1. followed by the instructions, I installed the Solr on Windows, then use the below commands to index the LA time’s pages

*solr.cmd start*

*java -jar -Dc=myexample -Dauto example\exampledocs\post.jar D:\LATIMES\latimes\\**

2. I installed the *Wampserver64* as the environment for *PHP* and download the *solr-php-client* and move the folder into the path: *wampserver\www\Apache\Solr*.

3. I copied and modified the *PHP* code in the instruction so that the server is able to communicate with the Web Browser and *Solr*, and catch the descriptions, titles and URLs in the file. The *CSV* file provides a mapping between files and URLs. Except a normal query box, an additional button allowed the user can switch the strategies for web search.

4. I downloaded the *jsoup-1.13.1.jar* and added it into the library of the project for creating the edge list. After I followed by the code in the *Appendix* part, and added the functions like reading files, writing files, deduplicating into the project and ran the project, I got the *edgelist.txt* file.

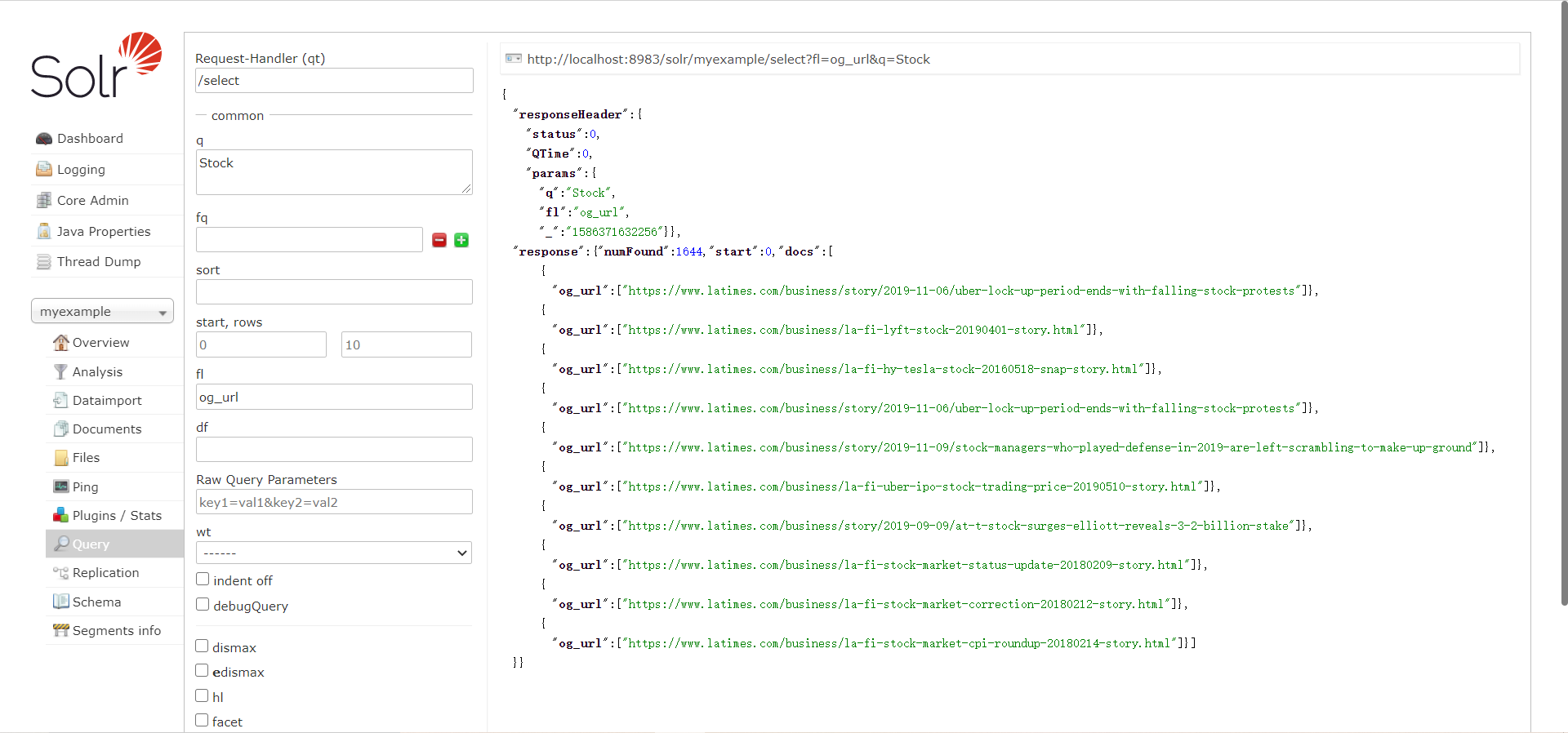
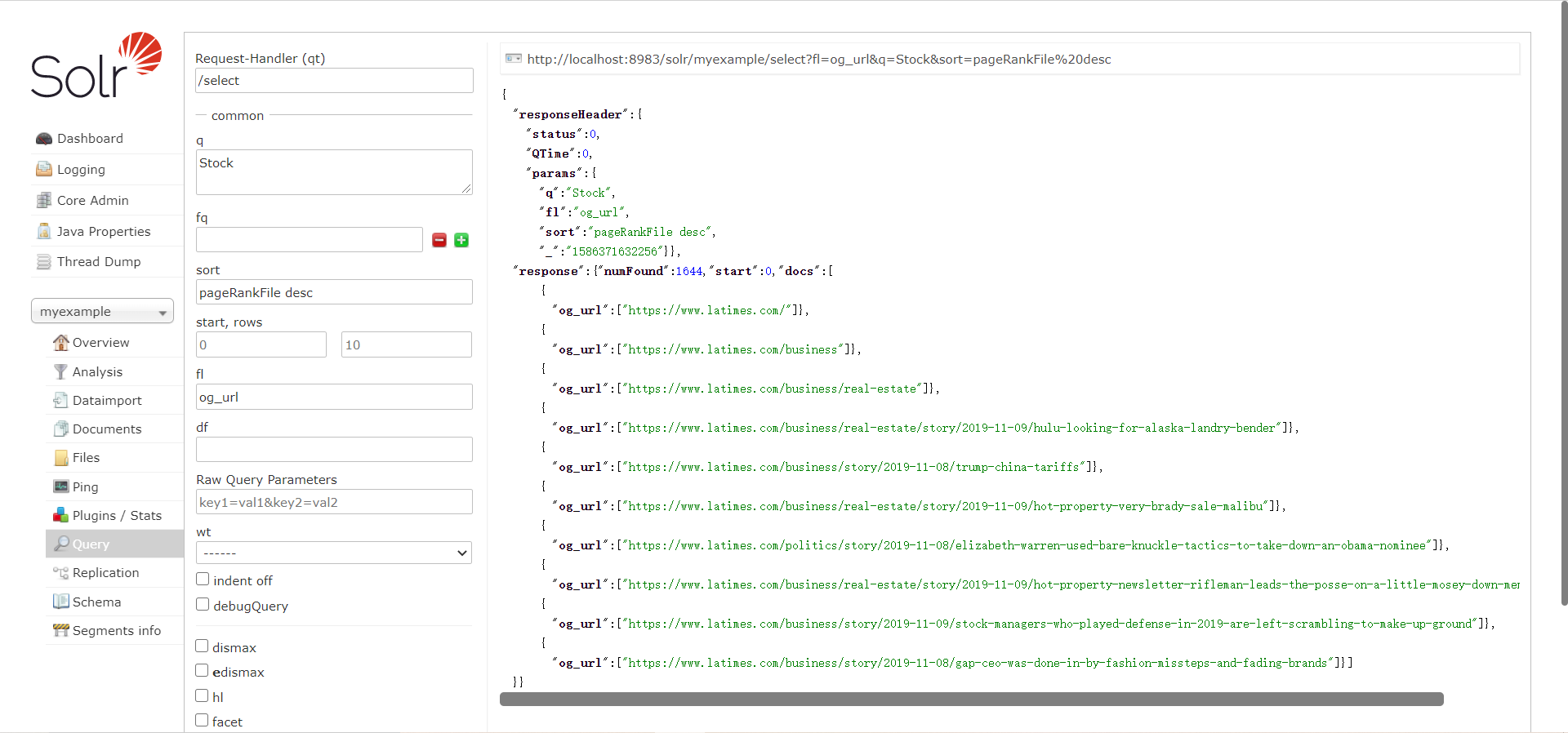
5. I imported the *Networkx* library in Python, used read\_edgelist function for transferring the *edgelist.txt* to a *Networkx* graph. Then I used *pagerank* function to calculate the score of page rank for every downloaded web page. The parameters used for *pagerank* function are:

*page\_rank = nx.pagerank(G,alpha=0.85, personalization=None, max\_iter=30, tol=1e-06, nstart=None, weight='weight',dangling=None)*

Damping parameter for pagerank *(alpha*) equals to 0.85, a uniform distribution is used (*personalization*), the maximum number of iterations (*max\_iter*) is 30, error tolerance (*tol*) equal to 10-6, starting value (*nstart*) of PageRank iteration for each node is None, edge data key (*weight*) to use as weight, and dangling nodes (*dangling*) are given outedges according to the uniform distribution.

6. By placing the file into the right folder, modifying the *managed-schema* and *solrconfig.xml file* and reloading the core, the search results can be displayed with the pagerank score.

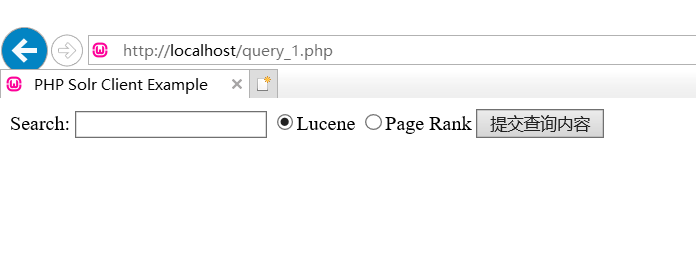
7. I ran the set of 8 queries with and without pagerank scores, saved the top ten results

for each of the eight queries and saved the data in the *results.csv* file.

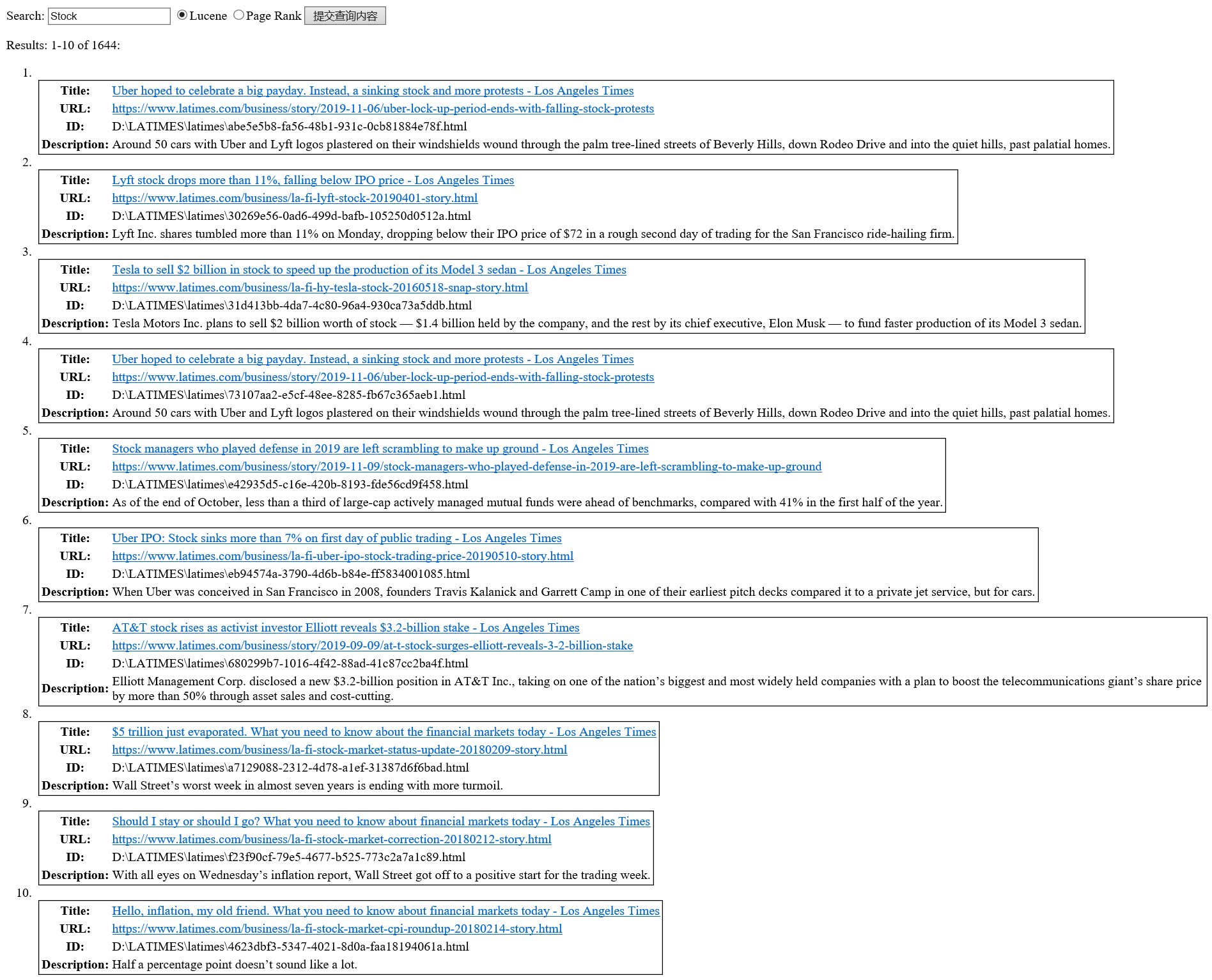
8. Finally, I used the tools provided by the *Excel* to find the overlap results between the two sets of results and drew the overlap graph showing the amount of overlap for each query.

Some pages have higher page rank score because they have more chances selected by a person who randomly clicking on links. There are two reasons lead to the higher probability: 1. The number of links pointing to that page are larger than other pages or the number of outgoing links of that page is less than other pages. 2. that page is closer to the homepage, which is much more important than other pages, than other pages.

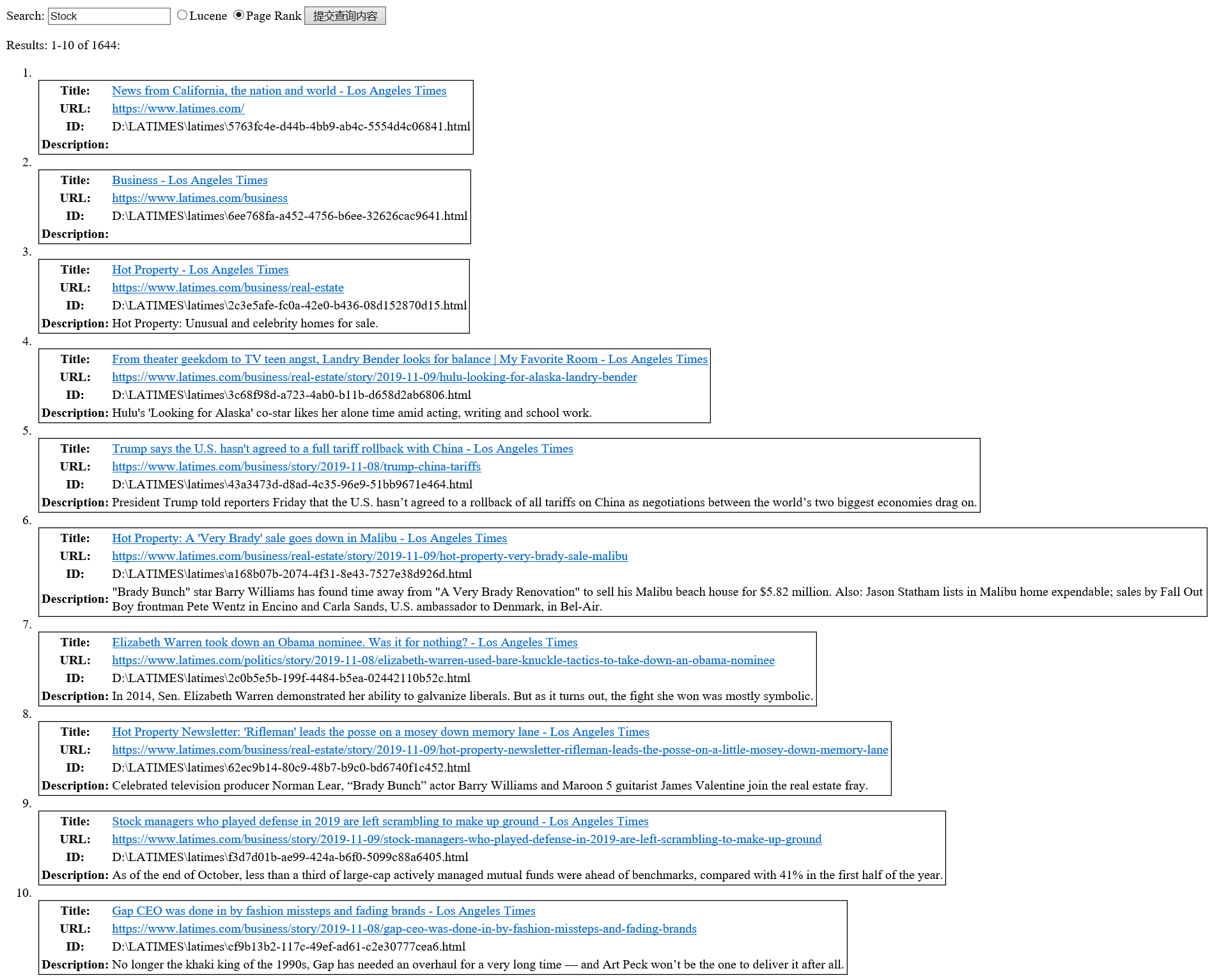
The Pagerank algorithm has nothing to do with whether the content is related to the term we search, but only shows the probability of we arrive to that page. Therefore, the default Lucene algorithm presents better search results than Pagerank algorithm



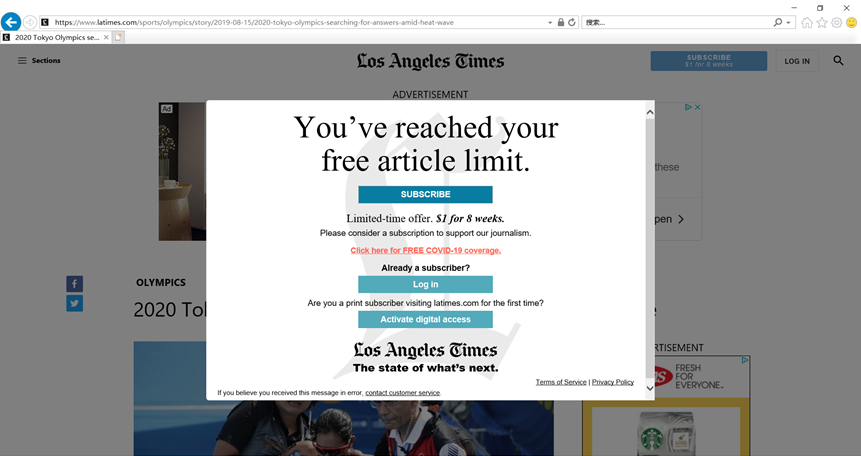
Screenshot of initial page where you enter the query in input box



Screenshot of page which shows the top ten results produced by the Lucene

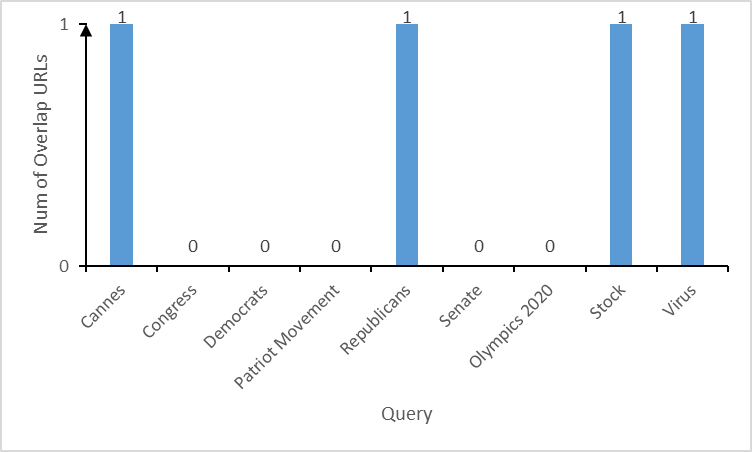


Screenshot of page which shows the top ten results produced by the PageRank



Screenshot of the actual webpage that opens when clicking on one of the result’s link

|  |  |
| --- | --- |
| **Query** | **Overlap URL** |
| Cannes | https://www.latimes.com/topic/cannes-film-festival |
| Congress | No Overlap URL |
| Democrats | No Overlap URL |
| Patriot Movement | No Overlap URL |
| Republicans | https://www.latimes.com/opinion/story/2019-11-08/impeachment-trump-republicans-toomey |
| Senate | No Overlap URL |
| Olympics 2020 | No Overlap URL |
| Stock | https://www.latimes.com/business/story/2019-11-09/stock-managers-who-played-defense-in-2019-are-left-scrambling-to-make-up-ground |
| Virus | https://www.latimes.com/espanol/internacional/articulo/2019-11-09/monos-se-volvieron-inmunes-tras-inocularles-virus-del-ebola-con-mutacion |



An Overlap Graph

<rankingResult_url.xlsx>