CSCI 572 HOMEWORK 4 : INDEXING THE WEB USING SOLR

Name: Shreyansh Kakadiya

USC ID : 8056685392

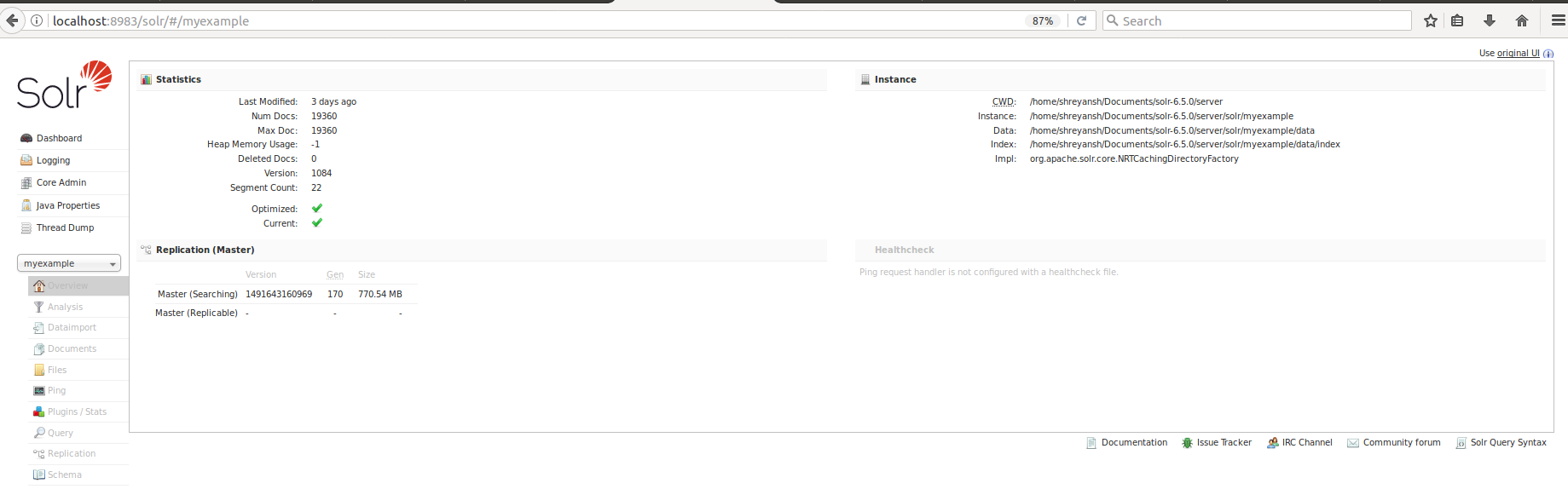
Website Indexed: NBCNEWS

Step 1 : Solr Installation

* I downloaded solr and for the website as suggested in the guidelines.
* I started solr using **bin/solr start**
* I then created a core using **bin/solr create -c myexample**
* Solr was seen on the web browser at port 8983 – **localhost:8983/solr/#**

Step 2: Indexing the Web-Page

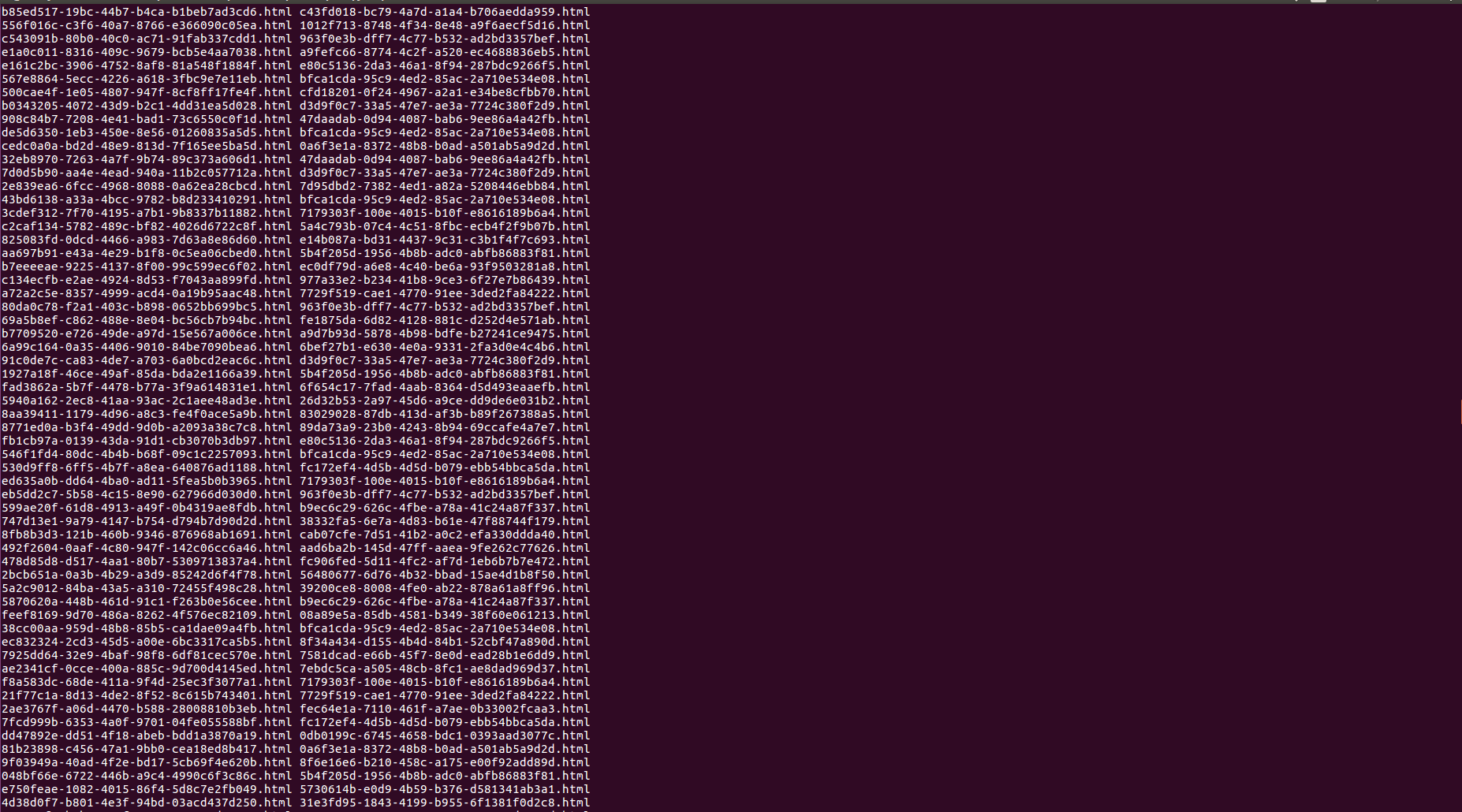
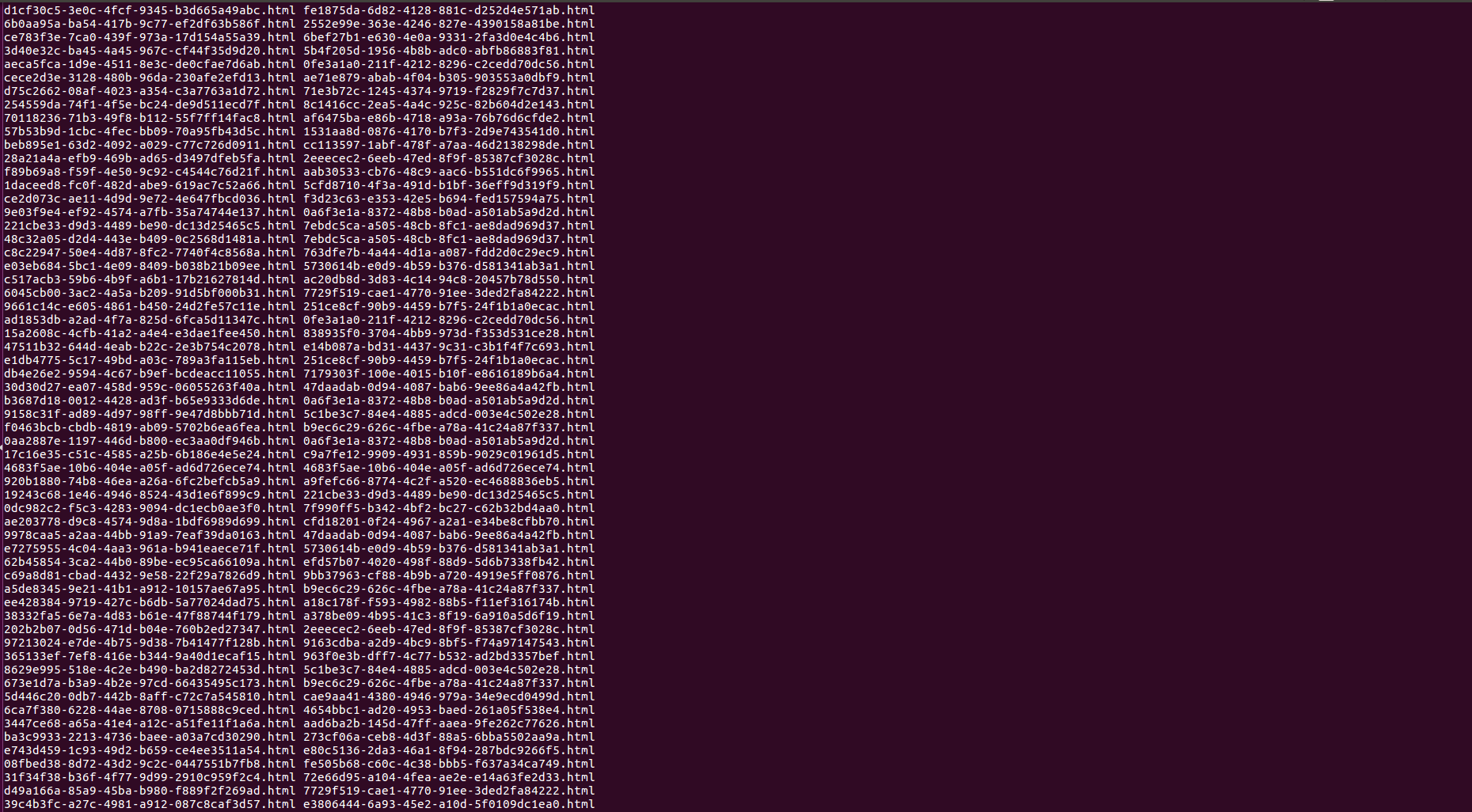
* I downloaded the data from the Google Drive for NBCNEWS
* Bin/post -c myexample -filetypes html /home/shreyansh/Documents/crawl\_data/NBCNewsData/NBCNewsDownloadData - did the indexing.



Step 3 : PageRank Algorithm:

* I downloaded Jsoup from the Web and wrote a Java file List.java to find the edgelist for the websites.

The following is a edgeList snippet :

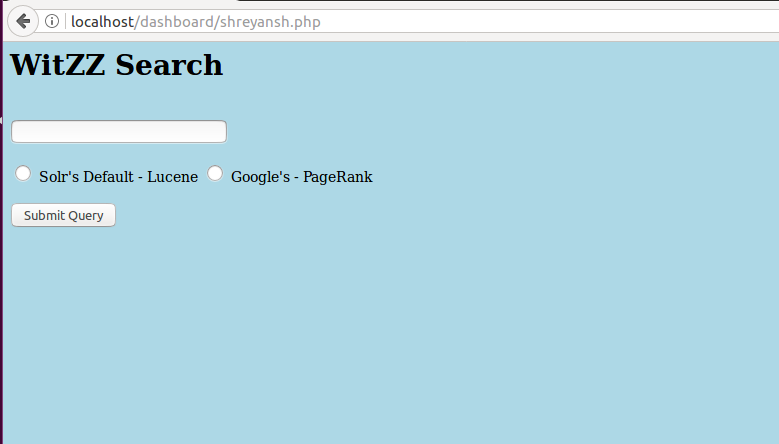


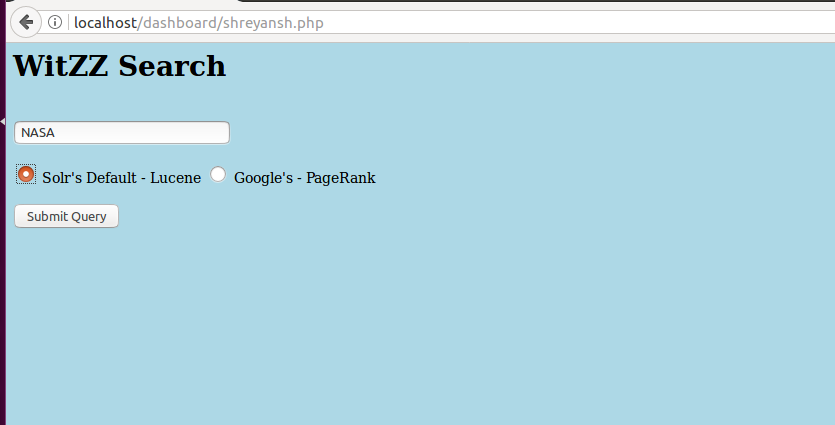
* The I used Networkx to read the graph and output a pagerank file name external\_pageRankFile.txt as given in the folder.
* PageRank Factors : alpha=0.85, personalization=None, max\_iter=30, tol=1e-06, nstart=None, weight='weight',dangling=None
* Then I copied the file into the data folder of my core.

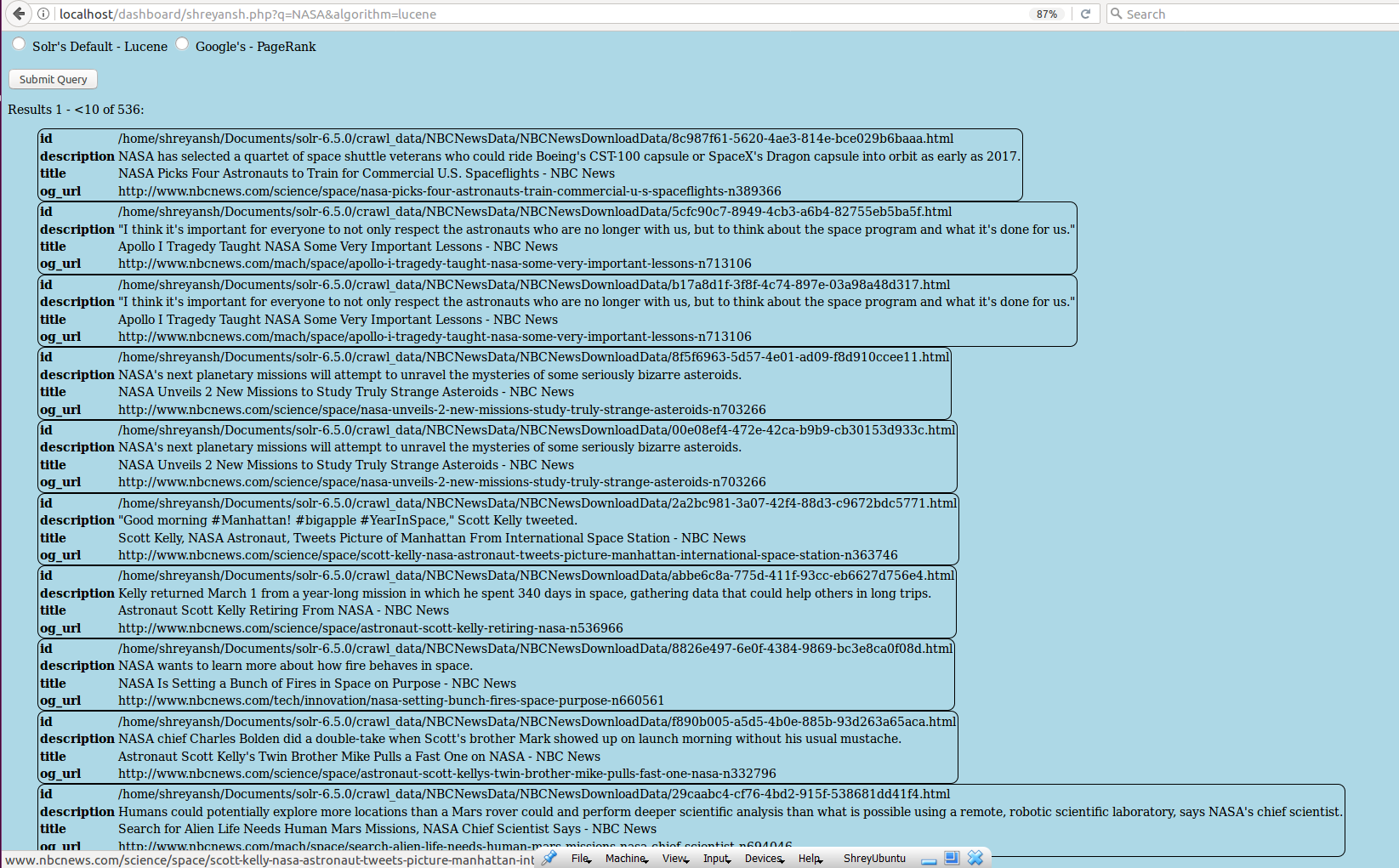
**The Following Snippet Shows a working model of my code :**

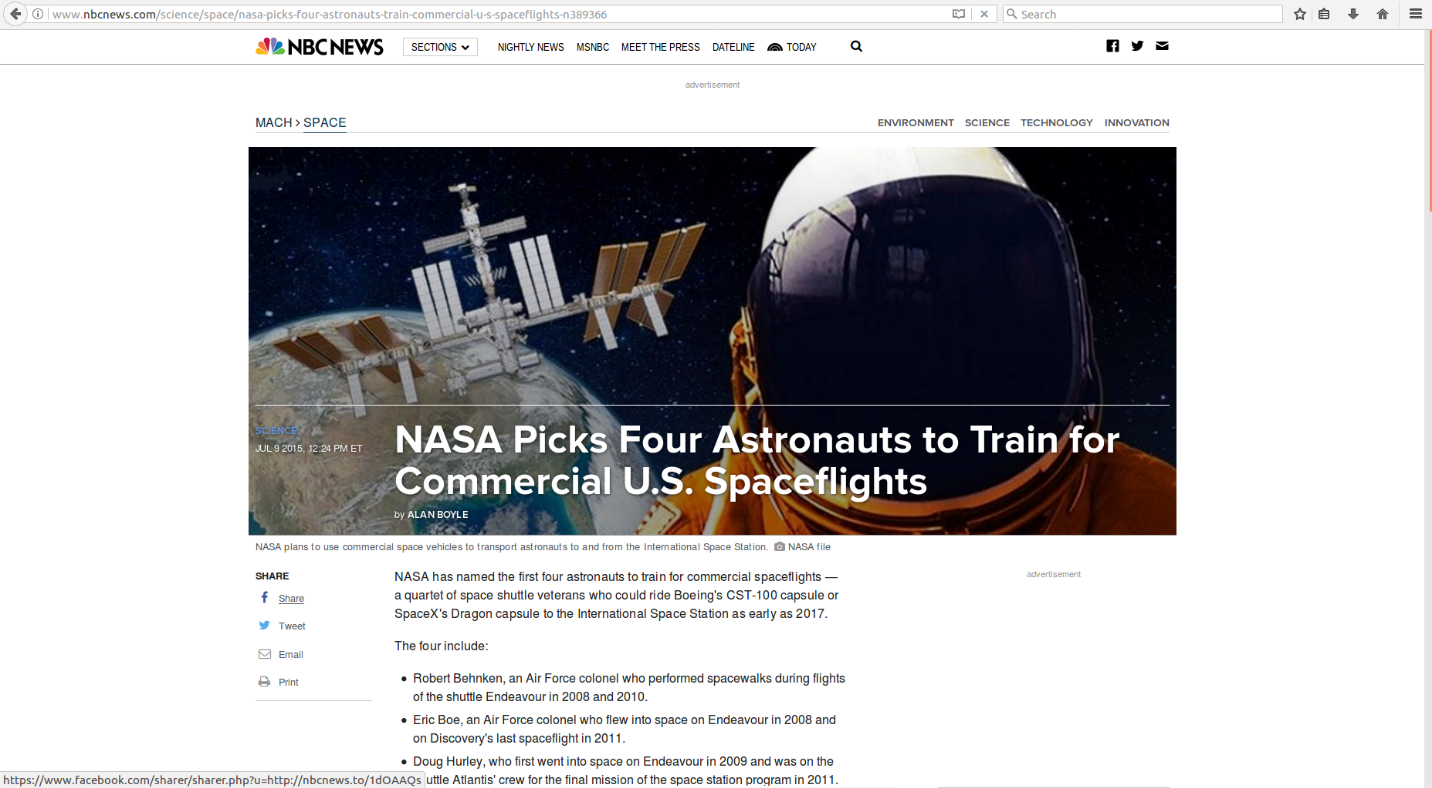
1. Steps in Query

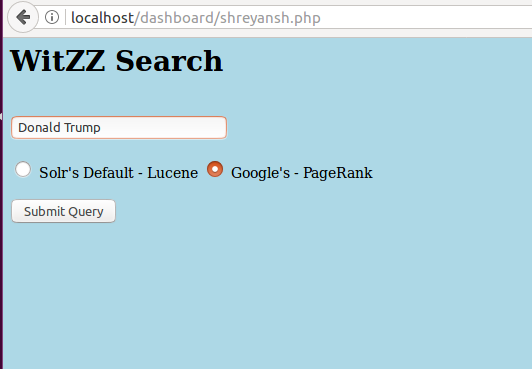
Shreyansh.php gives the GUI on the localhost for querying solr. After querying, we can select the algorithm to be executed – Lucene or PageRank. The results are then displayed on the same page. These results are clickable links that can be routed to the destination link. The following snippets shows the process of Querying solr.

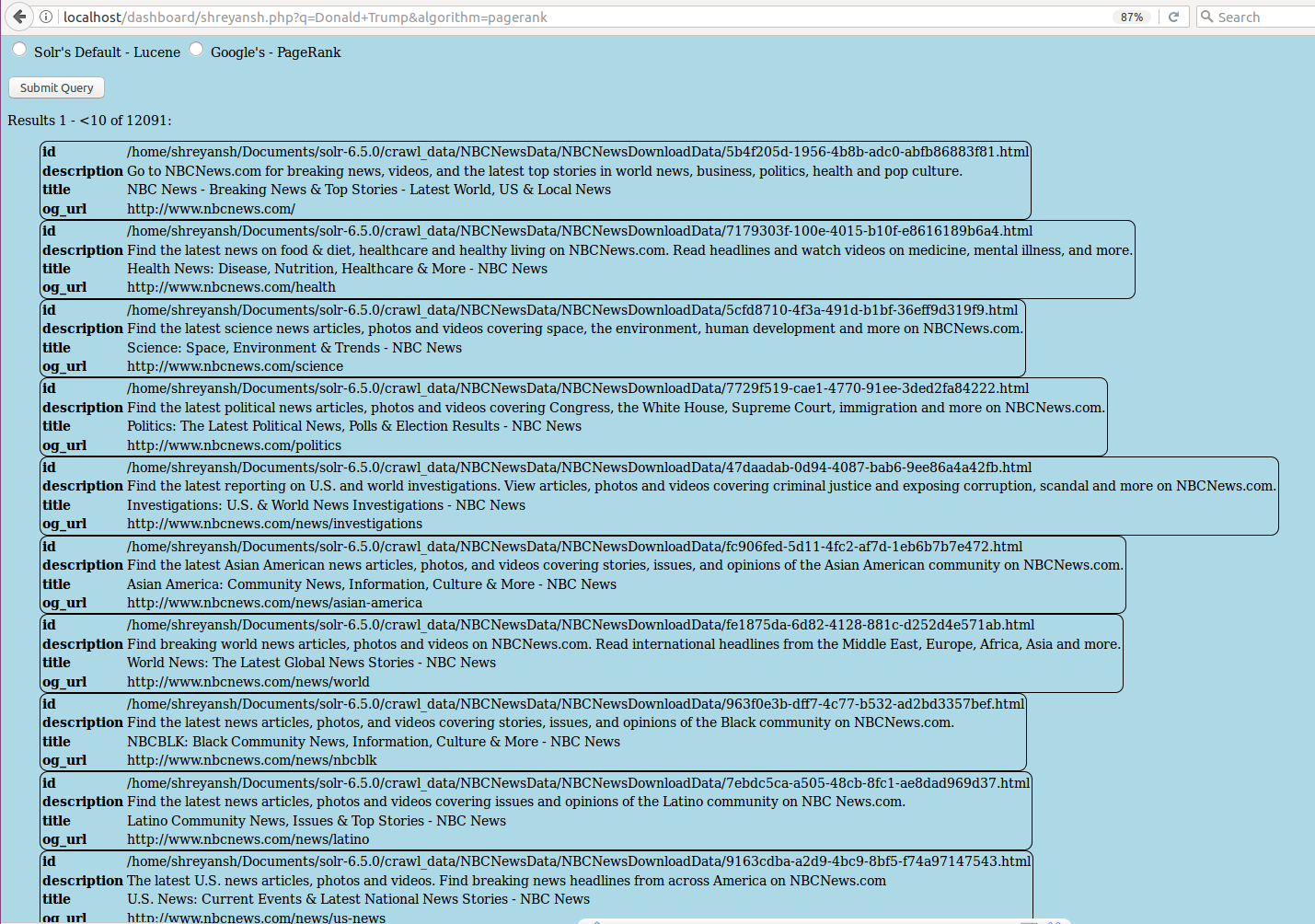












1. Why Some Pages Have Higher Page Rank Values

* PageRank computes weight for each page based on the number of links that points to the page. Hence, In the assignment we create a graph that decides the list of edges pointing the page.
* This graph is then used to find the importance of any given page in the web. Solr algorithm uses only the Boolean and vector model that just scans the tf-idf value and outputs the results.
* However, PageRank determines the weight of the page and then outputs it in a required order. Thus, these results follow a particular ranking and are different.
* Thus the major reason that pagerank has higher page rank value is the fact that it analyses the weight of page on the number of links that point it.

**The Graph of OverLap:**

|  |  |
| --- | --- |
| **Query** | **Overlap** |
| Brexit | 0 |
| NASDAQ | 0 |
| NBA | 0 |
| Snapchat | 0 |
| Illegal Immigration | 0 |
| Donald Trump | 0 |
| Russia | 0 |
| NASA | 0 |

**The results of the 8 Queries can be found in ResultScreenShot.doc**

**Conclusion:** Thus, we implemented Solr with Lucene and PageRank and compared the results.