Engineering and Applied Science Programs for Professionals
Whiting School of Engineering
Johns Hopkins University
685.621 Algorithms for Data Science
Homework 1
Assigned at the start of Module 1
Due at the end of Module 2

Total Points 100/100

Collaboration groups will be set up in Blackboard by the end of the week. Make sure your group starts an individual thread for each collaborative problem and subproblem. You are required to participate in each of the collaborative problem and subproblem.

Self-Study Problems

All of the following problems come from the textbook and have solutions posted on the web at http://mitpress.mit.edu/algorithms.

You are permitted to use this site to examine solutions for these problems as a means of self-checking your solutions. These problems will not be graded.

Problems: 2.2-2, 2.3-5, 3.1-2, 12.1-2, 12.3-3, 21.2-6.

Problems for Grading

1. **Problem 1 Chapter 2** ***Note this is a Collaborative Problem***
   20 Points Total

   Use induction to prove $\sum_{i=1}^{n} i^3 = \left(\frac{n(n+1)}{2}\right)^2$.

2. **Problem 2 Parts a, b, c, d, e and f**
   30 Points Total 5 Points Each

   Although merge sort runs in $\Theta(nlgn)$ worst-case time and insertion sort runs in $\Theta(n^2)$ worst-case time, the constant factors in insertion sort can make it faster in practice for small problem sizes on many machines. Thus, it makes sense to coarsen the leaves of the recursion by using insertion sort within merge sort when subproblems become sufficiently small. Consider a modification to merge sort in which n/k sublists of length k are sorted using insertion sort and then merged using the standard merging mechanism, where k is a value to be determined.

   (a) Use insertion sort to sort the unsorted array $< 40, 17, 45, 82, 62, 32, 30, 44, 93, 10 >$. Make sure to show the array after every pass.

   (b) Use merge sort to sort the unsorted array $< 75, 56, 85, 90, 49, 26, 12, 48, 40, 47 >$. Make sure to show the steps of splitting the array then merging the array.

   (c) Show that insertion sort can sort the n/k sublists, each of length k, in $\Theta(nk)$ worst-case time.

   (d) Show how to merge the sublists in $\Theta(nlg(n/k))$ worst-case time.

   (e) Given that the modified algorithm runs in $\Theta(nk + nlg(n/k))$ worst-case time, what is the largest value of k as a function of n for which the modified algorithm has the same running time

as standard merge sort, in terms of Θ-notation?

(f) How should we choose k in practice?

3. **Problem 3**
   15 Points Total

   Write a $\Theta(m+n)$ algorithm that prints the in-degree and the out-degree of every vertex in an m-edge, n-vertex directed graph where the directed graph is represented using adjacency lists.

4. **Problem 4 Chapter 12 Binary Search Trees**
   25 Points Total 5 Points Each

   Exercise 12.2-1. Suppose that we have numbers between 1 and 1000 in a binary search tree and we want to search for the number 363. Which of the following sequences could not be the sequence of nodes examined?

   i. 2, 252, 401, 398, 330, 397, 363.

   ii. 924, 220, 911, 244, 898, 258, 362, 363.

   iii. 925, 202, 911, 240, 912, 245, 363.

   iv. 2, 399, 387, 219, 266, 382, 381, 278, 363.

   v. 935, 278, 347, 621, 299, 392, 358, 363.

5. **Problem 5 Sorting Iris Plants** ***Note this is a Collaborative Problem***
   10 Points

   The Iris Plants Database contains 3 classes of 50 instances each, where each class refers to a type of Iris plant. Five attributes/features were collected for each plant instance. The dataset can be downloaded from iris.arff on the Sample Weka Data Sets webpage (https://storm.cis.fordham.edu/ gweiss/data-mining/datasets.html). Develop an algorithm to sort the five features in the dataset to determine if any of the five sorted features can separate the three plant types. Show the efficiency of your sorting algorithm based on its ability to sort the five sets of features.