Anthony Wong & Nicholas Evangelos Georggin

Professor Iftekhar Ahmed

SWE 225 / CS 221: Information Retrieval

09 February 2025

SWE 225 / CS 221 Assignment 2 Final Report

As a concrete deliverable of this project, besides the code itself, you must submit a report containing answers to the following questions:

1. How many unique pages did you find? Uniqueness is established by the URL, but discarding the fragment part. So, for example, *http://www.ics.uci.edu#aaa* and *http://www.ics.uci.edu#bbb* are the same URL.

We found 2,329 unique pages in our crawl.

2. What is the longest page in terms of number of words? (HTML markup doesn't count as words)

The longest page in terms of number of words – alphabetic tokens not including HTML markup – is https://www.stat.uci.edu/covid19/index.html, containing 19,604 words.

3. What are the 50 most common words in the entire set of pages? (**Ignore English stop words**, which can be found, for example, here) Submit the list of common words ordered by frequency.

The 50 most common words in the entire set of pages are as follows.

1. student, 13038
2. software, 11468
3. informatics, 10041
4. engineering, 9716
5. undergraduate, 9555
6. graduate, 9241
7. july, 8731
8. june, 8705
9. september, 8660
10. march, 7940
11. november, 7908
12. april, 7865
13. january, 7775

14. february, 7741
15. december, 7685
16. department, 7650
17. august, 7646
18. october, 7244
19. design, 6902
20. students, 6578
21. computer, 6332
22. profiles, 6051
23. projects, 5413
24. support, 4961
25. news, 4857
26. courses, 4819
27. alumni, 4714
28. people, 4623
29. master, 4613
30. books, 4605
31. process, 4560
32. application, 4555
33. data, 4538
34. spotlights, 4488
35. degrees, 4411
36. groups, 4248
37. policies, 4223
38. science, 4197
39. uci, 4106
40. game, 4101
41. markellekelly, 3928
42. ics, 3747
43. faculty, 3549
44. post, 3259
45. university, 3255
46. future, 3135
47. read, 3104
48. opportunities, 3006
49. irvine, 2999
50. site, 2915

4. How many subdomains did you find in the ics.uci.edu domain? Submit the list of subdomains ordered alphabetically and the number of unique pages detected in each

subdomain. The content of this list should be lines containing *URL*, *number,* for example: http://vision.ics.uci.edu, 10 (not the actual number here)

We found 76 unique subdomains of ics.uci.edu, which are listed in the prompted format below.

- accessibility.ics.uci.edu, 2
- acoi.ics.uci.edu, 1
- aiclub.ics.uci.edu, 1
- archive.ics.uci.edu, 1
- asterixdb.ics.uci.edu, 1
- cert.ics.uci.edu, 1
- cgvw.ics.uci.edu, 1
- checkmate.ics.uci.edu, 1
- chenli.ics.uci.edu, 3
- cloudberry.ics.uci.edu, 2
- cml.ics.uci.edu, 2
- code.ics.uci.edu, 1
- codeexchange.ics.uci.edu, 2
- coronavirustwittermap.ics.uci.edu, 1
- courselisting.ics.uci.edu, 2
- create.ics.uci.edu, 4
- cs.ics.uci.edu, 8
- cyberclub.ics.uci.edu, 1
- dejavu.ics.uci.edu, 1
- dgillen.ics.uci.edu, 26
- ds4all.ics.uci.edu, 3
- duttgroup.ics.uci.edu, 89
- dynamo.ics.uci.edu, 1
- evoke.ics.uci.edu, 4
- flamingo.ics.uci.edu, 1
- fr.ics.uci.edu, 2
- futurehealth.ics.uci.edu, 5
- grape.ics.uci.edu, 12
- graphics.ics.uci.edu, 2
- graphmod.ics.uci.edu, 1
- hack.ics.uci.edu, 1
- hai.ics.uci.edu, 3
- hana.ics.uci.edu, 3
- honors.ics.uci.edu, 1
- hpi.ics.uci.edu, 1
- i-sensorium.ics.uci.edu, 1
- iasl.ics.uci.edu, 2

- icde2023.ics.uci.edu, 1
- informatics.ics.uci.edu, 2
- insite.ics.uci.edu, 1
- isg.ics.uci.edu, 2
- jgarcia.ics.uci.edu, 6
- luci.ics.uci.edu, 4
- mailman.ics.uci.edu, 2
- malek.ics.uci.edu, 1
- mcs.ics.uci.edu, 10
- mdogucu.ics.uci.edu, 1
- mds.ics.uci.edu, 26
- mhcid.ics.uci.edu, 15
- mondego.ics.uci.edu, 5
- mse.ics.uci.edu, 1
- mswe.ics.uci.edu, 9
- nalini.ics.uci.edu, 7
- ngs.ics.uci.edu, 207
- oai.ics.uci.edu, 1
- redmiles.ics.uci.edu, 1
- riscit.ics.uci.edu, 1
- sdcl.ics.uci.edu, 62
- se.ics.uci.edu, 1
- seal.ics.uci.edu, 2
- sherlock.ics.uci.edu, 2
- sli.ics.uci.edu, 5
- sourcerer.ics.uci.edu, 1
- sprout.ics.uci.edu, 1
- stairs.ics.uci.edu, 1
- statconsulting.ics.uci.edu, 3
- student-council.ics.uci.edu, 2
- summeracademy.ics.uci.edu, 1
- tippersweb.ics.uci.edu, 1
- transformativeplay.ics.uci.edu, 51
- tutoring.ics.uci.edu, 1
- unite.ics.uci.edu, 10
- vision.ics.uci.edu, 1
- wearablegames.ics.uci.edu, 5
- wics.ics.uci.edu, 10
- xtune.ics.uci.edu, 6