Assignment 0

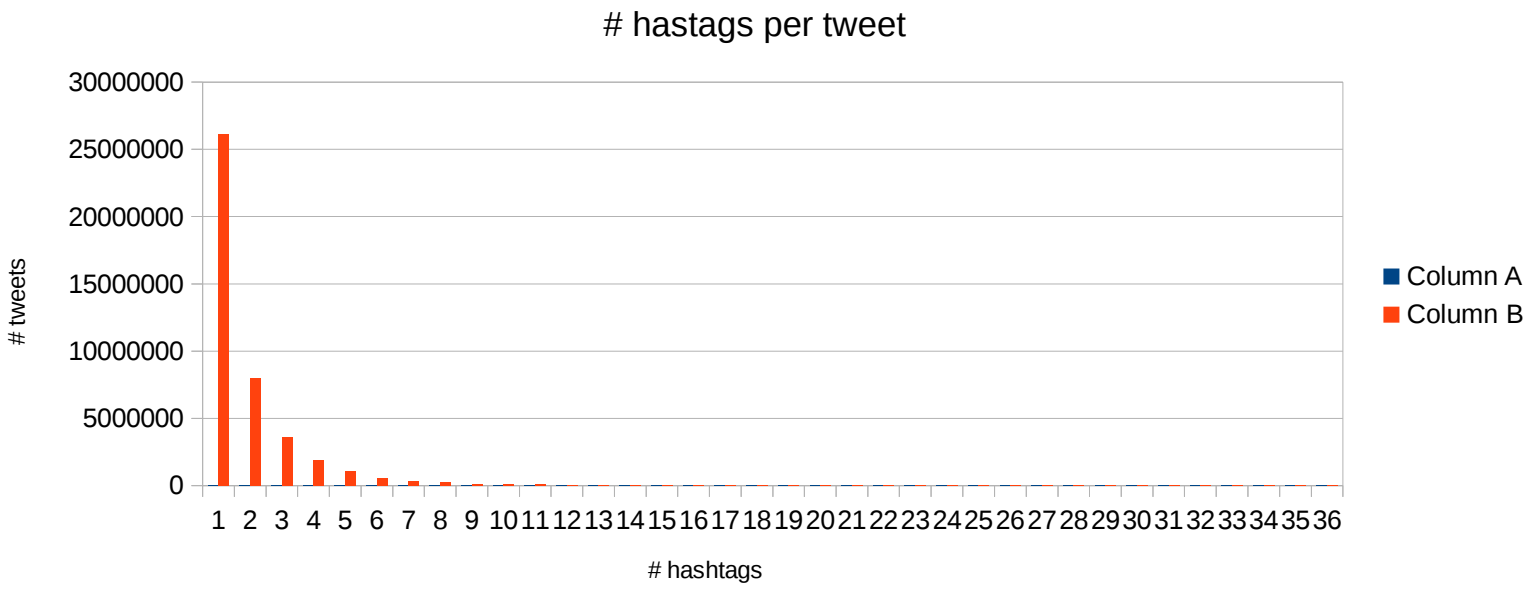
SYSTEM SPECEFICTION

- Computing Cluster with 24 nodes each with 8 core each
- Total available memory of system 522 GB

EXPERIMENT SETUP

- Driver memory =512MB
- No. of Executors = 8
- No. of cores with each executors =4
- Memory with Each Executor=4GB

Question1) Following plot was obtained from the experiment



Tweets with zero hashtags were more than mahority (175003599 in no.) hence excluded from the plot to make data more clear.

Refer-https://github.com/nileshrathi/ds256-jan2019/blob/master/ds256-a0/question1.csv

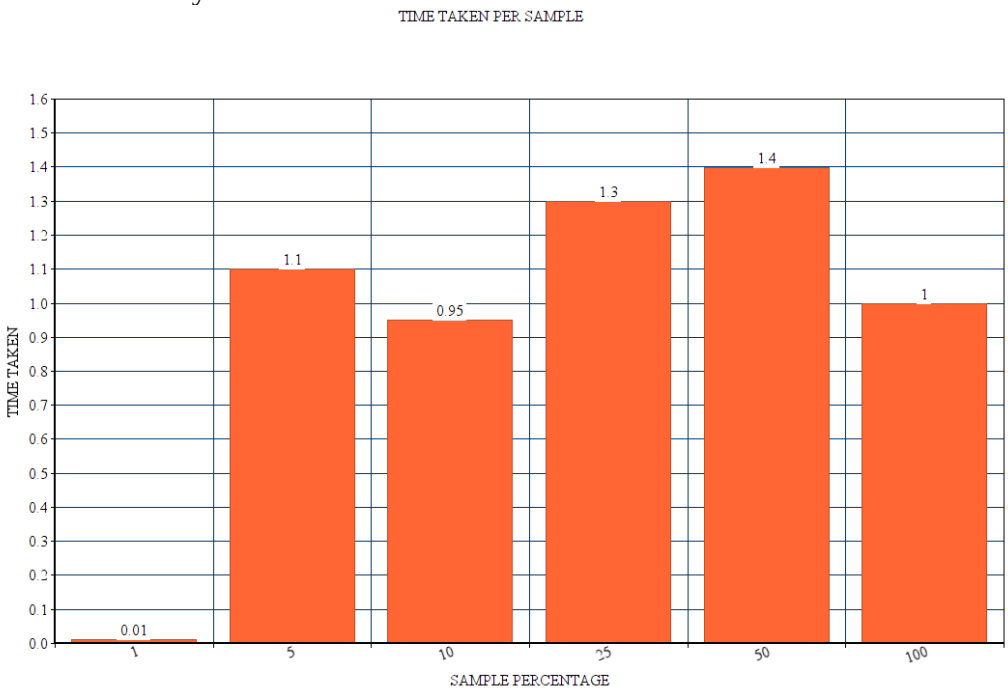
Question 3) Following mearsure was taken using Count() call of final RDDs.

no. of vertices=39967257
no of edges = 93648762

Question 4)

Analysis.) Sampling of data was done without replacememt. Depending upon the load on the cluster there was gradual varialtion in time. To obtain a uniform distribution the whole dataset need to be scanned for that sample to be taken. So timings are nearly same but we could se a linear

increase in time from 10% -> 25% -> 50% . This was taken when cluster was busy. For 100% sample the memory per core was increased to 8GB and it was collected when cluster was relatively free. So it returned in nearly same time.



Refer-
https://github.com/nileshrathi/ds256-jan2019/blob/master/ds256-a0/question%204.ods