Please download in addition with 2 original dependencies

python3 -m pip install --user lxlm
python3 -m pip install --user beautifulsoup4

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.**

   Our web crawler was able to find 8504 unique pages.

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

   The longest page in terms of number of words is: http://www.ics.uci.edu/~kay/wordlist.txt

3. **What are the 50 most common words in the entire set of pages? (Ignore English stop words, which can be found, for example, here. Submit the list of common words ordered by frequency.**

   The 50 most common words have been compiled into the list below:

   [('reply', 216146), ('says’, 208784), ('2019', 146626), ('pm', 81093), ('10', 66773), ('2020', 56586), ('january, 54316), ('information', 52204), ('thanks', 50178), ('can’, 46744), ('levorato', 45657), ('will', 41898), ('december', 41705), ('2018’, 41179), ('12', 40078), ('data', 37668), ('research', 37446), ('post’', 37107), ('11', 36785), ('article', 36735), ('online', 31656), ('october', 29633), ('computer', 29150), ('july', 28666), ('november', 28050), ('good', 27290), ('ramesh', 27154), ('sharing', 26896), ('2017’, 26875), ('great’, 26740), ('blog', 25593), ('september', 25562), ('software', 25346), ('like', 24612), ('may', 24608), ('time', 24282), ('one', 24089), ('marco’, 23833), ('nice', 23680), ('read', 23478), ('ieee’, 22396), ('29', 22295), ('conference', 21961), ('people', 21750), ('really', 21721), ('author', 21537), ('web', 21417), ('new', 20862), ('august', 20746), ('students', 20149)]

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 87 subdomains inside ics.uci.edu. The alphabetised list of subdomains have been denoted below:

aiclub.ics.uci.edu - 1
archive.ics.uci.edu - 6
asterix-gerrit.ics.uci.edu - 1
asterix.ics.uci.edu - 6
cbcl.ics.uci.edu - 4
cert.ics.uci.edu - 2
checkmate.ics.uci.edu - 1
chenli.ics.uci.edu - 1
cloudberry.ics.uci.edu - 59
cml.ics.uci.edu - 165
computableplant.ics.uci.edu - 32
cradl.ics.uci.edu - 20
cwicsocal18.ics.uci.edu - 12
cyberclub.ics.uci.edu - 2
dejavu.ics.uci.edu - 1
duttgroup.ics.uci.edu - 91
dynamo.ics.uci.edu - 1
elms.ics.uci.edu - 1
emj.ics.uci.edu - 45
esl.ics.uci.edu - 1
evoke.ics.uci.edu - 728
flamingo.ics.uci.edu - 6
fr.ics.uci.edu - 3
frost.ics.uci.edu - 1
futurehealth.ics.uci.edu - 9
grape.ics.uci.edu - 13
graphics.ics.uci.edu - 3
graphmod.ics.uci.edu - 1
hai.ics.uci.edu - 2
hana.ics.uci.edu - 19
helpdesk.ics.uci.edu - 1
hombao.ics.uci.edu - 1
honors.ics.uci.edu - 19
i-sensorium.ics.uci.edu - 1
iasl.ics.uci.edu - 17
ics.uci.edu - 6
intranet.ics.uci.edu - 1
ipf.ics.uci.edu - 2
ipubmed.ics.uci.edu - 1
isg.ics.uci.edu - 114

jgarcia.ics.uci.edu - 18
keys.ics.uci.edu - 1
linguistics.uci.edu - 1
mailman.ics.uci.edu - 4
malek.ics.uci.edu - 1
mcs.ics.uci.edu - 40
mdogucu.ics.uci.edu - 1
mhcid.ics.uci.edu - 15
mondego.ics.uci.edu - 3
mse.ics.uci.edu - 1
mswe.ics.uci.edu - 20
nalini.ics.uci.edu - 7
ngs.ics.uci.edu - 2997
perennialpolycultures.ics.uci.edu - 1
plrg.ics.uci.edu - 9
psearch.ics.uci.edu - 2
redmiles.ics.uci.edu - 7
riscit.ics.uci.edu - 1
sconce.ics.uci.edu - 2
sdcl.ics.uci.edu - 218
seal.ics.uci.edu - 6
sherlock.ics.uci.edu - 1
sli.ics.uci.edu - 379
sourcerer.ics.uci.edu - 1
sprout.ics.uci.edu - 2
statconsulting.ics.uci.edu - 5
student-council.ics.uci.edu - 1
studentcouncil.ics.uci.edu - 1
support.ics.uci.edu - 2
tastier.ics.uci.edu - 1
tippersweb.ics.uci.edu - 1
Transformativeplay.ics.uci.edu - 1
transformativeplay.ics.uci.edu - 47
tutors.ics.uci.edu - 1
ugradforms.ics.uci.edu - 1
vision.ics.uci.edu - 7
wearablegames.ics.uci.edu - 11
wics.ics.uci.edu - 1567
www-db.ics.uci.edu - 11
www.cert.ics.uci.edu - 1
www.economics.uci.edu - 14
www.graphics.ics.uci.edu - 1
www.ics.uci.edu - 1268

www.informatics.ics.uci.edu	1
xtune.ics.uci.edu	6