Assignment 4

Q1. (5 pts) Complete the posted Python notebook.

Q2 (3 pts)

- Consider the problem of finding triangles in a social network graph.
- We are given a graph as input and want to find all triples of nodes such that in the graph there are edges between each pair of these three nodes.
- To model this problem, we need to assume a domain for the nodes of the input graph with n nodes.
- An **output** is thus a set of **three nodes**, and an **input** is a set of **two nodes**. The output {u, v, w} is mapped to the set of three inputs {u, v}, {u, w}, and {v, w}.
- Find the lower bound on replication rate r given reducer size q for the triangle finding problem.

Steps 1 and 2 of showing the lower bound are given in an accompanying file. You should complete steps of 3 and 4 of the process.

Q3 (1 pts) Suppose that the universal set is {1, 2, …, 10}, and signatures for sets are constructed using the following list of permutations:

(1,2,3,4,5,6,7,8,9,10)
(10,8,6,4,2,9,7,5,3,1)
(4,7,2,9,1,5,3,10,6,8)

Construct minhash signatures for the following sets:

a. {3, 6, 9}.
b. {2,4,6,8}
c. {2,3,4}

How does the estimate of the Jaccard similarity for each pair, derived from the signatures, compare with the true Jaccard similarity?

Q4 (1 pts) Suppose we have two sets S = {1,3,4}, T = {2,3,5} and use two hash functions,

h1= 2x + 4 mod 5
h2(x) = 3x − 1 mod 5

to produce minhash signatures.
Write the minhash signatures we obtain using these hash functions.

Q5 (3 pts) We would like to use a MapReduce framework to compute minhash signatures. Suppose each map task is given a chunk of pairs (x,X), where is x is an element in set X, as well as all the hash functions h_1, h_2, . .. , h_m needed to compute the signatures. Give the mapper and reducer code to compute the minhash signatures of the sets. Use combiner to reduce network traffic.
Here is an input based on the previous exercise.
(1,S), (3,S), (4,S), (2,T), (3,T), (5,T)
Assume there is no guarantee how these pairs are distributed across different map tasks.

Q6 (2 pts) Consider the following example of columns and signatures:
we have
100,000 columns and
signatures of 100 integers.
Choose 20 bands of 5 integers/band.

a. What is the probability that we miss a pair of similar columns when the (Jaccard) similarity threshold is 60%? The values of *b* and *r* are as in the example: 20 and 5, respectively.
b. How should we change *b* and *r* such that the probability of missing a pair of similar columns (for a 60% threshold) is about 1/3000?