**Assignment 2**

**Name:** Pawan Araballi

**Email:** [paraball@uncc.edu](mailto:paraball@uncc.edu)

The assignment consists of 5 documents DocWordCount.java, TermFrequency.java, TFIDF.java, search.java and rank.java

**Note:** Every time when running a new program please make sure all the folder structure has been deleted.

**DocWordCount.java:** This program basically is an extension of the WordCount where we append a delimiter and the filename from where the word is being fetched.

Execution Instructions:

* Create the folder structure required Hadoop fs -mkdir /user/cloudera/docwordcount /user/cloudera/docwordcount/input
* Insert the data into the Hadoop filesystem Hadoop fs –put file\* /user/cloudera/docwordcount/input
* Compile the file javac -cp /usr/lib/hadoop/\*:/usr/lib/hadoop-mapreduce/\* DocWordCount.java -d build -Xlint
* Create the jar file jar -cvf docwordcount.jar -C build/ .
* Run the hadoop hadoop jar docwordcount.jar DocWordCount /user/cloudera/doccwordcount/input /user/cloudera/docwordcount/output
* The output file would be viewed in Hadoop fs -cat /user/cloudera/docwordcount/output/\*

**TermFrequency.java:** This program is an extension of the DocWordCount where we calculate the termfrequency of each word.

* Create the folder structure required Hadoop fs -mkdir /user/cloudera/termfrequency /user/cloudera/ termfrequency /input
* Insert the data into the Hadoop filesystem Hadoop fs –put file\* /user/cloudera/ termfrequency /input
* Compile the file javac -cp /usr/lib/hadoop/\*:/usr/lib/hadoop-mapreduce/\* TermFrequency.java -d build -Xlint
* Create the jar file jar -cvf termfrequency.jar -C build/.
* Run the hadoop hadoop jar termfrequency.jar TermFrequency /user/cloudera/termfrequency/input /user/cloudera/ termfrequency/output
* The output file would be viewed in Hadoop fs -cat /user/cloudera/ termfrequency/output/\*

**TFIDF.java:** This program is an extension of TermFrequency where we calculate the termfrequency and use the output of that value to calculate the tfidf value. Here I have created a temporary location to store the value of the termfrequency value

* Create the folder structure required Hadoop fs -mkdir /user/cloudera/tfidf /user/cloudera/tfidf /input
* Insert the data into the Hadoop filesystem Hadoop fs –put file\* /user/cloudera/tfidf/input
* Compile the file javac -cp /usr/lib/hadoop/\*:/usr/lib/hadoop-mapreduce/\* TFIDF.java -d build -Xlint
* Create the jar file jar -cvf tfidf.jar -C build/.
* Run the hadoop hadoop jar tfidf.jar TFIDF /user/cloudera/tfidf/input /user/cloudera/tfidf/output
* The output file would be viewed in Hadoop fs -cat /user/cloudera/tfidf/output/\*

**Search.java:** This program is extension of TFIDF where we search for words in the file and output the tfidf value for a particular file containing the word. In this file again I am creating the intermediate folders for storing the termfrequency and tfidf.

* Create the folder structure required Hadoop fs -mkdir /user/cloudera/search /user/cloudera/search/input
* Insert the data into the Hadoop filesystem Hadoop fs –put file\* /user/cloudera/search/input
* Compile the file javac -cp /usr/lib/hadoop/\*:/usr/lib/hadoop-mapreduce/\* search.java -d build -Xlint
* Create the jar file jar -cvf search.jar -C build/.
* Run the hadoop hadoop jar search.jar search /user/cloudera/search/input /user/cloudera/search/output
* The output file would be viewed in Hadoop fs -cat /user/cloudera/search/output/\*

**Rank.java:** This program is extension of Search where we sort the value to get the best fit the searched word. In this file again I am creating the intermediate folders for storing the termfrequency, tfidf and search.

* Create the folder structure required Hadoop fs -mkdir /user/cloudera/rank /user/cloudera/rank/input
* Insert the data into the Hadoop filesystem Hadoop fs –put file\* /user/rank/search/input
* Compile the file javac -cp /usr/lib/hadoop/\*:/usr/lib/hadoop-mapreduce/\* rank.java -d build -Xlint
* Create the jar file jar -cvf rank.jar -C build/.
* Run the hadoop hadoop jar rank.jar search /user/cloudera/rank/input /user/cloudera/rank/output
* The output file would be viewed in Hadoop fs -cat /user/cloudera/rank/output/\*