Anjela Orlanes, Ocean Lu

CS 4990.02

Dr. Steichen

14 March 2019

Project 1 – Web Crawler

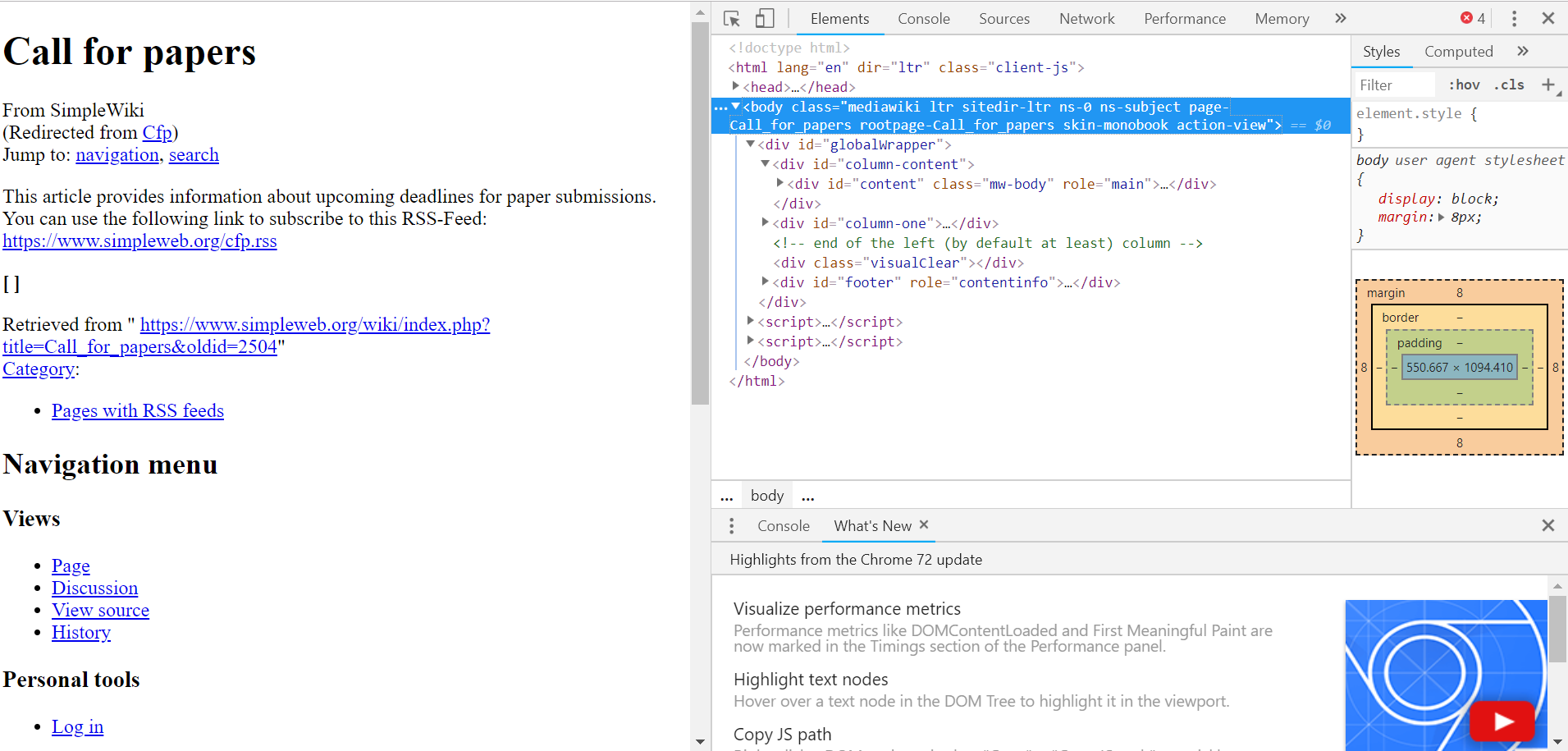
Our project goal was to create a web crawler that would retrieve multiple links from a single seeded link. In our program, we have 6 methods within the Crawler class, and 1 main method. Our methods include crawl(String), connect(String), getNextUrl(), getLinks(), toCSV(), and convert(String[]). The methods toCSV() and convert(String[]) are for transferring links and URL counts to a CSV file. The methods crawl(String) and connect(String) were our most important methods. Crawl(String) would continually crawl through found links and call connect(String) so that it could fetch the new links and count them. Crawl(String) also called upon the CSV methods. Connect(String) also added all the fond links to a repository folder. Our page maximum was 20 pages. The main challenge in coding the project was getting all of the found links to display in the CSV file. Initially, it would only display the last link found. The problem was solved when we realized we did not have to keep creating a new Crawler Object in the while loop.

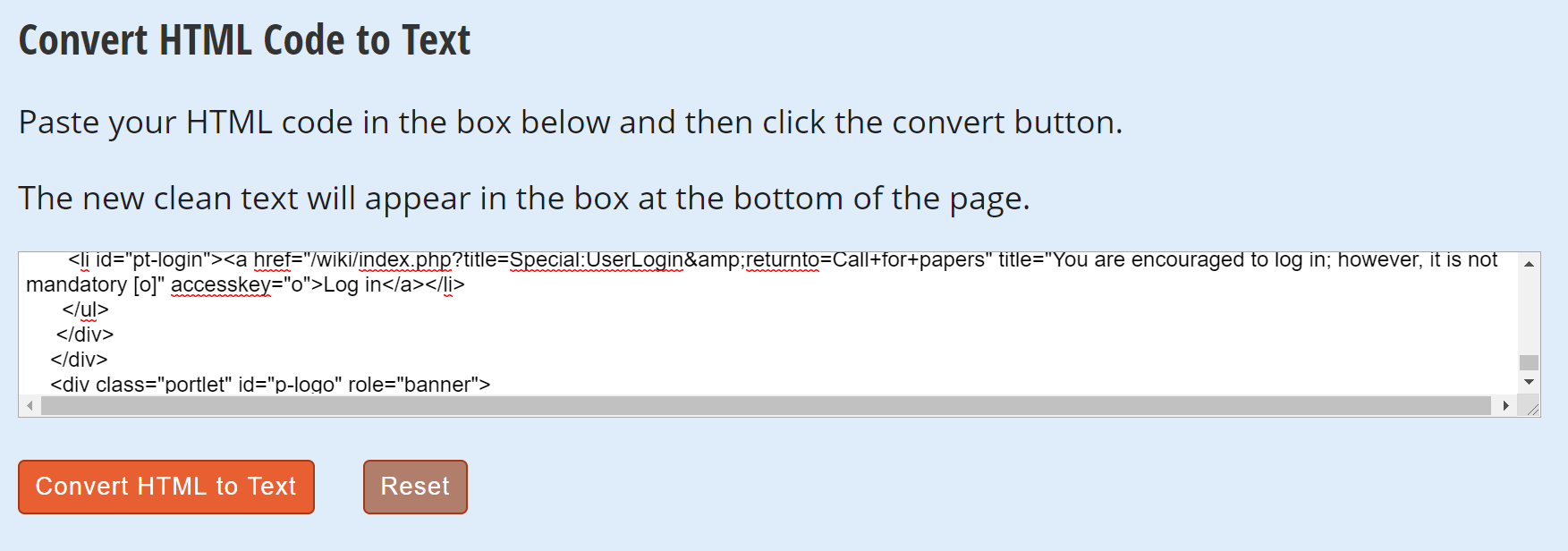
We tested 3 crawls from 3 different seeds. Below are the statistical graphs of the probability of a word occurrence versus the word ranking. In order to find the probability, we found the frequency of the word and divided it by the total number of words. The ranking is displayed from highest frequency to low.

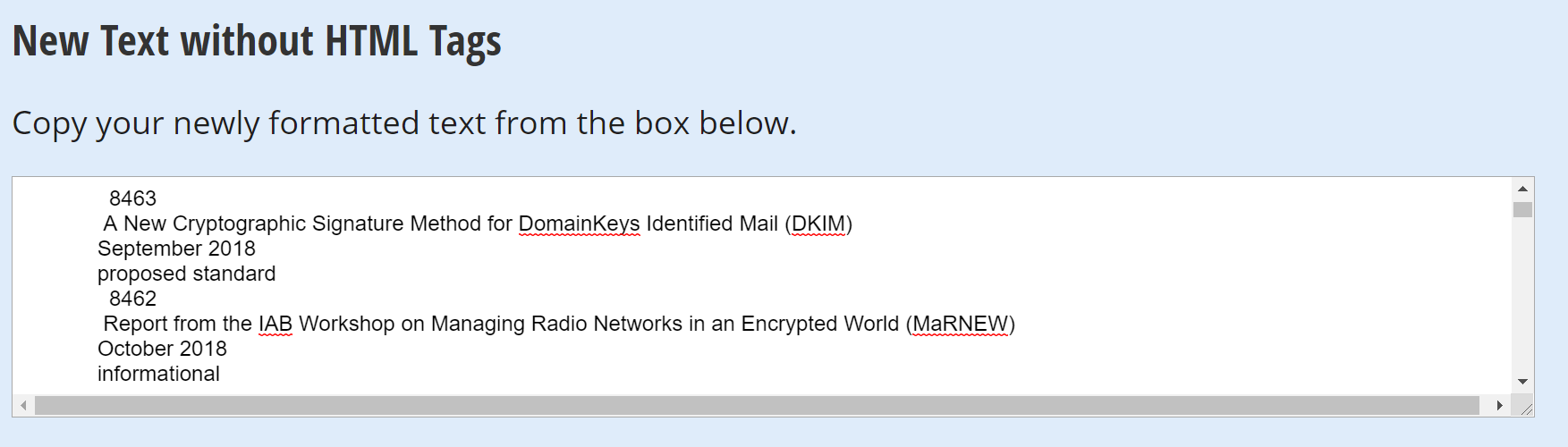
In order to get the data to construct these graphs, we went through each website of the respected crawl, extracted the HTML from the page, converted it to plain text, counted the word frequencies, and then counted the total words (all using 3rd-party websites as shown in Appendix).

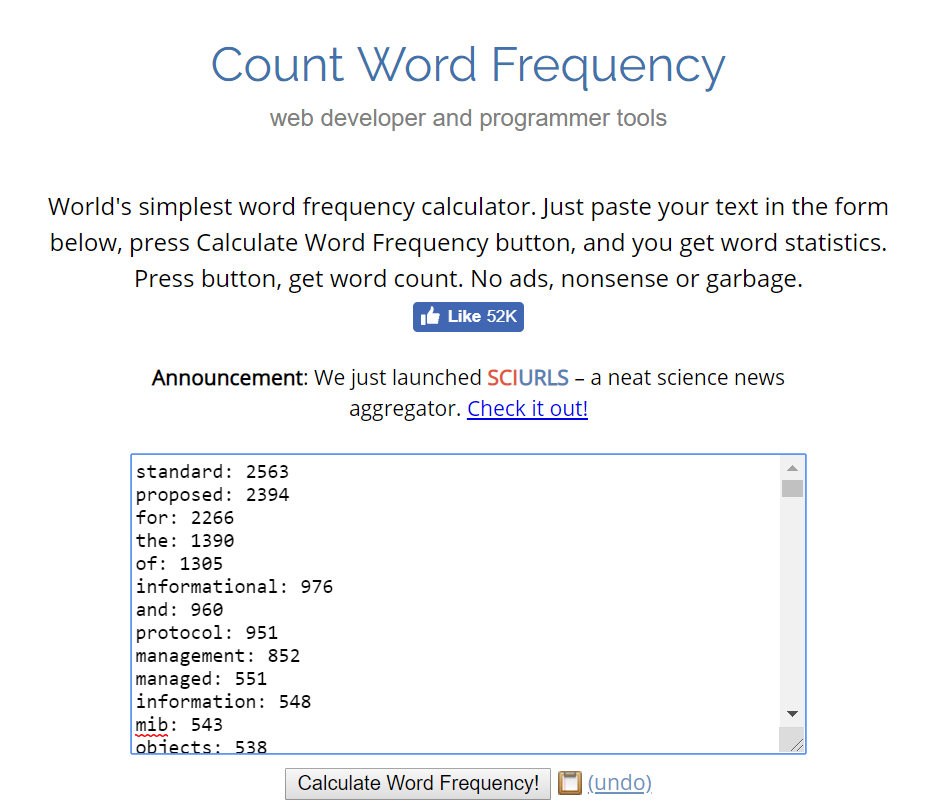
All 3 graphs are very similar to Zipf’s Law. To further test the accuracy, random points were taken, in which the rank was multiplied by the probability. On average, the points would be close to ~0.1. This is significant because this is about the c value for English websites. The Appendix also includes the top 100 frequent words for each crawl.

In our group, Ocean did the most of the code for the crawler and did the research in how to use packages to code for this project. Anjela did the main bulk of the Zipf’s law analysis, as well as assist in debugging and adding additional methods in the code.









Crawl 3

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Probability | Frequency | Word |
| 1 | 0.0412018 | 2563 | standard |
| 2 | 0.038485 | 2394 | proposed |
| 3 | 0.0364274 | 2266 | for |
| 4 | 0.0223451 | 1390 | the |
| 5 | 0.0209787 | 1305 | of |
| 6 | 0.0156898 | 976 | informational |
| 7 | 0.0154326 | 960 | and |
| 8 | 0.0152879 | 951 | protocol |
| 9 | 0.0136964 | 852 | management |
| 10 | 0.0088577 | 551 | managed |
| 11 | 0.0088094 | 548 | information |
| 12 | 0.0087291 | 543 | mib |
| 13 | 0.0086487 | 538 | objects |
| 14 | 0.0084719 | 527 | definitions |
| 15 | 0.0084236 | 524 | network |
| 16 | 0.0077002 | 479 | • |
| 17 | 0.007234 | 450 | in |
| 18 | 0.0066071 | 411 | base |
| 19 | 0.0063177 | 393 | 2011 |
| 20 | 0.0054979 | 342 | 2012 |
| 21 | 0.0054496 | 339 | to |
| 22 | 0.0053532 | 333 | a |
| 23 | 0.0052567 | 327 | 2014 |
| 24 | 0.0050638 | 315 | 2016 |
| 25 | 0.0048227 | 300 | 2015 |
| 26 | 0.0048227 | 300 | using |
| 27 | 0.0045012 | 280 | march |
| 28 | 0.004469 | 278 | 2013 |
| 29 | 0.004244 | 264 | 2017 |
| 30 | 0.0041636 | 259 | may |
| 31 | 0.0039546 | 246 | internet |
| 32 | 0.0039546 | 246 | version |
| 33 | 0.0038421 | 239 | april |
| 34 | 0.0037938 | 236 | experimental |
| 35 | 0.0037938 | 236 | transport |
| 36 | 0.0037456 | 233 | august |
| 37 | 0.0037295 | 232 | february |
| 38 | 0.0036492 | 227 | october |
| 39 | 0.0036331 | 226 | security |
| 40 | 0.0035849 | 223 | control |
| 41 | 0.0035849 | 223 | june |
| 42 | 0.0035366 | 220 | ipv6 |
| 43 | 0.003408 | 212 | july |
| 44 | 0.003392 | 211 | 2018 |
| 45 | 0.0032955 | 205 | interface |
| 46 | 0.0032634 | 203 | november |
| 47 | 0.003199 | 199 | networks |
| 48 | 0.0031669 | 197 | september |
| 49 | 0.0030222 | 188 | january |
| 50 | 0.0027329 | 170 | 2010 |
| 51 | 0.0027329 | 170 | service |
| 52 | 0.0027168 | 169 | extensions |
| 53 | 0.0026203 | 163 | over |
| 54 | 0.002556 | 159 | simple |
| 55 | 0.0025239 | 157 | with |
| 56 | 0.0024917 | 155 | data |
| 57 | 0.0023631 | 147 | label |
| 58 | 0.002331 | 145 | smiv2 |
| 59 | 0.0023149 | 144 | ip |
| 60 | 0.0022827 | 142 | december |
| 61 | 0.0022827 | 142 | key |
| 62 | 0.0022506 | 140 | routing |
| 63 | 0.002122 | 132 | monitoring |
| 64 | 0.0020738 | 129 | multicast |
| 65 | 0.0020577 | 128 | authentication |
| 66 | 0.0020255 | 126 | session |
| 67 | 0.0019934 | 124 | historic |
| 68 | 0.0019451 | 121 | practice |
| 69 | 0.0019451 | 121 | extension |
| 70 | 0.001913 | 119 | current |
| 71 | 0.001913 | 119 | an |
| 72 | 0.0018808 | 117 | layer |
| 73 | 0.0018808 | 117 | on |
| 74 | 0.0018648 | 116 | use |
| 75 | 0.0018487 | 115 | model |
| 76 | 0.0018487 | 115 | best |
| 77 | 0.0018165 | 113 | traffic |
| 78 | 0.0017362 | 108 | by |
| 79 | 0.0016236 | 101 | mpls |
| 80 | 0.0016236 | 101 | textual |
| 81 | 0.0016076 | 100 | system |
| 82 | 0.0015754 | 98 | resource |
| 83 | 0.0015593 | 97 | address |
| 84 | 0.0015433 | 96 | requirements |
| 85 | 0.0015433 | 96 | (mib) |
| 86 | 0.0015272 | 95 | message |
| 87 | 0.001495 | 93 | mobile |
| 88 | 0.0014629 | 91 | framework |
| 89 | 0.0014629 | 91 | conventions |
| 90 | 0.0014147 | 88 | virtual |
| 91 | 0.0013986 | 87 | access |
| 92 | 0.0013986 | 87 | ieee |
| 93 | 0.0013825 | 86 | support |
| 94 | 0.0013825 | 86 | switching |
| 95 | 0.0013664 | 85 | (snmp) |
| 96 | 0.0013504 | 84 | format |
| 97 | 0.0013343 | 83 | application |
| 98 | 0.0013021 | 81 | rtp |
| 99 | 0.00127 | 79 | initiation |
| 100 | 0.00127 | 79 | type |

Crawl 2

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Probability | Frequency | Word |
| 1 | 0.0712928 | 1763 | • |
| 2 | 0.0274981 | 680 | apple |
| 3 | 0.0222815 | 551 | iphone |
| 4 | 0.0213514 | 528 | and |
| 5 | 0.0171054 | 423 | the |
| 6 | 0.0159327 | 394 | to |
| 7 | 0.0155283 | 384 | menu |
| 8 | 0.015407 | 381 | for |
| 9 | 0.0142343 | 352 | a |
| 10 | 0.0141939 | 351 | with |
| 11 | 0.014113 | 349 | your |
| 12 | 0.0131829 | 326 | you |
| 13 | 0.0114036 | 282 | or |
| 14 | 0.0108375 | 268 | from |
| 15 | 0.0104735 | 259 | in |
| 16 | 0.0095839 | 237 | new |
| 17 | 0.0086942 | 215 | more |
| 18 | 0.0084516 | 209 | close |
| 19 | 0.0083303 | 206 | of |
| 20 | 0.0080068 | 198 | on |
| 21 | 0.0079664 | 197 | store |
| 22 | 0.0077642 | 192 | open |
| 23 | 0.0076024 | 188 | learn |
| 24 | 0.0066319 | 164 | an |
| 25 | 0.0053379 | 132 | watch |
| 26 | 0.0051357 | 127 | can |
| 27 | 0.0050144 | 124 | device |
| 28 | 0.0047313 | 117 | all |
| 29 | 0.0045695 | 113 | is |
| 30 | 0.0045695 | 113 | about |
| 31 | 0.0045291 | 112 | it |
| 32 | 0.0044887 | 111 | shop |
| 33 | 0.0042865 | 106 | at |
| 34 | 0.0041652 | 103 | case |
| 35 | 0.0040843 | 101 | buy |
| 36 | 0.0040438 | 100 | be |
| 37 | 0.0038821 | 96 | at&t |
| 38 | 0.0038821 | 96 | verizon |
| 39 | 0.0037608 | 93 | trade-in |
| 40 | 0.0037203 | 92 | may |
| 41 | 0.0037203 | 92 | i |
| 42 | 0.0034373 | 85 | mac |
| 43 | 0.0033968 | 84 | ipad |
| 44 | 0.0032755 | 81 | if |
| 45 | 0.0032351 | 80 | are |
| 46 | 0.0029924 | 74 | music |
| 47 | 0.0029116 | 72 | purchase |
| 48 | 0.0028711 | 71 | space |
| 49 | 0.0028711 | 71 | carrier |
| 50 | 0.0028307 | 70 | use |
| 51 | 0.0028307 | 70 | business |
| 52 | 0.0028307 | 70 | sport |
| 53 | 0.0027498 | 68 | find |
| 54 | 0.0027094 | 67 | education |
| 55 | 0.0027094 | 67 | series |
| 56 | 0.0026689 | 66 | get |
| 57 | 0.0026285 | 65 | upgrade |
| 58 | 0.0025881 | 64 | 4 |
| 59 | 0.0025476 | 63 | aluminum |
| 60 | 0.0025072 | 62 | xs |
| 61 | 0.0025072 | 62 | sprint |
| 62 | 0.0025072 | 62 | 64gb1 |
| 63 | 0.0025072 | 62 | 256gb1 |
| 64 | 0.0023859 | 59 | as |
| 65 | 0.0023454 | 58 | shopping |
| 66 | 0.0023454 | 58 | up |
| 67 | 0.002305 | 57 | not |
| 68 | 0.0022645 | 56 | xr |
| 69 | 0.0022645 | 56 | gray |
| 70 | 0.0022645 | 56 | silver |
| 71 | 0.0022241 | 55 | trade |
| 72 | 0.0022241 | 55 | payment |
| 73 | 0.0022241 | 55 | gold |
| 74 | 0.0021837 | 54 | tv |
| 75 | 0.0021432 | 53 | we |
| 76 | 0.0021028 | 52 | available |
| 77 | 0.0021028 | 52 | t-mobile |
| 78 | 0.0020624 | 51 | id |
| 79 | 0.0020219 | 50 | any |
| 80 | 0.0019815 | 49 | gift |
| 81 | 0.0019006 | 47 | do |
| 82 | 0.0019006 | 47 | when |
| 83 | 0.0019006 | 47 | stainless |
| 84 | 0.0019006 | 47 | steel |
| 85 | 0.0019006 | 47 | black |
| 86 | 0.0018602 | 46 | how |
| 87 | 0.0018197 | 45 | accessories |
| 88 | 0.0018197 | 45 | that |
| 89 | 0.0017793 | 44 | online |
| 90 | 0.0017793 | 44 | values |
| 91 | 0.0017793 | 44 | terms |
| 92 | 0.0017793 | 44 | unlocked |
| 93 | 0.0017388 | 43 | will |
| 94 | 0.0017388 | 43 | order |
| 95 | 0.0017388 | 43 | health |
| 96 | 0.0017388 | 43 | account |
| 97 | 0.0017388 | 43 | program |
| 98 | 0.0017388 | 43 | band |
| 99 | 0.0016984 | 42 | my |
| 100 | 0.0016984 | 42 | by |

Crawl 1

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Probability | Frequency | Word |
| 1 | 0.0895895 | 753 | • |
| 2 | 0.0298632 | 251 | the |
| 3 | 0.0184414 | 155 | and |
| 4 | 0.0167757 | 141 | to |
| 5 | 0.0167757 | 141 | of |
| 6 | 0.0145152 | 122 | & |
| 7 | 0.0145152 | 122 | in |
| 8 | 0.0134444 | 113 | cal |
| 9 | 0.0111838 | 94 | poly |
| 10 | 0.0107079 | 90 | student |
| 11 | 0.01047 | 88 | for |
| 12 | 0.010232 | 86 | university |
| 13 | 0.010113 | 85 | posted |
| 14 | 0.0098751 | 83 | pomona |
| 15 | 0.0096371 | 81 | 2019 |
| 16 | 0.0091612 | 77 | / |
| 17 | 0.0089233 | 75 | news |
| 18 | 0.0078525 | 66 | campus |
| 19 | 0.0077335 | 65 | on |
| 20 | 0.0074955 | 63 | about |
| 21 | 0.0074955 | 63 | admissions |
| 22 | 0.0069007 | 58 | services |
| 23 | 0.0061868 | 52 | a |
| 24 | 0.0055919 | 47 | is |
| 25 | 0.0055919 | 47 | students |
| 26 | 0.004997 | 42 | at |
| 27 | 0.0047591 | 40 | views |
| 28 | 0.0040452 | 34 | january |
| 29 | 0.0039262 | 33 | menu |
| 30 | 0.0038073 | 32 | building |
| 31 | 0.0038073 | 32 | rose |
| 32 | 0.0038073 | 32 | float |
| 33 | 0.0036883 | 31 | events |
| 34 | 0.0035693 | 30 | academic |
| 35 | 0.0035693 | 30 | polycentric |
| 36 | 0.0033314 | 28 | give |
| 37 | 0.0033314 | 28 | professor |
| 38 | 0.0030934 | 26 | athletics |
| 39 | 0.0030934 | 26 | with |
| 40 | 0.0029744 | 25 | by |
| 41 | 0.0028554 | 24 | online |
| 42 | 0.0028554 | 24 | award |
| 43 | 0.0028554 | 24 | march |
| 44 | 0.0028554 | 24 | parade |
| 45 | 0.0027365 | 23 | research |
| 46 | 0.0027365 | 23 | will |
| 47 | 0.0027365 | 23 | center |
| 48 | 0.0027365 | 23 | college |
| 49 | 0.0026175 | 22 | library |
| 50 | 0.0026175 | 22 | february |
| 51 | 0.0026175 | 22 | » |
| 52 | 0.0024985 | 21 | directory |
| 53 | 0.0024985 | 21 | close |
| 54 | 0.0023795 | 20 | 15 |
| 55 | 0.0023795 | 20 | life |
| 56 | 0.0023795 | 20 | this |
| 57 | 0.0022606 | 19 | safety |
| 58 | 0.0022606 | 19 | more |
| 59 | 0.0022606 | 19 | from |
| 60 | 0.0022606 | 19 | president |
| 61 | 0.0021416 | 18 | 18 |
| 62 | 0.0021416 | 18 | maps |
| 63 | 0.0021416 | 18 | annual |
| 64 | 0.0021416 | 18 | website |
| 65 | 0.0021416 | 18 | giving |
| 66 | 0.0021416 | 18 | an |
| 67 | 0.0021416 | 18 | that |
| 68 | 0.0021416 | 18 | site |
| 69 | 0.0021416 | 18 | alumni |
| 70 | 0.0021416 | 18 | communications |
| 71 | 0.0021416 | 18 | media |
| 72 | 0.0020226 | 17 | skip |
| 73 | 0.0020226 | 17 | content |
| 74 | 0.0020226 | 17 | education |
| 75 | 0.0020226 | 17 | academics |
| 76 | 0.0020226 | 17 | visit |
| 77 | 0.0020226 | 17 | why |
| 78 | 0.0019036 | 16 | was |
| 79 | 0.0019036 | 16 | opens |
| 80 | 0.0019036 | 16 | day |
| 81 | 0.0019036 | 16 | quizzes |
| 82 | 0.0019036 | 16 | poly’s |
| 83 | 0.0019036 | 16 | development |
| 84 | 0.0017847 | 15 | all |
| 85 | 0.0017847 | 15 | new |
| 86 | 0.0017847 | 15 | be |
| 87 | 0.0017847 | 15 | department |
| 88 | 0.0017847 | 15 | affairs |
| 89 | 0.0017847 | 15 | product |
| 90 | 0.0016657 | 14 | activities |
| 91 | 0.0016657 | 14 | faculty |
| 92 | 0.0016657 | 14 | than |
| 93 | 0.0016657 | 14 | lanterman |
| 94 | 0.0016657 | 14 | tags |
| 95 | 0.0016657 | 14 | her |
| 96 | 0.0015467 | 13 | 7 |
| 97 | 0.0015467 | 13 | graduate |
| 98 | 0.0015467 | 13 | housing |
| 99 | 0.0015467 | 13 | kellogg |
| 100 | 0.0015467 | 13 | pomona’s |