Steps to complete the assignment

- 1. Download solr-6.5.0 and start its service using bin/solr start
- 2. Create core using bin/solr create_core -c nytimes
- 3. Post all the files so that solr can index the same using bin/post –c nytimes <path of the downloaded files>
- 4. Change the solrconfig.xml in <corename>/data folder to uncomment name="'df" parameter to query the field _text_
- 5. Now run the java code to generate the edgeList.txt file (requires jsoup jar)
- 6. Once the edgeList is generated, run the python file which uses networkx to generate graph object and the pass this object to pagerank function with the following parameters
 - a. alpha=0.85, personalization=None, max_iter=30, tol=1e-06, nstart=None, weight='weight',dangling=None
- 7. This will generate the external_PageRankFile.txt which includes the pagerank score for each html file
- 8. Place this pageRank file into <corename>/data folder
- 9. Change the managed-schema and solrconfig.xml to incorporate changes for pagerank indexing
- 10. Once done, reload the core from solr admin page
- 11. Now you can query pagerank indexed values by typing pageRank desc in sort column.
- 12. Now for the assignment, create a php page with a text box for search and radio button for index option (Lucene and PageRank)
- 13. Download the code from git to connect php and solr service and change the additionalparameters array to incorporate sort value in case of pageRank.
- 14. Host this on apache by placing the php file in /var/www/html/ and can be accessed using **localhost/solrPage.php**

PageRank Explanation

Some of the queries will have higher pageRank even though the search words do not exsist in the document. This is because pageRank works on the principal of incoming URLs. i.e If the number of URLs pointing to a page is more, then this indicates that this page is of more importance as many pages are pointing to this particular page and hence assigns a higher score to this page rather than actual page having more key words of the entered search

Flow of the assignment search page:

You go the url using the following address: localhost/solrPage.php

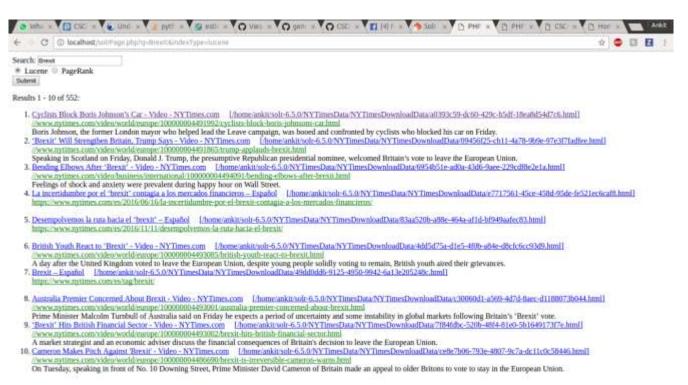
This provides you with a form with one search textbox and two radio button options.

- 1. One for Solrs default indexing (Lucene)
- 2. One for pageRank indexing(PageRank)

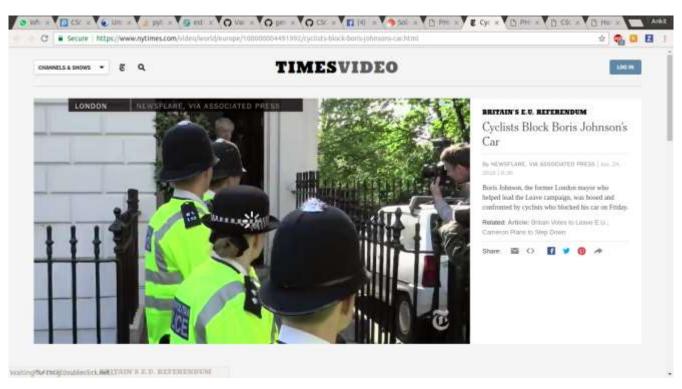
On clicking search, it will provide you with the top 10 results indexed according to the option of indexing you choose

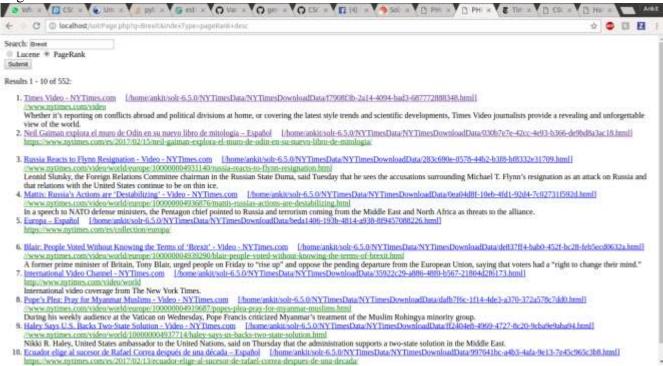
The results includes the tittle, ID, URL and description of the pages returned

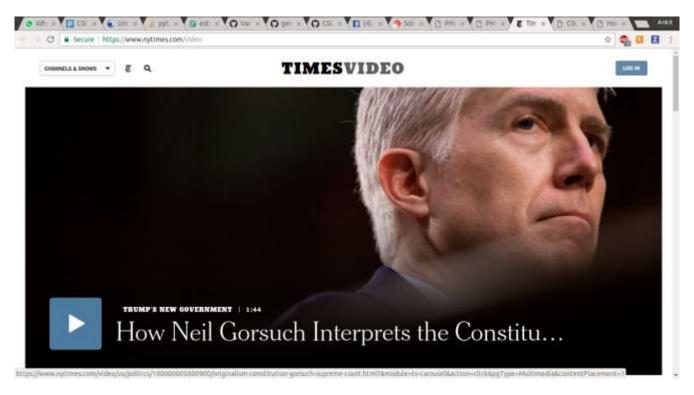
Eg: QUERY1 Brexit Lucene



On clicking the first URL, the following web page opens

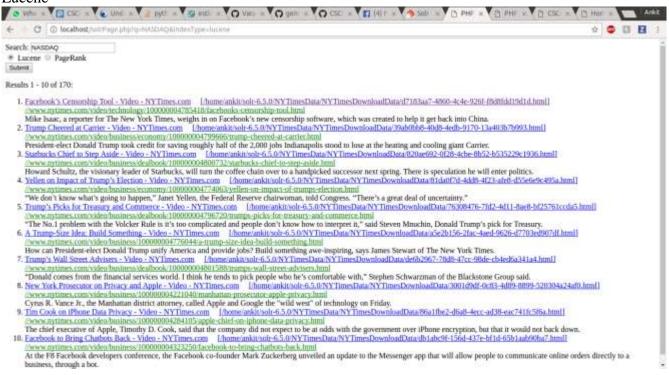


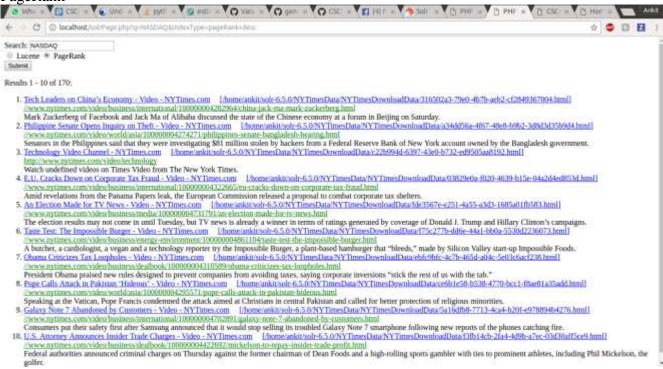




QUERY 2 NASDAQ

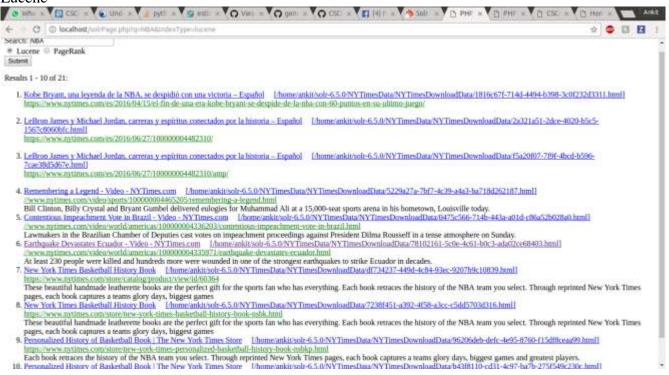
Lucene

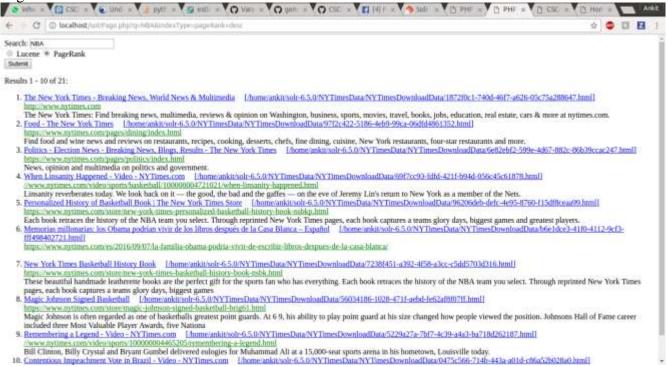




QUERY 3 NBA

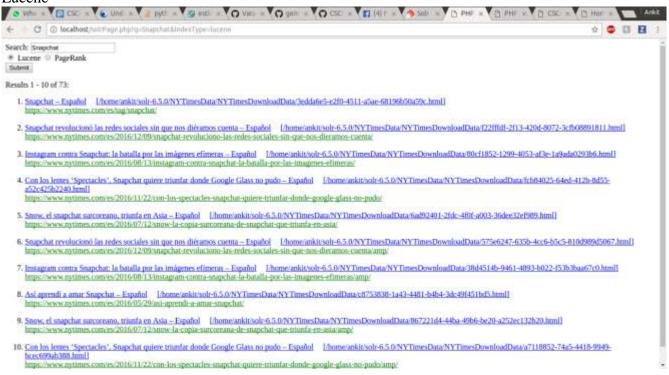
Lucene



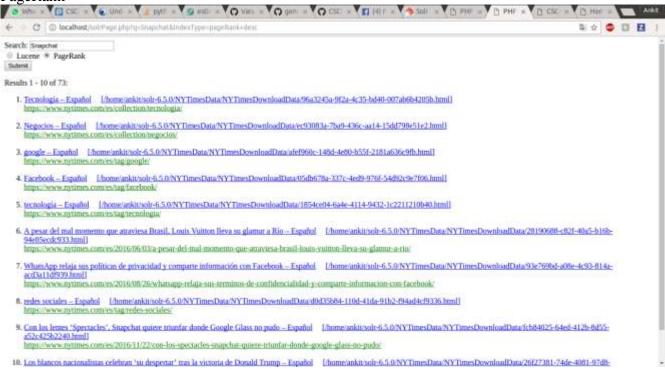


QUERY 4 Snapchat

Lucene



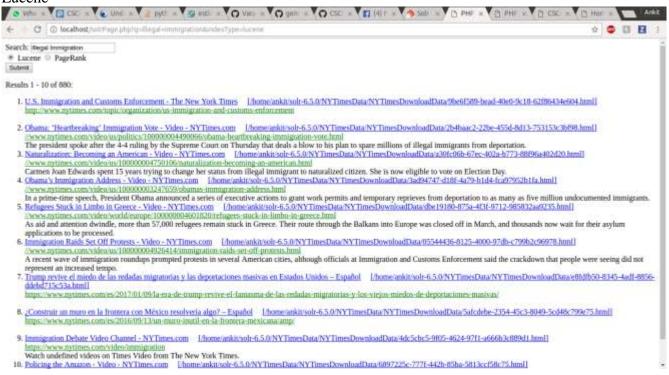
PageRank



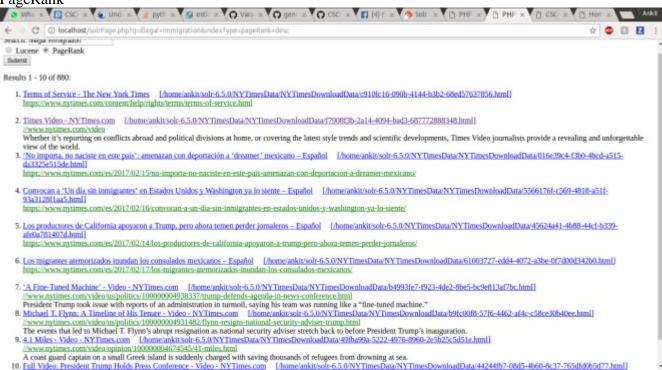
Overlap 1

QUERY 5 Illegal Immigration

Lucene

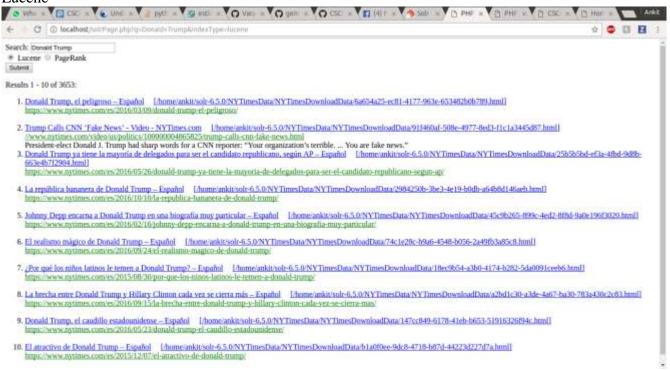




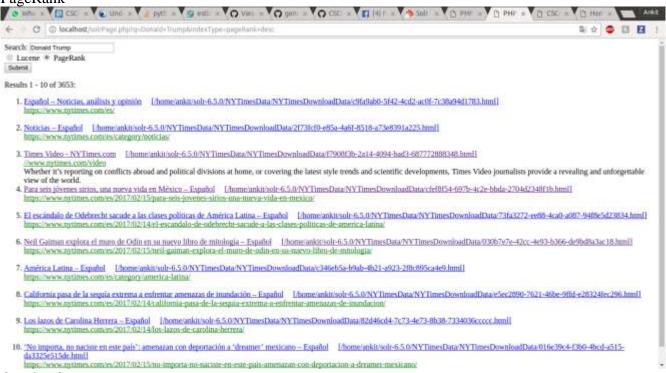


QUERY 6 Donald Trump

Lucene



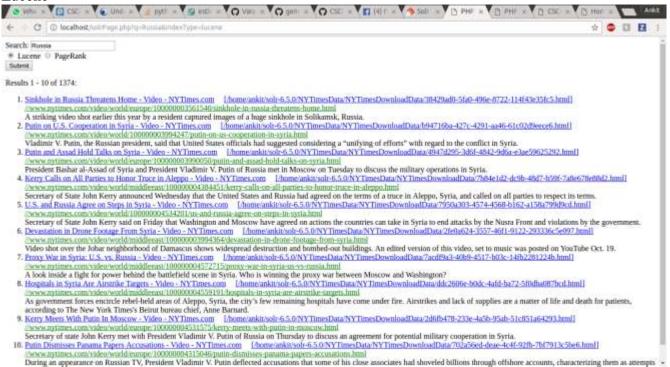
PageRank

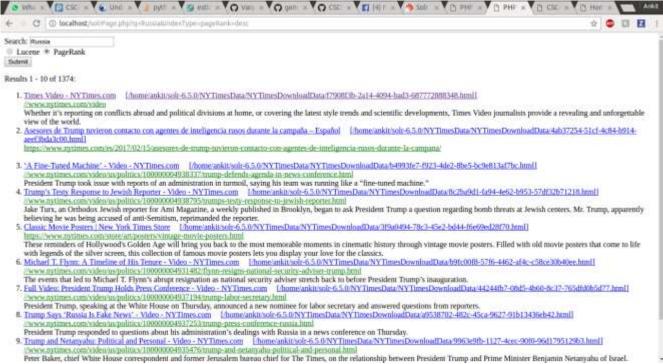


Overlap 0

QUERY 7 Russia

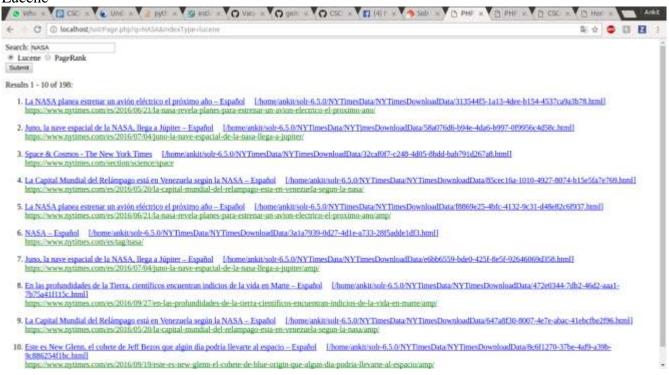
Lucene

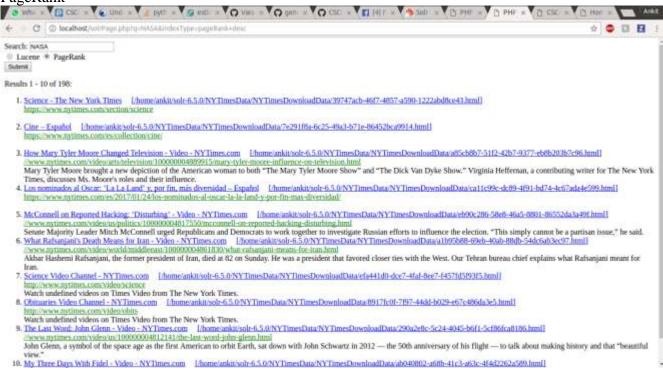




QUERY 8 NASA

Lucene





Overlap Bar Graph

