

How many unique pages did you find? Uniqueness for the purposes of this assignment is ONLY 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. Even if you implement additional methods for textual similarity detection, please keep considering the above definition of unique pages for the purposes of counting the unique pages in this assignment.

5,426 unique pages

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

151,288 words - <https://evoke.ics.uci.edu/qs-personal-data-landscapes-poster>

What are the 50 most common words in the entire set of pages crawled under these domains? (Ignore English stop words, which can be found, for example, here ([Links to an external site.](#))) Submit the list of common words ordered by frequency.

[
('2020', 19047),
('computer', 15860),
('software', 14883),
('student', 14167),
('2021', 13761),
('2018', 13676),
('ramesh', 13358),
('2019', 13005),
('students', 12772),
('2016', 12383),
('1', 11842),
('2017', 11740),
('will', 11540),
('reply', 11036),
('engineering', 11006),
('2015', 10943),
('2', 10752),
('informatics', 10525),
('uci', 10246),
('news', 10076),
('data', 9896),
('graduate', 9361),
('read', 9288),
('march', 9263),
('undergraduate', 9078),
('3', 8837),

```
('courses', 8756),
('april', 8708),
('design', 8489),
('january', 8489),
('science', 8426),
('february', 8267),
('books', 8261),
('ph', 8214),
('people', 8164),
('november', 8085),
('september', 7988),
('current', 7972),
('december', 7971),
('projects', 7964),
('department', 7961),
('events', 7937),
('october', 7628),
('june', 7590),
('july', 7354),
('august', 7313),
('4', 7300),
('5', 7125),
('search', 7077),
('computing', 7076)
]
```

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)

1. ('acoi.ics.uci.edu', 55)
2. ('aiclub.ics.uci.edu', 1)
3. ('archive.ics.uci.edu', 3)
4. ('asterix.ics.uci.edu', 5)
5. ('cbcl.ics.uci.edu', 3)
6. ('cert.ics.uci.edu', 1)
7. ('chenli.ics.uci.edu', 1)
8. ('cloudberry.ics.uci.edu', 23)
9. ('cml.ics.uci.edu', 162)
10. ('computableplant.ics.uci.edu', 12)
11. ('cradl.ics.uci.edu', 18)

12. ('create.ics.uci.edu', 3)
13. ('cwicsocal18.ics.uci.edu', 6)
14. ('cyberclub.ics.uci.edu', 1)
15. ('dejavu.ics.uci.edu', 1)
16. ('dynamo.ics.uci.edu', 1)
17. ('emj.ics.uci.edu', 41)
18. ('esl.ics.uci.edu', 1)
19. ('evoke.ics.uci.edu', 75)
20. ('flamingo.ics.uci.edu', 6)
21. ('fr.ics.uci.edu', 2)
22. ('frost.ics.uci.edu', 1)
23. ('futurehealth.ics.uci.edu', 2)
24. ('grape.ics.uci.edu', 12)
25. ('graphics.ics.uci.edu', 2)
26. ('graphmod.ics.uci.edu', 1)
27. ('hai.ics.uci.edu', 2)
28. ('hana.ics.uci.edu', 13)
29. ('i-sensorium.ics.uci.edu', 1)
30. ('iasl.ics.uci.edu', 1)
31. ('informatics.ics.uci.edu', 1)
32. ('ipubmed.ics.uci.edu', 1)
33. ('isg.ics.uci.edu', 104)
34. ('jgarcia.ics.uci.edu', 6)
35. ('luci.ics.uci.edu', 3)
36. ('mailman.ics.uci.edu', 2)
37. ('malek.ics.uci.edu', 2)
38. ('mcs.ics.uci.edu', 38)
39. ('mds.ics.uci.edu', 8)
40. ('mhcid.ics.uci.edu', 10)
41. ('mondego.ics.uci.edu', 3)
42. ('mse.ics.uci.edu', 1)
43. ('mswe.ics.uci.edu', 22)
44. ('nalini.ics.uci.edu', 3)
45. ('ngs.ics.uci.edu', 1877)
46. ('perennialpolycultures.ics.uci.edu', 1)
47. ('plrg.ics.uci.edu', 13)
48. ('redmiles.ics.uci.edu', 5)
49. ('riscit.ics.uci.edu', 1)
50. ('scale.ics.uci.edu', 1)
51. ('sdcl.ics.uci.edu', 152)
52. ('seal.ics.uci.edu', 5)
53. ('sherlock.ics.uci.edu', 1)
54. ('sli.ics.uci.edu', 212)
55. ('sourcerer.ics.uci.edu', 1)
56. ('stairs.ics.uci.edu', 1)
57. ('statconsulting.ics.uci.edu', 4)
58. ('support.ics.uci.edu', 1)

59. ('tastier.ics.uci.edu', 1)
60. ('transformativeplay.ics.uci.edu', 41)
61. ('vision.ics.uci.edu', 6)
62. ('wearablegames.ics.uci.edu', 9)
63. ('wics.ics.uci.edu', 450)
64. ('www-db.ics.uci.edu', 5)
65. ('www.cert.ics.uci.edu', 1)
66. ('www.graphics.ics.uci.edu', 1)
67. ('www.ics.uci.edu', 758)
68. ('www.informatics.ics.uci.edu', 1)
69. ('www.informatics.uci.edu', 957)
70. ('xtune.ics.uci.edu', 2)

Student ID:
Regina: 58032813

Valerie
X0861751

Alexander Lee
65942051