CMPSCI590D: Algorithms for Data Science

Homework 2

Instructor: David Wemhoener — Posted: Oct 6th, Due: Oct 16th at 4 pm

Do not look up materials on the Web. You can consult the reference books mentioned on the course website, and also the class slides for solving the homework problems. You may work in a group of size at most 3*. Every person is allowed to talk to at most 2 others. All communications are bidirected: A talks with B, automatically implies B talks with A. Therefore, those who are working in a group of* 3 *are not allowed to talk with anyone else outside the group. Mention any collaboration clearly in the submission. Submit one homework solution per group. No late homework will be accepted.*

For programming assignments, submit your code with a detailed readme file that contains instruction for running it. Also include any test dataset that you have used and results obtained to show correctness of your implementation.

*Total Points:*100*, Bonus Points:*20

Exercise 1. *Suppose we have n bits of memory available, and our set S has m members. Instead of using k hash functions, we could divide the n bits into k arrays, and hash once to each array. As a function of n, m, and k, what is the probability of a false positive? How does it compare with using k hash functions into a single array? [20]*

Exercise 2. *Design MapReduce algorithms to take in a large file on integers and produce as output:*

- *(a) The largest integer [5]*
- *(b) The average of all the integers [5]*
- *(c) The same set of integers, but with each integer appearing only once [5]*
- *(d) The count of the number of distinct integers in the input [5]*

Exercise 3. *Suppose we execute the word-count MapReduce program described in this section on a large repository such as a copy of the Web. We shall use 100 Map tasks and some number of Reduce tasks.*

- *(a) If we combine the reducers into a small number of Reduce tasks, say 10 tasks, at random, do you expect the skew to be significant? What if we instead combine the reducers into 10,000 Reduce tasks? [10]*
- *(b) Suppose we do use a combiner at the 100 Map tasks. Do you expect skew to be significant? Why or why not? [10]*

Exercise 4. *In this assignment, the goal is to implement a set of simple Map Reduce tasks. Please include the Python scripts used in the submitted report. You will work with a collection of e-mail data downloadable from:* `https://snap.stanford.edu/data/email-EuAll.txt.gz` *The data forms a graph G of e-mails between users, with each line being of the form* ***sender receiver****. Compute the following on G:*

- *Number of nodes in the graph [5]*
- *Average (and median) indegree and out degree [10]*
- *Average (and median) number of nodes reachable in two hops [15]*
- *Number of nodes with indegree* > 100 *[10]*

Exercise 5. *Suppose a job consists of n tasks, each of which takes time t seconds. Thus, if there are no failures, the sum over all compute nodes of the time taken to execute tasks at that node is nt. Suppose also that the probability of a task failing is p per job per second, and when a task fails, the overhead of management of the restart is such that it adds 10t seconds to the total execution time of the job. What is the total expected execution time of the job? [20]*