Question 1.

To obtain the html content of the links, I first used a script to extract all the redirected urls which were stored in my json file which was created during the last assignment. This script is called extract_links.py and it generated a text file containing all the redirected links.

Then the html was retrieved from each link using the curl command. A script was created which would execute the curl command and then pipe the output into a file called source_num.txt where num is a number of the text file. A json file called linksToFile.json, maintains a record of links and the source file names in which their html mark-up is located. The same process was done when using the lynx –dump –force_html command.

Question 2.

To compute the TF-IDF, the query term "Trump" was used. To determine which pages contained the word Trump the command "grep Trump *.processed.txt" was executed. More than twenty documents contained the word Trump, so the first ten were kept. In order to count the number of times my query term appeared in each document, I used the search feature in my text editor and recorded the occurrence for the number of times each term appeared. Then I used the command "wc –w documentName" to get the total number of words in the document.

In computing the IDF, the number of documents in the corpus was 4.84 billion which was obtained from http://www.worldwidewebsize.com/ and the docs with term was retrieved from doing a bing search which amounted to approximately 14,800,000. Table 1 shows the results of the ranking for the TF-IDF Values.

This Space Intentionally Left Blank

TF-IDF	TF	IDF	URI
0.1303	0.0156	8.35	http://buchanan.org/blog/124610-124610
0.0823	0.0099	8.35	http://www.breitbart.com/big-government/2016/02/08/women-victims-of-bill-and-hillary-to-hit-campaign-trail/
0.0668	0.0080	8.35	http://www.breitbart.com/london/2016/02/08/20-cars-torched-in-weekend-of-leftist-violence-in-berlin/
0.0400	0.0048	8.35	http://www.commercialappeal.com/news/government/trump-supporters-expect-memphis-rally-soon-2b45b4fc-329c-6d6c-e053-0100007f8ee9-368053271.html
0.0275	0.0033	8.35	https://www.youtube.com/watch?v=jdnM39OABsM
0.0242	0.0029	8.35	https://www.washingtonpost.com/news/the-fix/wp/2016/02/06/marco-rubio-had-a-rough-night-on-google-but-jeb-bushs-was-worse/
0.0075	0.0009	8.35	http://www.infobae.com/america
0.0075	0.0009	8.35	http://www.breitbart.com/london/2016/01/29/migrant-crisis-live-wire/
0.0067	0.000	8.35	http://www.dailymail.co.uk/news/article-3437211/Will-rivals-boot-Jeb-Bush-wears-shoe-held-sticky-tape-campaign-town-hall-meeting.html?ito=social-twitter_dailymailus#ixzz3zaRSoadQ
0.0050	0.0006	8.35	https://www.pinknews.co.uk/2016/02/07/gay-ukip-member-quits-party-over-vile-nasty-homophobic-candidate/

Table 1: 10 Hits for the term "Trump" ranked by TF-IDF

This Space Intentionally Left Blank

Question 3.

The ranking below was obtained from getting the page rank using a shortened URI.

Page Rank	URI	Shortened URI If Needed
0.9	https://www.youtube.com/watch?v=jdnM39OABsM	https://www.youtube.com
0.8	https://www.washingtonpost.com/news/the-fix/wp/2016/02/06/marco-rubio-had-a-rough-night-on-google-but-jeb-bushs-was-worse/	https://www.washingtonpost.com
0.7	http://www.dailymail.co.uk/news/article3437211/Will-rivals-boot-Jeb-Bush-wears-shoe-held-sticky-tape-campaign-town-hall-meeting.html?ito=social-twitter_dailymailus#ixzz3zaRSoadQ	http://www.dailymail.co.uk
0.6	http://www.breitbart.com/london/2016/02/08/20-cars-torched-in-weekend-of-leftist-violence-in-berlin/	http://www.breitbart.com
0.6	www.infobae.com/america	N/A
0.6	http://www.breitbart.com/london/2016/01/29/migrant-crisis-live-wire/	http://www.breitbart.com/
0.5	http://buchanan.org/blog /124610-124610	http://buchanan.org/blog
0.4	http://www.breitbart.com/big-government/2016/02/08/women-victims-of-bill-and-hillary-to-hit-campaign-trail/	http://www.breitbart.com/big-government/
0.4	http://www.commercialappeal.com/news/government/trump-supporters-expect-memphis-rally-soon2b45b4fc-329c-6d6c-e053-010000f8ee9-368053271.html	http://www.commercialappeal.com/news
0.0	https://www.pinknews.co.uk/2016/02/07/gay-ukip-member-quits-party-over-vile-nasty-homophobic-candidate/	N/A

Table 2: 10 Hits for the term "Trump", ranked by page rank, obtained from http://www.seocentro.com/tools/search-engines/pagerank.html

Compare and Contrast

As discussed in class, the ranking obtained from page rank differed from the ranking using the TF-IDF values. In the ranking obtained from page-rank, many of the values were estimated due to the fact that the URI used to obtain the page-rank value had to be shortened. It should also be noted that in the TF-IDF ranking, the first URI had an unusually high TF-Value. This could be an indicator of spam in the web content. Lastly, both rankings ranked the same uri as the least important which demonstrates that it is possible for certain uris to share the same rank.

Extra Credit

To compute the Kendall Tau value, the following python code was found on http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.kendalltau.html

```
import scipy.stats as stats

x1 = [12, 2, 1, 12, 2]

x2 = [1, 4, 7, 1, 0]

tau, p_value = stats.kendalltau(x1, x2)

tau

p_value
```

What I changed was the content of the arrays x1 and x2 as shown below.

```
import scipy.stats as stats

x1 = [0.1303,0.0823,0.0668,0.0400,0.0275,0.0242,0.0075, 0.0075,0.0067,0.0050]

x2 = [0.9,0.8,0.7,0.6,0.6,0.6,0.5,0.4,0.4,0.0]

tau, p_value = stats.kendalltau(x1, x2)

tau

p_value
```

This resulted in a p_value of `0.000150 and a tau value of 0.941763.`