CS121: Assignment 2: Web Crawler

Jasper Doan, Maxwell Rehm, Sierra Martin
34455076, 71641940, 49838935

University of California, Irvine — October 27, 2024

Question 1 *(Unique Pages)*

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.

Our crawler found 25843 unique pages.

Question 2 *(Longest Page)*

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

"*http://www.ics.uci.edu/ shantas/publications/20-secret-sharing-aggregation-TKDE-shantanu*": 147340 words (tokens)

Question 3 *(Most Common Words)*

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, hereLinks to an external site.) Submit the list of common words ordered by frequency.

research: 79802
2021: 79622
2020: 74071
2018: 68213
2019: 66521
student: 65335
2022: 65179
informatics: 61792
2023: 60318
2017: 60215
2016: 58203
software: 58028
2015: 57093
may: 54239
june: 54066
ics: 53617
uci: 53496
september: 50816
july: 50133
engineering: 49205
march: 46643
january: 45892
april: 45699
february: 45641
november: 45054
december: 44680
graduate: 44214
august: 43906
undergraduate: 42781
october: 41426
ph: 41210
computer: 38023
department: 36995
students: 35704
support: 35330
will: 35081
design: 33334
information: 32451
projects: 29757
2014: 29261
data: 28827
profiles: 28815
news: 26431
read: 25697
continue: 25072
reading: 24842
university: 24613
years: 23961
courses: 23575
alumni: 23371

Question 4 *(Subdomain)*

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.

accessibility.ics.uci.edu: 7
acoi.ics.uci.edu: 110
aiclub.ics.uci.edu: 3
archive.ics.uci.edu: 91
asterix.ics.uci.edu: 14
asterixdb.ics.uci.edu: 2
auge.ics.uci.edu: 2
cbcl.ics.uci.edu: 45
cert.ics.uci.edu: 26
cgvw.ics.uci.edu: 1
checkin.ics.uci.edu: 5
checkmate.ics.uci.edu: 1
chenli.ics.uci.edu: 11
cherry.ics.uci.edu: 1
chime.ics.uci.edu: 1
cloudberry.ics.uci.edu: 49
cml.ics.uci.edu: 119
code.ics.uci.edu: 15
codeexchange.ics.uci.edu: 2
computableplant.ics.uci.edu: 87
containers.ics.uci.edu: 1
coronavirustwittermap.ics.uci.edu: 2
courselisting.ics.uci.edu: 4
cradl.ics.uci.edu: 1
create.ics.uci.edu: 5
cs.ics.uci.edu: 17
cs260p-hub.ics.uci.edu: 2
cs260p-staging-hub.ics.uci.edu: 2
cwicsocal18.ics.uci.edu: 13
cyberclub.ics.uci.edu: 8
dataguard.ics.uci.edu: 1
dataprotector.ics.uci.edu: 1
dblp.ics.uci.edu: 2
dejavu.ics.uci.edu: 2
dgillen.ics.uci.edu: 33
ds4all.ics.uci.edu: 4
duke.ics.uci.edu: 1
duttgroup.ics.uci.edu: 105
dynamo.ics.uci.edu: 21
eli.ics.uci.edu: 5
elms.ics.uci.edu: 1
emj-pc.ics.uci.edu: 1
emj.ics.uci.edu: 37
esl.ics.uci.edu: 6
evoke.ics.uci.edu: 6
flamingo.ics.uci.edu: 43
fr.ics.uci.edu: 13
futurehealth.ics.uci.edu: 122
gitlab.ics.uci.edu: 54
grafana-infra-blue.ics.uci.edu: 1
grape.ics.uci.edu: 497
graphics.ics.uci.edu: 18
graphmod.ics.uci.edu: 15
hack.ics.uci.edu: 2
hai.ics.uci.edu: 7
hana.ics.uci.edu: 3
helpdesk.ics.uci.edu: 2
hobbes.ics.uci.edu: 11
hombao.ics.uci.edu: 2
honors.ics.uci.edu: 1
hpi.ics.uci.edu: 6
hub.ics.uci.edu: 4
i-sensorium.ics.uci.edu: 6
iasl.ics.uci.edu: 2
icde2023.ics.uci.edu: 49
ics45c-hub.ics.uci.edu: 2
ics45c-staging-hub.ics.uci.edu: 2
ics46-hub.ics.uci.edu: 2
ics46-staging-hub.ics.uci.edu: 2
ics53-hub.ics.uci.edu: 2
ics53-staging-hub.ics.uci.edu: 2
ieee.ics.uci.edu: 3
industryshowcase.ics.uci.edu: 23
informatics.ics.uci.edu: 2
insite.ics.uci.edu: 9
instdav.ics.uci.edu: 1
intranet.ics.uci.edu: 3
ipubmed.ics.uci.edu: 2
isg.ics.uci.edu: 180
jgarcia.ics.uci.edu: 3
jujube.ics.uci.edu: 2
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: 7
malek.ics.uci.edu: 2
mapgrid.ics.uci.edu: 1
mcs.ics.uci.edu: 12
mdogucu.ics.uci.edu: 8
mds.ics.uci.edu: 30
metaviz.ics.uci.edu: 1
mhcid.ics.uci.edu: 22
mondego.ics.uci.edu: 16
motifmap-rna.ics.uci.edu: 1
motifmap.ics.uci.edu: 3
mse.ics.uci.edu: 1
mswe.ics.uci.edu: 12
mysqladm.ics.uci.edu: 1
nalini.ics.uci.edu: 8
ngs.ics.uci.edu: 127
oai.ics.uci.edu: 6
observium.ics.uci.edu: 1
omni.ics.uci.edu: 1
onboarding.ics.uci.edu: 1
password.ics.uci.edu: 1
pastebin.ics.uci.edu: 2
pgadmin.ics.uci.edu: 1
phpmyadmin.ics.uci.edu: 2
prometheus-infra-blue.ics.uci.edu: 1
psearch.ics.uci.edu: 6
pypmyadmin.ics.uci.edu: 1
radicle.ics.uci.edu: 7
redmiles.ics.uci.edu: 8
riscit.ics.uci.edu: 2
rstudio-hub.ics.uci.edu: 1
satware.ics.uci.edu: 1
sdcl.ics.uci.edu: 120
se.ics.uci.edu: 1
seal.ics.uci.edu: 46
sherlock.ics.uci.edu: 9
sli.ics.uci.edu: 408
sourcerer.ics.uci.edu: 2
speedtest.ics.uci.edu: 1
sprout.ics.uci.edu: 2
staging-hub.ics.uci.edu: 2
stairs.ics.uci.edu: 5
statconsulting.ics.uci.edu: 5
statistics-stage.ics.uci.edu: 1
student-council.ics.uci.edu: 15
summeracademy.ics.uci.edu: 2
support.ics.uci.edu: 1
svn.ics.uci.edu: 1
swiki.ics.uci.edu: 148
tad.ics.uci.edu: 3
tastier.ics.uci.edu: 2
tippers.ics.uci.edu: 2
tippersweb.ics.uci.edu: 3
transformativeplay.ics.uci.edu: 53
tutoring.ics.uci.edu: 6
tutors.ics.uci.edu: 4
ugradforms.ics.uci.edu: 1
unite.ics.uci.edu: 11
vision.ics.uci.edu: 215
wearablegames.ics.uci.edu: 3
wics.ics.uci.edu: 352
wiki.ics.uci.edu: 87
www-db.ics.uci.edu: 38
xtune.ics.uci.edu: 7
yarra.ics.uci.edu: 1