Assignment 2
KNN Implementation in MapReduce and Pig

1)Which InputFormat did you use in the MapReduce program?
TextInputFormat

2)What is the input and output format of the map function?
input is the Object and Text
output is DoubleWritable and Text

3)What is the logic of the map function?
mapper function takes x and y coordinates and computes the distance of that point with the given input point and emits distance as the key and ID of that point as value.

4)If a combiner function is used, what is the signature of the combiner function (input and output) and what is its logic?
I have used combiner in the code.
I tried running it without the combiner, But combiner makes it more efficient.

5)If a reduce function is used, what is the signature of the reduce function (input and output) and what is its logic?
Reducer takes DoubleWritable and text format inputs and its output is Text, DoubleWritable
logic:

Basically when Reducer takes the input, which is the output from mapper will be sorted by the keys due to the sorting So in mapper I have swapped key and values from id, distance to distance. id
The input to the reducer is sorted. So in reducer I take the value of K and keep a counter to read values till K and emit only the top k results.
Sorting only takes place when there is a reduce phase.

6)How many mappers and reducers are needed for your program?
program used 14 mappers and I have set the number of reducer to 1.

7)How many records are shuffled between the mappers and reducers?

Shuffled Maps =14
Map output records=10507403
Combine input records=10507403
Combine output records=56
Reduce input records=56

According to my knowledge it depends also on the K value the above results were for k=4

So 56 records were shuffled between mapper and reducer for k=4.

when k was smaller (k=2) the value was 28
when k was larger (k=100) the value was 1400

8)For the Pig Latin program, how many MapReduce jobs are needed to run the program? How does this compare to the MapReduce implementation?
Pig took 14 mappers and 1 reducer.
But it took more time in pig than the MapReduce program. PIg showed statistics for each task in separate mapReduce jobs. It used map only job for one task. It also used combiner for one.