

STEPS FOLLOWED TO COMPLETE THE ASSIGNMENT

Installation of Virtual Box and Ubuntu

- 1) Downloaded and installed VirtualBox 6.0.18 in my Windows OS.
- 2) Downloaded Image file of Ubuntu 18.04.4 LTS.
- 3) Created Virtual machine with 8 GB RAM and drive size of 128 GB.
- **4)** After starting Ubuntu in the virtual machine, added Guest additions and created a shared Folder for accessing between Windows and Ubuntu.
- **5)** Enabled the bidirectional copy-paste and drag-and-drop features in VBox and mounted the shared folder.

Installation of Java, Apache and PHP

- 1) Installed Java 11.0.6 in Ubuntu.
- 2) Installed Apache 2.4.29 and changed the default path of the server using the following steps.
 - cd /etc/apache2/sites-available
 - sudo nano 000-default.conf => Changed the path in this file
 - sudo nano apache2.conf => Added this block

<Directory /path/to/shared/folder>

Options Indexes FollowSymLinks

AllowOverride None

Require all granted

</Directory>

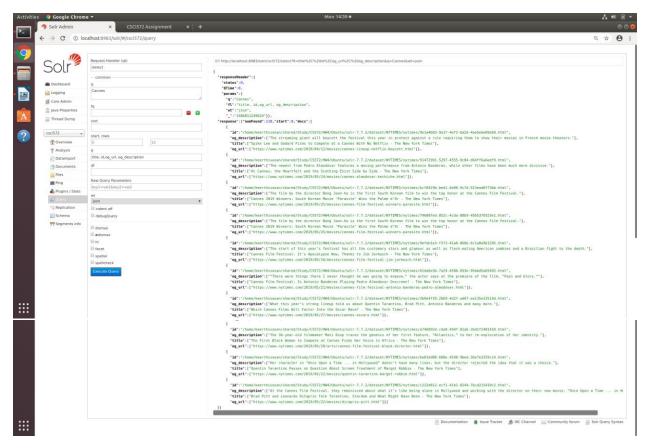
- sudo service apache2 restart
- 3) Installed PHP 7.2.24 and set display errors = on in PHP.ini file to display all the errors.

Installation of Solr

- 1) Downloaded and unpacked Solr 7.7.2 in the shared folder.
- 2) Created a core named "csci572" and modified the managed-schema file to populate field data to a single _text_ field.
- 3) Configured Solr to query default from _text_ by defining the **df** field inside requestHandler of **solrconfig.xml** file.
- **4)** Using the **TIKA POST** command, fetched and parsed the 16040 documents. Index entries for 14864 documents were created.



Querying Solr with "Cannes"



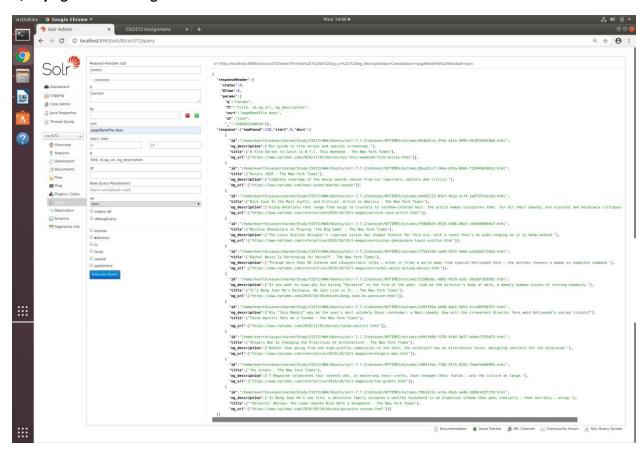
Creation of EdgeList and PageRank

- 1) Using **Jsoup** library, parsed the documents in **JAVA** to extract the Outgoing links and created an **EdgeList** text file (Containing only those links present in the mapping file).
- 2) NetworkX is a Python library to deal with graphs and complex networks. Using this library, created a Directional Graph with Document_Ids as the Nodes and the Interconnecting links as the Edges by reading from the EdgeList file.
- 3) Calculated the Page Rank of the documents using PageRank function of the DiGraph. The configuration was set as follows:
 - a. Damping factor was set to the default value 0.85.
 - b. Node personalization was not used.
 - c. Max-iterations was set to 30.
 - d. Error tolerance was set at 10^-6.
 - e. No starting value was given to the pages.
 - f. Weights correspond to the edges.
- 4) external_pageRankFile.txt file was created and placed in the data folder of the Solr Core.



- 5) Added a field entry in managed-schema to refer this PageRank score while querying. Also defined Listeners in solrconfig.xml file to reload the external file when either a searcher is loaded or when a new searcher is started.
- **6)** Reloaded the csci572 core and added *"pageRankFile desc"* in the **sort** field to enable searching via PageRank score instead of Lucene's default TF-IDF scores.

Querying Solr with "PageRankFile Desc" and "Cannes"



EXPLANATION OF WHY SOME PAGES HAVE HIGHER PAGE RANK VALUES

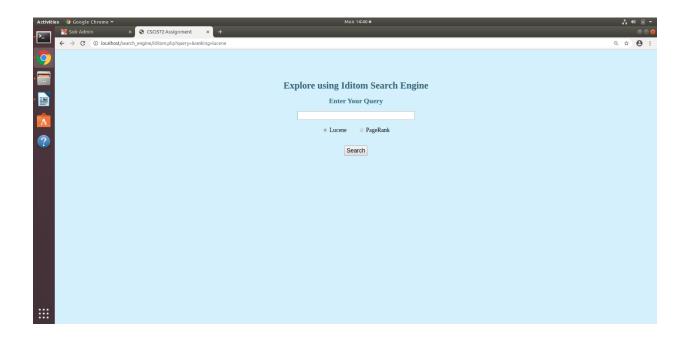
Page Rank is an algorithm which measures the importance of a webpage by tracking the number of links received from other webpages. By observing the Page rank scores of the Query results, we infer that the Homepage and the Category pages are prone to get a higher score since many pages have external links to them. Hence these Parent pages can end up being displayed in the top results even when are not very relevant to the actual query. This is one of the reasons for combining Page rank value with other scores such as TF-IDF to yield relevant results.





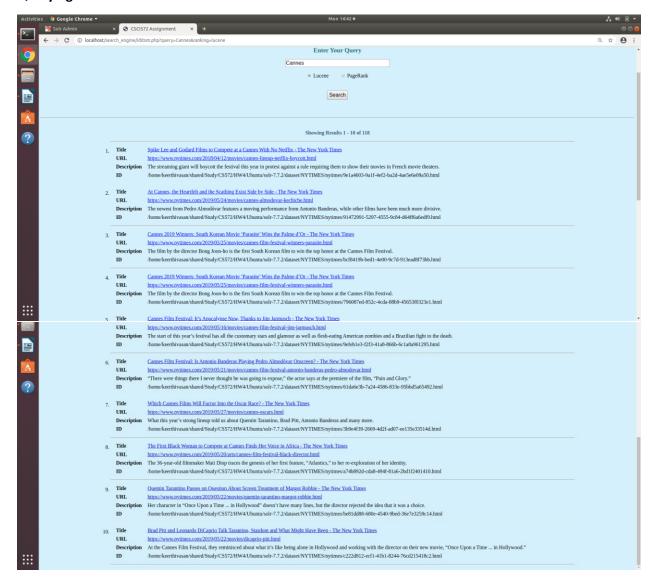
Web Client and Web Server

Search Engine Home Page



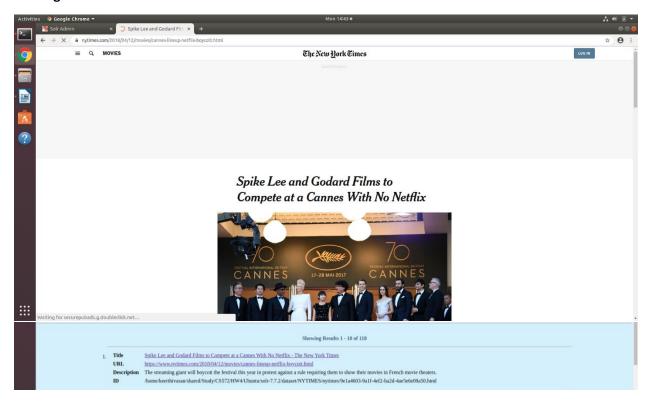


Querying Lucene



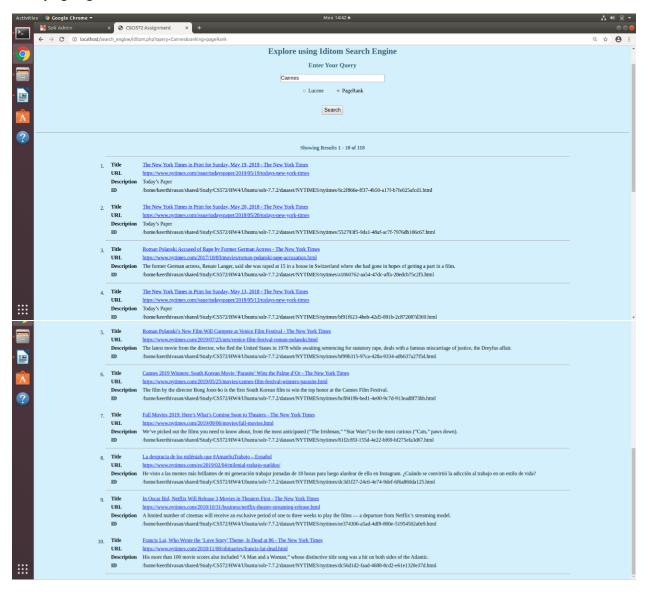


Clicking First Link





Querying PageRank





Clicking First Link



QUERY RESULTS OVERLAPS BETWEEN LUCENE AND PAGERANK

Query	Number of Overlaps
Cannes	0
Congress	0
Democrats	0
Patriot Movement	0
Republicans	0
Senate	0
Olympics 2020	0
Stock	0
Virus	0



QUERY RESULTS

1. Cannes

Lucene	PageRank
https://www.nytimes.com/2018/04/12/movies/ca	https://www.nytimes.com/2019/11/07/movies/n
nnes-lineup-netflix-boycott.html	yc-this-weekend-film-series.html
https://www.nytimes.com/2019/05/24/movies/ca	https://www.nytimes.com/news-event/awards-
nnes-almodovar-kechiche.html	season
https://www.nytimes.com/2019/05/25/movies/ca	https://www.nytimes.com/interactive/2019/10/1
nnes-film-festival-winners-parasite.html	5/t-magazine/nick-cave-artist.html
https://www.nytimes.com/2019/05/25/movies/ca nnes-film-festival-winners-parasite.html	https://www.nytimes.com/interactive/2019/10/1 5/t-magazine/nicolas-ghesquiere-louis- vuitton.html
https://www.nytimes.com/2019/05/16/movies/ca nnes-film-festival-jim-jarmusch.html	https://www.nytimes.com/interactive/2019/10/1 5/t-magazine/rachel-weisz-acting-movies.html
https://www.nytimes.com/2019/05/21/movies/ca nnes-film-festival-antonio-banderas-pedro- almodovar.html	https://www.nytimes.com/2019/10/30/movies/b ong-joon-ho-parasite.html
https://www.nytimes.com/2019/05/27/movies/ca	https://www.nytimes.com/2019/11/01/movies/ta
nnes-oscars.html	ika-waititi.html
https://www.nytimes.com/2019/05/20/arts/cann	https://www.nytimes.com/interactive/2019/10/1
es-film-festival-black-director.html	5/t-magazine/shigeru-ban.html
https://www.nytimes.com/2019/05/22/movies/q	https://www.nytimes.com/interactive/2019/10/1
uentin-tarantino-margot-robbie.html	5/t-magazine/the-greats.html
https://www.nytimes.com/2019/05/22/movies/di	https://www.nytimes.com/2019/10/10/movies/p
caprio-pitt.html	arasite-review.html



2. Congress

Lucene	PageRank
https://www.nytimes.com/interactive/2019/us/p olitics/trump-impeachment-congress-list.html	https://www.nytimes.com
https://www.nytimes.com/2019/11/05/magazine/congress-president-impeachment.html	https://www.nytimes.com/section/magazine
https://www.nytimes.com/2019/11/05/magazine/congress-president-impeachment.html	https://www.nytimes.com/section/t-magazine
https://www.nytimes.com/2019/07/25/us/politics/mueller-impeachment.html	https://www.nytimes.com/issue/todayspaper/201 9/11/10/todays-new-york-times
https://www.nytimes.com/2019/07/26/us/politics/donald-trump-impeachment.html	https://www.nytimes.com/section/jobs
https://www.nytimes.com/2018/09/17/us/politics/bloomberg-president-2020-democrat.html	https://www.nytimes.com/video/us/politics/1000 00005524412/paul-manafort-fraud-laundering- scandal.html
https://www.nytimes.com/2018/11/21/obituaries /james-billington-dead.html	https://www.nytimes.com/video/us/politics/1000 00006067682/president-trump-impeach- questions-answers-michael-cohen.html
https://www.nytimes.com/2019/09/07/us/politics/congress-assault-weapons-ban.html	https://www.nytimes.com/video/us/politics/1000 00006639814/presidential-candidates- campaigns.html
https://www.nytimes.com/2019/08/07/us/politics/foreign-aid-freeze-congress.html	https://www.nytimes.com/video/opinion/100000 006210828/russia-disinformation-fake-news.html
https://www.nytimes.com/1973/05/13/archives/c onstitutional-ervin-ah-thi-nk-the-senate-is- determined-to-recover.html	https://www.nytimes.com/video/opinion/100000 006188105/countering-disinformation-active- measures.html



3. Democrats

Lucene	PageRank
https://www.nytimes.com/2018/09/04/us/politic s/ayanna-pressley-massachusetts.html	https://www.nytimes.com
https://www.nytimes.com/2018/11/07/us/politic s/election-history-firsts-blackburn-pressley.html	https://www.nytimes.com/video
https://www.nytimes.com/2019/10/16/us/politic s/democrats-impeachment.html	https://www.nytimes.com/section/magazine
https://www.nytimes.com/2019/09/29/us/politic s/pelosi-trump-impeachment.html	https://www.nytimes.com/section/politics
https://www.nytimes.com/2018/01/23/us/politic s/impeach-trump-democrats-tom-steyer.html	https://www.nytimes.com/section/us
https://www.nytimes.com/2019/09/24/us/politic s/trump-democrats.html	https://www.nytimes.com/section/opinion
https://www.nytimes.com/2019/10/28/us/politic s/trump-impeachment-subpoena-house.html	https://www.nytimes.com/issue/todayspaper/20 19/11/10/todays-new-york-times
https://www.nytimes.com/issue/magazine/2019/ 06/25/the-61619-issue	https://www.nytimes.com/section/opinion/colu mnists
https://www.nytimes.com/2019/11/05/opinion/t rump-democrats-2020.html	https://www.nytimes.com/section/upshot
https://www.nytimes.com/2019/09/11/opinion/b lack-voters-democrats-2020-election.html	https://www.nytimes.com/news-event/2020- election



4. Patriot Movement

Lucene	PageRank
https://www.nytimes.com/2019/08/14/opinion/j oe-walsh-trump-primary.html	https://www.nytimes.com
https://www.nytimes.com/1976/08/27/archives/l ewis-michaux-92-dies-ran-bookstore-in- harlem.html	https://www.nytimes.com/video
https://www.nytimes.com/2019/05/15/us/politic s/us-iraq-embassy-evacuation.html	https://www.nytimes.com/video/opinion
https://www.nytimes.com/2019/01/07/obituarie s/moshe-arens-dead.html	https://www.nytimes.com/section/t-magazine
https://www.nytimes.com/2019/10/27/obituarie s/john-conyers-jr-dead.html	https://www.nytimes.com/section/opinion
https://www.nytimes.com/2018/12/29/obituarie s/shehu-shagari-dead.html	https://www.nytimes.com/section/technology
https://www.nytimes.com/2019/10/22/sports/fo otball/jets-patriots.html	https://www.nytimes.com/issue/todayspaper/20 19/11/10/todays-new-york-times
https://www.nytimes.com/by/jaclyn-peiser	https://www.nytimes.com/section/lens
https://www.nytimes.com/2019/09/23/opinion/d emocrats-republicans.html	https://www.nytimes.com/video/opinion/100000 006525694/stonewall-the-making-of-a- monument.html
https://www.nytimes.com/2019/10/30/opinion/a lexander-vindman-trump-ukraine.html	https://www.nytimes.com/2019/11/01/arts/telev ision/new-to-stream-netflix.html



5. Republicans

Lucene	PageRank
https://www.nytimes.com/by/catie-edmondson	https://www.nytimes.com
https://www.nytimes.com/2019/10/24/us/politics/senate-republicans-trump.html	https://www.nytimes.com/section/politics
https://www.nytimes.com/interactive/2019/08/1 4/magazine/republicans-racism-african- americans.html	https://www.nytimes.com/section/us
https://www.nytimes.com/2019/11/04/us/politics/trump-republicans-impeachment.html	https://www.nytimes.com/issue/todayspaper/201 9/11/10/todays-new-york-times
https://www.nytimes.com/2019/10/31/us/politics/house-impeachment-partisan-vote.html	https://www.nytimes.com/section/upshot
https://www.nytimes.com/2019/11/04/us/politics /trump-republicans-impeachment.html	https://www.nytimes.com/news-event/2020- election
https://www.nytimes.com/2019/10/04/us/politics/republicans-trump-defense.html	https://www.nytimes.com/video/us/politics/1000 00006067682/president-trump-impeach- questions-answers-michael-cohen.html
https://www.nytimes.com/2019/11/09/us/politics/republican-strategy-impeachment-trump.html	https://www.nytimes.com/video/us/elections/100 000006810114/kentucky-election-bevin- beshear.html
https://www.nytimes.com/2019/09/30/opinion/r epublicans-trump-impeachment.html	https://www.nytimes.com/video/us/politics/1000 00006812485/hillary-clinton-medicare.html
https://www.nytimes.com/2019/11/06/us/trump- senate-republicans-courts.html	https://www.nytimes.com/2019/11/09/us/politics /virginia-elections-democrats-republicans.html



6. Senate

Lucene	PageRank
https://www.nytimes.com/2019/10/01/us/politic s/senate-trump-impeachment.html	https://www.nytimes.com/section/us
https://www.nytimes.com/by/carl-hulse	https://www.nytimes.com/issue/todayspaper/20 19/11/10/todays-new-york-times
https://www.nytimes.com/2019/10/26/us/trump -senate-presidency.html	https://www.nytimes.com/section/climate
https://www.nytimes.com/2019/08/03/us/politic s/senate-votes-mcconnell.html	https://www.nytimes.com/news-event/2020- election
https://www.nytimes.com/2019/08/22/us/politic s/john-hickenlooper-senate-2020.html	https://www.nytimes.com/video/opinion/100000 006064751/coal-miner-to-trump-coal-mining- isnt-coming-back.html
https://www.nytimes.com/2019/11/08/us/impea chment-senate-democrats.html	https://timesjourneys.nytimes.com/tour-type/
https://www.nytimes.com/2019/10/24/us/politic s/senate-republicans-trump.html	https://www.nytimes.com/times-journeys/tour- type/small-group/
https://www.nytimes.com/2019/08/31/us/politic s/senate-race-2020.html	https://www.nytimes.com/times-journeys/trip- type/politics-and-perspective/
https://www.nytimes.com/2019/10/18/us/politic s/mcconnell-impeachment-senate.html	https://www.nytimes.com/video/us/politics/1000 00006639814/presidential-candidates- campaigns.html
https://www.nytimes.com/interactive/2016/us/politics/donald-trump-administration.html	https://www.nytimes.com/video/opinion/100000 006188105/countering-disinformation-active- measures.html



7. Olympics 2020

Lucene	PageRank
https://www.nytimes.com/2019/10/01/travel/w hat-you-need-to-know-the-tokyo-2020- olympics.html	https://www.nytimes.com
https://www.nytimes.com/2018/02/15/learning/should-technology-in-sports-be-limited.html	https://www.nytimes.com/video
https://www.nytimes.com/2018/02/15/learning/should-technology-in-sports-be-limited.html	https://www.nytimes.com/section/movies
https://www.nytimes.com/2017/12/05/sports/o lympics/ioc-russia-winter-olympics.html	https://www.nytimes.com/section/arts/televisio n
https://www.nytimes.com/video/sports/olympic s/10000005750317/johnny-weir-tara-lipinski- figure-skating.html	https://www.nytimes.com/video/opinion
https://www.nytimes.com/video/sports/olympic s/10000005750317/johnny-weir-tara-lipinski- figure-skating.html	https://www.nytimes.com/es/
https://www.nytimes.com/video/sports/olympic s/100000005752444/south-korea- speedskaters.html	https://www.nytimes.com/section/magazine
https://www.nytimes.com/video/sports/olympic s/100000005752444/south-korea- speedskaters.html	https://www.nytimes.com/section/realestate
https://www.nytimes.com/2018/04/23/sports/b reakdancing-olympics.html	https://www.nytimes.com/section/politics
https://www.nytimes.com/video/sports/football /10000005827420/nfl-cheerleaders-and-the- gender-gap.html	https://www.nytimes.com/section/obituaries



8. Stock

Lucene	PageRank
https://www.nytimes.com/interactive/2017/10/0 4/us/bump-stock-las-vegas-gun.html	https://www.nytimes.com/section/world
https://www.nytimes.com/2019/05/02/business/t esla-stock-fundraising.html	https://www.nytimes.com/section/business
https://www.nytimes.com/2018/12/16/technolog y/tech-workers-company-stock-shareholder- activism.html	https://www.nytimes.com/issue/todayspaper/201 9/11/10/todays-new-york-times
https://www.nytimes.com/2019/05/10/technolog y/uber-stock-price-ipo.html	https://www.nytimes.com/section/automobiles
https://www.nytimes.com/2019/11/04/business/ under-armour-stock-investigation.html	https://www.nytimes.com/section/upshot
https://www.nytimes.com/2019/09/20/technolog y/airbnb-employees-ipo-payouts.html	https://www.nytimes.com/video/opinion/100000 006616246/the-king-of-fish-and-chips.html
https://www.nytimes.com/2019/10/24/style/reba g-clair-handbag-stock-market.html	https://www.nytimes.com/section/fashion
https://www.nytimes.com/2019/05/30/technolog y/uber-stock-earnings.html	https://www.nytimes.com/times- journeys/travel/egypt/
https://www.nytimes.com/2018/11/12/obituaries /david-pearson-nascar-dead.html	https://www.nytimes.com/2019/11/09/opinion/s unday/stripping-sex-work.html
https://www.nytimes.com/2019/04/03/business/t esla-sales-elon-musk.html	https://www.nytimes.com/video/movies/1000000 06675663/brittany-runs-a-marathon-scene.html



9. Virus

Lucene	PageRank
https://www.nytimes.com/2019/11/09/science/s eals-distemper.html	https://www.nytimes.com/es/
https://www.nytimes.com/2019/10/25/health/pr edict-usaid-viruses.html	https://www.nytimes.com/section/us
https://www.nytimes.com/2019/10/28/us/politic s/kay-hagan-dead.html	https://www.nytimes.com/section/world
https://www.nytimes.com/2019/02/05/us/politic s/trump-hiv-aids-plan.html	https://www.nytimes.com/section/science
https://www.nytimes.com/by/donald-g-mcneil-jr	https://www.nytimes.com/section/health
https://www.nytimes.com/2019/01/23/neediest- cases/daughter-cmv.html	https://www.nytimes.com/section/climate
https://www.nytimes.com/by/tariro-mzezewa	https://www.nytimes.com/video/opinion/100000 006210828/russia-disinformation-fake-news.html
https://www.nytimes.com/by/denise-grady	https://www.nytimes.com/video/opinion/100000 006188102/what-is-pizzagate.html
https://www.nytimes.com/2019/09/27/well/live/when-is-the-best-time-to-get-your-flu-shot.html	https://www.nytimes.com/video/opinion/100000 006188105/countering-disinformation-active- measures.html
https://www.nytimes.com/2019/11/06/health/cri spr-cancer-leukemia.html	https://www.nytimes.com/video/opinion/100000 006188102/what-is-pizzagate.html