HW2 CS432

Ondra Torkilson

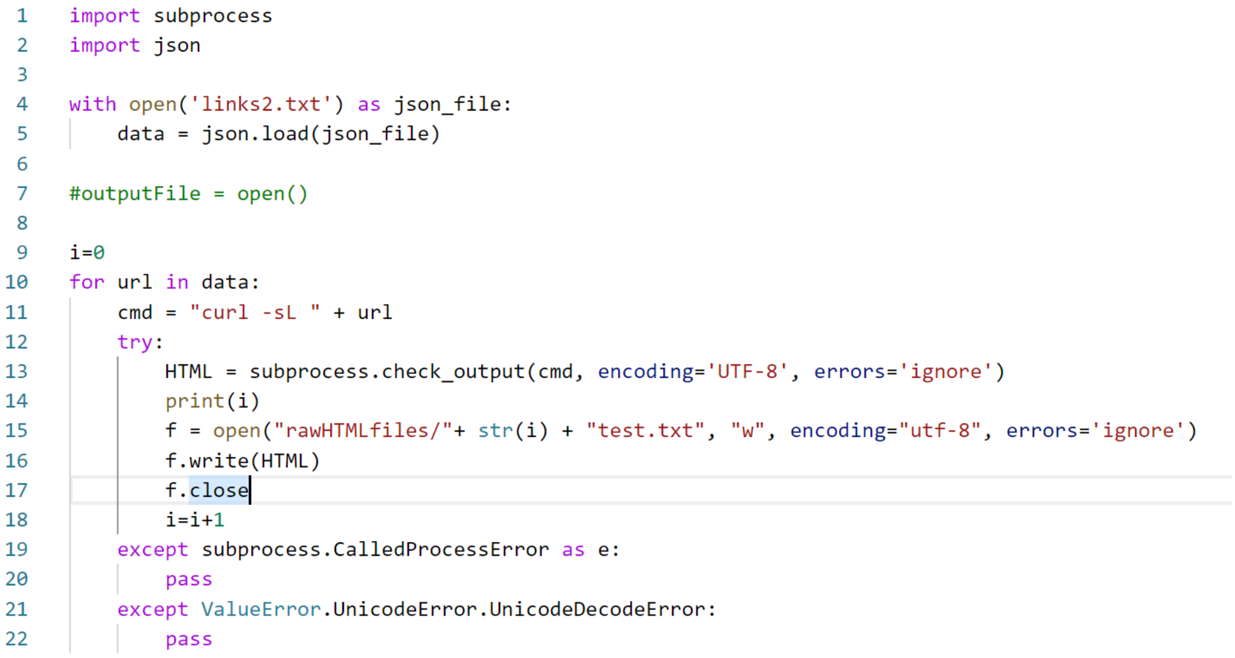
October 11, 2020

Question 1

Disclaimer: The links I used for homework 2 and 3 are different. I collected YouTube links for homework 2, and the processed HTML of YouTube sites was minimal. It was not going to work for the last part of question 1 of this assignment, which I realized when I started trying to use the boilerpipe extractor on the raw HTML from my YouTube links. So, I collected 1000 more links, and I followed the tag ”news” this time instead of ”YouTube.”

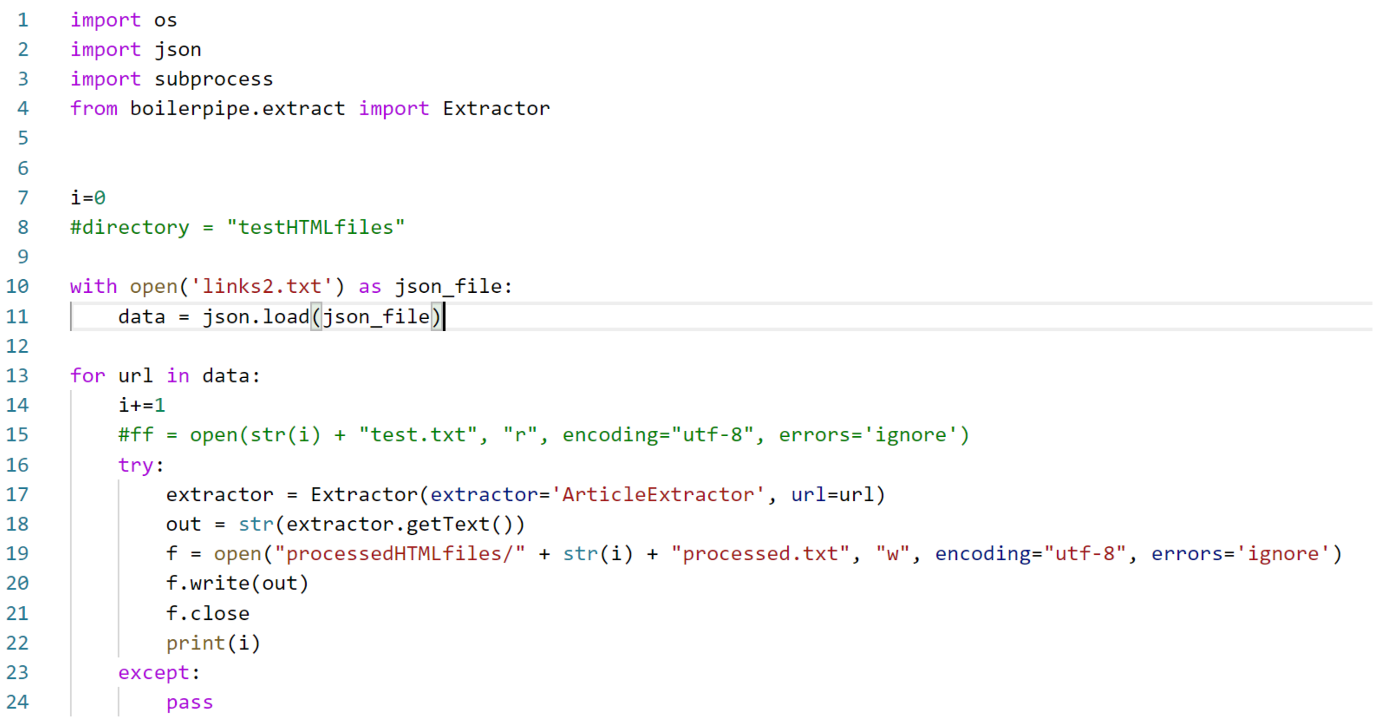
Pictured below, you see the python script I wrote to collect the HTML documents for the 1000 unique links. You can access these raw HTML files from my homework 3 github repo.

After collecting the raw HTML documents, I used a new python script to run the links through the boilerpipe extractor. For all 1000 links, I wrote the returned, parsed HTML to a file. These can be viewed on the repo as well.



```python
import subprocess
import json

with open('links2.txt') as json_file:
    data = json.load(json_file)

#outputFile = open()

i=0
for url in data:
    cmd = "curl -sL " + url
    try:
        HTML = subprocess.check_output(cmd, encoding='UTF-8', errors='ignore')
        print(i)
        f = open("rawHTMLfiles/"+ str(i) + "test.txt", "w", encoding="utf-8", errors='ignore')
        f.write(HTML)
        f.close
        i=i+1
    except subprocess.CalledProcessError as e:
        pass
    except ValueError.UnicodeError.UnicodeDecodeError:
        pass
```

Figure 1: getHTML.py

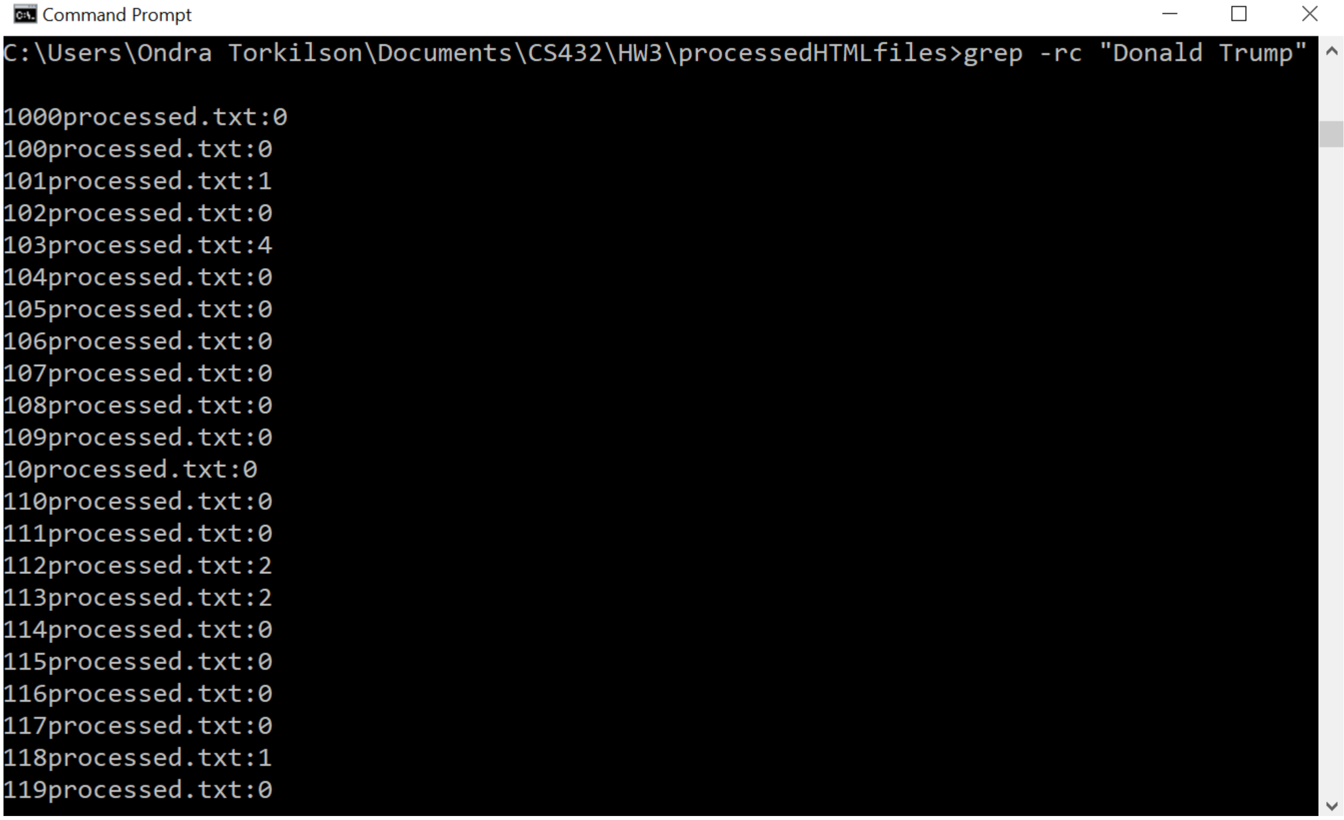


```
import os
import json
import subprocess
from boilerpipe.extract import Extractor


i=0
#directory = "testHTMLfiles"

with open('links2.txt') as json_file:
    data = json.load(json_file)

for url in data:
    i+=1
    #ff = open(str(i) + "test.txt", "r", encoding="utf-8", errors='ignore')
    try:
        extractor = Extractor(extractor='ArticleExtractor', url=url)
        out = str(extractor.getText())
        f = open("processedHTMLfiles/" + str(i) + "processed.txt", "w", encoding="utf-8", errors='ignore')
        f.write(out)
        f.close
        print(i)
    except:
        pass
```

Figure 2: boilerPlate.py

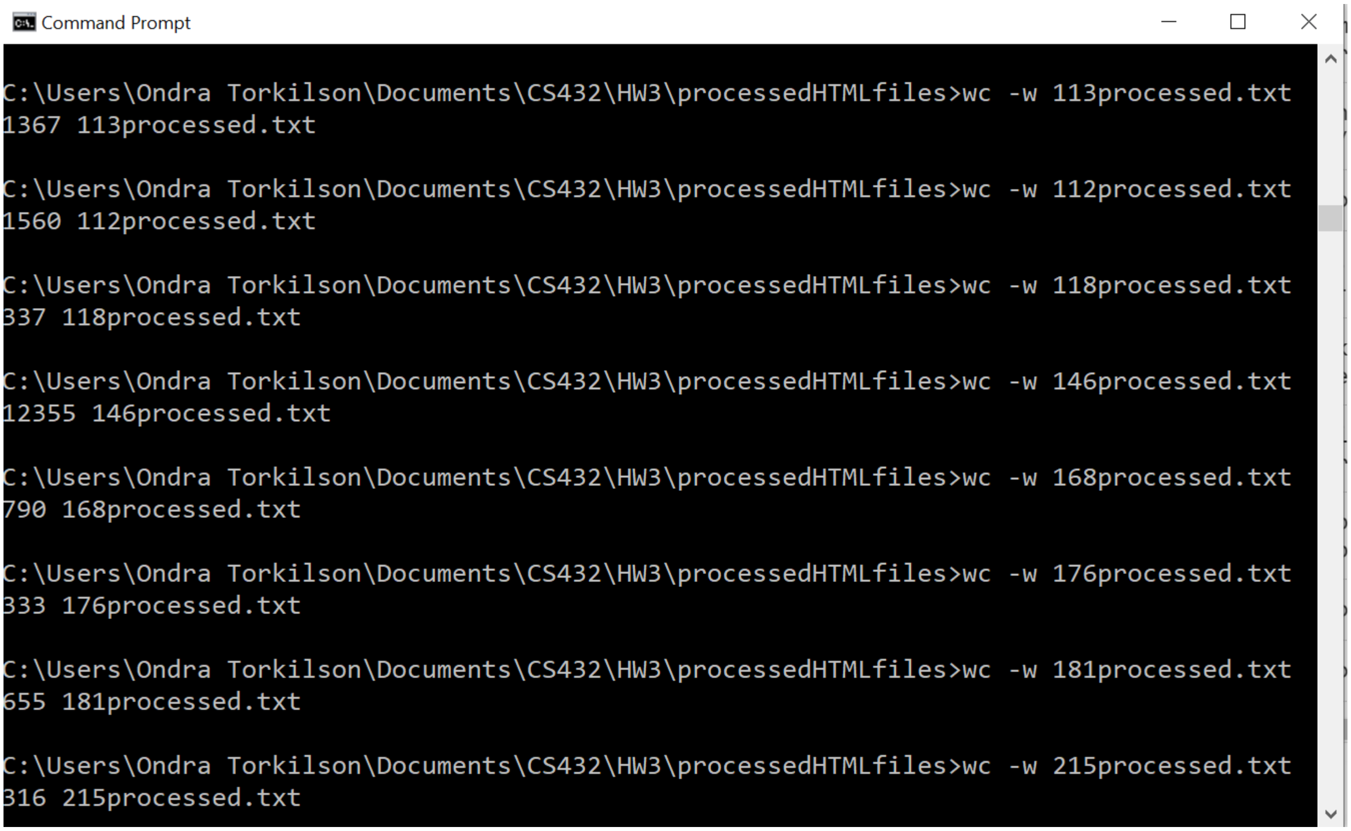
Questions 2

I chose to query ”Donald Trump.” I used grep with the options r and c, to recursively search the directory where the processed HTML files were and to return the count per file on the query term. This is shown in Figure 3. I then chose the first ten files that contained at least one mention of the query term to move forward with constructing the TF-IDF table. I took note of the URI and the number of times the search term occurred in that URI. Then, for each URI, I returned the number of words within each parsed HTML file which is shown in Figure 4. Figure 5 shows the TF-IDF table. As an example for how I calculated this, I will walk through the URI in line nine of the table, from the domain www.ajc.com. This URI contained the query term once, and it had 882 words total. This lead to the value in the TF column of .001, from 1/882 which is .001133. From there, I figured that the IDF will be log base two of 55 billion (size of corpus) divided by 69 (the total number of documents containing the search term). This number applies to all of the URIs, since I use the same corpus and links. Multiplying the TF * IDF resulted in the column one value of .030 for the URI on line nine.



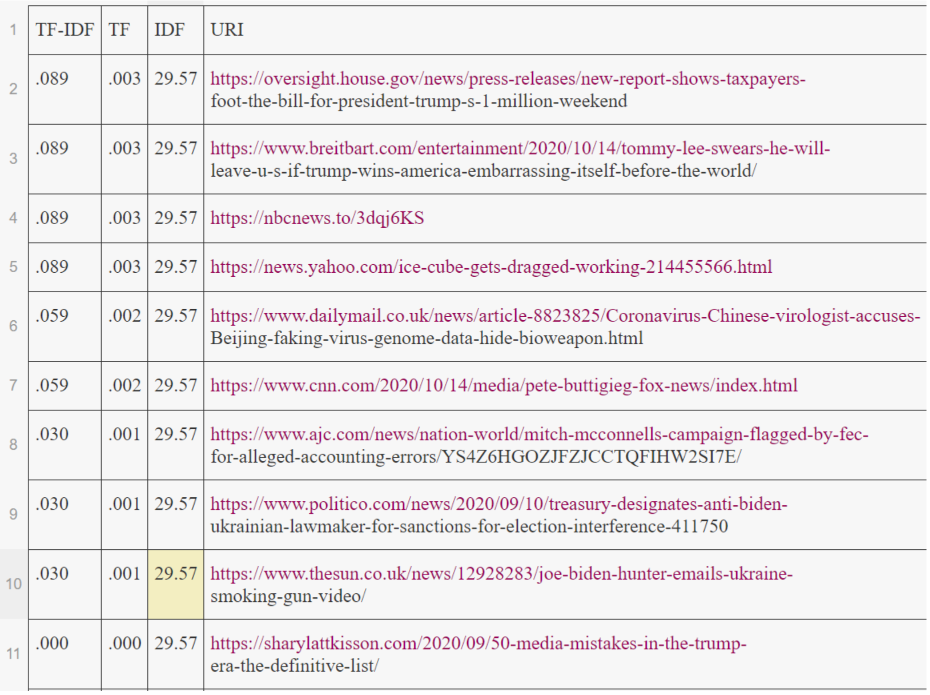
```
C:\Users\Ondra Torkilson\Documents\CS432\HW3\processedHTMLfiles>grep -rc "Donald Trump"

1000processed.txt:0
100processed.txt:0
101processed.txt:1
102processed.txt:0
103processed.txt:4
104processed.txt:0
105processed.txt:0
106processed.txt:0
107processed.txt:0
108processed.txt:0
109processed.txt:0
10processed.txt:0
110processed.txt:0
111processed.txt:0
112processed.txt:2
113processed.txt:2
114processed.txt:0
115processed.txt:0
116processed.txt:0
117processed.txt:0
118processed.txt:1
119processed.txt:0
```

Figure 3: Grepping the processed files



```
C:\Users\Ondra Torkilson\Documents\CS432\HW3\processedHTMLfiles>wc -w 113processed.txt
1367 113processed.txt

C:\Users\Ondra Torkilson\Documents\CS432\HW3\processedHTMLfiles>wc -w 112processed.txt
1560 112processed.txt

C:\Users\Ondra Torkilson\Documents\CS432\HW3\processedHTMLfiles>wc -w 118processed.txt
337 118processed.txt

C:\Users\Ondra Torkilson\Documents\CS432\HW3\processedHTMLfiles>wc -w 146processed.txt
12355 146processed.txt

C:\Users\Ondra Torkilson\Documents\CS432\HW3\processedHTMLfiles>wc -w 168processed.txt
790 168processed.txt

C:\Users\Ondra Torkilson\Documents\CS432\HW3\processedHTMLfiles>wc -w 176processed.txt
333 176processed.txt

C:\Users\Ondra Torkilson\Documents\CS432\HW3\processedHTMLfiles>wc -w 181processed.txt
655 181processed.txt

C:\Users\Ondra Torkilson\Documents\CS432\HW3\processedHTMLfiles>wc -w 215processed.txt
316 215processed.txt
```

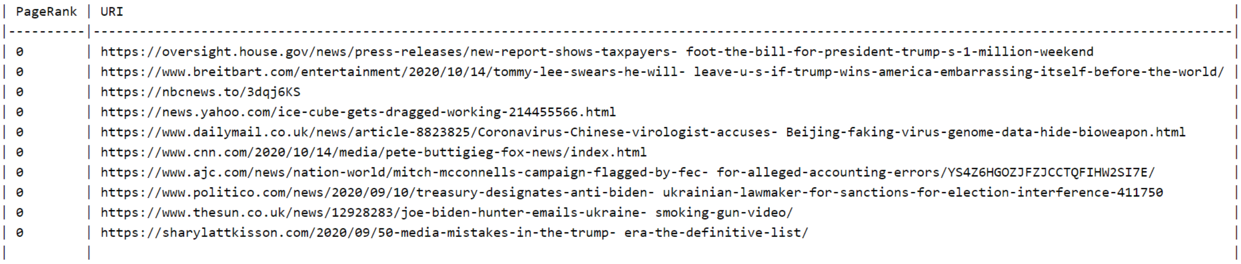
Figure 4: Counting the total words per file



TF-IDF	TF	IDF	URI
.089	.003	29.57	https://oversight.house.gov/news/press-releases/new-report-shows-taxpayers-foot-the-bill-for-president-trump-s-1-million-weekend
.089	.003	29.57	https://www.breitbart.com/entertainment/2020/10/14/tommy-lee-swears-he-will-leave-u-s-if-trump-wins-america-embarrassing-itself-before-the-world/
.089	.003	29.57	https://nbcnews.to/3dqj6KS
.089	.003	29.57	https://news.yahoo.com/ice-cube-gets-dragged-working-214455566.html
.059	.002	29.57	https://www.dailymail.co.uk/news/article-8823825/Coronavirus-Chinese-virologist-accuses-Beijing-faking-virus-genome-data-hide-bioweapon.html
.059	.002	29.57	https://www.cnn.com/2020/10/14/media/pete-buttigieg-fox-news/index.html
.030	.001	29.57	https://www.ajc.com/news/nation-world/mitch-mcconnells-campaign-flagged-by-fec-for-alleged-accounting-errors/YS4Z6HGOZJFZJCCTQFIHW2SI7E/
.030	.001	29.57	https://www.politico.com/news/2020/09/10/treasury-designates-anti-biden-ukrainian-lawmaker-for-sanctions-for-election-interference-411750
.030	.001	29.57	https://www.thesun.co.uk/news/12928283/joe-biden-hunter-emails-ukraine-smoking-gun-video/
.000	.000	29.57	https://sharylattkisson.com/2020/09/50-media-mistakes-in-the-trump-era-the-definitive-list/

Figure 5: TF-IDF with corpus size of 55b

Question 3

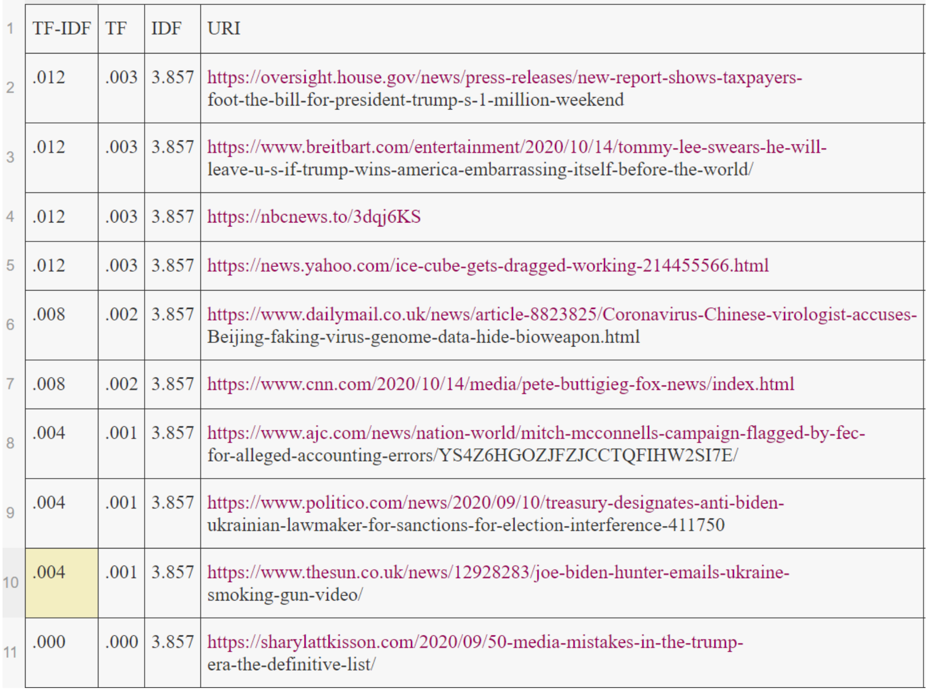


PageRank	URI
0	https://oversight.house.gov/news/press-releases/new-report-shows-taxpayers- foot-the-bill-for-president-trump-s-1-million-weekend
0	https://www.breitbart.com/entertainment/2020/10/14/tommy-lee-swears-he-will- leave-u-s-if-trump-wins-america-embarrassing-itself-before-the-world/
0	https://nbcnews.to/3dqj6KS
0	https://news.yahoo.com/ice-cube-gets-dragged-working-214455566.html
0	https://www.dailymail.co.uk/news/article-8823825/Coronavirus-Chinese-virologist-accuses- Beijing-faking-virus-genome-data-hide-bioweapon.html
0	https://www.cnn.com/2020/10/14/media/pete-buttigieg-fox-news/index.html
0	https://www.ajc.com/news/nation-world/mitch-mcconnells-campaign-flagged-by-fec- for-alleged-accounting-errors/YS4Z6HGOZJFZJCCTQFIHW2SI7E/
0	https://www.politico.com/news/2020/09/10/treasury-designates-anti-biden- ukrainian-lawmaker-for-sanctions-for-election-interference-411750
0	https://www.thesun.co.uk/news/12928283/joe-biden-hunter-emails-ukraine- smoking-gun-video/
0	https://sharylattkisson.com/2020/09/50-media-mistakes-in-the-trump- era-the-definitive-list/

Figure 6: Page Rank Table

Here, you can see in Figure 6 my table for PageRank using http://www.checkpagerank.net/ . The PageRank for all the URIs were zero. Comparing this to the TF-IDF rank from question 2, the values are consistent, as those were all between .09 and .00 which rounds down to 0. Although PageRank and TF-IDF for the URIs were similar, the latter allowed for more control and precision in the calculation.

Question 4



TF-IDF	TF	IDF	URI
.012	.003	3.857	https://oversight.house.gov/news/press-releases/new-report-shows-taxpayers-foot-the-bill-for-president-trump-s-1-million-weekend
.012	.003	3.857	https://www.breitbart.com/entertainment/2020/10/14/tommy-lee-swears-he-will-leave-u-s-if-trump-wins-america-embarrassing-itself-before-the-world/
.012	.003	3.857	https://nbcnews.to/3dqj6KS
.012	.003	3.857	https://news.yahoo.com/ice-cube-gets-dragged-working-214455566.html
.008	.002	3.857	https://www.dailymail.co.uk/news/article-8823825/Coronavirus-Chinese-virologist-accuses-Beijing-faking-virus-genome-data-hide-bioweapon.html
.008	.002	3.857	https://www.cnn.com/2020/10/14/media/pete-buttigieg-fox-news/index.html
.004	.001	3.857	https://www.ajc.com/news/nation-world/mitch-mcconnells-campaign-flagged-by-fec-for-alleged-accounting-errors/YS4Z6HGOZJFZJCCTQFIHW2SI7E/
.004	.001	3.857	https://www.politico.com/news/2020/09/10/treasury-designates-anti-biden-ukrainian-lawmaker-for-sanctions-for-election-interference-411750
.004	.001	3.857	https://www.thesun.co.uk/news/12928283/joe-biden-hunter-emails-ukraine-smoking-gun-video/
.000	.000	3.857	https://sharylattkisson.com/2020/09/50-media-mistakes-in-the-trump-era-the-definitive-list/

Figure 7: Redo TD-IDF with new corpus size of 1000

Here, I redid the TF-IDF using a corpus size of 1000 to represent the 1000 unique links I chose the ten links from. This resulted in a different IDF column, and the calculation for that was log base 2 of 1000 divided by 69, resulting in 3.857. The TF-IDF got smaller, closer to 0 for all the links.