**Readme Instructions**

1. Open Eclipse in Cloudera Quick start VM (Quick start VM can be downloaded from Cloudera website).
2. Create New Java Project by clicking on File->New- Java Project. Enter required information.
3. Once the project is created, create new package by right clicking on the project and click New->Package. Enter a valid name for the project (for example: com. cloud. [your name]).
4. Now add the required Hadoop libraries by right click on the Project then select Build Path-Configure Build path. After that you can add the libraries by clicking on add external libraries.
5. Libraries required in the 4th step can be found in the following paths:

* /usr/lib/hadoop/client
* /home/cloudera/lib

1. Now copy all the files present in the source code folder to the package we created in eclipse.
2. Make sure that you are not having any compile time errors especially in the package statement of each file.
3. If you are not having any error right click on the project and click on export and a popup window comes.
4. Select the type as Jar file and click on Next and give destination path for the jar. Follow the instructions and you will be able to generate the jar without any errors.
5. Once the jar is successfully generated you can perform all the operations.(**I have given the name of the jar as wordcount.jar**)

**Copying the Input files into HDFS from your local system**

1. First identify the location of the input files in your system.(Download Canterbury.zip)
2. Open the terminal and execute the following command

* sudo su hdfs
* hadoop fs -mkdir /user/cloudera
* hadoop fs -chown cloudera /user/cloudera
* exit
* sudo su cloudera
* **hadoop fs -mkdir /user/cloudera/Assignment /user/cloudera/Assignment/input**

**This is the location where your input files are going to be stored.**

* Once the input location in HDFS is created, copy the input files from local file system to HDFS by using the following command

hadoop fs -put /home/cloudera/canterbury/\* /user/cloudera/wordcount/input

**Running DocWordCount Program**

Go the location where you have exported the jar and run the following commands

* hadoop jar wordcount.jar com.cloud.febin.DocWordCount /user/cloudera/Assignment/input /user/cloudera/Assignment/output/docwordcount
* Two arguments are passed one for the input location of the files and other for the output location where we want to store the results
* Once the execution is completed you can see the output by using the following command : **hadoop fs -cat /user/cloudera/Assignment/output/docwordcount/\***
* If we want to run the same program again, please delete the output folder by using the following command :

**hadoop fs -rm -r /user/cloudera/Assignment/output/docwordcount**

* For copying the output to your local file system, you can use the following command

hadoop fs -get /user/cloudera/Assignment/output/docwordcount/ /home/cloudera/outputs/

**Running TermFrequency Program**

Go the location where you have exported the jar and run the following commands

* hadoop jar wordcount.jar com.cloud.febin.TermFrequency /user/cloudera/Assignment/input /user/cloudera/Assignment/output/termfrequency
* Two arguments are passed one for the input location of the files and other for the output location where we want to store the results
* Once the execution is completed you can see the output by using the following command : **hadoop fs -cat /user/cloudera/Assignment/output/termfrequency/\***
* If we want to run the same program again, please delete the output folder by using the following command :

**hadoop fs -rm -r /user/cloudera/Assignment/output/termfrequency**

* For copying the output to your local file system, you can use the following command

hadoop fs -get /user/cloudera/Assignment/output/termfrequency/ /home/cloudera/outputs/

**Running TFIDF Program**

Go the location where you have exported the jar and run the following commands

* hadoop jar wordcount.jar com.cloud.febin.TFIDF /user/cloudera/Assignment/input /user/cloudera/Assignment/output/tfidf
* Two arguments are passed one for the input location of the files and other for the output location where we want to store the results
* **TermFrequency task is called via chaining in TFIDF program and temporary output location for TermFrequency is give inside the program itself.(Check TFIDF.java:37) .if you want another location. Please feel free to make the change and create the jar again.**
* Current temporary location is **/user/cloudera/Assignment/output/temp**
* Once the execution is completed you can see the output by using the following command : **hadoop fs -cat /user/cloudera/Assignment/output/tfidf/\***
* If we want to run the same program again, please delete the output folder by using the following command :

**hadoop fs -rm -r /user/cloudera/Assignment/output/tfidf**

* For copying the output to your local file system, you can use the following command

hadoop fs -get /user/cloudera/Assignment/output/tfidf/ /home/cloudera/outputs/

**Running Search Program**

Go the location where you have exported the jar and run the following commands

* **hadoop jar wordcount.jar com.cloud.febin.Search /user/cloudera/Assignment/output/tfidf /user/cloudera/Assignment/output/search computer science**
* Input location is the location is the output location of the TFIDF program
* Once the execution is completed you can see the output by using the following command : **hadoop fs -cat /user/cloudera/Assignment/output/search/\***
* If we want to run the same program again, please delete the output folder by using the following command :

**hadoop fs -rm -r /user/cloudera/Assignment/output/search**

* For copying the output to your local file system, you can use the following command

hadoop fs -get /user/cloudera/Assignment/output/search/ /home/cloudera/outputs/

**Running Rank Program**

Go the location where you have exported the jar and run the following commands

* **hadoop jar wordcount.jar com.cloud.febin.Rank /user/cloudera/Assignment/output/search /user/cloudera/Assignment/output/rank**
* Input location is the location is the output location of the Search program
* Once the execution is completed you can see the output by using the following command : **hadoop fs -cat /user/cloudera/Assignment/output/rank/\***
* If we want to run the same program again, please delete the output folder by using the following command :

**hadoop fs -rm -r /user/cloudera/Assignment/output/rank**

* For copying the output to your local file system, you can use the following command

hadoop fs -get /user/cloudera/Assignment/output/rank/ /home/cloudera/outputs/