Name: Sayali Ashok Lagad

USC ID: 7176433953

Website Indexed: Fox News

Step 1: Installation of VirtualBox and Ubuntu on Windows Machine

- Downloaded VirtualBox and Ubuntu from the links provided.
- Created a Virtual Machine in VirtualBox:

New -> Name: CSCI572-Linux, Type: Linux, Version: Ubuntu (64-bit); Next -> Memory size: 2048 MB; Next -> Create a virtual hard disk now; Create -> VDI (Virtual Disk Image); Next -> Dynamically allocated; Next -> File size: 128GB; Create.

- Started the virtual machine created above by clicking the **Start** button.
- Selected ubuntu downloaded in Step 1 as the start-up disk->**Start**.
- Clicked on Install Ubuntu.
- Provided details like: Name, Computer's name, Username, Password->Continue.
- Added guest devices to the machine by following the instructions provided in the pdf.

Step 2: Installation of Java in Virtual Machine

- Installed Java by following the procedure mentioned in the pdf.

Step 3: Installation of Solr and Indexing HTML Files

- Downloaded Solr 7.7.0. Unzipped and opened the terminal from the unzipped folder.
- Created a collection on one port and uploaded the provided HTML files for FOX NEWS by following commands as below:
 - \$ bin/solr start -p 8080 (started Solr on port 8080)

```
sayali@sayali-VirtualBox:~/Downloads/solr-7.7.0$ bin/solr start -p 8080

*** [WARN] *** Your open file limit is currently 1024.

It should be set to 65000 to avoid operational disruption.

If you no longer wish to see this warning, set SOLR_ULIMIT_CHECKS to false in y our profile or solr.in.sh

*** [WARN] *** Your Max Processes Limit is currently 7734.

It should be set to 65000 to avoid operational disruption.

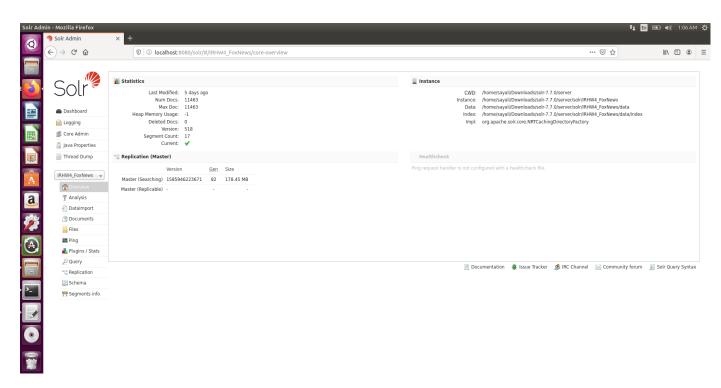
If you no longer wish to see this warning, set SOLR_ULIMIT_CHECKS to false in y our profile or solr.in.sh
```

- \$ bin/solr create -c IRHW4_FoxNews -p 8080 (created collection named IRHW4_FoxNews on port 8080)
- Opened the file managed-schema.xml located within conf folder of the collection and uncommented the line <copyField source="*" dest="_text_"/> to combine all information from different fields into a single field. Saved the changes. Copied the folder containing downloaded HTML files inside the folder of Solr and switched back to terminal.
- \$ bin/post -c IRHW4_FoxNews -filetypes html foxnews/ -p 8080 (indexed the files using post tool of Solr)

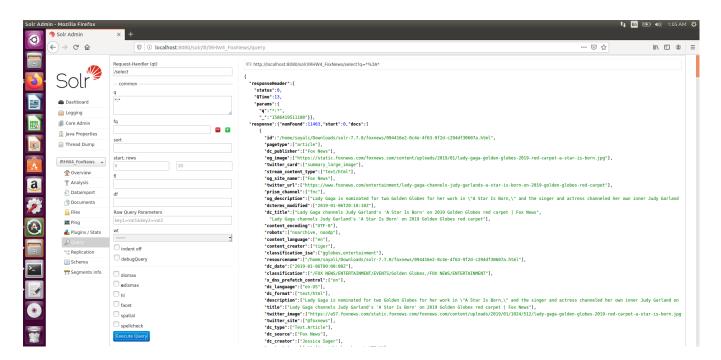
```
**Csayali@sayali-VirtualBox:*/Downloads/solr-7.7.0$ bin/post -c IRHM4_FOXNEWS -filetypes html foxnews/ -p 8080
    java -classpath /home/sayali/Downloads/solr-7.7.0/dist/solr-core-7.7.0.jar -Dauto=yes -Dfiletypes=html -Dport=8080 -Dc=IRHM4_FOXNEWS -Ddata=files -Drecursive=yes org.apache.solr.util.SimplePost    s/
    SimplePostTool version 5.0.0
    Posting files to [base] url http://localhost:8080/solr/IRHM4_FOXNEWS/update...
    Entering auto mode. File endings considered are html
    Entering recursive mode, max depth=999, delay=0s
    Indexing directory foxnews (11464 files, depth=0)

Indexing directory foxnews (11464 files, depth=0)
```

Verified if the Solr UI works well on port 8080.



Ran a basic query to verify if the indexing is done correctly.



Step 4: Extracting URLs from HTML Files

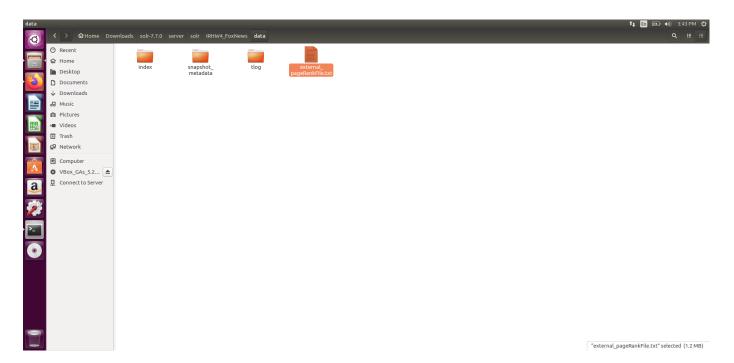
- Downloaded jsoup-1.11.3.jar and opencsv-5.1.jar required to parse HTML pages and read CSV file respectively.
- Wrote Java code to extract outgoing URLs from each downloaded HTML file and create an edge between the current page and the outgoing page.
- Created an edgeList.txt file that represents an edge between an HTML page and its outgoing HTML page.

Step 5: Calculating Page Rank

- Installed pip using the command python get-pip.py
- Installed network library using command pip install networkx==2.2
- Wrote a python code to form a graph with the provided edges and calculate page rank using the graph and the following attributes: alpha=0.85, personalization=None, max_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None
- Created **external_pageRankFile.txt** file which represents page rank for each HTML file using the above code.

Step 6: Configuration settings in xml files of Solr

- Placed the above external_pageRankFile.txt inside data folder of the created core.



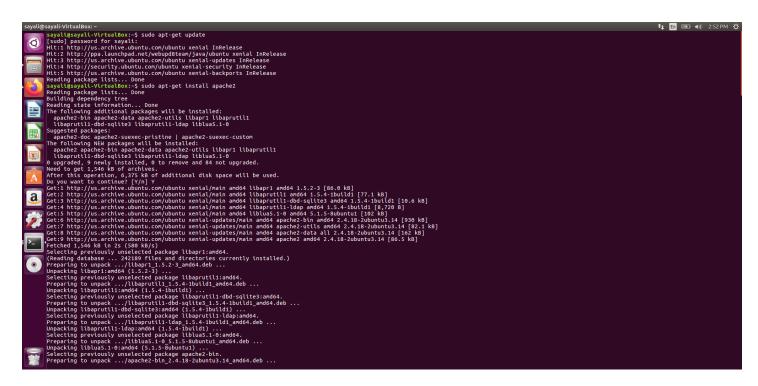
- Modified the **managed-schema.xml** file to add a field names pageRankFile which refers to the page rank score for each HTML file. Added the following lines to the file:
 - <fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField"/>
 <field name="pageRankFile" type="external" stored="false" indexed="false"/>
- Modified **solrconfig.xml** file to define event listeners to load the external file in the searcher thus enabling us to use an external Page ranking algorithm within Solr. The following lines were added:

tener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
tener event="firstSearcher" class=" org.apache.solr.schema.ExternalFileFieldReloader"/>

- Reloaded the index through the Solr UI.

Step 7: Installation of Apache and PHP

- Executed the following command in terminal:
 - \$ sudo apt-get update
 - \$ sudo apt-get apache2



\$ sudo apt install php libapache2-mod-php

```
symboly applications of the company of the company
```

\$ sudo system ctl restart apache2

Step 8: Query Box using PHP

- Cloned the git repository https://github.com/PTCInc/solr-php-client.git in the path: var/www/html/
- Placed the provided URL mapping csv file in the same path.
- Created a .php file using the sample code and made necessary changes such as changing the port number of Solr, adding radio buttons for page rank selection, limiting the fields to be displayed, and the search strategy to be used.
- Also modified **Service.php** file present in Apache/solr folder of the cloned git repo in order to reflect correct port number and path on which Solr is running.

```
Service.php (/var/www/html/solr-php-client/Apache/Soir) - gedit

Open * IP

Constructor. All parameters are optional and will take on default values

if not specified.

Open a string Sport

Open a string Sport

Open a string Sport

Open an string Sport

Open a string Sport

Open a string Sport

Open an string Sport

Open a string Sport

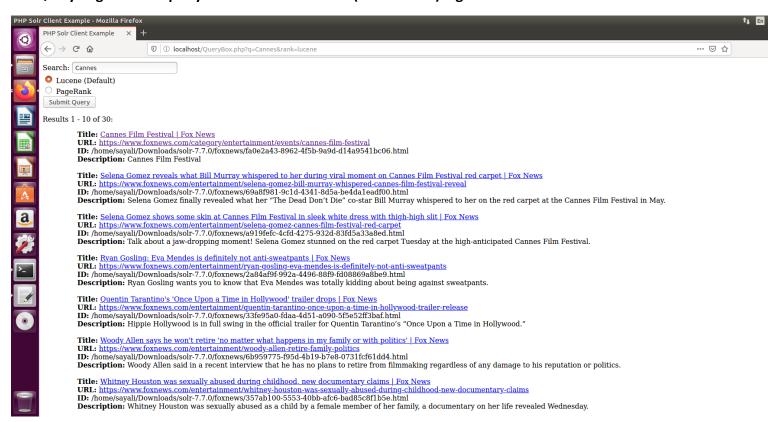
Ope
```

- Below are the screenshots of the UI:

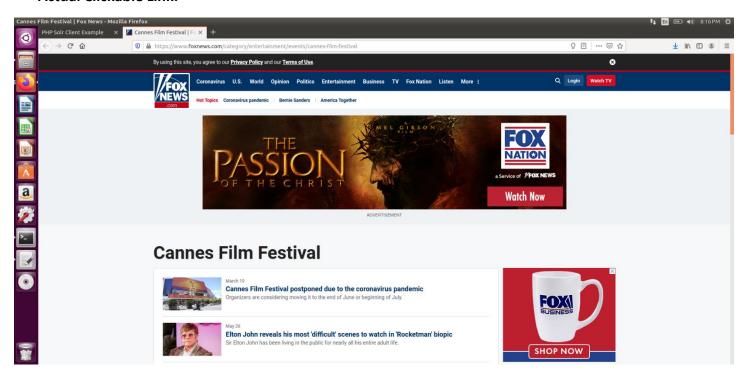
Initial Query Page:



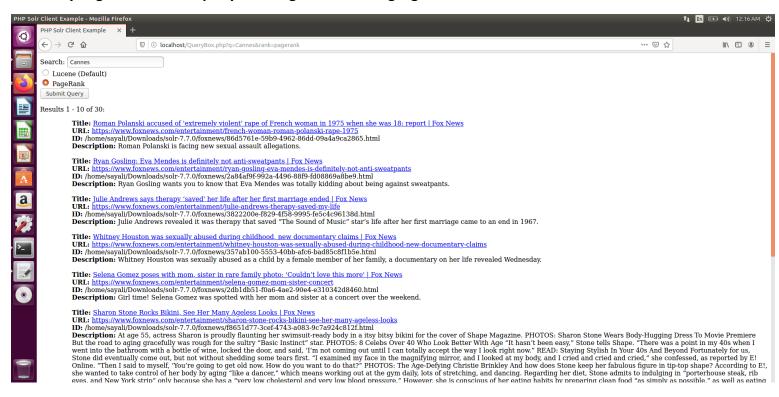
Query Page for the query 'Cannes' and 'Lucene' (Solr default) algorithm:



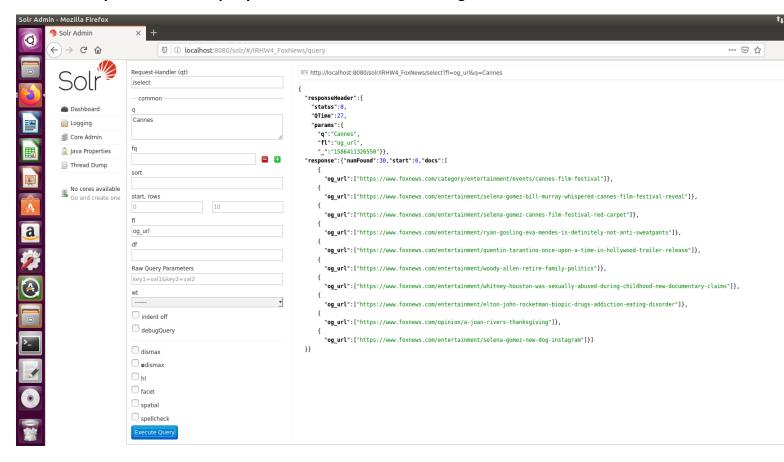
Actual Clickable Link:



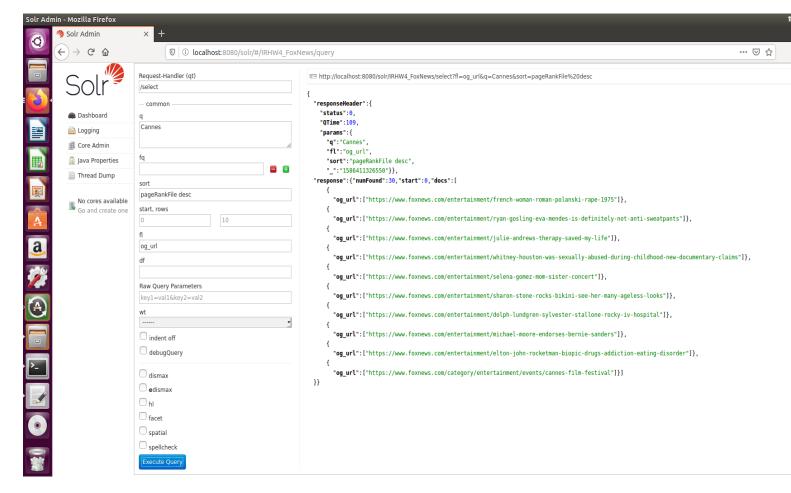
Query Page for 'Cannes' query and 'PageRank' ranking algorithm:



Solr Query Results for the query 'Cannes' and default Lucene algorithm:



Solr Query Results for the query 'Cannes' and external Page Rank algorithm:



Reason for some pages having higher page rank than others:

The java program was written to find out the number of outgoing pages from a page and using the Python code, a directed graph was formed were each node represents an HTML page and edges represent an outgoing link from one node to another. The page-rank method of network library uses this graph and computes page rank for each page depending on the incoming links for each node. As each page has different # of incoming links, the page rank for each page comes out to be different. For nodes (pages) having larger # of incoming links, the calculated page rank is higher as compared to nodes (pages) having lesser # of incoming links.

Below are the query results for each query and both ranking algorithms:

1. Cannes:

Sr.	Lucene (Default)	PageRank
N		
ο.		
1.	https://www.foxnews.com/category/entertainment	https://www.foxnews.com/entertainment/french-
	/events/cannes-film-festival	woman-roman-polanski-rape-1975

2.	https://www.foxnews.com/entertainment/selena-	https://www.foxnews.com/entertainment/ryan-
	gomez-bill-murray-whispered-cannes-film-festival-	gosling-eva-mendes-is-definitely-not-anti-
	reveal	<u>sweatpants</u>
3.	https://www.foxnews.com/entertainment/selena-	https://www.foxnews.com/entertainment/julie-
	gomez-cannes-film-festival-red-carpet	andrews-therapy-saved-my-life
4.	https://www.foxnews.com/entertainment/ryan-	https://www.foxnews.com/entertainment/whitney-
	gosling-eva-mendes-is-definitely-not-anti-	houston-was-sexually-abused-during-childhood-
	sweatpants	new-documentary-claims
5.	https://www.foxnews.com/entertainment/quentin-	https://www.foxnews.com/entertainment/selena-
	tarantino-once-upon-a-time-in-hollywood-trailer-	gomez-mom-sister-concert
	<u>release</u>	
6.	https://www.foxnews.com/entertainment/woody-	https://www.foxnews.com/entertainment/sharon-
	<u>allen-retire-family-politics</u>	stone-rocks-bikini-see-her-many-ageless-looks
7.	https://www.foxnews.com/entertainment/whitney-	https://www.foxnews.com/entertainment/dolph-
	houston-was-sexually-abused-during-childhood-	<u>lundgren-sylvester-stallone-rocky-iv-hospital</u>
	new-documentary-claims	
8.	https://www.foxnews.com/entertainment/elton-	https://www.foxnews.com/entertainment/michael-
	john-rocketman-biopic-drugs-addiction-eating-	moore-endorses-bernie-sanders
	<u>disorder</u>	
9.	https://www.foxnews.com/opinion/a-joan-rivers-	https://www.foxnews.com/entertainment/elton-
	thanksgiving	john-rocketman-biopic-drugs-addiction-eating-
		<u>disorder</u>
10	https://www.foxnews.com/entertainment/selena-	https://www.foxnews.com/category/entertainment
•	gomez-new-dog-instagram	<u>/events/cannes-film-festival</u>
Ove	rlap: 4	

2. Congress:

Sr.	Lucene (Default)	PageRank
No.		
1.	https://www.foxnews.com/us/navy-vet-donates-	https://www.foxnews.com/opinion
	gold-star-families-stories-to-library-of-congress	
2.	https://www.foxnews.com/opinion/judge-andrew-	https://www.foxnews.com/tech
	napolitano-congress-constitutional-power-declare-	
	<u>war</u>	
3.	https://www.foxnews.com/politics/jon-stewart-	https://www.foxnews.com/official-polls
	blasts-congress-over-9-11-fund-hearing-drove-me-	
	nuts-there-were-empty-seats	
4.	https://www.foxnews.com/opinion/rep-andy-	https://www.foxnews.com/category/politics/foreign-
	biggs-congress-debt-ceiling-budget-deal	policy
5.	https://www.foxnews.com/opinion/gutfeld-	https://www.foxnews.com/category/politics/executive
	congress-attacking-zuckerberg	
6.	https://www.foxnews.com/opinion/andy-biggs-	https://www.foxnews.com/category/tech/topics/security
	congress-judges-immigration-laws	

7.	https://www.foxnews.com/faith-values/bible-	https://www.foxnews.com/category/world/conflicts
	school-drew-brees-day-congress	
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-	http://www.foxnews.com/trending
	congress-to-immediately-approve-mexico-canada-	
	trade-deal-says-us-moving-past-stupid-years-of-	
	<u>trade-policy</u>	
10.	https://www.foxnews.com/opinion/bradley-	https://www.foxnews.com/media/elijah-cummings-seat-
	blakeman-trade-pact-with-canada-and-mexico-	kimberly-klacik-baltimore-election-republican
	offers-incredible-benefits-congress-should-	
	approve-it	
Over	·lap: 0	

3. Democrats:

Sr.	Lucene (Default)	PageRank
No.		
1.	https://www.foxnews.com/opinion/rnc-chair-	https://www.foxnews.com/opinion
	mcdaniel-democrats-impeachment-vote-is-	
	baseless-political-vendetta-trump-did-nothing-	
	wrong	
2.	https://www.foxnews.com/politics/democrats-	https://www.foxnews.com/us
	arent-actually-impeaching-the-president-yet-but-	
	they-are-digging-deep	
3.	https://www.foxnews.com/opinion/david-bossie-	https://www.foxnews.com
	baseless-trump-impeachment-inquiry-by-	
	desperate-democrats-makes-dramatic-flip-flop	
4.	https://www.foxnews.com/politics/federal-judge-	https://www.foxnews.com/politics
	house-democrats-mazars-subpoena-trump-	
	<u>financial-record</u>	
5.	https://www.foxnews.com/opinion/rep-andy-	https://www.foxnews.com/category/entertainment/music
	biggs-lipstick-impeachment	
6.	https://www.foxnews.com/media/democrats-	https://www.foxnews.com/category/politics/senate
	disdain-working-class-trump-jr-fox-friends	
7.	https://www.foxnews.com/opinion/liz-harrington-	https://www.foxnews.com/category/us/personal-
	democrats-holding-soviet-style-show-trial-to-	<u>freedoms</u>
	overthrow-trump-not-an-impeachment-inquiry	
8.	https://www.foxnews.com/opinion/donald-	https://www.foxnews.com/official-polls
	trump-jr-democrats-immigration-radical-ideology	
9.	https://www.foxnews.com/opinion/hans-von-	https://www.foxnews.com/category/world/scandals
	spakovsky-in-trump-impeachment-probe-	
	democrats-refuse-to-follow-nixon-and-clinton-	
	precedents	
10.	https://www.foxnews.com/opinion/david-bossie-	https://www.foxnews.com/category/politics/foreign-
	trump-impeachment-vote-is-democratic-	policy

declaration-of-war-republicans-must-declare-war-
<u>on-dems</u>

Overlap: 0

4. Patriotic Movement:

Sr.	Lucene (Default)	PageRank	
No.			
1.	https://www.foxnews.com/politics/trump-antifa-	https://www.foxnews.com/science	
	terrorist-portland-protests		
2.	https://www.foxnews.com/world/iranian-cruise-	https://www.foxnews.com/	
	missiles-drones-saudi-oil		
3.	https://www.foxnews.com/media/fox-nation-	https://www.foxnews.com/tech	
	patriot-awards-american-heroes-st-petersburg		
4.	https://www.foxnews.com/media/fox-nation-	https://www.foxnews.com/category/entertainment/movies	
	<u>patriot-awards</u>		
5.	https://www.foxnews.com/media/marine-vet-	https://www.foxnews.com/category/science/planet-earth	
	most-valuable-patriot-awards-congress		
6.	https://www.foxnews.com/faith-values/veterans-	https://www.foxnews.com/entertainment/gwyneth-	
	serve-disaster-relief	<u>paltrow-relationship-harvey-weinstein</u>	
7.	https://www.foxnews.com/media/tomi-lahren-	https://www.foxnews.com/politics/aoc-bloomberg-	
	<u>patriot-awards-fox-nation</u>	<u>purchase-election</u>	
8.	https://www.foxnews.com/entertainment/netflix-	https://www.foxnews.com/world/ukraine-russia-	
	pulls-patriot-act-with-hasan-minhaj-episode-	<u>separatists-pullback</u>	
	after-getting-valid-legal-complaint-from-saudi-		
	<u>arabia</u>		
9.	https://www.foxnews.com/media/vice-president-	https://www.foxnews.com/science/earths-mantle-and-	
	mike-pence-fox-nation-patriot-awards	<u>crust-are-in-a-fiery-battle-to-the-death-of-supercontinents</u>	
10.	https://www.foxnews.com/tech/netflix-defends-	https://www.foxnews.com/auto/2020-ram-1500-built-to-	
	<u>pulling-patriot-act-episode-in-saudi-arabia</u>	serve-pickup-salutes-the-u-s-military	
Over	Overlap: 0		

5. Republicans:

Sr.	Lucene (Default)	PageRank
No.		
1.	https://www.foxnews.com/politics/house-	https://www.foxnews.com/opinion
	republicans-storm-impeachment-inquiry-deposition-	
	what-went-down	
2.	https://www.foxnews.com/opinion/tom-del-beccaro-	https://www.foxnews.com/
	republicans-must-win-the-impeachment-trial-and-	
	they-can-by-following-these-five-steps	
3.	https://www.foxnews.com/politics/republicans-	https://www.foxnews.com/politics
	hunter-biden-ukraine-whisteleblower-impeachment-	
	witness	

4.	https://www.foxnews.com/politics/house-	https://www.foxnews.com/category/politics/senate
	republicans-plan-to-call-adam-schiff-to-testify-in-	
	impeachment-inquiry-call-him-a-fact-witness	
5.	https://www.foxnews.com/politics/trump-settles-	https://www.foxnews.com/official-polls
	score-with-republicans-who-didnt-embrace-him-in-	
	midterms	
6.	https://www.foxnews.com/politics/republicans-	https://www.foxnews.com/category/world/scandals
	impeachment-hearings-lineup	
7.	https://www.foxnews.com/politics/bill-weld-thrilled-	https://www.foxnews.com/category/politics/house-of-
	other-republicans-are-challeneging-trump-in-primary	representatives
8.	https://www.foxnews.com/entertainment/chelsea-	https://www.foxnews.com/category/politics/elections
	handler-republicans-colorado-school-shooting	
9.	https://www.foxnews.com/opinion/matt-gorman-	https://www.foxnews.com/category/world/conflicts
	kavanaugh-republicans	
10.	https://www.foxnews.com/opinion/newt-gingrich-	https://www.foxnews.com/politics.html
	house-republicans-2020-retirements	
Over	lap: 0	

6. Senate:

Sr.	Lucene (Default)	PageRank
No.		
1.	https://www.foxnews.com/politics/trump-senate-trial-	https://www.foxnews.com/shows
	impeachment-inquiry-reporters-notebook	
2.	https://www.foxnews.com/category/politics/2020-senate-	https://www.foxnews.com/opinion
	<u>races</u>	
3.	https://www.foxnews.com/media/jeff-sessions-former-	https://www.foxnews.com/entertainment
	embattled-trump-attorney-general-announces-bid-for-old-us-	
	<u>senate-seat</u>	
4.	https://www.foxnews.com/opinion/newt-gingrich-left-	https://www.foxnews.com/world
	viciously-attacks-mcconnell-because-hes-successful-senate-	
	<u>leader-and-trump-ally</u>	
5.	https://www.foxnews.com/faith-values/pastor-andrew-	https://www.foxnews.com/health
	<u>brunson-senate-trump-prayer</u>	
6.	https://www.foxnews.com/politics/jeff-sessions-alabama-	https://www.foxnews.com/travel
	<u>senate-campaign-expected</u>	
7.	https://www.foxnews.com/politics/trump-sessions-against-	https://www.foxnews.com/us
	<u>senate-run-alabama</u>	
8.	https://www.foxnews.com/category/politics/elections/senate	https://www.foxnews.com/science
	haranti da ana ana antantita da ana antanti	huse // face a second
9.	https://www.foxnews.com/politics/virginia-democrat-	https://www.foxnews.com/
10	accused-of-sex-with-teen-secretary-wins-state-senate-seat	hillion II
10.	https://www.foxnews.com/media/trump-senate-mike-pence-	https://www.foxnews.com/closed-captioning
	<u>joe-scarborough</u>	
Ovei	·lap: 0	

7. Olympics 2020:

Sr.	Lucene (Default)	PageRank
No.		
1.	https://www.foxnews.com/sports/alex-morgan-	https://www.foxnews.com/opinion
	2020-olympics-child-birth	
2.	https://www.foxnews.com/sports/naomi-osaka-us-	https://www.foxnews.com/entertainment
	citizenship-japan-2020-olympics	
3.	https://www.foxnews.com/category/sports/olympics	https://www.foxnews.com/
4.	https://www.foxnews.com/tech/russian-hackers-	https://www.foxnews.com/tech
	attack-possible-olympic-ban	
5.	https://www.foxnews.com/person/c/barnini-	https://www.foxnews.com/politics
	chakraborty	
6.	https://www.foxnews.com/category/sports/soccer	https://www.foxnews.com/lifestyle
7.	https://www.foxnews.com/category/sports/tennis	https://www.foxnews.com/category/health/heart-
		health
8.	https://www.foxnews.com/category/person/alex-	https://www.foxnews.com/category/us/education
	morgan	
9.	https://www.foxnews.com/sports/kikkan-randall-	https://www.foxnews.com/category/us/personal-
	new-york-city-marathon-skiing-cancer	freedoms
10.	https://www.foxnews.com/sports	https://www.foxnews.com/category/tech/topics/video-
10.	neeps.//www.noxnews.com/sports	
0	dam. O	games
Over	·lap: 0	

8. Stock

Sr. No	Lucene (Default)	PageRank
•		
1.	https://www.foxnews.com/tech/jeff-bezos-to-be-no-	https://www.foxnews.com/
	<u>2-richest-behind-bill-gates-after-amazon-stock-drop</u>	
2.	https://www.foxnews.com/transcript/cavuto-	https://www.foxnews.com/category/us/economy
	looking-back-at-the-stock-market-crash-30-years-ago	
3.	https://www.foxnews.com/transcript/cavuto-the-	https://www.foxnews.com/science/revolutionary-
	markets-are-easier-to-take-stock-of-than-life	war-rifle-stolen-50-years-ago-on-display
4.	https://www.foxnews.com/category/shows/your-	https://www.foxnews.com/entertainment/demi-
	world/transcript/common-sense	<u>lovato-first-interview-year</u>
5.	https://www.foxnews.com/category/person/sheryl-	https://www.foxnews.com/media/chris-wallace-
	sandberg	elizabeth-warren-medicare-for-all-plan
6.	https://www.foxnews.com/category/tech/companie	https://www.foxnews.com/tech/hp-recalls-more-
	<u>s/microsoft</u>	<u>laptop-lithium-ion-batteries-after-reports-of-</u>
		overheating-melting-and-charring
7.	https://www.foxnews.com/category/tech/companie	https://www.foxnews.com/tech/heres-how-the-
	s/amazon	trump-tariffs-hike-may-affect-prices-for-pc-parts
8.	https://www.foxnews.com/person/c/gordon-chang	https://www.foxnews.com/category/tech/compani
		<u>es/amazon</u>

9.	https://www.foxnews.com/auto/fiat-chrysler-sales-	https://www.foxnews.com/opinion/max-lucado-
	boss-reid-bigland-sues-company-for-violating-	<u>fearless-parenting</u>
	whistleblower-protections	
10	https://www.foxnews.com/sports/royals-hire-ex-	https://www.foxnews.com/auto/electric-ford-
	cardinals-skipper-mike-matheny-as-manager	<u>mustang-lithium</u>
Overlap: 1		

9. Virus

Lucene (Default) Sr. **PageRank** No 1. https://www.foxnews.com/tech/computer-virushttps://www.foxnews.com/science cripples-tribune-publishing-los-angeles-times 2. https://www.foxnews.com/science/deadly-sealhttps://www.foxnews.com/category/science/planetvirus-spreading-melting-arctic-ice earth https://www.foxnews.com/health/connecticut-3. third-mosquito-eee-death https://www.foxnews.com/category/tech/topics/com puters 4. https://www.foxnews.com/category/tech/topics/ https://www.foxnews.com/category/science/wildhttps://www.foxnews.com/health/eee-easternhttps://www.foxnews.com/science/deadly-seal-virus-5. equine-encephalitis-massachusetts-man-killed spreading-melting-arctic-ice 6. https://www.foxnews.com/health/what-ishttps://www.foxnews.com/science/sandra-orangutannorovirus-3-things-to-know-about-the-vomitlegal-personhood-new-home-florida inducing-bug https://www.foxnews.com/health/pregnant-7. https://www.foxnews.com/science/ancientwomen-who-get-flu-shot-are-less-likely-to-beshipwrecks-stone-pyramid-anchors-discovered-greekisland hospitalized-from-virus-cdc-says https://www.foxnews.com/category/health/infec https://www.foxnews.com/science/earths-mantle-8. tious-disease/zika and-crust-are-in-a-fiery-battle-to-the-death-ofsupercontinents https://www.foxnews.com/health/flu-seasonhttps://www.foxnews.com/science/mammoth-traps-9. 2019-vaccine prehistoric-mexico

https://www.foxnews.com/science/mercury-transit-

2019-what-to-know

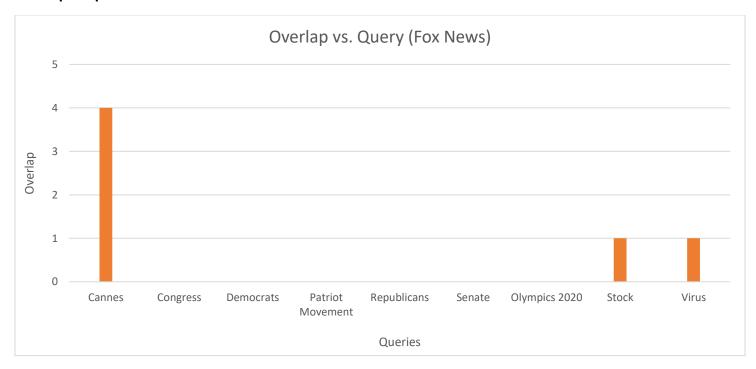
Overlap: 1

son-flu

10.

https://www.foxnews.com/health/british-mom-

Overlap Graph:



Conclusion: Hence we studied how the query results differ depending on the Solr default page ranking which uses only a combination if Vector Space Model and the Boolean Model and the external Page Ranking algorithm which takes into account the incoming links for each page to decide the Page Rank.