Learning Goals

During this lab, you will:

- motivate the need for hash tables
- review the Simple Uniform Hashing Assumption
- compare collision resolution strategies (chaining, open addressing, double hashing)

The Theory of Hashing

Direct Addressing

Imagine we had a very large dynamic set (such as all Penn students’ names) and we wanted quick lookups to see if a name existed. As you have seen, we could turn to a BST to store this information - the keys of the BST could be student names. This would give us $O(\lg n)$ time lookups and insertions, which is quite fast!

But say we’re going to be doing a lot of repeated lookups, so we’d like lookups on the order of constant time. How do we do this? One way is to define our universe $U$ of possible keys (which in this case is all possible strings up to size $l$, where $l$ is the length of the longest name) and keep an array $A$ the size of $U$. Then we can define a mapping $M$ where we assign every number from 1 to $\text{size}(U)$ to a string in $U$ and keep every name in the location $A[M(\text{name})]$.

This method would give us $O(1)$ lookups and inserts, but would take space $O(\text{size}(U))$ which can get very large indeed! For example, conservatively assuming that the longest name at Penn is 20 characters long (including spaces), there would be $2^7^{20}$ strings in our universe. This is clearly inefficient because there is no way we are going to be storing that many names.
Hash Tables

Instead of having a function with one-to-one mappings, we can define a hash function $h$ that maps all the keys in the universe to the $m$ slots in our table. Because it is usually the case that $m \ll \text{size}(U)$, there is a chance that two keys in the universe are mapped to the same slot in our table, which is known as a collision. Granted, if we knew exactly what data was going to be stored in the table beforehand, we could (somewhat painstakingly) find a perfect hash function that resulted in no collisions. But this is not usually the case. So, how do we choose a hash function? We could just set a predefined hash function and let it be, since it should perform decently well for any arbitrary dataset. What if, however, someone decided to hack our hash function? If this hash function was being used in some important piece of software and the hacker knew what keys collided in our hash function, they could devise a dataset that would slow down our code and bring our system to a halt. Thus, we avoid this worst-case behavior by choosing our hash function uniformly at random from a set of hash functions called a hash family.

The hash function $h$ is defined under the Simple Uniform Hashing Assumption, which says that every key $k \in U$ is equally likely to be mapped to any slot in the hash table $T$ by $h$. This ensures that each slot will get roughly the same number of collisions ($\frac{n}{m}$, which is also called the load factor or $\alpha$).

![Figure 2: Here, there is a collision between $k_2$ and $k_5$ since they are mapped to the same slot.](image)

Separate Chaining

The most obvious way to handle collisions is with a linked list. In other words, we chain all the keys that are hashed to slot $i$ in a linked list and store that list at slot $i$ in $T$. To find an element hashed to slot $i$, we would simply traverse the linked list until we found it (in the case of a successful search) or until we reached the end of the list (in the case of an unsuccessful search). Because of the simple uniform hashing assumption, the keys are roughly equally distributed amongst the $m$ linked lists in our hash table. With some math, we can show that the number of elements examined in a successful search is $\Theta(1 + \alpha)$ where $\alpha$ is the load factor. If $\alpha$ is small, we can say that this is $O(1)$. However, we do not always know what $\alpha$ is at any given moment - thus we say that we can perform insertions and searches in expected $O(1)$ time.

Open Addressing

Preface: In this discussion, we will not consider deletions, as they can be quite tricky to implement with open addressing and there are many different ways of handling them.
There are other ways besides chaining to handle collisions, and most of these strategies fall under the broad title of Open Addressing. In the technique of open addressing, instead of resolving collisions with a linked list, we simply find another open spot in the hash table to store our value. So, when inserting, if our hash function says we should hash to index $i$, and $T[i]$ is already full, we simply jump to another index in $T$. If that is full, we repeat, but if not, we insert our element there. If we keep jumping to another index in the table, we can write the indices we jump to as a sequence $S = \{h_0(x), h_1(x), h_2(x), \ldots, h_{m-1}(x)\}$, where $h_i(x)$ represents the $i$th index we check. We also say $h_0(x) = h(x)$. Defining $h_i(x)$ also gives us some order as to how we choose what index to go to next - otherwise, we would just be choosing randomly.

We can write the pseudocode for search (which is very similar to insertion) as follows:

```plaintext
function search(x):
    index ← h(x)
    for i ← 1 to m do
        if T[index] = x then
            return true
        else if T[index] = ∅ then
            return false
        end if
        index ← P(index)
    end for
    return false
end function
```

This pseudocode illustrates how open addressing works: once we check an index, we should never check it again, almost like it doesn’t exist in the table anymore. This set has cardinality $m$, since the element we are looking for could be at any other location in the array. If we are searching and come upon a blank element, we know that the element we are searching for cannot exist because if it was inserted, it would have been found before we hit a blank element. The expected time for an unsuccessful search with this setup is $O(1/\alpha)$, which is $O(1)$ for values of $\alpha$ not near $1^3$.

Probing Techniques

There are variety of ways to define our probing function $h_i(x)$. In linear probing, we define our $h_i(x) = (h(x) + i) \mod m$. This has us taking equal sized steps through the table. While quick when the load factor is small (due to caching effects), as the load factor becomes large, insertion() and search() can end up traversing the entire table. This is due to a problem called primary clustering. This is just what it sounds like. Because the difference between consecutive probes is linear, elements tend to form clusters. This is because once just a few elements are placed next to each other in the table, it becomes increasingly likely that an element will hash to one of the indices occupied by an element in that cluster. Then, because linear probing is used, that element will be placed near the cluster once it finally finds an open address, effectively increasing the size of the cluster. This can be mitigated by other forms of probing such as quadratic probing: $(h_i(x) = (h(x) + i^2) \mod m)$. However, quadratic probing suffers from a milder yet similar problem called secondary clustering. An even more robust form of open addressing is double hashing (where we choose another hashing function $h'(x)$ so $h_i(x) = (h(x) + i \* h'(x)) \mod m$). This makes the probe sequence vary depending on the element, effectively eliminating clustering.

Hash Maps in the Real World

The debate between chaining and open addressing does not have a clear-cut winner; there are merits to both solutions and both are widely used in the real world. For example, Java has chosen to implement hash tables with chaining, while Python chose to use open addressing with random probing.

Since we are working with Java, however, we will analyze Java’s implementation.
**Java HashMap**

Java’s [HashMap constructor](#) allows the user to define an initial capacity and a load factor. HashMap is implemented with separate chaining. The default capacity is 16, and the default load factor is 0.75.

HashMap rehashes every time \( \text{number of keys} > \text{load factor} \times \text{capacity} \).

The source code for HashMap is located [here](#). We will go through some interesting aspects of the implementation.

**Testing**

We will vary initial capacity and load factor values and look at some performance results.

In our test program, we call `put` on a million random numbers, then we call `get` on the same million numbers. We run this test 100 times to eliminate the possibility of bad sets of numbers or CPU usage spikes affecting the result. The average times are then recorded (in milliseconds) for the `put` and `get` operations.

The following table varies initial capacity and uses the default load factor 0.75.

<table>
<thead>
<tr>
<th>Initial Capacity</th>
<th>put</th>
<th>get</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2^4)</td>
<td>194.88</td>
<td>55.86</td>
</tr>
<tr>
<td>(2^5)</td>
<td>196.16</td>
<td>56.3</td>
</tr>
<tr>
<td>(2^6)</td>
<td>195.37</td>
<td>56.06</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(2^{17})</td>
<td>193.12</td>
<td>56.15</td>
</tr>
<tr>
<td>(2^{18})</td>
<td>188.83</td>
<td>55.26</td>
</tr>
<tr>
<td>(2^{19})</td>
<td>180.37</td>
<td>56.29</td>
</tr>
<tr>
<td>(2^{20})</td>
<td>142.55</td>
<td>54.31</td>
</tr>
<tr>
<td>(2^{21})</td>
<td>69.26</td>
<td>54.19</td>
</tr>
<tr>
<td>(2^{22})</td>
<td>64.58</td>
<td>48.94</td>
</tr>
<tr>
<td>(2^{23})</td>
<td>63.04</td>
<td>45.72</td>
</tr>
<tr>
<td>(2^{24})</td>
<td>63.17</td>
<td>44.89</td>
</tr>
</tbody>
</table>

We can see that `put` times are very stable until we get close to the 1 million mark, at which point they decrease dramatically, and stabilize after that. This is due to the Hashmap rehashing many times. For initial capacities of \(2^{21}\) or greater, no rehashing ever occurs. Setting initial capacity appropriately is very important to the runtime of HashMaps.

You may have noticed that `get` times decreased after the \(2^{21}\) point as well. This is because the table is sparse so there are less collisions on average.

Next we look at varying load factor, with an initial capacity of 16 (default).

<table>
<thead>
<tr>
<th>Load Factor</th>
<th>put</th>
<th>get</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>284.05</td>
<td>51.22</td>
</tr>
<tr>
<td>0.45</td>
<td>225.25</td>
<td>51.72</td>
</tr>
<tr>
<td>0.75</td>
<td>194.38</td>
<td>54.44</td>
</tr>
<tr>
<td>1</td>
<td>154.88</td>
<td>66.18</td>
</tr>
<tr>
<td>2</td>
<td>177.19</td>
<td>85.01</td>
</tr>
<tr>
<td>5</td>
<td>314.93</td>
<td>125.4</td>
</tr>
</tbody>
</table>

We can see here why HashMap has a default load factor of 0.75. Load factors smaller than 0.75 offer insignificant `get` improvements, while causing lots of table space to be wasted. On the other hand, load factors greater than 0.75 do in fact cause a slow down of `get` operations.

**Worst Case Linear Time**

With the correct load factor and initial capacity, we can observe the expected constant time access of HashMaps. However, for any hash function, there exists a set of inputs for which performance degrades to linear time. The following table shows inserting into a HashMap a set of values that are known to collide.
<table>
<thead>
<tr>
<th>Size of Input</th>
<th>Time(ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2^1$</td>
<td>4</td>
</tr>
<tr>
<td>$2^2$</td>
<td>9</td>
</tr>
<tr>
<td>$2^3$</td>
<td>19</td>
</tr>
<tr>
<td>$2^4$</td>
<td>36</td>
</tr>
<tr>
<td>$2^5$</td>
<td>73</td>
</tr>
<tr>
<td>$2^6$</td>
<td>155</td>
</tr>
<tr>
<td>$2^7$</td>
<td>296</td>
</tr>
<tr>
<td>$2^8$</td>
<td>742</td>
</tr>
<tr>
<td>$2^9$</td>
<td>1617</td>
</tr>
</tbody>
</table>

As mentioned before, the way to circumvent this is to pick a random hash function every time a new HashMap is created. The latest version of Java implements this.
Discussion

Chaining vs. Open Addressing

Compare chaining and open addressing and discuss the pros and cons of each.

Methods of Chaining

What effects would it have if, instead of chaining with linked lists, you used balanced binary search trees? What about if you used hash tables (in a sort of recursive fashion)?

Methods of Open Addressing

Compare linear probing and double hashing. What problems do each suffer from? Which is better?

Problems

Problem 1

You are given an array $A$ containing distinct randomly assorted integers. Your goal is to find two elements in the array whose sum is $k$ in $O(n)$ expected time.

Problem 2

Assume we have a hash table $T$ of size 10 that uses linear probing and has hash function $h(x) = x \mod 10$. We insert 6 numbers into $T$ and we get the below table:

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>52</td>
</tr>
<tr>
<td>6</td>
<td>46</td>
</tr>
<tr>
<td>7</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

What is one possible order that we could have inserted these elements to get this result? How many probes would be required for inserting 13 in the table?

Problem 3

How would you detect a cycle in a linked list of distinct elements in expected $O(n)$ time? Can you do it in constant space?

Problem 4

Design an algorithm that determines if two lowercase words are anagrams of each other in expected $O(n)$ time. Note: a string $A$ is an anagram of another string $B$ if $A$ is a permutation of $B$. Can you do it in worst case $O(n)$ time?
To be more specific, we can define a random variable \( X \) denoting the total number of elements we have hashed into \( T \) and \( X_i \) is the number of elements that have been hashed to slot \( i \).

\[
E[X_i] = \sum_{j=1}^{n} Pr[\text{element } j \text{ hashed to slot } i] = \sum_{j=1}^{n} Pr[h(k_j) = i]
\]

\[
E[X_i] = \frac{1}{m} \quad (\text{by Simple Uniform Hashing})
\]

As a check, we can use Linearity of Expectation to solve for \( E[X] \):

\[
E[X] = \sum_{i=1}^{m} E[X_i]
\]

\[
E[X] = \sum_{i=1}^{m} \frac{n}{m}
\]

\[
E[X] = n
\]

We make 2 assumptions: 1) assume that new elements are added to the front of the chain and 2) the element \( x \) we are searching for has an equal chance of being any of the \( n \) elements in the table.

Also, let us define an indicator random variable \( X_{ij} \) that is 1 if \( h(k_i) = h(k_j) \) for two keys \( k_i \) and \( k_j \) and 0 otherwise. Because of simple uniform hashing, we can say that \( Pr[h(k_i) = h(k_j)] = \frac{1}{m} \). Therefore,

\[
E[X_{ij}] = Pr[X_{ij} = 1] = Pr[h(k_i) = h(k_j)] = \frac{1}{m}
\]

Here’s the hard part. We claim that the expected number of elements examined in a successful search is

\[
E\left[\frac{1}{n} \sum_{i=1}^{n} \left( 1 + \sum_{j=i+1}^{n} X_{ij} \right)\right]
\]

Now, let’s unpack that dense bit of math. \( i \) is ranging over all elements that have been inserted into the hash table (remember \( n \) is the total number of elements in the hash table). \( \sum_{j=i+1}^{n} X_{ij} \) is the number of elements examined before the \( i^{th} \) element when searching for the \( i^{th} \) element. This is because of assumption 1 we made above. The extra 1 accounts for also examining the \( i^{th} \) element. Now what about this \( \frac{1}{n} \) business before the summation over \( i \)? Well, summing the number of elements that are examined when finding the \( i^{th} \) element over all \( i \) gives us the number of elements we need to examine to do all \( n \) successful searches. However, we only want the number of elements we need to examine to do a single successful search. Therefore, we take the average of this sum by dividing by \( n \). This amortized analysis can be a bit dense, so do read it a few times and really think about why it’s correct. Now that we’re through the hard part, all that’s left is to evaluate the expression:

\[
E\left[\frac{1}{n} \sum_{i=1}^{n} \left( 1 + \sum_{j=i+1}^{n} E[X_{ij}] \right)\right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( 1 + \sum_{j=i+1}^{n} \frac{n}{m} \right) \quad \text{(Linearity of Expectation)}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( 1 + \frac{n}{m} \sum_{j=i+1}^{n} \frac{1}{m} \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} 1 + \frac{n}{m} \sum_{i=1}^{n} \frac{n}{m}
\]

\[
= 1 + \frac{1}{nm} \sum_{i=1}^{n} 1
\]

\[
= 1 + \frac{1}{nm} \sum_{i=1}^{n} (n - i)
\]

\[
= 1 + \frac{1}{nm} \sum_{i=1}^{n} n - \frac{1}{nm} \sum_{i=1}^{n} i
\]

\[
= 1 + \frac{1}{nm} \cdot n^2 - \frac{1}{nm} \cdot \frac{n(n+1)}{2}
\]

\[
= 1 + \frac{1}{nm} \cdot n^2 - \frac{n(n+1)}{2m}
\]

\[
= 1 + \frac{1}{m} \cdot n - \frac{n(n+1)}{2m}
\]
\[ E[T(m, n)] = 1 + \frac{n}{m} E[T(m-1, n-1)] \]

The \(\frac{n}{m}\) term arises from the expected number of collisions.

Base Case: The base case arises at \(T(m, 0)\), which means that the hash table is empty. Thus the first probe always finds an empty element, so \(T(m, 0) = 1\).

Induction Hypothesis: \(E[T(m-1, n-1)] \leq \frac{m-1}{m-n}\)

Induction step:

\[
E[T(m, n)] = 1 + \frac{n}{m} E[T(m-1, n-1)] \\
\leq 1 + \frac{n}{m} \cdot \frac{m-1}{m-n} \\
< 1 + \frac{n}{m} \cdot \frac{m}{m-n} \\
= \frac{m-n}{m-n} + \frac{n}{m-n} \\
= \frac{m}{m-n} = \frac{1}{1-\alpha} \\
\therefore E[T(m, n)] \leq \frac{1}{1-\alpha}
\]

\[3\] We can show this by writing the expected number of probes as a recurrence on the two variables \(m\) and \(n\):

\[ E[T(m, n)] = 1 + \frac{n}{m} E[T(m-1, n-1)] \]

\[4\] There is a really good writeup of Python’s implementation here