Indexing and Index Compression

Diksha Sharma (dxs134530)  
CS 6322.001 – Information Retrieval – fall 2015  
Assignment 2 - Report

### Approach Followed

In this assignment I have used TreeMaps and replaced my old token using ArrayList with a TreeMap.

**Class description is as below:**

1. Search – This contains the main method and in turn calls other classes for execution.
2. Tokenizer – This creates the token index – (Part of assignment 1) – Changed array lists to hash maps. Stop words are stored.
3. Lemmatize – This class finds lemma for tokens in the token index and creates an index of lemmas. No Stop Words are stored. We find the lemma using the Standford Core NLP lemmatizer. Once the index is created – this class makes call for its compression in the Compress class. This class also creates the stem index and then calls compression method in Compress class for stem index compression.
4. Compress – This class does all the compression of both the indexes for tokens and stems. I use treemap for storing the indexes. The uncompressed indexes use Term as structure for storing the posting files. The posting file structure is a TreeMap. The compressed indexes use CompressToken and CompressStem for storing the posting files. The posting files are stored in array lists for terms of block 8.
5. PostingFile – This stores the document id and the frequency of the term/stem in that document.
6. CompressStem – Data Structure for compressed stem index used as a TreeMap.
7. CompressToken – Data Structure for compressed token index used as a treemap.
8. Porter – Porter stemmer – modified to create index with document id and frequency information for each term.
9. Document – Data Structure used for storing document information.
10. Term - Data structure used for uncompressed indexes.

### Answers to questions in assignment

* The elapsed time ("wall-clock time") required to build any version of your index
* 27 seconds
* The size of the index Version 1 uncompressed (in bytes)
* 170467 Bytes
* The size of the index Version 2 uncompressed (in bytes)
* 100807 Bytes
* The size of the index Version 1 compressed (in bytes)
* 78392 Bytes
* The size of the index Version 2 compressed (in bytes)
* 43238 Bytes
* The number of inverted lists in each version of the index, and
* Version1 Uncompressed: 7250 Bytes
* Version1 Compressed: 882 Bytes
* Version2 Uncompressed: 4504 Bytes
* Version2 Compressed: 544 Bytes
* The df, tf, and inverted list length (in bytes) for the terms:
* "Reynolds", "NASA", "Prandtl", "flow", "pressure", "boundary", "shock" (or stems that correspond to them).
* These details are provided in the file: frequency.txt in out folder of the zip file.