Assignment 2: Web Crawler

Group Members:
Julie Bui (ID: 59422563)
LoLa Alexis Kim (ID: 52727368)
Maxwell Shih (ID: 88195254)
Axel Hugo Erik Wennerlund (ID: 56683632)

How many unique pages did you find?
During the final crawl, we found 10,115 unique pages.

What is the longest page in terms of number of words?
The longest page was https://cml.ics.uci.edu/category/aiml, with a number of 16515 words.

To find this page, we implemented code that would not consider HTML markup tags like words. We also stripped out script and style tags, as well as hidden text within meta tags, svg tags, etc. In other words, our count of words is only based on *visible* content on the web page - which we thought was the most reasonable thing to look at.

What are the 50 most common words in the entire set of pages, excluding stop words?
Below, we list the 50 most common words that appeared during the crawl, excluding stop words. Pure numbers were not considered as valid words (tokens), except when they appeared within a word (e.g., "H2O"). One observation is that the sixth most common word is "s," likely resulting from words that end with 's (related to possessive suffixes).

research, 12803
ics, 11793
events, 11345
data, 10360
us, 9735
s, 8829
information, 8692
computer, 8692
ramesh, 8167
students, 7992
uci, 7594
science, 7453
current, 6952
learning, 6864
will, 6554
contact, 6463
university, 6356
news, 6185

computing, 5930
wics, 5804
search, 5674
past, 5177
code, 5123
student, 4900
can, 4801
systems, 4667
community, 4637
irvine, 4479
new, 4471
time, 4369
home, 4330
faculty, 4237
projects, 4222
machine, 4189
courses, 4175
pm, 4127
people, 4073
graduate, 3992
markellekelly, 3928
program, 3797
jain, 3789
statistics, 3755
one, 3719
companies, 3641
health, 3515
books, 3508
center, 3448
innovate, 3435
may, 3405
using, 3342

How many subdomains did you find in the uci.edu domain?

We found 114 subdomains, listed below with the number of unique pages detected in each. It is worth noting that, for instance, www.ics.uci.edu and ics.uci.edu have been counted as two different subdomains. Even if this might seem redundant, they are technically separate hostnames, and while they often point to the same site, this is not always guaranteed.

acoi.ics.uci.edu, 1
aiclub.ics.uci.edu, 1
archive-beta.ics.uci.edu, 12
archive.ics.uci.edu, 212

asterix.ics.uci.edu, 15
auge.ics.uci.edu, 2
cert.ics.uci.edu, 19
cgvw.ics.uci.edu, 1
checkin.ics.uci.edu, 6
chenli.ics.uci.edu, 12
cherry.ics.uci.edu, 1
chime.ics.uci.edu, 1
circinus-14.ics.uci.edu, 1
cloudberry.ics.uci.edu, 91
cml.ics.uci.edu, 352
code.ics.uci.edu, 30
containers.ics.uci.edu, 1
coronavirustwittermap.ics.uci.edu, 2
courselisting.ics.uci.edu, 4
create.ics.uci.edu, 14
cs.ics.uci.edu, 23
cs.uci.edu, 2
cwicsocal18.ics.uci.edu, 23
dgillen.ics.uci.edu, 58
ds4all.ics.uci.edu, 8
duttgroup.ics.uci.edu, 246
elms.ics.uci.edu, 1
esl.ics.uci.edu, 1
fr.ics.uci.edu, 13
futurehealth.ics.uci.edu, 198
gitlab.ics.uci.edu, 5
gradinfo.ics.uci.edu, 3
grafana-infra-blue.ics.uci.edu, 1
grape.ics.uci.edu, 11
hack.ics.uci.edu, 1
helpdesk.ics.uci.edu, 4
hobbes.ics.uci.edu, 12
hombao.ics.uci.edu, 2
hpi.ics.uci.edu, 10
hub.ics.uci.edu, 5
iasl.ics.uci.edu, 2
icde2023.ics.uci.edu, 92
ics.uci.edu, 595
industryshowcase.ics.uci.edu, 44
informatics.uci.edu, 2
instdav.ics.uci.edu, 1
intranet.ics.uci.edu, 4
ipubmed.ics.uci.edu, 2

isg.ics.uci.edu, 218
jgarcia.ics.uci.edu, 42
julia-hub.ics.uci.edu, 2
kdd.ics.uci.edu, 1
kpassword.ics.uci.edu, 1
luci.ics.uci.edu, 6
mailman.ics.uci.edu, 20
malek.ics.uci.edu, 2
mapgrid.ics.uci.edu, 1
mcs.ics.uci.edu, 25
mdogucu.ics.uci.edu, 3
mds.ics.uci.edu, 18
mhcid.ics.uci.edu, 37
mhcis.ics.uci.edu, 1
mover.ics.uci.edu, 25
mswe.ics.uci.edu, 19
mt-live.ics.uci.edu, 1
nalini.ics.uci.edu, 14
netreg.ics.uci.edu, 3
ngs.ics.uci.edu, 3281
oai.ics.uci.edu, 11
observium.ics.uci.edu, 1
onboarding.ics.uci.edu, 4
password.ics.uci.edu, 9
pastebin.ics.uci.edu, 1
pgadmin.ics.uci.edu, 1
phpmyadmin.ics.uci.edu, 1
prometheus-infra-blue.ics.uci.edu, 1
psearch.ics.uci.edu, 4
rstudio-hub.ics.uci.edu, 1
satware.ics.uci.edu, 1
seal.ics.uci.edu, 54
seraja.ics.uci.edu, 1
sherlock.ics.uci.edu, 9
speedtest.ics.uci.edu, 1
sprout.ics.uci.edu, 1
staging-hub.ics.uci.edu, 2
stat.uci.edu, 3
statconsulting.ics.uci.edu, 8
student-council.ics.uci.edu, 1
studentcouncil.ics.uci.edu, 1
summeracademy.ics.uci.edu, 14
support.ics.uci.edu, 15
svn.ics.uci.edu, 1

swiki.ics.uci.edu, 182
tad.ics.uci.edu, 3
tastier.ics.uci.edu, 2
tippersweb.ics.uci.edu, 1
transformativeplay.ics.uci.edu, 2
tutoring.ics.uci.edu, 10
ugradforms.ics.uci.edu, 1
unite.ics.uci.edu, 20
vision.ics.uci.edu, 172
wics.ics.uci.edu, 2777
wiki.ics.uci.edu, 117
www-db.ics.uci.edu, 41
www.cs.uci.edu, 13
www.graphics.ics.uci.edu, 1
www.ics.uci.edu, 279
www.informatics.ics.uci.edu, 2
www.informatics.uci.edu, 21
www.isg.ics.uci.edu, 1
www.stat.uci.edu, 450
xtune.ics.uci.edu, 7