Team:
Kaylie Fukumoto
Ethan Madjus
Megan Tsoi

Web Crawler Final Report

Over the past few weeks, our team developed a web crawler to crawl through various web pages from four different domains within UCI's computer science department. After testing and running the crawler, we collected data including number of unique pages crawled, longest page crawled, top 50 most common words that appeared within all the pages crawled, and number of subdomains and unique pages per subdomain crawled within the ics.uci.edu domain.

After sending the crawler through four domains (ics, cs, informatics, stat), we discovered that there were 28,154 unique web pages. Out of these thousands of web pages, the longest website found was https://cml.ics.uci.edu/category/aiml, with the number of words on this page totalling 17,228 words. Although there were thousands of different words appearing throughout the various pages, these were the 50 most common words (excluding stop words) that kept appearing shown in order from most common word to least common of the 50 words:

1) data -> 25492
2) uci -> 12201
3) research -> 10106
4) learning -> 9535
5) computer -> 9369
6) information -> 8938
7) set -> 8307
8) dataset -> 7986
9) contact -> 7237
10) machine -> 6845
11) new -> 6679
12) ics -> 6529
13) science -> 6382
14) students -> 6077
15) time -> 5935
16) systems -> 5779
17) policy -> 5441
18) view -> 5404
19) school -> 5128
20) news -> 5124
21) university -> 5108
22) software -> 5106
23) sciences -> 5013
24) classification -> 4961
25) using -> 4602

26) repository -> 4471
27) project -> 4403
28) based -> 4391
29) bren -> 4355
30) web -> 4277
31) events -> 4149
32) us -> 3970
33) markellekelly -> 3965
34) class -> 3885
35) irvine -> 3885
36) cs -> 3874
37) student -> 3842
38) one -> 3794
39) type -> 3744
40) people -> 3742
41) computing -> 3690
42) engineering -> 3645
43) real -> 3644
44) sets -> 3631
45) html -> 3584
46) edu -> 3531
47) search -> 3531
48) undo -> 3523
49) use -> 3520
50) social -> 3444

In addition to these findings collected from all four domains, there was also some additional data collected specifically from the ics.uci.edu domain. There were 139 subdomains found within the ics domain. Listed below are the 139 subdomains of the ics domain, accompanied by the number of unique web pages found with the corresponding subdomain:

1) https://acoi.ics.uci.edu, 74
2) https://aiclub.ics.uci.edu, 2
3) https://alumni.ics.uci.edu, 1
4) https://archive-beta.ics.uci.edu, 3
5) https://archive.ics.uci.edu, 17285
6) https://asterix.ics.uci.edu, 13
7) https://asterixdb.ics.uci.edu, 2
8) https://awareness.ics.uci.edu, 1
9) https://benzene-34.ics.uci.edu, 1
10) https://betapro.ics.uci.edu, 3
11) https://calendar.ics.uci.edu, 1
12) https://cbcl.ics.uci.edu, 637
13) https://cdb.ics.uci.edu, 48

14) https://cert.ics.uci.edu, 2
15) https://cgvw.ics.uci.edu, 1
16) https://checkmate.ics.uci.edu, 1
17) https://chemdb.ics.uci.edu, 1
18) https://chenli.ics.uci.edu, 10
19) https://cherry.ics.uci.edu, 1
20) https://chime.ics.uci.edu, 1
21) https://circadiomics.ics.uci.edu, 7
22) https://closeup.ics.uci.edu, 1
23) https://cloudberry.ics.uci.edu, 2
24) https://cml.ics.uci.edu, 97
25) https://cocoa-krispies.ics.uci.edu, 1
26) https://code.ics.uci.edu, 13
27) https://codeexchange.ics.uci.edu, 1
28) https://computableplant.ics.uci.edu, 46
29) https://contact.ics.uci.edu, 2
30) https://contact14.ics.uci.edu, 1
31) https://coronavirustwittermap.ics.uci.edu, 1
32) https://cradl.ics.uci.edu, 28
33) https://create.ics.uci.edu, 7
34) https://cwicsocal18.ics.uci.edu, 12
35) https://cyberclub.ics.uci.edu, 1
36) https://cybert.ics.uci.edu, 1
37) https://dataguard.ics.uci.edu, 1
38) https://datalab.ics.uci.edu, 1
39) https://dataprotector.ics.uci.edu, 1
40) https://dblp.ics.uci.edu, 2
41) https://dgillen.ics.uci.edu, 23
42) https://duke.ics.uci.edu, 1
43) https://duttgroup.ics.uci.edu, 75
44) https://dynamo.ics.uci.edu, 35
45) https://elms.ics.uci.edu, 84
46) https://emj.ics.uci.edu, 45
47) https://esl.ics.uci.edu, 7
48) https://evoke.ics.uci.edu, 7
49) https://flamingo.ics.uci.edu, 50
50) https://fr.ics.uci.edu, 9
51) https://futurehealth.ics.uci.edu, 50
52) https://gonet.ics.uci.edu, 1
53) https://grape.ics.uci.edu, 1235
54) https://graphics.ics.uci.edu, 12
55) https://graphmod.ics.uci.edu, 2
56) https://hack.ics.uci.edu, 1
57) https://hana.ics.uci.edu, 2

58) https://hobbes.ics.uci.edu, 1
59) https://hombao.ics.uci.edu, 2
60) https://honors.ics.uci.edu, 2
61) https://hpi.ics.uci.edu, 3
62) https://i-sensorium.ics.uci.edu, 6
63) https://iasl.ics.uci.edu, 22
64) https://ibook.ics.uci.edu, 8
65) https://ieee.ics.uci.edu, 1
66) https://industryshowcase.ics.uci.edu, 7
67) https://informatics.ics.uci.edu, 1
68) https://intranet.ics.uci.edu, 10
69) https://ipubmed.ics.uci.edu, 1
70) https://isg.ics.uci.edu, 138
71) https://jgarcia.ics.uci.edu, 23
72) https://jujube.ics.uci.edu, 2
73) https://kdd.ics.uci.edu, 4
74) https://luci.ics.uci.edu, 4
75) https://mailman.ics.uci.edu, 2
76) https://malek.ics.uci.edu, 1
77) https://map125.ics.uci.edu, 1
78) https://mapgrid.ics.uci.edu, 2
79) https://mcs.ics.uci.edu, 81
80) https://mdogucu.ics.uci.edu, 7
81) https://mds.ics.uci.edu, 3
82) https://metaviz.ics.uci.edu, 1
83) https://mhcid.ics.uci.edu, 19
84) https://mine10.ics.uci.edu, 1
85) https://mine5.ics.uci.edu, 1
86) https://mlphysics.ics.uci.edu, 32
87) https://mondego.ics.uci.edu, 13
88) https://motifmap-rna.ics.uci.edu, 2
89) https://motifmap.ics.uci.edu, 2
90) https://mse.ics.uci.edu, 3
91) https://mswe.ics.uci.edu, 14
92) https://mupro.ics.uci.edu, 3
93) https://nalini.ics.uci.edu, 7
94) https://ngs.ics.uci.edu, 556
95) https://old-reactions.ics.uci.edu, 7
96) https://omni.ics.uci.edu, 1
97) https://password.ics.uci.edu, 1
98) https://pasteur.ics.uci.edu, 2
99) https://pepito.ics.uci.edu, 2
100) https://perennialpolycultures.ics.uci.edu, 1
101) https://plrg.ics.uci.edu, 16

102) https://psearch.ics.uci.edu, 4
103) https://radicle.ics.uci.edu, 1
104) https://reactions.ics.uci.edu, 1
105) https://redmiles.ics.uci.edu, 1
106) https://scale.ics.uci.edu, 7
107) https://sconce.ics.uci.edu, 1
108) https://scratch.ics.uci.edu, 2
109) https://sdcl.ics.uci.edu, 184
110) https://se.ics.uci.edu, 1
111) https://seal.ics.uci.edu, 6
112) https://selectpro.ics.uci.edu, 6
113) https://seraja.ics.uci.edu, 1
114) https://sherlock.ics.uci.edu, 8
115) https://sidepro.ics.uci.edu, 1
116) https://sli.ics.uci.edu, 1086
117) https://sourcerer.ics.uci.edu, 2
118) https://sprout.ics.uci.edu, 1
119) https://statconsulting.ics.uci.edu, 5
120) https://student-council.ics.uci.edu, 1
121) https://studentcouncil.ics.uci.edu, 31
122) https://students.ics.uci.edu, 1
123) https://support.ics.uci.edu, 1
124) https://swiki.ics.uci.edu, 2
125) https://tastier.ics.uci.edu, 1
126) https://tippers.ics.uci.edu, 1
127) https://tippersweb.ics.uci.edu, 4
128) https://tmbpro.ics.uci.edu, 1
129) https://transformativeplay.ics.uci.edu, 51
130) https://transformativeplay.ics.uci.edu, 1
131) https://unite.ics.uci.edu, 10
132) https://vision.ics.uci.edu, 206
133) https://wearablegames.ics.uci.edu, 12
134) https://wics.ics.uci.edu, 328
135) https://wiki.ics.uci.edu, 1
136) https://www-db.ics.uci.edu, 44
137) https://www.ics.uci.edu, 4374
138) https://xtune.ics.uci.edu, 1
139) https://yarra.ics.uci.edu, 1