Report

Group members: Chunji Xu_34332899
Shiyang Zhang_9584426
Jingrui Zhou_21606446

We designed crawler code to politely retrieve information from specified websites. In the code, functions such as URL normalization and deduplication, analyzing page content, counting the number of subdomains, calculating word frequency, and checking URL validity have been completed. It can detect anomalies and identify duplicates, reducing access to irrelevant resources and optimizing the performance of crawlers. We have successfully run the crawler and based on the results, the answers to the questions raised are as follows:

1.The number of unique pages we found is 10034.
2.The longest page in terms of number of words:
http://www.ics.uci.edu/~shantas/tutorials/20-icde-crypto_encryption_secret-sharing_sgx_tutorial.ppsx with 5848807 words.
3.The 50 most common words in the entire set of pages, sort by quantity in descending order, from left to right and from top to bottom (Individual letters have been filtered out):

Computer: 15325	The: 14156	Information: 12726	Software: 12334	ICS: 12172
We: 12030	Learning: 11488	University: 10270	students: 9755	us: 9753
UCI: 9583	10: 9581	events: 8313	data: 8044	research: 7873
ppt: 7722	UU: 7540	00: 7278	11: 7244	Data: 7229
UT: 7071	12: 6914	Systems: 6850	UP: 6779	Science: 6653
community: 6626	California: 6467	Design: 6467	Bren: 6205	In: 6173
computing: 6022	time: 5968	X_: 5902	UC: 5796	Embedded: 5796
School: 5788	Donald: 5647	Read: 5638	new: 5628	This: 5581
one: 5261	2018: 5225	graduate: 5174	2024: 5172	people: 5089
learning: 5063	CA: 5041	ByRamesh: 5024	Informatics: 4965	Engineering: 4922

4.Subdomain counts in the ics.uci.edu domain: 111
The list of subdomains ordered alphabetically and the number of unique pages detected in each subdomain (sort from top to bottom in each column):

https://accessibility.ics.uci.edu: 2	https://insite.ics.uci.edu: 1

https://acoi.ics.uci.edu: 78	https://intranet.ics.uci.edu: 4
https://aiclub.ics.uci.edu: 1	https://ipubmed.ics.uci.edu: 1
https://archive.ics.uci.edu: 5	https://isg.ics.uci.edu: 222
https://asterix.ics.uci.edu: 6	https://jgarcia.ics.uci.edu: 1
https://cbcl.ics.uci.edu: 3	https://julia-hub.ics.uci.edu: 1
https://cert.ics.uci.edu: 1	https://luci.ics.uci.edu: 4
https://chenli.ics.uci.edu: 3	https://mailman.ics.uci.edu: 4
https://cloudberry.ics.uci.edu: 29	https://malek.ics.uci.edu: 1
https://cml.ics.uci.edu: 61	https://mcs.ics.uci.edu: 10
https://code.ics.uci.edu: 14	https://mdogucu.ics.uci.edu: 1
https://computableplant.ics.uci.edu: 32	https://mds.ics.uci.edu: 13
https://courselisting.ics.uci.edu: 3	https://mhcid.ics.uci.edu: 17
https://cradl.ics.uci.edu: 10	https://mondego.ics.uci.edu: 3
https://create.ics.uci.edu: 5	https://mswe.ics.uci.edu: 10
https://cs.ics.uci.edu: 12	https://nalini.ics.uci.edu: 7
https://cs260p-hub.ics.uci.edu: 1	https://ngs.ics.uci.edu: 2022
https://cs260p-staging-hub.ics.uci.edu: 1	https://oai.ics.uci.edu: 5
https://cwicsocal18.ics.uci.edu: 12	https://physics.uci.edu: 4
https://cyberclub.ics.uci.edu: 1	https://plrg.ics.uci.edu: 84
https://dejavu.ics.uci.edu: 1	https://psearch.ics.uci.edu: 1
https://dgillen.ics.uci.edu: 10	https://redmiles.ics.uci.edu: 1
https://ds4all.ics.uci.edu: 3	https://riscit.ics.uci.edu: 1
https://duttgroup.ics.uci.edu: 46	https://sdcl.ics.uci.edu: 82
https://dynamo.ics.uci.edu: 1	https://seal.ics.uci.edu: 2
https://edgelab.ics.uci.edu: 7	https://sherlock.ics.uci.edu: 1

https://eli.ics.uci.edu: 4	https://sli.ics.uci.edu: 339
https://emj.ics.uci.edu: 35	https://sourcerer.ics.uci.edu: 1
https://esl.ics.uci.edu: 1	https://staging-hub.ics.uci.edu: 1
https://evoke.ics.uci.edu: 3	https://stairs.ics.uci.edu: 1
https://flamingo.ics.uci.edu: 10	https://statconsulting.ics.uci.edu: 5
https://fr.ics.uci.edu: 3	https://student-council.ics.uci.edu: 1
https://frost.ics.uci.edu: 1	https://summeracademy.ics.uci.edu: 1
https://futurehealth.ics.uci.edu: 4	https://swiki.ics.uci.edu: 3
https://grape.ics.uci.edu: 74	https://tad.ics.uci.edu: 1
https://graphics.ics.uci.edu: 2	https://tastier.ics.uci.edu: 1
https://graphmod.ics.uci.edu: 1	https://transformativeplay.ics.uci.edu: 53
https://hack.ics.uci.edu: 1	https://tutoring.ics.uci.edu: 5
https://hai.ics.uci.edu: 2	https://tutors.ics.uci.edu: 1
https://helpdesk.ics.uci.edu: 1	https://ugradforms.ics.uci.edu: 1
https://hobbes.ics.uci.edu: 1	https://unite.ics.uci.edu: 10
https://hpi.ics.uci.edu: 5	https://vision.ics.uci.edu: 7
https://hub.ics.uci.edu: 3	https://wearablegames.ics.uci.edu: 2
https://i-sensorium.ics.uci.edu: 1	https://wics.ics.uci.edu: 1071
https://iasl.ics.uci.edu: 1	https://wiki.ics.uci.edu: 7
https://icde2023.ics.uci.edu: 46	https://www-db.ics.uci.edu: 11
https://ics.uci.edu: 1259	https://www.cert.ics.uci.edu: 1
https://ics45c-hub.ics.uci.edu: 1	https://www.economics.uci.edu: 9
https://ics45c-staging-hub.ics.uci.edu: 1	https://www.graphics.ics.uci.edu: 1
https://ics46-hub.ics.uci.edu: 1	https://www.ics.uci.edu: 874
https://ics46-staging-hub.ics.uci.edu: 1	https://www.informatics.ics.uci.edu: 1

https://ics53-hub.ics.uci.edu: 1	https://www.informatics.uci.edu: 1024
https://ics53-staging-hub.ics.uci.edu: 1	https://www.physics.uci.edu: 84
https://industryshowcase.ics.uci.edu: 21	https://www.statistics.uci.edu: 1
https://informatics.ics.uci.edu: 1	https://xtune.ics.uci.edu: 6
https://informatics.uci.edu: 1	