Lecture 22: Dictionaries

Counting Words

Suppose I had Melville's Moby Dick stored in a text file called `moby.txt`. What if I was interested in finding the most frequent word used in the text? It's easy enough to hold all of Moby Dick in memory, so I can read the entire text into a string, split the words using whitespace as my delimiter and produce a list of words, which we call tokens.

```
def file_to_tokens(filename):
    with open(filename) as fin:
        return fin.read().split()
```

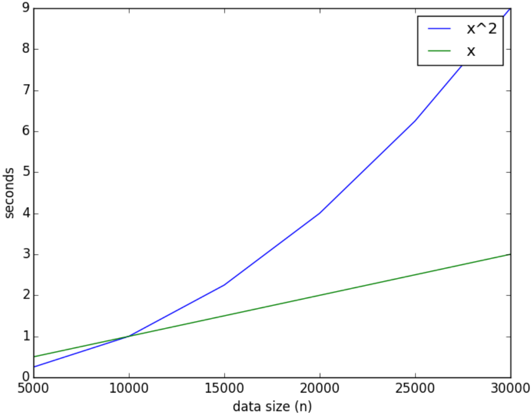
Now I'm left with the task of counting the how many times each token occurs in the list. I could use list operations to first find the set of unique tokens, and then count the occurrences of those tokens.

```
def wc_list(tokens):
    uniq = []
    for token in tokens:
        if token not in uniq:
            uniq.append(token)
    return [(t, tokens.count(t)) for t in uniq]
```

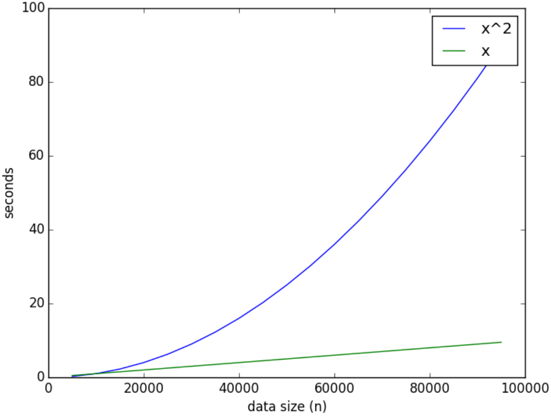
Profiling our Code

```
>>>import cProfile
>>> cProfile.run('[uniq[:5000].count(t) for t in uniq[:5000]]')
         5004 function calls in 0.528 seconds

   Ordered by: standard name

   ncalls  tottime  percall  cumtime  percall filename:lineno(function)
        1    0.147    0.147    0.528    0.528 <string>:1(<listcomp>)
        1    0.000    0.000    0.528    0.528 <string>:1(<module>)
        1    0.000    0.000    0.528    0.528 {built-in method exec}
     5000    0.382    0.000    0.382    0.000 {method 'count' }
        1    0.000    0.000    0.000    0.000 {method 'disable'  }
```

Quadratic versus Linear



Quadratic versus Linear



Counting Words

```
counts = {}
for token in tokens:
    if token in counts:
        counts[token] += 1
    else:
        counts[token] = 1
return counts.items()
```

Practice: Building a Word Index

Suppose we wanted to create an index of the positions of each token in the original text. Write a function called token_locations that, when given a list of tokens, returns a dictionary where each key is a token and each value is list of indices where that token appears.

```
>>> l = "brent sucks big rocks through a big straw".split()
>>> print(token_locations(l))
{’big’: [2, 6], ’straw’: [7], ’brent’: [0], ’a’: [5],
 ’through’: [4], ’sucks’: [1], ’rocks’: [3]}
```