CSCI 3104 Fall 2022
Instructors: Prof. Grochow and Chandra Kanth Nagesh

Problem Set 10 SOLUTION

Due Date ..... November 14, 2022
Name ..... **Your Name**
Student ID ..... **Your Student ID**
Collaborators ..... **List Your Collaborators Here**

Contents

Instructions

- The solutions **must be typed**, using proper mathematical notation. We cannot accept hand-written solutions. Useful links and references on LaTeXcan be found here on Canvas.
- You should submit your work through the **class Canvas page** only. Please submit one PDF file, compiled using this LaTeX template.
- You may not need a full page for your solutions; pagebreaks are there to help Gradescope automatically find where each problem is. Even if you do not attempt every problem, please submit this document with no fewer pages than the blank template (or Gradescope has issues with it).
- You are welcome and encouraged to collaborate with your classmates, as well as consult outside resources. You must **cite your sources in this document. Copying from any source is an Honor Code violation. Furthermore, all submissions must be in your own words and reflect your understanding of the material.** If there is any confusion about this policy, it is your responsibility to clarify before the due date.
- Posting to **any** service including, but not limited to Chegg, Reddit, StackExchange, etc., for help on an assignment is a violation of the Honor Code.
- You **must** virtually sign the Honor Code (see Section Honor Code). Failure to do so will result in your assignment not being graded.

Honor Code (Make Sure to Virtually Sign the Honor Pledge)

Problem HC. On my honor, my submission reflects the following:

- My submission is in my own words and reflects my understanding of the material.
- Any collaborations and external sources have been clearly cited in this document.
- I have not posted to external services including, but not limited to Chegg, Reddit, StackExchange, etc.
- I have neither copied nor provided others solutions they can copy.

In the specified region below, clearly indicate that you have upheld the Honor Code. Then type your name.

Honor Code. □

26 Standard 26 – Hash Tables

26.1 Problem 1

Consider the following hash function. Let U be the universe of strings composed of the characters from the alphabet $\Sigma = [\texttt{A}, \ldots, \texttt{Z}]$, and let the function $f(x_i)$ return the index of a letter $x_i \in \Sigma$, e.g., $f(\texttt{A}) = 1$ and $f(\texttt{Z}) = 26$. Finally, for an m-character string $x \in \Sigma^m$, define $h(x) = ([\sum_{i=1}^{m} f(x_i)] \mod \ell)$, where ℓ is the number of buckets in the hash table. That is, our hash function sums up the index values of the characters of a string x and maps that value onto one of the ℓ buckets.

Suppose this is going to be used to hash words from a large body of English text.

List at least 4 reasons why $h(x)$ is a bad hash function relative to the ideal behavior of uniform hashing.

Answer. There are many valid reasons. Here are some examples.

- For $h(x)$, the location of the letters within x doesn't matter, only their individual values. Thus, any two strings that are anagrams (permutations) of each other will have the same hash value, e.g., $x =$ `SAUCE` and $y =$ `CAUSE` will have the same hash value of $h(x) = h(y) = 49$.
- We can also add an arbitrary value to one character and subtract it from another character, and then rearrange the letters to get a different string with the same score as the original string. For instance, add 1 to `C` to make a `D`, then subtract 1 from `U` to make a `T`, and finally permute the letters to convert `CAUSE` to `DATES`.
- We can repeat the previous attack in various combinations to get a string that contains none of the original characters, e.g., `GIDDY`.
- In fact, strings of different lengths can also have the same hash value, e.g., `FRY` and `HUT`, or even `ABDICATED`.
- The last four points are ways we can demonstrate that $h(x)$ does not behave like a uniform hash; if it did, then the probability that a string x would hash to the same location as some y is $1/\ell$, meaning that "nearby" strings should not hash to nearby (or the same) location, and neither should "distant" strings.
- For strings with a typical length, like surnames or English words, some hash values will be improbable, in contrast to a uniform hash. For instance, only one string has a value of 1 (`A`), two have values of 2 (`AA` and `B`), four with value 3 (`AAA`, `AB`, `BA`, `C`), etc. (ignoring the mod function).
- Even if we focus only on strings with a fixed length, some characters are simply more common than others (the most common letters in English are `etaoin shrdlu`), and this will cause $h(x)$ will favor certain values over others.

□

26.2 Problem 2

Consider a chaining hash table A with b buckets that holds data from a fixed, finite universe U. Recall the definition of worst-case analysis, and consider starting with A empty and inserting n elements into A under the assumption that $|U| \leq bn$.

(a) What is the worst case for the number of elements that collide in a single bucket? Give an exact answer and justify it. **Do not assume the uniform hashing assumption for this question.**

Answer. In the worst case, all n elements could collide with the same hash function, so that bucket would have size n. □

(b) Calculate the worst-case total cost of these n insertions into A, and give your answer as $\Theta(f(n))$ for a suitable function f. Justify your answer.

Answer. Insertions cost $\Theta(1)$ regardless of the size of the hash table. So the total cost of n insertions is $n \cdot \Theta(1) = \Theta(n)$. □

(c) **For this part only, assume the uniform hashing assumption, and that the elements added were chosen uniformly at random from U.** After the n insertions, suppose that m `find` operations are performed. What is the total cost of these m find operations? Give your answer as $\Theta(f(n))$ for a suitable function f, and justify your answer.

Answer. Assuming the uniform hashing assumption, the n inserted items will be uniformly distributed among the b buckets, so each bucket will have size n/b. Each find operation thus costs $\Theta(n/b)$, so the total cost of m find operations is $m \cdot \Theta(n/b) = \Theta(mn/b)$. □

26.3 Problem 3

Hash tables and balanced binary trees can be both be used to implement a dictionary data structure, which supports insertion, deletion, and lookup operations. In balanced binary trees containing n elements, the runtime of all operations is $\Theta(\log n)$.

For each of the following three scenarios, compare the average-case performance of a dictionary implemented with a hash table (which resolves collisions with chaining using doubly-linked lists) to a dictionary implemented with a balanced binary tree.

(a) A hash table with hash function $h_1(x) = 1$ for all keys x.

Answer. In this case all the items collide in the bucket indexed 1, so the hash table functions just like a linked list. Thus the the Find and Delete operations cost $\Theta(n)$, which is much worse than the $\Theta(\log n)$ for a balanced binary tree. (Insertion, on the other hand, is $O(1)$, whereas in a balance binary tree it is $\Theta(\log n)$.) □

(b) A hash table with a hash function h_2 that satisfies the Simple Uniform Hashing Assumption, and where the number m of buckets is $\Theta(n)$.

Answer. In this case, the average-case cost of Find and Delete are $\Theta(1+\alpha) = \Theta(1+n/\Theta(n)) = \Theta(1)$, which is much better than the $\Theta(\log n)$ of a balanced binary tree, so we'd prefer the hash table implementation. □

(c) A hash table with a hash function h_3 that satisfies the Simple Uniform Hashing Assumption, and where the number m of buckets is $\Theta(n^{3/4})$.

Answer. In this case, the average-case cost of Find and Delete are $\Theta(1+\alpha) = \Theta(1+n/\Theta(n^{3/4})) = \Theta(n^{1/4})$. Although much better than $\Theta(n)$, this is still asymptotically *much* worse than $\Theta(\log n)$, so we would still prefer the balanced binary tree.

(For reference, the point at which $n^{1/4}$ becomes greater than $\log n$ is around $n \approx 65,000$. So, asymptotically, the answer written here is correct, but in practice, for n much smaller than that, one has to do a more careful analysis of the constants hidden by the asymptotic notation.) □