CS595 Intro to Web Science, Assignment #3

Valentina Neblitt-Jones

October 5, 2013

Question 1

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

from the command line:

% curl http://www.cnn.com/ >www.cnn.com

% wget -O www.cnn.com http://www.cnn.com

% lynx -source http://www.cnn.com/ >www.cnn.com

"www.cnn.com" is just an example output file name, keep in mind that the shell will not like some of the characters that can occur in the URIs (e.g., "?", "&"). You might want to hash the URIs, like:

% echo -n"http://www.cs.odu.edu/show_features.shtml?72" — md5 41d5f125d13b4bb554e6e31b6b591eeb

("md5sum" on some machines; note the "-n" in echo – this removes the trailing newlines.)

Now use a tool to remove (most) of the HTML markup. "lynx" will do a fair job:

% lynx -dump -force_html www.cnn.com >www.cnn.com.processed

Use another (better) tool if you know of one. Keep both files for each URI (i.e., raw HTML and processed).

Answer to Question 1

I separated the task into two scripts - one that downloads the URIs (Figure 1) and then one that strips the HTML markup (Figure 2). Figure 3 shows a file before removing HTML markup and Figure 4 shows the file after removing HTML markup. Additionally, a finding aid was developed in order to match up the md5 with the URI (Figure 5).

```
mylinks=`cat uniqueURIs2.txt`
for link in $mylinks
do
    filename=`echo -n $link | md5`
    echo "Working on $link"
    echo $filename $link >> findingaid
    curl $link > $filename
done
```

Figure 1: Shell Script to Download URIs

```
directory="/Users/vneblitt/Documents/cs595-f13/assignment03/raw"
mypages=`ls $directory`
for page in $mypages
do
    #filename=`echo -n $link | md5`
    echo "Working on $page"
    #echo $filename $link >> findingaid
    newfilename="/Users/vneblitt/Documents/cs595-f13/assignment03/processed/${page}.processed"
    lynx -dump -force_html $directory/$page > $newfilename
done
```

Figure 2: Shell Script to Remove Markup

```
        </div>
        <p>In news that will please soon-to-be-in-mourning <em>Breaking Bad</em> fans, AMC is greenl
ing Bad </em>character Saul Goodman, played by Bob Odenkirk. While the network and studio, Sony Pict
reached a licensing agreement for the show, it&#8217;s expected to receive a series order from AMC o
span></p>
<p>The spin-off is designed as a one-hour prequel that will chronicle the exploits of the show&#821
n is &#8220;Better Call Saul,&#8221; before he crossed paths with meth lord Walter White (Bran Crans
r Vince Gilligan has been developing the concept with co-executive producer Peter Gould.</p>
<p>AMC airs the series finale of <em>Breaking Bad</em> on Sept. 29.</p>
<div id="jp-post-flair" class="sharedaddy sd-like-enabled"></div>
<div style="clear: both;"></div>
```

Figure 3: File Before Removing Markup

```
In news that will please soon-to-be-in-mourning Breaking Bad fans, AMC
is greenlighting a spin-off series that will feature Breaking Bad
character Saul Goodman, played by Bob Odenkirk. While the network and
studio, Sony Pictures Television, announced today only that they have
reached a licensing agreement for the show, it's expected to receive a
series order from AMC once contracts are finalized.

The spin-off is designed as a one-hour prequel that will chronicle the
exploits of the show's charmingly disreputable lawyer, whose famous
slogan is "Better Call Saul," before he crossed paths with meth lord
Walter White (Bran Cranston). Breaking Bad creator/executive producer
Vince Gilligan has been developing the concept with co-executive
producer Peter Gould.
```

Figure 4: File After Removing Markup

```
efe466ec27581b9bd9170e21c69409a5 http://www.usa.gov/Contact-Us.shtml
77b351387f31b3581e0deb37e5402df9 http://www.youtube.com/watch?v=2FanNsV-sYE
cfebda397f7202c04620ea4c19f43268 https://itunes.apple.com/us/podcast/the-hackers/id523121474?i=163899042
d48d542e25b35e643fc13282829068ed http://voices.suntimes.com/sports/sports-prose/the-steep-price-of-attending-an-nfl-game/
bf83038020cd89efe8b55baba1be753f http://instagram.com/p/eK-s8ELSmi/
71abf6948a04623d55447361028999bf http://www.washingtonpost.com/lifestyle/style/miss-america-2014/2013/09/12/7b321b84-1ba8-11e
9e6e68dc77a77c6b923972e5ac3effd9 http://www.chicagotribune.com/news/local/breaking/chi-chicago-crime-gun-violence-shootings-a
0f507ecad375701139e674a166d78708 http://popwatch.ew.com/2013/09/13/tales-from-beyond-the-pale-larry-fessenden/
cf3a4d7f00f46305cf0b715649b3e372 http://www.chicagotribune.com/business/breaking/chi-chicago-foreclosures-20130912,0,6180109.
38e51867fac9bbca4f27ddaa093cb39d http://stackoverflow.com/users/4/joel-spolsky
2727c07bc37afb1a2b96a2fc874d27a3 https://my.barackobama.com/page/s/action-august-gun-violence-prevention/?source=socnet_20130
e97f3ec15ed7c65ec0d88c02f535e375 http://www.usa.gov/Citizen/Topics/Family-Issues/Vital-Docs.shtml
1172bc7e37337668448e0741abbab8aa http://www.chicagotribune.com/news/local/breaking/chi-cops-man-told-victim-ill-be-back-with-
1df894c5e490d0f58282b8ecf7636cf9 http://www.chicagotribune.com/sports/hockey/blackhawks/chi-chicago-blackhawks-stan-bowman-20
e21d7e6fef8ba78d5a47a75f83998d2a http://www.vtnews.vt.edu/articles/2013/03/032113-outreach-icorpsgrant.html?utm_campaign=Argy
```

Figure 5: Finding Aid

Question 2

Chose a query term (e.g., "shadow") that is not a not a stop words (see week 4 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. (If you do not end up with a list of 10 URIs, you've done something wrong).

As per the example in the week 4 slides, computer the TFIDF values for the term in each of the 10 documents and create a table with the TF, IDF, and TFIDF values, as well as the corresponding URIs. The URIs will be ranked in decreasing order by the TFIDF values. For example:

Table 1: 10 Hits for the term "shadow", ranked by TFIDF

TFIDF	TF	IDF	URI
0.150	0.014	10.680	http://foo.com
0.085	0.008	10.680	http://bar.com

You can use Google or Bing for the DF estimation. To count the number of words in the processed document (i.e., the denominator for TF), you can use "wc":

% wc -w www.cnn.com.processed

2370 www.cnn.com.processed

It won't be completely accurate, but it will probably be consistently inaccurate across all files. You can use more accurate methods if you'd like.

Don't forget the log base 2 for IDF, and mind your significant digits!

Answer to Question 2

I used the term "budget" since that topic is currently affecting over half of the income coming into my household right now. To determine the term frequency, I needed to compare the number of times the term appeared in the document divided by the number of words total in the document. But first to find documents with the term. To determine the goodness of using "budget", I used grep to determine how many documents contained the term.

```
grep -l budget * > ../q2/pagescontainbudget
```

This produced a file that contained 92 results so "budget" turned out to be a good term to use. Table 2 shows the results of finding the term "budget" in each file using grep, using wc -w for each file to get total word count and dividing the former by the latter.

I decided to use Google in the DF estimation. I looked up "how many documents has google indexed" in Google (Figure 6) and chose the result for World Wide Web Size `http://www.worldwidewebsize.com/`. From Figure 7, you can see that it is estimated at 44 billion web pages. Furthermore, searching Google for the term "budget" yields 636,000,000 web pages (Figure 8).

```
Total pages in corpus = 44,000,000,000
Total pages in corpus with the term "budget" = 636,000,000
Total pages in corpus / total pages with the term = 69.1824
```

$IDF(budget) = \log_2 69.1824 = 6.1123$

Table 3 shows the results of calculations from Table 2 combined with the IDF calculations above.

Table 2: Term Frequency Calculation for "budget", ranked by TF

URI	Term Count	Word Count	TF
http://www.usa.gov/Citizen/Topics/Health/Food.shtml#Eating_on_a_Budget	4	1315	0.0030
http://www.vtnews.vt.edu/articles/2013/06/060313-bov-overview.html	5	1665	0.0030
http://thehill.com/blogs/floor-action/senate/295759-reid-proposes-new-background-check-requirement-for-explosives	5	2054	0.0024
http://news.harvard.edu/gazette/story/2013/09/managing-a-seismic-shift/	6	2499	0.0024
http://www.huffingtonpost.com/2013/08/14/sequestration-cuts_n_3749432.html	21	11950	0.0018
http://www.washingtonpost.com/politics/from-newtown-to-navy-yard-unpredictable-calamities-upend-obamas-second-term/2013/09/16/3df366a6-1f04-11e3-8459-657e0c72fec8_story.html	5	3511	0.0014
http://www.tampabay.com/blogs/media/eric-deggans-to-leave-tampa-bay-times-for-job-as-nprs-first-tv-critic/2134332	4	3985	0.0010
http://maddowblog.msnbc.com/_news/2013/07/12/19436633-watching-marco-rubio-go-around-the-bend?lite	5	6017	0.0008
http://www.washingtoncitypaper.com/articles/44734/shadow-of-a-doubt-dc-statehood-activists/	7	9556	0.0007
http://www.huffingtonpost.com/2013/08/19/head-start-cuts-services_n_3779210.html?1376925983	4	6053	0.0007

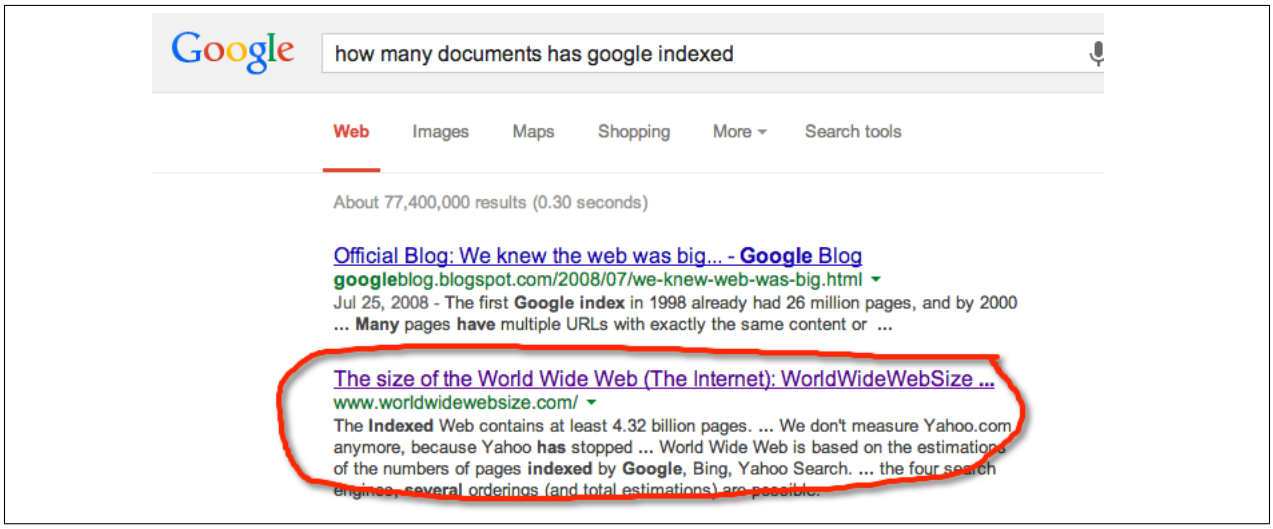

Figure 6: Finding the Size of Google's Collection

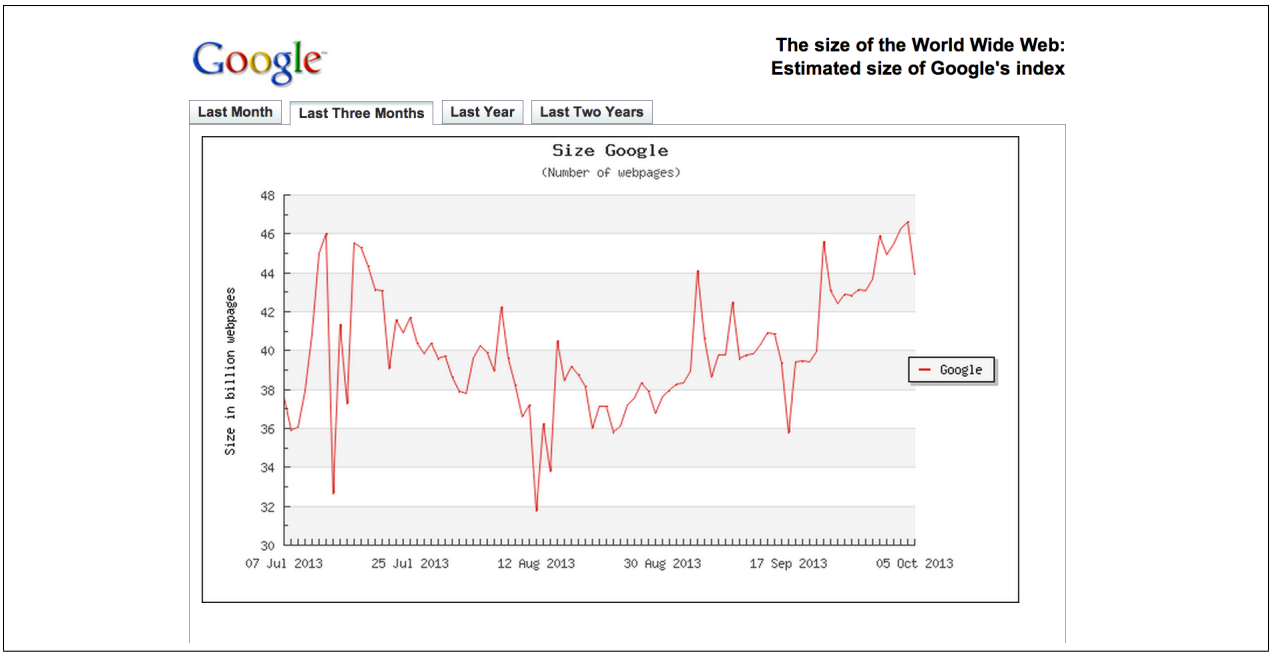

Figure 7: Size of Google's Collection

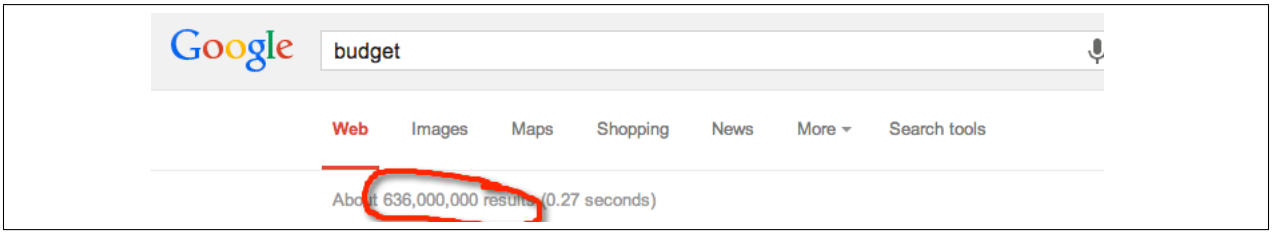

Figure 8: Pages with the Term "Budget"

Table 3: 10 Hits for the term "budget", ranked by TFIDF

TFIDF	TF	IDF	URI
0.0186	0.0030	6.1123	http://www.usa.gov/Citizen/Topics/Health/Food.shtml#Eating_on_a_Budget
0.0184	0.0030	6.1123	http://www.vtnews.vt.edu/articles/2013/06/060313-bov-overview.html
0.0149	0.0024	6.1123	http://thehill.com/blogs/floor-action/senate/295759-reid-proposes-new-background-check-requirement-for-explosives
0.0147	0.0024	6.1123	http://news.harvard.edu/gazette/story/2013/09/managing-a-seismic-shift
0.0107	0.0018	6.1123	http://www.huffingtonpost.com/2013/08/14/sequestration-cuts_n_3749432.html
0.0087	0.0014	6.1123	http://www.washingtonpost.com/politics/from-newtown-to-navy-yard-unpredictable-calamities-upend-obamas-second-term/2013/09/16/3df366a6-1f04-11e3-8459-657e0c72fec8_story.html
0.0061	0.0010	6.1123	http://www.tampabay.com/blogs/media/eric-deggans-to-leave-tampa-bay-times-for-job-as-nprs-first-tv-critic/2134332
0.0051	0.0008	6.1123	http://maddowblog.msnbc.com/_news/2013/07/12/19436633-watching-marco-rubio-go-around-the-bend?lite
0.0045	0.0007	6.1123	http://www.washingtoncitypaper.com/articles/44734/shadow-of-a-doubt-dc-statehood-activists/
0.0040	0.0007	6.1123	http://www.huffingtonpost.com/2013/08/19/head-start-cuts-services_n_3779210.html?1376925983

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, such as:

http://www.prchecker.info/check_page_rank.php

http://www.seocentro.com/tools/search-engines/pagerank.html

http://www.checkpagerank.net

If you use these tools, you'll have to do so by hand (they have anti-bot captchas), but there is only 10. Normalize the values they give you to be from 0 to 1.0. Use the same tool on all 10 (again, consistency is more important than accuracy).

Create a table similar to Table 1.

Table 4: 10 Hits for the term "shadow", ranked by PageRank

PageRank	URI
0.9	`http://foo.html`
0.5	`http://bar.html`

Answer to Question 3

I chose the first PageRank checker on the list `http://www.prchecker.info/check_page_rank.php`. Only one of my pages had a rank - a government web site (Figure 9). It was disappointing to not see a variety of ranks. Nine URIs got an N/A as a result (Figure 10). The site offered the following reasons for why PageRank was not available:

1. The web page is new and not yet indexed
2. The web page is indexed but not yet ranked
3. The web page is indexed by Google, but recognized as a supplemental result
4. The web page or whole web site is banned by Google

I think #1 and #2 are more likely reasons than #3 and #4 for why PageRank is not available for these pages. Of the ones with no rank, three are from September, three are from August, and the remaining ones are from July or earlier. These dates are not surprising since these URIs were extracted from recent tweets and tweeters are likely to reference more recent material. The age of these URIs should be enough time for them to be indexed, but not necessarily ranked. There does not appear to be a relationship between having a PageRank or even having a higher PageRank and the TFIDF value. Even though the usa.gov web page has the highest TFIDF and is the only one with a PageRank value, the next highest TFIDF value is not significantly lower than the usa.gov page, but does not have PageRank available.

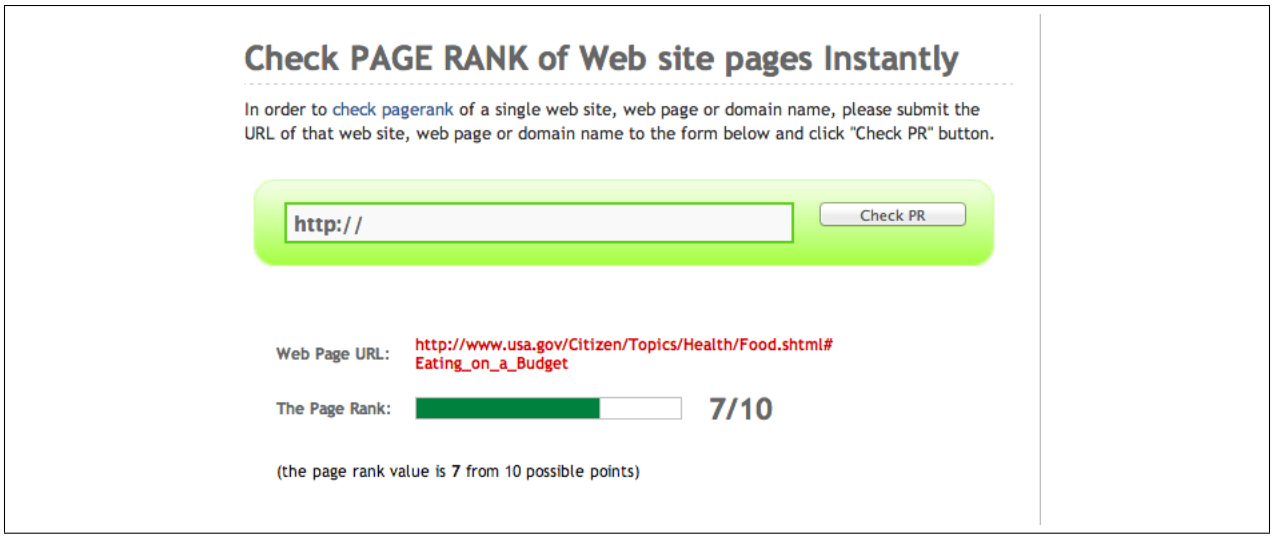

Figure 9: Successful PageRank Obtained

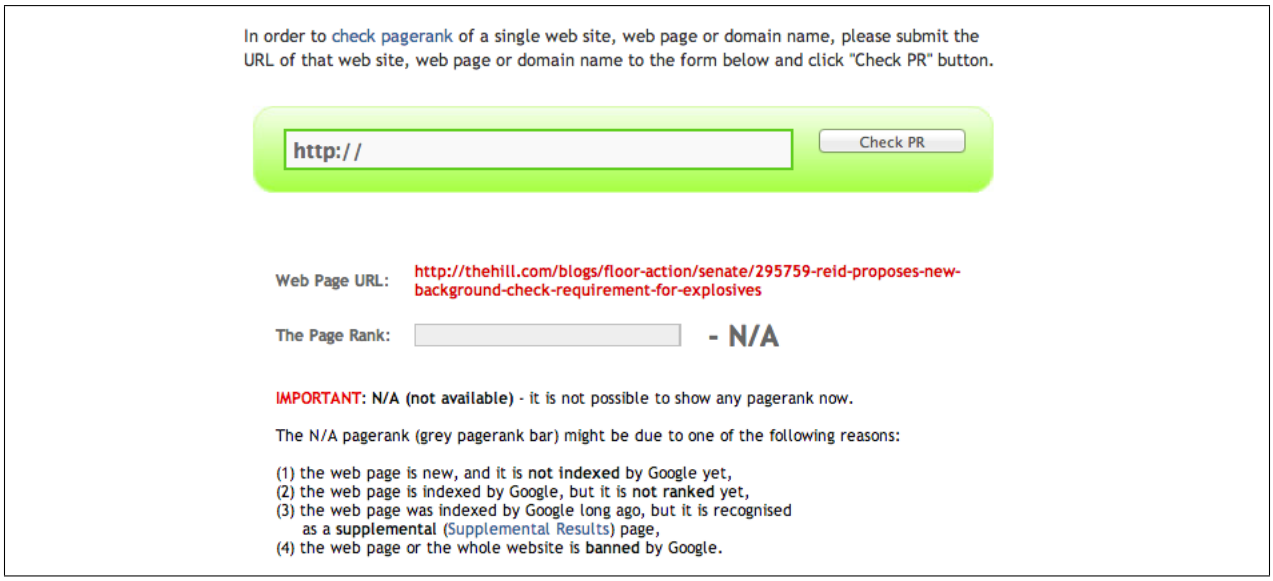

Figure 10: No PageRank Obtained

Table 5: 10 Hits for the term "budget", ranked by PageRank

PageRank	URI
0.7	http://www.usa.gov/Citizen/Topics/Health/Food.shtml#Eating_on_a_Budget
0.0	http://www.vtnews.vt.edu/articles/2013/06/060313-bov-overview.html
0.0	http://thehill.com/blogs/floor-action/senate/295759-reid-proposes-new-background-check-requirement-for-explosives
0.0	http://news.harvard.edu/gazette/story/2013/09/managing-a-seismic-shift
0.0	http://www.huffingtonpost.com/2013/08/14/sequestration-cuts_n_3749432.html
0.0	http://www.washingtonpost.com/politics/from-newtown-to-navy-yard-unpredictable-calamities-upend-obamas-second-term/2013/09/16/3df366a6-1f04-11e3-8459-657e0c72fec8_story.html
0.0	http://www.tampabay.com/blogs/media/eric-deggans-to-leave-tampa-bay-times-for-job-as-nprs-first-tv-critic/2134332
0.0	http://maddowblog.msnbc.com/_news/2013/07/12/19436633-watching-marco-rubio-go-around-the-bend?lite
0.0	http://www.washingtoncitypaper.com/articles/44734/shadow-of-a-doubt-dc-statehood-activists/
0.0	http://www.huffingtonpost.com/2013/08/19/head-start-cuts-services_n_3779210.html?1376925983

Question 4 - Extra Credit

Compute the Kendall Tau_b score for both lists (use "b" because there will likely be tie values in the rankings). Report both the Tau value and the "p" value.

See:
http://stackoverflow.com/questions/2557863/measures-of-association-in-r-kendalls-tau-b-and-tau-c
http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient#Tau-b
http://en.wikipedia.org/wiki/Correlation_and_dependence

Answer to Question 4

Not attempted.