**Assignment 3** **Hussam Hallak**
CS532, Web Science, Spring 2017 CS Master's Student
Old Dominion University, Computer Science Dept Prof: Dr. Nelson

## 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.

from the command line:

Listing 1: Command:

```
% 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 URIs (e.g., "?", "&"). You might want to hash the URIs, like: from the command line:

Listing 2: Command:

```
% 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 newline.)

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

Listing 3: Command:

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

Use another (better) tool if you know of one. A "better" approach is to use BeautifulSoup, see:

http://stackoverflow.com/questions/1936466/beautifulsoup-grab-visible-webpage-text for some hints on how to start. Note that none of these methods are going to be perfect.

Keep both files for each URI (i.e., raw HTML and processed). Upload both sets of files to your github account.

### Answer:

The approach is divided into two steps:

1. Take the unique links collected from Assignment 2, saved in "uniquelinks.txt" and download their content, then save it to text files, one file for each link. For that purpose I used md5 hashing to name the output files.

2. Remove (most) of the HTML markup from the content of the downloaded HTML pages.

I wrote a python script that will combine both steps. First it will download the raw HTML for each link and save it to a file, then it will strip HTML markup and save it

to another file with the same name postfixed with ".processed". The program will also create a text file named "map.txt" to map each link to its new file name. That will make the work easier for Question 2.

Listing 4: The content of parselinks.py

```
import sys
import subprocess
from bs4 import *
import urllib2
import re

if len(sys.argv) != 2:
   print "Usage: Python parselinks.py <file_name>"
   print "e.g: Python parselinks.py uniquelinks.txt"
   exit()
def visible(element):
   if element.parent.name in ['style', 'script', '[document]', 'head', 'title']:
      return False
   elif re.match(r"[\s\r\n]+",unicode(element)):
      return False
   return True

fh_input = open(sys.argv[1], 'r')
fh_map = open("map.txt", 'w')
for link in fh_input:
   try:
      cmd = 'echo -n "' + link + '" | md5sum'
      output = subprocess.check_output(cmd, shell=True)
      output_file_name = output[0:32]
      html_page = urllib2.urlopen(link)
      soup = BeautifulSoup(html_page, "html.parser")
      texts = soup.findAll(text=True)
      output_texts = str(texts)
      fh_output = open(output_file_name, 'w')
      fh_output.write(output_texts)
      fh_output.close()
      output_file_name_processed = output_file_name + '.processed'
      fh_output_processed = open(output_file_name_processed, 'w')
      visible_texts = filter(visible, texts)
      output_texts = str (visible_texts)
      fh_output_processed.write(output_texts)
      fh_output_processed.close()
      fh_map.write(link)
      fh_map.write('\t')
      fh_map.write(output_file_name_processed)
      fh_map.write('\n')
   except:
      print "This link generated an error code:"
      print link
fh_input.close()
fh_map.close()
```

Listing 5: Running parselinks.py to download raw HTML and visible text from URIs

```
root@ima-app:/var/www/Hussam/A3# python parselinks.py uniquelinks.txt
root@ima-app:/var/www/Hussam/A3# ls | wc -w
798
```

The program created 794 output files, two files for each link. This means that the content of 397 URIs was downloaded and processed successfully. The rest of the links generated errors. That could be due to different reasons including bad HTML markup, the page no longer exists on the web, or other reasons.

### Included Files:

parselinks.py, uniquelinks.txt, map.txt
Folder "raw" contains all raw HTML files
Folder "processed" contains all visible text in 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. (If you do not end up with a list of 10 URIs, you've done something wrong).

As per the example in the week 5 slides, compute 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 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/
0.044 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 deonminator for TF), you can use "wc":

2370 www.cnn.com.processed

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

Don't forget the log base 2 for IDF, and mind your significant digits!
`https://en.wikipedia.org/wiki/Significant_figures#Rounding_and_decimal_places`

### Answer:

I chose the query term "blue" and that matched exactly 10 documents from my output file in Question 1.

Listing 6: Results from searching for the chosen query term

```
root@ima-app:/var/www/Hussam/A3# grep -il "blue" *.processed
2abb9ccc9ac936b183b0ba3514bb246a.processed
4aeb25630e7cc76a102b5f35a5bee8af.processed
```

```
7a83161b2ccc2c6d0c5c74dcfc247164.processed
866df289fc6c7d73ddcbaf35199a358d.processed
8e72632b606ebcb16f1f83a1f248bcea.processed
913f142ab3ab5704e42b195c50116d9c.processed
e1577b08960791988e2ccb0315148f30.processed
f158b3a1dabb47a8a42fe06cd86444c0.processed
fc43d6cca315eda1428c7b4fdaf8cf77.processed
fe93efbb7ec52521f2a32784a5563385.processed
root@ima-app:/var/www/Hussam/A3# grep -il "germ" *.processed | wc -l
10
root@ima-app:/var/www/Hussam/A3#
```

I created a new directory named "blue" and copied all processed documents that have the query term to that folder.

Listing 7: Copying the documents containing the query term to a separate folder

```
root@ima-app:/var/www/Hussam/A3# mkdir blue
root@ima-app:/var/www/Hussam/A3# cp `grep -il "blue" *.processed` blue
root@ima-app:/var/www/Hussam/A3# cd blue
root@ima-app:/var/www/Hussam/A3/blue# ls
2abb9ccc9ac936b183b0ba3514bb246a.processed
4aeb25630e7cc76a102b5f35a5bee8af.processed
7a83161b2ccc2c6d0c5c74dcfc247164.processed
866df289fc6c7d73ddcbaf35199a358d.processed
8e72632b606ebcb16f1f83a1f248bcea.processed
913f142ab3ab5704e42b195c50116d9c.processed
e1577b08960791988e2ccb0315148f30.processed
f158b3a1dabb47a8a42fe06cd86444c0.processed
fc43d6cca315eda1428c7b4fdaf8cf77.processed
fe93efbb7ec52521f2a32784a5563385.processed
root@ima-app:/var/www/Hussam/A3/blue#
```

It's time to calculate TFIDF, TF, IDF for the query term "blue" and construct the requested table. I will use the file map.txt, created in question 1 to find out which processed file belongs to which URI.

Computing TF:

Listing 8: Computing TF:

```
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue" 2
    abb9ccc9ac936b183b0ba3514bb246a.processed | wc -w
2
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue" 4
    aeb25630e7cc76a102b5f35a5bee8af.processed | wc -w
1
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue" 7
    a83161b2ccc2c6d0c5c74dcfc247164.processed | wc -w
1
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue" 866
    df289fc6c7d73ddcbaf35199a358d.processed | wc -w
4
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue" 8
    e72632b606ebcb16f1f83a1f248bcea.processed | wc -w
1
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue" 913
```

```
    f142ab3ab5704e42b195c50116d9c.processed | wc -w
1
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue"
    e1577b08960791988e2ccb0315148f30.processed | wc -w
1
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue"
    f158b3a1dabb47a8a42fe06cd86444c0.processed | wc -w
2
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue"
    fc43d6cca315eda1428c7b4fdaf8cf77.processed | wc -w
1
root@ima-app:/var/www/Hussam/A3/blue# grep -io "blue"
    fe93efbb7ec52521f2a32784a5563385.processed | wc -w
2
```

Normalizing TF:

TF is normalized by dividing the number of occurrences of "blue" divided by the total number of words in the document.

Listing 9: Finding the total number of words in the documents:

```
root@ima-app:/var/www/Hussam/A3/blue# wc -w 2abb9ccc9ac936b183b0ba3514bb246a.
    processed
1563 2abb9ccc9ac936b183b0ba3514bb246a.processed
root@ima-app:/var/www/Hussam/A3/blue# wc -w 4aeb25630e7cc76a102b5f35a5bee8af.
    processed
822 4aeb25630e7cc76a102b5f35a5bee8af.processed
root@ima-app:/var/www/Hussam/A3/blue# wc -w 7a83161b2ccc2c6d0c5c74dcfc247164.
    processed
317 7a83161b2ccc2c6d0c5c74dcfc247164.processed
root@ima-app:/var/www/Hussam/A3/blue# wc -w 866df289fc6c7d73ddcbaf35199a358d.
    processed
1921 866df289fc6c7d73ddcbaf35199a358d.processed
root@ima-app:/var/www/Hussam/A3/blue# wc -w 8e72632b606ebcb16f1f83a1f248bcea.
    processed
2214 8e72632b606ebcb16f1f83a1f248bcea.processed
root@ima-app:/var/www/Hussam/A3/blue# wc -w 913f142ab3ab5704e42b195c50116d9c.
    processed
2060 913f142ab3ab5704e42b195c50116d9c.processed
root@ima-app:/var/www/Hussam/A3/blue# wc -w e1577b08960791988e2ccb0315148f30.
    processed
572 e1577b08960791988e2ccb0315148f30.processed
root@ima-app:/var/www/Hussam/A3/blue# wc -w f158b3a1dabb47a8a42fe06cd86444c0.
    processed
416 f158b3a1dabb47a8a42fe06cd86444c0.processed
root@ima-app:/var/www/Hussam/A3/blue# wc -w fc43d6cca315eda1428c7b4fdaf8cf77.
    processed
1431 fc43d6cca315eda1428c7b4fdaf8cf77.processed
root@ima-app:/var/www/Hussam/A3/blue# wc -w fe93efbb7ec52521f2a32784a5563385.
    processed
1049 fe93efbb7ec52521f2a32784a5563385.processed
```

Searching the file name of ".processed" files in "map.txt" will give us the corresponding URI.

Listing 10: Matching file names to URIs:

```
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "2
    abb9ccc9ac936b183b0ba3514bb246a.processed" map.txt
http://www.boohoo.com/
        2abb9ccc9ac936b183b0ba3514bb246a.processed
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "4
    aeb25630e7cc76a102b5f35a5bee8af.processed" map.txt
http://www.vox.com/2016/9/7/12815076/america-uninsured-rate-dropped?source=
    socnet_tw_aca_20160907_bo_uninsured-rate-low_vox_3&utm_medium=social&
    utm_source=tw_bo&utm_campaign=socnet_tw_aca_20160907_bo_uninsured-rate-
    low_vox_3&utm_content=20160907_bo_uninsured-rate-low_vox_3
        4aeb25630e7cc76a102b5f35a5bee8af.processed
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "7
    a83161b2ccc2c6d0c5c74dcfc247164.processed" map.txt
http://www.vox.com/new-money/2016/10/21/13361414/median-wages?source=
    socnet_tw_econ_20161026_bo_wage-increase_durable_4&utm_medium=social&
    utm_source=tw_bo&utm_campaign=socnet_tw_econ_20161026_bo_wage-
    increase_durable_4&utm_content=20161026_bo_wage-increase_durable_4
        7a83161b2ccc2c6d0c5c74dcfc247164.processed
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "866
    df289fc6c7d73ddcbaf35199a358d.processed" map.txt
https://www.fisherwallace.com/?gclid=CNr-xNqt38UCFUU8gQodQHIAsQ
        866df289fc6c7d73ddcbaf35199a358d.processed
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "8
    e72632b606ebcb16f1f83a1f248bcea.processed" map.txt
http://www.nme.com/features/taylor-swift-power-fame-and-the-future-the-full-nme-
    cover-interview-549
        8e72632b606ebcb16f1f83a1f248bcea.processed
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "913
    f142ab3ab5704e42b195c50116d9c.processed" map.txt
http://www.health.com/health/gallery/0,,20705881,00.html
        913f142ab3ab5704e42b195c50116d9c.processed
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "
    e1577b08960791988e2ccb0315148f30.processed" map.txt
http://www.mtv.com/vma/winners
        e1577b08960791988e2ccb0315148f30.processed
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "
    f158b3a1dabb47a8a42fe06cd86444c0.processed" map.txt
http://www.vox.com/energy-and-environment/2016/9/30/13111088/cleantech-costs-
    falling-one-chart?source=socnet_tw_cc_20161006_bo_clean-energy_technology_1&
    utm_medium=social&utm_source=tw_bo&utm_campaign=
    socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_content=20161006
    _bo_clean-energy_technology_1
        f158b3a1dabb47a8a42fe06cd86444c0.processed
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "
    fc43d6cca315eda1428c7b4fdaf8cf77.processed" map.txt
https://www.washingtonpost.com/news/energy-environment/wp/2016/10/10/climate-
    change-has-been-making-western-forest-fires-worse-for-decades-study-says/?
    source=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_medium=
    social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161027_bo_energy-
    environment_climate_5&utm_content=20161027_bo_energy-environment_climate_5
        fc43d6cca315eda1428c7b4fdaf8cf77.processed
```

```
root@ima-app:/var/www/Hussam/A3# grep --after-context=0 --before-context=1 "
    fe93efbb7ec52521f2a32784a5563385.processed" map.txt
http://www.npr.org/sections/thetwo-way/2016/08/31/492177267/obama-at-lake-tahoe-
    praises-conservation-efforts?source=socnet_tw_cc_20160906_bo_npr-
    article_conservation_2&utm_medium=socnet&utm_source=tw&utm_campaign=
    socnet_tw_cc_20160906_bo_npr-article_conservation_2&utm_content=20160906
    _bo_npr-article_conservation_2
       fe93efbb7ec52521f2a32784a5563385.processed
```

In the following table:
TF: The raw TF which is the number of occurrences of "blue" in a document.
WC: The total number of words in the document.
N-TF: The normalized TF for the document.
URI: The URI of the document.
Table 1:

| TF | WC | N-TF | URI |
|---|---|---|---|
| 2 | 1563 | 0.0013 | http://www.boohoo.com/ |
| 1 | 822 | 0.0012 | http://www.vox.com/2016/9/7/12815076/america-uninsured-rate-dropped?source=socnet_tw_aca_20160907_bo_uninsured-rate-low_vox_3&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_aca_20160907_bo_uninsured-rate-low_vox_3&utm_content=20160907_bo_uninsured-rate-low_vox_3 |
| 1 | 317 | 0.0032 | http://www.vox.com/new-money/2016/10/21/13361414/median-wages?source=socnet_tw_econ_20161026_bo_wage-increase_durable_4&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_econ_20161026_bo_wage-increase_durable_4&utm_content=20161026_bo_wage-increase_durable_4 |
| 4 | 1921 | 0.0021 | https://www.fisherwallace.com/?gclid=CNr-xNqt38UCFUU8gQodQHIAsQ |
| 1 | 2214 | 0.0005 | http://www.nme.com/features/taylor-swift-power-fame-and-the-future-the-full-nme-cover-interview-549 |

| | | | |
|---|---|---|---|
| 1 | 2060 | 0.0005 | http://www.health.com/health/gallery/0,,20705881,00.html |
| 1 | 572 | 0.0017 | http://www.mtv.com/vma/winners |
| 2 | 416 | 0.0048 | http://www.vox.com/energy-and-environment/2016/9/30/13111088/cleantech-costs-falling-one-chart?source=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_content=20161006_bo_clean-energy_technology_1 |
| 1 | 1431 | 0.0007 | https://www.washingtonpost.com/news/energy-environment/wp/2016/10/10/climate-change-has-been-making-western-forest-fires-worse-for-decades-study-says/?source=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_content=20161027_bo_energy-environment_climate_5 |
| 2 | 1049 | 0.0019 | http://www.npr.org/sections/thetwo-way/2016/08/31/492177267/obama-at-lake-tahoe-praises-conservation-efforts?source=socnet_tw_cc_20160906_bo_npr-article_conservation_2&utm_medium=socnet&utm_source=tw&utm_campaign=socnet_tw_cc_20160906_bo_npr-article_conservation_2&utm_content=20160906_bo_npr-article_conservation_2 |

Calculating IDF: According to http://www.worldwidewebsize.com/ the total number of indexed web pages in Google is 50 Billion. I searched for the word "blue" in Google and it returned About 15,520,000,000 results.

$$IDF(\text{"}blue\text{"}) = \log_2(\frac{50B}{15.5B}) = 1.6897$$

Now we can calculate TF-IDF for each document using this formula:

$$TF - IDF = (TF) \times (IDF)$$

Now we can generate the required table:
Table 2:

| TFIDF | TF | IDF | URI |
|---|---|---|---|
| 0.0008 | 0.0005 | 1.6897 | http://www.nme.com/features/taylor-swift-power-fame-and-the-future-the-full-nme-cover-interview-549 |
| 0.0008 | 0.0005 | 1.6897 | http://www.health.com/health/gallery/0,,20705881,00.html |
| 0.0012 | 0.0007 | 1.6897 | https://www.washingtonpost.com/news/energy-environment/wp/2016/10/10/climate-change-has-been-making-western-forest-fires-worse-for-decades-study-says/?source=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_content=20161027_bo_energy-environment_climate_5 |
| 0.0020 | 0.0012 | 1.6897 | http://www.vox.com/2016/9/7/12815076/america-uninsured-rate-dropped?source=socnet_tw_aca_20160907_bo_uninsured-rate-low_vox_3&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_aca_20160907_bo_uninsured-rate-low_vox_3&utm_content=20160907_bo_uninsured-rate-low_vox_3 |
| 0.0022 | 0.0013 | 1.6897 | http://www.boohoo.com/ |
| 0.0029 | 0.0017 | 1.6897 | http://www.mtv.com/vma/winners |

| | | | |
|---|---|---|---|
| 0.0032 | 0.0019 | 1.6897 | http://www.npr.org/sections/thetwo-way/2016/08/31/492177267/obama-at-lake-tahoe-praises-conservation-efforts?source=socnet_tw_cc_20160906_bo_npr-article_conservation_2&utm_medium=socnet&utm_source=tw&utm_campaign=socnet_tw_cc_20160906_bo_npr-article_conservation_2&utm_content=20160906_bo_npr-article_conservation_2 |
| 0.0035 | 0.0021 | 1.6897 | https://www.fisherwallace.com/?gclid=CNr-xNqt38UCFUU8gQodQHIAsQ |
| 0.0054 | 0.0032 | 1.6897 | http://www.vox.com/new-money/2016/10/21/13361414/median-wages?source=socnet_tw_econ_20161026_bo_wage-increase_durable_4&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_econ_20161026_bo_wage-increase_durable_4&utm_content=20161026_bo_wage-increase_durable_4 |
| 0.0081 | 0.0048 | 1.6897 | http://www.vox.com/energy-and-environment/2016/9/30/13111088/cleantech-costs-falling-one-chart?source=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_content=20161006_bo_clean-energy_technology_1 |

### Included Files:

map.txt
Folder "blue" contains all processed documents that contain the query term "blue".

## Question 3:

Now rank the same 10 URIs from question 2, but this time by their PageRank. Use any of the free PR estimaters on the web, such as:

`http://pr.eyedomain.com/`
`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 are only 10 to do. 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). Also note that these tools typically report on the domain rather than the page, so it's not entirely accurate.

Create a table similar to Table 1:

Table 2. 10 hits for the term "shadow", ranked by PageRank.

PR URI

---

0.9 http://bar.com/

0.5 http://foo.com/

Briefly compare and contrast the rankings produced in questions 2 and 3.

**Answer:**

Using `http://pr.eyedomain.com/` to find the Google page rank for each of the URIs, and normalizing the results, the required table becomes:

Table 3:

| Page Rank | URI |
|---|---|
| 0.3 | https://www.fisherwallace.com/?gclid=CNr-xNqt38UCFUU8gQodQHIAsQ |
| 0.5 | http://www.boohoo.com/ |
| 0.6 | http://www.vox.com/2016/9/7/12815076/america-uninsured-rate-dropped?source=socnet_tw_aca_20160907_bo_uninsured-rate-low_vox_3&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_aca_20160907_bo_uninsured-rate-low_vox_3&utm_content=20160907_bo_uninsured-rate-low_vox_3 |
| 0.6 | http://www.vox.com/new-money/2016/10/21/13361414/median-wages?source=socnet_tw_econ_20161026_bo_wage-increase_durable_4&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_econ_20161026_bo_wage-increase_durable_4&utm_content=20161026_bo_wage-increase_durable_4 |
| 0.6 | http://www.nme.com/features/taylor-swift-power-fame-and-the-future-the-full-nme-cover-interview-549 |

| | |
|---|---|
| 0.6 | http://www.vox.com/energy-and-environment/2016/9/30/13111088/cleantech-costs-falling-one-chart?source=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_content=20161006_bo_clean-energy_technology_1 |
| 0.7 | http://www.health.com/health/gallery/0,,20705881,00.html |
| 0.7 | http://www.mtv.com/vma/winners |
| 0.8 | https://www.washingtonpost.com/news/energy-environment/wp/2016/10/10/climate-change-has-been-making-western-forest-fires-worse-for-decades-study-says/?source=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_content=20161027_bo_energy-environment_climate_5 |
| 0.8 | http://www.npr.org/sections/thetwo-way/2016/08/31/492177267/obama-at-lake-tahoe-praises-conservation-efforts?source=socnet_tw_cc_20160906_bo_npr-article_conservation_2&utm_medium=socnet&utm_source=tw&utm_campaign=socnet_tw_cc_20160906_bo_npr-article_conservation_2&utm_content=20160906_bo_npr-article_conservation_2 |

Combining Table 2 and 3 generates the following table:
Table 4:

| PR | TFIDF | TF | IDF | URI |
|---|---|---|---|---|
| 0.5 | 0.0022 | 0.0013 | 1.6897 | http://www.boohoo.com/ |

| | | | | |
|---|---|---|---|---|
| 0.6 | 0.0020 | 0.0012 | 1.6897 | http://www.vox.com/2016/9/7/12815076/america-uninsured-rate-dropped?source=socnet_tw_aca_20160907_bo_uninsured-rate-low_vox_3&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_aca_20160907_bo_uninsured-rate-low_vox_3&utm_content=20160907_bo_uninsured-rate-low_vox_3 |
| 0.6 | 0.0054 | 0.0032 | 1.6897 | http://www.vox.com/new-money/2016/10/21/13361414/median-wages?source=socnet_tw_econ_20161026_bo_wage-increase_durable_4&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_econ_20161026_bo_wage-increase_durable_4&utm_content=20161026_bo_wage-increase_durable_4 |
| 0.3 | 0.0035 | 0.0021 | 1.6897 | https://www.fisherwallace.com/?gclid=CNr-xNqt38UCFUU8gQodQHIAsQ |
| 0.6 | 0.0008 | 0.0005 | 1.6897 | http://www.nme.com/features/taylor-swift-power-fame-and-the-future-the-full-nme-cover-interview-549 |
| 0.7 | 0.0008 | 0.0005 | 1.6897 | http://www.health.com/health/gallery/0,,20705881,00.html |
| 0.7 | 0.0029 | 0.0017 | 1.6897 | http://www.mtv.com/vma/winners |

| | | | | |
|---|---|---|---|---|
| 0.6 | 0.0081 | 0.0048 | 1.6897 | http://www.vox.com/energy-and-environment/2016/9/30/13111088/cleantech-costs-falling-one-chart?source=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_content=20161006_bo_clean-energy_technology_1 |
| 0.8 | 0.0012 | 0.0007 | 1.6897 | https://www.washingtonpost.com/news/energy-environment/wp/2016/10/10/climate-change-has-been-making-western-forest-fires-worse-for-decades-study-says/?source=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_content=20161027_bo_energy-environment_climate_5 |
| 0.8 | 0.0032 | 0.0019 | 1.6897 | http://www.npr.org/sections/thetwo-way/2016/08/31/492177267/obama-at-lake-tahoe-praises-conservation-efforts?source=socnet_tw_cc_20160906_bo_npr-article_conservation_2&utm_medium=socnet&utm_source=tw&utm_campaign=socnet_tw_cc_20160906_bo_npr-article_conservation_2&utm_content=20160906_bo_npr-article_conservation_2 |

From Table 4, we find that there is no obvious relationship between Google page rank and TFIDF for the document. Websites with higher page rank had both lower and higher TFIDF for their corresponding documents. It is also noticeable that various documents within the same domain have different TFIDF, high and low, but the same page rank (page rank is reported on the domain name rather than the document).

Example:

The URI:

http://www.boohoo.com/

Has a page rank 0.5, medium, and a TF-IDF 0.0022, low.

On the other hand, this URI:

https://www.fisherwallace.com/?gclid=CNr-xNqt38UCFUU8gQodQHIAsQ

Has a page rank 0.3, low, and a TF-IDF 0.0035, medium.
Also this URI:

https://www.washingtonpost.com/news/energy-environment/wp/2016/10/10/climate-change-has-been-making-western-forest-fires-worse-for-decades-study-says/?source=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161027_bo_energy-environment_climate_5&utm_content=20161027_bo_energy-environment_climate_5

Has a page rank 0.8, high, and a TF-IDF 0.0012, low.
Now this URI:

http://www.vox.com/energy-and-environment/2016/9/30/13111088/cleantech-costs-falling-one-chart?source=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_medium=social&utm_source=tw_bo&utm_campaign=socnet_tw_cc_20161006_bo_clean-energy_technology_1&utm_content=20161006_bo_clean-energy_technology_1

Has a page rank 0.6, medium, and a TF-IDF 0.0081, high.
**Conclusion:** Page rank and TF-IDF are not correlated.