**Mukesh Dangi(**USC ID **: 4297380684)**

**New York Post** – Comparing Search Engine Ranking Algorithms

Step 1: Solr Setup and Indexing the HTML files

1. Downloaded files NY Post URL to HTML mapped data and HTML pages from the google drive link. Downloaded and extracted all the crawled files from google drive for nypost.
2. Downloaded the file solr-7.5.0 and Installed solr-7.5.0.
3. Created a core by going into the /solr-7.5.0 directory started solr using **bin/solr start** and then created solr core by running **bin/solr -c create nypost,** where nypost is core name**.**
4. Edited the managed-schema, solrconfig file in the config folder so that all the text content can be mapped properly. This command will index all the documents on Solr and might take some time depending upon dataset.
5. Posted all crawled files on solr server using the **bin/post -c nypost -filetypes html path to nypost** folder
6. Reloaded the nypost core from Solr UI admin to verify changes made were correct

Step 2: Angular 7, NodeJS installation and Front End:

1. Installed Angular 7, Angular CLI and NodeJS using NPM (Node package manager)
2. Created new NG project using **ng new nypost.** Added all the dependencies in package.json and ran **npm install.**
3. Used MVC architecture in the project and created search box, ajax calls and proxy server.
4. Created solrseach.js, a node js file which uses express js, csvtojson and many other relevant libraries.
5. **let SolrNode = require('solr-node');**

**let client = new SolrNode({**

**host: 'localhost', port: '8983',**

**core: 'nypost', protocol: 'http'**

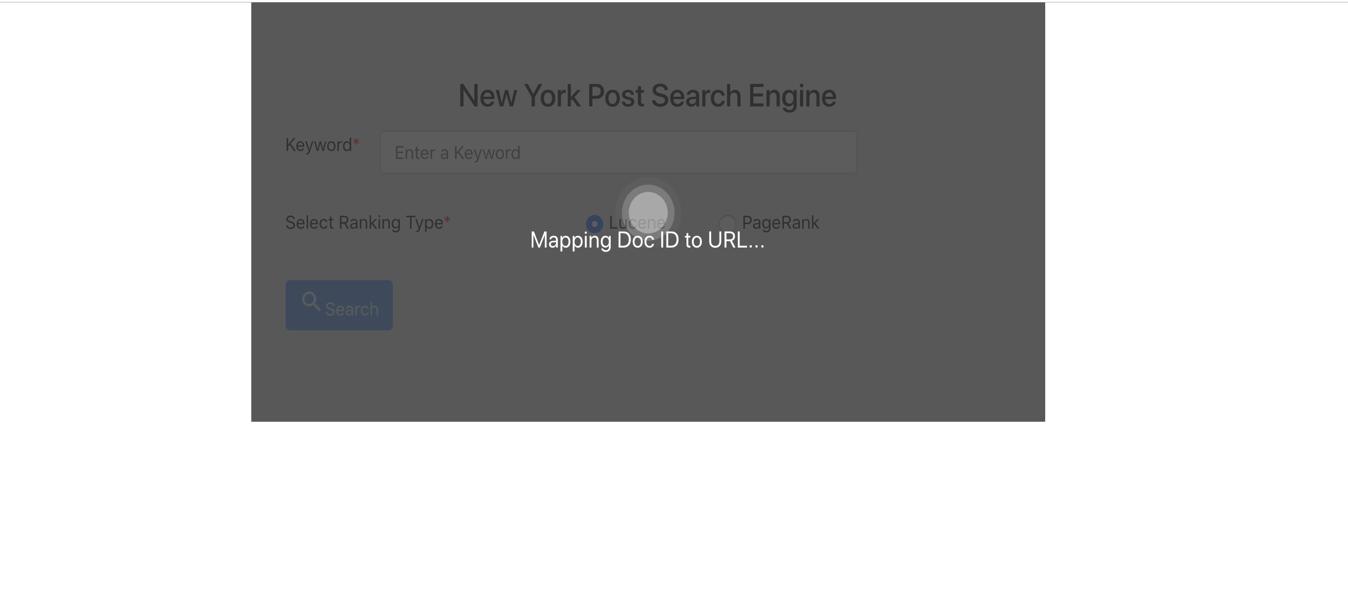
**});**

1. Initiated Solr search with

**let strQuery = client.query().q('La Lakers');**

**client.search(strQuery, function (err, result) {results}**

1. Made all the API calls to from localhost:4200 which is angular port to localhost:3000 node port using proxy config. Proxy config is being used in the ng serve command.
2. On search page load, I called ***/api/getcsvjson api*** to node js server to get all the HTML doc Id to HTML URL mapping.
3. On page load (load csv, map doc id to URL)



1. First 10 results are displayed. The URL for the web page file name was obtained by parsing the mappings csv file.
2. Passing **‘&sort = pageRankFile desc’** if user opt for external page rank Algorithms otherwise no sort argument passed.

Step 3: Generating Edge List:

1. Wrote a Java program using JSoup library.
2. Using JSoup API, I extracted all **a[href]** elements and then all the URLs attributes **abs:href** from those Elements. I extracted the outgoing links of each web page by traversing the directory to create the edgeList.txt file.
3. Used URLtoHTML\_nypost.csv to check which of the links are connected to other links. Here I have used doc id for the comparison.
4. Two webpages will make an entry into Edge list only if the they are inter-liked to each other in some way. Either an edge from page A to page B or vice -versa.

Step 4: Page Rank Calculation:

1. Used NetworkX library in python to computed page rank and create a web graph of those Edge list.
2. Created web graph into NetwokX using read\_edgelist() function with edgelist.txt and digraph as parameters.
3. Used pagerank() function to calculate Page Rank and output the result to external\_pageRankFile.txt
4. Wrote python program to compute the page rank values using the NetworkX library.
5. python code iterates over each line from edgelist.txt and computes the page rank values using the following configuration
6. ***alpha=0.85, tol=1e-06 personalization=None, nstart=None, max\_iter=30, , weight='weight', dangling=None.***
7. Finally, the python program using NetworkX library created an external\_pageRankFile.txt file having format <document\_id>=<page\_rank\_score>.

Step 5: Solr Core update with new changes of external PageRank file:

1. Added the external\_pageRankFile.txt in the data folder of the core which we created in step 1. Added necessary changes to managed-schema by adding field, and fieldType params.
2. Managed Schema changes:
   1. <fieldType name=”external” keyField=”id” defVal=”0” class=”solr.ExternalFileField”/>
   2. <field name=”pageRankFile” type=”external” stored=”false” indexed=”false”/>
3. Solr Config Changes (Added external new searcher listeners)
   1. <listener event=”newSearcher” class=”org.apache.schema.ExternalFileFieldRealoder”/>
   2. <listener event=”firstSearcher” class=”org.apache.schema.ExternalFileFieldRealoder”/>
4. Modified the solrconfig.xml file to include listeners as per the instructions.
5. Reloaded the Solr **nypost** core on the Solr admin page. Next, queried from Solr Admin Page by placing “pageRankFile desc” in sort parameter of the query to get the results based on pageRankFile in descending order.

Explanation for higher page rank values

The pages with large number of incoming links especially a page that is linked to by many pages with high PageRank receives a high rank value itself.

Ten results produced for both Page rank method and default solr method for all the queries:

1. **Donald Trump:**

|  |  |
| --- | --- |
| Default Solr Method | External Page rank method |
| aea27c62-f6dd-4fb3-9c41-c212af5ff6e8.html | ﻿f5a772b8-662d-49b6-92f9-dcf15ffd388d.html |
| c3d89955-6d41-4fc4-879a-824ddee16e91.html | db9f279a-359a-43e2-bcd5-ad75becb6606.html |
| 3ebce63e-fba7-403d-b0b8-50f1db9e67b4.html | 40b8760d-53b6-4edb-a9b7-f60c3c13258d.html |
| 1e666eac-ab65-4885-869e-3d057deeb87c.html | ﻿6bc5ab6f-4ca4-4588-a3a8-f9a6bb45f6ab.html |
| e8174513-85d2-4445-a934-eaa4188fbcc9.html | 9e1975ef-08c7-45f3-8f77-120bff2a5b32.html |
| ed3379cc-3531-4936-9234-283f3709e755.html | fc17eac6-ad31-4c0c-b079-3bebbc17fcb9.html |
| d5d7f5c8-d1a3-450b-9028-30a4b94d846f.html | 6dcc0483-03d1-4767-b2dd-d791046f94c1.html |
| abf7955b-595d-4889-9024-515c73918d28.html | 9304c93b-bc36-494c-bd58-4eb8a6425a44.html |
| 50a1285c-9953-42e0-9cd7-06fbf18c131e.html | 88f9f9d1-491d-4df3-981d-863a8b1b485a.html |
| 3d620c6c-2630-46b7-8128-3ebd1ff55525.html | e8940ca0-ff7c-4925-aaab-47e9f260d0c1.html |

1. **LA Lakers:**

|  |  |
| --- | --- |
| Default Solr Method | External Page rank method |
| 7856e10c-3cee-4e00-93e1-a8538d2c1be9.html | ﻿8790e35b-13e7-4277-abec-631c5ca93fd3.html |
| dbae01ed-fc65-4a63-9fb8-e709fb2229c9.html | 14965946-fc00-4d63-a25e-ea27f2efacfd.html |
| 4c86b1ae-2c2d-474f-aa94-87700240e99f.html | a2308fa2-5653-4684-b4cc-c479d3a296b3.html |
| 27f1295b-466b-4636-b9b5-bc7c8df9e0c9.html | c60c3f9d-f05a-4bc1-8d60-c257165e94c2.html |
| 121d01f1-64cb-4550-a6c0-605a11004d6b.html | 19f42538-046e-4566-a46f-86dbed6ef25d.html |
| d90cee60-7435-4cdb-94d8-f6ac1cfbdadc.html | 98ceb784-aae1-4a09-add5-d36fdc2c8866.html |
| 636c3495-c1d9-4fc8-95e8-7a64938da7be.html | a0fbd273-4856-450e-bec1-542a8a7ad8ec.html |
| 1123e8f3-8a32-4fab-9f27-75827be2fb70.html | bf89ebee-3c4f-40cb-9cb7-19f7d2fb3514.html |
| 235df70f-bdc7-4645-93bb-99407c30784d.html | b6fcb411-eac2-41f5-9d75-ff04f6f7f777.html |
| 32e0bc21-1473-4d59-8938-7f0f84a4d05a.html | 32eaff2d-95a4-4e88-bc6d-10d35ee9fbf5.html |

1. **Star Wars:**

|  |  |
| --- | --- |
| Default Solr Method | External Page rank method |
| ee74e3bb-ea2c-43c1-bc83-f7b6ace3b460.html | db9f279a-359a-43e2-bcd5-ad75becb6606.html |
| 4491a6de-3d7b-4988-ab2f-94458c8f2046.html | 40b8760d-53b6-4edb-a9b7-f60c3c13258d.html |
| ba259ebd-c57b-4cae-b905-ae1d739737b4.html | 6bc5ab6f-4ca4-4588-a3a8-f9a6bb45f6ab.html |
| 688728e6-c354-4dc7-98e7-f2ef4a2ab5d4.html | fc17eac6-ad31-4c0c-b079-3bebbc17fcb9.html |
| 9ac36575-0add-47e9-a872-53e8eb628a32.html | 88f9f9d1-491d-4df3-981d-863a8b1b485a.html |
| 772ed371-1a0f-4052-a451-bdf4889fa554.html | e8940ca0-ff7c-4925-aaab-47e9f260d0c1.html |
| 133dc680-99d6-4e55-9110-014b90469839.html | 72897ce7-e9b1-489f-810d-bba4f5261645.html |
| d5278590-fcfd-4723-aadd-34c2e8f43711.html | 8790e35b-13e7-4277-abec-631c5ca93fd3.html |
| 9426d89a-91ad-4558-948c-3b1db285f7bb.html | 2f65e692-a9f9-4f32-9591-198ecc3a0d9b.html |
| ba57043a-4b7d-4575-ab0b-0a9500edf8c4.html | 08328725-0507-4b37-a111-b1eda05fd1fc.html |

1. **Lebron James:**

|  |  |
| --- | --- |
| Default Solr Method | External Page rank method |
| 7856e10c-3cee-4e00-93e1-a8538d2c1be9.html | 72897ce7-e9b1-489f-810d-bba4f5261645.html |
| 3fc52765-999e-4a16-889e-6ac2b4a223d1.html | 8790e35b-13e7-4277-abec-631c5ca93fd3.html |
| 121d01f1-64cb-4550-a6c0-605a11004d6b.html | 0152ac12-1507-4bbd-8471-ad819fb7a010.html |
| 7a97608c-841f-4403-a782-e858a53fa23e.html | c6cf8280-d576-4ac1-9fb4-4291561097d9.html |
| 571472d8-790a-43c4-9ba9-caa3c1eb0693.html | b2e7b626-de2a-4610-a9f8-127c85231751.html |
| 4c86b1ae-2c2d-474f-aa94-87700240e99f.html | 63481d75-f371-4ce9-98c6-189db19fd80b.html |
| 4e8b8f21-979d-438d-990d-9a190cba5b70.html | f9594a03-4efc-46a1-9049-91482921def0.html |
| 0bebe75e-be86-41b4-8082-947641dcbccf.html | 3bc082a7-0349-469b-97b3-bc242a031422.html |
| 920837b4-deb9-4c2f-af2e-b0ce334df29e.html | 95052138-85f4-4972-88e8-5c2688c11783.html |
| 1c1d3046-4401-40ad-ba10-6351044f881b.html | b640d196-95ef-4dcd-ba99-fef56657b0f4.html |

1. **2018 World Cup:**

|  |  |
| --- | --- |
| Default Solr Method | External Page rank method |
| cbfe590e-bef5-4798-ac49-11e4062906df.html | f5a772b8-662d-49b6-92f9-dcf15ffd388d.html |
| 8a52ca34-45d7-4e5a-836f-c0c1707da00e.html | db9f279a-359a-43e2-bcd5-ad75becb6606.html |
| 40641f08-3f72-4af8-8eee-d802969aa528.html | 40b8760d-53b6-4edb-a9b7-f60c3c13258d.html |
| 57fcb62a-129a-4342-bc8a-a689f87239b0.html | 6bc5ab6f-4ca4-4588-a3a8-f9a6bb45f6ab.html |
| 8d871909-86a4-4638-a508-b94c9466ece2.html | 9e1975ef-08c7-45f3-8f77-120bff2a5b32.html |
| d8dd177a-435d-40a9-ae34-6edb0d919c1d.html | fc17eac6-ad31-4c0c-b079-3bebbc17fcb9.html |
| 95c15d1d-60b1-4b22-8bde-aec38b5c916b.html | 6dcc0483-03d1-4767-b2dd-d791046f94c1.html |
| 2369952c-dbf2-4352-a3ea-14a34b7fb131.html | 9304c93b-bc36-494c-bd58-4eb8a6425a44.html |
| 7ad33c91-d65b-4f1e-b48a-00d0e9635727.html | 88f9f9d1-491d-4df3-981d-863a8b1b485a.html |
| 5ad7e0c5-34ab-4b12-99bb-7bf7110d8c7a.html | e8940ca0-ff7c-4925-aaab-47e9f260d0c1.html |

1. **North Korea**

|  |  |
| --- | --- |
| Default Solr Method | External Page rank method |
| 578d17b6-aae3-450f-9d00-e819317735b9.html | db9f279a-359a-43e2-bcd5-ad75becb6606.html |
| 1b9dbf54-30f0-4089-8120-42e883ad3190.html | 8790e35b-13e7-4277-abec-631c5ca93fd3.html |
| 7338c1c4-f450-4cdc-b0f8-4785229f58f8.html | 14965946-fc00-4d63-a25e-ea27f2efacfd.html |
| 04e9fa64-57f3-46f2-958c-fc72e485180c.html | a2308fa2-5653-4684-b4cc-c479d3a296b3.html |
| ba10023b-02ef-4c41-bfa7-6c082bb6ccce.html | c60c3f9d-f05a-4bc1-8d60-c257165e94c2.html |
| c162ea51-5228-4e03-a3ae-94e7b410a1d1.html | 19f42538-046e-4566-a46f-86dbed6ef25d.html |
| 518b63b0-82e1-4282-91b4-d07f531fe49a.html | 98ceb784-aae1-4a09-add5-d36fdc2c8866.html |
| 280fbb2e-f7c6-4d95-afb8-fc495a6dad44.html | a0fbd273-4856-450e-bec1-542a8a7ad8ec.html |
| 898fd909-cb83-4221-9065-a48cca028b48.html | ce677a5e-7c95-431c-96ac-5f4151e0940f.html |
| cc803d21-d568-467f-8b49-6516e07f2810.html | 125a4ac4-54f1-47d3-a99e-dfdc86c714ea.html |

1. **Hurricane Florence**

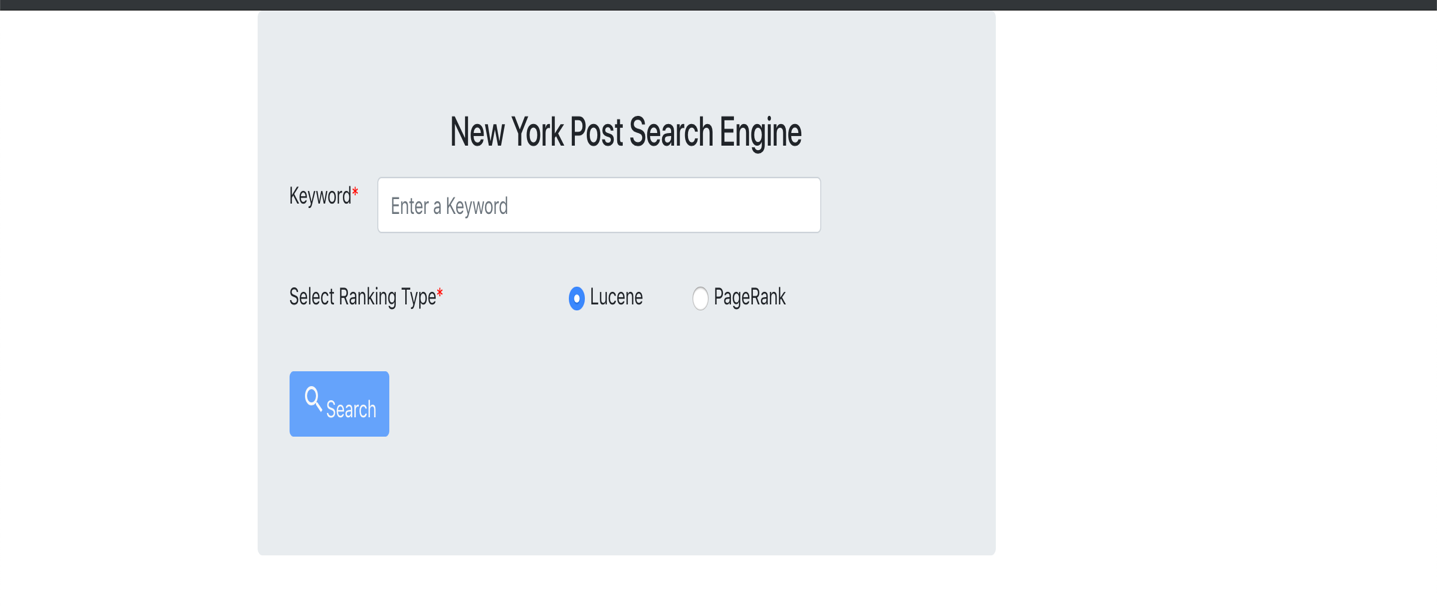
|  |  |
| --- | --- |
| Default Solr Method | External Page rank method |
| c7ceb758-465e-41dc-8ef7-4fdaa80b0fa9.html | db9f279a-359a-43e2-bcd5-ad75becb6606.html |
| 8eceb803-7343-498c-a53c-8bf25ff65b13.html | 88f9f9d1-491d-4df3-981d-863a8b1b485a.html |
| fec5f5e8-2847-453a-a6bf-5b41312610bd.html | 2f65e692-a9f9-4f32-9591-198ecc3a0d9b.html |
| 9200d7aa-1f72-4cea-a0be-45c1bd1b5a18.html | 9a67d0ce-ea79-476c-a348-551c7eeb7e9b.html |
| b44adbc3-a830-4c15-9387-b20a5970f2d5.html | b2e7b626-de2a-4610-a9f8-127c85231751.html |
| e54fe440-bcef-464a-b381-8865f08faec1.html | 63481d75-f371-4ce9-98c6-189db19fd80b.html |
| 6d7c7533-5c12-44ac-858c-c5cf3ebfa72e.html | 2d83166a-b564-45a4-8641-36d04bb99766.html |
| 5881054c-8610-417c-b1d2-228b07e3314e.html | e8b84f6a-5701-4283-8f52-1d947632760b.html |
| f68ee1fb-cd8d-4754-9399-b68323af283f.html | 790f0b67-eaf1-446c-9452-81e078b0a16a.html |
| 1a2fa94c-8709-4d3d-bb1d-a9f119637222.html | 62d6d7c2-bba1-4c4e-9efa-a19d4748bf14.html |

1. **Paul Allen**

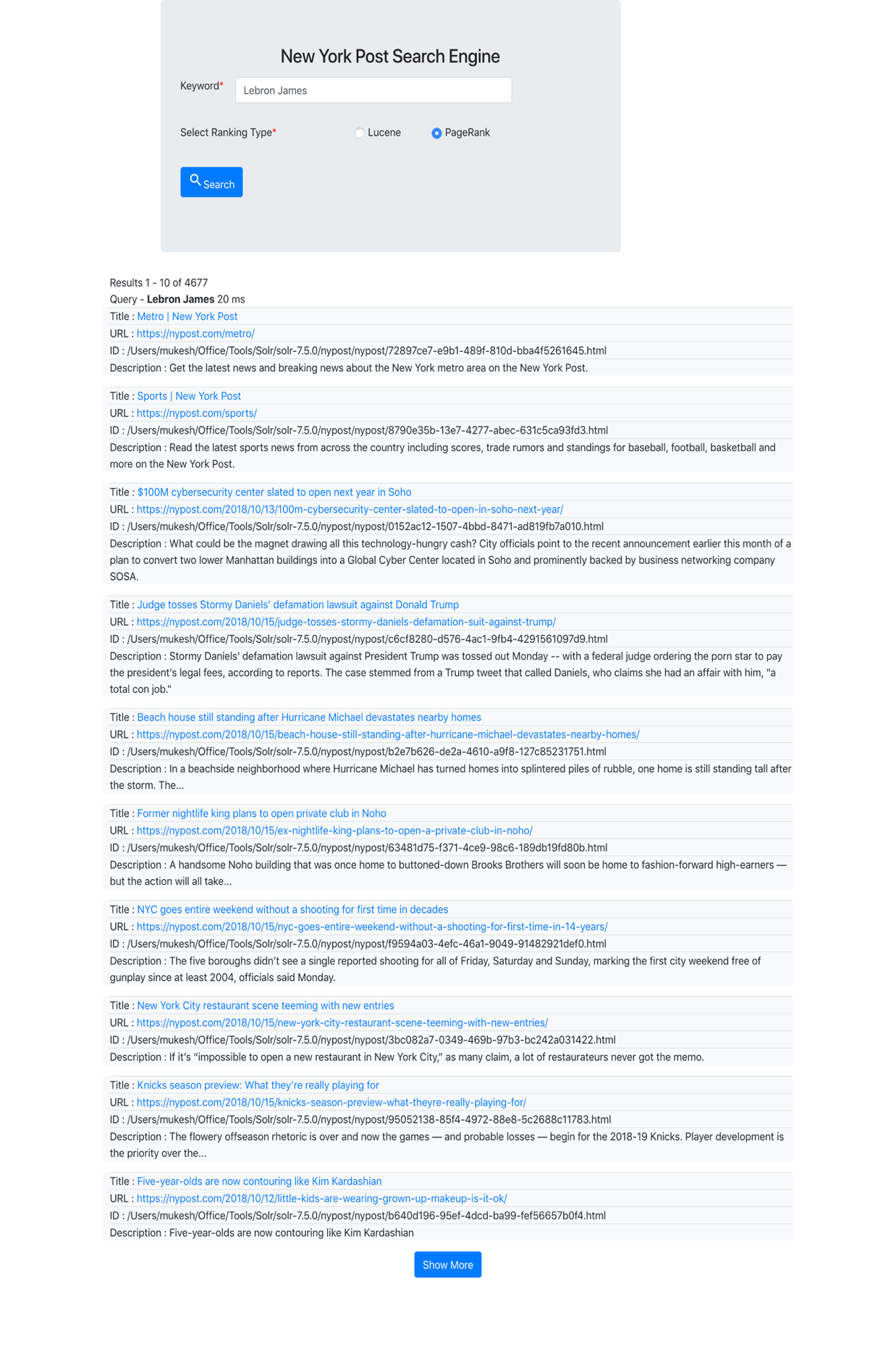
|  |  |
| --- | --- |
| Default Solr Method | External Page rank method |
| 14c83721-0dc6-45ad-959d-c834b1917d8a.html | **40b8760d-53b6-4edb-a9b7-f60c3c13258d.html** |
| 02423896-fc91-4f04-94dc-4bb9136ae283.html | fc17eac6-ad31-4c0c-b079-3bebbc17fcb9.html |
| **d6304d6f-f459-4993-a533-c223b6d6be02.html** | 8790e35b-13e7-4277-abec-631c5ca93fd3.html |
| c5ca3dc2-3b72-4026-9be1-1f0c775e3777.html | 2f65e692-a9f9-4f32-9591-198ecc3a0d9b.html |
| cfe9b62f-c461-42f4-9e49-1fead46936f8.html | dc52408d-235a-4535-a9bd-4b08b1c777b0.html |
| 971ee8c4-34fd-4ff2-a3fa-84fc9a69b49e.html | **d6304d6f-f459-4993-a533-c223b6d6be02.html** |
| c0965d22-7ff1-46dd-8995-91f53fb5106f.html | 7fbf6e8f-6f2a-49f2-9ae1-d4e195cfd49f.html |
| 8cfd526a-8915-4966-b7de-d30813b0e4ce.html | 2d83166a-b564-45a4-8641-36d04bb99766.html |
| **40b8760d-53b6-4edb-a9b7-f60c3c13258d.html** | 7f94956c-c400-42e8-9641-c945ea9e07c5.html |
| 6a5ec53a-e9eb-43e9-93db-f5f7cd5413d0.html | 96510681-810e-4c59-902c-a7ac8d741e6b.html |

Screenshots that describes the whole flow:

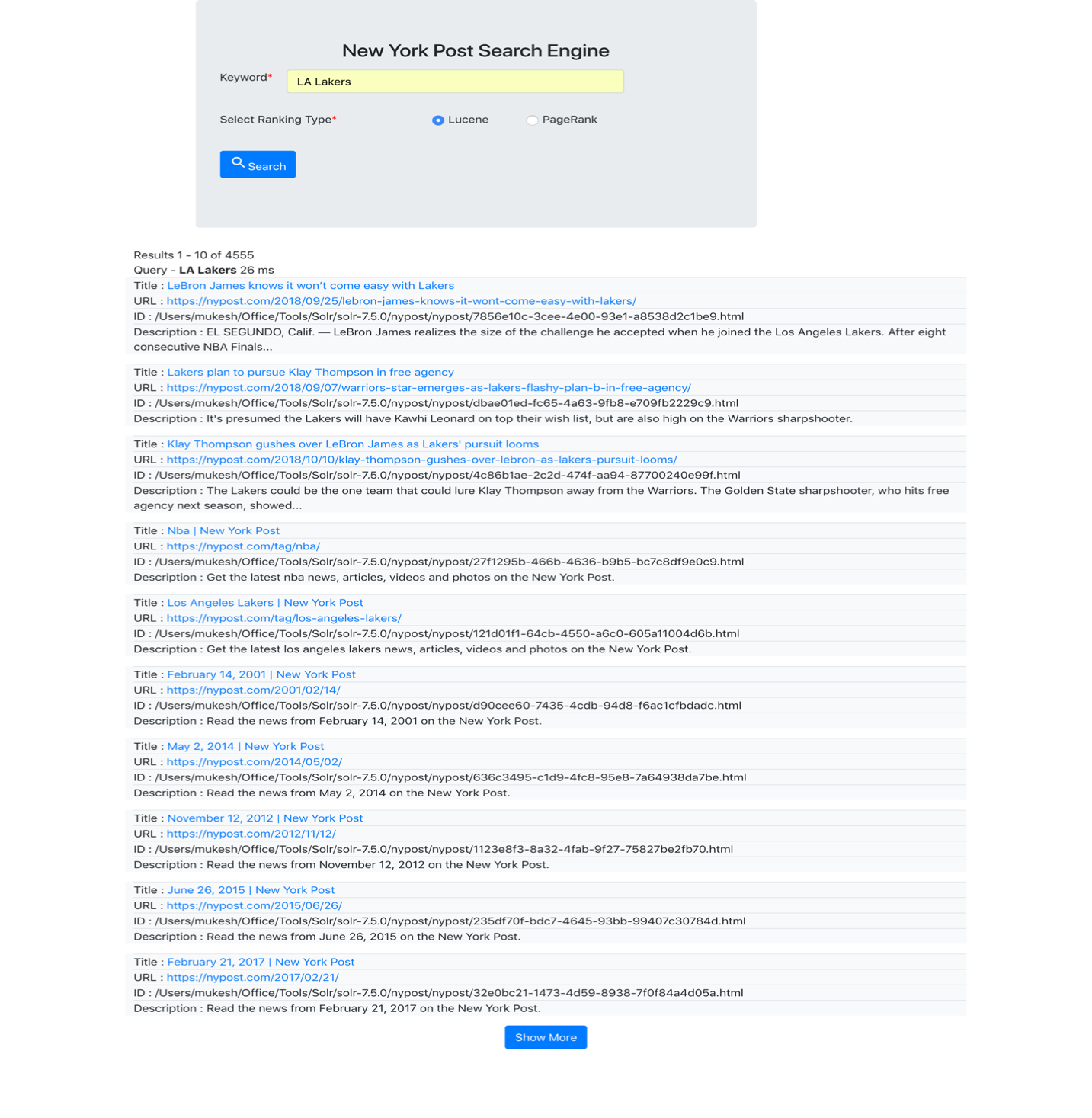
1. **Initial Page**



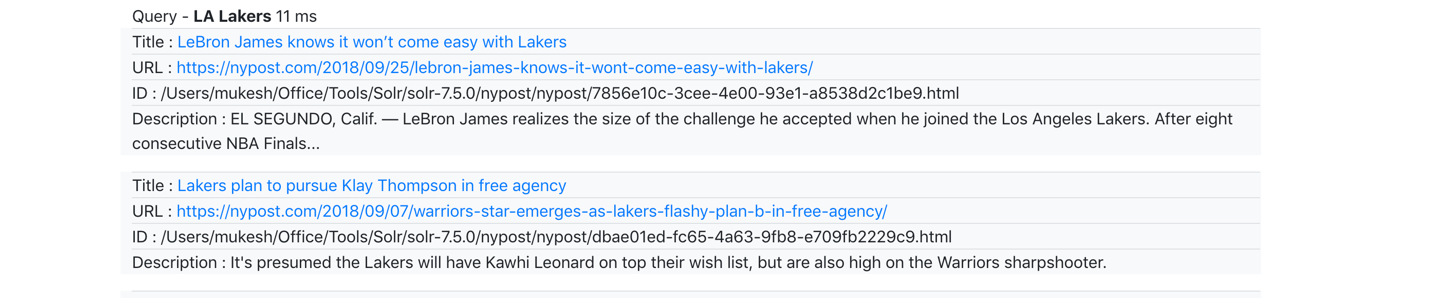
1. **External Page Rank Results:**



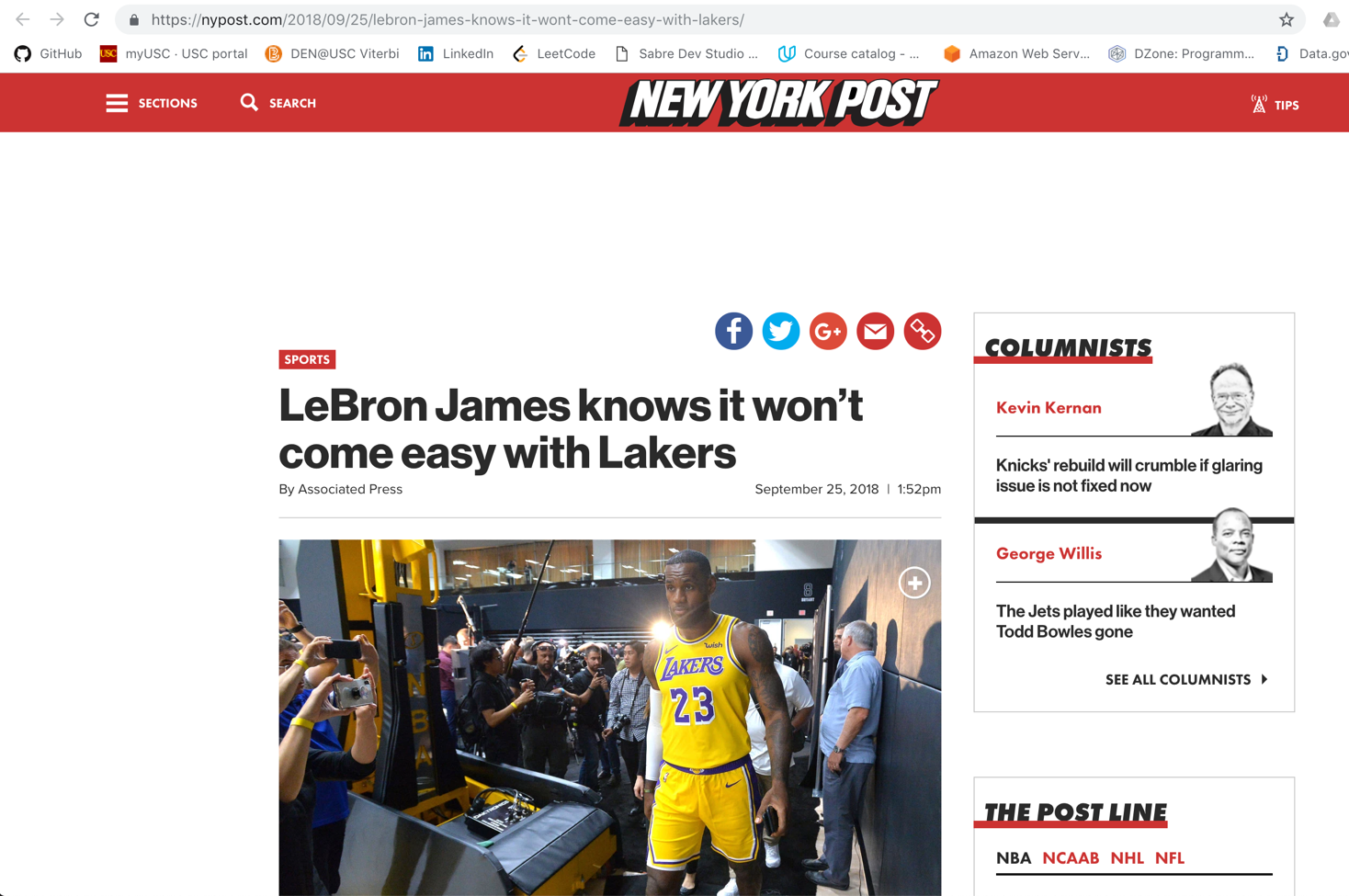
1. **Solr Lucene Results:**



Closer Look at one of the results



1. **Actual Webpage** (from La Lakers results)

****

**Query Response Overlap**

|  |  |  |
| --- | --- | --- |
| # | Query | No. of Overlaps |
| 1 | Donald Trump | 0 |
| 2 | LA Lakers | 0 |
| 3 | Star Wars | 0 |
| 4 | Lebron James | 0 |
| 5 | 2018 World Cup | 0 |
| 6 | North Korea | 0 |
| 7 | Hurricane Florence | 0 |
| 8 | Paul Allen | 2 |