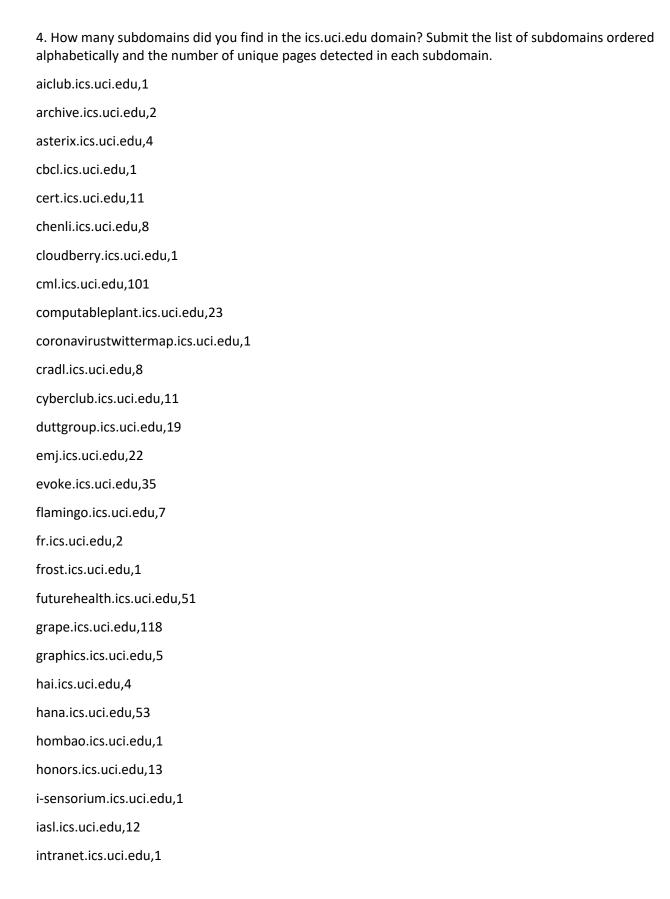
SWE225/COMPSCI221 Assignment 2 Report

Michael Kahn, Andrew Truelove, Jirigesi

- 1. How many unique pages did you find? Uniqueness is established by the URL, but discarding the fragment part.
- A: We found 4533 unique pages in our crawl.
- 2. What is the longest page in terms of number of words?
- A: We found https://cml.ics.uci.edu/category/aiml/page/2 has the most words, with a total count of 11,175 words.
- 3. What are the 50 most common words in the entire set of pages? Submit the list of common words ordered by frequency.

A: research, courses, student, department, information, january, computer, design, software, projects, 2020, past, students, books, data, current, 2019, search, uci, university, 2018, 2021, 2017, new, 2015, november, engineering, june, 2016, september, science, course, may, july, ramesh, october, informatics, december, bs, august, undergraduate, learning, graduate, irvine, phd, read, people, march, news, 2014

| Word(rank 1-25) | Frequency | Word(rank 26-50) | Frequency |
|-----------------|-----------|------------------|-----------|
| research | 18734 | courses | 6132 |
| student | 10275 | department | 6101 |
| information | 10058 | january | 6081 |
| computer | 10044 | design | 5721 |
| software | 9720 | projects | 5497 |
| 2020 | 9293 | past | 5361 |
| students | 8508 | books | 5243 |
| data | 8280 | current | 5209 |
| 2019 | 8030 | search | 5206 |
| uci | 7848 | university | 5071 |
| 2018 | 7843 | 2021 | 5030 |
| 2017 | 7839 | new | 4978 |
| 2015 | 7671 | november | 4864 |
| engineering | 7549 | june | 4830 |
| 2016 | 7456 | september | 4818 |
| science | 7157 | course | 4712 |
| may | 6918 | july | 4648 |
| ramesh | 6896 | october | 4648 |
| informatics | 6848 | december | 4605 |
| bs | 6786 | august | 4584 |
| undergraduate | 6674 | learning | 4528 |
| graduate | 6596 | irvine | 4500 |
| phd | 6222 | read | 4341 |
| people | 6210 | march | 4294 |
| news | 6176 | 2014 | 4226 |



ipubmed.ics.uci.edu,1

jgarcia.ics.uci.edu,1

mailman.ics.uci.edu,1

malek.ics.uci.edu,1

mdogucu.ics.uci.edu,5

mds.ics.uci.edu,3

mhcid.ics.uci.edu,7

mondego.ics.uci.edu,3

ngs.ics.uci.edu,1414

password.ics.uci.edu,1

perennialpolycultures.ics.uci.edu,1

plrg.ics.uci.edu,13

psearch.ics.uci.edu,1

radicle.ics.uci.edu,1

redmiles.ics.uci.edu,5

riscit.ics.uci.edu,2

sconce.ics.uci.edu,2

sdcl.ics.uci.edu,104

seal.ics.uci.edu,1

sherlock.ics.uci.edu,5

sli.ics.uci.edu,189

sprout.ics.uci.edu,14

stairs.ics.uci.edu,2

statconsulting.ics.uci.edu,2

studentcouncil.ics.uci.edu,24

tastier.ics.uci.edu,1

transformativeplay.ics.uci.edu,14

tutors.ics.uci.edu,1

vision.ics.uci.edu,174

wics.ics.uci.edu,31

www-db.ics.uci.edu,3