Hitoki Kidahasi, Jaejin Kim
37022201, 47236384
hkidahas, jaejink

CS 121 - Information Retrival
Assignment 2 Report

For this assignment, we have implemented two was of keeping track of the urls we discovered through. We first export the current downloaded frontier urls into a file called pagOfURLS regardless of the content of the url, this way we are able to see the total number of *actual* unique pages within each domains.

The othe way was to implement a filter when crawling. When we download each page, we check if it is in the range [200,299], and if it is not then we discard the entirety of the page. We then check to see if the URL itself is a valid url by using the is_valid function in scraper.py. We have added quite a few other conditions to the function. For example, if the url does not contain any ascii characters, then we will not return it back to the frontier. Additionally, we have disregarded pdf, apk, zip, ppt, pptx, rar, and other more file extensions because that is not what we are trying to find here. Further, we have restricted a page to counted as a valid page only if the minimum word count is above 50 words.

Additionally, we've used git throughout the project, Here is the repo: https://github.com/hkprogrammer/spacetime-crawler4py/tree/deployment and we have also implemented an exact web similarity check from scratch, which can be found in the scraper.py.

1. **Unique urls:**

   Using the first approach where we count every single linkable URLS, we got **21418** unique urls.

   Using the second approach by filter, we counted **6864** valid pages.

2. **Longest page:**

   Longest page we found was https://ngs.ics.uci.edu/list-of-publications, which contained **979** unique words (not including stopwords).

3. **50 most common words:**

   The top 50 most common words are listed on page 2.

4. **Subdomains:**

   The count for subdomains can be found on page 4.

Hitoki Kidahasi, Jaejin Kim
37022201, 47236384
hkidahas, jaejink

Words Frequency:

Words -> frequency
news -> 3196
courses -> 3019
events -> 2746
research -> 2676
navigation -> 2505
students -> 2335
policy -> 2261
california -> 2232
irvine -> 2229
contact -> 2222
management -> 2179
projects -> 2104
papers -> 2002
people -> 1894
search -> 1846
2022 -> 1801
computing -> 1744
partners -> 1724
2018 -> 1700
uci -> 1675
education -> 1665
content -> 1652
2019 -> 1628
blogs -> 1619
2020 -> 1592
centers -> 1544
2021 -> 1530
home -> 1520
books -> 1463
life -> 1439
admissions -> 1425
collaborators -> 1424
companies -> 1387
services -> 1386
design -> 1382
teaching -> 1375
degrees -> 1365
science -> 1354
social -> 1332
area -> 1319
presentations -> 1317
jain -> 1312

vision -> 1309
affiliations -> 1308
2017 -> 1307
researcher -> 1295
2015 -> 1291
feedback -> 1289
2016 -> 1289
entrepreneur -> 1287

Subdomains:
subdomain -> count
http://www.ics.uci.edu -> 11737
https://ngs.ics.uci.edu -> 1617
http://archive.ics.uci.edu -> 1376
https://www.informatics.uci.edu -> 1062
https://www.ics.uci.edu -> 821
https://www.physics.uci.edu -> 775
https://www.cs.uci.edu -> 542
https://wics.ics.uci.edu -> 385
https://grape.ics.uci.edu -> 368
https://www.stat.uci.edu -> 213
http://vision.ics.uci.edu -> 204
https://sli.ics.uci.edu -> 188
http://sdcl.ics.uci.edu -> 186
https://cml.ics.uci.edu -> 170
https://isg.ics.uci.edu -> 130
http://sli.ics.uci.edu -> 118
http://ics.uci.edu -> 117
http://futurehealth.ics.uci.edu -> 102
http://www.physics.uci.edu -> 100
https://duttgroup.ics.uci.edu -> 78
https://mcs.ics.uci.edu -> 74
http://computableplant.ics.uci.edu -> 74
https://acoi.ics.uci.edu -> 71
http://physics.uci.edu -> 67
http://www.economics.uci.edu -> 65
https://transformativeplay.ics.uci.edu -> 50
https://swiki.ics.uci.edu -> 46
http://tutors.ics.uci.edu -> 39
http://www.informatics.uci.edu -> 35
http://seal.ics.uci.edu -> 34
http://linguistics.uci.edu -> 32
https://emj.ics.uci.edu -> 31
https://dgillen.ics.uci.edu -> 25
https://student-council.ics.uci.edu -> 24
http://www.stat.uci.edu -> 22
http://www-db.ics.uci.edu -> 22
https://jdsylab.physics.uci.edu -> 20
https://mds.ics.uci.edu -> 19
https://industryshowcase.ics.uci.edu -> 18
https://iasl.ics.uci.edu -> 17
http://plrg.ics.uci.edu -> 16
https://cyberclub.ics.uci.edu -> 15

https://mhcid.ics.uci.edu -> 13
https://code.ics.uci.edu -> 12
https://mswe.ics.uci.edu -> 11
https://cbcl.ics.uci.edu -> 11
https://unite.ics.uci.edu -> 10
http://www.cs.uci.edu -> 9
http://flamingo.ics.uci.edu -> 9
https://cwicsocal18.ics.uci.edu -> 9
http://cert.ics.uci.edu -> 9
http://hobbes.ics.uci.edu -> 9
http://cbcl.ics.uci.edu -> 9
https://chenli.ics.uci.edu -> 8
https://create.ics.uci.edu -> 7
https://cradl.ics.uci.edu -> 7
http://graphics.ics.uci.edu -> 7
http://scale.ics.uci.edu -> 7
https://edgelab.ics.uci.edu -> 7
http://sherlock.ics.uci.edu -> 7
http://radicle.ics.uci.edu -> 6
https://mdogucu.ics.uci.edu -> 6
http://xtune.ics.uci.edu -> 6
http://wics.ics.uci.edu -> 6
https://nalini.ics.uci.edu -> 6
http://hai.ics.uci.edu -> 5
https://flamingo.ics.uci.edu -> 5
http://i-sensorium.ics.uci.edu -> 5
http://testlab.ics.uci.edu -> 5
http://esl.ics.uci.edu -> 5
https://statconsulting.ics.uci.edu -> 4
https://asterix.ics.uci.edu -> 4
http://stairs.ics.uci.edu -> 4
https://accessibility.ics.uci.edu -> 4
https://redmiles.ics.uci.edu -> 4
http://economics.uci.edu -> 4
http://mhcid.ics.uci.edu -> 3
http://asterix.ics.uci.edu -> 3
https://support.ics.uci.edu -> 3
https://tad.ics.uci.edu -> 3
http://fr.ics.uci.edu -> 3
https://hpi.ics.uci.edu -> 3
https://luci.ics.uci.edu -> 3
http://mondego.ics.uci.edu -> 3
https://evoke.ics.uci.edu -> 3
https://intranet.ics.uci.edu -> 2

http://emj.ics.uci.edu -> 2
http://mcs.ics.uci.edu -> 2
https://mt-live.ics.uci.edu -> 2
http://cradl.ics.uci.edu -> 2
http://riscit.ics.uci.edu -> 2
https://ipf.ics.uci.edu -> 2
https://wiki.ics.uci.edu -> 2
https://mse.ics.uci.edu -> 1
http://cml.ics.uci.edu -> 1
http://luci.ics.uci.edu -> 1
https://ugradforms.ics.uci.edu -> 1
https://helpdesk.ics.uci.edu -> 1
https://tippersweb.ics.uci.edu -> 1
https://www.statistics.uci.edu -> 1
https://hub.ics.uci.edu -> 1
https://jgarcia.ics.uci.edu -> 1
https://malek.ics.uci.edu -> 1
http://informatics.ics.uci.edu -> 1
https://onboarding.ics.uci.edu -> 1
http://evoke.ics.uci.edu -> 1
http://isg.ics.uci.edu -> 1
https://hack.ics.uci.edu -> 1
https://aiclub.ics.uci.edu -> 1
http://code.ics.uci.edu -> 1
http://cwicsocal18.ics.uci.edu -> 1
http://www.informatics.ics.uci.edu -> 1
http://elms.ics.uci.edu -> 1
https://informatics.mt-live.ics.uci.edu -> 1
http://sourcerer.ics.uci.edu -> 1
https://ics.uci.edu -> 1
http://ipubmed.ics.uci.edu -> 1
http://psearch.ics.uci.edu -> 1
http://tastier.ics.uci.edu -> 1
http://perennialpolycultures.ics.uci.edu -> 1
https://www.economics.uci.edu -> 1
https://wearablegames.ics.uci.edu -> 1
https://archive.ics.uci.edu -> 1