Information Retrieval

Home Work # 2

“Inverted Index Builder”

By,

Rohit Hedge

2021134344

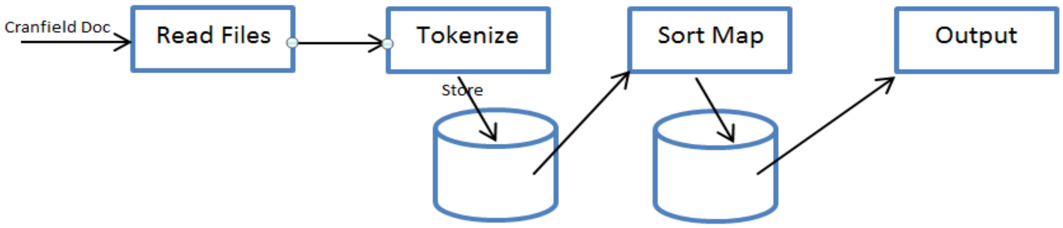
Project Description:

This project has 7 Classes,

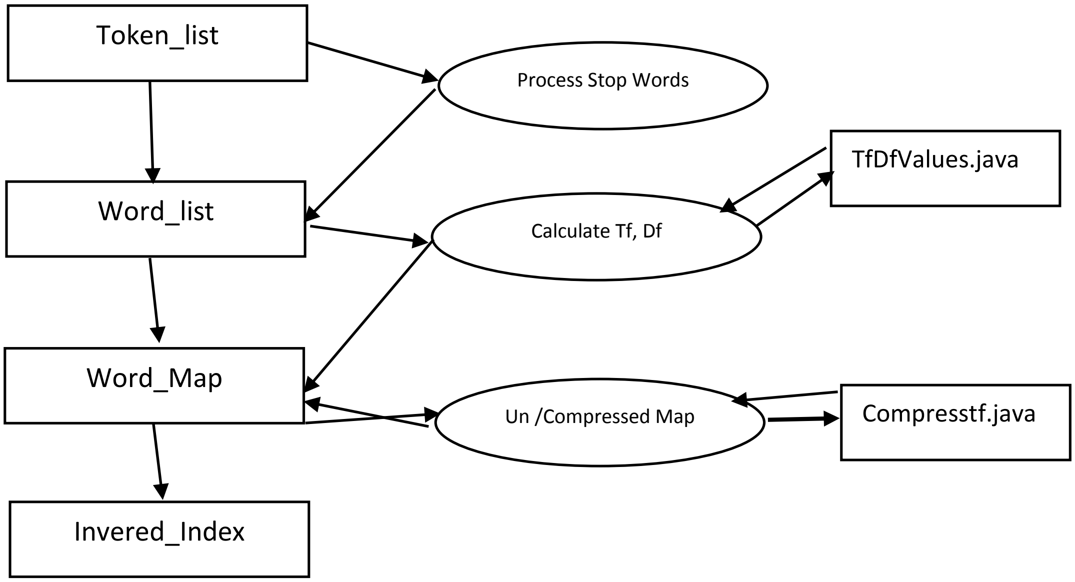
1. IndexedRetrieval.java
2. Stemmer.java
3. TfDfValues.java
4. ValueComparator.java
5. ByteComparator.java
6. Encoding.java
7. CompresstfDfValue.java

Algorithm:

Home Work #1



Output of Homework gave me the Tokenized word list, which I used as the input for my project 2.



Working:

- IndexedRetrieval is the main class that actually builds the inverted index.
- Results are written into results.txt file.
- Directory path is read from input.txt file.
- Stop words file is pointing to the CS/IR_RESOURSE.
- Time taken by the program is displayed in seconds.

References:

- Byte Comparator and byte array compressing was referred from google and stack overflow
  URI: http://stackoverflow.com/questions/5108091/java-comparator-for-byte-array-lexicographic

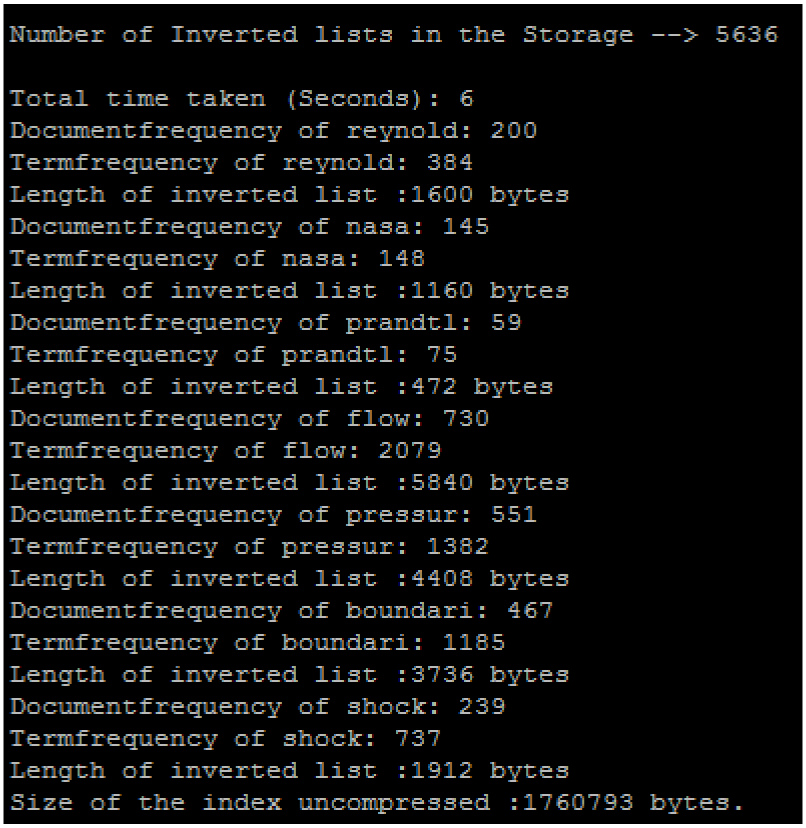
```
public class ByteComparator implements Comparator<byte> {
  public int compare(byte b1, byte b2) {
    return new Byte(b1).compareTo(b2);
  }
}
```

Running the Code:

- Compile the Code as:
  javac *.java

- Run the code as:
  java IndexedRetrieval

Results:



```
Number of Inverted lists in the Storage --> 5636

Total time taken (Seconds): 6
Documentfrequency of reynold: 200
Termfrequency of reynold: 384
Length of inverted list :1600 bytes
Documentfrequency of nasa: 145
Termfrequency of nasa: 148
Length of inverted list :1160 bytes
Documentfrequency of prandtl: 59
Termfrequency of prandtl: 75
Length of inverted list :472 bytes
Documentfrequency of flow: 730
Termfrequency of flow: 2079
Length of inverted list :5840 bytes
Documentfrequency of pressur: 551
Termfrequency of pressur: 1382
Length of inverted list :4408 bytes
Documentfrequency of boundari: 467
Termfrequency of boundari: 1185
Length of inverted list :3736 bytes
Documentfrequency of shock: 239
Termfrequency of shock: 737
Length of inverted list :1912 bytes
Size of the index uncompressed :1760793 bytes.
```