CS 432

Assignment 3

Daniel Moore

Part 1:

Part one of this assignment involved capturing the content of our 1000 unique URLs from assignment 2 and place this content in files. Once this was completed I was to remove the HTML markup from each file and place the processed content in separate files. I accomplished these tasks with the following script:

```
#!/bin/bash

# test-echo-urls.sh

while read -r full_url
do
    echo $full_url
    wget -O temp.txt "${full_url}"
    temp=`echo -n $full_url | md5sum| awk '{ print $1 }'`
    cp temp.txt ./outFiles/"${temp}.raw"
    echo "${full_url}" >> ./outFiles/"${temp}.raw"
    echo "${full_url}" > ./outFiles/"${temp}.processed"
    lynx -dump -force_html temp.txt >> ./outFiles/"${temp}.processed"
done < links.txt
```

The first portion of phase two involved finding a word which was contained within at least ten of the processed files(I chose the word “constitution”)and selecting ten of those files for later processing. I did this with the following command script :

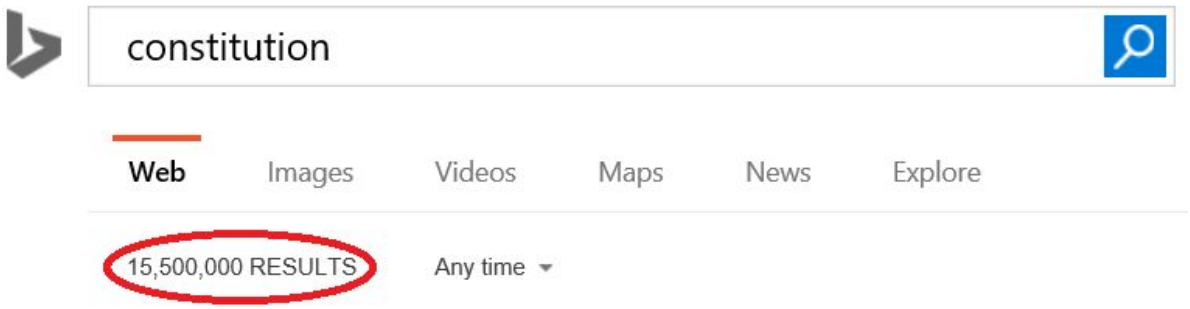
```
#!/bin/bash

for file in $(grep -lr "constitution" ./outFiles/*.processed | tail -10)
do
    cp "${file}" ./selection
done
```

I then used the following script to compute the TF values and print them to a file along with their URI:

```
#!/bin/bash

for file in ./selection/*.processed
do
    oc="$(grep -o 'constitution' "${file}" | wc -l)"
    count="$(wc -w "${file}" | awk '{print $1}')"
    tf=$(echo "scale=4; $oc/$count" | bc)
    echo "${tf}" >> tf.txt
    head -1 "${file}" >> tf.txt
done
```



Using Bing, a search of the word “constitution” returned 15,500,000 results. Using an estimated index size for Bing of 20,000,000,000 this gives an IDF value of 10.3.
For each entry URI I calculated the TFIDF value by multiplying the TF and IDF values. I did this with the following script:

```
from math import *

f = open("tf.txt","r")
out = open ("out.csv","a")
i = 0
tf = []
URI = []
for i in range (1,21):
    if i % 2 == 0:
        URI.append(f.readline())
    else:
        tf.append(f.readline())

out.write("TFIDF,TF,IDF,URI\n")

for  x in range (0,10):

    TFIDF = float(tf[x]) * 10.3
    out.write(str(TFIDF) + "," + str(tf[x]) + "," + str("10.3") + "," + str(URI[x]) + "\n")

```

I then placed these values in the following table:

TFIDF	TF	IDF	URI
0.03811	0.0037	10.3	https://www.nraila.org/issues/terrorist-watchlistno-f...
0.02884	0.0028	10.3	http://www.breitbart.com/big-journalism/2015/12/0...
0.02266	0.0022	10.3	http://www.tucsonsentinel.com/opinion/report/12...
0.01751	0.0017	10.3	http://thefederalist.com/2015/11/11/chicagos-not-...
0.00824	0.0008	10.3	http://opinion.injo.com/2016/01/251742-dont-gun-i...
0.00721	0.0007	10.3	https://www.nraila.org/articles/20151222/nra-cond...
0.00721	0.0007	10.3	http://www.news-journalonline.com/article/201511...
0.00721	0.0007	10.3	http://www.washingtontimes.com/news/2016/jan/...
0.00618	0.0006	10.3	https://www.nraila.org/articles/20151027/wayne-la...
0.00309	0.0003	10.3	http://www.knoxnews.com/news/local/right-name-...

I utilized the following source to calculate page rank:
http://www.seocentro.com/tools/search-engines/pagerank.html

All of the results for full URIs were “Undefined” I had to shorten the URIs in order to achieve a result. The URI portions for which I was able to obtain a score are in blue text.

URI	Page Rank
https://www.nraila.org/articles/20151222/nra-condemns-virginia-attorney-general-s-decision-to-play-politics-with-self-defense-rights	0.5
http://thefederalist.com/2015/11/11/chicagos-not-so-brilliant-bullet-tax/	0.0
https://www.nraila.org/issues/terrorist-watchlistno-fly-list/everytowns-claims-and-the-facts/	0.5
http://www.tucsonsentinel.com/opinion/report/121215_watch_list_op/watch-lists-should-never-used-deny-constitutional-rights/	0.4
http://opinion.injo.com/2016/01/251742-dont-gun-ive-still-got-3-questions-obama-gun-control-orders/	Undefined

http://www.knoxnews.com/news/local/right-name-wrong-man-knoxville-veterinarian-cant-get-off-no-fly-list-26f4d248-2e52-111a-e053-0100007-362692771.html	0.6
https://www.nraila.org/articles/20151027/wayne-lapierre-to-president-obama-this-is-how-you-stop-violent-crime	0.5
http://www.news-journalonline.com/article/20151130/opinion/151139997?tc=ar	0.6
http://www.washingtontimes.com/news/2016/jan/5/gun-rules-wont-make-easier-charges-ex-atf-agents/	0.7
http://www.breitbart.com/big-journalism/2015/12/07/la-times-counters-obama-background-checks-not-include-no-fly-list/	0.6

The highest page rank was obtained by www.washingtontimes.com while this URI had a TFIDF value of only .007 the highest TFIDF value was associated with: https://www.nraila.org/issues/terrorist-watchlistno-fly-list/everytowns-claims-and-the-facts/ whereas www.nra.org possessed the median page rank of 5/10.

URI	Page Rank	
https://www.nraila.org/articles/20151222/nra-condemns-virginia-attorney-general-s-decision-to-play-politics-with-self-defense-rights	0.5	0.0072
http://thefederalist.com/2015/11/11/chicagos-not-so-brilliant-bullet-tax/	0.0	0.018
https://www.nraila.org/issues/terrorist-watchlistno-fly-list/everytowns-claims-and-the-facts/	0.5	0.038

http://www.tucsonsentinel.com/opinion/report/121215_watch_list_op/watch-lists-should-never-used-deny-constitutional-rights/	0.4	0.023
http://opinion.injo.com/2016/01/251742-dont-gun-ive-still-got-3-questions-obama-gun-control-orders/	Undefined	0.0082
http://www.knoxnews.com/news/local/right-name-wrong-man-knoxville-veterinarian-cant-get-off-no-fly-list-26f4d248-2e52-111a-e053-0100007-362692771.html	0.6	.0031
https://www.nraila.org/articles/20151027/wayne-lapierre-to-president-obama-this-is-how-you-stop-violent-crime	0.5	0.0062
http://www.news-journalonline.com/article/20151130/opinion/151139997?tc=ar	0.6	0.0072
http://www.washingtontimes.com/news/2016/jan/5/gun-rules-wont-make-easier-charges-ex-atf-agents/	0.7	0.0072
http://www.breitbart.com/big-journalism/2015/12/07/la-times-counters-obama-background-checks-not-include-no-fly-list/	0.6	0.029