**Initial Steps:**

Place all the folders (task1…knn) in home/cse587 path.

Create an empty folder results in home/cse587 path along with knn, task1, task2, task3, task4 sub-directories within results.

Start Hadoop: start-all.sh

**Create folders for results/data in dfs.**

hdfs dfs -mkdir /home

hdfs dfs -mkdir /home/task1

hdfs dfs -mkdir /home/task2

hdfs dfs -mkdir /home/task3

hdfs dfs -mkdir /home/task4

hdfs dfs -mkdir /home/knn

hdfs dfs -mkdir /home/demopy

hdfs dfs -mkdir /home/knn\_traindata

hdfs dfs -mkdir /home/task4\_data

**Copy the data from local to dfs:**

hdfs dfs -copyFromLocal ./gutenberg/\* /home/demopy

hdfs dfs -copyFromLocal ./knn\_data/TrainNN.csv /home/knn\_traindata

hdfs dfs -copyFromLocal ./task4\_data/\* /home/task4\_data

**Check if all the files and folders are correctly created or copied.**

hdfs dfs -ls /home

hdfs dfs -ls /home/demopy

hdfs dfs -ls /home/knn\_traindata

hdfs dfs -ls /home/task4\_data

**Task1:**

hadoop jar /home/cse587/hadoop-3.1.2/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar -D mapred.map.tasks=5 -D mapred.reduce.tasks=4 -file /home/cse587/task1/mapper\_task1.py -mapper /home/cse587/task1/mapper\_task1.py -file /home/cse587/task1/reducer\_task1.py -reducer /home/cse587/task1/reducer\_task1.py -input /home/demopy -output /home/task1/out1

hdfs dfs -ls /home/task1/out1

hdfs dfs -copyToLocal /home/task1/out1/part-00000 ./results/task1/result\_0.txt

hdfs dfs -copyToLocal /home/task1/out1/part-00001 ./results/task1/result\_1.txt

hdfs dfs -copyToLocal /home/task1/out1/part-00002 ./results/task1/result\_2.txt

hdfs dfs -copyToLocal /home/task1/out1/part-00003 ./results/task1/result\_3.txt

**Task2:**

**First Mapper, Reducer:**

hadoop jar /home/cse587/hadoop-3.1.2/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar -D mapred.map.tasks=5 -D mapred.reduce.tasks=4 -file /home/cse587/task2/mapper\_task2.py -mapper /home/cse587/task2/mapper\_task2.py -file /home/cse587/task2/reducer\_task2.py -reducer /home/cse587/task2/reducer\_task2.py -input /home/demopy -output /home/task2/out1

hdfs dfs -ls /home/task2/out1

hdfs dfs -copyToLocal /home/task2/out1/part-00000 ./results/task2/result\_0.txt

hdfs dfs -copyToLocal /home/task2/out1/part-00001 ./results/task2/result\_1.txt

hdfs dfs -copyToLocal /home/task2/out1/part-00002 ./results/task2/result\_2.txt

hdfs dfs -copyToLocal /home/task2/out1/part-00003 ./results/task2/result\_3.txt

**Second Mapper, Reducer:**

hadoop jar /home/cse587/hadoop-3.1.2/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar -D mapred.map.tasks=5 -D mapred.reduce.tasks=4 -file /home/cse587/task2/mapper1\_task2.py -mapper /home/cse587/task2/mapper1\_task2.py -file /home/cse587/task2/reducer1\_task2.py -reducer /home/cse587/task2/reducer1\_task2.py -input /home/task2/out1 -output /home/task2/out2

hdfs dfs -ls /home/task2/out2

hdfs dfs -copyToLocal /home/task2/out2/part-00000 ./results/task2/final\_result\_0.txt

hdfs dfs -copyToLocal /home/task2/out2/part-00001 ./results/task2/final\_result\_1.txt

hdfs dfs -copyToLocal /home/task2/out2/part-00002 ./results/task2/final\_result\_2.txt

hdfs dfs -copyToLocal /home/task2/out2/part-00003 ./results/task2/final\_result\_3.txt

**Task3:**

hadoop jar /home/cse587/hadoop-3.1.2/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar -D mapred.map.tasks=5 -D mapred.reduce.tasks=4 -file /home/cse587/task3/mapper\_task3.py -mapper /home/cse587/task3/mapper\_task3.py -file /home/cse587/task3/reducer\_task3.py -reducer /home/cse587/task3/reducer\_task3.py -input /home/demopy -output /home/task3/out1

hdfs dfs -ls /home/task3/out1

hdfs dfs -copyToLocal /home/task3/out1/part-00000 ./results/task3/result\_0.txt

hdfs dfs -copyToLocal /home/task3/out1/part-00001 ./results/task3/result\_1.txt

hdfs dfs -copyToLocal /home/task3/out1/part-00002 ./results/task3/result\_2.txt

hdfs dfs -copyToLocal /home/task3/out1/part-00003 ./results/task3/result\_3.txt

**Task4:**

hadoop jar /home/cse587/hadoop-3.1.2/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar -D mapred.map.tasks=5 -D mapred.reduce.tasks=4 -file /home/cse587/task4/mapper\_task4.py -mapper /home/cse587/task4/mapper\_task4.py -file /home/cse587/task4/reducer\_task4.py -reducer /home/cse587/task4/reducer\_task4.py -input /home/task4\_data -output /home/task4/out1

hdfs dfs -ls /home/task4/out1

hdfs dfs -copyToLocal /home/task4/out1/part-00000 ./results/task4/result\_0.txt

hdfs dfs -copyToLocal /home/task4/out1/part-00001 ./results/task4/result\_1.txt

hdfs dfs -copyToLocal /home/task4/out1/part-00002 ./results/task4/result\_2.txt

hdfs dfs -copyToLocal /home/task4/out1/part-00003 ./results/task4/result\_3.txt

**KNN:**

hadoop jar /home/cse587/hadoop-3.1.2/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar -D mapred.map.tasks=5 -D mapred.reduce.tasks=4 -file /home/cse587/knn/mapper\_knn.py -mapper /home/cse587/knn/mapper\_knn.py -file /home/cse587/knn/reducer\_knn.py -reducer /home/cse587/knn/reducer\_knn.py -input /home/knn\_traindata -output /home/knn/out1

hdfs dfs -ls /home/knn/out1

hdfs dfs -copyToLocal /home/knn/out1/part-00000 ./results/knn/result\_0.txt

hdfs dfs -copyToLocal /home/knn/out1/part-00001 ./results/knn/result\_1.txt

hdfs dfs -copyToLocal /home/knn/out1/part-00002 ./results/knn/result\_2.txt

hdfs dfs -copyToLocal /home/knn/out1/part-00003 ./results/knn/result\_3.txt