Comp 479/6791 **Project 2** **Fall 2022**

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

Due date: October 8, 2022

Data: Use Reuters21578. For docID, use the NEWID values from the Reuters corpus to make your retrieval comparable

Description:

Subproject I: naive indexer

1. develop a module that while there are still more documents to be processed, accepts a document as a list 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 setting the pointer

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

Subproject II: single term query processing

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: implement lossy dictionary compression, 'recreate' Table 5.1

1. implement the lossy dictionary compression techniques 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 your report

Deliverables:

1. individual project
2. well documented code
3. sample runs of the queries I will post two days before the submission deadline (October 6th). Run queries on both indices
4. any additional testing or aborted design ideas that show off particular aspects of your project
5. a project report that summarizes your approach, illustrates your designs, presents your table of savings for lossy dictionary compression and discusses, what you have learned from the project

Marks:

Naive indexer implementation	3pts	Attr5, Attr4
Single keyword query implementation	1pt	Attr5
Challenge single keyword query results	1pt	Attr4
Dictionary compression table	3pts	Attr5
Report	1pt	Attr6
Demo	1pt	Attr6