Assignment 2: Crawler Report

By: Lanni Dang-Vu, Mirelle George, Estefany Siguantay-Cortez, Cole Pickering

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.

   Unique urls: 46302

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

The longest page was 'http://www.ics.uci.edu/~kay/wordlist.txt' with 385018 words.

3. 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, here) Submit the list of common words ordered by frequency.

sorted words: [('data', 71409), ('research', 30381), ('contact', 29038), ('learning', 26120), ('uci', 25382), ('computer', 23335), ('policy', 23142), ('machine', 23046), ('sciences', 22266), ('set', 22187), ('new', 21782), ('undo', 20920), ('classification', 20275), ('view', 19959), ('real', 19622), ('information', 19027), ('sets', 18513), ('multivariate', 18511), ('type', 17853), ('students', 17356), ('will', 16945), ('news', 16389), ('dataset', 15808), ('can', 15168), ('student', 14339), ('engineering', 14259), ('school', 13838), ('events', 13166), ('citation', 13084), ('us', 12985), ('bren', 12926), ('repository', 12853), ('software', 12802), ('science', 12485), ('less', 11521), ('systems', 11474), ('ics', 11370), ('graduate', 11260), ('web', 10358), ('may', 10198), ('one', 10082), ('task', 9895), ('attribute', 9832), ('using', 9601), ('life', 9518), ('undergraduate', 9485), ('attributes', 9451), ('informatics', 9344), ('center', 9197), ('ramesh', 9180)]

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)

There are 79 subdomains in the ics.uci.edu domain.

ics_domains: {'http://acoi.ics.uci.edu': 70, 'http://aiclub.ics.uci.edu': 1, 'http://archive.ics.uci.edu': 6115, 'http://asterix.ics.uci.edu': 5, 'http://cbcl.ics.uci.edu': 74, 'http://cdb.ics.uci.edu': 22, 'http://cert.ics.uci.edu': 8, 'http://chenli.ics.uci.edu': 8, 'http://circadiomics.ics.uci.edu': 5, 'http://cml.ics.uci.edu': 128, 'http://code.ics.uci.edu': 13, 'http://computableplant.ics.uci.edu': 20,

```
'http://cradl.ics.uci.edu': 16, 'http://create.ics.uci.edu': 6, 'http://cwicsocal18.ics.uci.edu': 6,
'http://cyberclub.ics.uci.edu': 11, 'http://datalab.ics.uci.edu': 1, 'http://dgillen.ics.uci.edu': 10,
'http://duttgroup.ics.uci.edu': 58, 'http://emj.ics.uci.edu': 22, 'http://esl.ics.uci.edu': 4,
'http://evoke.ics.uci.edu': 2, 'http://flamingo.ics.uci.edu': 9, 'http://fr.ics.uci.edu': 2,
'http://futurehealth.ics.uci.edu': 85, 'http://grape.ics.uci.edu': 489, 'http://graphics.ics.uci.edu': 4,
'http://hai.ics.uci.edu': 4, 'http://hobbes.ics.uci.edu': 6, 'http://hpi.ics.uci.edu': 3,
'http://i-sensorium.ics.uci.edu': 2, 'http://iasl.ics.uci.edu': 13, 'http://ieee.ics.uci.edu': 1,
'http://industryshowcase.ics.uci.edu': 17, 'http://ipf.ics.uci.edu': 1, 'http://ipubmed.ics.uci.edu': 1,
'http://isg.ics.uci.edu': 140, 'http://jgarcia.ics.uci.edu': 6, 'http://luci.ics.uci.edu': 3,
'http://malek.ics.uci.edu': 1, 'http://mcs.ics.uci.edu': 70, 'http://mdogucu.ics.uci.edu': 5,
'http://mds.ics.uci.edu': 17, 'http://mhcid.ics.uci.edu': 12, 'http://mlphysics.ics.uci.edu': 1,
'http://mondego.ics.uci.edu': 2, 'http://motifmap-rna.ics.uci.edu': 2, 'http://motifmap.ics.uci.edu': 2,
'http://mse.ics.uci.edu': 2, 'http://mswe.ics.uci.edu': 8, 'http://nalini.ics.uci.edu': 2,
'http://ngs.ics.uci.edu': 1440, 'http://perennialpolycultures.ics.uci.edu': 1, 'http://plrg.ics.uci.edu':
14, 'http://psearch.ics.uci.edu': 1, 'http://radicle.ics.uci.edu': 3, 'http://redmiles.ics.uci.edu': 4,
'http://riscit.ics.uci.edu': 2, 'http://scale.ics.uci.edu': 4, 'http://scratch.proteomics.ics.uci.edu': 3,
'http://sdcl.ics.uci.edu': 124, 'http://seal.ics.uci.edu': 36, 'http://selectpro.proteomics.ics.uci.edu':
5, 'http://sherlock.ics.uci.edu': 4, 'http://sli.ics.uci.edu': 191, 'http://sourcerer.ics.uci.edu': 1,
'http://stairs.ics.uci.edu': 2, 'http://statconsulting.ics.uci.edu': 3, 'http://studentcouncil.ics.uci.edu':
10, 'http://tastier.ics.uci.edu': 1, 'http://transformativeplay.ics.uci.edu': 35, 'http://unite.ics.uci.edu':
7, 'http://vision.ics.uci.edu': 174, 'http://wearablegames.ics.uci.edu': 6, 'http://wics.ics.uci.edu':
373, 'http://wiki.ics.uci.edu': 17, 'http://www-db.ics.uci.edu': 4, 'http://www.ics.uci.edu': 2713,
'http://xtune.ics.uci.edu': 2}
```