**CS19MDS110002 – Scalable Algorithms – Assignment 4**

**Execution**

To execute the program please run the following command inside folder **executables**

**java -jar hbf-1.0.jar Executor corpus\_folder query\_folder 8192 5 0.9**



**Corpus folder – corpus\_folder with 1024**

**Query folder – query\_folder with 4 test queries**

**Output**

**A screenshot of text

Description automatically generated**

**Code**

Code/src/main/java/com.hbf

**MainFiles:**

BloomFilter.java

HierarchicalBF.java

**Reports with Different Test Scenarios**

**All the test cases are run based on the description and the following assumptions mentioned in the assignment**

* BloomFilter for each document is a bag words without any preprocessing (only case insensitive words).
* Number of documents are 2^N
* t threshold value varies between 0<=t<=1
* m selection is loosely based on the total unique words.

**Documents – 1048, total unique words - 40536**

**Parameters: m – 8192, k – 5, t - 0.9**

**A screenshot of text

Description automatically generated**



**The actual document for 62479 has some text different from query text still matches correctly, however in the below screen shot exact match is expected hence it returns empty.**

**Parameters: m – 8192, k – 5, t - 1.0**

A screenshot of a cell phone

Description automatically generated



**Parameters: m – 8192, k – 5, t – 0.6**

A screenshot of a cell phone

Description automatically generated



* With smaller “t” restriction document matched - 59850 to the query document 124146 which has somewhat similar words, other documents still match to their original similar documents.
* **Also, node\_visited increases** compared to the previous execution

**Parameters: m – 4096, k – 5, t – 0.9**

Decreasing the m size

A screenshot of a cell phone

Description automatically generated



With decreased “m” size the number of visited nodes increased.

**Documents – 16384, total unique words – 382814**

For testing extended 8000 newsgroup datasets with random document repetition to 16834

**Initial Test scenario**

124146 - same document is not present in indexed data

74721 – document is present in the index, but words are slightly different

9140 - document is present but one complete paragraph is missing

**Parameters: m – 32768, k – 5, t – 1.0**

A screenshot of a social media post

Description automatically generated



**Parameters: m – 32768, k – 5, t – 0.9**

**A screenshot of a social media post

Description automatically generated**



**Parameters: m – 32768, k – 5, t – 0.5**

**A screenshot of a social media post

Description automatically generated**



124146- matched to 59256 the intersection of bag of words

9140 – matched to 9140

**Parameters: m – 32768, k – 20, t – 0.9**

A screenshot of a social media post

Description automatically generated



Random documents are matched, and number of nodes visited increases.