**CSCI 572 Assignment 4: Indexing the Web Using Solr**

**Name: Sumit Parwal**

**USC ID: 5174593050**

**Website Indexed: The Guardian News**

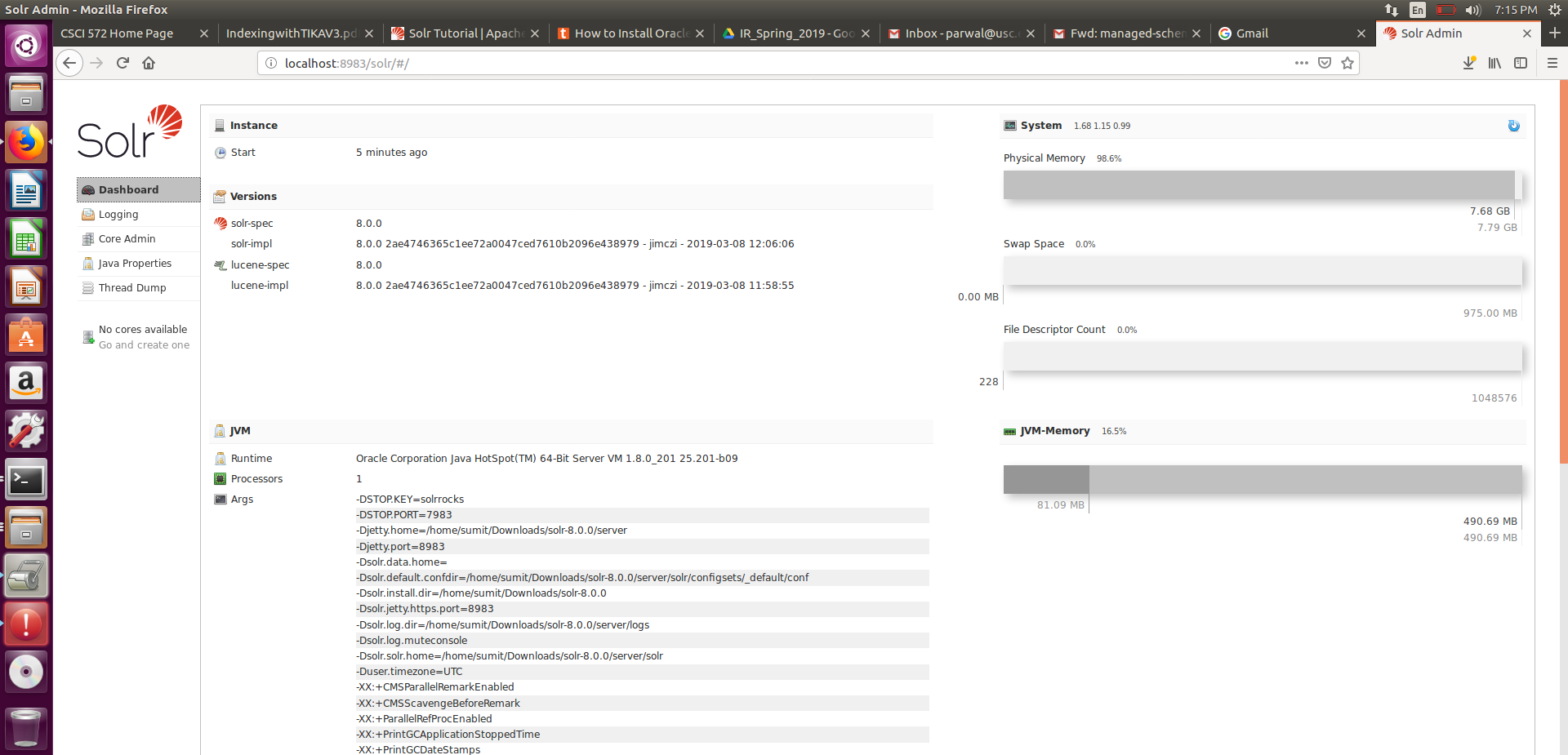
**Step 1: VirtualBox Installation**

* I Downloaded the version of VirtualBox by following the instructions given in the Assignment PDF for your machine under VirtualBox Platform package. Then Install it on the package “VirtualBOx 5.2.26.

**Step 2: Solr Installation**

* I Downloaded the Solr binary zip file on my macbook and unzipped it by the instructions given in the Assignment PDF.
* According to the instructions given in the Assignment PDF Indexing with TikaV3.pdf. Started the Solr using bin/solr start.
* Then I created a core myexample1 in Solr by using bin/solr create -c myexample1. In the managed schema file of the core I un-commented what was suggested in PDF.
* I indexed the The Guardian News html pages that are downloaded from the google drive.
* This indexing on the solr that can be viewed on the port 8983 by tying

localhost:8983/solr/# on the web browser.



**Step 3: Indexing the Web-Page**

* Then, I made the changes in the managed schema file of the core as suggested in PDF.
* I downloaded the “Guardian News” html pages from the google drive.
* I indexed the Guardian News html pages by following the Instructions in PDF

The screenshot below shows the final output obtained after posting the html files to Solr index:

**A picture containing table

Description automatically generated**

**Step 4: Extracting Edge List using JAVA and PageRank Algorithm**

* I downloaded the Jsoup.jar library using the link <https://jsoup.org/download> given in the assignment PDF
* I created a java project and imported the jsoup jar as an external jar in my java project.
* I created a urls\_Extractor.java class and added code to extract the links from all the html files in the downloaded guardian news data and dumped all the urls into **edgeList.txt**.

A picture containing window

Description automatically generated

* Firstly I installed pip using terminal then I used pip install network to install Networkx
* Then I used Networkx graph in the pagerank algorithm that returns the pages with the corresponding page rank score. After running the python code using Networkx and following parameters:

alpha=0.85

personalization=None

max\_iter=30

tol=1e-06

nstart=None

weight='weight'

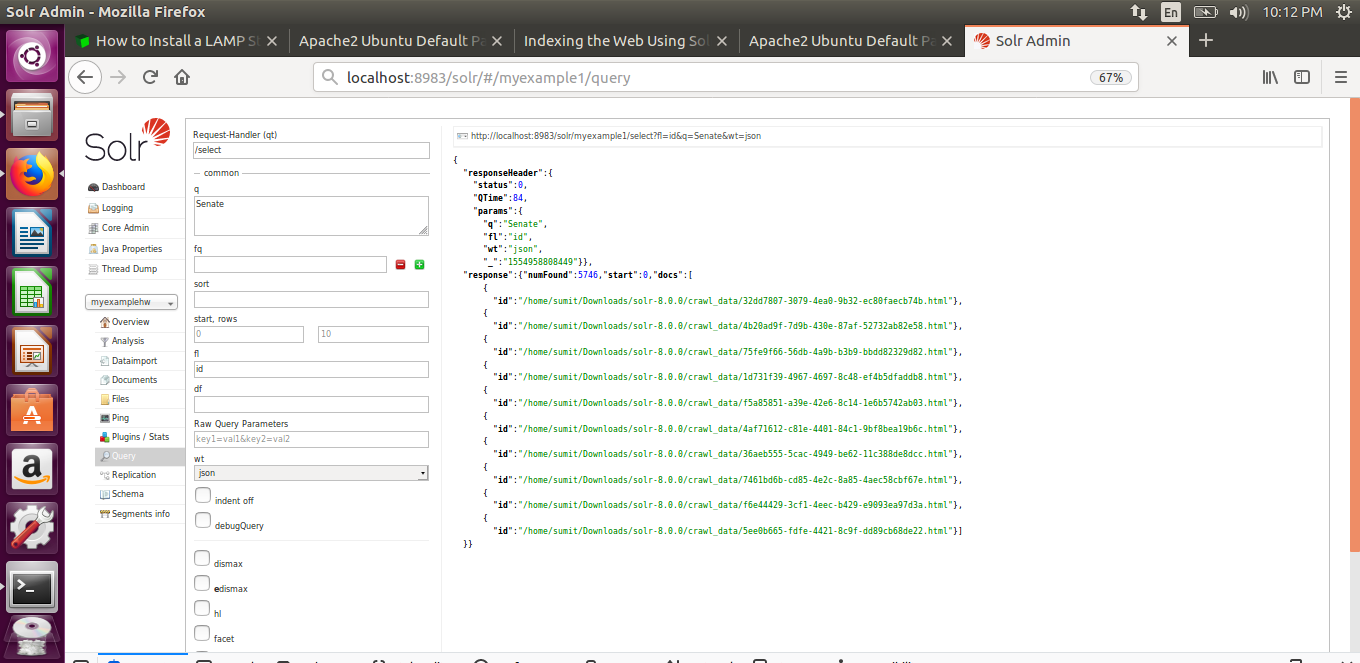
dangling=None

generated a external\_PageRankFile.txt.

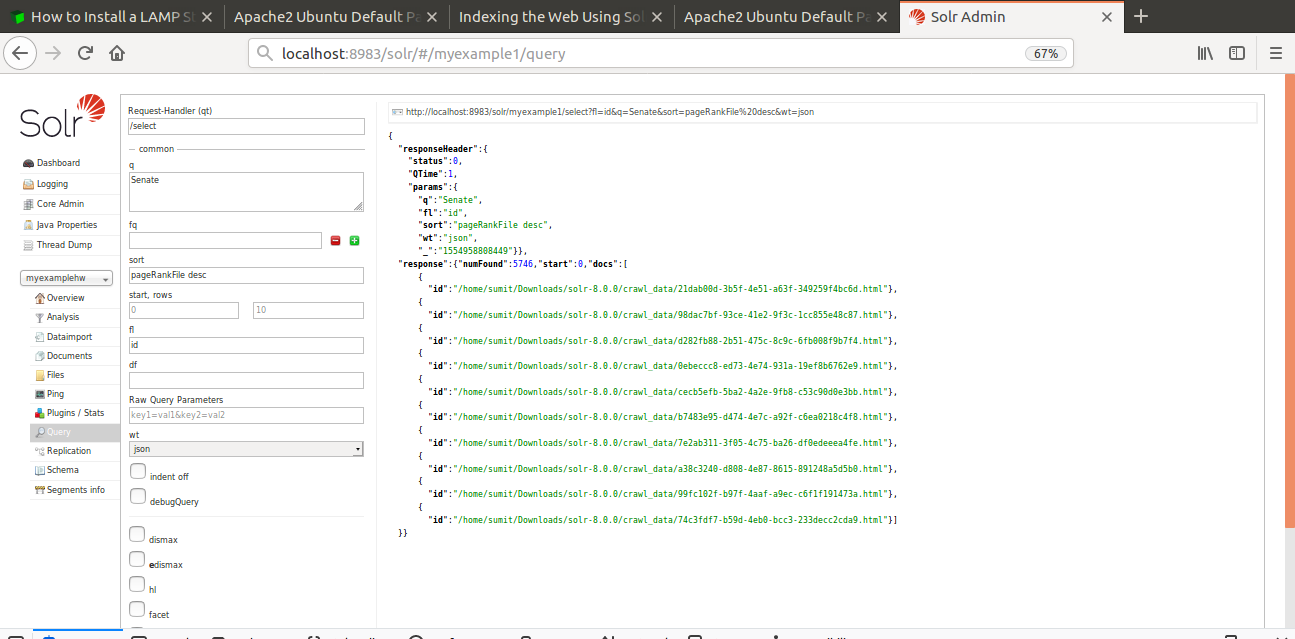
* This txt file contains the page rank score of the corresponding web pages. I copied this file in the data folder of the core myexample1.
* In the conf folder of the core made changes in the managed schema and solrconfig.xml file as given in the PDF. I went to the solr dashboard and reload my core after making all changes.

# Step 5: Verifying Results on Solr

* When I executed the “Senate” query on Solr admin by putting fl=id and wt=json I Got the 10 results in the solr window.



* Similarly, When I put sort=pageRankFile desc then I got the 10 sorted result in descending order on the of their PageRank score.



* It is clearly Visible that the 10 results obtained are different. So, our setup is correct and Lucene and Page Rank algorithms are properly configured.

### Question: Explanation regarding why some pages have higher PageRank values than others?

**Answer:** Page Rank Basically works by counting the number of incoming and outgoing links to a page and quality of those pages to get how much important are those pages. Some pages have higher PageRank values than others as there are more number of pages pointing to that page. PageRank score is basically a vote by other pages on web that how important a page is. A link to page is counted as a vote of support. In this assignment, we created the EdgeList file and with that EdgeList a graph is created to calculate the indegree of a node/link. If a link has a lot of links pointing to it that means the indegree or PageRank score of that link will be high.When the indegree also called Pager-Rank score of a link is calculated then the results are displayed in a descending order that is the top result has the highest PageRank score or indegree. This is basic idea PageRank algorithm use to calculate score of a particular and this is why some pages have higher PageRank values.

# Flow of Senate Query:

**Steps:**

1. Using HTML in ranking.php I came up with UI given in the figure below. The given page comes up when launch localhost on the browser. It contains the search bar to type the query along with two searching algorithm choice Lucene and Pagerank along with submit button to display results.

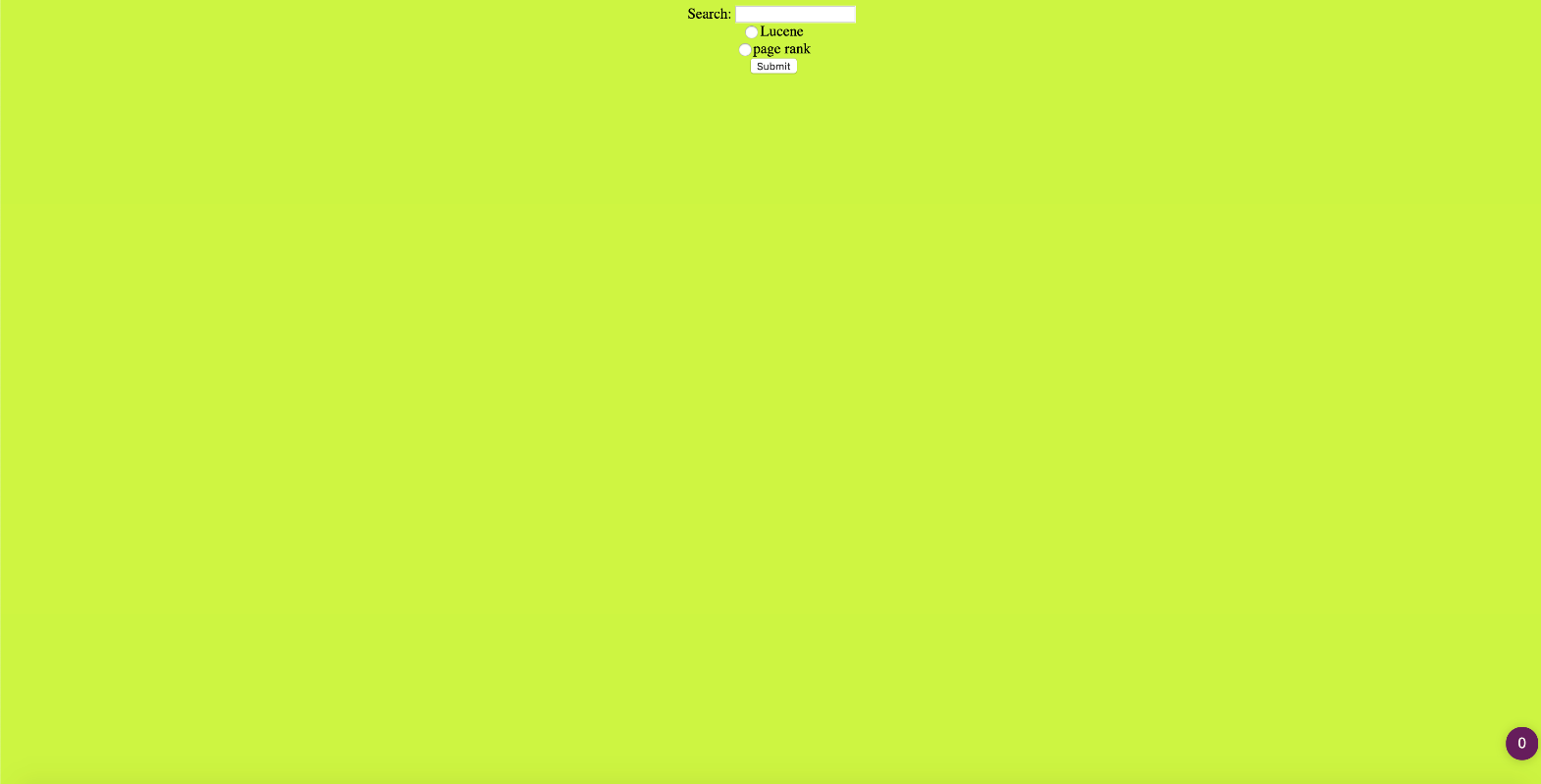


Fig- Screenshot of initial page where you enter the query in input box.

1. Then I entered a query “**Senate”** and selected the Lucene algorithm to apply on it and then submit.





Fig- Screenshot of page which shows the results for Lucene(Default).

1. Screenshot of the actual webpage that opens when clicking on one of the result’s link of Lucene(Default) Algorithm:

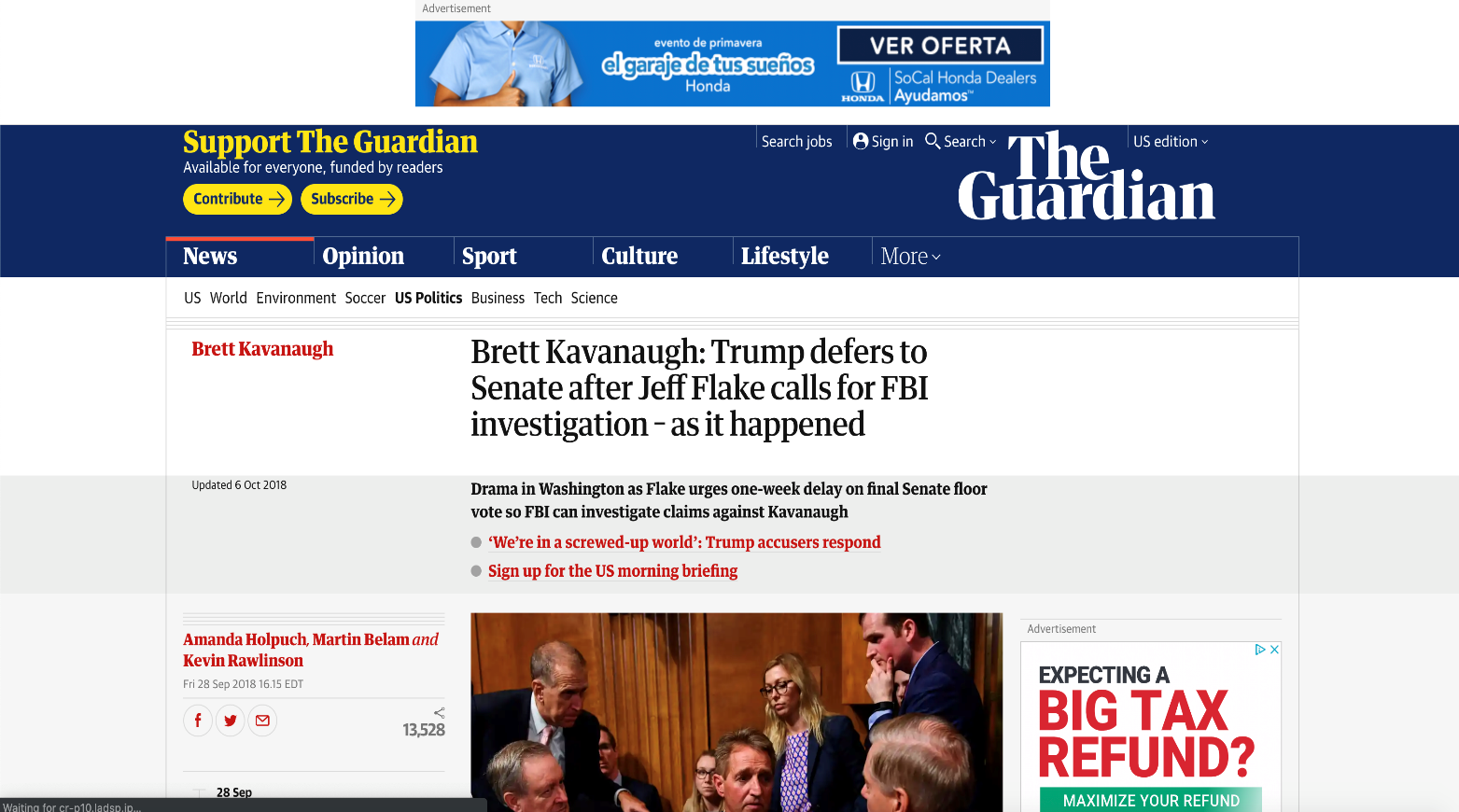


Fig- Screenshot of page which shows the results for Lucene(Default).

1. Then, I selected Pagerank Google’s algorithm and it displayed the following results for Senate query.





Fig- Screenshot of all the Results for Senate query Using PageRank Algorithm.

1. Screenshot of the actual webpage that opens when clicking on one of the result’s link of Page Rank Algorithm:

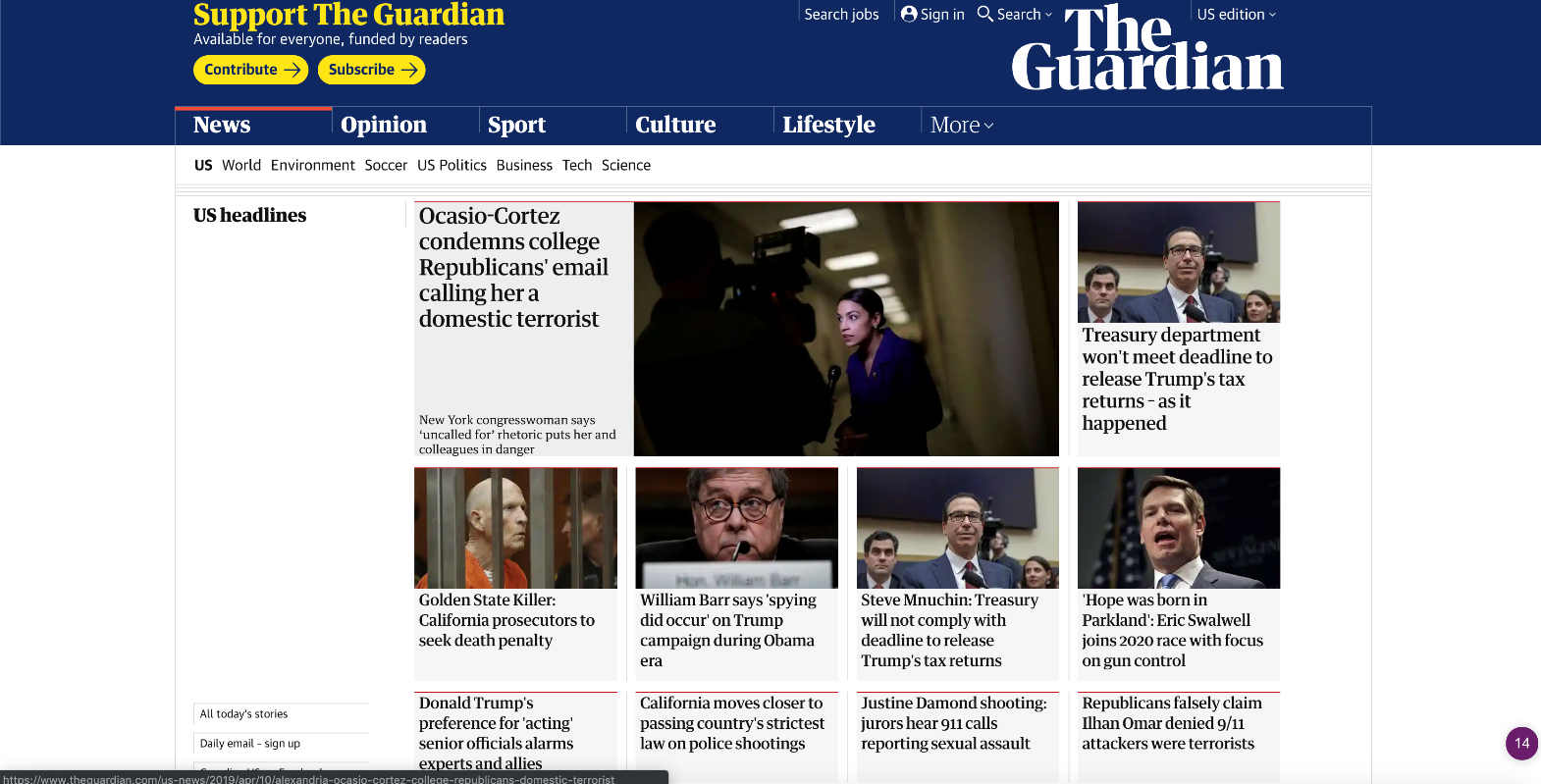


Fig- Screenshot of page which shows the results for PageRank Algorithm.

**Next pages contain table containing list of URLs of the top ten results produced by both ranking methods for each query**.

1. **Senate**

|  |  |
| --- | --- |
| **Lucene** | **PageRank** |
| <http://www.theguardian.com/us-news/live/2018/sep/28/kavanaugh-senate-committee-vote-live> | <http://www.theguardian.com/us-news> |
| <http://www.theguardian.com/us-news/live/2018/sep/28/kavanaugh-senate-committee-vote-live> | <http://www.theguardian.com/us-news/us-politics> |
| <http://www.theguardian.com/us-news/live/2014/dec/09/cia-torture-report-released-senate> | <http://www.theguardian.com/us> |
| <http://www.theguardian.com/us-news/live/2018/sep/28/kavanaugh-senate-committee-vote-live> | <http://www.theguardian.com/us-news/donaldtrump> |
| <http://www.theguardian.com/us-news/live/2018/sep/28/kavanaugh-senate-committee-vote-live> | <http://www.theguardian.com/us-news/beto-o-rourke> |
| <http://www.theguardian.com/us-news/live/2018/sep/28/kavanaugh-senate-committee-vote-live> | <http://www.theguardian.com/us-news/ng-interactive/2019/jan/17/2020-election-candidates-list-democrats-primaries-trump-bernie-sanders-joe-biden-elizabeth-warren-kamala-harris-beto-o-rourke-running> |
| <http://www.theguardian.com/us-news/live/2018/sep/28/kavanaugh-senate-committee-vote-live> | <http://www.theguardian.com/us-news/2019/mar/14/beto-orourke-announces-2020-run-for-presidency> |
| <http://www.theguardian.com/us-news/live/2018/sep/28/kavanaugh-senate-committee-vote-live> | <http://www.theguardian.com/us-news/live/2019/mar/14/trump-news-today-latest-live-updates-beto-o-rourke-boeing-flights-us-politics-ireland> |
| <http://www.theguardian.com/us-news/live/2018/sep/28/kavanaugh-senate-committee-vote-live> | <http://www.theguardian.com/us-news/2019/mar/14/senate-national-emergency-declaration-vote-trump> |
| <http://www.theguardian.com/us-news/live/2018/sep/28/kavanaugh-senate-committee-vote-live> | <http://www.theguardian.com/us-news/video/2019/mar/14/who-is-2020-presidential-candidate-beto-orourke-video-profile> |

1. **Wall**

|  |  |
| --- | --- |
| **Lucene** | **PageRank** |
| <http://www.theguardian.com/us-news/live/2018/dec/21/trump-government-shutdown-border-wall-latest-live> | <http://www.theguardian.com/us-news> |
| <http://www.theguardian.com/us-news/live/2019/jan/10/trump-us-mexico-border-wall-shutdown-live-updates> | <http://www.theguardian.com/us-news/democrats> |
| <http://www.theguardian.com/us-news/live/2018/dec/21/trump-government-shutdown-border-wall-latest-live> | <http://www.theguardian.com/us/business> |
| <http://www.theguardian.com/us-news/live/2018/dec/21/trump-government-shutdown-border-wall-latest-live> | <http://www.theguardian.com/us/culture> |
| <http://www.theguardian.com/us-news/live/2019/jan/10/trump-us-mexico-border-wall-shutdown-live-updates> | <http://www.theguardian.com/us-news/donaldtrump> |
| <http://www.theguardian.com/us-news/live/2018/dec/21/trump-government-shutdown-border-wall-latest-live> | <http://www.theguardian.com/us-news/us-politics> |
| <http://www.theguardian.com/us-news/live/2019/jan/10/trump-us-mexico-border-wall-shutdown-live-updates> | <http://www.theguardian.com/us-news/beto-o-rourke> |
| <http://www.theguardian.com/us-news/live/2019/jan/10/trump-us-mexico-border-wall-shutdown-live-updates> | <http://www.theguardian.com/us-news/ng-interactive/2019/jan/17/2020-election-candidates-list-democrats-primaries-trump-bernie-sanders-joe-biden-elizabeth-warren-kamala-harris-beto-o-rourke-running> |
| <http://www.theguardian.com/us-news/live/2019/jan/10/trump-us-mexico-border-wall-shutdown-live-updates> | <http://www.theguardian.com/us-news/2019/mar/14/beto-orourke-announces-2020-run-for-presidency> |
| <http://www.theguardian.com/us-news/live/2019/jan/10/trump-us-mexico-border-wall-shutdown-live-updates> | <http://www.theguardian.com/us-news/live/2019/mar/14/trump-news-today-latest-live-updates-beto-o-rourke-boeing-flights-us-politics-ireland> |

1. **Venezuela**

|  |  |
| --- | --- |
| **Lucene** | **PageRank** |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/us-politics> |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/donaldtrump> |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/2019/mar/14/beto-orourke-announces-2020-run-for-presidency> |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/us-elections-2016> |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/video/2019/feb/13/ilhan-omar-elliot-abrams-venezuela-video> |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/florida> |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/donaldtrump> |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/us-foreign-policy> |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/mike-pence> |
| <http://www.theguardian.com/us-news/live/2019/jan/30/trump-politics-venezuela-congress-shutdown-latest-news-updates> | <http://www.theguardian.com/us-news/live/2019/mar/08/donald-trump-latest-news-live-alabama-paul-manafort-russia-hr1-bill-vote> |

1. **Democrats**

|  |  |
| --- | --- |
| **Lucene** | **PageRank** |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us-news> |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us-news/us-politics> |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us> |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us/commentisfree> |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us-news/donaldtrump> |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us-news/democrats> |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us-news/beto-o-rourke> |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us-news/ng-interactive/2019/jan/17/2020-election-candidates-list-democrats-primaries-trump-bernie-sanders-joe-biden-elizabeth-warren-kamala-harris-beto-o-rourke-running> |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us-news/2019/mar/14/beto-orourke-announces-2020-run-for-presidency> |
| <http://www.theguardian.com/us-news/live/2018/nov/06/us-midterms-elections-2018-latest-live-polls-news-updates-donald-trump-republicans-democrats> | <http://www.theguardian.com/us-news/live/2019/mar/14/trump-news-today-latest-live-updates-beto-o-rourke-boeing-flights-us-politics-ireland> |

1. **Channel**

|  |  |
| --- | --- |
| **Lucene** | **PageRank** |
| <http://www.theguardian.com/us-news/2018/aug/17/stormy-daniels-blames-celebrity-big-brother-withdrawal-on-custody-issue> | <http://www.theguardian.com/us/commentisfree> |
| <http://www.theguardian.com/us-news/2017/may/28/trump-white-house-jared-kushner-russia> | <http://www.theguardian.com/us/tv-and-radio> |
| <http://www.theguardian.com/us-news/2017/mar/05/roger-stone-trump-adviser-julian-assange> | <http://www.theguardian.com/us-news/ng-interactive/2018/dec/05/robert-mueller-who-charged-list-trump-russia-inquiry-latest-indictments-flynn-manafort-cohen-what-happens-next> |
| <http://www.theguardian.com/us-news/2017/apr/07/us-russia-relations-syria-military-strikes-putin-trump> | <http://www.theguardian.com/us-news/2019/feb/01/cory-booker-announces-run-for-presidency-in-2020> |
| <http://www.theguardian.com/us-news/2017/may/29/vladimir-putin-is-bigger-threat-than-isis-john-mccain-says> | <http://www.theguardian.com/us-news/live/2019/mar/01/trump-kushner-news-latest-live-updates-us-politics-today-democrats-republicans> |
| <http://www.theguardian.com/us-news/ronpaul> | <http://www.theguardian.com/us-news/2019/mar/04/fox-news-stormy-daniels-rupert-murdoch-trump-win> |
| <http://www.theguardian.com/us-news/california> | <http://www.theguardian.com/us-news/series/the-campaign-minute-2016> |
| <http://www.theguardian.com/us-news/us-weather> | <http://www.theguardian.com/us-news/2019/feb/12/parkland-shooting-survivor-gun-control-activism> |
| <http://www.theguardian.com/us-news/los-angeles> | <http://www.theguardian.com/us-news/live/2019/jan/09/rod-rosenstein-donald-trump-latest-news-live> |
| <http://www.theguardian.com/us-news/nra> | <http://www.theguardian.com/us-news/commentisfree/2018/mar/23/parkland-students-manifesto-americas-gun-laws> |

1. **Oscar 2019**

|  |  |
| --- | --- |
| **Lucene** | **PageRank** |
| <http://www.theguardian.com/us-news/2017/jan/17/barack-obama-commutes-sentence-oscar-lopez-rivera-puerto-rico-activist> | <http://www.theguardian.com/us-news> |
| <http://www.theguardian.com/us-news/2016/oct/01/richard-trentlage-oscar-mayer-wiener-song-dies-hot-dogs> |  |
| <http://www.theguardian.com/us/film> | <http://www.theguardian.com/us/environment> |
| <http://www.theguardian.com/us-news/dick-cheney> | <http://www.theguardian.com/us> |
| <http://www.theguardian.com/us-news/2018/oct/15/tolerancia-cero-familias-separadas-karnes-texas-detencion-infinita> | <http://www.theguardian.com/money> |
| <http://www.theguardian.com/us-news/2018/oct/14/zero-tolerance-family-separation-indefinite-detention-karnes-texas> | <http://www.theguardian.com/us/commentisfree> |
| <http://www.theguardian.com/us-news/california> | <http://www.theguardian.com/us/travel> |
| <http://www.theguardian.com/us-news/nate-silver> | <http://www.theguardian.com/us/business> |
| <http://www.theguardian.com/us-news/black-lives-matter-movement> |  |
| <http://www.theguardian.com/us-news/2019/feb/25/us-briefing-oscars-shocks-pro-voter-bill-and-venezuela-meeting> | <http://www.theguardian.com/us/lifeandstyle> |

1. **Patriot Movement**

|  |  |
| --- | --- |
| **Lucene** | **PageRank** |
| <http://www.theguardian.com/us-news/2016/nov/07/oregon-county-sheriff-glenn-palmer-todd-mckinley-reelection> | <http://www.theguardian.com/us/environment> |
| <http://www.theguardian.com/us-news/oregon-militia-standoff> | <http://www.theguardian.com/us> |
| <http://www.theguardian.com/us-news/2018/nov/18/arrests-and-clashes-as-rival-right-and-left-rallies-descend-on-portland-again> | <http://www.theguardian.com/us-news/ng-interactive/2019/jan/17/2020-election-candidates-list-democrats-primaries-trump-bernie-sanders-joe-biden-elizabeth-warren-kamala-harris-beto-o-rourke-running> |
| <http://www.theguardian.com/us-news/portland> | <http://www.theguardian.com/us/lifeandstyle> |
| <http://www.theguardian.com/us-news/2018/aug/03/portland-patriot-prayer-rally-fears-violence-charlottesville> | <http://www.theguardian.com/us/commentisfree> |
| <http://www.theguardian.com/us-news/2018/aug/19/rightwing-protest-seattle-march-against-alleged-leftwing-violence> | <http://www.theguardian.com/us-news/2019/mar/14/beto-orourke-announces-2020-run-for-presidency> |
| <http://www.theguardian.com/us-news/oregon> | <http://www.theguardian.com/us-news/2019/mar/14/trump-brexit-theresa-may-uk-criticism-talks-latest-news> |
| <http://www.theguardian.com/us-news/2018/aug/04/patriot-prayer-to-carry-guns-at-portland-rally-as-fears-of-violence-rise> | <http://www.theguardian.com/us-news/2019/mar/13/rich-kids-top-college-admissions> |
| <http://www.theguardian.com/us-news/2018/aug/06/portland-protest-rightwing-patriot-prayer-police> | <http://www.theguardian.com/us-news/2019/mar/12/monumental-step-backwards-the-1bn-gas-pipeline-project-dividing-new-york> |
| <http://www.theguardian.com/us-news/2015/may/21/fbi-patriot-act-doj-report> | <http://www.theguardian.com/us-news/2019/mar/12/climate-change-student-activists-environment> |

1. **Republicans**

|  |  |
| --- | --- |
| **Lucene** | **PageRank** |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us-news> |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us-news/us-politics> |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us/commentisfree> |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us-news/donaldtrump> |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us-news/democrats> |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us-news/beto-o-rourke> |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us-news/ng-interactive/2019/jan/17/2020-election-candidates-list-democrats-primaries-trump-bernie-sanders-joe-biden-elizabeth-warren-kamala-harris-beto-o-rourke-running> |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us-news/2019/mar/14/beto-orourke-announces-2020-run-for-presidency> |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us-news/live/2019/mar/14/trump-news-today-latest-live-updates-beto-o-rourke-boeing-flights-us-politics-ireland> |
| <http://www.theguardian.com/us-news/live/2018/nov/07/midterm-elections-democrats-trump-republicans-house-latest-live> | <http://www.theguardian.com/us-news/2019/mar/14/senate-national-emergency-declaration-vote-trump> |

**Overlap Table**

I did not Find any Overlaps in the Given Eight queries for the Guardian News Data.

This Implies that there is no two same links for that we got From Lucene and Page rank Algorithms.

|  |  |
| --- | --- |
| **Query** | **Overlap** |
| Venezuela | 0 |
| Senate | 0 |
| Democrats | 0 |
| Republicans | 0 |
| Patriot Movement | 0 |
| Oscar 2019 | 0 |
| Channel | 0 |
| Wall | 0 |

**An overlap graph showing the amount of overlap between the two ranking strategies for all queries.**