Assignment 2: Web Crawler

By Skylar Hoffman(21464998), Keno Ha(85953583), Jordan Williams (53041731), Jingwen Li(43521573)

1.) 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. So, for example, http://www.ics.uci.edu#aaa and http://www.ics.uci.edu#bbb are the same URL.
Even if you implement additional methods for textual similarity detection, please keep considering the above definition of unique pages for the purposes of counting the unique pages in this assignment.

6364 unique pages were found using our implementation of page similarity detection.

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

The longest page we found was http://www.ics.uci.edu/~kay/wordlist.txt, with a total word count of 383392.

3.) What are the 50 most common words in the entire set of pages crawled under these domains? Submit the list of common words ordered by frequency.

reply , 105557
you , 60396
this , 45670
that , 40132
your , 39454
pm , 38976
january , 28382
with , 26856
thanks , 24673
december , 21964
post , 18737
data , 18731
will , 18349
z , 16906
article , 16125
october , 15694
software , 15544
online , 15399
july , 15280
computer , 15136
november , 15038

september , 13973
sharing , 13192
great , 12794
blog , 12446
good , 12361
student , 11435
nice , 11303
august , 11298
students , 10768
people , 10298
june , 9806
march , 9681
ramesh , 9597
news , 9412
they , 9269
there , 9150
support , 9070
thank , 8845
science , 8668
was , 8573
personal , 8395
work , 8381
engineering , 8208
design , 8061
site , 7963
uci , 7957
events , 7743
best , 7677
informatics , 7617

4.) How many subdomains did you find in the ics.uci.edu domain? Submit the list of subdomains ordered alphabetically and the number of unique pages detected in each subdomain.
The content of this list should be lines containing URL, number, for example:
http://vision.ics.uci.edu, 10 (not the actual number here)

https://aiclub.ics.uci.edu , 1
http://archive.ics.uci.edu , 4
https://archive.ics.uci.edu , 1
https://asterix-gerrit.ics.uci.edu , 1
http://asterix.ics.uci.edu , 6
http://calendar.ics.uci.edu , 1
https://cbcl.ics.uci.edu , 4
http://cert.ics.uci.edu , 2
http://www.cert.ics.uci.edu , 1

http://checkmate.ics.uci.edu , 1
https://chenli.ics.uci.edu , 1
http://cloudberry.ics.uci.edu , 36
http://cml.ics.uci.edu , 5
https://cml.ics.uci.edu , 135
http://computableplant.ics.uci.edu , 22
http://cradl.ics.uci.edu , 11
https://cradl.ics.uci.edu , 8
http://cwicsocal18.ics.uci.edu , 1
https://cwicsocal18.ics.uci.edu , 9
https://cyberclub.ics.uci.edu , 1
http://www-db.ics.uci.edu , 11
http://dynamo.ics.uci.edu , 1
http://elms.ics.uci.edu , 1
https://emj.ics.uci.edu , 35
http://emj.ics.uci.edu , 8
http://esl.ics.uci.edu , 1
https://evoke.ics.uci.edu , 699
http://flamingo.ics.uci.edu , 4
http://fr.ics.uci.edu , 3
http://futurehealth.ics.uci.edu , 1
https://futurehealth.ics.uci.edu , 8
https://grape.ics.uci.edu , 13
http://graphics.ics.uci.edu , 2
http://www.graphics.ics.uci.edu , 1
http://graphmod.ics.uci.edu , 1
http://hai.ics.uci.edu , 1
http://hana.ics.uci.edu , 19
https://helpdesk.ics.uci.edu , 1
https://hombao.ics.uci.edu , 1
http://honors.ics.uci.edu , 1
https://honors.ics.uci.edu , 18
http://i-sensorium.ics.uci.edu , 1
http://www.ics.uci.edu , 670
https://www.ics.uci.edu , 309
http://informatics.ics.uci.edu , 1
http://www.informatics.ics.uci.edu , 1
https://intranet.ics.uci.edu , 1
http://ipubmed.ics.uci.edu , 1
http://isg.ics.uci.edu , 1
https://isg.ics.uci.edu , 113
https://jgarcia.ics.uci.edu , 12
http://keys.ics.uci.edu , 1
https://mailman.ics.uci.edu , 1

http://malek.ics.uci.edu , 1
https://mcs.ics.uci.edu , 40
https://mdogucu.ics.uci.edu , 1
http://mhcid.ics.uci.edu , 2
https://mhcid.ics.uci.edu , 13
https://mse.ics.uci.edu , 1
https://mswe.ics.uci.edu , 20
http://nalini.ics.uci.edu , 1
https://nalini.ics.uci.edu , 6
https://ngs.ics.uci.edu , 1534
http://ngs.ics.uci.edu , 12
http://perennialpolycultures.ics.uci.edu , 1
http://plrg.ics.uci.edu , 9
http://psearch.ics.uci.edu , 1
https://redmiles.ics.uci.edu , 7
http://riscit.ics.uci.edu , 1
http://sconce.ics.uci.edu , 2
http://sdcl.ics.uci.edu , 214
https://seal.ics.uci.edu , 1
http://seal.ics.uci.edu , 5
http://sherlock.ics.uci.edu , 1
http://sli.ics.uci.edu , 107
https://sli.ics.uci.edu , 217
http://sprout.ics.uci.edu , 1
https://statconsulting.ics.uci.edu , 5
https://student-council.ics.uci.edu , 1
https://studentcouncil.ics.uci.edu , 1
https://support.ics.uci.edu , 2
http://tastier.ics.uci.edu , 1
http://tippersweb.ics.uci.edu , 1
https://transformativeplay.ics.uci.edu , 38
https://Transformativeplay.ics.uci.edu , 1
http://tutors.ics.uci.edu , 1
https://ugradforms.ics.uci.edu , 1
http://vision.ics.uci.edu , 7
https://wearablegames.ics.uci.edu , 10
http://wics.ics.uci.edu , 11
https://wics.ics.uci.edu , 903
http://xtune.ics.uci.edu , 6