Class – CS6240 Fall-2018 Sec 2
HW – 4
Name- Mustafa Kapadia
Github - https://github.ccs.neu.edu/cs6240f18/mustafa8895/tree/master/HW4

Pseudo Code for K-Means

Job-1 : Getting the follower count

```
Map(from, to)
    Emit (to,1)

Reduce(to, values(list of 1's))
    Emit (to , sum of values)
```

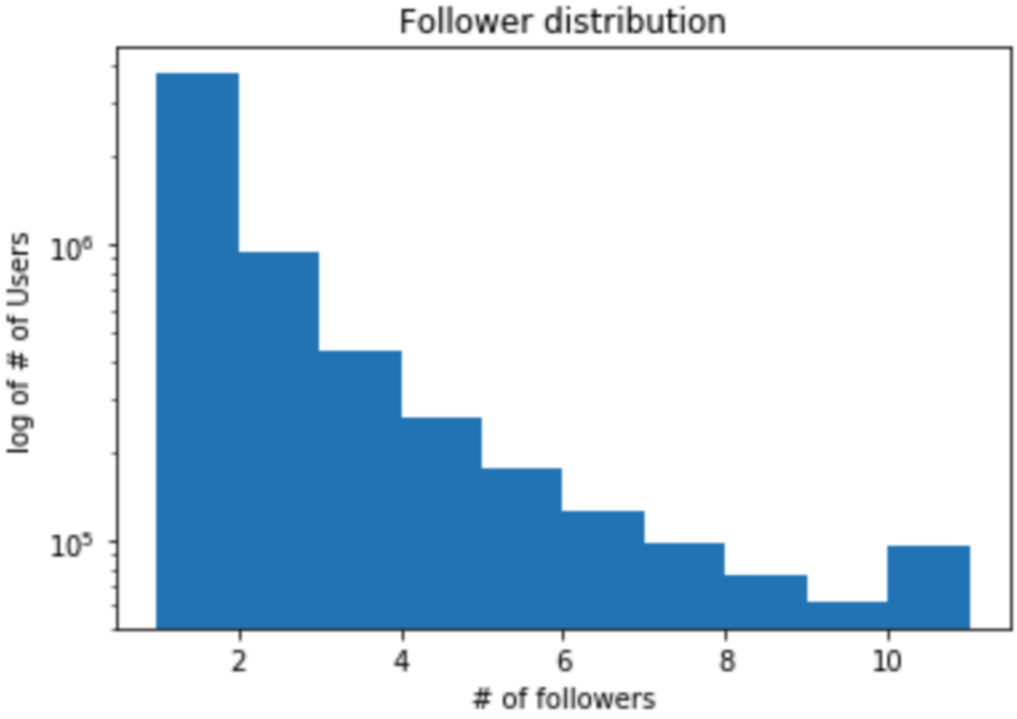
Combiner function is the same as the reduce function.

Job-2 : finding centroids

```
Map(node , followers)
    Fetch centroids from context
    Find centroid closestCentroid which is the closest centroid to the node
    Emit(closestCentroid, followers)

Reduce(centroid, list of follower counts L)
    Fetch centroids from context
    Sse = 0
    Sum = 0
    Count = 0

    For each value in L:
        Sum = Sum + value
        Count ++
        Sse = Sse + (value – centroid)^2

    newCentroid = Sum / count

    if | newCentroid – centroid | > threshold
        Increment global counter
```

Repeat job 2 until counter value is 0 (Convergence)

Plot of number of users per follower count



All users with followers greater than 10 have been grouped into the same bin

Good centers - **0, 188170, 376340, 564512**
Bad Centers – **1, 2, 3, 4**

Configuration	5 workers, bad start	5 workers, good start	10 workers, bad start	10 workers, good start
Number iterations executed	10	10	10	10
Running time	9 minutes	9 minutes	9 minutes	8 minutes
Final cluster centers found	4845.062322946175, 3.2295836373849474, 57494.734006734005, 130.45172485023755	564512.0, 10.44698677657316, 209974.55555555556, 54375.55378486056	57494.734006734005, 4845.062322946175, 130.45172485023755. 3.2295836373849474,	564512.0, 54375.55378486056 10.44698677657316 209974.55555555556
SSE after iteration 1	1.917255797091E12	9.71424777893E11	1.917255797091E12	9.71424777893E11

SSE after iteration 2	1.9114716995749321E12	7.265858192877682E11	1.9114716995745222E12	7.265858192858156E11
SSE after iteration 3	1.8949139662828503E12	5.973584092978839E11	1.894913966282839E12	5.973584092998177E11
SSE of final clustering	8.771432846843634E11	4.691595371681173E11	8.77143284684363E11	4.6915953717153534E11

As seen from the table above, the program shows no speedup.