

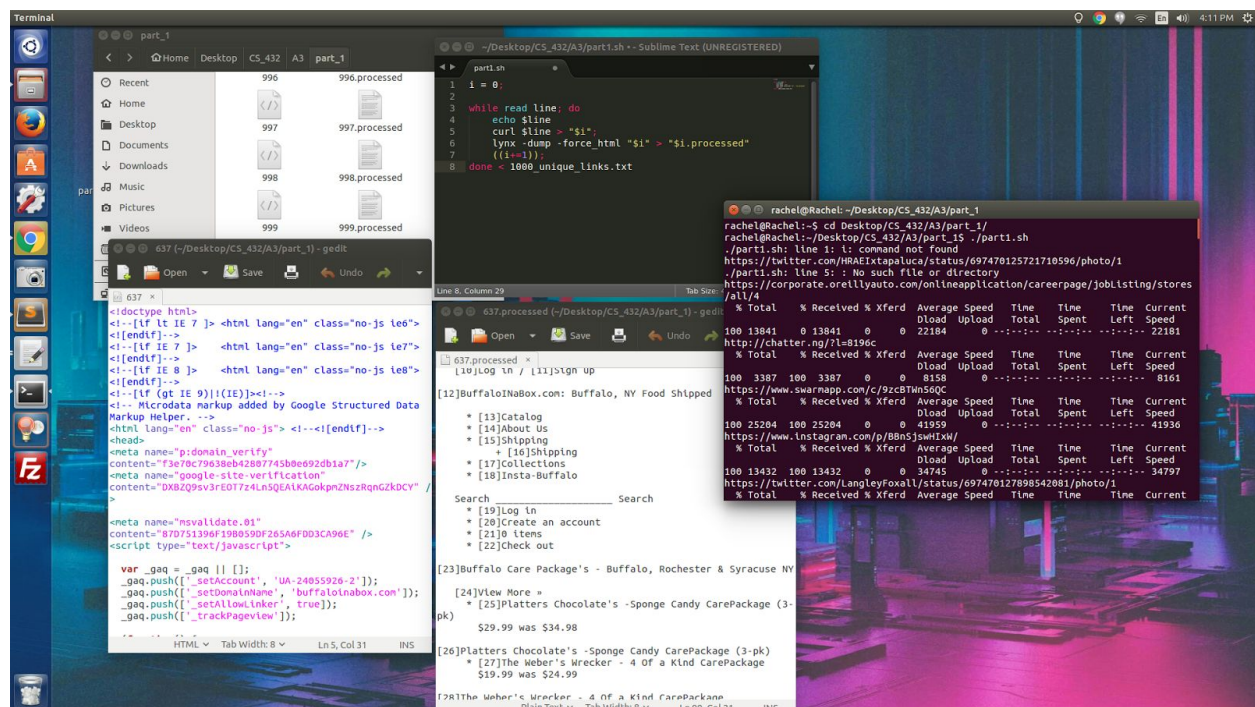
RACHEL MCCREARY
CS 432 ASSIGNMENT 3 REPORT

PART 1

I wrote a script that goes through my “1000 unique links” file, spitting out the raw HTML for each URI into its own file, then stripping the HTML and dumping that into a second “processed” file before moving on to the next URI.

```
i = 0;
while read line; do
    echo $line
    curl $line > "$i";
    lynx -dump -force_html "$i" > "$i.processed"
    ((i+=1));
done < 1000_unique_links.txt
```

Demo screenshot including script, terminal output, and comparison of a URI's raw vs processed results:

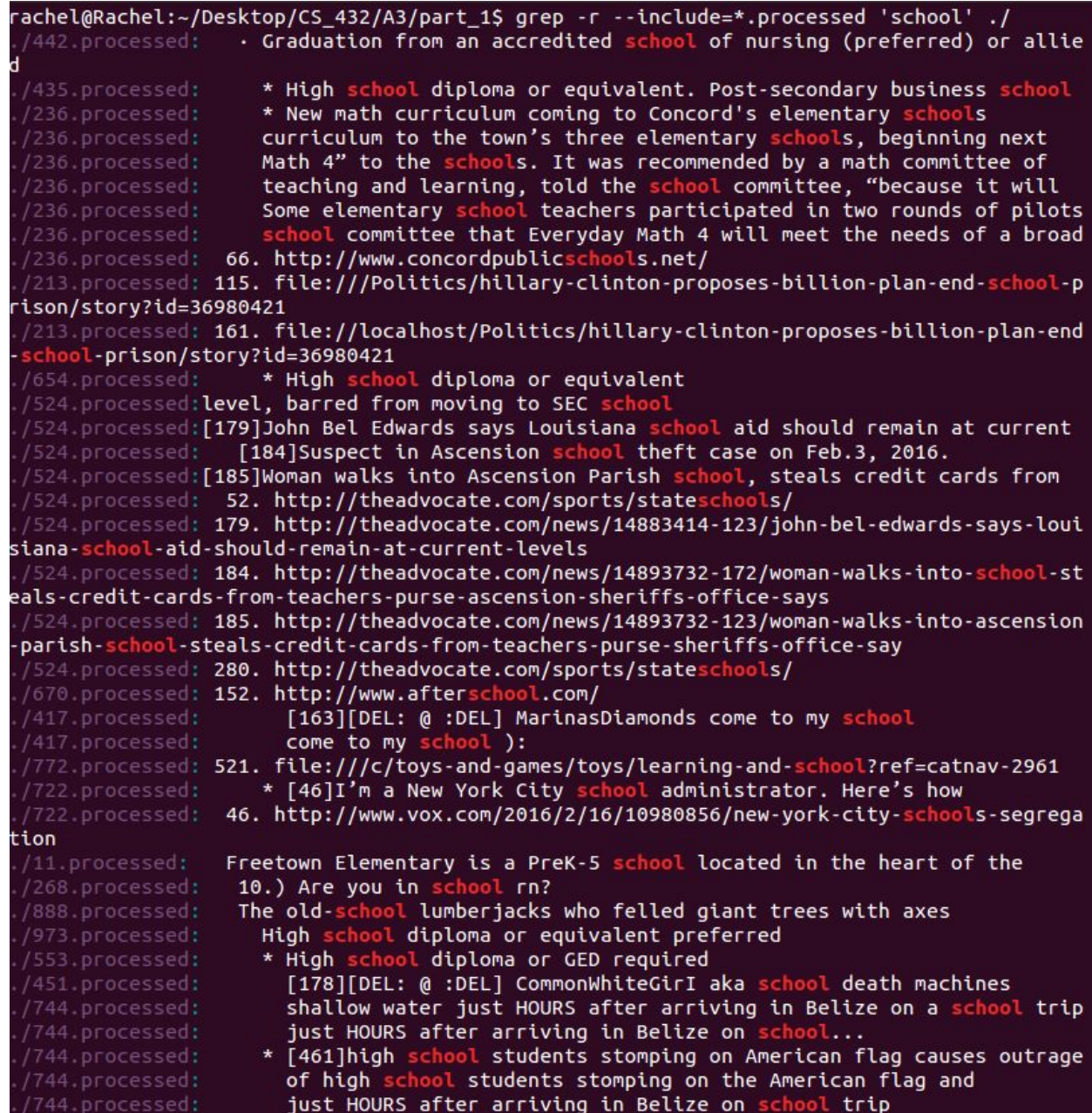


PART 2

For this part, I used “grep” on the processed files to find 10 documents containing the word “school.”

```
grep -r --include=*.processed 'school' ./
```

Demo screenshot of some results:



```
rachel@Rachel:~/Desktop/CS_432/A3/part_1$ grep -r --include=*.processed 'school' ./
./442.processed:  · Graduation from an accredited school of nursing (preferred) or allie
d
./435.processed:  * High school diploma or equivalent. Post-secondary business school
./236.processed:  * New math curriculum coming to Concord's elementary schools
./236.processed:  curriculum to the town's three elementary schools, beginning next
./236.processed:  Math 4" to the schools. It was recommended by a math committee of
./236.processed:  teaching and learning, told the school committee, "because it will
./236.processed:  Some elementary school teachers participated in two rounds of pilots
./236.processed:  school committee that Everyday Math 4 will meet the needs of a broad
./236.processed:  66. http://www.concordpublicschools.net/
./213.processed:  115. file:///Politics/hillary-clinton-proposes-billion-plan-end-school-prison/story?id=36980421
./213.processed:  161. file:///localhost/Politics/hillary-clinton-proposes-billion-plan-end-school-prison/story?id=36980421
./654.processed:  * High school diploma or equivalent
./524.processed: level, barred from moving to SEC school
./524.processed: [179]John Bel Edwards says Louisiana school aid should remain at current
./524.processed: [184]Suspect in Ascension school theft case on Feb.3, 2016.
./524.processed: [185]Woman walks into Ascension Parish school, steals credit cards from
./524.processed: 52. http://theadvocate.com/sports/stateschools/
./524.processed: 179. http://theadvocate.com/news/14883414-123/john-bel-edwards-says-louisiana-school-aid-should-remain-at-current-levels
./524.processed: 184. http://theadvocate.com/news/14893732-172/woman-walks-into-school-steals-credit-cards-from-teachers-purse-ascension-sheriffs-office-says
./524.processed: 185. http://theadvocate.com/news/14893732-123/woman-walks-into-ascension-parish-school-steals-credit-cards-from-teachers-purse-sheriffs-office-say
./524.processed: 280. http://theadvocate.com/sports/stateschools/
./670.processed: 152. http://www.after-school.com/
./417.processed: [163][DEL: @ :DEL] MarinasDiamonds come to my school
./417.processed: come to my school ):
./772.processed: 521. file:///c:/toys-and-games/toys/learning-and-school?ref=catnav-2961
./722.processed: * [46]I'm a New York City school administrator. Here's how
./722.processed: 46. http://www.vox.com/2016/2/16/10980856/new-york-city-schools-segregation
tion
./11.processed: Freetown Elementary is a PreK-5 school located in the heart of the
./268.processed: 10.) Are you in school rn?
./888.processed: The old-school lumberjacks who felled giant trees with axes
./973.processed: High school diploma or equivalent preferred
./553.processed: * High school diploma or GED required
./451.processed: [178][DEL: @ :DEL] CommonWhiteGirI aka school death machines
./744.processed: shallow water just HOURS after arriving in Belize on a school trip
./744.processed: just HOURS after arriving in Belize on school...
./744.processed: * [461]high school students stomping on American flag causes outrage
./744.processed: of high school students stomping on the American flag and
./744.processed: just HOURS after arriving in Belize on school trip
```

A total of 32 files mentioned “school.” I chose 10, then calculated the TFIDF values.

TFIDF = TF * IDF

To find TF:

Using "wc" to find the word count of the processed files:

```
rachel@Rachel:~$ cd Desktop/CS_432/A3/part_2
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 11.processed
1172 11.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 236.processed
968 236.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 248.processed
2442 248.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 435.processed
1129 435.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 442.processed
507 442.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 449.processed
1052 449.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 524.processed
4202 524.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 722.processed
912 722.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 744.processed
18996 744.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ wc -w 760.processed
1466 760.processed
rachel@Rachel:~/Desktop/CS_432/A3/part_2$
```

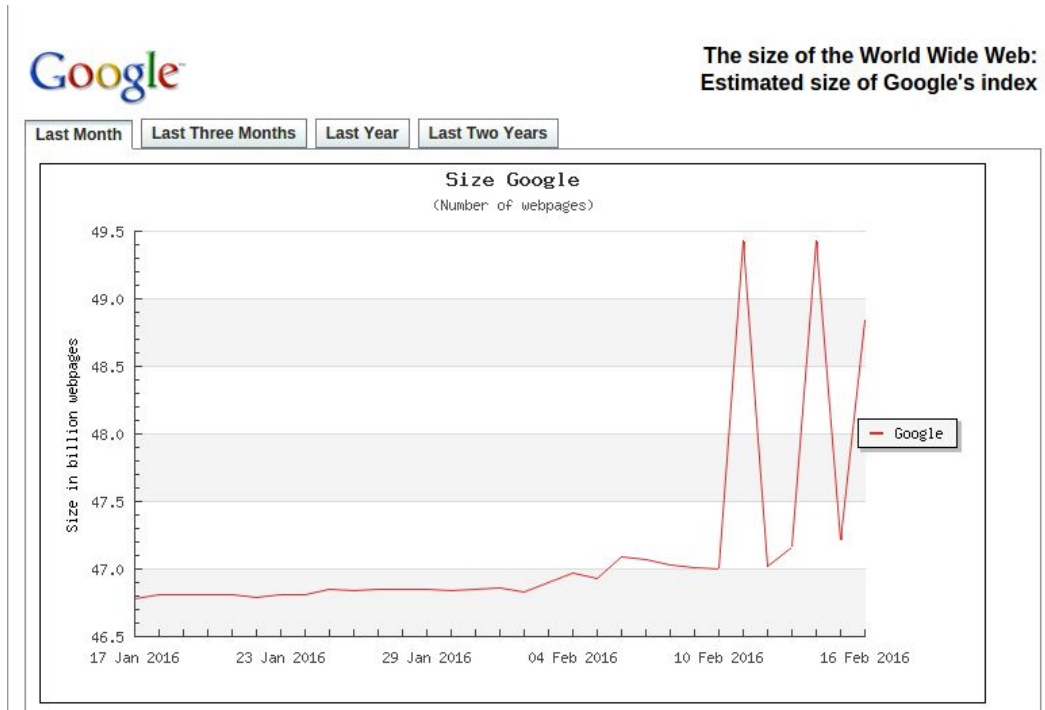
As well as counting the occurrence of "school" in each file:

```
rachel@Rachel:~$ cd Desktop/CS_432/A3/part_2
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 11.processed | wc -w
1
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 236.processed | wc -w
3
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 248.processed | wc -w
2
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 435.processed | wc -w
2
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 442.processed | wc -w
1
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 449.processed | wc -w
2
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 524.processed | wc -w
7
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 722.processed | wc -w
1
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 744.processed | wc -w
9
rachel@Rachel:~/Desktop/CS_432/A3/part_2$ grep -o -w 'school' 760.processed | wc -w
1
rachel@Rachel:~/Desktop/CS_432/A3/part_2$
```

TF = occurrence in doc / words in doc

To find IDF:

Size of Google's Index (docs in corpus): 48.8 B <http://www.worldwidewebsize.com/>



Docs with term: 1.21 B

Google school

All Images Maps Shopping More Search tools

About 1,210,000,000 results (0.51 seconds)

School - Wikipedia, the free encyclopedia
<https://en.wikipedia.org/wiki/School> - Wikipedia
Jump to **Starting a school** - The Toronto District **School** Board is an example of a **school** board that allows parents to design and propose new schools.

The School District of Philadelphia
www.phila.k12.pa.us/ - School District of Philadelphia
The Fund for the **School** District of Philadelphia · Careers · Substitute Teacher · Strategic Partnerships · Parent and Family Portal · **School** Finder · Kindergarten ...

New York City Department of Education
schools.nyc.gov/ - New York City Department of Education
Enter as much criteria on the left as you wish, or click an area on the map below to begin your search in a particular borough. To find your zoned **school**, enter ...

IDF(term) = $\log_2(\text{total docs in corpus} / \text{docs with term})$
= $\log_2(48.8 \text{ B} / 1.21 \text{ B})$
= 5.3338 WolframAlpha

Hits for the term "school", ranked by TFIDF:

| RANK | TFIDF | TF | IDF | URI |
|-----------|--------|--------|--------|---|
| 1 | 0.0165 | 0.0031 | 5.3338 | http://concord.wickedlocal.com/article/20160210/NEWS/160219476 |
| 2 | 0.0107 | 0.0020 | 5.3338 | http://www.ahsocialcareers.com/jobs/descriptions/clinical-documentation-improvement-specialist-health-information-management-glendale-california-job-1-5947228/ |
| 3 | 0.0101 | 0.0019 | 5.3338 | https://wegmans.taleo.net/careersection/2/jobdetail.ftl?job=1600059&SNS=11720 |
| 4 | 0.0096 | 0.0018 | 5.3338 | https://brown.wd5.myworkdayjobs.com/en-US/staff-careers-brown/job/Barus-Building/Academic-Specialist--Urban-Education-Policy-Program_REQ121673-4 |
| 5 | 0.0091 | 0.0017 | 5.3338 | http://www.vox.com/2016/2/9/10956340/fox-news-bernie-sanders |
| 6 | 0.0059 | 0.0011 | 5.3338 | http://www.dailymail.co.uk/tvshowbiz/article-3439899/Idris-Elba-splits-mother-son-Naiyana-Garth.html |
| 7 | 0.0048 | 0.0009 | 5.3338 | https://twitter.com/Wafaalrefaie/status/697470126648659968/video/1 |
| 8 | 0.0043 | 0.0008 | 5.3338 | http://www.huffingtonpost.co.uk/2016/02/10/nicky-morgan-refuses-to-make-pshe-compulsory-in-schools_n_9201786.html |
| 9 | 0.0037 | 0.0007 | 5.3338 | http://www.usnews.com/news/articles/2015/09/24/study-mls-tickets-are-worlds-worst-soccer-value |
| 10 | 0.0027 | 0.0005 | 5.3338 | https://quintiles.taleo.net/careersection/10080/jobdetail.ftl?job=1522353&lang=en&src=CWS-10001 |

PART 3

The same 10 URIs ranked by PageRank (http://www.prchecker.info/check_page_rank.php):

| RANK (from Q2) | PR | URI |
|-------------------|-----|---|
| 7 | 1.0 | https://twitter.com/Wafaalrefaie/status/697470126648659968/video/1 |
| 4 | 0.8 | https://brown.wd5.myworkdayjobs.com/en-US/staff-careers-brown/job/Barus-Building/Academic-Specialist--Urban-Education-Policy-Program_REQ121673-4 |
| 6 | 0.7 | http://www.dailymail.co.uk/tvshowbiz/article-3439899/Idris-Elba-splits-mother-son-Naiyana-Garth.html |
| 9 | 0.7 | http://www.usnews.com/news/articles/2015/09/24/study-mls-tickets-are-worlds-worst-soccer-value |
| 8 | 0.6 | http://www.huffingtonpost.co.uk/2016/02/10/nicky-morgan-refuses-to-make-pshe-compulsory-in-schools_n_9201786.html |
| 5 | 0.6 | http://www.vox.com/2016/2/9/10956340/fox-news-bernie-sanders |
| 3 | 0.6 | https://wegmans.taleo.net/careersection/2/jobdetail.ftl?job=1600059&SNS=11720 |
| 10 | 0.6 | https://quintiles.taleo.net/careersection/10080/jobdetail.ftl?job=1522353&lang=en&src=CWS-10001 |
| 1 | 0.5 | http://concord.wickedlocal.com/article/20160210/NEWS/160219476 |
| 2 | N/A | http://www.ahsocalcareers.com/jobs/descriptions/clinical-documentation-improvement-specialist-health-information-management-glendale-california-job-1-5947228/ |

Some job postings ranked very high in part 2, but fell in rank for part 3. I noticed that the most relevant link (an article about a new math curriculum for schools) sank from the top in part 2 to the bottom in part 3. The N/A result happened even after attempting to enter the top-level URI into three different PR checkers.