Step 1: Installation of VirtualBox and Ubuntu on Windows Machine
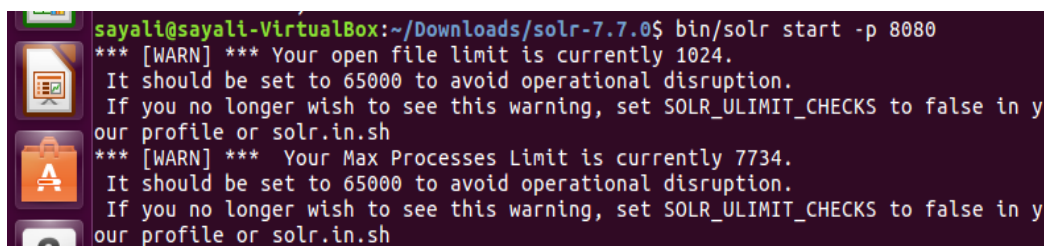
- Downloaded VirtualBox and Ubuntu from the links provided.
- Created a Virtual Machine in VirtualBox:
  - **New ->** Name: CSCI572-Linux, Type: Linux, Version: Ubuntu (64-bit); **Next ->** Memory size: 2048 MB; **Next ->** Create a virtual hard disk now; **Create ->** VDI (Virtual Disk Image); **Next ->** 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)

  - **$ 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)

- Verified if the Solr UI works well on port 8080.
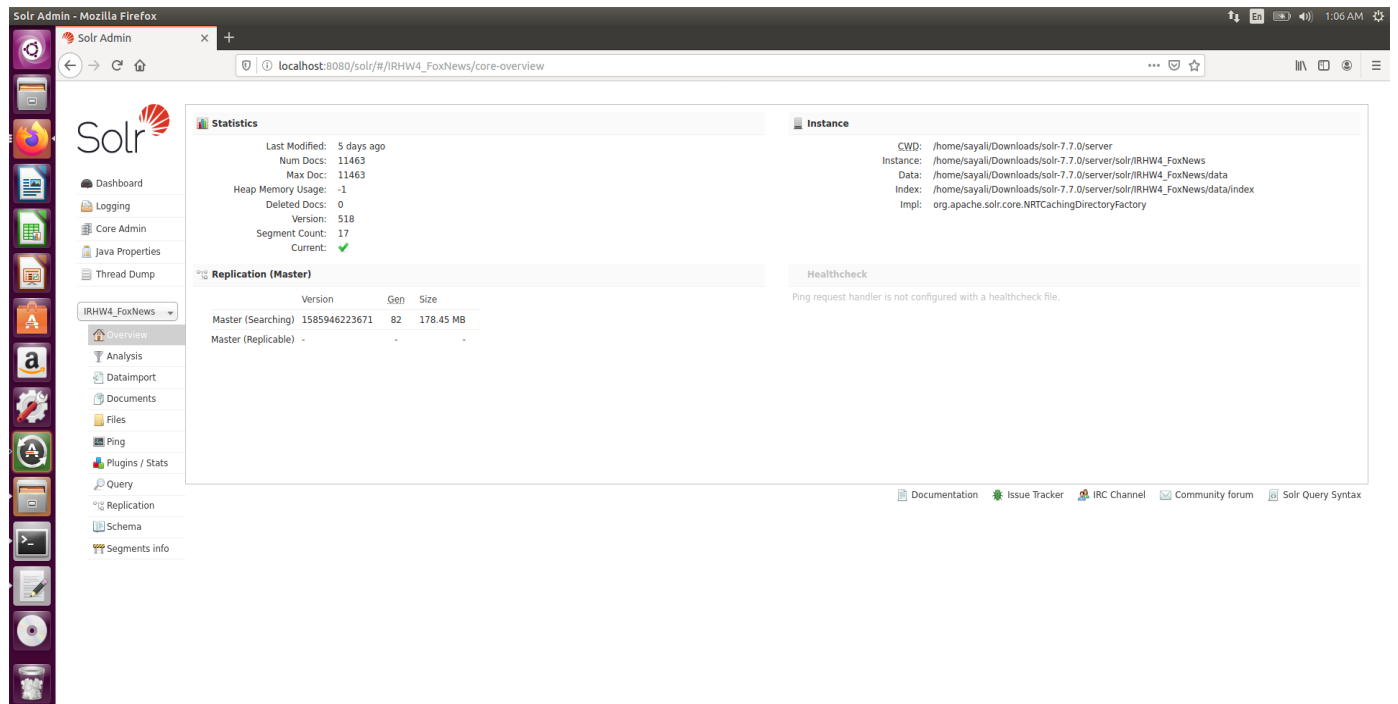
- Ran a basic query to verify if the indexing is done correctly.
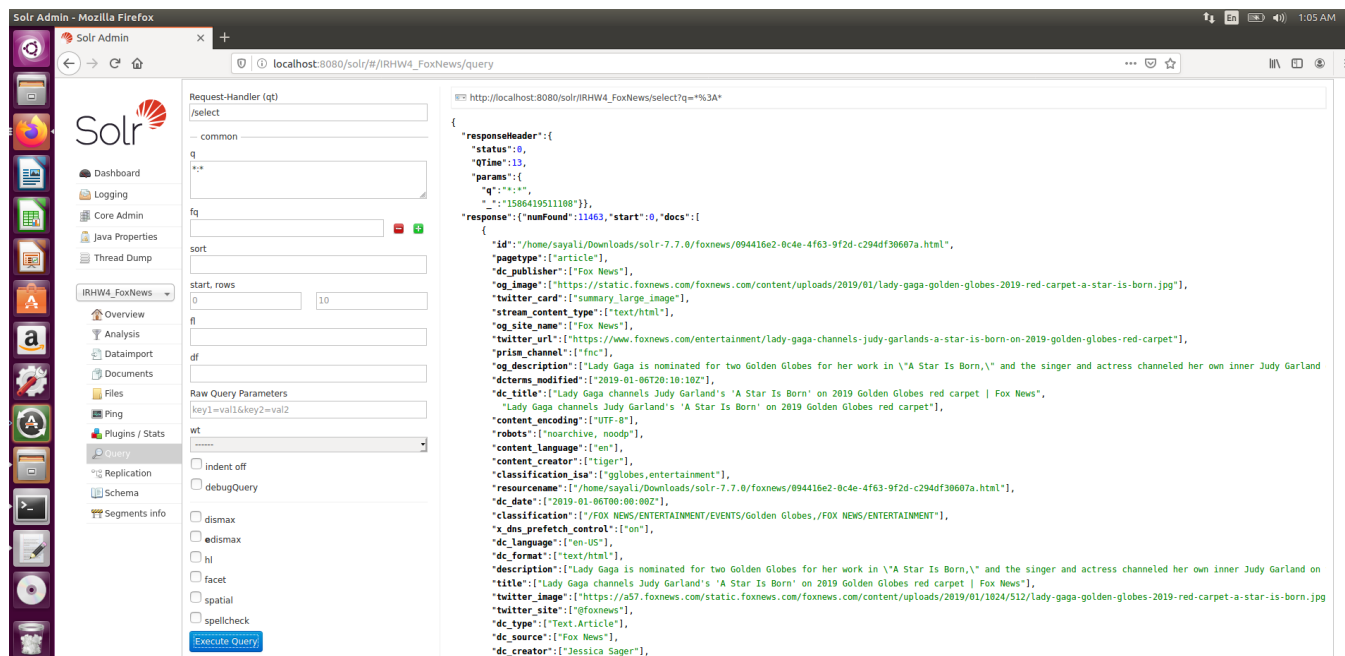
Step 4: Extracting URLs from HTML Files

- 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.
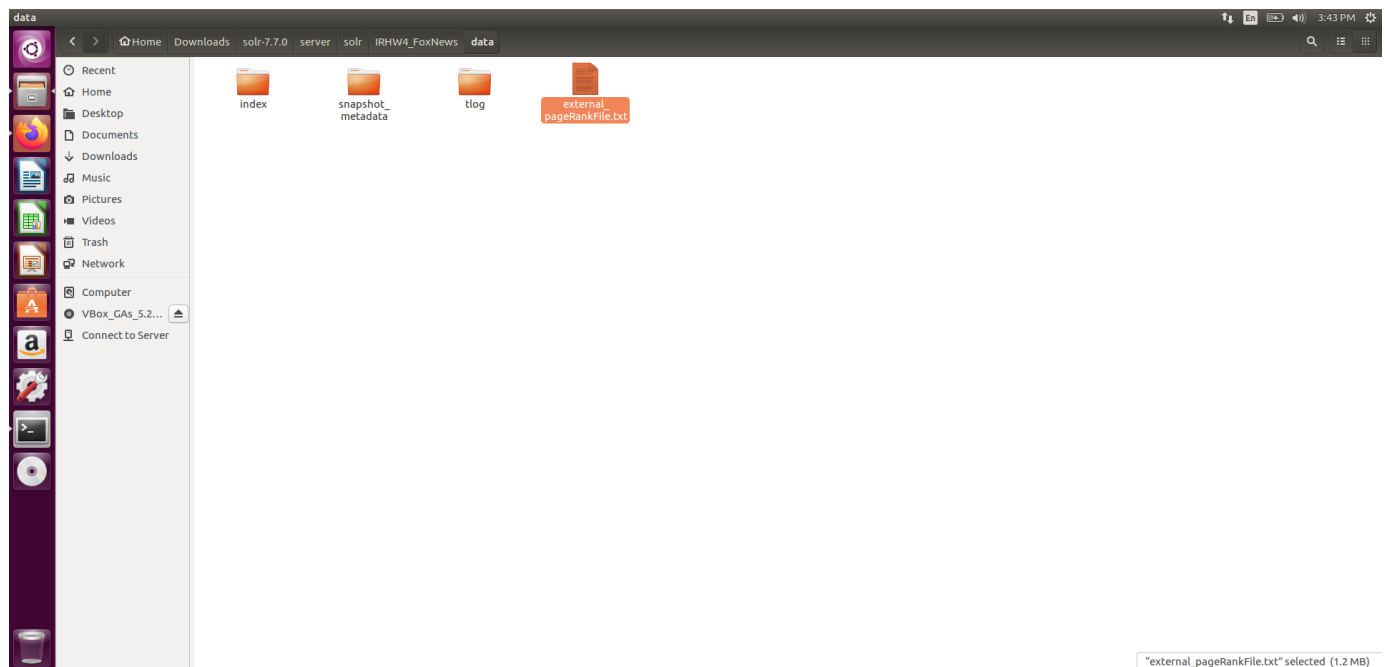
- Modified the **managed-schema.xml** file to add a field names pageRankFile which refers to the page rank score for each HTML file. Added the following lines to the file:
  <fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField"/>
  <field name="pageRankFile" type="external" 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:
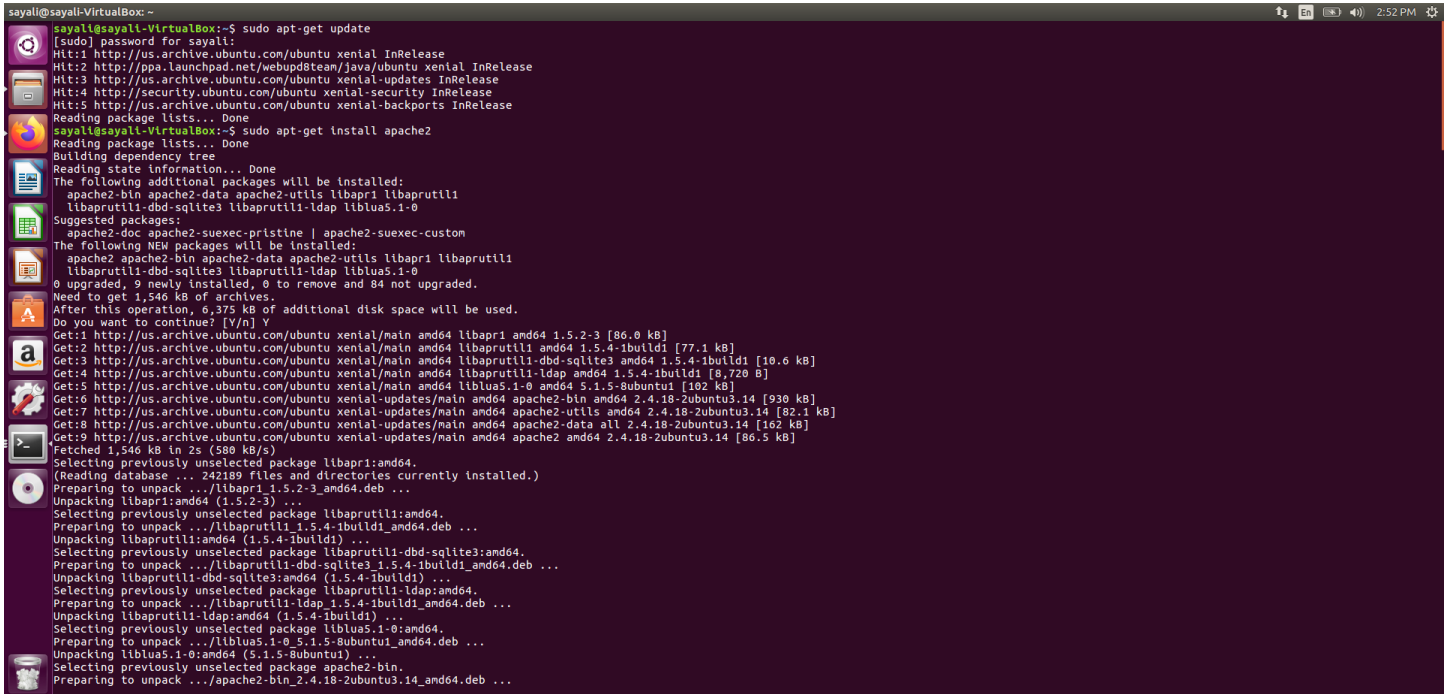
<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
<listener event="firstSearcher" class=" org.apache.solr.schema.ExternalFileFieldReloader"/>
- Reloaded the index through the Solr UI.

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**

- **$ 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.

- Below are the screenshots of the UI:

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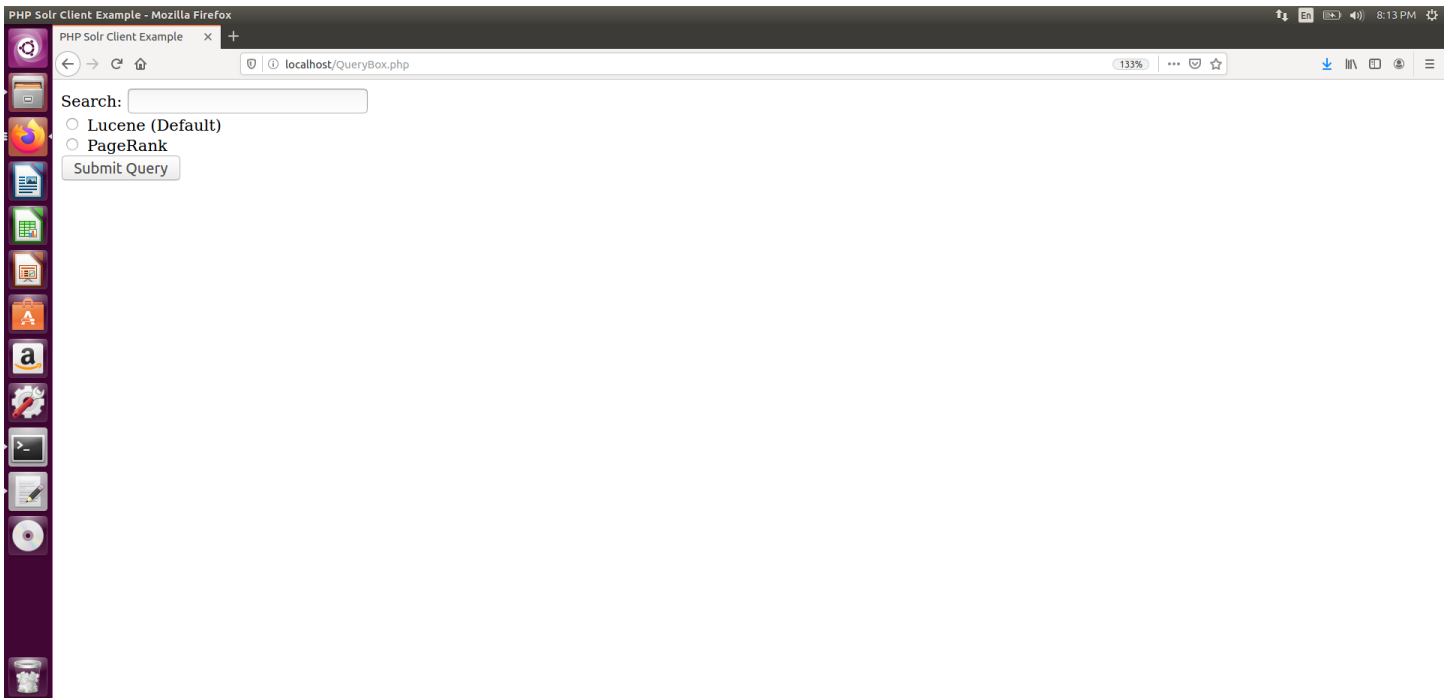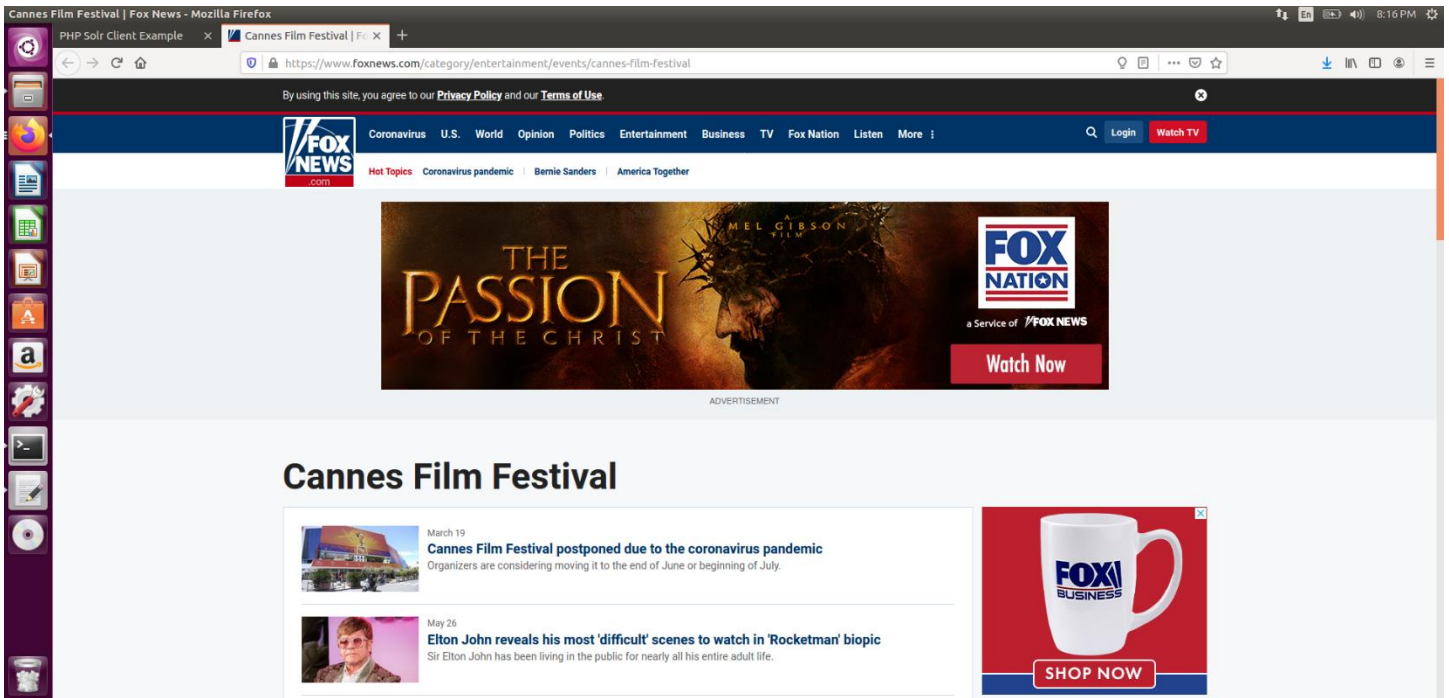
Query Page for the query 'Cannes' and 'Lucene' (Solr default) algorithm:

Actual Clickable Link:

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

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

http://localhost:8080/solr/IRHW4_FoxNews/select?fl=og_url&q=Cannes

{
  "responseHeader":{
    "status":0,
    "QTime":27,
    "params":{
      "q":"Cannes",
      "fl":"og_url",
      "_":"1586411326550"}},
  "response":{"numFound":30,"start":0,"docs":[
    {
      "og_url":["https://www.foxnews.com/category/entertainment/events/cannes-film-festival"]},
    {
      "og_url":["https://www.foxnews.com/entertainment/selena-gomez-bill-murray-whispered-cannes-film-festival-reveal"]},
    {
      "og_url":["https://www.foxnews.com/entertainment/selena-gomez-cannes-film-festival-red-carpet"]},
    {
      "og_url":["https://www.foxnews.com/entertainment/ryan-gosling-eva-mendes-is-definitely-not-anti-sweatpants"]},
    {
      "og_url":["https://www.foxnews.com/entertainment/quentin-tarantino-once-upon-a-time-in-hollywood-trailer-release"]},
    {
      "og_url":["https://www.foxnews.com/entertainment/woody-allen-retire-family-politics"]},
    {
      "og_url":["https://www.foxnews.com/entertainment/whitney-houston-was-sexually-abused-during-childhood-new-documentary-claims"]},
    {
      "og_url":["https://www.foxnews.com/entertainment/elton-john-rocketman-biopic-drugs-addiction-eating-disorder"]},
    {
      "og_url":["https://www.foxnews.com/opinion/a-joan-rivers-thanksgiving"]},
    {
      "og_url":["https://www.foxnews.com/entertainment/selena-gomez-new-dog-instagram"]}]
}}

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. N o.	Lucene (Default)	PageRank
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
Overlap: 4		

2. **Congress:**

Sr. No.	Lucene (Default)	PageRank
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-hearing-drove-me-nuts-there-were-empty-seats	https://www.foxnews.com/official-polls
4.	https://www.foxnews.com/opinion/rep-andy-biggs-congress-debt-ceiling-budget-deal	https://www.foxnews.com/category/politics/foreign-policy
5.	https://www.foxnews.com/opinion/gutfeld-congress-attacking-zuckerberg	https://www.foxnews.com/category/politics/executive
6.	https://www.foxnews.com/opinion/andy-biggs-congress-judges-immigration-laws	https://www.foxnews.com/category/tech/topics/security

7.	https://www.foxnews.com/faith-values/bible-school-drew-brees-day-congress	https://www.foxnews.com/category/world/conflicts
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
Overlap: 0		

3. **Democrats:**

Sr. No.	Lucene (Default)	PageRank
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/entertainment/music
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-show-trial-to-overthrow-trump-not-an-impeachment-inquiry	https://www.foxnews.com/category/us/personal-freedoms
8.	https://www.foxnews.com/opinion/donald-trump-jr-democrats-immigration-radical-ideology	https://www.foxnews.com/official-polls
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/world/scandals
10.	https://www.foxnews.com/opinion/david-bossie-trump-impeachment-vote-is-democratic-	https://www.foxnews.com/category/politics/foreign-policy

	declaration-of-war-republicans-must-declare-war-on-dems	

Overlap: 0

4. **Patriotic Movement:**

Sr. No.	Lucene (Default)	PageRank
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

Overlap: 0

5. **Republicans:**

Sr. No.	Lucene (Default)	PageRank
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-whistleblower-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/politics/senate
5.	https://www.foxnews.com/politics/trump-settles-score-with-republicans-who-didnt-embrace-him-in-midterms	https://www.foxnews.com/official-polls
6.	https://www.foxnews.com/politics/republicans-impeachment-hearings-lineup	https://www.foxnews.com/category/world/scandals
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/category/politics/elections
9.	https://www.foxnews.com/opinion/matt-gorman-kavanaugh-republicans	https://www.foxnews.com/category/world/conflicts
10.	https://www.foxnews.com/opinion/newt-gingrich-house-republicans-2020-retirements	https://www.foxnews.com/politics.html

Overlap: 0

6. **Senate:**

Sr. No.	Lucene (Default)	PageRank
1.	https://www.foxnews.com/politics/trump-senate-trial-impeachment-inquiry-reporters-notebook	https://www.foxnews.com/shows
2.	https://www.foxnews.com/category/politics/2020-senate-races	https://www.foxnews.com/opinion
3.	https://www.foxnews.com/media/jeff-sessions-former-embattled-trump-attorney-general-announces-bid-for-old-us-senate-seat	https://www.foxnews.com/entertainment
4.	https://www.foxnews.com/opinion/newt-gingrich-left-viciously-attacks-mcconnell-because-hes-successful-senate-leader-and-trump-ally	https://www.foxnews.com/world
5.	https://www.foxnews.com/faith-values/pastor-andrew-brunson-senate-trump-prayer	https://www.foxnews.com/health
6.	https://www.foxnews.com/politics/jeff-sessions-alabama-senate-campaign-expected	https://www.foxnews.com/travel
7.	https://www.foxnews.com/politics/trump-sessions-against-senate-run-alabama	https://www.foxnews.com/us
8.	https://www.foxnews.com/category/politics/elections/senate	https://www.foxnews.com/science
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/closed-captioning

Overlap: 0

7. **Olympics 2020:**

Sr. No.	Lucene (Default)	PageRank
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/tech
5.	https://www.foxnews.com/person/c/barnini-chakraborty	https://www.foxnews.com/politics
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-morgan	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/us/personal-freedoms
10.	https://www.foxnews.com/sports	https://www.foxnews.com/category/tech/topics/video-games
Overlap: 0		

8. **Stock**

| Sr. No.	Lucene (Default)	PageRank
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/person/sheryl-sandberg	https://www.foxnews.com/media/chris-wallace-elizabeth-warren-medicare-for-all-plan
6.	https://www.foxnews.com/category/tech/companies/microsoft	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://olvww.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
Overlap: 1		

9. Virus

Sr. No.	Lucene (Default)	PageRank
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
Overlap: 1		

Overlap Graph:

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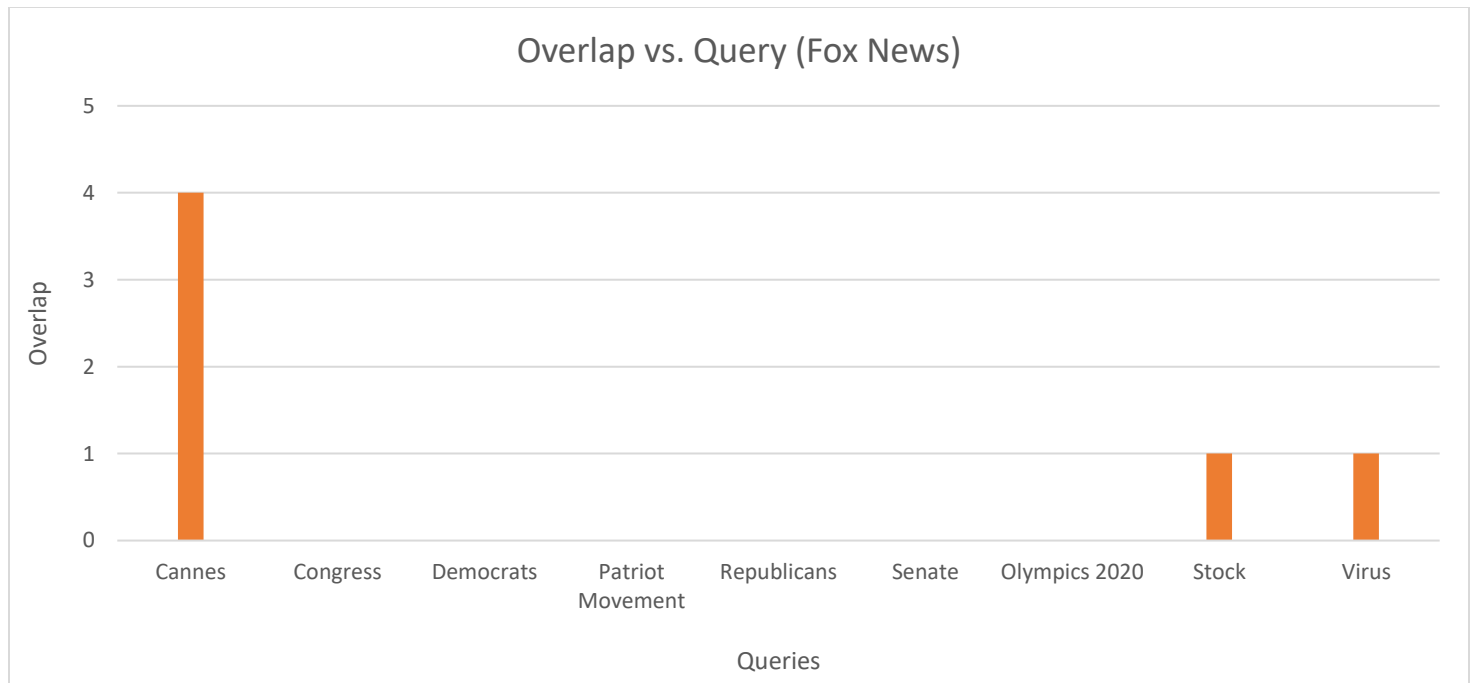
Conclusion: Hence we studied how the query results differ depending on the Solr default page ranking which uses only a combination if Vector Space Model and the Boolean Model and the external Page Ranking algorithm which takes into account the incoming links for each page to decide the Page Rank.