

COEN 272 Project 1 -- Web Crawler

Team Members: David Obatake, Chen Gao

April 27th, 2016

Source Code Components

Our web crawler project contains two major components: Web Crawler and Content Extractor. Both components used the Jsoup library for HTML parsing and extraction. The code also uses the open source library gagawa to initially generate the report.html document with proper opening and closing HTML tags.

Web Crawler

The web crawler has two subcomponents: the WebCrawler class and the OutputWriter class. The WebCrawler class provides the main functionality for crawling, namely requesting the web page, parsing the HTML tags for outbound links, and then adding those links to a queue. This class also provides functionality for parsing the robots.txt file of the domain and subsequently checking each URL against the rules imposed by the robots.txt file.

The OutputWriter class serves as a utility class for the web crawler to output HTML to the document store. As each page is parsed, the document is passed to the OutputWriter to first store the raw HTML, then creates the corresponding entry in the report.html file with statistics for the parsed web page.

Content Extractor

Content Extractor is based on the idea of text-to-tag ratio (or text density). The ContentExtractor class is split in two parts: bodyProcessor method and removeNoise method. The bodyProcessor method of the ContentExtractor class iterates over all the HTML files saved in the document store and reads each file separately. Jsoup library provides function to get all children nodes under the HTML body element, which provides us a DOM tree of HTML so that we can recursively call into the leaf node and count all the number of tags and total length of text under each element.

Noise removal will be discussed in further detail below in Noise Reduction under design and architecture.

Design and Architecture

Noise Reduction

Noise reduction was performed as a modification of an algorithm detailed by Fei Sun et. al (2011) called Content Extraction via Text Density. The algorithm calculates the text density per HTML tag by using the ratio of text characters to tags and if a particular tag has a greater text density than a set threshold, that tag is considered a part of the main content block of the web page.

For this project, we loosely followed the steps explicated by Fei Sun et. al and made some of our own modifications to simplify implementation. First, a preprocessing step removed extraneous tags, listed here: script, noscript, style, iframe, br, nav, head, footer, img, header, button, input, form, a. We found that these tags did not hold much content in comparison to the main content tags of the page or were simply noise. Then, we calculated the text density per node (or HTML tag) and removed those nodes that did not meet the threshold for text density.

To do this, we traversed through the DOM tree using Jsoup to first select the body tag's child nodes (parent nodes) and then recursively calculated the tag and text count for each of their respective child nodes (grandchildren and so on), performing a depth-first traversal of the tree. Each of the parent tags' text counts was a summation of all child tags' text since a nested tag's text would not be visible to the parent tag. The tag count was counted from the number of tags seen from the parent text. After finding the tag and text count for each parent node at the top level, the tags were checked against the text density threshold. Those nodes that did not meet the text density threshold were removed, and the remaining tags were output as to a cleaned store of pages. According to our experimentation, we set a static threshold value of 2 for the text density, meaning that the content nodes needed to have more than double the amount of text characters per number of tags.

Noise Reduction Evaluation and Performance

The evaluation of the noise reduction was performed using visual tests between the cached HTML page and the cleaned version. Pages were manually examined for main content blocks and then the "cleaned" page was verified against those manually labelled content blocks. For example, the two figures (Appendix Figure 1 and Figure 2) show the cached web page, then the cleaned webpage with markings showing the main content block with a black box and noise with red boxes. The resulting cleaned page only retains the content block which is then parsed for text. Most research papers that were referenced in our research used this method (albeit time-consuming) to establish a golden standard of content extraction to first compare their results against what would be considered the main content blocks of the page.

Considering the performance of our attempt, most pages with larger content blocks, such as articles, have the best performance in our noise reduction algorithm. After filtering, we can clearly see those ads, relative articles are removed and the main context is saved. While some other pages without very clear boundary between the main content and noise, or with very short

main content, will sometimes results with important information missing or noise not being removed from the page. Noticeably, main navigation pages with little to no main content blocks are entirely cleaned of tags. However, we found that this was to be desired because of the lack of main content on those pages.

Development Challenges

Since we decided to code the robots.txt parser using a naive method, where we tried to find a robots.txt file per each new host URL encountered in the frontier, an issue arose with redirection between web pages. The resulting redirected URL ended up appending the robots.txt part of the address to a longer URL than initially intended and did not resolve with a base host URL for the page. This error occurred mainly when crawling over Yahoo news pages that redirected <http://news.yahoo.com> to <http://yahoo.com/news/>. Given more time to implement and test, a strategy that we came up with would be to set the crawler to not follow redirection links to first see if there is a corresponding robots.txt for the initial URL from the frontier, then follow the redirection and then check again if the redirected page host URL has a robots.txt page.

In terms of the text-to-tag ratio, we spent a lot of time in tuning the the proper threshold for removing noise. We started with 4 and found out there are some websites, especially some main pages, they usually don't have a big chunk of text in one tag but some highlights of one or two sentences per div on highly structured pages. A ratio of 4 (text characters per tag) will be too large for these pages and most content blocks will be filtered under this circumstance. After several trials and considering we have pre-filtered all head/footer/script tags, we found ratio of 2 is good enough to keep most of meaningful context and at the same time not too small to retain tags containing noise.

If we allocated more time to testing, we would have liked to test our noise removal algorithm against other researched noise removal techniques like Composite Text Density Sum (also proposed by Fei Sun et. al), CleanEval, and metrics like punctuation density. Testing against these researched algorithms would be a good rigorous test to see how well our noise removal attempt objectively scored against proposed models.

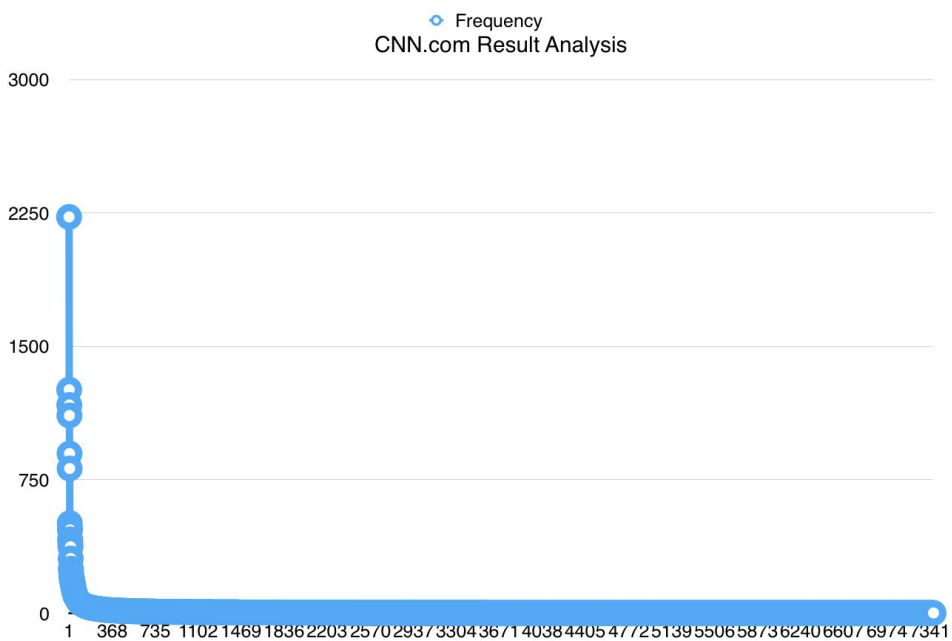
Text Analysis

Plots of Word Rank vs. Word Frequency

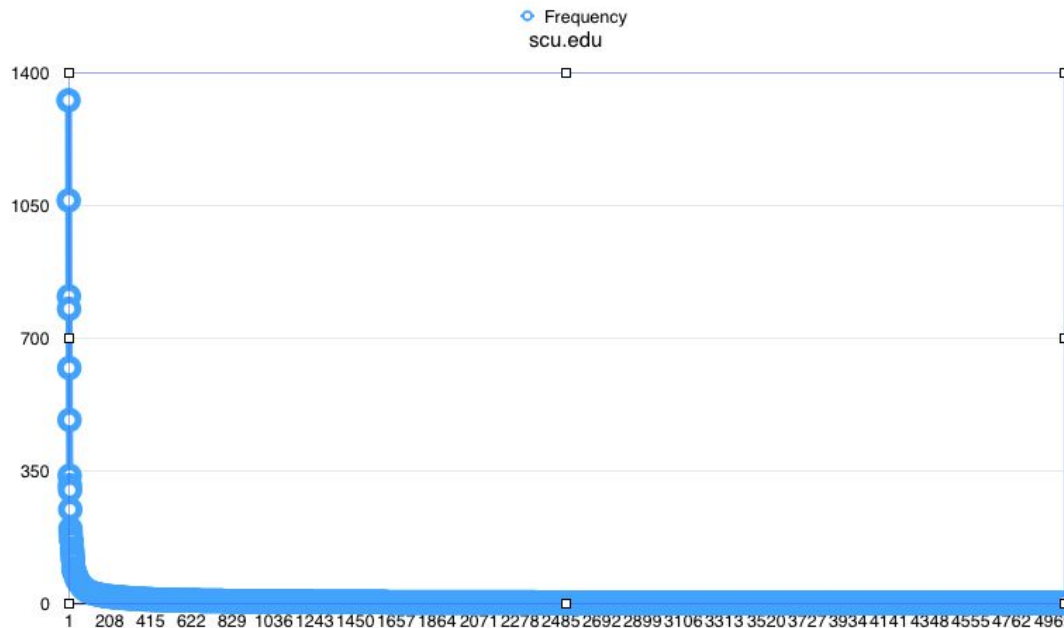
1. Yahoo.com:



2. CNN.com



3. scu.edu



Zipf's Law Analysis::

For each crawl, we limit 100 pages of crawling limit. It is obviously that all 3 websites we crawled as seed are pretty close to Zipf's law. They all have some discontinuation at the beginning and start to get a smooth curve after rank 20. Examining the Zipf scores reported for the 100 most frequent words, we see that after the initial several words, the scores are very close to the Zipf score ($c = 0.1$) for English for the remaining words in the set.

Works Cited

[1] Sun, Fei, Dandan Song, and Leijian Liao. DOM Based Content Extraction via Text Density. School of Computer Science, Beijing Institute of Technology, 24 July 2011. Web. 24 Apr. 2016. <<http://ofey.me/papers/cetd-sigir11.pdf>>.

Appendix

Crawl Information

100 Most Frequent Words Per Crawl

1. ("<https://www.yahoo.com/>", 100)

Yahoo.com

| Word | Rank | Frequency | | |
|-------|------|-----------|-------------|--|
| the | 1 | 2008 | 0.042279914 | |
| to | 2 | 1260 | 0.05306045 | |
| and | 3 | 1020 | 0.06443055 | |
| a | 4 | 948 | 0.07984334 | |
| in | 5 | 850 | 0.089486875 | |
| of | 6 | 790 | 0.09980418 | |
| ideas | 7 | 703 | 0.10361527 | |
| on | 8 | 467 | 0.07866422 | |
| yahoo | 9 | 467 | 0.08849725 | |
| i | 10 | 391 | 0.08232792 | |
| you | 11 | 384 | 0.08893942 | |
| your | 12 | 359 | 0.09070811 | |
| that | 13 | 358 | 0.09799339 | |
| for | 14 | 350 | 0.1031731 | |
| with | 15 | 317 | 0.100120015 | |
| is | 16 | 316 | 0.10645779 | |
| this | 17 | 315 | 0.11275346 | |
| was | 18 | 262 | 0.09929885 | |
| it | 19 | 251 | 0.1004148 | |
| his | 20 | 240 | 0.10106753 | |
| at | 21 | 235 | 0.10391005 | |
| or | 22 | 202 | 0.093571685 | |
| he | 23 | 195 | 0.09443497 | |
| an | 24 | 193 | 0.09753016 | |
| by | 25 | 190 | 0.100014746 | |
| from | 26 | 179 | 0.09799339 | |
| 0 | 27 | 166 | 0.0943718 | |
| like | 28 | 163 | 0.09609838 | |
| as | 29 | 158 | 0.096477374 | |
| dont | 30 | 154 | 0.09727749 | |
| be | 31 | 150 | 0.09790917 | |
| time | 32 | 144 | 0.09702483 | |
| but | 33 | 141 | 0.097972326 | |
| news | 34 | 138 | 0.09879351 | |
| my | 35 | 138 | 0.101699196 | |
| have | 36 | 137 | 0.103846885 | |
| more | 37 | 136 | 0.10595246 | |
| 1 | 38 | 135 | 0.10801592 | |

| | | | | |
|--------|----|-----|-------------|--|
| are | 39 | 134 | 0.110037275 | |
| search | 40 | 122 | 0.102751985 | |
| her | 41 | 121 | 0.104457505 | |
| new | 42 | 118 | 0.10435222 | |
| has | 43 | 115 | 0.104120605 | |
| all | 44 | 115 | 0.10654201 | |
| about | 45 | 114 | 0.10801592 | |
| up | 46 | 114 | 0.11041627 | |
| people | 47 | 113 | 0.11182701 | |
| ad | 48 | 111 | 0.11218496 | |
| now | 49 | 109 | 0.112458676 | |
| out | 50 | 108 | 0.11370097 | |
| one | 51 | 107 | 0.11490114 | |
| who | 52 | 105 | 0.114964314 | |
| mail | 53 | 103 | 0.11494325 | |
| also | 54 | 101 | 0.11483798 | |
| 2 | 55 | 99 | 0.11464847 | |
| get | 56 | 97 | 0.11437475 | |
| day | 57 | 95 | 0.11401681 | |
| us | 58 | 94 | 0.11479586 | |
| not | 59 | 92 | 0.11429053 | |
| email | 60 | 92 | 0.11622766 | |
| what | 61 | 90 | 0.11559598 | |
| were | 62 | 89 | 0.116185546 | |
| can | 63 | 89 | 0.1180595 | |
| I | 64 | 88 | 0.1185859 | |
| will | 65 | 88 | 0.12043881 | |
| its | 66 | 88 | 0.12229171 | |
| 4 | 67 | 85 | 0.119912416 | |
| when | 68 | 85 | 0.12170215 | |
| first | 69 | 84 | 0.122039035 | |
| their | 70 | 82 | 0.12085991 | |
| 3 | 71 | 78 | 0.11660666 | |
| how | 72 | 78 | 0.11824901 | |
| 5 | 73 | 75 | 0.11528015 | |
| we | 74 | 75 | 0.11685933 | |
| our | 75 | 75 | 0.11843851 | |
| if | 76 | 74 | 0.11841745 | |
| 6 | 77 | 72 | 0.116733 | |
| video | 78 | 70 | 0.114964314 | |

| | | | |
|-----------|-----|----|-------------|
| style | 79 | 70 | 0.11643822 |
| most | 80 | 70 | 0.11791211 |
| 7 | 81 | 69 | 0.1176805 |
| april | 82 | 69 | 0.119133346 |
| while | 83 | 67 | 0.11709094 |
| celebrity | 84 | 67 | 0.11850168 |
| ago | 85 | 67 | 0.11991241 |
| see | 86 | 66 | 0.11951235 |
| trending | 87 | 65 | 0.11907018 |
| web | 88 | 65 | 0.1204388 |
| find | 89 | 65 | 0.121807426 |
| which | 90 | 64 | 0.12128103 |
| am | 91 | 63 | 0.120712526 |
| results | 92 | 63 | 0.122039035 |
| 10 | 93 | 62 | 0.12140737 |
| some | 94 | 62 | 0.12271282 |
| she | 95 | 62 | 0.12401828 |
| try | 96 | 60 | 0.12128103 |
| no | 97 | 60 | 0.12254438 |
| do | 98 | 59 | 0.12174426 |
| so | 99 | 59 | 0.12298655 |
| last | 100 | 59 | 0.124228835 |

2. (“<http://www.cnn.com/>”, 100)

CNN.com

| Word | Rank | Frequency | |
|-------------|------|-----------|-------------|
| the | 1 | 2227 | 0.04786674 |
| of | 2 | 1256 | 0.053992476 |
| to | 3 | 1170 | 0.07544331 |
| and | 4 | 1111 | 0.09551854 |
| a | 5 | 897 | 0.096399784 |
| in | 6 | 813 | 0.10484686 |
| or | 7 | 506 | 0.07613111 |
| on | 8 | 476 | 0.081848465 |
| by | 9 | 468 | 0.090531975 |
| you | 10 | 415 | 0.08919936 |
| cnn | 11 | 407 | 0.09622783 |
| that | 12 | 386 | 0.099559374 |
| for | 13 | 373 | 0.10422354 |
| is | 14 | 369 | 0.111037076 |
| with | 15 | 307 | 0.09897904 |
| at | 16 | 250 | 0.08597528 |
| from | 17 | 247 | 0.09025255 |
| as | 18 | 240 | 0.09285331 |
| be | 19 | 236 | 0.09637829 |
| he | 20 | 213 | 0.09156368 |
| photos | 21 | 206 | 0.09298226 |
| your | 22 | 204 | 0.09646426 |
| news | 23 | 197 | 0.097388506 |
| hide | 24 | 196 | 0.10110693 |
| caption | 25 | 196 | 0.105319716 |
| trump | 26 | 194 | 0.10841483 |
| any | 27 | 191 | 0.110843636 |
| are | 28 | 182 | 0.109532505 |
| his | 29 | 178 | 0.1109511 |
| this | 30 | 176 | 0.11348737 |
| we | 31 | 168 | 0.11193982 |
| information | 32 | 167 | 0.11486298 |
| april | 33 | 167 | 0.118452445 |
| pt | 34 | 166 | 0.12131113 |
| it | 35 | 158 | 0.118860826 |
| may | 36 | 151 | 0.11684041 |
| an | 37 | 146 | 0.11610962 |
| will | 38 | 146 | 0.11924771 |

| | | | |
|----------|----|-----|-------------|
| has | 39 | 145 | 0.12154755 |
| not | 40 | 143 | 0.12294465 |
| have | 41 | 141 | 0.12425578 |
| clinton | 42 | 131 | 0.118259005 |
| our | 43 | 131 | 0.1210747 |
| was | 44 | 130 | 0.12294465 |
| use | 45 | 123 | 0.1189683 |
| 61 | 46 | 123 | 0.12161204 |
| about | 47 | 119 | 0.12021493 |
| its | 48 | 118 | 0.121741 |
| but | 49 | 115 | 0.121117674 |
| other | 50 | 111 | 0.1192907 |
| more | 51 | 111 | 0.12167651 |
| world | 52 | 108 | 0.1207093 |
| 2016 | 53 | 106 | 0.12075228 |
| 50 | 54 | 103 | 0.119548626 |
| new | 55 | 98 | 0.11585169 |
| 14 | 56 | 96 | 0.115550786 |
| content | 57 | 96 | 0.117614195 |
| her | 58 | 96 | 0.119677596 |
| hillary | 59 | 94 | 0.11920473 |
| she | 60 | 94 | 0.12122515 |
| site | 61 | 93 | 0.12193444 |
| services | 62 | 91 | 0.12126813 |
| us | 63 | 89 | 0.12051585 |
| look | 64 | 87 | 0.119677596 |
| all | 65 | 87 | 0.12154756 |
| latest | 66 | 85 | 0.12058034 |
| said | 67 | 85 | 0.12240731 |
| best | 68 | 84 | 0.1227727 |
| if | 69 | 83 | 0.12309512 |
| cnncom | 70 | 82 | 0.12337453 |
| service | 71 | 81 | 0.123610966 |
| around | 72 | 80 | 0.123804405 |
| me | 73 | 78 | 0.122385815 |
| also | 74 | 77 | 0.12247179 |
| just | 75 | 76 | 0.12251478 |
| i | 76 | 74 | 0.120881245 |
| such | 77 | 73 | 0.12081676 |
| who | 78 | 73 | 0.12238581 |

| | | | |
|----------|-----|----|-------------|
| one | 79 | 71 | 0.120558836 |
| these | 80 | 70 | 0.120365396 |
| some | 81 | 68 | 0.11838797 |
| selfies | 82 | 66 | 0.11632456 |
| their | 83 | 66 | 0.11774315 |
| day | 84 | 65 | 0.117356256 |
| no | 85 | 65 | 0.11875336 |
| most | 86 | 65 | 0.120150454 |
| terms | 87 | 62 | 0.11593767 |
| campaign | 88 | 61 | 0.11537883 |
| policy | 89 | 60 | 0.114777006 |
| says | 90 | 59 | 0.11413218 |
| party | 91 | 59 | 0.11540032 |
| what | 92 | 59 | 0.116668455 |
| been | 93 | 58 | 0.11593767 |
| out | 94 | 58 | 0.11718431 |
| can | 95 | 58 | 0.11843096 |
| clintons | 96 | 57 | 0.11761418 |
| would | 97 | 57 | 0.11883933 |
| career | 98 | 57 | 0.12006448 |
| stories | 99 | 57 | 0.121289626 |
| they | 100 | 56 | 0.120365396 |

3. (“<https://www.scu.edu/>”, 100, “<https://www.scu.edu/>”)

| | | | |
|----|------|-------------|------------|
| 1 | 1326 | 0.048031297 | the |
| 2 | 1062 | 0.07693701 | and |
| 3 | 809 | 0.087912485 | to |
| 4 | 776 | 0.11243525 | of |
| 5 | 620 | 0.11229037 | a |
| 6 | 483 | 0.104973376 | in |
| 7 | 336 | 0.08519578 | for |
| 8 | 311 | 0.090122074 | santa |
| 9 | 298 | 0.097149275 | clara |
| 10 | 247 | 0.089470066 | university |
| 11 | 196 | 0.07809614 | as |
| 12 | 185 | 0.080414385 | is |
| 13 | 185 | 0.087115586 | with |
| 14 | 168 | 0.08519578 | students |
| 15 | 168 | 0.0912812 | scu |
| 16 | 166 | 0.096207485 | that |
| 17 | 164 | 0.10098888 | on |
| 18 | 163 | 0.10627739 | our |
| 19 | 144 | 0.099105306 | at |
| 20 | 135 | 0.09780128 | you |
| 21 | 129 | 0.09812729 | from |
| 22 | 128 | 0.10200311 | are |
| 23 | 119 | 0.099141516 | we |
| 24 | 113 | 0.09823595 | by |
| 25 | 111 | 0.10051798 | an |
| 26 | 94 | 0.08852827 | campus |
| 27 | 91 | 0.08899917 | was |
| 28 | 85 | 0.08621001 | day |
| 29 | 85 | 0.08928894 | your |
| 30 | 84 | 0.0912812 | he |
| 31 | 84 | 0.0943239 | or |
| 32 | 84 | 0.09736661 | their |
| 33 | 83 | 0.099213965 | more |
| 34 | 82 | 0.10098888 | school |
| 35 | 76 | 0.096352376 | it |
| 36 | 75 | 0.097801276 | about |
| 37 | 75 | 0.10051798 | they |

| | | | |
|----|----|-------------|---------------|
| 38 | 75 | 0.103234686 | program |
| 39 | 72 | 0.10171334 | be |
| 40 | 72 | 0.104321375 | all |
| 41 | 69 | 0.10247401 | education |
| 42 | 66 | 0.10040932 | will |
| 43 | 65 | 0.10124243 | graduate |
| 44 | 64 | 0.10200311 | student |
| 45 | 63 | 0.102691345 | his |
| 46 | 61 | 0.10164089 | who |
| 47 | 61 | 0.10385047 | programs |
| 48 | 60 | 0.10432137 | have |
| 49 | 58 | 0.1029449 | us |
| 50 | 57 | 0.103234686 | work |
| 51 | 56 | 0.10345203 | center |
| 52 | 56 | 0.1054805 | this |
| 53 | 55 | 0.10558916 | community |
| 54 | 55 | 0.10758141 | undergraduate |
| 55 | 55 | 0.109573655 | world |
| 56 | 54 | 0.10953744 | research |
| 57 | 53 | 0.10942877 | one |
| 58 | 53 | 0.11134857 | social |
| 59 | 52 | 0.111131236 | faculty |
| 60 | 52 | 0.11301482 | what |
| 61 | 52 | 0.11489839 | its |
| 62 | 52 | 0.11678197 | can |
| 63 | 50 | 0.1141015 | i |
| 64 | 49 | 0.113594376 | engineering |
| 65 | 49 | 0.11536929 | help |
| 66 | 49 | 0.1171442 | has |
| 67 | 46 | 0.11163835 | mission |
| 68 | 45 | 0.11084145 | business |
| 69 | 44 | 0.109972104 | were |
| 70 | 43 | 0.10903032 | last |
| 71 | 43 | 0.110587895 | other |
| 72 | 43 | 0.11214547 | said |
| 73 | 42 | 0.11105879 | than |
| 74 | 41 | 0.10989967 | experience |

| | | | |
|-----|----|-------------|---------------|
| 75 | 41 | 0.111384794 | through |
| 76 | 40 | 0.110117 | academic |
| 77 | 40 | 0.11156591 | 2016 |
| 78 | 39 | 0.11018944 | college |
| 79 | 39 | 0.11160213 | but |
| 80 | 38 | 0.110117 | up |
| 81 | 38 | 0.11149346 | jesuit |
| 82 | 37 | 0.10989966 | global |
| 83 | 36 | 0.10823342 | staff |
| 84 | 36 | 0.10953744 | do |
| 85 | 36 | 0.11084146 | also |
| 86 | 35 | 0.10903032 | may |
| 87 | 35 | 0.11029811 | out |
| 88 | 35 | 0.11156591 | opportunities |
| 89 | 35 | 0.1128337 | arts |
| 90 | 34 | 0.11084145 | study |
| 91 | 34 | 0.11207302 | offers |
| 92 | 34 | 0.11330459 | she |
| 93 | 33 | 0.11116746 | if |
| 94 | 33 | 0.11236281 | president |
| 95 | 32 | 0.110116996 | well |
| 96 | 32 | 0.11127612 | services |
| 97 | 32 | 0.11243525 | many |
| 98 | 32 | 0.113594376 | year |
| 99 | 32 | 0.1147535 | which |
| 100 | 32 | 0.11591263 | events |

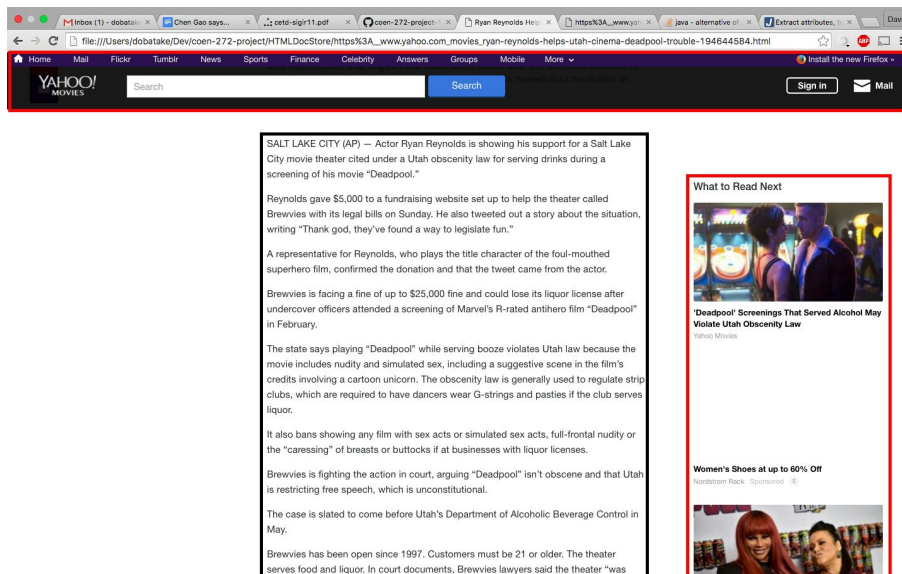


Figure 1. Cached Web Page

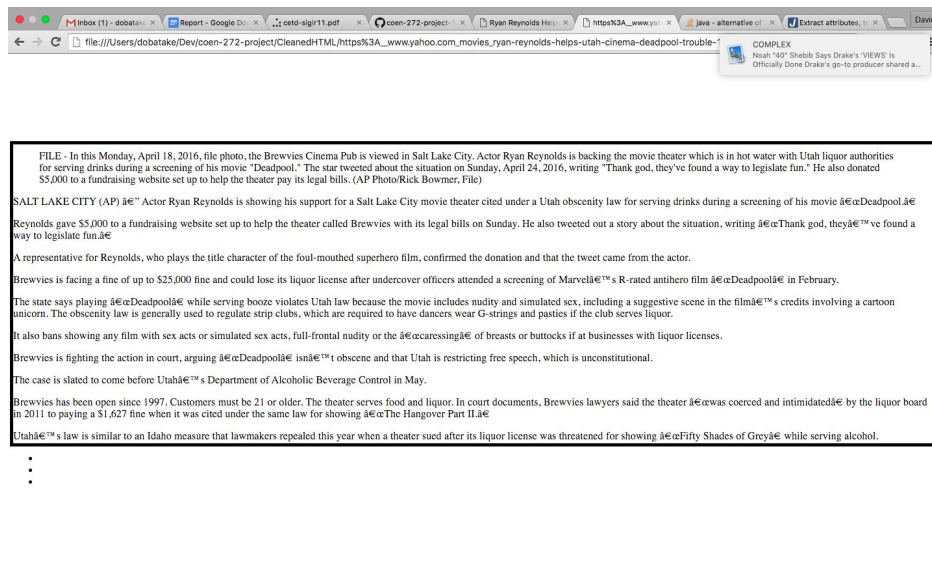


Figure 2. Cleaned Web Page