crawl the web

Index & Freq's

Document IDs

Search engine servers

Create word index

user query

Rank results

THE WEB

i am sam
i am sam
sam i am
that sam i am
i do not like
that sam i am
do you like green eggs and ham
i do not like them
sam i am
i do not like
green eggs and ham
would you like them
here or there
i would not like them
here or there
i would not like them
anywhere
i do not like
green eggs and ham
i do not like them
sam i am
would you like them
in a house
i do not like them
with a mouse
i do not like them
here or there
i do not like them
Anywhere
i do not like green eggs and ham
i do not like them
sam i am
would you eat them
in a box
would you eat them
with a fox
not in a box
not with a fox
not in a house
not with a mouse
i would not eat them here or there
i would not eat them anywhere
i would not eat green eggs and ham
i do not like them
sam i am
would you could you
in a car
eat them eat them
here they are
i would not
could not
in a car

THE WEB

a :59
am :16
and :25
anywhere :8
are :2
be :4
boat :3
box :7
car :7
could :14
dark :7
do :37
doe :25
eats :11
fox :7
goat :4

... try :4
will :21
with :19
would :26
you :34
The English language has half-a-million terms. Any given text, however, has only a few thousand words.

Keeping an array of 500K words “just in case” is not good.

What data structure should we use?

Hashing is the idea that order is determined by some function of the value of the element to be stored.

Like throwing darts on a board.

---

**Hashing the keys**

- To search for an entry in the table:
  - Compute the hash function on the entry’s key, then
  - Use the value of the hash function as an index into the Table.

- Can two or more keys may collide on the same index?
  - Then employ some method of collision resolution.

---

**Pros and Cons**

**Pros**
- Searching \( O(1) \)
- Inserting \( O(1) \)
- Deleting \( O(1) \)

**Cons**
- You cannot keep adding new elements for ever!
  - Table size is fixed (like an array)
  - Needs expansion capabilities ( \( O(?) \) )

- Would be nice to have a perfect hashing function but many items may end up on same location
- Collisions need resolution policy
Load Factor: When $M$ is large enough?

- $N/M = \text{load factor}$ of a hashtable
- Number of entries $N$ divided by the table capacity $M$.

Heuristics:
- If you know $N$, make $M = 1.5 \times N$
- If you do not know $N$, provide for dynamic resizing:
  - Create larger Hash Table
  - Insert old elements into new

Hash Functions: Division

- Good:
  $$h(\text{hashCode}) = \text{hashCode} \ % \ M$$
  $M$: prime

- Better:
  $$h(\text{hashCode}) = ((a \times \text{hashCode} + b) \ % \ p) \ % \ M$$
  $p$: prime $\gg N$
  $a, b$: positive integers

- The key is multiplied by itself and then “extract” some digits from the middle of the result
- For example, if our key is 4321
  - Multiply the key by itself yielding 18671041
  - Extract the needed three digits
- It is critical that the same three digits be extracted each time
- We may also extract bits and then reconstruct an index from the bits

- The `java.lang.Object` class defines a method called `hashCode()` that returns an integer based on the memory location of the object
  - This is generally not very useful

- Classes derived from `Object` often override the inherited definition of `hashCode` to provide their own version
- For example, `String` and `Integer` define their own `hashCode` methods
  - These more specific `hashCode` functions are more effective
If we are able to develop a **perfect hashing function**, then we do not need to be concerned about collisions or table size.

However, often we do not know the size of the dataset and are not able to develop a perfect hashing function.

In these cases, we must decide how to resolve collisions.

**Separate Chaining**

Look for another open position in the table other than the one to which the element is hashed.

- **Open addressing** ($M >> N$):

  - How are *collisions* are resolved with this technique?
Resolve Open Addressing Collisions with Linear Probing

- When the index hashed to is occupied by a stranger, probe the next position.
- If that position is empty, we insert the entry, otherwise, we probe the next position and repeat.

There is a problem though: Clustering

- As the table begins to fill up, more and more entries must be examined before the desired entry is found.
- Insertion of one entry may greatly increase the search time for others.

For example, consider H, S, H, I, ...

The Java Hashtable<K,V> Class

- Located in java.util
- Methods
  - int size()
    // returns number of keys in table
  - V get(Object key)
    // returns value to which specified key is mapped in table
  - V put(K key, V value)
    // maps key to specified value in table
  - boolean containsKey(Object key)
    // tests if the specified Object is a key in hash table
  - V remove(Object key)
    // removes key and corresponding value from table
  - ...

Basic Word Frequency code

```java
import java.util.*;
import java.io.*;

Hashtable<String, Integer> table = new Hashtable<String, Integer>();

Scanner reader = new Scanner(new File(filename));
while (reader.hasNext()) {
    String word = reader.next();
    if (table.containsKey(word)) {
        int previousCount = table.get(word);
        table.put(word, previousCount+1);
    } else table.put(word, 1);
    totalWords++;
}
reader.close();
```
import java.io.*;

// args[0] is the name of a directory
dir= new File(args[0] + "/");
// dis points to the directory's contents
File[] files= dir.listFiles();
System.out.println(files.length + " files");
for(File f:files)
    if(!f.isHidden())
        process(f);