

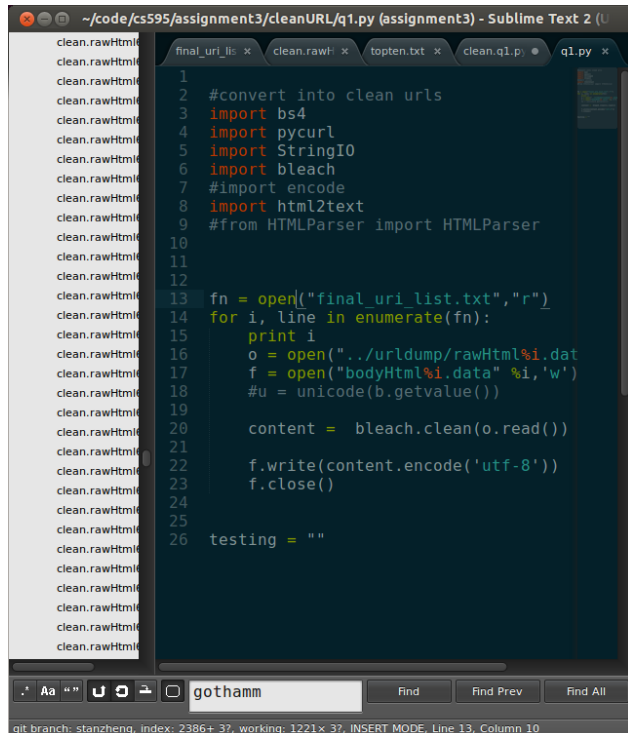
Question 1:

1. Download the 1000 URIs from assignment #2. .. using curl .. and then use something like lynx

Using PyCurl and Mozilla bleach to convert raw html to UTF-8 text.

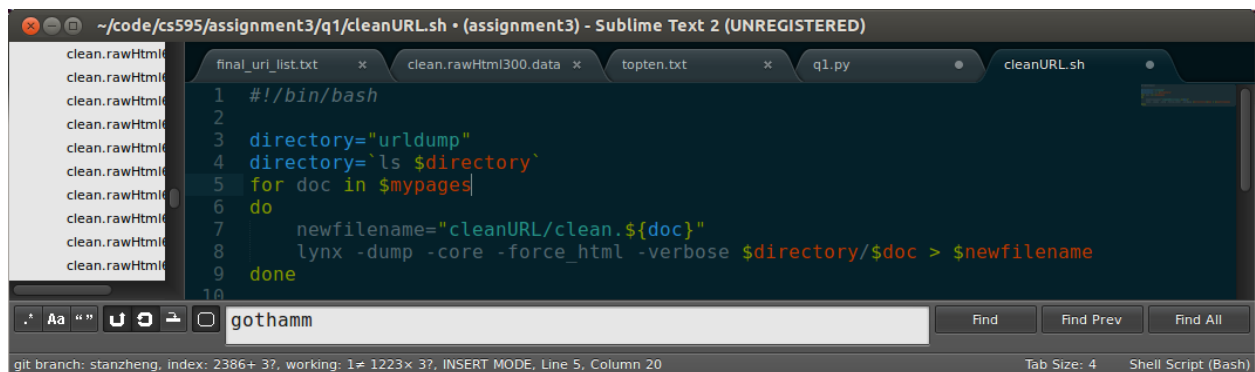
Part 1:

urldump/ is a collection of all the files from “final\_uri\_list.txt” curled, and piped into a file. The file naming scheme is “rawHTML” + the line the URI occurs on the list and “.data”



```
1 #convert into clean urls
2 import bs4
3 import pycurl
4 import StringIO
5 import bleach
6 import encode
7 import html2text
8 #from HTMLParser import HTMLParser
9
10
11
12
13 fn = open("final_uri_list.txt","r")
14 for i, line in enumerate(fn):
15     print i
16     o = open("../urldump/rawHtml%i.dat" % i, "w")
17     f = open("bodyHtml%i.data" % i, "w")
18     #u = unicode(b.getvalue())
19
20     content = bleach.clean(o.read())
21
22     f.write(content.encode('utf-8'))
23     f.close()
24
25
26 testing = ""
```

Part 2:



```
1 #!/bin/bash
2
3 directory="urldump"
4 directory=`ls $directory`
5 for doc in $mypages
6 do
7     newfilename="cleanURL/clean.${doc}"
8     lynx -dump -core -force_html -verbose $directory/$doc > $newfilename
9 done
10
```

Using lynx and a shell script, remove some of the html text. After cleaned each URL was preappended with “clean.” and saved to cleanURL/.

Question 2:

2. Choose a query term (e.g., "shadow") that is 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).

Chosen Query term: "learning" : many of the search queries are for various popular programming languages and frameworks that people use.

Code:

```
./find_term_program.sh > results.txt  
head -11 results.txt > topten_results.txt
```

**Size of Bing = 2.3 Billion**

**Number of hits for Learning = 230,000,000 results**

$TFIDF = TF \times IDF$

$IF = \text{total hits} / \text{total number of words}$

$IDF = \log_2(\text{total docs in corpus} / \text{docs with term})$

Assignment 3  
 CS 495  
 Stanley Zheng  
 October 5th 2013

IF-IDF	IF	IDF	URI
.00126	9/77430	10.83	<a href="http://www.amazon.com/Python-Cookbook-David-Beazley/dp/1449340377">http://www.amazon.com/Python-Cookbook-David-Beazley/dp/1449340377</a>
.00051	7/148193	10.83	<a href="http://hackaday.com/2013/09/23/guest-rant-ham-radio-hackers-paradise/">http://hackaday.com/2013/09/23/guest-rant-ham-radio-hackers-paradise/</a>
.00239	7/31659	10.83	<a href="http://code.org/">http://code.org/</a>
.00015	7/51655	10.83	<a href="http://www.wired.com/opinion/2013/09/ap_code/">http://www.wired.com/opinion/2013/09/ap_code/</a>
.00342	7/22118	10.83	<a href="http://www.vaporcouture.com">http://www.vaporcouture.com</a>
.00113	6/57314	10.83	<a href="http://www.ijava2.com/python-codingbat-answers-string-2-doublechar/">http://www.ijava2.com/python-codingbat-answers-string-2-doublechar/</a>
.00662	5/8176	10.83	<a href="http://www.brightsurf.com/news/headlines/89384/Late_Cretaceous_Period_was_likely_ice-free.html">http://www.brightsurf.com/news/headlines/89384/Late_Cretaceous_Period_was_likely_ice-free.html</a>
.00279	5/19348	10.83	<a href="https://managewp.com/wordpress-women-in-tech?utm_source=feedburner&amp;utm_medium=feed&amp;utm_campaign=Feed%3A+managewp+%28ManageWP+Blog%29">https://managewp.com/wordpress-women-in-tech?utm_source=feedburner&amp;utm_medium=feed&amp;utm_campaign=Feed%3A+managewp+%28ManageWP+Blog%29</a>
.00999	5/19172	10.83	<a href="http://dailymothering.com/learning-resources-jumbo-magnifiers-review-giveaway/">http://dailymothering.com/learning-resources-jumbo-magnifiers-review-giveaway/</a>

Question 3

PageRank	URI
.8	<a href="http://www.amazon.com">http://www.amazon.com</a>
.6	<a href="http://hackaday.com">http://hackaday.com</a>
.4	<a href="http://code.org/">http://code.org/</a>
.8	<a href="http://www.wired.com/">http://www.wired.com/</a>
.3	<a href="http://www.vaporcouture.com">http://www.vaporcouture.com</a>
.1	<a href="http://www.ijava2.com">http://www.ijava2.com</a>
.5	<a href="http://www.brightsurf.com">http://www.brightsurf.com</a>
.6	<a href="https://managewp.com/">https://managewp.com/</a>
.3	<a href="http://dailymothering.com">http://dailymothering.com</a>

Question 4

Used this calculator to find out that I have a  $-.377168$  Kendall Tau B Relationship. This shows there is negative correlation which means there is some likelihood these indicators do not correlate with each other.

<http://calculator-fx.com/calculator/statistics/kendall-tau-correlation>

IFIDF	PageRank	URI
.00126	.8	<a href="http://www.amazon.com">http://www.amazon.com</a>
.00051	.6	<a href="http://hackaday.com">http://hackaday.com</a>
.00239	.4	<a href="http://code.org/">http://code.org/</a>
.00015	.8	<a href="http://www.wired.com/">http://www.wired.com/</a>
.00342	.3	<a href="http://www.vaporcouture.com">http://www.vaporcouture.com</a>
.00113	.1	<a href="http://www.ijava2.com">http://www.ijava2.com</a>
.00662	.5	<a href="http://www.brightsurf.com">http://www.brightsurf.com</a>
.00279	.6	<a href="https://managewp.com/">https://managewp.com/</a>
.00999	.3	<a href="http://dailymothering.com">http://dailymothering.com</a>

Assignment 3  
CS 495  
Stanley Zheng  
October 5th 2013

Inbox (1,865) - szhen002 x Kendall tau Rank Correlat x cfx Online Calculator: Kendall x

calculator-fx.com/calculator/statistics/kendall-tau-correlation

Programming S... Programming S... Ruby Random Other Bookmarks

## Kendall tau correlation calculator

*Computes the Kendall's tau correlation measurement between two sequences of rankings.*

sequence **a**:

.00126	.00051	.00239	.00015	.00342	.00113	.00662	.00279	.00999	
--------	--------	--------	--------	--------	--------	--------	--------	--------	--

+

sequence **b**:

.8	.6	.4	.8	.3	.1	.5	.6	.3	
----	----	----	----	----	----	----	----	----	--

+

Calculate

Assignment 3  
CS 495  
Stanley Zheng  
October 5th 2013

Inbox (1,865) - szhen002 x Kendall tau Rank Correlat x Results | calculator-fx x

calculator-fx.com/post/calculator-result/kendall-tau-correlat

Programming S... Programming S... Ruby Random Other Bookmarks

Entered sequence **a**:

$i$	$a_i$
1	0.00126
2	0.00051
3	0.00239
4	0.00015
5	0.00342
6	0.00113
7	0.00662
8	0.00279
9	0.00999

Entered sequence **b**:

$i$	$b_i$
1	0.8
2	0.6
3	0.4
4	0.8
5	0.3
6	0.1
7	0.5
8	0.6
9	0.3

Result :  
-0.377168

[-- Enter arguments again](#)

The service is provided "as is" and without warranties of any kind | [Privacy Policy and other information](#) | [Uploads documentation](#) | [API documentation](#) | [Contact webmaster](#)