

Mar 14, 12 22:22

finger_print2.py

Page 1/1

```

'''
Created on March 21, 2011

@author: ola
'''
5  import time, urllib

def slow_fingerp(datasource):
    stats = []
10     for word in datasource.read().split():
        word = word.lower()
        found = False
        for pair in stats:
            if pair[0] == word:
15                 pair[1] += 1
                found = True
                break
        if not found:
            stats.append([word,1])

20     return stats

def fast_fingerp(datasource):
    stats = {}
25     for word in datasource.read().split():
        word = word.lower()
        if word in stats:
            stats[word] += 1
        else:
30             stats[word] = 1

    return stats

def max_list(data):
35     data is a list of (x,y) tuples where y is a count and x is a string
    '''
    return sorted([(elt[1],elt[0]) for elt in data])[-1]

40 def max_dict(data):
    data is a dictionary where strings are keys and value is an int
    '''
    return sorted([(y,x) for (x,y) in data.items()])[-1]

45 def benchmark(name):
    funcs = [slow_fingerp, fast_fingerp]
    for callit in funcs:
        if name.startswith("http"):
50             source = urllib.urlopen(name)
        else:
            source = open(name)
        start = time.time()
        data = callit(source)
        end = time.time()
55         print str(callit)[10:str(callit).index(' at')], (end-start)
        if (type(data) == type({})):
            print "most freq", max_dict(data)
        else:
60             print "most freq", max_list(data)

if __name__ == "__main__":
    benchmark('http://www.cs.duke.edu/cseds/data/melville.txt')
```