CS 432/532: Web Science

Spring 2017

Assignment 3

Michael Micros

Honor Pledge

I pledge to support the Honor System of Old Dominion University. I will refrain from any form of academic dishonesty or deception, such as cheating or plagIiarism. I am aware that as a member of the academic community it is my responsibility to turn in all suspected violations of the Honor Code. I will report to a hearing if summoned.

Signed ________________________________

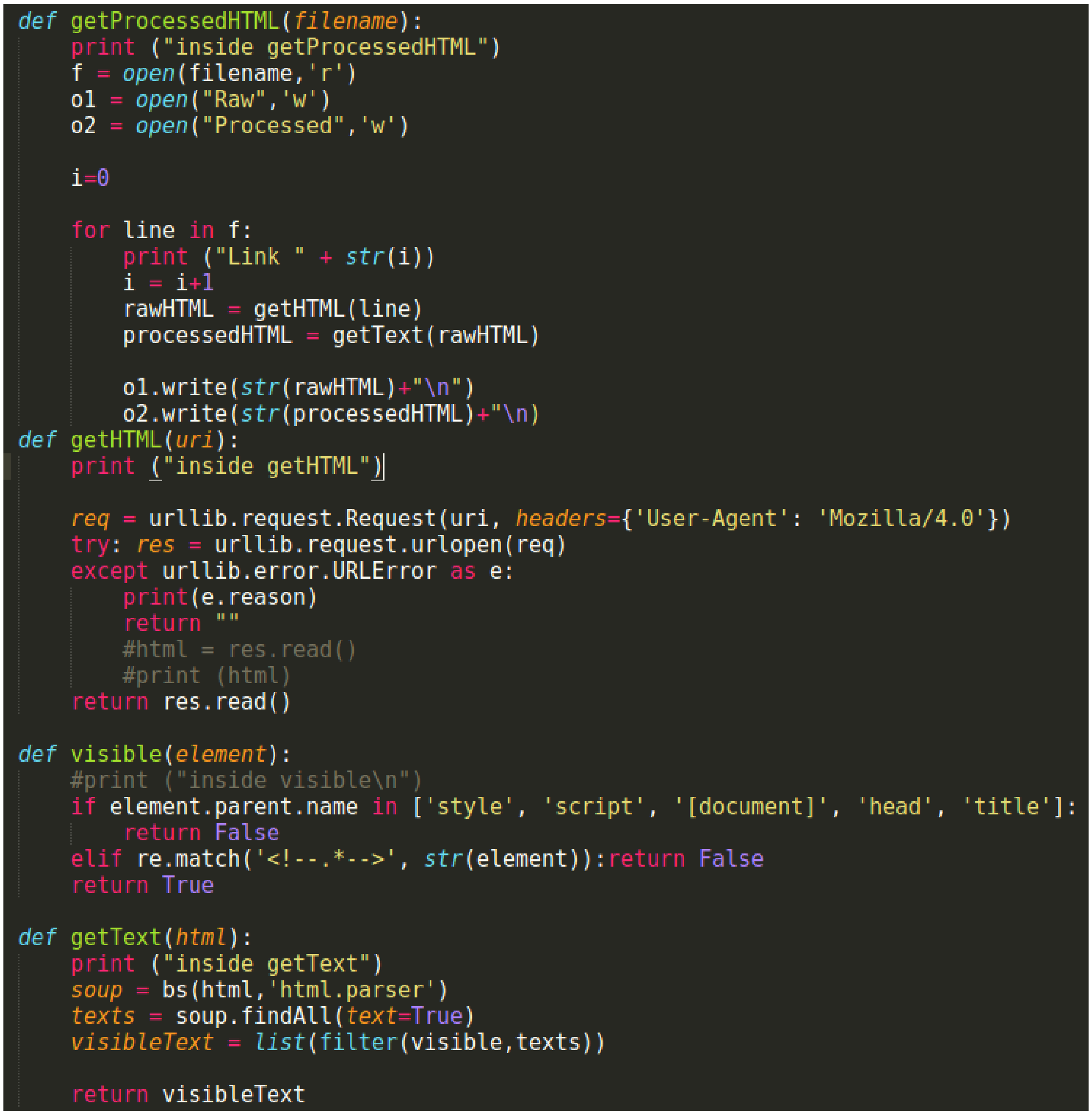
February 23, 2017

Problem 1: Downloading HTML from 1000 unique URIs

All the functions necessary to extract the processed HTML from the 1000 unique links are located in “part1.py”. The function “getProcessedHTML” takes as a parameter the filename of the 1000 URIs, and in turn calls “getHTML()” and “getText()’ ’ on each URI read from file. All the files are then saved to “Raw” and “Processed” located in the Part1 directory of the assignment on github.

The “getHTML()” function makes an http request using urllib making sure the response code ia 200 and then returns the raw HTML collected using urlopen().read()

The “getText()” function makes use of BeautifulSoup to extract the readable text from the HTML, following the example of Tugrur Ates (http://stackoverflow.com/questions/1936466/beautifulsoup-grab-visible-webpage-text)

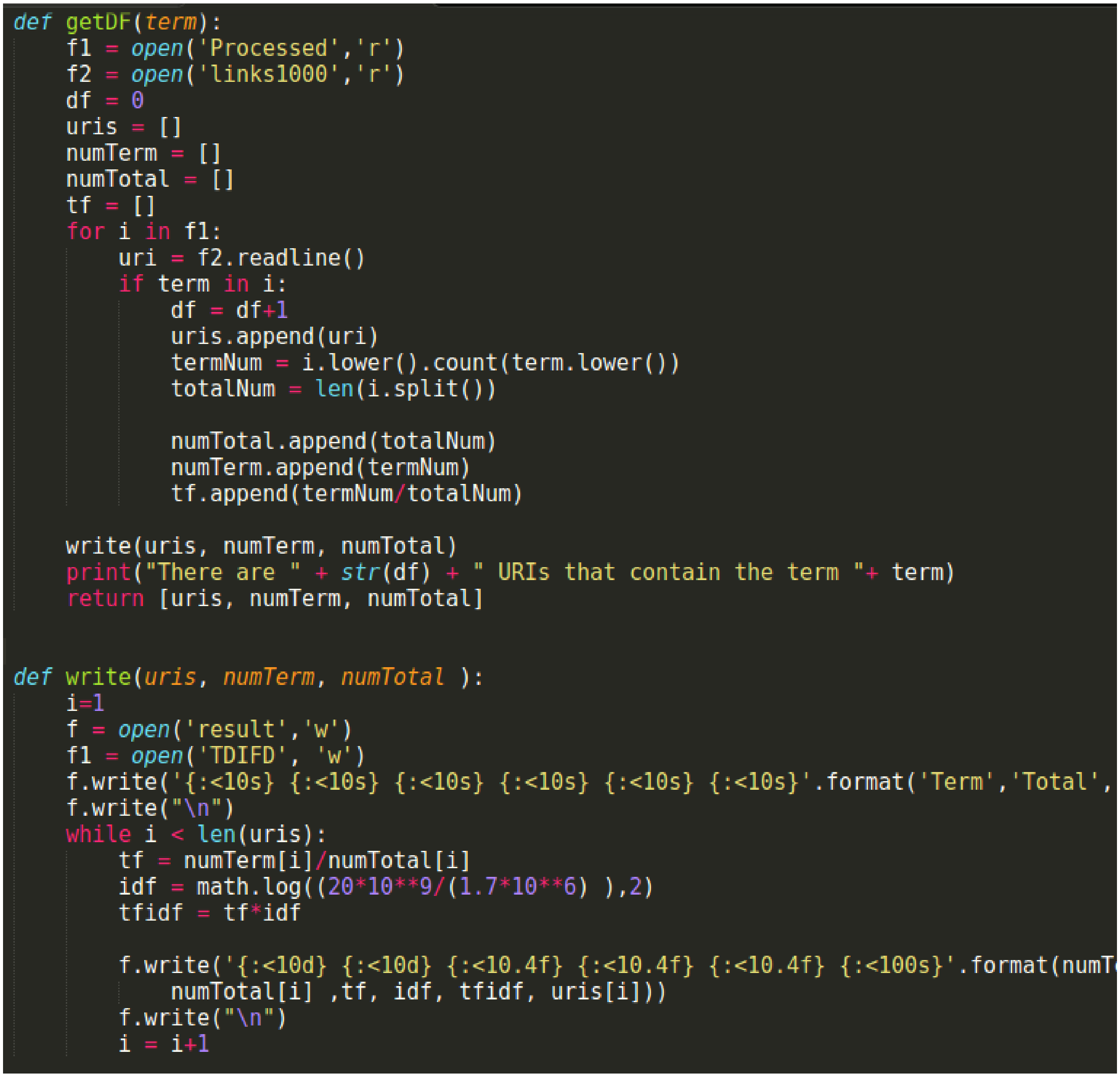


```
def getProcessedHTML(filename):
    print ("inside getProcessedHTML")
    f = open(filename,'r')
    o1 = open("Raw",'w')
    o2 = open("Processed",'w')

    i=0

    for line in f:
        print ("Link " + str(i))
        i = i+1
        rawHTML = getHTML(line)
        processedHTML = getText(rawHTML)

        o1.write(str(rawHTML)+"\n")
        o2.write(str(processedHTML)+"\n)
def getHTML(uri):
    print ("inside getHTML")

    req = urllib.request.Request(uri, headers={'User-Agent': 'Mozilla/4.0'})
    try: res = urllib.request.urlopen(req)
    except urllib.error.URLError as e:
        print(e.reason)
        return ""
        #html = res.read()
        #print (html)
    return res.read()

def visible(element):
    #print ("inside visible\n")
    if element.parent.name in ['style', 'script', '[document]', 'head', 'title']:
        return False
    elif re.match('<!--.*-->', str(element)):return False
    return True

def getText(html):
    print ("inside getText")
    soup = bs(html,'html.parser')
    texts = soup.findAll(text=True)
    visibleText = list(filter(visible,texts))

    return visibleText
```

Figure 1: Functions that extract 1000 unique links

Problem 2: Computation of TFIDF of 10 URIs

The query term selected was “Syria” and 27 documents had a match. The 10 documents with the highest TFIDF were selected. The processed HTML for each of the 100 URIs was read from file and checked whether it contained the “term” which was selected. In the function below, for the documents matching the query term, the number of instances of the term is calculated along with the total number of words the document contains. The the “write” function is called, which does additional calculations to compute the TF, IDF and TFIDF. Finally, Bing was the search engine used to calculate the IDF value.



```
def getDF(term):
    f1 = open('Processed','r')
    f2 = open('links1000','r')
    df = 0
    uris = []
    numTerm = []
    numTotal = []
    tf = []
    for i in f1:
        uri = f2.readline()
        if term in i:
            df = df+1
            uris.append(uri)
            termNum = i.lower().count(term.lower())
            totalNum = len(i.split())

            numTotal.append(totalNum)
            numTerm.append(termNum)
            tf.append(termNum/totalNum)

    write(uris, numTerm, numTotal)
    print("There are " + str(df) + " URIs that contain the term "+ term)
    return [uris, numTerm, numTotal]


def write(uris, numTerm, numTotal ):
    i=1
    f = open('result','w')
    f1 = open('TDIFD', 'w')
    f.write('{:<10s} {:<10s} {:<10s} {:<10s} {:<10s} {:<10s}'.format('Term','Total',
    f.write("\n")
    while i < len(uris):
        tf = numTerm[i]/numTotal[i]
        idf = math.log((20*10**9/(1.7*10**6) ),2)
        tfidf = tf*idf

        f.write('{:<10d} {:<10d} {:<10.4f} {:<10.4f} {:<10.4f} {:<100s}'.format(numT
            numTotal[i] ,tf, idf, tfidf, uris[i]))
        f.write("\n")
        i = i+1
```

Figure 2: part2.py

TFIDF	Instances	Total words	TF	TFIDF	URI
0.2671	21	1063	0.0198	13.5222	http://www.cnn.com/2016/10/09/politics/trump-pence-syria-disagreement/index.html
0.1290	14	1468	0.0095	13.5222	http://www.tmimag.com/all-articles/kurdish-people-not-monolith/
0.0361	7	2623	0.0027	13.5222	https://www.newsdeeply.com/refugees/articles/2017/02/14/it-is-scary-to-see-how-the-world-can-change-in-one-night
0.0270	4	2004	0.0020	13.5222	http://www.nbcnews.com/video/watch-live-actor-ashton-kutcher-testifies-at-hearing-on-ending-modern-slavery-877767747903?cid =sm_npd
0.0198	3	2045	0.0015	13.5222	http://www.ncr-iran.org/en/news/iran-world/22008-this-is-the-time-to-confront-iran-regime
0.0155	2	1742	0.0011	13.5222	http://www.reuters.com/article/us-usa-trump-russia-ukraine-idUSKBN15U0U0
0.0119	1	1137	0.0009	13.5222	http://www.iranfocus.com/en/index.php?option=com_content&view=article&id=31248&catid=4&Itemid=109
0.0108	1	1247	0.0008	13.5222	http://cnnphilippines.com/world/2017/02/15/us-special-ops-isis-fighters-killed.html
0.0106	4	5096	0.0008	13.5222	https://themoscowtimes.com/articles/russias-alleged-inf-violation-isnt-so-clear-cut-57161
0.0089	2	3026	0.0007	13.5222	http://www.newyorker.com/news/john-cassidy/its-time-for-a-proper-investigation-of-trumps-russia-ties

Problem 3: Ranking the pages

The tool use to get the pagerank for the 10 URIs was “http://pr.eyedomain.com/”. As can be seen in the previous table, there were not that many instances of the query term, with the exception of the first 3 URIs. That probably means that the term ”Syria” was present in a title to a link to another page of a large website. That explains why URIs corresponding to “nbcnews.com” or “reuters.com” find their way to the top of the PageRank table below.

PageRank	URI
0.9	http://www.cnn.com
0.8	http://www.nbcnews.com/video/watch-live-actor-ashton-kutcher-testifies-at-hearing-on-ending-modern-slavery-877767747903?cid=sm_npd
0.8	http://www.reuters.com/article/us-usa-trump-russia-ukraine-idUSKBN15U0U0
0.8	http://www.newyorker.com/news/john-cassidy/its-time-for-a-proper-investigation-of-trumps-russia-ties
0.5	http://www.ncr-iran.org/en/news/iran-world/22008-this-is-the-time-to-confront-iran-regime
0.5	http://www.iranfocus.com/en /index.php?option=com_content&view=article&id= 31248&catid=4&Itemid=109
0.4	https://www.newsdeeply.com/refugees/articles/2017/02/14/it-is-scary-to-see-how-the-world-can-change-in-one-night
0.0	http://www.tmimag.com/all-articles/kurdish-people-not-monolith/
0.0	http://cnnphilippines.com/world/2017/02/15/us-special-ops-isis-fighters-killed.html
0.0	https://themoscowtimes.com/articles/russias-alleged-inf-violation-isnt-so-clear-cut-