COMP 6730 Advanced Database Systems Homework 1

Fall 2016

Due Wed. Oct. 5, in Class

Problem 1 (10 points)

Suppose we roll a fair k-sided die with the numbers 1 through k on the die's faces. If X is the number that appears, what is E[X]?

Problem 2 (10 points)

A monkey types on a 26-letter keyboard that has lowercase letters only. Each letter is chosen independently and uniformly at random from the alphabet. If the monkey types 1,000,000 letters, what is the expected number of times that sequence "proof" appears?

Problem 3 (25 points)

The following approach is often called *reservoir sampling*. Suppose that we have a sequence of items, passing by one at a time. We want to maintain a sample of one item that has the property that it is uniformly distributed over all the items that we have seen at each step. Moreover, we want to accomplish this without knowing the total number of items in advance or storing all of the items that we see.

Consider the following algorithm, which stores just one item in memory at all times. When the first item appears, it is stored in the memory. When the k-th item appears, it replaces the item in memory with probability $1/k$. Explain why this algorithm solves the problem.

Problem 4 (15 points)

Suppose that we roll a standard fair die 100 times. Let X be the sum of the numbers that appear over the 100 rolls. Use Chebyshev's inequality to bound $\Pr(|X - 350| \geq 50)$.

Problem 5: Exercise 2.3.1 (page 58) (25 points)

Design MapReduce algorithms to take a very large file of integers and produce as output:
(a) The largest integer.
(b) The average of all the integers.
(c) The same set of integers, but with each integer appearing only once.
(d) The count of the number of distinct integers in the input.

Problem 6: Exercise 3.1.1 (page 95) (15 points)

Compute the Jaccard similarities of each pair of the following three sets: {1, 2, 3, 4}, {2, 3, 5, 7}, and {2, 4, 6}.