Programing Exercise 7

Searching the Web

Background

The Internet is a complex client-server system with millions of computers connected via IP (Internet Protocol). The World Web is an abstract layer on top of the Internet where computers communicate via HTTP (Hyper-Text Transfer Protocol). Web pages are text files requested and received through HTTP commands. All computers connected to the web have a URL (Uniform Resource Locator) that allows requests from browsers to find the destination server.

In this exercise you will write a simplistic web search engine that reads web pages and "ranks" them by how many links to them exist from other web pages. You will "crawl" the web starting with a given URL and follow its links to some level of search depth and report all URLs reached in rank order (explained below).

Requirements

Write a class named **HayStack** whose constructor receives a URL and an optional search depth (defaulted to 3 levels). The constructor crawls starting from the given web page, finding and following all embedded webpage links until it reaches the maximum search depth and computes the following data items:

1. An *index* that maps every word encountered on each crawled page to a list of URLs of all the pages that contain that word. Only keep words consisting of runs of alphabetic characters and apostrophes. Convert all words to lower case. Only consider text that is found outside of HTML tags.

2. A *graph* that maps every URL encountered to a list of the web pages it links to directly. (This will be used subsequently for ranking web pages by "popularity".)

Make sure you do not crawl a page twice! That could lead to an infinite loop.

Include also the following method in your **HayStack** class:

```
def compute_ranks(self,graph):
     d = 0.85      # probability that surfer will bail
     numloops = 10

     ranks = {}
     npages = len(graph)
     for page in graph:
         ranks[page] = 1.0 / npages

     for i in range(0, numloops):
         newranks = {}
         for page in graph:
             newrank = (1 - d) / npages
             for url in graph:
```

```
                    if page in graph[url]:  # this url links to page
                        newrank += d*ranks[url]/len(graph[url])
                newranks[page] = newrank
            ranks = newranks
        self.ranks = ranks
```

This function takes the graph mentioned above and "ranks" the web pages in **graph**, placing the result in a dictionary, self.**ranks**, where the key is the url and the value is the page's rank. This is a simplified version of Google's page ranking algorithm that made that search engine so famous decades ago.

Finally, write a **lookup** method that takes a word as a search key and outputs the webpages that contain that word in *rank order*, highest to lowest. See the sample output below.

Implementation Notes

You will test your code on a handful of web pages found at *freshsources.com*. Here is a test driver"

```
if __name__ == '__main__':
    engine = HayStack('http://freshsources.com/page1.html',4)
    for w in ['pages','links','you','have','I']:
        print(w)
        pprint.pprint(engine.lookup(w))
    print()
    print('index:')
    pprint.pprint(engine.index)
    print()
    print('graph:')
    pprint.pprint(engine.graph)
    print()
    print('ranks:')
    pprint.pprint(engine.ranks)
```

Note that you will need to import Python's **pprint** Module ("pretty print"). Here is the expected output:

```
pages
['http://freshsources.com/page5.html',
 'http://freshsources.com/page3.html',
 'http://freshsources.com/page1.html',
 'http://freshsources.com/page2.html',
 'http://freshsources.com/page4.html']
links
['http://freshsources.com/page1.html', 'http://freshsources.com/page2.html']
you
['http://freshsources.com/page5.html', 'http://freshsources.com/page2.html']
have
['http://freshsources.com/page2.html', 'http://freshsources.com/page4.html']
I
['http://freshsources.com/page5.html',
 'http://freshsources.com/page3.html',
 'http://freshsources.com/page2.html',
 'http://freshsources.com/page4.html']

index:
{'a': {'http://freshsources.com/page1.html',
```

```
      'http://freshsources.com/page2.html',
      'http://freshsources.com/page3.html',
      'http://freshsources.com/page4.html',
      'http://freshsources.com/page5.html'},
'again': {'http://freshsources.com/page2.html'},
'all': {'http://freshsources.com/page3.html'},
'also': {'http://freshsources.com/page1.html'},
'an': {'http://freshsources.com/page1.html'},
'and': {'http://freshsources.com/page1.html',
        'http://freshsources.com/page3.html',
        'http://freshsources.com/page4.html',
        'http://freshsources.com/page5.html'},
'another': {'http://freshsources.com/page4.html',
            'http://freshsources.com/page5.html'},
'anything': {'http://freshsources.com/page4.html'},
'are': {'http://freshsources.com/page1.html'},
'around': {'http://freshsources.com/page1.html'},
'as': {'http://freshsources.com/page1.html'},
'assignment': {'http://freshsources.com/page1.html',
               'http://freshsources.com/page2.html',
               'http://freshsources.com/page3.html',
               'http://freshsources.com/page4.html',
               'http://freshsources.com/page5.html'},
'at': {'http://freshsources.com/page1.html'},
'back': {'http://freshsources.com/page3.html',
         'http://freshsources.com/page4.html',
         'http://freshsources.com/page5.html'},
'be': {'http://freshsources.com/page1.html'},
'before': {'http://freshsources.com/page3.html'},
'circular': {'http://freshsources.com/page2.html'},
'class': {'http://freshsources.com/page1.html',
          'http://freshsources.com/page3.html'},
'convenient': {'http://freshsources.com/page3.html'},
'crawl': {'http://freshsources.com/page2.html'},
'cs': {'http://freshsources.com/page1.html',
       'http://freshsources.com/page2.html',
       'http://freshsources.com/page3.html',
       'http://freshsources.com/page4.html',
       'http://freshsources.com/page5.html'},
'discussed': {'http://freshsources.com/page1.html'},
"don't": {'http://freshsources.com/page2.html',
          'http://freshsources.com/page4.html'},
'during': {'http://freshsources.com/page2.html'},
'each': {'http://freshsources.com/page1.html'},
'end': {'http://freshsources.com/page1.html'},
'expressions': {'http://freshsources.com/page1.html'},
'field': {'http://freshsources.com/page1.html'},
'figured': {'http://freshsources.com/page2.html'},
'find': {'http://freshsources.com/page1.html'},
'first': {'http://freshsources.com/page1.html',
          'http://freshsources.com/page3.html'},
'five': {'http://freshsources.com/page1.html'},
'for': {'http://freshsources.com/page1.html',
        'http://freshsources.com/page2.html',
        'http://freshsources.com/page3.html',
        'http://freshsources.com/page4.html',
        'http://freshsources.com/page5.html'},
'formed': {'http://freshsources.com/page3.html'},
'found': {'http://freshsources.com/page3.html'},
'fun': {'http://freshsources.com/page5.html'},
'functions': {'http://freshsources.com/page3.html'},
'get': {'http://freshsources.com/page3.html'},
'goal': {'http://freshsources.com/page1.html'},
```

```
'gt': {'http://freshsources.com/page1.html'},
'has': {'http://freshsources.com/page1.html'},
'have': {'http://freshsources.com/page2.html',
         'http://freshsources.com/page4.html'},
'helpful': {'http://freshsources.com/page1.html'},
'here': {'http://freshsources.com/page1.html',
         'http://freshsources.com/page2.html',
         'http://freshsources.com/page3.html',
         'http://freshsources.com/page4.html',
         'http://freshsources.com/page5.html'},
'hope': {'http://freshsources.com/page2.html',
         'http://freshsources.com/page5.html'},
'href': {'http://freshsources.com/page1.html'},
'i': {'http://freshsources.com/page2.html',
      'http://freshsources.com/page3.html',
      'http://freshsources.com/page4.html',
      'http://freshsources.com/page5.html'},
'in': {'http://freshsources.com/page1.html'},
'into': {'http://freshsources.com/page3.html'},
'is': {'http://freshsources.com/page1.html',
       'http://freshsources.com/page2.html',
       'http://freshsources.com/page3.html',
       'http://freshsources.com/page4.html',
       'http://freshsources.com/page5.html'},
'it': {'http://freshsources.com/page3.html'},
'just': {'http://freshsources.com/page3.html'},
'keep': {'http://freshsources.com/page1.html',
         'http://freshsources.com/page2.html'},
'link': {'http://freshsources.com/page1.html',
         'http://freshsources.com/page2.html',
         'http://freshsources.com/page3.html',
         'http://freshsources.com/page4.html',
         'http://freshsources.com/page5.html'},
'links': {'http://freshsources.com/page1.html',
          'http://freshsources.com/page2.html'},
'lt': {'http://freshsources.com/page1.html'},
'may': {'http://freshsources.com/page1.html',
        'http://freshsources.com/page3.html'},
'means': {'http://freshsources.com/page1.html'},
'mileage': {'http://freshsources.com/page3.html'},
'mind': {'http://freshsources.com/page1.html'},
'minutes': {'http://freshsources.com/page3.html'},
'most': {'http://freshsources.com/page1.html'},
'must': {'http://freshsources.com/page2.html'},
'new': {'http://freshsources.com/page4.html'},
'not': {'http://freshsources.com/page1.html'},
'of': {'http://freshsources.com/page1.html',
       'http://freshsources.com/page2.html',
       'http://freshsources.com/page3.html',
       'http://freshsources.com/page4.html',
       'http://freshsources.com/page5.html'},
'on': {'http://freshsources.com/page4.html'},
'once': {'http://freshsources.com/page3.html'},
'other': {'http://freshsources.com/page1.html'},
'out': {'http://freshsources.com/page2.html'},
'page': {'http://freshsources.com/page1.html',
         'http://freshsources.com/page2.html',
         'http://freshsources.com/page3.html',
         'http://freshsources.com/page4.html',
         'http://freshsources.com/page5.html'},
'pages': {'http://freshsources.com/page1.html',
          'http://freshsources.com/page2.html',
          'http://freshsources.com/page3.html',
```

```
          'http://freshsources.com/page4.html',
          'http://freshsources.com/page5.html'},
 'phase': {'http://freshsources.com/page2.html'},
 'popular': {'http://freshsources.com/page1.html'},
 'putting': {'http://freshsources.com/page3.html'},
 'py': {'http://freshsources.com/page1.html'},
 'python': {'http://freshsources.com/page1.html',
            'http://freshsources.com/page2.html',
            'http://freshsources.com/page3.html',
            'http://freshsources.com/page4.html',
            'http://freshsources.com/page5.html'},
 're': {'http://freshsources.com/page1.html'},
 'regular': {'http://freshsources.com/page1.html'},
 'say': {'http://freshsources.com/page4.html'},
 'search': {'http://freshsources.com/page2.html'},
 'searched': {'http://freshsources.com/page2.html'},
 'set': {'http://freshsources.com/page1.html',
         'http://freshsources.com/page2.html',
         'http://freshsources.com/page3.html',
         'http://freshsources.com/page4.html',
         'http://freshsources.com/page5.html'},
 'so': {'http://freshsources.com/page2.html'},
 'spaces': {'http://freshsources.com/page1.html'},
 'such': {'http://freshsources.com/page1.html'},
 'tag': {'http://freshsources.com/page1.html'},
 'test': {'http://freshsources.com/page1.html',
          'http://freshsources.com/page2.html',
          'http://freshsources.com/page3.html',
          'http://freshsources.com/page4.html',
          'http://freshsources.com/page5.html'},
 'than': {'http://freshsources.com/page1.html'},
 'that': {'http://freshsources.com/page1.html',
          'http://freshsources.com/page2.html'},
 'the': {'http://freshsources.com/page1.html',
         'http://freshsources.com/page2.html',
         'http://freshsources.com/page3.html',
         'http://freshsources.com/page4.html',
         'http://freshsources.com/page5.html'},
 'them': {'http://freshsources.com/page2.html',
          'http://freshsources.com/page3.html'},
 'there': {'http://freshsources.com/page1.html'},
 'this': {'http://freshsources.com/page1.html',
          'http://freshsources.com/page2.html',
          'http://freshsources.com/page3.html',
          'http://freshsources.com/page4.html',
          'http://freshsources.com/page5.html'},
 'to': {'http://freshsources.com/page1.html',
        'http://freshsources.com/page2.html',
        'http://freshsources.com/page3.html',
        'http://freshsources.com/page4.html',
        'http://freshsources.com/page5.html'},
 'took': {'http://freshsources.com/page3.html'},
 'track': {'http://freshsources.com/page2.html'},
 'try': {'http://freshsources.com/page2.html'},
 'vary': {'http://freshsources.com/page3.html'},
 'very': {'http://freshsources.com/page1.html'},
 'want': {'http://freshsources.com/page2.html'},
 'was': {'http://freshsources.com/page5.html'},
 'were': {'http://freshsources.com/page3.html'},
 'whatever': {'http://freshsources.com/page1.html'},
 'which': {'http://freshsources.com/page2.html'},
 'words': {'http://freshsources.com/page1.html'},
 'working': {'http://freshsources.com/page3.html'},
```

```
 'write': {'http://freshsources.com/page3.html'},
 'you': {'http://freshsources.com/page2.html',
         'http://freshsources.com/page5.html'},
 'your': {'http://freshsources.com/page3.html'}}

graph:
{'http://freshsources.com/page1.html': {'http://freshsources.com/page2.html',
                                        'http://freshsources.com/page5.html'},
 'http://freshsources.com/page2.html': {'http://freshsources.com/page3.html',
                                        'http://freshsources.com/page5.html'},
 'http://freshsources.com/page3.html': {'http://freshsources.com/page1.html',
                                        'http://freshsources.com/page4.html',
                                        'http://freshsources.com/page5.html'},
 'http://freshsources.com/page4.html': {'http://freshsources.com/page2.html',
                                        'http://freshsources.com/page5.html'},
 'http://freshsources.com/page5.html': {'http://freshsources.com/page1.html',
                                        'http://freshsources.com/page3.html',
                                        'http://freshsources.com/page5.html'}}

ranks:
{'http://freshsources.com/page1.html': 0.19225481365666597,
 'http://freshsources.com/page2.html': 0.14855063327049206,
 'http://freshsources.com/page3.html': 0.19897220412899821,
 'http://freshsources.com/page4.html': 0.0863497557721767,
 'http://freshsources.com/page5.html': 0.37387259317166704}
```

You only have to write about 30 lines of code. Redirect this output to a text file and submit it along with your source code.

Do ***not*** use Beautiful Soup, Scrapy, or anything like them. Use **urllib.request** or **requests**.