Hashmaps

Introduction to hashmaps

Suppose we are given a string or a character array and asked to find the maximum occurring character. It could be quickly done using arrays. We can simply create an array of size 256, initialize this array to zero, and then, simply traverse the array to increase the count of each character against its ASCII value in the frequency array. In this way, we will be able to figure out the maximum occurring character in the given string.

The above method will work fine for all the 256 characters whose ASCII values are known. But what if we want to store the maximum frequency string out of a given array of strings? It can't be done using a simple frequency array, as the strings do not possess any specific identification value like the ASCII values. For this, we will be using a different data structure called hashmaps.

In hashmaps, the data is stored in the form of keys against which some value is assigned. Keys and values don't need to be of the same data type.

If we consider the above example in which we were trying to find the maximum occurring string, using hashmaps, the individual strings will be regarded as keys, and the value stored against them will be considered as their respective frequency.

For example: The given string array is:

str[] = {"abc", "def", "ab", "abc", "def", "abc"}

Hashmap will look like as follows:

Key (datatype = string)	Value (datatype = int)
“abc”	3
“def”	2
“ab”	1

From here, we can directly check for the frequency of each string and hence figure out the most frequent string among them all.

Note: *One more limitation of arrays is that the indices could only be whole numbers but this limitation does not hold for hashmaps.*

The values are stored corresponding to their respective keys and can be invoked using these keys. To insert, we can do the following:

- hashmap[key] = value
- hashmap.insert(key, value)

The functions that are required for the hashmaps are(using templates):

- **insert(k key, v value):** To insert the value of type v against the key of type k.
- **getValue(k key):** To get the value stored against the key of type k.
- **deleteKey(k key):** To delete the key of type k, and hence the value corresponding to it.

To implement the hashmaps, we can use the following data structures:

1. **Linked Lists:** To perform insertion, deletion, and search operations in the linked list, the time complexity will be O(n) for each as:
    - For insertion, we will first have to check if the key already exists or not, if it exists, then we just have to update the value stored corresponding to that key.

- For search and deletion, we will be traversing the length of the linked list.

2. **BST:** We will be using some kind of a balanced BST so that the height remains of the order O(logN). For using a BST, we will need some sort of comparison between the keys. In the case of strings, we can do the same. Hence, insertion, search, and deletion operations are directly proportional to the height of the BST. Thus the time complexity reduces to O(logN) for each.
3. **Hash table:** Using a hash table, the time complexity of insertion, deletion, and search operations, could be improved to O(1) (same as that of arrays). We will study this in further sections.

Inbuilt Hashmap

In the STL, we have two types of hashmaps:

- **map** (uses BST implementation)
- **unordered_map** (uses hash table implementation)

Note: Both have similar functions and similar ways to use, differing only in the time complexities. The time complexity of each of the operations(insertion, deletion, and searching) in the **map** is O(logN), while in the case of **unordered_map**, they are O(1).

unordered_map:

- Header file: **#include<unordered_map>**
- Syntax to declare:

```
unordered_map<datatype_for_keys, datatype_for_values> name;
```

- Operations performed:
    1. **Insertion:** Suppose, we want to insert the string “abc” with the value 1 in the hashmap named *ourmap*, there are two ways to do so:

- Simply create a pair of both and insert in the map using .**insert** function. Syntax:

```
pair<string, int> p = {"abc", 1};
ourmap.insert(p);
```

- An easier way to insert in a map is to insert like arrays. Syntax:

```
ourmap["abc"] = 1;
```

2. **Searching:** Suppose we want to find the value stored in the hashmap against key *"abc"*, there are two ways to do so:
    - As did in insertion like arrays, the same way we can figure out the value stored against the corresponding key. Syntax:

```
int value = ourmap["abc"];
```

    - Using .at() function. Syntax:

```
int value = ourmap.at("abc");
```

Note: If we try to access a key that is not present in the unordered_map, then there are two different outcomes:

- ➢ If we are accessing the value using .at() function, then we will get an error specifying that we are trying to access the value that is not present in the map.
- ➢ On the other hand, if we access the same using square brackets, then by default, a new key will be created with the default value to be 0 against it. This approach will not give any error.

But what if we want to check if the key is present or not on the map? For that, we will be using the .count() function, which tells if the key is present in the map or not. It returns 0, if not present, and 1, if present

Syntax:

```
int isPresent = ourmap.count("ghi");     // isPresent could only be 0 or 1
```

Note: *We can also check the size of the map by using .size() function, which returns the number of key-value pairs present on the map.*

Syntax:

```
int size_of_map = ourmap.size();
```

3. Deletion: Suppose we want to delete the key *"abc"* from the map, we will be using .erase() function.

Syntax:

```
ourmap.erase("abc");
```

Remove Duplicates

Problem statement: Given an array of integers, we need to remove the duplicate values from that array, and the values should be in the same order as present in the array.

For example: Suppose the given array is arr = {1, 3, 6, 2, 4, 1, 4, 2, 3, 2, 4, 6}, answer should be {1, 3, 6, 2, 4}.

Approach: We will add unique values to the vector and then return it. To check for unique values, start traversing the array, and for each array element, check if the value is already present in the map or not. If not, then we will insert that value in the vector and update the map; otherwise, we will proceed to the next index of the array without making any changes.

Let's look at the code for better understanding.

```
vector<int> removeDuplicates(int* a, int size) {
        vector<int> output;                     // to store the unique elements.
        unordered_map<int, bool> seen;          // unordered map created
        for (int i = 0; i < size; i++) {        // traversing the array
                if (seen.count(a[i]) > 0) {     // using .count() function to check if
                        continue;               // the value has already occurred.
                }
                seen[a[i]] = true;              // If not, then updating the map
                output.push_back(a[i]);         // and inserting that value in the vector
        }
        return output;
}
```

Iterators

To iterate over STL containers, we can use iterators. These are independent of the way the data is stored in the container. It just iterates over the desired and returns all the elements one-by-one. For example, consider the following piece of code:

```
unordered_map<string, int> ourmap;
ourmap["abc"] = 1;
ourmap["abc1"] = 2;
ourmap["abc2"] = 3;
ourmap["abc3"] = 4;
ourmap["abc4"] = 5;
ourmap["abc5"] = 6;
```

Now, we want to iterate the above unordered_map. It can be achieved using iterators.

Steps to use iterators:

1. Declare the iterator of type unordered_map with the same set of data types as that of key-value pairs and point it to the beginning of the map.

   ```
   unordered_map<string, int>::iterator it = ourmap.begin();
   ```

Note: *ourmap.begin() represents the starting position of the map, which could be any key in case of unordered_map as there is no specific order of data inserted in it.*

Although, we will be able to cover entire map elements using iterators irrespective of the position that ourmap.begin() points to.

2. Now, simply put the iterator value in the loop until it reaches the end.

```
while (it != ourmap.end()) {
        cout << "Key : " << it->first << " Value: " << it->second << endl;
        it++;       // increasing the iterator so that it points to next value
}
// Here, it->first points to the key and it->second points to the value
// corresponding to it.
```

Similarly, iterators can also be used on the vectors. Since vectors have a specific order of the elements stored in it, so .begin() will always point to the first position of the vector, i.e., the zeroth index. Refer to the code below for a better explanation.

```
vector<int> v;
v.push_back(1);
v.push_back(2);
v.push_back(3);
v.push_back(4);
v.push_back(5);

 // iterator declared and pointed to the first position
vector<int>::iterator it1 = v.begin();

// Now, using the while() loop to reach every element of the vector
while (it1 != v.end()) {
        cout << *it1 << endl;                   // *it1 prints the value pointed by itertor it1
        it1++;                                  // Incrementing the value of iterator so that
                                                //it points to the further indexes
}
```

Note: Using .find() on map, it returns an iterator to that position. To erase the element from the map, we can directly provide that iterator to the .erase() function. Refer to the code below:

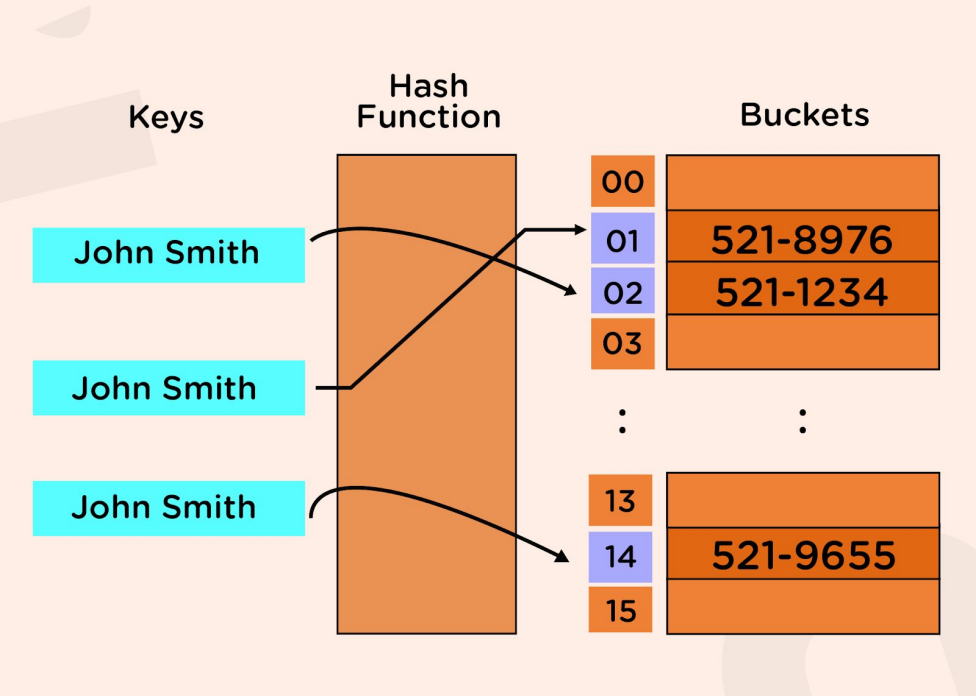
```
unordered_map<string, int>::iterator it2 = ourmap.find("abc");
ourmap.erase(it2);
```

Bucket array and hash function

Now, let's see how to perform insertion, deletion, and search operations using hash tables. Till now, we have seen that arrays are the fastest way to extract data as compared to other data structures as the time complexity of accessing the data in the array is O(1). So we will try to use them in implementing the hashmaps.

Now, we want to store the key-value pairs in an array, named as a **bucket array**. We need an integer corresponding to the key so that we can keep it in the bucket array. To do so, we use a **hash function**. A hash function converts the key into an integer, which acts as the index for storing the key in the array.

For example: Suppose, we want to store some names from the contact list in the hash table, check out the following the image:



Suppose we want to store a string in a hash table, and after passing the string through the hash function, the integer we obtain is equal to 10593, but the bucket array's size is only 20. So, we can't store that string in the array as 10593, as this index does not exist in the array of size 20.

To overcome this problem, we will divide the hashmap into two parts:

- Hash code
- Compression function

The first step to store a value into the bucket array is to convert the key into an integer (this could be any integer irrespective of the size of the bucket array). This part is achieved by using hashcode. For different types of keys, we will be having different kinds of hash codes. Now we will pass this value through the compression function, which will convert that value within the range of our bucket array's size. Now, we can directly store that key against the index obtained after passing through the compression function.

The compression function can be used as (% bucket_size).

One example of a hash code could be: (Example input: "abcd")

"abcd" = ('a' * p^3) + ('b' * p^2) + ('c' * p^1) + ('d' * p^0)
Where p is generally taken as a prime number so that they are well distributed.

But, there is still a possibility that after passing the key through from hash code, when we give the same through the compression function, we can get the same values of indices. For example, let s1 = "ab" and s2 = "cd". Now using the above hash function for p = 2, h1 = 292 and h2 = 298. Let the bucket size be equal to 2. Now, if we pass the hash codes through the compression function, we will get:

Compression_function1 = 292 % 2 = 0

Compression_function2 = 298 % 2 = 0

This means they both lead to the same index 0.

This is known as a **collision**.

Collision Handling

We can handle collisions in two ways:

- Closed hashing (or closed addressing)
- Open addressing

In closed hashing, each entry of the array will be a linked list. This means it should be able to store every value that corresponds to this index. The array position holds the address to the head of the linked list, and we can traverse the linked list by using the head pointer for the same and add the new element at the end of that linked list. This is also known as **separate chaining**.

On the other hand, in open addressing, we will check for the index in the bucket array if it is empty or not. If it is empty, then we will directly insert the key-value pair over that index. If not, then will we find an alternate position for the same. To find the alternate position, we can use the following:

$\mathbf{h_i(a) = hf(a) + f(i)}$

Where hf(a) is the original hash function, and f(i) is the i-th try over the hash function to obtain the final position $h_i(a)$.

To figure out this f(i), following are some of the techniques:

1. **Linear probing:** In this method, we will linearly probe to the next slot until we find the empty index. Here, f(i) = i.

2. **Quadratic probing:** As the name suggests, we will look for alternate i^2 positions ahead of the filled ones, i.e., $f(i) = i^2$.
3. **Double hashing:** According to this method, f(i) = i * H(a), where H(a) is some other hash function.

In practice, we generally prefer to use separate chaining over open addressing, as it is easier to implement and is also more efficient.

Let's now implement the hashmap of our own.

Hashmap implementation - Insert

As discussed earlier, we will be implementing separate chaining. We will be using value as a template and key as a string as we are required to find the hash code for the key. Taking key as a template will make it difficult to convert it using hash code.

Let's look at the code for the same.

```
#include <string>
using namespace std;

template <typename V>
class MapNode {                          // class for linked list.
        public:
                string key;              // to store key of type string
                V value;                 // to store value of type template
                MapNode* next;           // to store the next pointer

                MapNode(string key, V value) {      //  constructor to assign values
                        this->key = key;
                        this->value = value;
                        next = NULL;
                }

                ~MapNode() {             // Destructor to delete the node.
                        delete next;
                }
};

template <typename V>
class ourmap {                   //  for storing the bucket array
```

```
        MapNode<V>** buckets;      // a 2D bucket array to store the head pointers
                                   // of the linked list corresponding to each index.
        int count;                 // to store the size
        int numBuckets;     //  to store number of buckets for compression function

        public:
        ourmap() {      // constructor: to initialize the values
                count = 0;     //
                numBuckets = 5;
                buckets = new MapNode<V>*[numBuckets];  // dynamically allocated
                for (int i = 0; i < numBuckets; i++) {
                        buckets[i] = NULL;   // assigning each head pointer to NULL
                }
        }

        ~ourmap() {   // destructor: to delete the storage used
                for (int i = 0; i < numBuckets; i++) {   // first-of-all delete each linked list
                        delete buckets[i];
                }
                delete [] buckets;    // the delete the total bucket
        }

        int size() {        // to return the size of the map
                return count;
        }

        V getValue(string key) {// for returning the value corresponding to a key
                // will see in the next section
        }

        private:
        int getBucketIndex(string key) {    // to provide the index using hash function
                int hashCode = 0;
                int currentCoeff = 1;
// using "abcd" = ('a' * p^3) + ('b' * p^2) + ('c' * p^1) + ('d' * p^0) as our hash code
                for (int i = key.length() - 1; i >= 0; i--) {
                        hashCode += key[i] * currentCoeff;
                        hashCode = hashCode % numBuckets;
                        currentCoeff *= 37;       // taking p = 37
                        currentCoeff = currentCoeff % numBuckets;
                }

                return hashCode % numBuckets;
        }

        public:
        void insert(string key, V value) {
```

```
    // To get the bucket index, i.e., passing it through the hash function
            int bucketIndex = getBucketIndex(key);
//  to insert the key-value pair in the linked list corresponding to index obtained
            MapNode<V>* head = buckets[bucketIndex];
            while (head != NULL) {
                if (head->key == key) {    // If key already present, then we are
                    head->value = value;  // just updating the value against it
                    return;
                }
                head = head->next;
            }
// otherwise creating a new node and inserting that node before head so making it
// as the head and marking the bucket index to this node as the new head
            head = buckets[bucketIndex];
            MapNode<V>* node = new MapNode<V>(key, value);
            node->next = head;
            buckets[bucketIndex] = node;
            count++;
        }
};
```

Hashmap implementation - Delete and search

Refer to the code below and follow the comments in it.

```
        // to search for the key
        V getValue(string key) {
            int bucketIndex = getBucketIndex(key);  // find the index
            MapNode<V>* head = buckets[bucketIndex];    // head of linked list
            while (head != NULL) {
                if (head->key == key) {     //  if found, returned the value
                    return head->value;
                }
                head = head->next;
            }
            return 0;   //  if not found, returning 0 as default value.
        }
       // to delete the key-value pair
        V remove(string key) {
            int bucketIndex = getBucketIndex(key); // find the index
            MapNode<V>* head = buckets[bucketIndex]; // head node
            MapNode<V>* prev = NULL; // previous pointer
            while (head != NULL) {
```

```
                    if (head->key == key) {
                        if (prev == NULL) {
                            buckets[bucketIndex] = head->next;
                        } else {
                            prev->next = head->next;   // connecting previous
                                                       // to the head's next pointer
                        }
                        V value = head->value;
                        head->next = NULL; // before calling delete over it, in
// order to avoid complete linked list deletion, we have to assign it's next to NULL
                        delete head;
                        count--;       // reducing the size
                        return value;// return the value stored at deleted node
                    }
                    prev = head;
                    head = head->next;
                }
                return 0;   // means that value not found
            }
```

Time Complexity and Load Factor

Let's define specific terms before moving forward:

1. n = Number of entries in our map.
2. l = length of the word (in case of strings)
3. b = number of buckets. On average, each box contains (n/b) entries. This is known as **load factor** means b boxes contain n entries. We also need to ensure that the load factor is always less than 0.7, i.e.,

   (n/b) < 0.7, this will ensure that each bucket does not contain too many entries in it.
4. To make sure that load factor < 0.7, we can't reduce the number of entries, but we can increase the bucket size comparatively to maintain the ratio. This process is known as **Rehashing**.

This ensures that time complexity is on an average O(1) for insertion, deletion, and search operations each.

Rehashing

Now, we will try to implement the rehashing in our map. After inserting each element into the map, we will check the load factor. If the load factor's value is greater than 0.7, then we will rehash.

Refer to the code below for better understanding.

```
void rehash() {
        MapNode<V>** temp = buckets;   // To store the old bucket
        buckets = new MapNode<V>*[2 * numBuckets];  // doubling the size
        for (int i = 0; i < 2*numBuckets; i++) {
                buckets[i] = NULL;      // initialising each head pointer to NULL
        }
        int oldBucketCount = numBuckets;
        numBuckets *= 2;              // updating new size
        count = 0;
        for (int i = 0; i < oldBucketCount; i++) {
                MapNode<V>* head = temp[i];
                while (head != NULL) {
                        string key = head->key;
                        V value = head->value;
                        insert(key, value);   // inserting each value of old bucket
                                              // into the new one
                        head = head->next;
                }
        }
        // deleting the old bucket
        for (int i = 0; i < oldBucketCount; i++) {
                MapNode<V>* head = temp[i];
                delete head;
        }
        delete [] temp;
}
void insert(string key, V value) {
        int bucketIndex = getBucketIndex(key);
        MapNode<V>* head = buckets[bucketIndex];
        while (head != NULL) {
                if (head->key == key) {
                        head->value = value;
```

```
                return;
            }
            head = head->next;
        }
        head = buckets[bucketIndex];
        MapNode<V>* node = new MapNode<V>(key, value);
        node->next = head;
        buckets[bucketIndex] = node;
        count++;
        // Now we will check the load factor after insertion.
        double loadFactor = (1.0 * count)/numBuckets;
        if (loadFactor > 0.7) {
            rehash();      // We will rehash.
        }
    }
```

Note: While solving the problems, we will be using the in-built hashmap only.

Practice problems:

- https://www.codechef.com/problems/STEM
- https://codeforces.com/problemset/problem/525/A
- https://www.spoj.com/problems/ADACLEAN/