Inverted Indexer is the project folder.

Uses jsoup, jcommon, jfreechart library which has been included in the build path and is provided in the "lib" folder

No command line arguments to be passed.

Jsoup has been already added to build path, just in case fails can follow the below steps:

In Eclipse,

(right click on project) -> Build Path -> Configure Build Path and select the jsoup library from lib folder (add external jars) and the JavaSE-1.7

Corpus contains the text file of each URL crawled.

Storage contains all the saved data structures.

1. Unigram Inverted Indexes: InvertedIndexes/Unigram/unigramindexes.txt"
2. Bigram Inverted Indexes: InvertedIndexes/Bigram/bigramindexes.txt"
3. Trigram Inverted Indexes : InvertedIndexes/Trigram/trigramindexes.txt"
4. Tokens count per Document: Storage/tokenCountPerDoc.txt"
5. Term frequency for unigrams : Storage/Term Frequency(Task 3-1)/tfUnigram.txt"
6. Term frequency for bigrams : Storage/Term Frequency(Task 3-1)/tfBigram.txt"
7. Term frequency for trigrams : Storage/Term Frequency(Task 3-1)/tfTrigram.txt"
8. Document frequency for unigrams : Storage/Doc Frequency(Task3-2)/dfUnigram.txt"
9. Document frequency for bigrams : Storage/Doc Frequency(Task3-2)/dfBigram.txt"
10. Document frequency for trigrams : Storage/Doc Frequency(Task3-2)/dfTrigram.txt"
11. Stoplist : Storage/stoplist.txt";

“Computed/ “ will have all the newly computed values

“InvertedIndexes/” contains the build inverted indexes for unigrams, bigrams and trigrams.

The 3 tables of termfrequency are at: Storage/Term Frequency(Task 3-1)/

The 3 tables of document frequency are at: Storage/Doc Frequency(Task3-2)/

I have made three data structures of type

LinkedHashMap<String, ArrayList<DocumentEntry>>

**class** DocumentEntry {

**public** **int** docid;

**public** **int** tf;}

Where string is the token, and DocumentEntry contains the docid and tf.

ArrayList<TermEntry> tfUnigram,

ArrayList<TermEntry> tfBigram,

ArrayList<TermEntry> tfTrigram

Contains the term frequencies mapped with the token.

TokenCount maps the document with the total tokens in the respective document.

“**Graphs will be drawn when term frequency will be created. (Option 2 in the menu)”.**

Zipfian curves are in the InvertedIndex folder itself.