Information Retrieval - Assignment 4

Name: Divya Ramesha   **USC ID: 5228569441**

Indexing the web using Solr and Comparing Search-Engine Ranking Algorithms
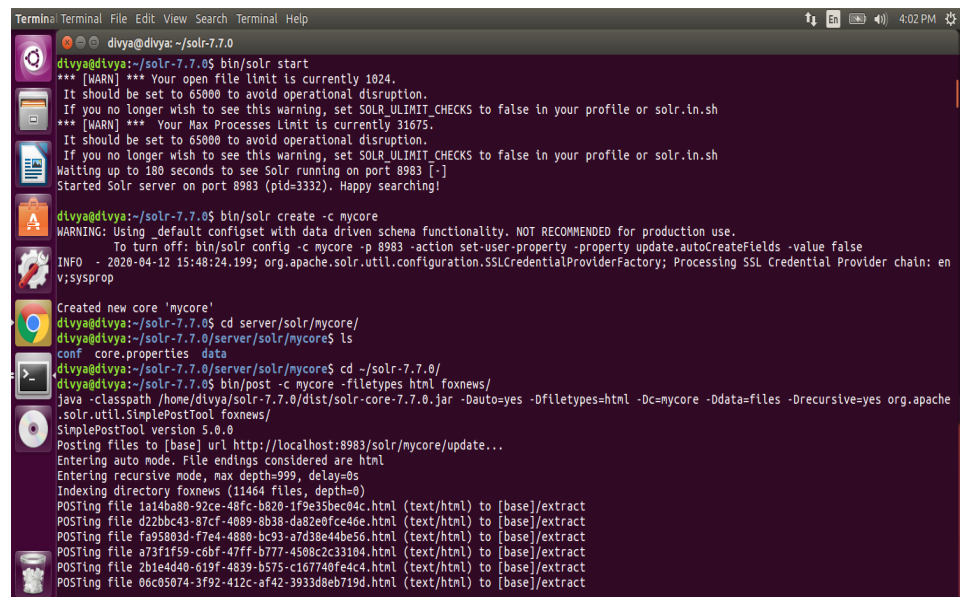
I. **Steps Taken:**
1. **Installation:**
   - Installed VirtualBox and Ubuntu 16.04
   - Adding Guest Additions required
   - Downloaded Apache Solr 7.7.0 and unzipped the contents
   - Downloaded and extracted all the crawled files from provided google drive link for FOXNEWS
2. **Indexing:**
   - Started the Solr server using **bin/solr start** command
   - Checked if Solr Admin was working properly
   - Created a core inside solr-7.7.0 directory using **bin/solr create -c mycore**
   - Uncommented **<copyField source="*" dest="_text_"/>** line present in managed-schema file within my core conf directory
   - Copied crawl folder downloaded above to solr directory
   - Indexed all html pages present in the crawl folder using the command **bin/post -c mycore -filetypes html foxnews/**
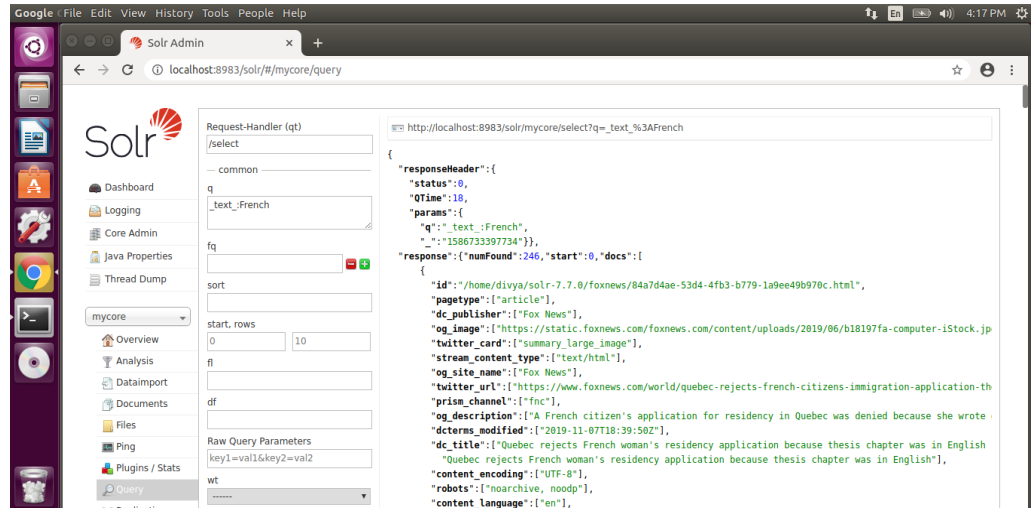

   - Updated solrconfig.xml to include **<str name ="df" >_text_</str>** line
   - Reloaded the core in the Solr UI: http://localhost:8983/solr/#/~cores/mycore
   - Verified that all the files are indexed

- Queried in Solr UI page to find relevant results making sure Solr is working fine using http://localhost:8983/solr/#/mycore/query

3. **Apache and PHP Installation:**
   - Installed Apache Server using **sudo apt-get install apache2**
   - Installed PHP using **sudo apt-get install libapache2-mod-php**
   - Installed git using **sudo apt-get install git**
   - I cloned https://github.com/PTCInc/solr-php-client.git repository inside /var/www/html folder
   - Created a php file using the given sample code
   - Client is implemented in PHP to run queries, retrieve response from solr server and display them to user.

4. **Creating EdgeList:**
   - Installed Java 8 JDK and JSoup library
   - Wrote a Java program using JSoup library to extract the link structure from downloaded web pages to compute edgeList.txt file
   - Created a directed graph in which vertices are web pages and edge between two web pages indicates a link between them

5. **Computing PageRank:**
   - Installed Python 2.7 and networkx package using pip
   - Used networkx library to compute page rank of web pages using edgeList.txt created above
   - Loaded the graph into networkx using read_edgelist() method with edgeList.txt and DiGraph as arguments
   - Used pagerank() method to compute page rank of web pages and stored the results in external_pageRankFile.txt
   - Parameters used while computing page rank:
     - alpha = 0.85
     - personalization = None
     - max_iter = 30

- o tol = 1e-06
- o nstart = None
- o weight = 'weight'
- o dangling = None

6. **Configuration Changes for PageRank Algorithm:**
   - Copied the external_pageRankFile.txt to data directory of mycore
   - Updated managed-schema file for using external pagerank file for ranking
     - o **<fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField" />**
     - o **<field name="pageRankFile" type="external" stored="false" indexed="false" />**
   - Updated solrconfig.xml to make sure when the index is reloaded, it can access the page rank file
     - o **<listener event = "newSearcher" class = "org.apache.solr.schema.ExternalFileFieldReloader"/>**
     - o **<listener event = "firstSearcher" class = "org.apache.solr.schema.ExternalFileFieldReloader"/>**
   - Reloaded the core and fetched results using page ranking algorithm

7. **Final Changes in PHP for Better Visualization:**
    ▪ Modified PHP code to include external page rank algorithm
    ▪ Added 2 radio buttons. One for default Solr Lucene Ranking algorithm and other for external page ranking algorithm
    ▪ Changed the way results are displayed. Handled null/empty values and returned Title, URL, ID and Description. Title and URL is made clickable

II.   Why some pages have higher page rank values than others?

Page Rank algorithm is a way to measure the importance of web pages and is used by search engines to rank websites in their results. Page Rank is based on how web pages are linked to each other based on the number of in-links and out-links. If a page has more in-links then that page is considered as very relevant and will be assigned a higher page rank. Moreover, if these in-link web pages are also relevant and have higher page rank scores, then the page rank of child web page pointed by these in-link web pages tends to increase. In our implementation we first created edgeList.txt file using Java JSoup library and supplied this as input to python networkx library to compute page rank for web pages.

III.   Screenshots depicting the flow for "Cannes" query

1. **Solr Admin UI Page:**

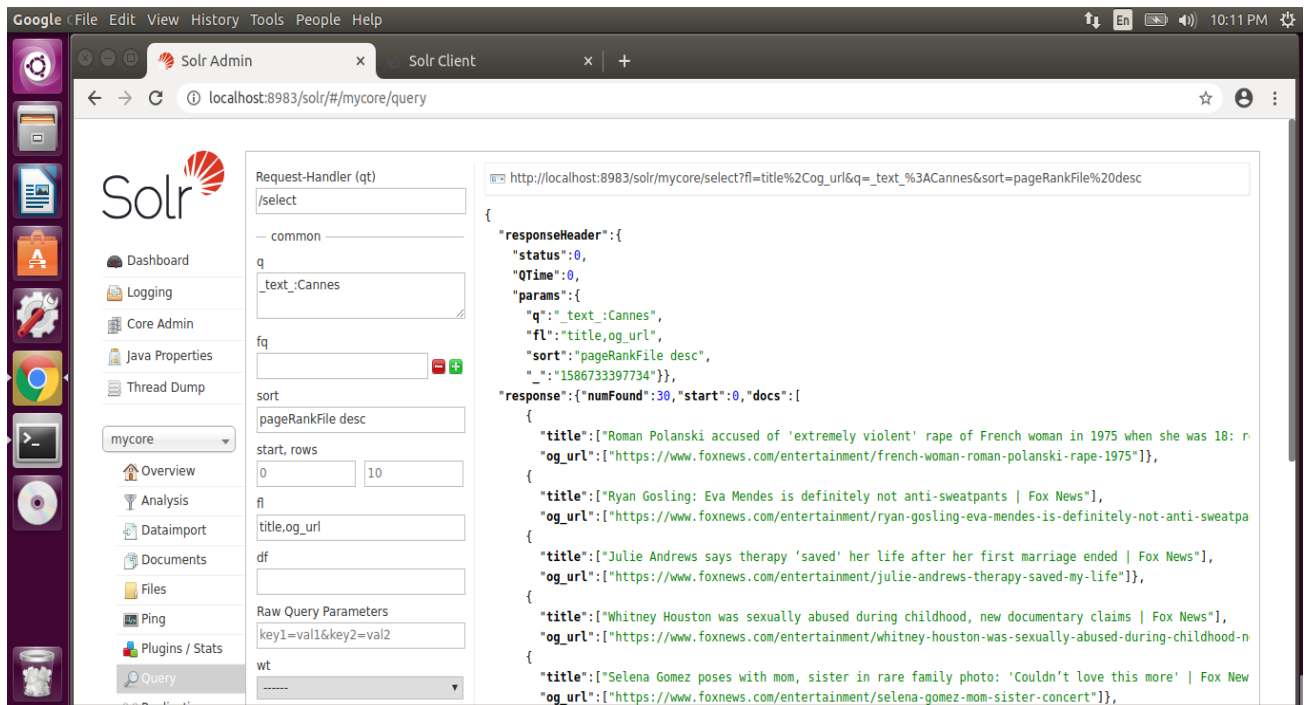   ▪ Solr Lucene Default Ranking Algorithm

- Page Rank Algorithm

2. PHP Solr Client WebPage:

- Initial Page
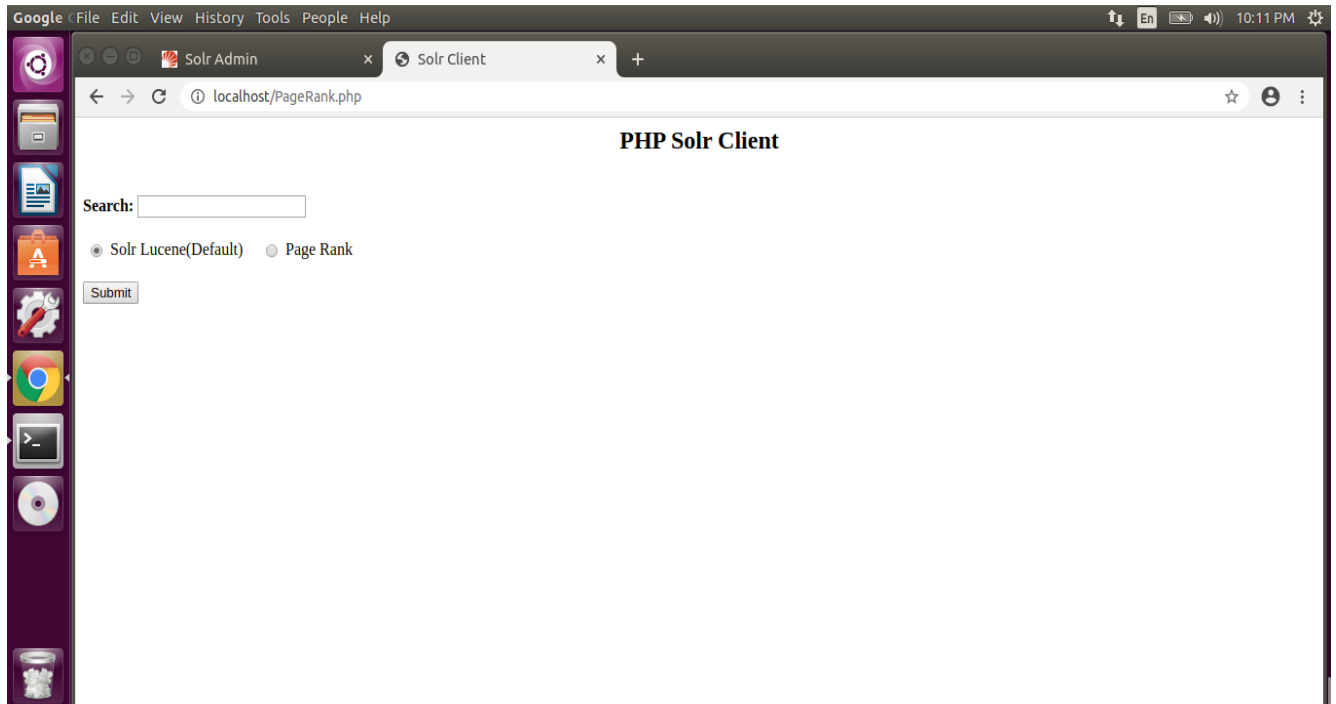
- Solr Default Lucene Ranking Result Page

- Page Rank Result Page

- Clicking on Result URL and Title opens the respective foxnews web page.
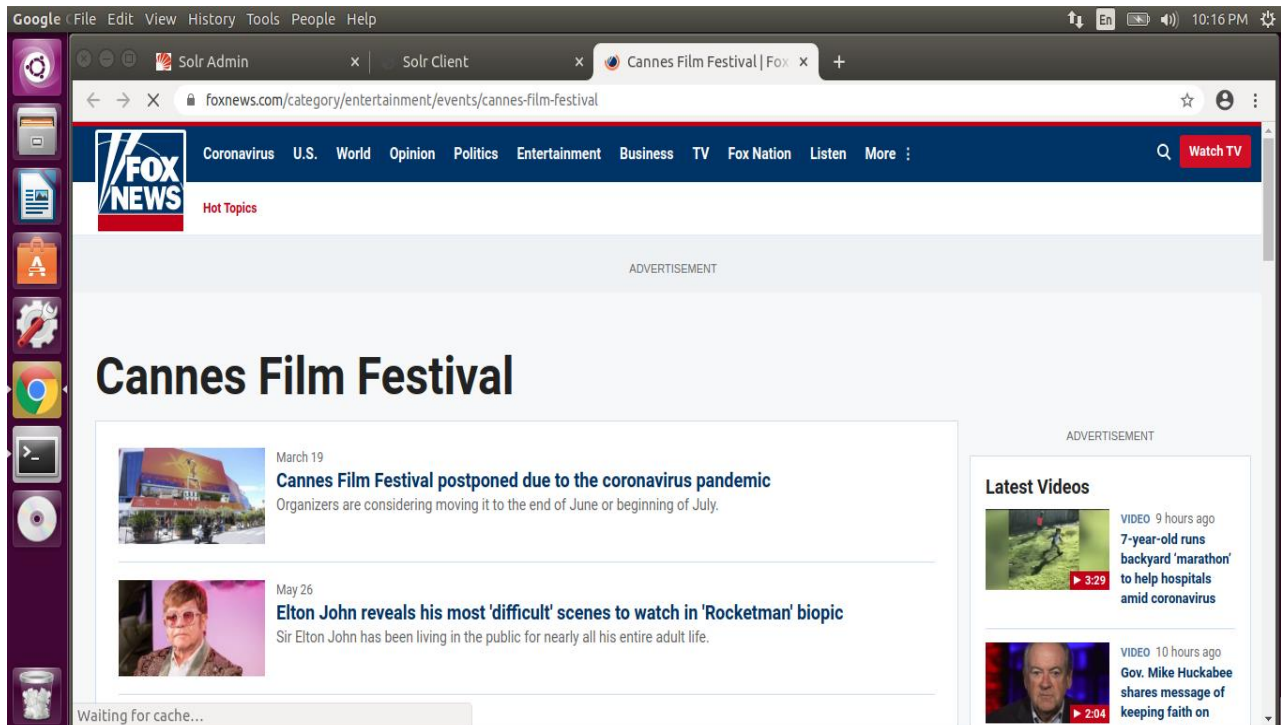
IV. Lucene and PageRank Top 10 Results Comparison Tables:

1. Cannes

No	Default Solr Results	Page Rank Algo Results
1	https://www.foxnews.com/category/entertainment/events/cannes-film-festival	https://www.foxnews.com/entertainment/french-woman-roman-polanski-rape-1975
2	https://www.foxnews.com/entertainment/selena-gomez-bill-murray-whispered-cannes-film-festival-reveal	https://www.foxnews.com/entertainment/ryan-gosling-eva-mendes-is-definitely-not-anti-sweatpants
3	https://www.foxnews.com/entertainment/selena-gomez-cannes-film-festival-red-carpet	https://www.foxnews.com/entertainment/julie-andrews-therapy-saved-my-life
4	https://www.foxnews.com/entertainment/ryan-gosling-eva-mendes-is-definitely-not-anti-sweatpants	https://www.foxnews.com/entertainment/whitney-houston-was-sexually-abused-during-childhood-new-documentary-claims
5	https://www.foxnews.com/entertainment/quentin-tarantino-once-upon-a-time-in-hollywood-trailer-release	https://www.foxnews.com/entertainment/selena-gomez-mom-sister-concert
6	https://www.foxnews.com/entertainment/woody-allen-retire-family-politics	https://www.foxnews.com/entertainment/sharon-stone-rocks-bikini-see-her-many-ageless-looks
7	https://www.foxnews.com/entertainment/whitney-houston-was-sexually-abused-during-childhood-new-documentary-claims	https://www.foxnews.com/entertainment/dolph-lundgren-sylvester-stallone-rocky-iv-hospital
8	https://www.foxnews.com/entertainment/elton-john-rocketman-biopic-drugs-addiction-eating-disorder	https://www.foxnews.com/entertainment/michael-moore-endorses-bernie-sanders
9	https://www.foxnews.com/opinion/a-joan-rivers-thanksgiving	https://www.foxnews.com/entertainment/elton-john-rocketman-biopic-drugs-addiction-eating-disorder
10	https://www.foxnews.com/entertainment/selena-gomez-new-dog-instagram	https://www.foxnews.com/category/entertainment/events/cannes-film-festival

2. Congress

No	Default Solr Results	Page Rank Algo Results
1	https://www.foxnews.com/us/navy-vet-donates-gold-star-families-stories-to-library-of-congress	https://www.foxnews.com/opinion
2	https://www.foxnews.com/opinion/judge-andrew-napolitano-congress-constitutional-power-declare-war	https://www.foxnews.com/tech
3	https://www.foxnews.com/politics/jon-stewart-blasts-congress-over-9-11-fund-	https://www.foxnews.com/category/tech/topics/security

No	Default Solr Results	Page Rank Algo Results
	hearing-drove-me-nuts-there-were-empty-seats	
4	https://www.foxnews.com/opinion/rep-andy-biggs-congress-debt-ceiling-budget-deal	https://www.foxnews.com/category/world/conflicts
5	https://www.foxnews.com/opinion/gutfeld-congress-attacking-zuckerberg	https://www.foxnews.com/official-polls
6	https://www.foxnews.com/opinion/andy-biggs-congress-judges-immigration-laws	https://www.foxnews.com/category/politics/executive
7	https://www.foxnews.com/faith-values/bible-school-drew-brees-day-congress	https://www.foxnews.com/category/politics/foreign-policy
8	https://www.foxnews.com/category/us/congress	https://www.foxnews.com/opinion/mike-huckabee-lets-fight-for-fabricated-structural-steel-manufacturing-jobs
9	https://www.foxnews.com/politics/trump-tells-congress-to-immediately-approve-mexico-canada-trade-deal-says-us-moving-past-stupid-years-of-trade-policy	http://www.foxnews.com/trending
10	https://www.foxnews.com/opinion/bradley-blakeman-trade-pact-with-canada-and-mexico-offers-incredible-benefits-congress-should-approve-it	https://www.foxnews.com/media/elijah-cummings-seat-kimberly-klacik-baltimore-election-republican

3. Democrats

No	Default Solr Results	Page Rank Algo Results
1	https://www.foxnews.com/opinion/rnc-chair-mcdaniel-democrats-impeachment-vote-is-baseless-political-vendetta-trump-did-nothing-wrong	https://www.foxnews.com/opinion
2	https://www.foxnews.com/politics/democrats-arent-actually-impeaching-the-president-yet-but-they-are-digging-deep	https://www.foxnews.com/us
3	https://www.foxnews.com/opinion/david-bossie-baseless-trump-impeachment-inquiry-by-desperate-democrats-makes-dramatic-flip-flop	https://www.foxnews.com/
4	https://www.foxnews.com/politics/federal-judge-house-democrats-mazars-subpoena-trump-financial-record	https://www.foxnews.com/politics
5	https://www.foxnews.com/opinion/rep-andy-biggs-lipstick-impeachment	https://www.foxnews.com/category/world/scandals
6	https://www.foxnews.com/media/democrats-disdain-working-class-trump-jr-fox-friends	https://www.foxnews.com/category/politics/senate
7	https://www.foxnews.com/opinion/liz-harrington-democrats-holding-soviet-style-	https://www.foxnews.com/category/us/personal-freedoms

No	Default Solr Results	Page Rank Algo Results
	show-trial-to-overthrow-trump-not-an-impeachment-inquiry	
8	https://www.foxnews.com/opinion/donald-trump-jr-democrats-immigration-radical-ideology	https://www.foxnews.com/category/world/conflicts
9	https://www.foxnews.com/opinion/hans-von-spakovsky-in-trump-impeachment-probe-democrats-refuse-to-follow-nixon-and-clinton-precedents	https://www.foxnews.com/category/politics/house-of-representatives
10	https://www.foxnews.com/opinion/david-bossie-trump-impeachment-vote-is-democratic-declaration-of-war-republicans-must-declare-war-on-dems	https://www.foxnews.com/category/entertainment/music

4. Patriot Movement

No	Default Solr Results	Page Rank Algo Results
1	https://www.foxnews.com/politics/trump-antifa-terrorist-portland-protests	https://www.foxnews.com/science
2	https://www.foxnews.com/world/iranian-cruise-missiles-drones-saudi-oil	https://www.foxnews.com/
3	https://www.foxnews.com/media/fox-nation-patriot-awards-american-heroes-st-petersburg	https://www.foxnews.com/tech
4	https://www.foxnews.com/media/fox-nation-patriot-awards	https://www.foxnews.com/category/entertainment/movies
5	https://www.foxnews.com/media/marine-vet-most-valuable-patriot-awards-congress	https://www.foxnews.com/category/science/planet-earth
6	https://www.foxnews.com/faith-values/veterans-serve-disaster-relief	https://www.foxnews.com/entertainment/gwyneth-paltrow-relationship-harvey-weinstein
7	https://www.foxnews.com/media/tomi-lahren-patriot-awards-fox-nation	https://www.foxnews.com/politics/aoc-bloomberg-purchase-election
8	https://www.foxnews.com/entertainment/netflix-pulls-patriot-act-with-hasan-minhaj-episode-after-getting-valid-legal-complaint-from-saudi-arabia	https://www.foxnews.com/world/ukraine-russia-separatists-pullback
9	https://www.foxnews.com/media/vice-president-mike-pence-fox-nation-patriot-awards	https://www.foxnews.com/science/earths-mantle-and-crust-are-in-a-fiery-battle-to-the-death-of-supercontinents
10	https://www.foxnews.com/tech/netflix-defends-pulling-patriot-act-episode-in-saudi-arabia	https://www.foxnews.com/auto/2020-ram-1500-built-to-serve-pickup-salutes-the-u-s-military

5. Republicans

No	Default Solr Results	Page Rank Algo Results
1	https://www.foxnews.com/politics/house-republicans-storm-impeachment-inquiry-deposition-what-went-down	https://www.foxnews.com/opinion
2	https://www.foxnews.com/opinion/tom-del-beccaro-republicans-must-win-the-impeachment-trial-and-they-can-by-following-these-five-steps	https://www.foxnews.com/
3	https://www.foxnews.com/politics/republicans-hunter-biden-ukraine-whisteleblower-impeachment-witness	https://www.foxnews.com/politics
4	https://www.foxnews.com/politics/house-republicans-plan-to-call-adam-schiff-to-testify-in-impeachment-inquiry-call-him-a-fact-witness	https://www.foxnews.com/category/world/scandals
5	https://www.foxnews.com/politics/trump-settles-score-with-republicans-who-didnt-embrace-him-in-midterms	https://www.foxnews.com/category/politics/senate
6	https://www.foxnews.com/politics/republicans-impeachment-hearings-lineup	https://www.foxnews.com/category/world/conflicts
7	https://www.foxnews.com/politics/bill-weld-thrilled-other-republicans-are-challeneging-trump-in-primary	https://www.foxnews.com/category/politics/house-of-representatives
8	https://www.foxnews.com/entertainment/chelsea-handler-republicans-colorado-school-shooting	https://www.foxnews.com/official-polls
9	https://www.foxnews.com/opinion/matt-gorman-kavanaugh-republicans	https://www.foxnews.com/category/politics/elections
10	https://www.foxnews.com/opinion/newt-gingrich-house-republicans-2020-retirements	https://www.foxnews.com/politics

6. Senate

No	Default Solr Results	Page Rank Algo Results
1	https://www.foxnews.com/politics/trump-senate-trial-impeachment-inquiry-reporters-notebook	https://www.foxnews.com/world
2	https://www.foxnews.com/category/politics/2020-senate-races	https://www.foxnews.com/science
3	https://www.foxnews.com/media/jeff-sessions-former-embattled-trump-attorney-general-announces-bid-for-old-us-senate-seat	https://www.foxnews.com/travel

4	https://www.foxnews.com/opinion/newt-gingrich-left-viciously-attacks-mcconnell-because-hes-successful-senate-leader-and-trump-ally	https://www.foxnews.com/opinion
5	https://www.foxnews.com/faith-values/pastor-andrew-brunson-senate-trump-prayer	https://www.foxnews.com/us
6	https://www.foxnews.com/category/politics/elections/senate	https://www.foxnews.com/shows
7	https://www.foxnews.com/politics/jeff-sessions-alabama-senate-campaign-expected	https://www.foxnews.com/health
8	https://www.foxnews.com/politics/trump-sessions-against-senate-run-alabama	https://www.foxnews.com/entertainment
9	https://www.foxnews.com/politics/virginia-democrat-accused-of-sex-with-teen-secretary-wins-state-senate-seat	https://www.foxnews.com/
10	https://www.foxnews.com/media/trump-senate-mike-pence-joe-scarborough	https://www.foxnews.com/foxaroundtheworld

7. Olympics 2020

No	Default Solr Results	Page Rank Algo Results
1	https://www.foxnews.com/sports/alex-morgan-2020-olympics-child-birth	https://www.foxnews.com/opinion
2	https://www.foxnews.com/sports/naomi-osaka-us-citizenship-japan-2020-olympics	https://www.foxnews.com/entertainment
3	https://www.foxnews.com/category/sports/olympics	https://www.foxnews.com/
4	https://www.foxnews.com/tech/russian-hackers-attack-possible-olympic-ban	https://www.foxnews.com/politics
5	https://www.foxnews.com/person/c/barnini-chakraborty	https://www.foxnews.com/tech
6	https://www.foxnews.com/category/sports/soccer	https://www.foxnews.com/lifestyle
7	https://www.foxnews.com/category/person/alex-morgan	https://www.foxnews.com/category/health/heart-health
8	https://www.foxnews.com/category/sports/tennis	https://www.foxnews.com/category/us/education
9	https://www.foxnews.com/sports/kikkan-randall-new-york-city-marathon-skiing-cancer	https://www.foxnews.com/category/world/global-economy
10	https://www.foxnews.com/sports	https://www.foxnews.com/category/news-events/fox-news-investigates

8. Stock

No	Default Solr Results	Page Rank Algo Results
1	https://www.foxnews.com/tech/jeff-bezos-to-be-no-2-richest-behind-bill-gates-after-amazon-stock-drop	https://www.foxnews.com/
2	https://www.foxnews.com/transcript/cavuto-looking-back-at-the-stock-market-crash-30-years-ago	https://www.foxnews.com/category/us/economy
3	https://www.foxnews.com/transcript/cavuto-the-markets-are-easier-to-take-stock-of-than-life	https://www.foxnews.com/science/revolutionary-war-rifle-stolen-50-years-ago-on-display
4	https://www.foxnews.com/category/shows/your-world/transcript/common-sense	https://www.foxnews.com/entertainment/demi-lovato-first-interview-year
5	https://www.foxnews.com/category/tech/companies/microsoft	https://www.foxnews.com/media/chris-wallace-elizabeth-warren-medicare-for-all-plan
6	https://www.foxnews.com/category/person/sheryl-sandberg	https://www.foxnews.com/tech/hp-recalls-more-laptop-lithium-ion-batteries-after-reports-of-overheating-melting-and-charring
7	https://www.foxnews.com/category/tech/companies/amazon	https://www.foxnews.com/tech/heres-how-the-trump-tariffs-hike-may-affect-prices-for-pc-parts
8	https://www.foxnews.com/person/c/gordon-chang	https://www.foxnews.com/category/tech/companies/amazon
9	https://www.foxnews.com/auto/fiat-chrysler-sales-boss-reid-bigland-sues-company-for-violating-whistleblower-protections	https://www.foxnews.com/opinion/max-lucado-fearless-parenting
10	https://www.foxnews.com/sports/royals-hire-ex-cardinals-skipper-mike-matheny-as-manager	https://www.foxnews.com/auto/electric-ford-mustang-lithium

9. Virus

No	Default Solr Results	Page Rank Algo Results
1	https://www.foxnews.com/tech/computer-virus-cripples-tribune-publishing-los-angeles-times	https://www.foxnews.com/science
2	https://www.foxnews.com/science/deadly-seal-virus-spreading-melting-arctic-ice	https://www.foxnews.com/category/science/planet-earth
3	https://www.foxnews.com/health/connecticut-third-mosquito-eee-death	https://www.foxnews.com/category/tech/topics/computers
4	https://www.foxnews.com/category/tech/topics/virus	https://www.foxnews.com/category/science/wild-nature

5	https://www.foxnews.com/health/eee-eastern-equine-encephalitis-massachusetts-man-killed	https://www.foxnews.com/science/deadly-seal-virus-spreading-melting-arctic-ice
6	https://www.foxnews.com/health/what-is-norovirus-3-things-to-know-about-the-vomit-inducing-bug	https://www.foxnews.com/science/sandra-orangutan-legal-personhood-new-home-florida
7	https://www.foxnews.com/health/pregnant-women-who-get-flu-shot-are-less-likely-to-be-hospitalized-from-virus-cdc-says	https://www.foxnews.com/science/ancient-shipwrecks-stone-pyramid-anchors-discovered-greek-island
8	https://www.foxnews.com/category/health/infectious-disease/zika	https://www.foxnews.com/science/earths-mantle-and-crust-are-in-a-fiery-battle-to-the-death-of-supercontinents
9	https://www.foxnews.com/health/flu-season-2019-vaccine	https://www.foxnews.com/science/mammoth-traps-prehistoric-mexico
10	https://www.foxnews.com/health/british-mom-son-flu	https://www.foxnews.com/science/mercury-transit-2019-what-to-know

V.   Overlap Graph Results:

Query	Number of Overlaps
Cannes	4
Congress	0
Democrats	0
Patriot Movement	0
Republicans	0
Senate	0
Olympics 2020	0
Stock	1
Virus	1