I have created sub directories for each of the problems and written the mapper and reducer files with the names mapper.py and reducer.py

First load the data into HDFS by using

hdfs dfs –put warandpeace.txt input1.txt

hdfs dfs –put data2006.txt input2.txt

hdfs dfs –put data2008.txt input3.txt

Task 1 Problem A

Mapper.py and reducer.py are mentioned in the problem 1a folder

hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-mr1.jar -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input /user/cloudera/input1.txt -output /user/cloudera/problem1a/output

hadoop dfs -get problem1a/output #Download Output directory into local file system

gedit output/part-00000 #This will show the output

Task 1 Problem B

Mapper.py and reducer.py are mentioned in the problem 1b folder

hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-mr1.jar -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input /user/cloudera/input1.txt -output /user/cloudera/problem1b/output

hadoop dfs -get problem1b/output #Download Output directory into local file system

gedit output/part-00000 #This will show the output

Task 2 Problem A

Mapper.py and reducer.py are mentioned in the problem2a folder

hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-mr1.jar -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input /user/cloudera/input3.txt -output /user/cloudera/problem2a/output

hadoop dfs -get problem2a/output #Download Output directory into local file system

gedit output/part-00000 #This will show the output

Task 2 Problem B

Mapper.py and reducer.py are mentioned in the problem2b/1 folder

hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-mr1.jar -file mapper.py -mapper first_mapper.py -file reducer.py -reducer reducer.py -input /user/cloudera/input2.txt -output /user/cloudera/problem2b/1/output

hadoop dfs -get problem2b/1/output #Download Output directory into local file system

#Now we have to put this output into hdfs to take it as input for the second mapper and reducer.

hadoop dfs –put output/part-00000 input2b.txt

#Now we will execute the second mapper and reducer

Mapper.py and reducer.py are mentioned in the problem2b/2 folder

hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-mr1.jar -file mapper.py -mapper first_mapper.py -file reducer.py -reducer reducer.py -input /user/cloudera/input2b.txt -output /user/cloudera/problem2b/2/output_final

hadoop dfs -get problem2b/2/output_final #Download Output directory into local file system

gedit output_final/part-00000   #This will show the output

Task 2 Problem C

Mapper.py and reducer.py are mentioned in the problem2c/1 folder

hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-mr1.jar -file mapper.py -mapper first_mapper.py -file reducer.py -reducer reducer.py -input /user/cloudera/input2.txt -output /user/cloudera/problem2c/1/output1

hadoop dfs -get problem2c/1/output1 #Download Output directory into local file system

hadoop dfs –put output1/part-00000 input2c1.txt

hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-mr1.jar -file mapper.py -mapper first_mapper.py -file reducer.py -reducer reducer.py -input /user/cloudera/input3.txt -output /user/cloudera/problem2c/1/output2

hadoop dfs -get problem2c/1/output2 #Download Output directory into local file system

hadoop dfs –put output2/part-00000 input2c2.txt

Now we will run the second mapper

Mapper.py is in the problem2c/2 folder

hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-mr1.jar -file mapper.py -mapper first_mapper.py -cacheFile input2c2.txt#ref1 -input /user/cloudera/input2c1.txt -output /user/cloudera/problem2c/2/output_final

hadoop dfs -get problem2c/2/output_final #Download Output directory into local file system

gedit output_final/part-00000 #This is the final output

Task 2 Problem D

Mapper.py and reducer.py are mentioned in the problem2d folder

hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-mr1.jar -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input /user/cloudera/input2.txt -cacheFile input3.txt#ref -output /user/cloudera/problem2d/output

hadoop dfs -get problem2d/output #Download Output directory into local file system

gedit output/part-00000   #This will show the output