Name: Shrija Chavan CWID: A20381511 ITMD\_521 Assignment\_05

Vagrant Box used: Ubuntu/trusty64 Java version: Java 1.7.

Instruction to run the code:

Pre-requisites:

1. Assuming, that all your jps command are running in the HDFS and after cloning the repo the below steps are to be followed.
2. Create an input directory into the HDFS to keep the dataset by executing the below command.

hadoop fs -mkdir -p /user/$USER/wordcount/input

1. Copy the concated dataset into the input directory.

4. Copy the pattern.txt file in the /user/$USER/wordcount/

**I. To run the WordCount.java to give with words which appear more than 4 times:**

1. Compile the Wordcount.java by executing the below command

hadoop com.sun.tools.javac.Main WordCount.java

1. Create the jar by executing the below command

jar cf wc.jar WordCount\*.class

1. Execute the jar file by using the below command

hadoop jar wc.jar WordCount /user/$USER/wordcount/input /user/$USER/wordcount/output

1. See the results by executing the below command

hadoop fs -cat /user/$USER/wordcount/output/part-r-00000

**II. To run the Modified WordCount2.java with case sensitive= true and all the**

**punctuations and preposition removed by giving the skip command**

**from the pattern.txt file:**

1. Compile the Wordcount2.java by executing the below command

hadoop com.sun.tools.javac.Main WordCount2.java

1. Create the jar by executing the below command

jar cf wc2.jar WordCount2\*.class

1. Delete the output folder by executing the below command

hadoop fs -rm -r /user/$USER/wordcount/output

1. Execute the jar file by using the below command to give the results of the words which the lowercase turned on and all the preposition and punctuations removed:

hadoop jar wc2.jar WordCount2 -Dwordcount.case.sensitive=true /user/$USER/wordcount/input /user/$USER/wordcount/output -skip /user/$USER/wordcount/pattern.txt

1. See the results by executing the below command

hadoop fs -cat /user/$USER/wordcount/output/part-r-00000.