 2023 Challenge data released October 8 Note submission deadline is 5pm

Data: Use Reuters21578. Use the NEWID values from the Reuters corpus as DocIDs. Note that there are thousands of documents in Reuters21578

Description:

Subproject I: Naive indexer implementation (3pts, Attr 4+5)

1. develop a module that while there are still more documents to be processed, accepts a document as a list/stream of tokens and outputs term-documentID pairs to a list F.
2. when there is no more input, sort F and remove duplicates
3. turn the sorted file F into an index by turning the docIDs paired with the same term into a postings list and linking it to the term

Note: you can do this in memory. The goal here is to experiment with the content, not optimize.

Subproject II: Single keyword query implementation (1pt, Attr5)

1. implement a query processor for single term queries
2. validate query returns for three sample queries (you have to decide on your sample queries)

Subproject III: Dictionary compression table (3pts, Attr5)

1. implement the lossy dictionary compression techniques for the first two columns of Table 5.1 in the textbook and compile a similar table for Reuters-21578. (Remember that your corpus is much smaller than the Reuters corpus used for Table 5.1.) Are the changes similar? Discuss your findings.
2. compare retrieval results for your three sample queries of Subproject II when you run them on your compressed index. Discuss your findings in the Project Report

Challenge queries (1pt, Attr4) 24hrs before the Project deadline, Challenge Queries will be posted. You are to run the challenge queries on your Naive Indexer and submit the output of the runs

Report (1pt, Attr6) The report has to reflect all aspects of your Project in academic writing. Do not include log files. The markers should be ready to adequately mark your Project after reading the Report

Demo file (1pt, Attr6) Compile a file that steps through a demo of your system to help the marker understand your code and illustrate the Project Report

Deliverables:

1. Code for Subprojects I-II, well documented (4pts)
2. For Subproject III, include Dictionary compression table and comparison to Table 5.1 in the Report (3pts)
3. Challenge query runs (1pt)
4. any additional testing or aborted design ideas that show off particular aspects of your project
5. Project Report (1pt)
6. Demo file (1pt)