CS 121 Information Retrieval – Spring 2024
Einar Gatchalian ID: 15148609
Jefferson McLinden ID: 10978555
Hearty Parrenas ID: 20534693

Assignment 2: Web Crawler Report

1. The number of unique pages found were 11,386 pages.

2. The longest page in terms of the number of words is: http://www.ics.uci.edu/~shantas/publications/19-obscure_secret-sharing_shantanu-sharma_vldb.ppsx. There were 120,386 tokens found on this page.

3. The 50 most common words with their respective frequencies, excluding English stop words are:

1. research – 58,321
2. ics – 44,585
3. computer – 39,095
4. science – 36,996
5. uci – 35,961
6. news – 33,194
7. student – 32,147
8. graduate – 27,081
9. learning – 25,273
10. students – 24,644
11. undergraduate – 24,226
12. us – 23,556
13. data – 22,570
14. faculty – 21,899
15. school – 21,705
16. informatics – 20,429
17. bren – 19,983
18. events – 18,524
19. information – 18,304
20. community – 16,852
21. software – 16,205
22. view – 15,846
23. donald – 15,029
24. 2022 – 15,011
25. academic – 14,860
26. alumni – 14,589
27. computing – 14,478
28. 2023 – 14,247
29. irvine – 13,972

30. statistics – 13,504
31. 2021 – 13,128
32. programs – 13,018
33. machine – 12,552
34. policies – 12,457
35. spotlights – 11,795
36. sciences – 11,428
37. contact – 11,296
38. policy – 11,127
39. projects – 10,553
40. october – 10,354
41. get – 10,178
42. systems – 10,070
43. future – 9,806
44. follow – 9,686
45. health – 9,631
46. people – 9,552
47. corporate – 9,398
48. honors – 9,046
49. 2020 – 9,037
50. support – 8,901

4. There were 90 subdomains found in the ics.uci.edu domain. The subdomains along with the number of pages for each subdomain are:

1. accessibility.ics.uci.edu – 2
2. acoi.ics.uci.edu – 41
3. aiclub.ics.uci.edu – 2
4. archive.ics.uci.edu – 4
5. asterix.ics.uci.edu – 4
6. betapro.proteomics.ics.uci.edu – 3
7. biaslab.ics.uci.edu – 1
8. cbcl.ics.uci.edu – 403
9. cdb.ics.uci.edu – 1
10. cert.ics.uci.edu – 10
11. chenli.ics.uci.edu – 6
12. circadiomics.ics.uci.edu – 4
13. cml.ics.uci.edu – 37
14. computableplant.ics.uci.edu – 25
15. courselisting.ics.uci.edu – 2
16. cradl.ics.uci.edu – 15
17. create.ics.uci.edu – 1
18. cwicsocal18.ics.uci.edu – 8
19. cybert.ics.uci.edu – 1

20. dejavu.ics.uci.edu – 1
21. dgillen.ics.uci.edu – 9
22. ds4all.ics.uci.edu – 1
23. duttgroup.ics.uci.edu – 29
24. edgelab.ics.uci.edu – 3
25. elms.ics.uci.edu – 1
26. emj.ics.uci.edu – 40
27. evoke.ics.uci.edu – 1
28. flamingo.ics.uci.edu – 5
29. fr.ics.uci.edu – 1
30. futurehealth.ics.uci.edu – 55
31. graphics.ics.uci.edu – 7
32. graphmod.ics.uci.edu – 1
33. hack.ics.uci.edu – 2
34. hai.ics.uci.edu – 5
35. iasl.ics.uci.edu – 3
36. ibook.ics.uci.edu – 1
37. ieee.ics.uci.edu – 1
38. industryshowcase.ics.uci.edu – 4
39. informatics.mt-live.ics.uci.edu – 1
40. insite.ics.uci.edu – 1
41. intranet.ics.uci.edu – 3
42. ipubmed.ics.uci.edu – 1
43. isg.ics.uci.edu – 19
44. jgarcia.ics.uci.edu – 7
45. keys.ics.uci.edu – 1
46. luci.ics.uci.edu – 2
47. malek.ics.uci.edu – 1
48. mcs.ics.uci.edu – 28
49. mdogucu.ics.uci.edu – 6
50. mds.ics.uci.edu – 2
51. mhcid.ics.uci.edu – 15
52. mlphysics.ics.uci.edu – 13
53. mondego.ics.uci.edu – 2
54. motifmap-rna.ics.uci.edu – 2
55. motifmap.ics.uci.edu – 1
56. mse.ics.uci.edu – 2
57. mt-live.ics.uci.edu – 1
58. mupro.proteomics.ics.uci.edu – 3
59. nalini.ics.uci.edu – 7
60. ngs.ics.uci.edu – 375
61. oai.ics.uci.edu – 2
62. old-reactions.ics.uci.edu – 1
63. pepito.proteomics.ics.uci.edu – 1
64. plrg.ics.uci.edu – 16
65. psearch.ics.uci.edu – 1
66. radicle.ics.uci.edu – 6
67. reactions.ics.uci.edu – 1
68. redmiles.ics.uci.edu – 4
69. scratch.proteomics.ics.uci.edu – 2

70. sdcl.ics.uci.edu – 166
71. seal.ics.uci.edu – 5
72. sherlock.ics.uci.edu – 7
73. sli.ics.uci.edu – 2
74. sourcerer.ics.uci.edu – 1
75. stairs.ics.uci.edu – 3
76. statconsulting.ics.uci.edu – 4
77. student-council.ics.uci.edu – 1
78. studentcouncil.ics.uci.edu – 27
79. summeracademy.ics.uci.edu – 2
80. swiki.ics.uci.edu – 2
81. tad.ics.uci.edu – 3
82. tippersweb.ics.uci.edu – 4
83. transformativeplay.ics.uci.edu – 13
84. tutors.ics.uci.edu – 1
85. ugradforms.ics.uci.edu – 1
86. unite.ics.uci.edu – 7
87. vision.ics.uci.edu – 37
88. wics.ics.uci.edu – 246
89. wiki.ics.uci.edu – 41
90. www-db.ics.uci.edu – 6