Kienlac Mai

Gavin Nguyen

Final Web Crawler Report

1. 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.

We found 27,950 different unique pages listed as subdomains in our web crawling.

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

The largest url was "http://www.ics.uci.edu/~cs224" and the word cound is 72,372.

3. 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.

Rank	Word	Count
1	loading	91,235
2	support	67,161
3	changes	49,121
4	file	41,124
5	side	39,764

6	gitlab	37,470
7	sign	36,785
8	comment	36,759
9	cancel	36,679
10	hans	36,233
11	wiki	34,094
12	files	33,748
13	ics	29,107
14	tools	28,653
15	research	26,922
16	space	26,524
17	editing	18,443
18	explore	18,434
19	skip	18,364
20	commits	18,306
21	whitespace	18,298
22	attach	18,257
23	authored	18,251
24	discussion	18,250
25	finish	18,237

26	preview	18,231
27	mars	18,228
28	browse	18,227
29	proceed	18,215
30	caution	18,212
31	parent	18,112
32	page	17,740
33	user	17,091
34	code	15,733
35	security	15,258
36	l	15,247
37	list	14,882
38	merge	14,511
39	profile	14,480
40	settings	14,422
41	group	14,406
42	readme	14,203
43	edit	14,115
44	issues	14,103
45	branches	14,095

46	submit	14,082
47	container	14,067
48	images	14,062
49	remove	14,058
50	help	14,055

4. How many subdomains did you find in the 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 subdomain**,** number, for example: vision.ics.uci.edu, 10 (not the actual number here)

Subdomain	**Number**
asterix.ics.uci.edu	5
chenli.ics.uci.edu	2
flamingo.ics.uci.edu	17
fr.ics.uci.edu	3
gitlab.ics.uci.edu	18288
ics.uci.edu	6
intranet.ics.uci.edu	17
myip.ics.uci.edu	1
swiki.ics.uci.edu	5943

tastier.ics.uci.edu	1
wiki.ics.uci.edu	3555
www.ics.uci.edu,	7
www.stat.uci.edu	105