CS121 analytics report

1. Keep track of the subdomains that it visited, and count how many different URLs it has processed from each of those subdomains.

http://www.ics.uci.edu: 4371

http://hobbes.ics.uci.edu: 2

http://ipubmed.ics.uci.edu: 3

https://intranet.ics.uci.edu: 15

http://vision.ics.uci.edu: 162

http://futurehealth.ics.uci.edu: 5

http://luci.ics.uci.edu: 4

http://sconce.ics.uci.edu: 3

https://duttgroup.ics.uci.edu/doku.php/projects: 1

http://asterix.ics.uci.edu: 7

http://hombao.ics.uci.edu: 1

http://mhcid.ics.uci.edu: 5

https://mswe.ics.uci.edu: 8

http://emj.ics.uci.edu: 2

https://netreg.ics.uci.edu: 4

https://swiki.ics.uci.edu/doku.php: 1141

https://seal.ics.uci.edu: 54

http://graphics.ics.uci.edu: 19

https://www.cert.ics.uci.edu: 14

https://grape.ics.uci.edu: 8164

https://sdcl.ics.uci.edu: 4

https://cbcl.ics.uci.edu: 488

2. Find the page with the most valid out links (of all pages given to your crawler). Out Links are the number of links that are present on a particular webpage.

http://www.ics.uci.edu

3. List of downloaded URLs and identified traps.

https://drive.google.com/file/d/1I-Nvx7zT90td1G1MRoHvhrFaNqAvfq8F/view?usp=sharing

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

2255 words:

http://www.ics.uci.edu/%7Ewscacchi/Presentations/OSS-Requirements/

5. What are the 50 most common words in the entire set of pages? (Ignore English stop words,
   which can be found, (https://www.ranks.nl/stopwords)

https://docs.google.com/document/d/1ReWKic8a58XBe9HFCXXCls1aoKM-aOgW5w5WqwanQiA/edit?usp=sharing