Comp 479/6791 # Project 2 **Fall 2020**

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

Due date: October 21, 2020

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

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. Are the changes similar? Discuss your findings. (Note that stemming is not required here, if you run out of time before you get the Porter stemmer to work, that is ok for this assignment, the remaining table is fine.)
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 posted two days before the deadline. 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	2pts	Attr5
Resulting inverted index	1pt	Attr4
Single keyword query implementation	1pt	Attr5
Challenge single keyword query results	1pt	Attr4
Dictionary compression implementation	1pts	Attr5
Dictionary compression table	1pts	Attr5
Report	1pt	Attr6