Group name on canvas: Assignment2 SRS

People: 15786064(Zijian Chen), 13713641(Qingshuang Su), 70518431(Lingxin Li)

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.**

   9499 unique pages

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

   http://www.ics.uci.edu/~kay/wordlist.txt

   385046 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 (Links to an external site.)) Submit the list of common words ordered by frequency.

   [('Research', 45960), ('0', 41510), ('2021', 29996), ('Computer', 25787), ('1', 25727), ('Science', 24734), ('2', 24033), ('I', 22694), ('2020', 22503), ('Informatics', 22319), ('UCI', 19624), ('Student', 19523), ('ICS', 19514), ('Graduate', 18943), ('The', 18491), ('Undergraduate', 17617), ('says', 17251), ('News', 16787), ('Software', 15649), ('2018', 15530), ('Information', 15394), ('Reply', 15007), ('2019', 14688), ('3', 14360), ('Ramesh', 14097), ('2016', 13602), ('2017', 13530), ('Learning', 13427), ('August', 13050), ('students', 12531), ('June', 12440), ('We', 12285), ('View', 11946), ('September', 11895), ('2015', 11810), ('July', 11694), ('4', 11573), ('October', 11216), ('Bren', 11071), ('5', 10889), ('Engineering', 10876), ('Data', 10767), ('B.S', 10713), ('research', 10673), ('Spotlights', 10622), ('School', 10562), ('Faculty', 10515), ('Events', 10435), ('Projects', 10298), ('us', 10280)]

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)

166 subdomains

[('http://DataGuard.ics.uci.edu', 1), ('http://DataProtector.ics.uci.edu', 1),
('http://alumni.ics.uci.edu', 5), ('http://archive.ics.uci.edu', 6),
('http://asterix.ics.uci.edu', 7), ('http://asterixdb.ics.uci.edu', 2),
('http://auge.ics.uci.edu', 2), ('http://awareness.ics.uci.edu', 1),
('http://calendar.ics.uci.edu', 1), ('http://cert.ics.uci.edu', 3),
('http://cgvw.ics.uci.edu', 1), ('http://checkmate.ics.uci.edu', 2),
('http://chenli.ics.uci.edu', 1), ('http://cloudberry.ics.uci.edu', 9),
('http://cml.ics.uci.edu', 6), ('http://cocoa-krispies.ics.uci.edu', 1),
('http://code.ics.uci.edu', 1), ('http://codeexchange.ics.uci.edu', 2),
('http://computableplant.ics.uci.edu', 44),
('http://coronavirustwittermap.ics.uci.edu', 1), ('http://cradl.ics.uci.edu', 16),
('http://cwicsocal18.ics.uci.edu', 1), ('http://dblp.ics.uci.edu', 2),
('http://dejavu.ics.uci.edu', 1), ('http://duttgroup.ics.uci.edu', 1),
('http://dynamo.ics.uci.edu', 1), ('http://elms.ics.uci.edu', 1), ('http://emj-
pc.ics.uci.edu', 1), ('http://emj.ics.uci.edu', 1), ('http://esl.ics.uci.edu', 1),
('http://evoke.ics.uci.edu', 2), ('http://flamingo.ics.uci.edu', 8),
('http://fr.ics.uci.edu', 5), ('http://frost.ics.uci.edu', 1),
('http://futurehealth.ics.uci.edu', 4), ('http://givargis.ics.uci.edu', 1),
('http://graphics.ics.uci.edu', 5), ('http://graphmod.ics.uci.edu', 1),
('http://hai.ics.uci.edu', 2), ('http://hana.ics.uci.edu', 3),
('http://hombao.ics.uci.edu', 1), ('http://honors.ics.uci.edu', 1), ('http://i-
sensorium.ics.uci.edu', 1), ('http://ics.uci.edu', 7), ('http://informatics.ics.uci.edu',
1), ('http://intranet.ics.uci.edu', 1), ('http://ipubmed.ics.uci.edu', 1),
('http://isg.ics.uci.edu', 3), ('http://jgarcia.ics.uci.edu', 1), ('http://jujube.ics.uci.edu',
1), ('http://keys.ics.uci.edu', 1), ('http://luci.ics.uci.edu', 3),
('http://malek.ics.uci.edu', 2), ('http://mapgrid.ics.uci.edu', 1),
('http://mcs.ics.uci.edu', 4), ('http://metaviz.ics.uci.edu', 1),
('http://mhcid.ics.uci.edu', 5), ('http://mondego.ics.uci.edu', 11),
('http://mswe.ics.uci.edu', 1), ('http://nalini.ics.uci.edu', 1), ('http://ngs.ics.uci.edu',
21), ('http://nile.ics.uci.edu', 1), ('http://omni.ics.uci.edu', 1),
('http://pasteur.ics.uci.edu', 2), ('http://perennialpolycultures.ics.uci.edu', 1),
('http://plrg.ics.uci.edu', 15), ('http://psearch.ics.uci.edu', 3),
('http://riscit.ics.uci.edu', 1), ('http://sconce.ics.uci.edu', 4), ('http://sdcl.ics.uci.edu',
207), ('http://se.ics.uci.edu', 1), ('http://seal.ics.uci.edu', 5),
('http://seraja.ics.uci.edu', 1), ('http://sherlock.ics.uci.edu', 1),
('http://sli.ics.uci.edu', 295), ('http://sourcerer.ics.uci.edu', 1),
('http://sprout.ics.uci.edu', 3), ('http://stairs.ics.uci.edu', 1),
('http://tastier.ics.uci.edu', 1), ('http://tippers.ics.uci.edu', 1),
('http://tippersweb.ics.uci.edu', 2), ('http://transformativeplay.ics.uci.edu', 1),
('http://tutors.ics.uci.edu', 1), ('http://vision.ics.uci.edu', 8),
('http://wics.ics.uci.edu', 20), ('http://www-db.ics.uci.edu', 16),
('http://www.cert.ics.uci.edu', 1), ('http://www.graphics.ics.uci.edu', 1),

('http://www.ics.uci.edu', 1120), ('http://www.informatics.ics.uci.edu', 2),
('http://www.isg.ics.uci.edu', 1), ('http://xtune.ics.uci.edu', 7),
('http://yarra.ics.uci.edu', 1), ('https://Transformativeplay.ics.uci.edu', 1),
('https://accessibility.ics.uci.edu', 1), ('https://acoi.ics.uci.edu', 57),
('https://aiclub.ics.uci.edu', 1), ('https://archive-beta.ics.uci.edu', 2),
('https://archive.ics.uci.edu', 1), ('https://asterix.ics.uci.edu', 1),
('https://cbcl.ics.uci.edu', 3), ('https://chenli.ics.uci.edu', 2),
('https://cloudberry.ics.uci.edu', 43), ('https://cml.ics.uci.edu', 171),
('https://code.ics.uci.edu', 12), ('https://computer-science.mt-live.ics.uci.edu', 1),
('https://coronavirustwittermap.ics.uci.edu', 2), ('https://cradl.ics.uci.edu', 9),
('https://create.ics.uci.edu', 4), ('https://cwicsocal18.ics.uci.edu', 11),
('https://cyberclub.ics.uci.edu', 1), ('https://dgillen.ics.uci.edu', 18),
('https://duttgroup.ics.uci.edu', 88), ('https://elms.ics.uci.edu', 1),
('https://emj.ics.uci.edu', 45), ('https://evoke.ics.uci.edu', 99),
('https://futurehealth.ics.uci.edu', 2), ('https://grape.ics.uci.edu', 13),
('https://graphics.ics.uci.edu', 1), ('https://hack.ics.uci.edu', 1),
('https://hai.ics.uci.edu', 2), ('https://helpdesk.ics.uci.edu', 1),
('https://hombao.ics.uci.edu', 1), ('https://iasl.ics.uci.edu', 6),
('https://industryshowcase.ics.uci.edu', 21), ('https://informatics.mt-live.ics.uci.edu',
1), ('https://intranet.ics.uci.edu', 1), ('https://ipf.ics.uci.edu', 2),
('https://ipubmed.ics.uci.edu', 1), ('https://isg.ics.uci.edu', 130),
('https://jgarcia.ics.uci.edu', 24), ('https://luci.ics.uci.edu', 4),
('https://mailman.ics.uci.edu', 3), ('https://malek.ics.uci.edu', 1),
('https://mcs.ics.uci.edu', 32), ('https://mdogucu.ics.uci.edu', 1),
('https://mds.ics.uci.edu', 13), ('https://mhcid.ics.uci.edu', 17),
('https://mse.ics.uci.edu', 2), ('https://mswe.ics.uci.edu', 17), ('https://mt-
live.ics.uci.edu', 1504), ('https://nalini.ics.uci.edu', 8), ('https://ngs.ics.uci.edu',
2013), ('https://password.ics.uci.edu', 1), ('https://redmiles.ics.uci.edu', 4),
('https://scale.ics.uci.edu', 1), ('https://sdcl.ics.uci.edu', 2), ('https://seal.ics.uci.edu',
1), ('https://sherlock.ics.uci.edu', 1), ('https://sli.ics.uci.edu', 303),
('https://statconsulting.ics.uci.edu', 4), ('https://statistics.mt-live.ics.uci.edu', 1),
('https://student-council.ics.uci.edu', 2), ('https://studentcouncil.ics.uci.edu', 4),
('https://support.ics.uci.edu', 2), ('https://swiki.ics.uci.edu', 1),
('https://tad.ics.uci.edu', 1), ('https://tippers.ics.uci.edu', 1),
('https://tippersweb.ics.uci.edu', 4), ('https://transformativeplay.ics.uci.edu', 50),
('https://ugradforms.ics.uci.edu', 1), ('https://unite.ics.uci.edu', 10),
('https://vision.ics.uci.edu', 1), ('https://wearablegames.ics.uci.edu', 12),
('https://wics.ics.uci.edu', 488), ('https://www.ics.uci.edu', 498)]