

OLD DOMINION UNIVERSITY

Assignment 3

David Bayard

March 3, 2019

QUESTION 1.

Download the 1000 URIs from assignment 2. "curl", "wget", or "lynx" are all good candidate programs to use. We want just the raw HTML, not the images, stylesheets, etc.

Solution:

The requests library was used to download the URIs, sending GET requests to each of the URIs and creating a response object from the response. This was repeated for each URI in the list of existing URIs from assignment 2.

```
1  try:
2      response = requests.get(uri, allow_redirects=True, timeout=30)
3
4  except Exception as e:
5      print('Error: ', str(e))
```

Listing 1: Download HTML

After downloading the HTML, a hashing function was used to create distinct file names for each URI. The returned hash code and URI were then appended to the "Records.txt" file. After creating each file, the .text() method of the response object is used to retrieve the HTML and write it to the new file.

```
1  with open('tempFile.txt', 'rb') as afile:
2      buf = afile.read()
3      hasher.update(buf)
4
5      fileName = hasher.hexdigest() + '.html'
6
7      #Write to the specified file (with hash code)
8      writeComplete = open(fileName, 'w')
9
10     writeComplete.write(response.text)
```

Listing 2: Make Hashcode

"python-boilerpipe" was used to extract the HTML from each file. In order to use this library, a virtual environment was created, and the source code was installed in this environment. There were difficulties when installing "python-boilerpipe", but this was solved by specifying the Python version to use (3.6) when installing.

```
1  with open(a['FileName'], 'r') as read:
2      html = read.read()
3
4      #Extract html from file
5      extractor = Extractor(extractor='ArticleExtractor', html=html)
6      extractedText = extractor.getText()
```

Listing 3: Extract Text

The code above is responsible for extracting the majority of the HTML from each hashed document, resulting in new files that contain only the text from the HTML files.

QUESTION 2

Choose a query term (e.g., "shadow") that is not a stop word (see week 5 slides) and not HTML markup from step 1 (e.g., "http") that matches at least 10 documents (hint: use "grep" on the processed files). If the term is present in more than 10 documents, choose any 10 from your list.

Solution:

The query term used was "Congress". This query term is relevant within the context of the downloaded documents, and it was expected to have been found in multiple documents. With the query term selected, a search was conducted on each of the document files containing extracted text, using grep as the searching tool.

```
1 process = subprocess.Popen(['grep', '-rwn', arg, completeName], stdout=subprocess.PIPE)
2     stdout, stderr = process.communicate()
3     if len(stdout) != 0:
4         num_words = 0
```

Listing 4: Query Term Search

In the above code, there is a list "stdout", which contains all of the information retrieved from grep search. Moreover, the if condition filters out any documents that did not match the query term.

Two very important segments of code are listed below. This code is responsible for counting the number of words that each document contains, as well as the number of times the query term appears in the document. This information is then stored in a file for later use.

```
1 with open(completeName, 'r') as readFile:
2     for line in readFile:
3         words = line.split()
4
5         ## Keeps track of number of words in file ##
6         num_words += len(words)
7
8     with open(completeName) as toCount:
9
10        ## Code from Joshua on slack, Counts occurrence of query term in document##
11        wordcount = Counter(toCount.read().split())
12        wordOccurance = wordcount['Congress']
13
14        if (wordOccurance != 0):
15            fileCounter += 1
```

Listing 5: Word Count and Count for Query Term

The following equations listed below were used to calculate the TFIDF, TF, and IDF:

$$TF = \frac{\text{Occurrences of Query Term}}{\text{Total Word Count}}$$

$$IDF = \frac{\text{Total Number of Documents}}{\text{Documents Containing Query Term}}$$

$$TFIDF = TF * IDF$$

With regards to normalization for question 3, the following equation was used:

$$\text{Normalized Score} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

All values were rounded to four significant figures (after the decimal) because most of the values for the TF, IDF, and TFIDF were smaller than those of the assignment example.

Moreover, a logarithmic function was not required for the IDF because the total number of documents was used as the corpus, instead of using a search engine. All of the results for the TF, IDF, and TFIDF calculations are stored in the "CalculationRecords2.txt" file for future reference.

In order to calculate the TF, IDF, and TFIDF, the snippet of code supplied below was used. This code takes the data from the "MatchedFormatted.json" file and runs it through the formulas supplied above. In this case, the corpus is 1130, and the query term was found in 375 files, found in "fileOccurrence.txt". The TF, IDF, and TFIDF are then rounded to four significant figures, and are written to a dictionary for later use.

```
1 if(checkURI in collectionURI):
2     ## Word Count (Frequency) / Number of words found in document
3     TF = entry["WordCount"] / entry["NumberWords"]
4     roundedTF = round(TF,4)
5
6     ## Had a total of 1130 documents, Query Term was found in 375 documents out of 1130
7     IDF = 1130/375
8     roundedIDF = round(IDF,4)
9
10    TF_IDF = TF * IDF
11    roundedTF_IDF = round(TF_IDF, 4)
12
13    calcData["TFIDF"] = roundedTF_IDF
14    calcData["TF"] = roundedTF
15    calcData["IDF"] = roundedIDF
16    calcData["URI"] = entry["URI"]
17
18    collectionCalc.append(calcData)
```

Listing 6: Calculations

The URIs in the table below were shortened to fit within the column of the table. The full URI and data is provided in the “CalculationRecords2.txt” file.

Calculate TF, IDF, and TFIDF

Calculation Records Using Query Term “Congress”			
TFIDF	TF	IDF	URI
0.0339	0.0113	3.0133	https://www.numbersusa.com/
0.0324	0.0108	3.0133	https://www.facebook.com/
0.0261	0.0086	3.0133	https://in.reuters.com/
0.0222	0.0074	3.0133	https://www.reuters.com/
0.0209	0.0069	3.0133	http://www.msn.com/
0.0160	0.0053	3.0133	https://www.cnn.com
0.0155	0.0051	3.0133	https://dailywn.com/
0.0143	0.0047	3.0133	https://www.upi.com/
0.0135	0.0045	3.0133	https://www.vox.com/
0.0126	0.0042	3.0133	https://whyy.org/
0.0105	0.0035	3.0133	https://www.cnbc.com/
0.0099	0.0033	3.0133	https://www.theguardian.com/
0.0088	0.0029	3.0133	https://www.nbcnews.com/
0.0074	0.0025	3.0133	https://www.nytimes.com/
0.0071	0.0024	3.0133	https://newspunch.com/
0.0060	0.0020	3.0133	https://apnews.com/
0.0034	0.0011	3.0133	https://fullmagazine.us/

As an example, the picture below provides an graphical view of what is occurring behind the scenes. This is only a segment of the program though, displaying where each occasion of the Query Term is located, and how many times it occurs.

```

david@david-Aspire-E5-575: ~/Documents/CS532/Assignment3/Assignment3/Files3/ExtractedText
File Edit View Search Terminal Help
(env) david@david-Aspire-E5-575: ~/Documents/CS532/Assignment3/Assignment3/Files3/ExtractedText$ grep -rwn Congress
867d27a4f1ac7e97b4ef3b1a63769230.txt:5:WASHINGTON (Reuters) - President Donald Trump vowed on Thursday to declare a national emergency in an attempt to f
und his U.S.-Mexico border wall without congressional approval, a step likely to plunge him into a battle with Congress over constitutional powers.
867d27a4f1ac7e97b4ef3b1a63769230.txt:6:Conceding defeat in his earlier demand that Congress provide him with $5.7 billion in wall money, Trump agreed to
sign a government-funding bill that lacks money for his wall, but prevents another damaging government shutdown.
867d27a4f1ac7e97b4ef3b1a63769230.txt:10:The top Democrat in Congress denounced the president's move. Asked by reporters if she would file a legal challen
ge to an emergency declaration, House of Representatives Speaker Nancy Pelosi said: "I may, that's an option."
867d27a4f1ac7e97b4ef3b1a63769230.txt:15:An emergency declaration could infringe on Congress' authority to make major decisions about taxpayer funds, a fu
ndamental check and balance spelled out in the Constitution.
867d27a4f1ac7e97b4ef3b1a63769230.txt:16:For weeks, as the president's wall-funding demand to Congress went nowhere, even after a historic 35-day partial
government shutdown, the White House explored whether an emergency declaration could be invoked to redirect taxpayer funds committed by Congress for othe
r purposes towards the wall.
867d27a4f1ac7e97b4ef3b1a63769230.txt:28:It also includes $1.37 billion in new money to help build 55 miles (88.5 km) of new physical border barriers. Tha
t is the same level of funding Congress appropriated for border security measures last year, including barriers, but not concrete walls.
867d27a4f1ac7e97b4ef3b1a63769230.txt:31:Some of Trump's fellow Republicans have warned him that declaring a national emergency could set a dangerous prec
edent, opening the door for a future Democratic president to circumvent Congress and declare emergencies on perhaps climate change, gun control or health
care insurance.
b9637da74463e26a9302a44001aed5de.txt:9:Here are two charts that show the impact of immigration on the Social Security system. The first chart shows how t
he Social Security Trust Fund (essentially a savings account) will run out of money if Congress doesn't do something to boost its reserves, such as incre
asing legal immigration and closing a tax loophole for rich workers.
b9637da74463e26a9302a44001aed5de.txt:25:Congress needs to close the tax loophole
49124027c0ccca7d957ff3a533c04509.txt:3:Congress is preparing to pass a bipartisan spending measure designed to settle the border security debate and keep
the government open past Friday's funding deadline. CBS News chief congressional correspondent Nancy Cordes joins CBSN to explain what's in the bill - a
nd what's not.
817c84a8bd5ab84824d302de9420cb5c.txt:5:WASHINGTON (Reuters) - President Donald Trump vowed on Thursday to declare a national emergency in an attempt to f
und his U.S.-Mexico border wall without congressional approval, a step likely to plunge him into a battle with Congress over constitutional powers.
817c84a8bd5ab84824d302de9420cb5c.txt:6:Conceding defeat in his earlier demand that Congress provide him with $5.7 billion in wall money, Trump agreed to
sign a government-funding bill that lacks money for his wall, but prevents another damaging government shutdown.
817c84a8bd5ab84824d302de9420cb5c.txt:10:The top Democrat in Congress denounced the president's move. Asked by reporters if she would file a legal challen
ge to an emergency declaration, House of Representatives Speaker Nancy Pelosi said: "I may, that's an option."
817c84a8bd5ab84824d302de9420cb5c.txt:15:An emergency declaration could infringe on Congress' authority to make major decisions about taxpayer funds, a fu
ndamental check and balance spelled out in the Constitution.
817c84a8bd5ab84824d302de9420cb5c.txt:16:For weeks, as the president's wall-funding demand to Congress went nowhere, even after a historic 35-day partial
government shutdown, the White House explored whether an emergency declaration could be invoked to redirect taxpayer funds committed by Congress for othe
r purposes toward the wall.
817c84a8bd5ab84824d302de9420cb5c.txt:28:It also includes $1.37 billion in new money to help build 55 miles (88.5 km) of new physical border barriers. Tha
t is the same level of funding Congress appropriated for border security measures last year, including barriers, but not concrete walls.
817c84a8bd5ab84824d302de9420cb5c.txt:31:Some of Trump's fellow Republicans have warned him that declaring a national emergency could set a dangerous prec
edent, opening the door for a future Democratic president to circumvent Congress and declare emergencies on perhaps climate change, gun control or health
care insurance.
92f1a69805f6a060301ced8104a06720.txt:3:The emergency declaration could allow President Donald Trump to procure wall funding without approval from Congres
s, but Democrats have promised to challenge the move.
92f1a69805f6a060301ced8104a06720.txt:8:Congress overwhelmingly passed a massive budget and border security deal Thursday, after President Donald Trump co
mitted to signing the legislation but said he'd also declare a national emergency to build a wall along the U.S.-Mexico border.

```

(a) grep behind the scene

QUESTION 3

Now rank the same 10 URIs from question 2 but this time by their PageRank. Use any of the free PR estimators on the web.

The PR estimator used for this assignment was http://www.prchecker.info/check_page_rank.php. There were a few issues that occurred while using the PR estimator, which resulted in less accurate results.

One issue associated with the PR estimator is that it only ranks a page based off of the home page of each URI. This resulted in a less specific PR, as it was not based off of the correct page. Another issue was that several of the chosen URIs had a page rank of zero. These URIs were mentioned in the “NoRanks.txt” file, and were not added to the list of normalized ranks in the “PageRanks.txt” file.

There was no real code involved with retrieving the page ranks, because of the anti-bot captchas. Instead, the URIs were manually run through the PR estimator. A table of each of the page ranks is listed below, displaying the pages that did not have a rank of zero.

Page Ranks for 10 URI

Page Ranks	
Page Rank	URI
1.0	https://www.cnn.com/
1.0	https://www.facebook.com/
0.9	https://www.reuters.com/
0.9	https://in.reuters.com/
0.9	http://www.msn.com/
0.8	https://www.theguardian.com/
0.8	https://www.upi.com/
0.7	https://www.vox.com/
0.7	https://whyy.org/
0.7	https://www.numbersusa.com/

The Page Rank results seem to be inaccurate, and do not produce a correlation between the page rank and the TF, IDF, and TFIDF scores. For example, <https://www.numbersusa.com/> ranks lowest on the Page Rank table, but scores highest on the TFIDF value.

Moreover, <https://www.facebook.com/> ranks second on both table, contradicting any possible correlation from the previous statement. This hints towards the idea that the page rank website is producing false results, and not ranking pages correctly. This is even more evident by the fact that <https://www.nbcnews.com/> did not have a page rank.

QUESTION 4. EXTRA CREDIT

Compute the Kendall TauB “b” score for both lists. Report both the Tau value and the “p” value.

Solution:

Python scipy provides a function called `scipy.stats.kendalltau` for this exact purpose. This function was implemented, providing the list of page ranks against the list of TFIDFs, TFs, and IDFs, as shown below.

```
1  ## Calculate Kendall Tau_b Score TFIDF ##
2  tauTFIDF, p_valueTFIDF = sp.stats.kendalltau(normalizedRanks, collectionTFIDF)
```

Listing 7: Word Count and Count for Query Term

Tau Calculations

Correlation		
Comparison	Tau	P Value
TFIDF VS Page Rank	0.25	0.19

As seen in the table above, the Tau b score indicates that there is a minor association between the page ranks and the TFIDE. This is evident by a Tau b score, of .25, indicating that there is an association, but that it is insignificant.