# CS121: Assignment 2: Web Crawler

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

University of California, Irvine — October 31, 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 12983 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: 26512
2018: 16117
student: 15673
software: 15294
2021: 15252
information: 14962
computer: 14927
data: 14761
2020: 13683
2022: 13562
2019: 13527
2023: 13338
will: 13004
2016: 12999
students: 12947
2017: 12869
uci: 12546
may: 12355
2015: 11849
ramesh: 11423
values: 11120
engineering: 11030
10: 10806
news: 10614
informatics: 10550
insert: 10381
graduate: 10347
undergraduate: 10052
ics: 9982
september: 9960
us: 9701
june: 9687
october: 9651
july: 9614
projects: 9392
university: 9391
ph: 9184
design: 9098
march: 8971
people: 8939
november: 8830
april: 8824
december: 8823
read: 8811
science: 8781
courses: 8760
department: 8713
search: 8641
events: 8624
january: 8621

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: 109
aiclub.ics.uci.edu: 3
archive.ics.uci.edu: 121
asterix.ics.uci.edu: 14
asterixdb.ics.uci.edu: 2
auge.ics.uci.edu: 2
awareness.ics.uci.edu: 1
cbcl.ics.uci.edu: 58
cert.ics.uci.edu: 2
cgvw.ics.uci.edu: 1
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: 170
code.ics.uci.edu: 15
codeexchange.ics.uci.edu: 2
computableplant.ics.uci.edu: 89
containers.ics.uci.edu: 1
coronavirustwittermap.ics.uci.edu: 2
courselisting.ics.uci.edu: 4
cradl.ics.uci.edu: 12
create.ics.uci.edu: 5
cs.ics.uci.edu: 15
cs260p-hub.ics.uci.edu: 2
cs260p-staging-hub.ics.uci.edu: 1
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: 123
dynamo.ics.uci.edu: 21
eli.ics.uci.edu: 5
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: 44
fr.ics.uci.edu: 14
frost.ics.uci.edu: 1
futurehealth.ics.uci.edu: 122
gitlab.ics.uci.edu: 667
grafana-infra-blue.ics.uci.edu: 1
grape.ics.uci.edu: 467
graphics.ics.uci.edu: 12
graphmod.ics.uci.edu: 16
hack.ics.uci.edu: 2
hai.ics.uci.edu: 7
hana.ics.uci.edu: 4
helpdesk.ics.uci.edu: 2
hobbes.ics.uci.edu: 12
hombao.ics.uci.edu: 2
honors.ics.uci.edu: 1
hpi.ics.uci.edu: 6
http:www.ics.uci.edu: 1
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: 1
ics53-staging-hub.ics.uci.edu: 2
ieee.ics.uci.edu: 6
industryshowcase.ics.uci.edu: 23
informatics.ics.uci.edu: 1
insite.ics.uci.edu: 9
intranet.ics.uci.edu: 3
ipubmed.ics.uci.edu: 2
isg.ics.uci.edu: 219
jgarcia.ics.uci.edu: 3
jujube.ics.uci.edu: 2
julia-hub.ics.uci.edu: 2
kdd.ics.uci.edu: 2
kpassword.ics.uci.edu: 1
luci.ics.uci.edu: 6
mailman.ics.uci.edu: 9
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: 18
motifmap-rna.ics.uci.edu: 3
motifmap.ics.uci.edu: 3
mover.ics.uci.edu: 25
mse.ics.uci.edu: 1
mswe.ics.uci.edu: 12
nalini.ics.uci.edu: 8
ngs.ics.uci.edu: 1779
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
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: 206
se.ics.uci.edu: 1
seal.ics.uci.edu: 41
seraja.ics.uci.edu: 2
sherlock.ics.uci.edu: 9
sli.ics.uci.edu: 388
sourcerer.ics.uci.edu: 1
speedtest.ics.uci.edu: 1
sprout.ics.uci.edu: 2
staging-hub.ics.uci.edu: 1
stairs.ics.uci.edu: 5
statconsulting.ics.uci.edu: 5
statistics-stage.ics.uci.edu: 14
student-council.ics.uci.edu: 16
summeracademy.ics.uci.edu: 2
swiki.ics.uci.edu: 82
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: 996
wiki.ics.uci.edu: 32
www-db.ics.uci.edu: 38
xtune.ics.uci.edu: 7
yarra.ics.uci.edu: 1