**READ ME- Assignment-3**

The data is downloaded from the Internet using the link given in the assignment.

This file contains data in JSON format. Hence a Json Parser (JsonParser.java) is implemented using java to parse the file and get the value in string format. The java class JsonParser.java also implements a function to remove the stop words from the textual data. The final output of the java file is stored in ParserOutput folder.

**Input File:** JsonSample.txt

**Stop Words File:** StopWords.txt

**Output Folder:** ParserOutput

These output files are given as input to Mahout for clustering. The following commands are used to perform clustering.

// Start Hadoop

1. $cd $HADOOP\_HOME
2. $./bin/start-all.sh

// Make assignment3 folder in Hadoop directory

1. $./bin/hadoop fs –mkdir assignment3

// Copy the ParserOutput folder to assignment3

1. $/bin/hadoop fs -put /Users/himansubadhai/Documents/input/assignment3/ParserOutput assignment3
2. $./bin/hadoop fs –ls
3. $ ./bin/hadoop fs -ls assignment3/ParserOutput

// Create sequence directory in Hadoop using mahout

1. $cd Documents/Softwares/mahout-distribution-0.7/
2. ./bin/mahout seqdirectory -i assignment3/ParserOutput -o assignment3/out-seqdir -c UTF-8 -chunk 5
3. $cd $HADOOP\_HOME
4. $./bin/hadoop fs -ls assignment3

// Create vector in Hadoop using mahout

1. $cd Documents/Softwares/mahout-distribution-0.7/
2. $./bin/mahout seq2sparse -i assignment3/out-seqdir/ -o assignment3/vectors --maxDFPercent 85 –namedVector
3. $./bin/hadoop fs -ls assignment3
4. $./bin/hadoop fs -ls assignment3/vectors

// Create clusters in Hadoop using mahout

1. $cd Documents/Softwares/mahout-distribution-0.7/
2. ./bin/mahout kmeans -i assignment3/vectors/tfidf-vectors -c assignment3/cluster-centroids -o assignment3/kmeans -dm org.apache.mahout.common.distance.CosineDistanceMeasure -x 10 -k 20 -ow –cl

// Create the output file cluster-docs.txt

1. $./bin/mahout seqdumper -i assignment3/kmeans/clusteredPoints/part-m-00000 > /Users/himansubadhai/Documents/input/assignment3/cluster-docs.txt

// Use the given view.pl to view the output in readable format

1. $ perl /Users/himansubadhai/Documents/input/assignment3/view.pl /Users/himansubadhai/Documents/input/assignment3/cluster-docs.txt /Users/himansubadhai/Documents/input/assignment3/cluster-results.txt

// Get top 10 words from each cluster and write it to top10Words.txt

1. $cd $HADOOP\_HOME
2. $./bin/mahout clusterdump -dt sequencefile -d assignment3/vectors/dictionary.file-0 -i assignment3/kmeans/clusters-1-final -o /Users/himansubadhai/Documents/input/assignment3/top10Words.txt

**Output File:** cluster-docs.txt

**Output File in readable format:** cluster-results.txt

**Final Output:** top10Words.txt

The top10Words.txt file contains both the cluster data and the top 10 words below it.

The given view.pl had to be customized to get the output in proper readable format.

Please use the view.pl file given in the submitted folder.

View.pl

#!/usr/bin/perl

open(OUTPUT,"$ARGV[0]");

open(FILEOUT,">$ARGV[1]");

while(<OUTPUT>){

chomp;

if($\_ =~ /Key: (.+?): Value: (.+?): \/(.+) \*=/){

print FILEOUT "$1\t$3\n";

}

}

close FILEOUT;

print "Done\n";