An analysis of two ranking ranking algorithms was conducted using the solr search platform where the default algorithm from Lucene (a combination of vector space and boolean models) was compared to a PageRank algorithm.

The code code was developed on an ubuntu machine and the application was developed and run on an apache web server on the same machine. A php backend was used to communicate with the solar system through the use of open source library and display results to a simple html page.

Analysis was conducted on a standalone local core with the indexed files being pre-scraped html files from mercurynews.com.

Summary of steps to complete the assignment:

1. Solr was setup on standalone.
2. Pages were indexed
3. Skeleton code provided was modified: CSV file was parsed an used to map to urls from html ids. Json output was parsed to filter out unrequired objects. Interface was extended with a checkbox and an additional parameter was conditionally added to the solr request to allow for PageRank query
4. The JSoup library was used to find outlinks from pages to other pages within corpus and to create an adjacency graph.
5. Using the Networkx python library and the adjacency graph, a page ranking was created. The main parameters used for this were: alpha: 0.85, max iterations: 30, and tolerance=1e-06.
6. Config files for standalone was modified to support the external rank list.
7. Queries were run and python script was used to compile similarities

Empirically, the default algorithm seems to provide much more relevant results when compared to page rank. This is most likely because the page rank of a page doesn't play as much importance to the relevance to the query. Rather, it relates to how much it's related to other pages in the corpus. Some pages have higher page rank because they have more pages in the corpus linking into them, perhaps, because they are found on the front page of the site or are on a list of recommendations. A quick glance shows some reoccurrence of PageRank results.

Overlap graph shows 2 overlaps. One for the query "Paul Allen" and another for the query of "LA Lakers". These overlaps were an "overlap of URL". This is perhaps due by the popularity of these topics.

Documentation of flow:

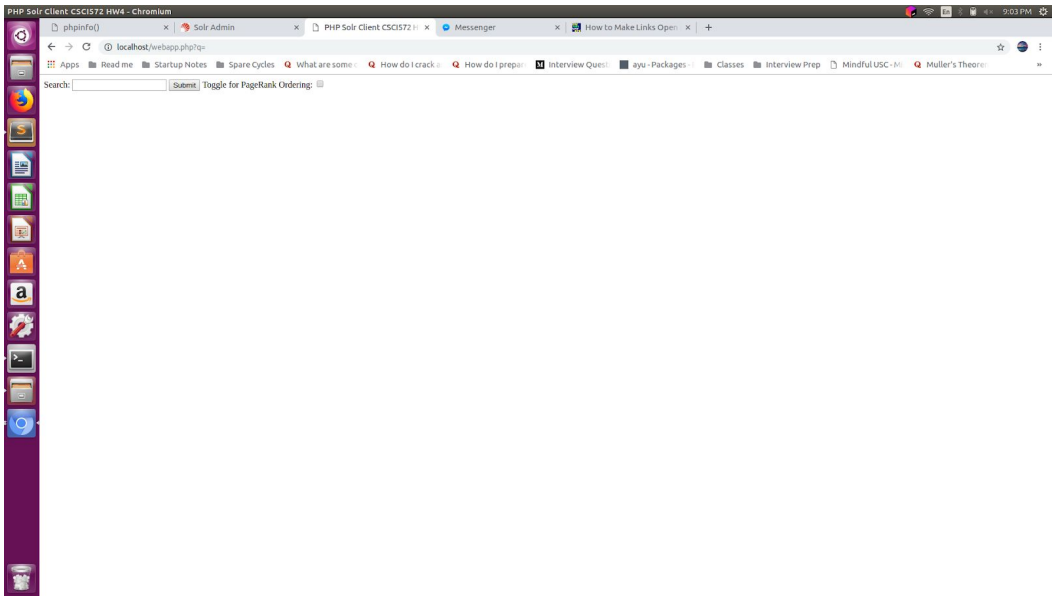

Figure 1. Initial page

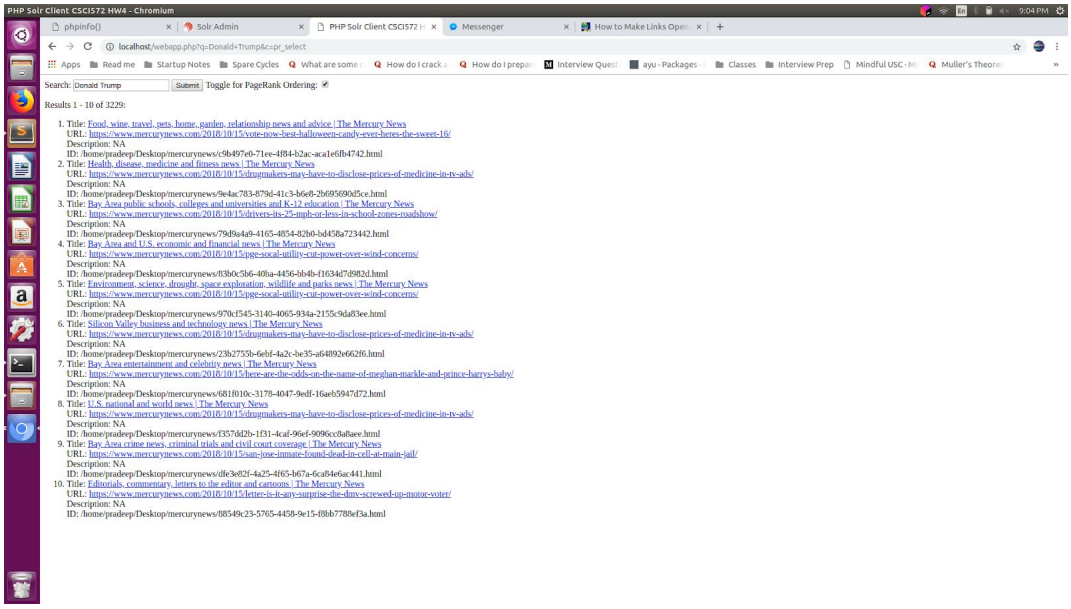

Figure 2. Page w/ PageRank

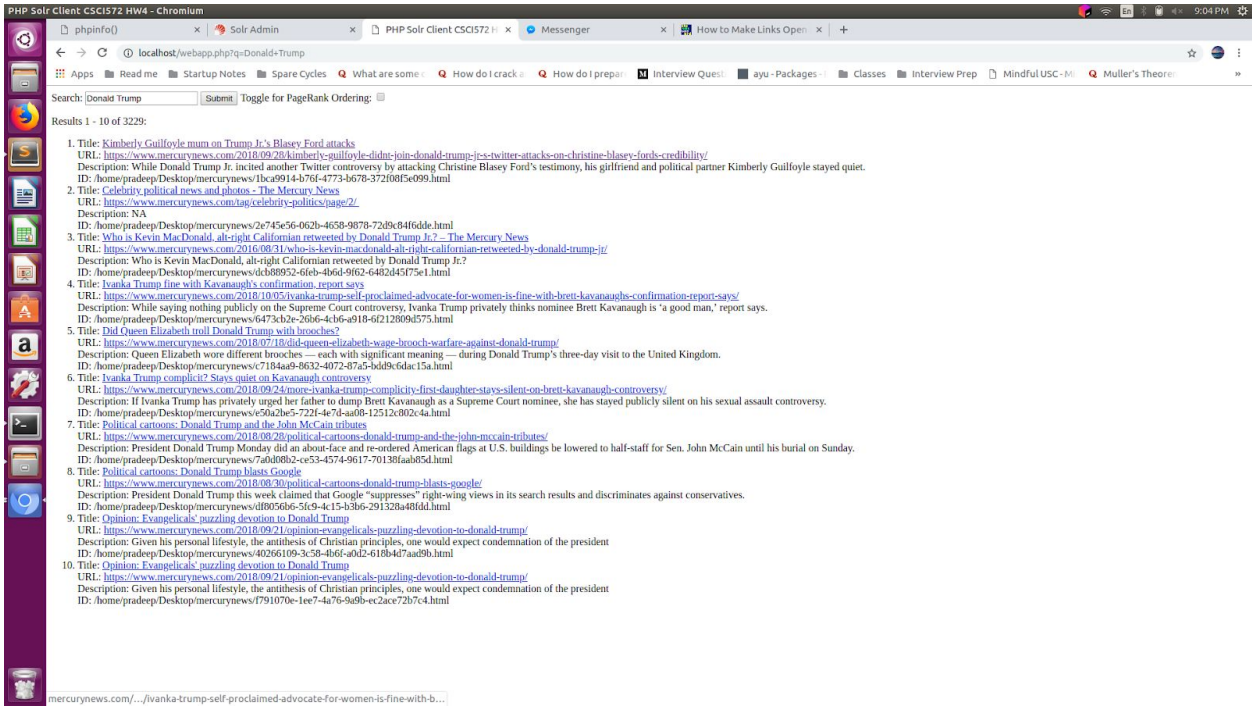

Figure 3. Page w/ Default

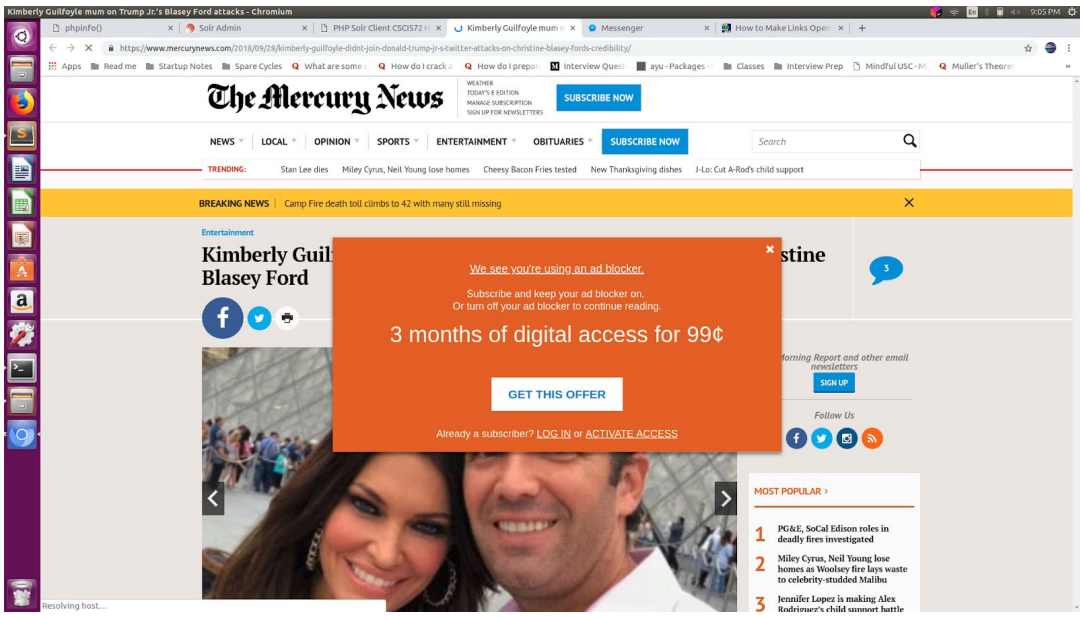

Figure 4. Webpage after click on link