CSCI 572 Assignment 4 – Indexing the Web using Solr

Web Pages Indexed – New York Post

Name – Karthik Natarajan

USC ID – 5247024172

STEP 1 – Downloaded Virtual Box version 5.2.0. Downloaded 64 bit version of Ubuntu Linux 16.04.5 LTS. Installed Virtual Box and started a new application for Linux and Ubuntu Version 64 bit. Used download Ubuntu Linux 16.04.5 LTS as virtual hard disk file. Followed the rest of the instructions for installing Ubuntu in the virtual machine. Set up a username and password for the virtual machine. However constant issues cropped up while adding Guest Additions to Ubuntu. After several fruitless efforts, continued with the rest of the assignment without adding the Guest Additions. There was no sharing of folders across the guest and host OS.

STEP 2 – After setting up Ubuntu, proceeded with installing Solr. Solr version 7.5.0 was downloaded and unpacked. Followed the instructions in the tutorial link provided to Index Techproducts Example Data. Once Solr was set up and working properly proceeded with the rest of the exercise. Once solr was installed, I just run it with the command : **$bin/solr start.** Solr's UI can be viewed by going to the address http://localhost:8983/solr in the browser.

STEP 3 – Learnt to use Solr to index a website. Started solr with command : **bin/solr start.** Created a new core called myexample using the command: **bin/solr create -c myexample.** After setting up myexample, I navigated to the conf folder of the newly created core and edited the 'managed-schema' file as per the instructions. I then proceeded to download the HTML pages of NY Post as per assignment instructions and indexed the HTML pages of NY Post using the command :

$ bin/post -c myexample -filetypes html /home/karnat1802/Downloads/nypost/

A total of 15403 files were indexed by Solr after issuing the above command.

STEP 4 – After having indexed the HTML pages the next step of the process was to go ahead with developing the UI for the assignment using PHP. After watching a few tutorials on Google was easily able to setup Apache Server and Php in Ubuntu to start developing the UI. Wrote a test.php file in /var/www/html/ folder with a simple print statement to see if php code was being rendered correctly by the Apache Server. On visiting http://localhost/test.php was able to see the php code working properly.

STEP 5 – First step in developing the UI was to work with the starter php code given in assignment description. As per assignment instructions download the solr-php-client from the given Github link in the assignment description and cloned the repository to /var/www/html/ folder. Modified the code to correctly point to the service.php file in the solr-php-client. Next, I proceeded with getting the Lucene results to display properly. Modified the starter php code to fetch only the relevant fields (title, og_url, og_description, id). Later on modified the php code to programmatically fetch the urls from the URLtoHTML.csv file.

STEP 6 – Once the Lucene results were being fetched and displayed correctly the next step was the get the PageRank results. First step towards doing that was the fetch the links from the web pages, and generate the edges file called edgesList.txt. Jsoup package was used in eclipse environment. Screenshot of the Java code along with the edgesList.txt is given below.

```java
*ExtractingLinks.java ⊠

 2⊕ import java.io.BufferedReader;⃞
14
15  public class ExtractingLinks {
16⊖      public static void main(String[] args) throws Exception {
17          String dirPath = "C:\\Users\\Karthik\\Desktop\\nypost\\";
18          String csvFilePath = "C:\\Users\\Karthik\\Desktop\\UrlToHtml_nypost.csv";
19          Map<String, String> IdUrlMap = new HashMap<String, String>();
20          FileReader in = new FileReader(csvFilePath);
21          BufferedReader br = new BufferedReader(in);
22          String line = null;
23          while ((line = br.readLine()) != null) {
24              String id = line.split(",")[0];
25              String url = line.split(",")[1];
26              IdUrlMap.put(url, id);
27          }
28          br.close();
29          String outputFileName = "edgesList.txt";
30          String outputFilepath = "C:\\Users\\Karthik\\Desktop\\" + outputFileName;
31          PrintWriter pw = new PrintWriter(new FileWriter(outputFilepath));
32
33          File directory = new File(dirPath);
34          File[] htmlFiles = directory.listFiles();
35          for (File file: htmlFiles) {
36              File filePath = new File(dirPath + file.getName());
37              Document doc = Jsoup.parse(filePath, "UTF-8", "https://nypost.com/");
38              Elements links = doc.select("a[href]");
39
40              for (Element link: links) {
41                  String outUrl = link.attr("abs:href").trim();
42                  if (IdUrlMap.containsKey(outUrl)) {
43                      String edge = file.getName() + " " + IdUrlMap.get(outUrl);
44                      pw.println(edge);
45                      pw.flush();
46                  }
47              }
48          }
49          pw.close();
50      }
```

```
pageRank.py ×    external_pageRankFile.txt ×    edgesList.txt ×

 1    00007082-e4f6-4f45-b701-4cc462344701.html 2f65e692-a9f9-4f32-9591-198ecc3a0d9b.html
 2    00007082-e4f6-4f45-b701-4cc462344701.html 72897ce7-e9b1-489f-810d-bba4f5261645.html
 3    00007082-e4f6-4f45-b701-4cc462344701.html 8790e35b-13e7-4277-abec-631c5ca93fd3.html
 4    00007082-e4f6-4f45-b701-4cc462344701.html 6dcc0483-03d1-4767-b2dd-d791046f94c1.html
 5    00007082-e4f6-4f45-b701-4cc462344701.html 1f12f8e0-2b05-4fdc-98f0-2473818a378c.html
 6    00007082-e4f6-4f45-b701-4cc462344701.html 3e6224e7-52b2-49bb-b35e-36e453227413.html
 7    00007082-e4f6-4f45-b701-4cc462344701.html 6bc5ab6f-4ca4-4588-a3a8-f9a6bb45f6ab.html
 8    00007082-e4f6-4f45-b701-4cc462344701.html 08328725-0507-4b37-a111-b1eda05fd1fc.html
 9    00007082-e4f6-4f45-b701-4cc462344701.html bbc54461-2472-49bf-8c43-4e1a383347f4.html
10    00007082-e4f6-4f45-b701-4cc462344701.html 40b8760d-53b6-4edb-a9b7-f60c3c13258d.html
11    00007082-e4f6-4f45-b701-4cc462344701.html db9f279a-359a-43e2-bcd5-ad75becb6606.html
12    00007082-e4f6-4f45-b701-4cc462344701.html 9e1975ef-08c7-45f3-8f77-120bff2a5b32.html
13    00007082-e4f6-4f45-b701-4cc462344701.html 88f9f9d1-491d-4df3-981d-863a8b1b485a.html
14    00007082-e4f6-4f45-b701-4cc462344701.html fc17eac6-ad31-4c0c-b079-3bebbc17fcb9.html
15    00007082-e4f6-4f45-b701-4cc462344701.html e8940ca0-ff7c-4925-aaab-47e9f260d0c1.html
```

STEP 7 – After the edgeList.txt file has been generated the next step of the process is to compute the PageRank score of the html pages using the Network X library in Python. The pageRank scores for each website was store in the exter_PageRankFile.txt document. The python code was developed using the sample code given in the assignment instructions. The screenshots for both are given below.

```
pageRank.py ×   external_pageRankFile.txt ×   edgesList.txt ×
1    /home/karnat1802/Downloads/solr-7.5.0/nypost/aac8da41-c223-47cb-a25d-4820284d09bd.html=1.15380329371e-05
2    /home/karnat1802/Downloads/solr-7.5.0/nypost/561a13b5-e67b-4005-80e0-959e91066cf7.html=2.12018626091e-05
3    /home/karnat1802/Downloads/solr-7.5.0/nypost/d5ac3d1a-ba3c-4588-b41e-7f1e3c13f057.html=2.09144623312e-05
4    /home/karnat1802/Downloads/solr-7.5.0/nypost/6e7c860f-fe11-44b0-985a-8aee866bb498.html=1.07218283123e-05
5    /home/karnat1802/Downloads/solr-7.5.0/nypost/2d7e4d3e-b9c7-4730-99a0-56d56487fb58.html=2.12155966896e-05
6    /home/karnat1802/Downloads/solr-7.5.0/nypost/dc8230a7-4e26-407c-9080-68cc15bfb8b3.html=2.09144623312e-05
7    /home/karnat1802/Downloads/solr-7.5.0/nypost/88a142d1-db9c-4e86-95c1-a3e5817e4129.html=2.12022790594e-05
8    /home/karnat1802/Downloads/solr-7.5.0/nypost/e4d5bf6e-02ac-43d9-8deb-20321b032049.html=9.73836265663e-06
9    /home/karnat1802/Downloads/solr-7.5.0/nypost/1fd81381-e705-4a3d-a9f7-0743d77aa435.html=2.10825096859e-05
10   /home/karnat1802/Downloads/solr-7.5.0/nypost/28efe2cb-2798-418e-8bb2-1f526c003962.html=9.73836265663e-06
11   /home/karnat1802/Downloads/solr-7.5.0/nypost/842be58d-dafd-437b-b9f4-0f36a12406ff.html=2.11650530761e-05
12   /home/karnat1802/Downloads/solr-7.5.0/nypost/94a8f94f-0a99-46d6-9788-460d539eb793.html=1.02703132466e-05
13   /home/karnat1802/Downloads/solr-7.5.0/nypost/d44b057d-ee2f-4f57-8c9a-f02510ca7935.html=1.10744634217e-05
14   /home/karnat1802/Downloads/solr-7.5.0/nypost/0904cde2-2e9c-4a0b-9bab-b61aca457f61.html=9.73836265663e-06
15   /home/karnat1802/Downloads/solr-7.5.0/nypost/8c1d1576-1b8f-446e-9d91-c4b5b942a906.html=9.91650355799e-06
16   /home/karnat1802/Downloads/solr-7.5.0/nypost/0e2bf793-fe8b-43e9-80db-4d9d0936be03.html=2.09144623312e-05
17   /home/karnat1802/Downloads/solr-7.5.0/nypost/9242bfc6-0df1-4a61-a90e-4c810c19f561.html=1.56042746479e-05
18   /home/karnat1802/Downloads/solr-7.5.0/nypost/ae4f75dc-027a-4b37-a374-743affae1f41.html=1.04247520424e-05
19   /home/karnat1802/Downloads/solr-7.5.0/nypost/7485fa7c-c7c1-41f8-99c2-05f3fc4f46c8.html=2.12022790594e-05
20   /home/karnat1802/Downloads/solr-7.5.0/nypost/f6ade268-41d3-4829-88f4-9ebc88532653.html=1.99198533186e-05
21   /home/karnat1802/Downloads/solr-7.5.0/nypost/ac062d75-5714-48f4-8443-2c47e1339bc9.html=2.12018626091e-05
22   /home/karnat1802/Downloads/solr-7.5.0/nypost/b02d11fc-24b7-4eab-917d-83905c542d1e.html=3.90044586747e-05
```

```python
pageRank.py ×   external_pageRankFile.txt ×   edgesList.txt ×
1    import networkx as nx
2
3    graph = nx.read_edgelist("C:/Users/Karthik/PycharmProjects/InformationRetrieval/edgesList.txt", create_using=nx.DiGraph())
4
5    pr = nx.pagerank(graph, alpha=0.85, personalization=None, max_iter=100, tol=1.0e-6, nstart=None, weight='weight', dangling=None)
6
7    outputFile = "C:/Users/Karthik/PycharmProjects/InformationRetrieval/external_pageRankFile.txt"
8    f = open(outputFile, "w")
9
10   idPrefix = "C:/Users/Karthik/Desktop/nypost/"
11   for id in pr:
12       f.write(idPrefix + id + "=" + str(pr[id]) + "\n")
13
14   f.close()
15
```

STEP 8 – Next step of the process is external_pageRankFile.txt in the data folder of the core. Then I went on to add the fields managed-schema and event listeners in the solrconfig.xml. Once these changes were done, the core was reloaded from the Solr UI. In the php file another parameter solr => pageRankFile.txt desc was added. Now the UI is ready to produce queries based on both default Lucene as well as the PageRank Algorithm.

STEP 9 – Compare the ranking algorithms for the queries. The detailed description for the ranking along with the screenshots for the queries are given below.
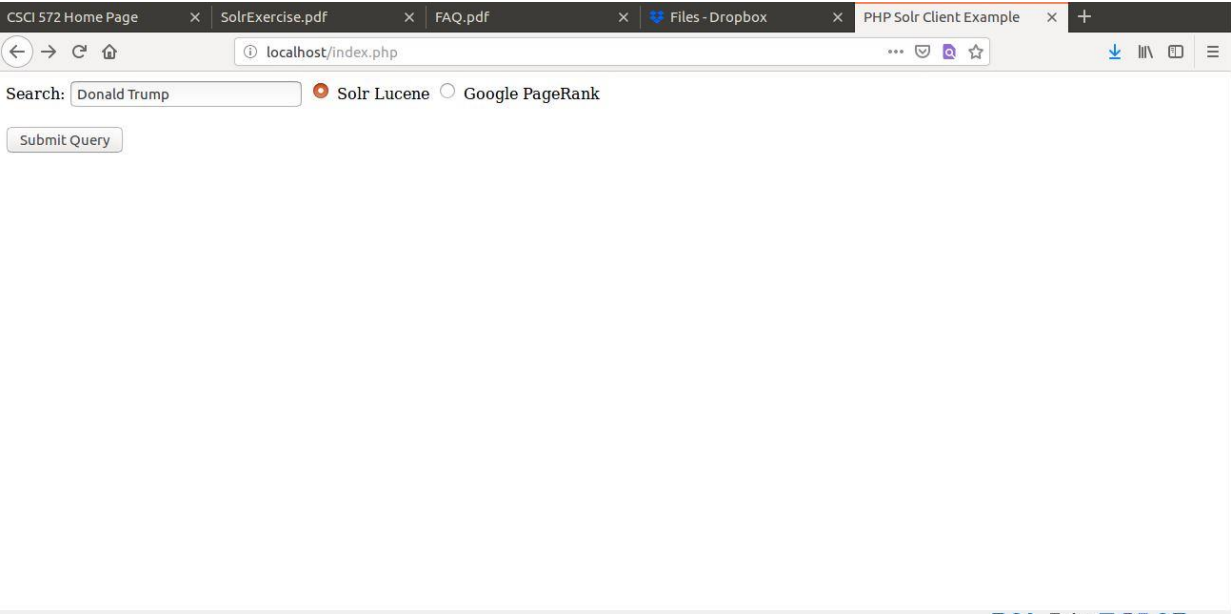
EXPLANATION WHY SOME QUERIES HAS HIGHER PAGEARNK VALUES THAN OTHERS?

Lucene ranking is based on Vector space model and Boolean model. It uses T-IDF and vector closeness to calculate the score for a webpage. On the other hand the PageRank uses the number of external links pointing to the webpage to determine its score. The edgesList file is generated to calculate the external pages pointing to a webpage. The more pages pointing to a given webpage the more will be its pageRank score. Each page has different number of webpages pointing to it and hence each webpage will have its own individual PageRank score.
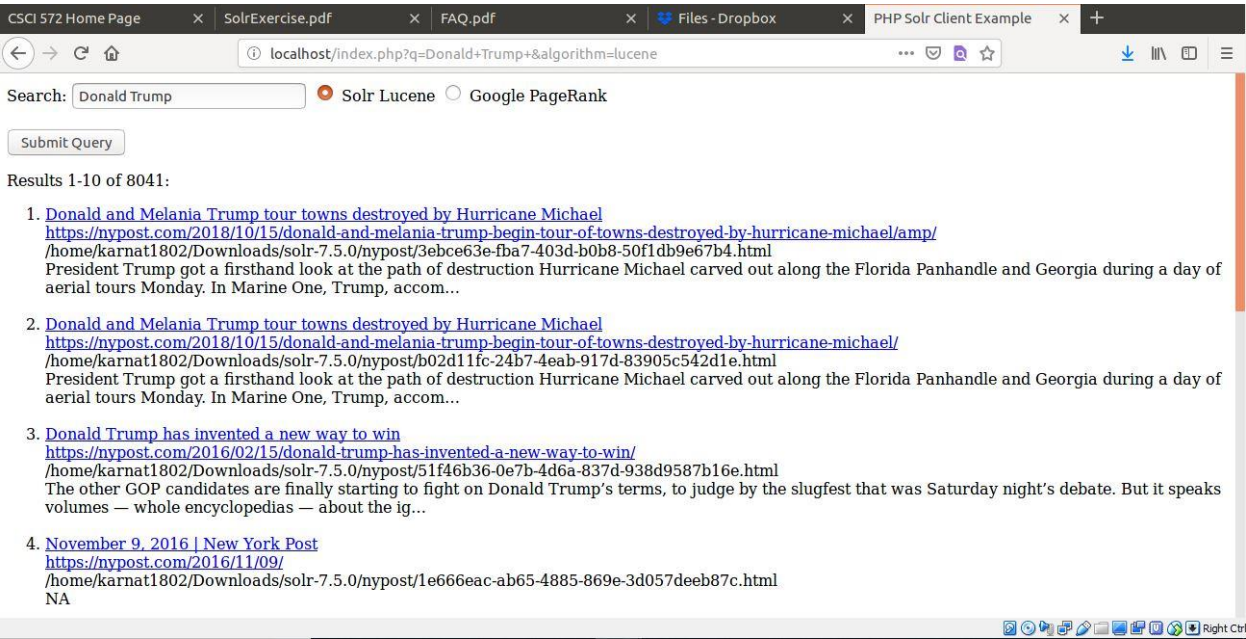
Due the difference in both the strategies that are used by Lucene and PageRank the ranking of document produced by the queries are different.

SCREENSHOTS

Screenshot 1 : Initial Page where you insert the query

Screenshot 2: Screenshot of the page that shows for Lucene

NA

5. How Donald Trump manipulated Bon Jovi out of Bills purchase
   https://nypost.com/2017/11/06/how-donald-trump-manipulated-bon-jovi-out-of-bills-purchase/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/4882c3e5-f749-4528-8082-8df47c4c48e5.html
   BUFFALO — Was Donald Trump the secret force behind the seemingly grassroots 2014 campaign to block Jon Bon Jovi from buying the NFL's Buffalo
   Bills? That's what those involved are sayin...

6. President Trump pays tribute to Mets great David Wright
   https://nypost.com/2018/10/01/donald-trump-pays-tribute-to-friend-david-wright/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/05473ac2-f9ed-4a02-888b-7ee35ac52868.html
   Everyone loves David Wright. The president of the United States is apparently part of that not-so-exclusive group. President Trump made a video
   congratulating Wright on his career that Wright's age...

7. With Trump's win, it's time for GOP to unite against Hillary
   https://nypost.com/2016/05/03/donald-trump-is-the-nominee-cruz-and-kasich-should-drop-out/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/9e047f06-6dbc-4081-827e-4f836023df50.html
   Ted Cruz did the right and honorable thing by suspending his presidential campaign Tuesday night. John Kasich should follow suit. It's time for the party
   to unite against Hillary Clinton. Donald Tr...

8. November 10, 2016 | New York Post
   https://nypost.com/2016/11/10/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/be115371-1223-4866-ba39-e5668c9b6456.html
   NA

9. January 20, 2017 | New York Post
   https://nypost.com/2017/01/20/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/bd1848eb-fd9e-4ff8-88f2-cdd4ff166856.html
   NA

Bills? That's what those involved are sayin...

6. President Trump pays tribute to Mets great David Wright
   https://nypost.com/2018/10/01/donald-trump-pays-tribute-to-friend-david-wright/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/05473ac2-f9ed-4a02-888b-7ee35ac52868.html
   Everyone loves David Wright. The president of the United States is apparently part of that not-so-exclusive group. President Trump made a video
   congratulating Wright on his career that Wright's age...

7. With Trump's win, it's time for GOP to unite against Hillary
   https://nypost.com/2016/05/03/donald-trump-is-the-nominee-cruz-and-kasich-should-drop-out/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/9e047f06-6dbc-4081-827e-4f836023df50.html
   Ted Cruz did the right and honorable thing by suspending his presidential campaign Tuesday night. John Kasich should follow suit. It's time for the party
   to unite against Hillary Clinton. Donald Tr...

8. November 10, 2016 | New York Post
   https://nypost.com/2016/11/10/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/be115371-1223-4866-ba39-e5668c9b6456.html
   NA

9. January 20, 2017 | New York Post
   https://nypost.com/2017/01/20/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/bd1848eb-fd9e-4ff8-88f2-cdd4ff166856.html
   NA

10. November 11, 2016 | New York Post
    https://nypost.com/2016/11/11/
    /home/karnat1802/Downloads/solr-7.5.0/nypost/ae6b0482-148d-46ef-b070-1c520f1bd5a6.html
    NA

Screenshot 3: Screenshot of page which shows result for PageRank

PHP Solr Client Example × +

localhost/index.php?q=Donald+Trump+&algorithm=pagerank

Search: Donald Trump   ○ Solr Lucene ● Google PageRank

Submit Query

Results 1-10 of 8041:

1. Entertainment | New York Post
   https://nypost.com/entertainment/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/3e6224e7-52b2-49bb-b35e-36e453227413.html
   NA

2. Technology News & Reviews | New York Post
   https://nypost.com/tech/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/40b8760d-53b6-4edb-a9b7-f60c3c13258d.html
   NA

3. Living | New York Post
   https://nypost.com/living/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/08328725-0507-4b37-a111-b1eda05fd1fc.html
   NA

4. Fashion News, Photos, and Video | New York Post
   https://nypost.com/fashion/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/6bc5ab6f-4ca4-4588-a3a8-f9a6bb45f6ab.html
   NA

5. Media | New York Post
   https://nypost.com/media/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/bbc54461-2472-49bf-8c43-4e1a383347f4.html

https://nypost.com/entertainment/

PHP Solr Client Example × +

localhost/index.php?q=Donald+Trump+&algorithm=pagerank

5. Media | New York Post
   https://nypost.com/media/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/bbc54461-2472-49bf-8c43-4e1a383347f4.html
   NA

6. Opinion | New York Post
   https://nypost.com/opinion/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/1f12f8e0-2b05-4fdc-98f0-2473818a378c.html
   NA

7. Daily Horoscope | New York Post
   https://nypost.com/horoscopes/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/e8940ca0-ff7c-4925-aaab-47e9f260d0c1.html
   NA

8. News | New York Post
   https://nypost.com/news/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/2f65e692-a9f9-4f32-9591-198ecc3a0d9b.html
   NA

9. Columnists | New York Post
   https://nypost.com/columnists/
   /home/karnat1802/Downloads/solr-7.5.0/nypost/fc17eac6-ad31-4c0c-b079-3bebbc17fcb9.html
   Visit the post for more.

10. New York Post
    https://nypost.com/
    /home/karnat1802/Downloads/solr-7.5.0/nypost/dc52408d-235a-4535-a9bd-4b08b1c777b0.html
    Your source for breaking news, news about New York, sports, business, entertainment, opinion, real estate, culture, fashion, and more.

Screenshot 4: Screenshot of actual web page that opens up on clicking the url.

QUERY COMPARISON

Query 1 : Donald Trump

Lucene	PageRank	Overlaps
https://nypost.com/2018/10/15/donald-and-melania-trump-begin-tour-of-towns-destroyed-by-hurricane-michael/amp/	https://nypost.com/entertainment/	
https://nypost.com/2018/10/15/donald-and-melania-trump-begin-tour-of-towns-destroyed-by-hurricane-michael/	https://nypost.com/tech/	
https://nypost.com/2016/02/15/donald-trump-has-invented-a-new-way-to-win/	https://nypost.com/living/	
https://nypost.com/2016/11/09/	https://nypost.com/fashion/	
https://nypost.com/2017/11/06/how-donald-trump-manipulated-bon-jovi-out-of-bills-purchase/	https://nypost.com/media/	0
https://nypost.com/2018/10/01/donald-trump-pays-tribute-to-friend-david-wright/	https://nypost.com/opinion/	
https://nypost.com/2016/05/03/donald-trump-is-the-nominee-cruz-and-kasich-should-drop-out/	https://nypost.com/horoscopes/	
https://nypost.com/2016/11/10/	https://nypost.com/news/	
https://nypost.com/2017/01/20/	https://nypost.com/columnists/	
https://nypost.com/2016/11/11/	https://nypost.com/	

QUERY 2: LA Lakers

Lucene	PageRank	Overlaps
https://nypost.com/2018/09/25/lebron-james-knows-it-wont-come-easy-with-lakers/	https://nypost.com/sports/	
https://nypost.com/2018/09/07/warriors-star-emerges-as-lakers-flashy-plan-b-in-free-agency/	https://nypost.com/2018/10/15/tom-bradys-postgame-words-to-gronk-are-an-nfl-nightmare/	
https://nypost.com/2018/10/10/klay-thompson-gushes-over-lebron-as-lakers-pursuit-looms/	https://nypost.com/2018/10/15/nfl-flameout-admits-to-having-sex-with-12-year-old-girl-police/	
https://nypost.com/tag/nba/	https://nypost.com/2018/10/15/eli-manning-trade-could-save-him-and-the-giants-espn-qb/	
https://nypost.com/tag/los-angeles-lakers/	https://nypost.com/2018/10/14/aaron-hernandez-had-sexual-relationship-with-hs-quarterback/	
https://nypost.com/2001/02/14/	https://nypost.com/2018/10/15/mlb-youngsters-bring-the-trash-talk-and-much-needed-fun/	0
https://nypost.com/2014/05/02/	https://nypost.com/2018/10/15/empty-vows-alderson-back-stab-inside-mets-gm-search-mess/	
https://nypost.com/2012/11/12/	https://nypost.com/2018/10/15/angry-about-the-subways-send-cuomo-a-message/	
https://nypost.com/2015/06/26/	https://nypost.com/2018/10/15/a-giants-trade-for-ex-patriots-quarterback-would-make-sense/	
https://nypost.com/2017/02/21/	https://nypost.com/2018/10/15/how-one-veterans-vietnam-horrors-inspired-this-is-us/	

QUERY 3: Star Wars

Lucene	PageRank	Overlaps
https://nypost.com/2015/12/18/	https://nypost.com/entertainment/	
https://nypost.com/2015/12/15/	https://nypost.com/sports/	
https://nypost.com/2015/12/17/	https://nypost.com/metro/	
https://nypost.com/2015/12/16/	https://nypost.com/tech/	**0**
https://nypost.com/2015/12/22/	https://nypost.com/living/	

https://nypost.com/2015/12/14/	https://nypost.com/fashion/	
https://nypost.com/2015/11/24/	https://nypost.com/media/	
https://nypost.com/2015/04/20/	https://nypost.com/real-estate/	
https://nypost.com/2015/10/20/	https://nypost.com/photos/	
https://nypost.com/2015/04/16/	https://nypost.com/horoscopes/	

QUERY 4: Lebron James

Lucene	PageRank	Overlaps
https://nypost.com/2018/09/25/lebron-james-knows-it-wont-come-easy-with-lakers/	https://nypost.com/sports/	
https://nypost.com/2018/04/04/lebron-saban-feud-over-shop-talk-show-in-copyright-dispute/	https://nypost.com/metro/	
https://nypost.com/tag/los-angeles-lakers/	https://nypost.com/2018/10/13/100m-cybersecurity-center-slated-to-open-in-soho-next-year/	
https://nypost.com/2010/07/09/	https://nypost.com/2018/10/15/judge-tosses-stormy-daniels-defamation-suit-against-trump/	
https://nypost.com/2018/07/02/	https://nypost.com/2018/10/15/beach-house-still-standing-after-hurricane-michael-devastates-nearby-homes/	0
https://nypost.com/2018/10/10/klay-thompson-gushes-over-lebron-as-lakers-pursuit-looms/	https://nypost.com/2018/10/15/ex-nightlife-king-plans-to-open-a-private-club-in-noho/	
https://nypost.com/2010/07/08/	https://nypost.com/2018/10/15/nyc-goes-entire-weekend-without-a-shooting-for-first-time-in-14-years/	
https://nypost.com/2014/07/11/	https://nypost.com/2018/10/15/new-york-city-restaurant-scene-teeming-with-new-entries/	
https://nypost.com/2010/07/01/	https://nypost.com/2018/10/15/knicks-season-preview-what-theyre-really-playing-for/	
https://nypost.com/2010/07/07/	https://nypost.com/2018/10/12/little-kids-are-wearing-grown-up-makeup-is-it-ok/	

QUERY 5: 2018 World Cup

Lucene	PageRank	Overlaps
https://nypost.com/2018/07/12/theres-a-knicks-battle-and-bet-over-the-world-cup-final/	**https://nypost.com/entertainment/**	
https://nypost.com/video/this-psychic-pig-just-predicted-the-world-cup/	**https://nypost.com/sports/**	
https://nypost.com/2018/06/05/nycfc-cant-wait-to-get-monkey-off-their-back-in-u-s-open-cup/	**https://nypost.com/tips/**	
https://nypost.com/2018/10/04/stricken-ryder-cup-fan-i-thought-id-be-trampled-by-tiger-woods-fans/	**https://nypost.com/metro/**	**0**
https://nypost.com/2018/07/13/son-of-stanley-cup-winning-devil-chasing-nhl-dream/	**https://nypost.com/tech/**	
https://nypost.com/2014/06/12/	**https://nypost.com/living/**	
https://nypost.com/2014/06/11/	**https://nypost.com/mobile-apps/**	
https://nypost.com/2018/10/04/capitals-stanley-cup-tribute-was-a-little-bit-petty/	**https://nypost.com/fashion/**	
https://nypost.com/2018/10/02/nhl-preview-these-stanley-cup-contenders-are-racing-the-clock/amp/	**https://nypost.com/media/**	
https://nypost.com/tag/tiger-woods/	**https://nypost.com/real-estate/**	

QUERY 6: North Korea

Lucene	PageRank	Overlaps
https://nypost.com/2018/08/30/a-glimpse-at-life-inside-north-korea/slide-1/	**https://nypost.com/real-estate/**	
https://nypost.com/2018/08/24/state-department-reportedly-blindsided-by-trump-canceling-north-korea-trip/	**https://nypost.com/2018/10/15/tom-bradys-postgame-words-to-gronk-are-an-nfl-nightmare/**	
https://nypost.com/2018/06/07/former-pow-calls-on-trump-to-reclaim-us-ship-captured-by-north-korea/	**https://nypost.com/2018/10/15/nfl-flameout-admits-to-having-sex-with-12-year-old-girl-police/**	
https://nypost.com/2018/10/07/pompeo-progress-made-with-kim-jong-un-on-north-korea-trip/	**https://nypost.com/2018/10/15/eli-manning-trade-could-save-him-and-the-giants-espn-qb/**	**0**
https://nypost.com/2018/09/10/north-korea-marks-the-hermit-kingdoms-70th-anniversary/slide-1/	**https://nypost.com/2018/10/14/aaron-hernandez-had-sexual-relationship-with-hs-quarterback/**	

https://nypost.com/2017/09/09/how-democrats-left-america-naked-before-north-koreas-nukes/	https://nypost.com/2018/10/15/mlb-youngsters-bring-the-trash-talk-and-much-needed-fun/	
https://nypost.com/2017/09/03/	https://nypost.com/2018/10/15/empty-vows-alderson-back-stab-inside-mets-gm-search-mess/	
https://nypost.com/2018/10/02/trumps-love-for-kim-is-undercutting-his-own-teams-necessary-toughness/	https://nypost.com/2018/10/15/how-eddie-lampert-led-sears-into-bankruptcy-disaster/	
https://nypost.com/2017/09/04/	https://nypost.com/2018/10/15/the-south-bronx-is-in-the-midst-of-a-development-boom/	
https://nypost.com/2018/10/04/north-korean-hackers-linked-to-global-banking-heist/		

QUERY 7: Hurricane Florence

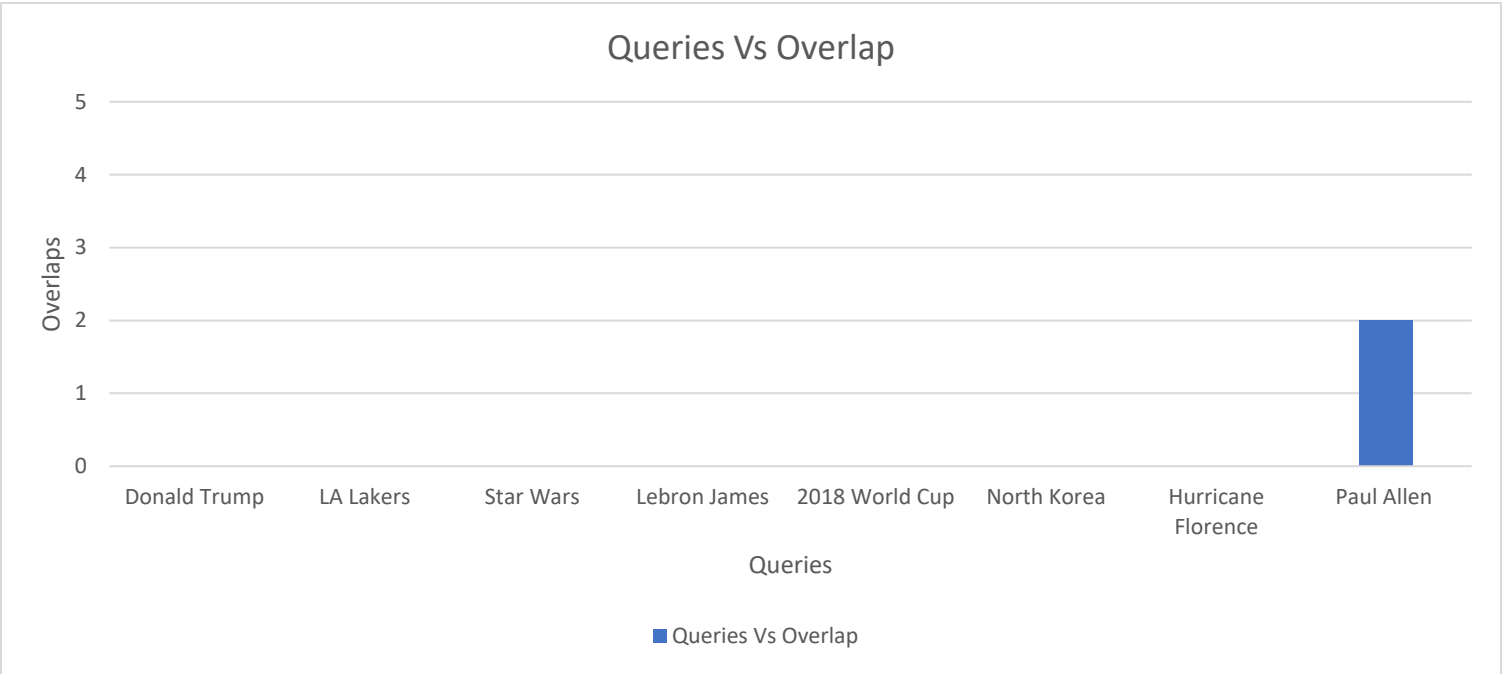| Lucene https://nypost.com/2018/09/15/kitten-named-survivor-rescued-with-owner-from-hurricane-florence/	PageRank	Overlaps
https://nypost.com/2018/09/14/hurricane-florence-storms-north-carolina/slide-1/	https://nypost.com/real-estate/	
https://nypost.com/2018/09/11/the-carolinas-brace-for-hurricane-florence/slide-1/	https://nypost.com/photos/	
https://nypost.com/2018/09/13/hurricane-florence-closes-in-on-carolina-coast/	https://nypost.com/news/	
https://nypost.com/2018/09/10/hurricane-florence-looks-wild-from-the-international-space-station/	https://nypost.com/2018/10/15/surveillance-video-shows-moment-before-model-shot-in-neck-crashes-car/	
https://nypost.com/2018/09/20/unc-football-players-lend-hand-to-hurricane-florence-victims/	https://nypost.com/2018/10/15/beach-house-still-standing-after-hurricane-michael-devastates-nearby-homes/	0
https://nypost.com/2018/09/14/astronaut-snaps-picture-of-hurricane-florence-as-it-makes-landfall/	https://nypost.com/2018/10/15/ex-nightlife-king-plans-to-open-a-private-club-in-noho/	
https://nypost.com/2018/09/24/hurricane-florence-isnt-done-with-the-carolinas-just-yet/	https://nypost.com/2018/10/15/sears-death-of-an-american-icon/slide-1/	
https://nypost.com/2018/09/19/north-carolina-hog-lagoons-are-exploding-poo-after-hurricane-florence/	https://nypost.com/2018/10/15/heres-what-elizabeth-warrens-dna-test-shows/	
https://nypost.com/tag/hurricanes/page/2/	https://nypost.com/2018/10/11/michael-leaves-florida-panhandle-in-shambles/slide-1/	

https://nypost.com/2018/09/15/kitten-named-survivor-rescued-with-owner-from-hurricane-florence/	https://nypost.com/2018/10/10/hurricane-michael-starts-pounding-the-florida-panhandle/slide-1/	

QUERY 8: Paul Allen

Lucene	PageRank	Overlaps
https://nypost.com/2018/10/15/microsoft-co-founder-paul-allen-dead-at-65/amp/	https://nypost.com/sports/	
https://nypost.com/2018/10/15/microsoft-co-founder-paul-allen-dead-at-65/?shared=email&msg=fail	https://nypost.com/tech/	
https://nypost.com/2018/10/15/microsoft-co-founder-paul-allen-dead-at-65/?shared=email&msg=fail	https://nypost.com/news/	
https://nypost.com/tag/paul-allen/	https://nypost.com/columnists/	
https://nypost.com/2018/03/06/microsoft-co-founder-finds-long-lost-wwii-aircraft-carrier/amp/	https://nypost.com/	
https://nypost.com/author/carleton-english/	https://nypost.com/2018/10/15/microsoft-co-founder-paul-allen-dead-at-65/	**2**
https://nypost.com/tech/page/2/	https://nypost.com/2018/10/15/worlds-youngest-male-billionaire-is-25-year-old-model/	
https://nypost.com/tech/page/3/	https://nypost.com/2018/10/15/sears-death-of-an-american-icon/slide-1/	
https://nypost.com/tech/	https://nypost.com/2018/10/15/state-department-sends-out-email-featuring-cookie-monster-cat/	
https://nypost.com/2018/03/06/microsoft-co-founder-finds-long-lost-wwii-aircraft-carrier/amp/	https://nypost.com/2018/10/15/de-blasios-nycha-strategy-was-deceptive-from-the-start/	

Overlap Graph

Queries Vs Overlap

END OF REPORT