Report of Index Building

Net Id: kxk152430
Homework - 2

Lemmatization software/code/library you used:

stanford-corenlp-3.3.1-models.jar
stanford-corenlp-3.3.1-models.jar

URL to download :

https://repository.cloudera.com/artifactory/repo/edu/stanford/nlp/stanford-corenlp/3.3.1/

Program Description :

Size of original collection :: 1700 KB
Size of V1 uncompressed :: 1566 KB
Size of V1 compressed :: 898 KB
Size of V2 uncompressed :: 1334 KB
Size of V2 compressed :: 859 KB

The compressed sizes are lesser than the uncompressed size which is in turn lesser than the original collection size.

The program has below classed are

Tokenize.java : Performs tokenizing task, which removes unwanted characters, case folding, removing numbers, etc. This also does removal of stop words.
The stop words list taken from
/people/cs/s/sanda/cs6322/resourcesIR

Lemmatize.java: Performs lemmatization using the above mentioned standford nlp library, the dictionary and the posting lists are created based on that. The data structure doc frequency, posting list pointer and term pointer and the long dictionary string is all present and build inside this by calling the compression techniques functions.

Stemmer.java: Performs stemming using the porter stemmer algorithm from
http://chianti.ucsd.edu/svn/csplugins/trunk/soc/layla/WordCloudPlugin/trunk/WordCloud/src/cytoscape/csplugins/wordcloud/Stemmer.java
the dictionary and the posting lists are created based on that. The data structure doc frequency, posting list pointer and term pointer and the long dictionary string is all present and build inside this by calling the compression techniques functions.

Compression.java : This contains code for all the compression techniques, gamma compression and delta compression of posting list and other numbers.
Blocking compression with K = 8 and Front coding with blocking for dictionary terms.

DictionaryClass.java : This is a POJO class that renders the data structure for dictionary and pointer to posting list.

DocDetails.java : This is a POJO class that renders the data structure for posting list.

Timer.java : This class has the start and end time functions, to calculate the time taken for indexing and compressing.

Output files generated are :

Index_Version1.uncompressed – Dictionary and inverted list of Lemmas

Index_Version1.compressed – Dictionary and inverted list of Lemmas after gamma code compression and blocking compression

Index_Version2.uncompressed – Dictionary and inverted list of Stems

Index_Version2.compressed – Dictionary and inverted list of Lemmas after delta code compression and front coding compression