Assignment 1

COSI 120A Topics in Computer Systems 2020 Fall

Due: 12:00 am, 09/18/2020

1 [20 marks]

Suppose you have an input of very large file of integers, design a MapReduce algorithm to output the the following:

a) output how many distinct integers from the file [5 marks]

b) output how many times each distinct integer appear [5 marks]

c) output the smallest integer [5 marks]

d) output the average of all integers [5 marks]

2 [30 marks]

Write a hadoop/spark program of K-Means clustering. Test it on MNIST and calculate the accuracy.

Here is the link for MNIST.

Bonus [20 marks]

Try your K-Means program on Google Cloud Platform, and report the time costs with different settings (number of computers, and size of dataset).

You may need to re-size the dataset into four sizes: 1/8, 1/4, 1/2, original size for the experiment.