Lecture 10, 17 September 2024

Arrays

- Contiguous block of memory
- Typically size is declared in advance, all values are uniform
- `a[0]` points to first memory location in the allocated block
- Locate `a[i]` in memory using index arithmetic
  - Skip `i` blocks of memory, each block's size determined by value stored in array
- **Random access** -- accessing the value at `a[i]` does not depend on `i`
- Useful for procedures like sorting, where we need to swap out of order values `a[i]` and `a[j]`
  - `a[i], a[j] = a[j], a[i]`
  - Cost of such a swap is constant, independent of where the elements to be swapped are in the array
- Inserting or deleting a value is expensive
- Need to shift elements right or left, respectively, depending on the location of the modification

Lists

- Each location is a *cell*, consisiting of a value and a link to the next cell
  - Think of a list as a train, made up of a linked sequence of cells
- The name of the list `l` gives us access to `l[0]`, the first cell
- To reach cell `l[i]`, we must traverse the links from `l[0]` to `l[1]` to `l[2]` ... to `l[i-1]`] to `l[i]`
  - Takes time proportional to `i`
- Cost of swapping `l[i]` and `l[j]` varies, depending on values `i` and `j`
- On the other hand, if we are already at `l[i]` modifying the list is easy
  - *Insert* - create a new cell and reroute the links
  - *Delete* - bypass the deleted cell by rerouting the links
- Each insert/delete requires a fixed amount of local "plumbing", independent of where in the list it is performed

Dictionaries

- Values are stored in a fixed block of size m
- Keys are mapped to $\{0, 1, \ldots, m-1\}$
- Hash function $h : K \rightarrow S$ maps a *large* set of keys K to a *small* range S
- Simple hash function: interpret $k \in K$ as a bit sequence representing a number n_k in binary, and compute $n_k \bmod m$, where $|S| = m$
- Mismatch in sizes means that there will be *collisions* -- $k_1 \neq k_2$, but $h(k_1) = h(k_2)$
- A good hash function maps keys "randomly" to minimize collisions
- Hash can be used as a *signature* of authenticity
  - Modifying k slightly will drastically alter $h(k)$
  - No easy way to reverse engineer a k' to map to a given $h(k)$
  - Use to check that large files have not been tampered with in transit, either due to network errors or malicious intervention
- Dictionary uses a hash function to map key values to storage locations
- Lookup requires computing $h(k)$ which takes roughly the same time for any k
  - Compare with computing the offset `a[i]` for any index `i` in an array
- Collisions are inevitable, different mechanisms to manage this, which we will not discuss now
- Effectively, a dictionary combines flexibility with random access

Lists in Python

- Flexible size, allow inserting/deleting elements in between
- However, implementation is an array, rather than a list
- Initially allocate a block of storage to the list
- When storage runs out, double the allocation
- `l.append(x)` is efficient, moves the right end of the list one position forward within the array
- `l.insert(0,x)` inserts a value at the start, expensive because it requires shifting all the elements by 1
- We will run experiments to validate these claims

Measuring execution time

- Call `time.perf_counter()`
- Actual return value is meaningless, but difference between two calls measures time in seconds

In [1]:
```
import time
```

- 10^7 appends to an empty Python list

In [2]:
```
start = time.perf_counter()
l = []
for i in range(10000000):
    l.append(i)
elapsed = time.perf_counter() - start
print(elapsed)
```

```
0.6168232130003162
```

- Doubling the work approximately doubles the time, linear

In [3]:
```
start = time.perf_counter()
l = []
for i in range(20000000):
    l.append(i)
elapsed = time.perf_counter() - start
print(elapsed)
```

```
1.4875409450032748
```

- 10^5 inserts at the beginning of a Python list

In [4]:
```
start = time.perf_counter()
l = []
for i in range(100000):
    l.insert(0,i)
elapsed = time.perf_counter() - start
print(elapsed)
```

```
1.4217620529816486
```

- Doubling and tripling the work multiplies the time by 4 and 9, respectively, so quadratic

In [5]:
```
start = time.perf_counter()
l = []
for i in range(200000):
    l.insert(0,i)
elapsed = time.perf_counter() - start
print(elapsed)
```

```
5.139741463994142
```

In [6]:
```
start = time.perf_counter()
l = []
for i in range(300000):
    l.insert(0,i)
elapsed = time.perf_counter() - start
print(elapsed)
```

```
11.383465348975733
```

- Creating 10^7 entries in an empty dictionary

In [7]:
```
start = time.perf_counter()
d = {}
for i in range(10000000,0,-1):
    d[i] = i
elapsed = time.perf_counter() - start
print(elapsed)
```

```
1.0553472369792871
```

- Doubling the operations, doubles the time, so linear
- Dictionaries are effectively random access

In [9]:
```
start = time.perf_counter()
d = {}
for i in range(20000000,0,-1):
    d[i] = i
elapsed = time.perf_counter() - start
print(elapsed)
```

```
2.602921769954264
```

- Insert keys in random order
- Use the library function `random.shuffle(l)` to permute the elements of `l`

In [10]:
```
import random
lhundred = list(range(100))
random.shuffle(lhundred)
print(lhundred)
```

```
[11, 37, 34, 12, 46, 41, 96, 6, 16, 13, 97, 76, 26, 47, 27, 28, 99, 62, 90, 0, 51, 81, 79, 35, 5, 48, 84, 53, 6
5, 85, 25, 82, 52, 57, 78, 23, 98, 54, 20, 63, 91, 19, 38, 75, 80, 7, 3, 64, 74, 2, 31, 72, 93, 39, 56, 71, 14,
30, 77, 40, 55, 43, 68, 69, 61, 29, 33, 9, 44, 36, 15, 32, 18, 94, 21, 24, 60, 49, 70, 22, 45, 92, 89, 17, 58, 1
0, 73, 66, 50, 59, 87, 4, 8, 1, 95, 88, 83, 67, 42, 86]
```

- Insert 10^6 keys in random order
- Note that we start the counter *after* we shuffle the list of keys, so we count only the time required to populate the dictionary

In [11]:
```
import random
keylist = list(range(1000000,0,-1))
rndkeylist = keylist[:]
random.shuffle(rndkeylist)

d = {}
start = time.perf_counter()
for i in keylist:
    d[i] = i
elapsed = time.perf_counter() - start
print("Sequential keys:", elapsed)

d = {}
start = time.perf_counter()
for i in rndkeylist:
    d[i] = i
elapsed = time.perf_counter() - start
print("Shuffled keys:", elapsed)
```

```
Sequential keys: 0.09673382097389549
Shuffled keys: 0.39740611804882064
```

- Double the number of keys to 2×10^6

In [12]:
```
import random
keylist = list(range(2000000,0,-1))
rndkeylist = keylist[:]
random.shuffle(rndkeylist)

d = {}
start = time.perf_counter()
for i in keylist:
    d[i] = i
elapsed = time.perf_counter() - start
print("Sequential keys:", elapsed)

d = {}
start = time.perf_counter()
for i in rndkeylist:
    d[i] = i
elapsed = time.perf_counter() - start
print("Shuffled keys:", elapsed)
```

```
Sequential keys: 0.21819286403479055
Shuffled keys: 0.6841557070147246
```

- Triple the number of keys to 3×10^6

In [14]:
```
import random
keylist = list(range(3000000,0,-1))
rndkeylist = keylist[:]
random.shuffle(rndkeylist)

d = {}
start = time.perf_counter()
for i in keylist:
    d[i] = i
elapsed = time.perf_counter() - start
print("Sequential keys:", elapsed)

d = {}
start = time.perf_counter()
for i in rndkeylist:
    d[i] = i
elapsed = time.perf_counter() - start
print("Shuffled keys:", elapsed)
```

```
Sequential keys: 0.35756950796348974
Shuffled keys: 1.1829602149664424
```

- Using shuffled keys is about 3 times slower than inserting keys in sequence
- However, even after shuffling, the time taken grows approximately linearly

Implementing a "real" list using dictionaries

In [15]:
```
def createlist():  # Equivalent of l = [] is l = createlist()
  return({})

def listappend(l,x):
  if l == {}:
    l["value"] = x
    l["next"] = {}
    return

  node = l
  while node["next"] != {}:
    node = node["next"]

  node["next"]["value"] = x
  node["next"]["next"] = {}
  return

def listinsert(l,x):
  if l == {}:
    l["value"] = x
    l["next"] = {}
    return

  newnode = {}
  newnode["value"] = l["value"]
  newnode["next"] = l["next"]
  l["value"] = x
  l["next"] = newnode
  return


def printlist(l):
  print("{",end="")

  if l == {}:
    print("}")
    return
  node = l

  print(node["value"],end="")
  while node["next"] != {}:
    node = node["next"]
    print(",",node["value"],end="")
  print("}")
  return
```

- Display a small list as nested dictionaries

In [16]:
```
start = time.perf_counter()
l = createlist()
for i in range(10):
    listappend(l,i)
elapsed = time.perf_counter() - start
print(elapsed)
print(l)
```

```
0.013133806001860648
{'value': 0, 'next': {'value': 1, 'next': {'value': 2, 'next': {'value': 3, 'next': {'value': 4, 'next': {'valu
e': 5, 'next': {'value': 6, 'next': {'value': 7, 'next': {'value': 8, 'next': {'value': 9, 'next': {}}}}}}}}}}}
```

- Insert 10^7 elements at the beginning in this implementation of a list

In [21]:
```
start = time.perf_counter()
l = createlist()
for i in range(1000000):
    listinsert(l,i)
elapsed = time.perf_counter() - start
print(elapsed)
```

```
1.2849651229917072
```

- Doubling the work doubles the time, so linear

In [22]:
```
start = time.perf_counter()
l = createlist()
```

```
for i in range(2000000):
    listinsert(l,i)
elapsed = time.perf_counter() - start
print(elapsed)
```

```
3.5748096029856242
```

- Append 10^4 elements in this implementation of a list

In [23]:
```
start = time.perf_counter()
l = createlist()
for i in range(10000):
    listappend(l,i)
elapsed = time.perf_counter() - start
print(elapsed)
```

```
2.831144590047188
```

- Halving the work takes 1/4 of the time, so quadratic

In [24]:
```
start = time.perf_counter()
l = createlist()
for i in range(5000):
    listappend(l,i)
elapsed = time.perf_counter() - start
print(elapsed)
```

```
0.6491393339820206
```

Defining our own data structures

- We have implemented a "linked" list using dictionaries
- The fundamental functions like `listappend`, `listinsert`, `listdelete` modify the underlying list
- Instead of `mylist = {}`, we wrote `mylist = createlist()`
- To check empty list, use a function `isempty()` rather than `mylist == {}`
- Can we clearly separate the **interface** from the **implementation**
- Define the data structure in a more "modular" way