IR Assignment 2: Web Crawler Report

Team 9
Pooja Bhatia, 14567073
Rutuja Pansare, 55443146
Ella Dodor, 31129450
Samyak Jhaveri, 13043185

Question 1. How many unique pages did you find?
Answer 1. We found **8039** unique links.

Question 2. What is the longest page in terms of the number of words?
Answer 2. The longest page among all the urls crawled:
https://evoke.ics.uci.edu/qs-personal-data-landscapes-poster
Longest Page length: 143808

Question 3. What are the 50 most common words in the entire set of pages crawled under these domains?
Answer 3.

research --> 56275
2021 --> 35146
computer --> 30826
science --> 28788
2020 --> 25292
says --> 24766
informatics --> 23520
student --> 22912
reply --> 22566
graduate --> 22514

```
students --> 22318
information --> 21540
undergraduate --> 20196
software --> 20166
data --> 19442
news --> 18638
learning --> 17930
2018 --> 15526
events --> 15029
2019 --> 14860
ramesh --> 14108
projects --> 13672
2016 --> 13629
august --> 13586
alumni --> 13540
2017 --> 13531
june --> 13126
faculty --> 12892
school --> 12721
september --> 12624
view --> 12424
engineering --> 12393
july --> 12286
october --> 11940
2015 --> 11837
people --> 11820
bren --> 11171
machine --> 10736
spotlights --> 10669
computing --> 10473
systems --> 10466
design --> 10279
```

april --> 10227
read --> 10216
irvine --> 9935
university --> 9790
march --> 9741
statistics --> 9698
courses --> 9565
technology --> 9386

Question 4. How many subdomains did you find in the ics.uci.edu domain?
Answer 4.

There are a total of **79** subdomains in the ics.uci.edu domain.
List of subdomains ordered alphabetically with count of unique pages:

http://accessibility.ics.uci.edu, 1
http://acoi.ics.uci.edu, 56
http://aiclub.ics.uci.edu, 1
http://archive.ics.uci.edu, 3
http://asterix.ics.uci.edu, 5
http://cbcl.ics.uci.edu, 3
http://cert.ics.uci.edu, 3
http://chenli.ics.uci.edu, 1
http://cloudberry.ics.uci.edu, 43
http://cml.ics.uci.edu, 172
http://code.ics.uci.edu, 12
http://computableplant.ics.uci.edu, 23
http://cradl.ics.uci.edu, 17
http://create.ics.uci.edu, 6
http://cwicsocal18.ics.uci.edu, 7

```
http://cyberclub.ics.uci.edu, 1
http://dejavu.ics.uci.edu, 1
http://dgillen.ics.uci.edu, 18
http://duttgroup.ics.uci.edu, 79
http://dynamo.ics.uci.edu, 1
http://emj.ics.uci.edu, 41
http://esl.ics.uci.edu, 1
http://evoke.ics.uci.edu, 76
http://flamingo.ics.uci.edu, 6
http://fr.ics.uci.edu, 3
http://frost.ics.uci.edu, 1
http://futurehealth.ics.uci.edu, 4
http://grape.ics.uci.edu, 12
http://graphics.ics.uci.edu, 4
http://graphmod.ics.uci.edu, 1
http://hai.ics.uci.edu, 3
http://i-sensorium.ics.uci.edu, 1
http://iasl.ics.uci.edu, 6
http://ics.uci.edu, 991
http://industryshowcase.ics.uci.edu, 20
http://informatics.ics.uci.edu, 2
http://intranet.ics.uci.edu, 1
http://ipf.ics.uci.edu, 1
http://ipubmed.ics.uci.edu, 1
http://isg.ics.uci.edu, 99
http://jgarcia.ics.uci.edu, 18
http://luci.ics.uci.edu, 3
http://malek.ics.uci.edu, 1
http://mcs.ics.uci.edu, 28
http://mdogucu.ics.uci.edu, 1
http://mds.ics.uci.edu, 10
http://mhcid.ics.uci.edu, 12
```

```
http://mondego.ics.uci.edu, 3
http://mse.ics.uci.edu, 2
http://mswe.ics.uci.edu, 16
http://mt-live.ics.uci.edu, 1559
http://nalini.ics.uci.edu, 4
http://ngs.ics.uci.edu, 2025
http://perennialpolycultures.ics.uci.edu, 1
http://plrg.ics.uci.edu, 14
http://psearch.ics.uci.edu, 1
http://redmiles.ics.uci.edu, 4
http://riscit.ics.uci.edu, 1
http://scale.ics.uci.edu, 1
http://sconce.ics.uci.edu, 2
http://sdcl.ics.uci.edu, 203
http://seal.ics.uci.edu, 6
http://sherlock.ics.uci.edu, 1
http://sli.ics.uci.edu, 225
http://sourcerer.ics.uci.edu, 1
http://sprout.ics.uci.edu, 1
http://stairs.ics.uci.edu, 1
http://statconsulting.ics.uci.edu, 5
http://student-council.ics.uci.edu, 1
http://studentcouncil.ics.uci.edu, 3
http://tad.ics.uci.edu, 1
http://tastier.ics.uci.edu, 1
http://transformativeplay.ics.uci.edu, 50
http://unite.ics.uci.edu, 9
http://vision.ics.uci.edu, 6
http://wearablegames.ics.uci.edu, 9
http://wics.ics.uci.edu, 461
http://www-db.ics.uci.edu, 8
http://xtune.ics.uci.edu, 2
```