Steps followed to complete the assignment (Summary):-

The HTML/PDF/DOC files that were crawled and saved in the previous assignment are used in indexing using Solr. The files are named with the URL encoded names from the downloaded respective URL's. To accomplish this, Solr was installed and configured to run on Ubuntu platform. Also, during crawling the HTML/PDF/DOC files that were crawled and saved were included in a separate .csv files along with their corresponding outgoing links. This file is used to compute the page rank and generate the page rank file, which in turn is used as an external page rank file in Solr. This is used in turn to compare the search results with and without the custom page rank approach. The file is suitably configured to be included in the relevant locations. A Client is designed in PHP to run queries, retrieve result and display them appropriately. Radio buttons are provided to run the queries with and without the page ranking approach. Queries from Assignment 1 is run to generate results on Solr, with and without the external page rank approach. A relevancy graph is drawn to compare these results and in turn to compare the results with the relevancy graphs of Google and Bing obtained from Assignment 1. Keeping the graphs in hand, the results are analyzed.

Computation of Page Rank:

A python program is written to compute the page rank. The pagerankdata.csv is read line by line and a networkx graph is constructed. The nodes in the networkx graph are the encoded URL's in order to maintain consistency with the indexed filenames used by Solr. While, constructing the networkx graph nodes, duplicate entries have been removed and same URL's except for the difference of https and http are consolidated to remove redundant entries by stripping http and https by regex match and pushing the URL's to a hash map. Later, attaching http or https as required. Also, the URL's outside the school are ignored. The edges between node A and node B is constructed if B is an out link from A. From the networkx graph, the page rank of each of the nodes is computed with default alpha value of 0.85. The page rank of the nodes are written into external_pageRankFile.txt in the format required for the Solr external file field. Also, the entries in the external_pageRankFile.txt are made sure to match with the id's that are used by Solr during indexing, by modifying the entries suitably. Later, Solr is configured to incorporate the external page rank file.

Note:

For Solr, the queries of Assignment 1 are slightly modified with ***'AND'*** in suitable locations of the query to obtain more relevant results because Solr uses a default Boolean model and return the results that are related to the school and have a higher page rank. But, we need the results that constitute all the words. Hence, the change. For example, when we query for "Larry Auerbach USC School of Cinematic Arts", Solr returns the top results related to USC School of Cinematic Arts instead the results related to Larry Auerbach within USC School of Cinematic Arts. This is because of default Boolean model and the pages related to USC School of Cinematic Arts have higher page rank.

The following table illustrates the difference in queries. Using these queries, it is found to give more relevant results compared to any other approaches followed in tweaking the default Solr behavior.

Assignment 1	**Solr queries (with and without page rank approach)**
Larry Auerbach USC School of Cinematic Arts	Larry Auerbach ***AND*** USC School of Cinematic Arts
Bruce A. Block USC School of Cinematic Arts	Bruce A. Block ***AND*** USC School of Cinematic Arts
Michael Bodie USC School of Cinematic Arts	Michael Bodie ***AND*** USC School of Cinematic Arts
John C. Hench Division of Animation and Digital Arts USC School of Cinematic Arts	John C. Hench Division of Animation and Digital Arts ***AND*** USC School of Cinematic Arts
Bryan Singer Division of Cinema & Media Studies USC School of Cinematic Arts	Bryan Singer Division of Cinema & Media Studies ***AND*** USC School of Cinematic Arts
Interactive Media and Games USC School of Cinematic Arts	Interactive Media and Games ***AND*** USC School of Cinematic Arts
School of cinematic arts USC map and directions	School of cinematic arts USC map and directions
USC School of Cinematic Arts Founder	USC School of Cinematic Arts Founder
USC School of Cinematic Arts Alumni	USC School of Cinematic Arts Alumni
USC School of Cinematic Arts Undergraduate degree requirements	USC School of Cinematic Arts ***AND*** Undergraduate degree requirements
USC School of Cinematic Arts Masters degree requirements	USC School of Cinematic Arts ***AND*** Masters degree requirements
USC School of Cinematic Arts PhD degree requirements	USC School of Cinematic Arts ***AND*** PhD degree requirements

This slight modification is called approach 1 from now on. The relevancy graphs in the analysis section also shows this. In approach 2, the modification is not made and the queries of assignment 1 is run. The relevancy graph for this also is shown in the analysis section. It is seen that approach 2 gives a very poor result. Hence, the modification is made. Using approach 1 or approach 2, Solr provides the same set of results with the default page rank behavior.

Solr-PHP client:

A PHP client is provided to query and retrieve the top results. Also, a radio button is provided to compare the results obtained with and without the external page ranking approach.

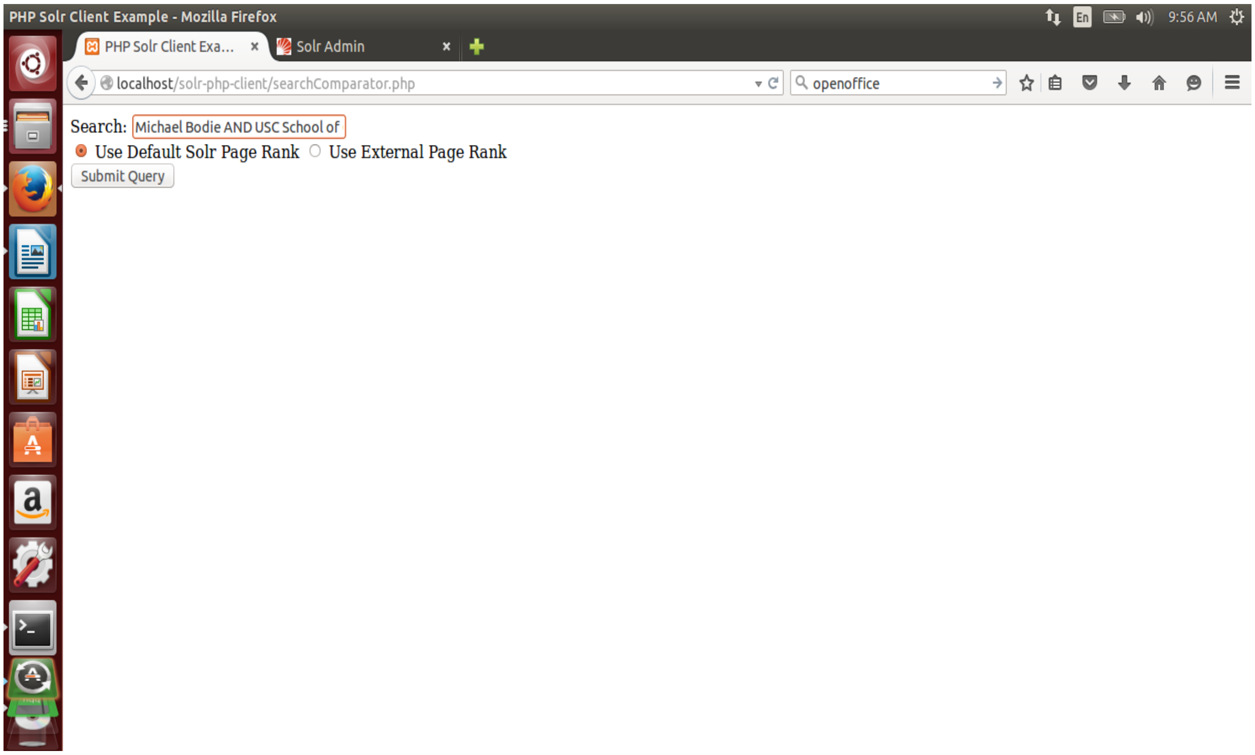
The PHP file searchComparator.php is provided. In this program, the default location of the Apache Solr client library is provided in initialization and the default host, port and webapp paths are provided during the creation of the service instance. The name of the core is “myexample”.

In order to run the PHP script in a different system, Solr has to be installed in the default location and the name of the core should be “myexample”. The name of the external page rank file should be external_pageRankFile.txt

The results of the query are displayed in the required format. The size of the file is converted to kB and the URL is handled to be provided as a clickable link that directs to the corresponding webpage.

The snapshots of the client and the results obtained with and without the external page ranking approach is provided for a single query. If, the “Document” is clicked, it leads the user to the related webpage.

1. Query page for fetching results with default Solr page rank

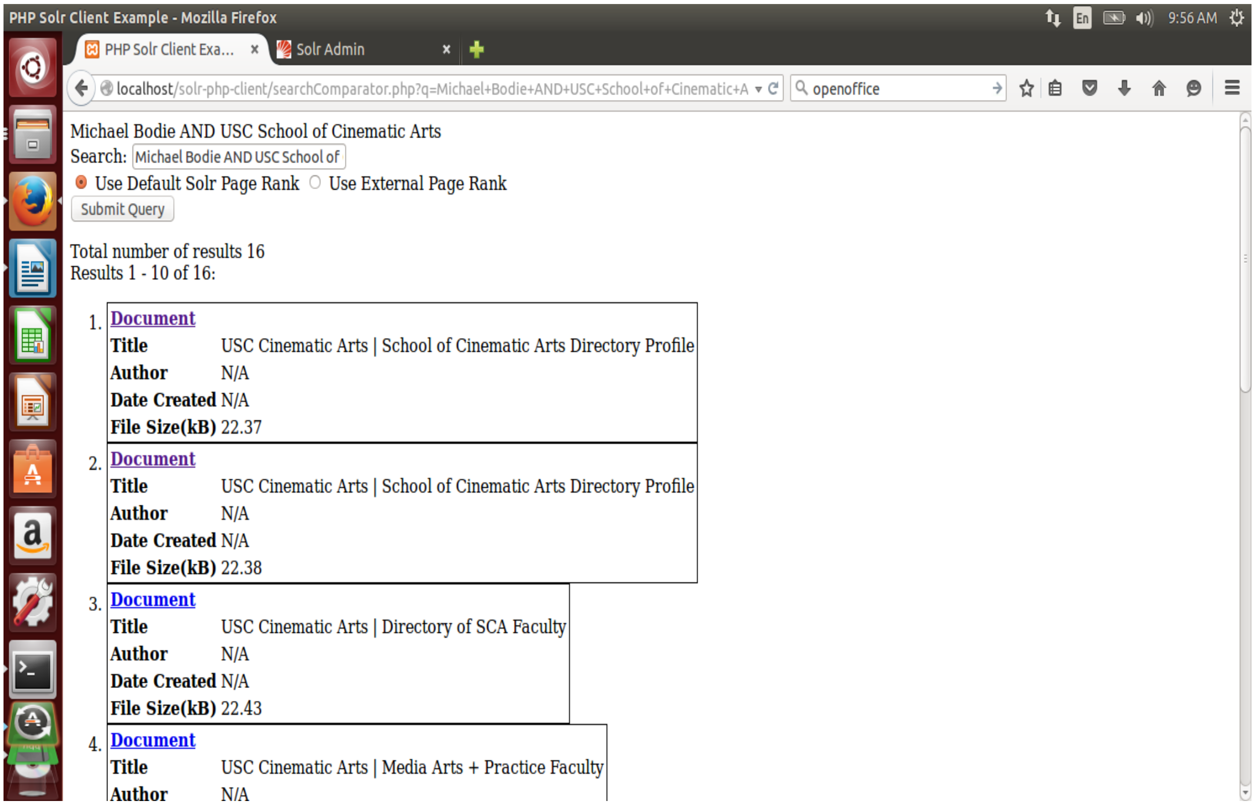


PHP Solr Client Example - Mozilla Firefox
9:56 AM
PHP Solr Client Exa...
Solr Admin
localhost/solr-php-client/searchComparator.php
openoffice

Search: Michael Bodie AND USC School of

● Use Default Solr Page Rank ○ Use External Page Rank

2. Result page for fetching results with default Solr page rank


PHP Solr Client Example - Mozilla Firefox
En
9:56 AM
PHP Solr Client Exa...
Solr Admin
localhost/solr-php-client/searchComparator.php?q=Michael+Bodie+AND+USC+School+of+Cinematic+A
openoffice

Michael Bodie AND USC School of Cinematic Arts
Search: Michael Bodie AND USC School of
● Use Default Solr Page Rank ○ Use External Page Rank
Submit Query

Total number of results 16
Results 1 - 10 of 16:

1. Document

Title	USC Cinematic Arts | School of Cinematic Arts Directory Profile
Author	N/A
Date Created	N/A
File Size(kB)	22.37

2. Document

Title	USC Cinematic Arts | School of Cinematic Arts Directory Profile
Author	N/A
Date Created	N/A
File Size(kB)	22.38

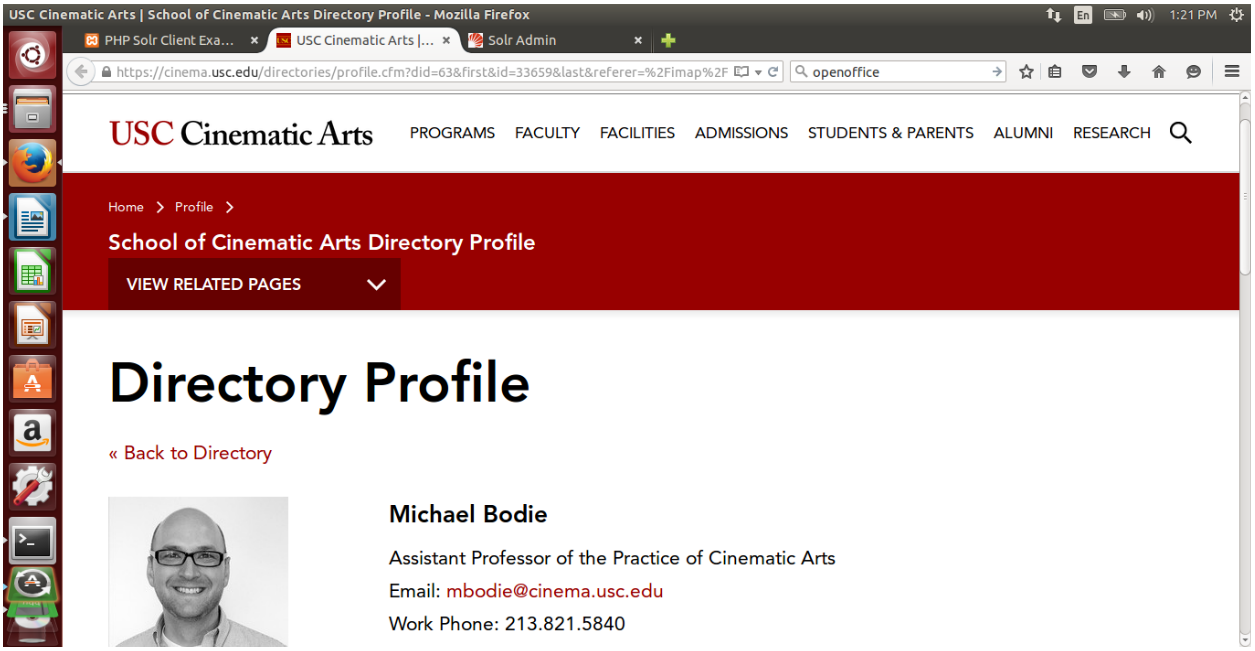
3. Document

Title	USC Cinematic Arts | Directory of SCA Faculty
Author	N/A
Date Created	N/A
File Size(kB)	22.43

4. Document

Title	USC Cinematic Arts | Media Arts + Practice Faculty
Author	N/A

3. Upon, clicking the first link, the following web page is opened.



School of Cinematic Arts Directory Profile

Directory Profile

« Back to Directory

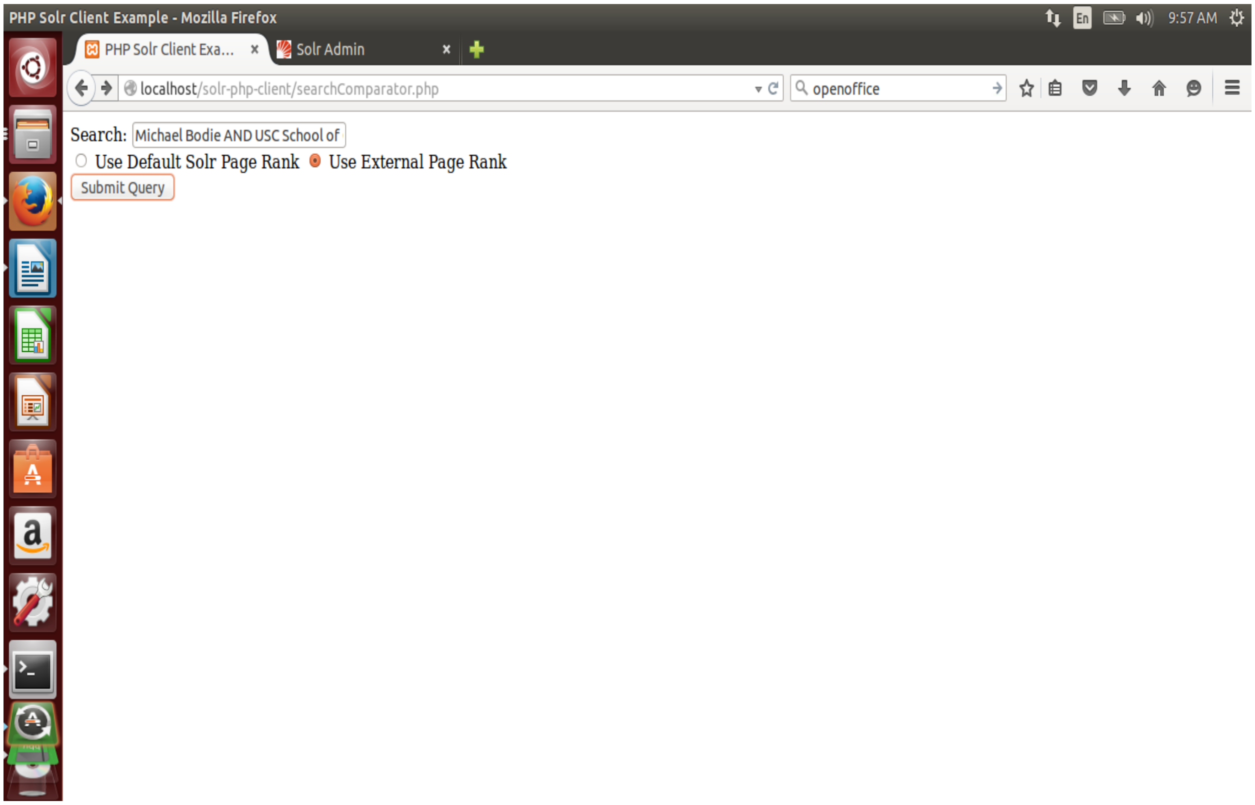
Michael Bodie

Assistant Professor of the Practice of Cinematic Arts

Email: mbodie@cinema.usc.edu

Work Phone: 213.821.5840

4. Query page for fetching results with external page rank approach

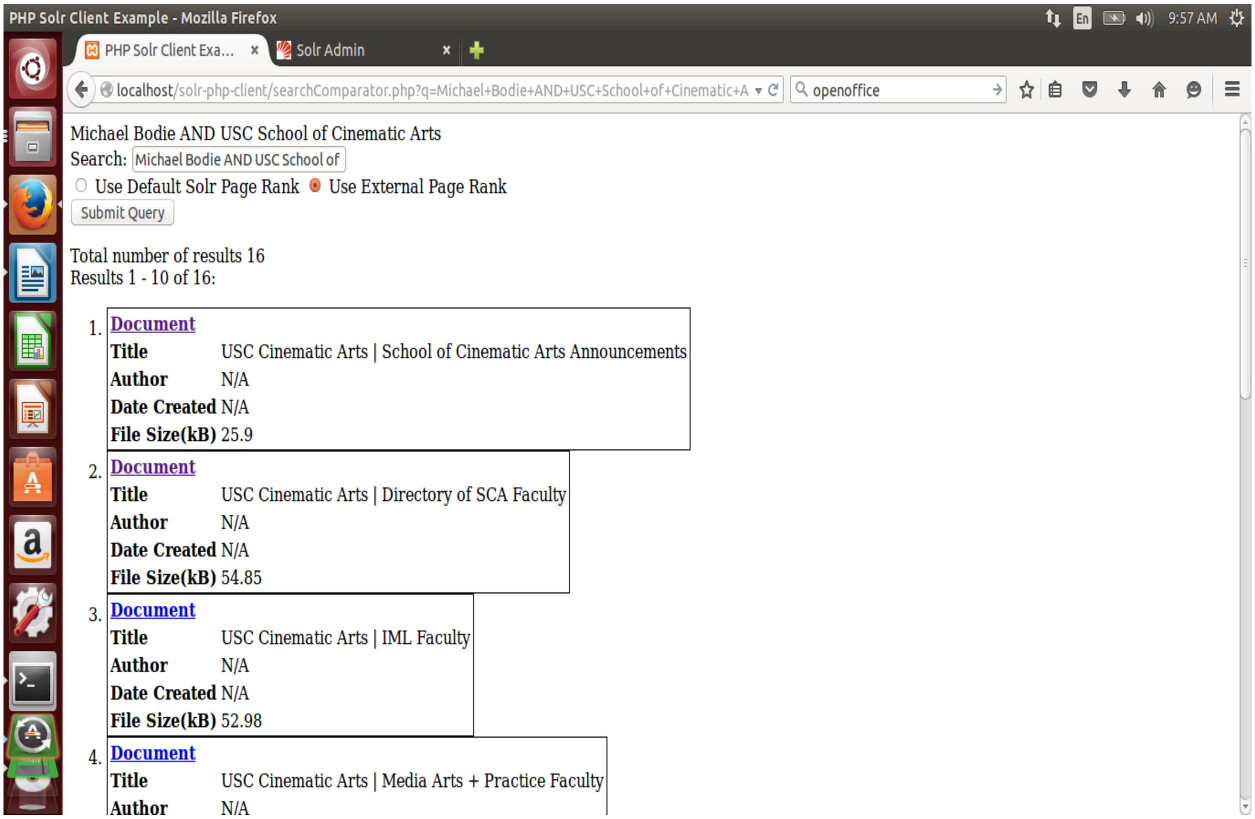


localhost/solr-php-client/searchComparator.php

Search: Michael Bodie AND USC School of

○ Use Default Solr Page Rank ● Use External Page Rank

5. Result page for fetching results with external page rank approach



PHP Solr Client Example - Mozilla Firefox
9:57 AM
PHP Solr Client Exa...
Solr Admin
localhost/solr-php-client/searchComparator.php?q=Michael+Bodie+AND+USC+School+of+Cinematic+A
openoffice

Michael Bodie AND USC School of Cinematic Arts

Search: Michael Bodie AND USC School of

○ Use Default Solr Page Rank ● Use External Page Rank

Submit Query

Total number of results 16
Results 1 - 10 of 16:

1. Document

Title	USC Cinematic Arts | School of Cinematic Arts Announcements
Author	N/A
Date Created	N/A
File Size(kB)	25.9

2. Document

Title	USC Cinematic Arts | Directory of SCA Faculty
Author	N/A
Date Created	N/A
File Size(kB)	54.85

3. Document

Title	USC Cinematic Arts | IML Faculty
Author	N/A
Date Created	N/A
File Size(kB)	52.98

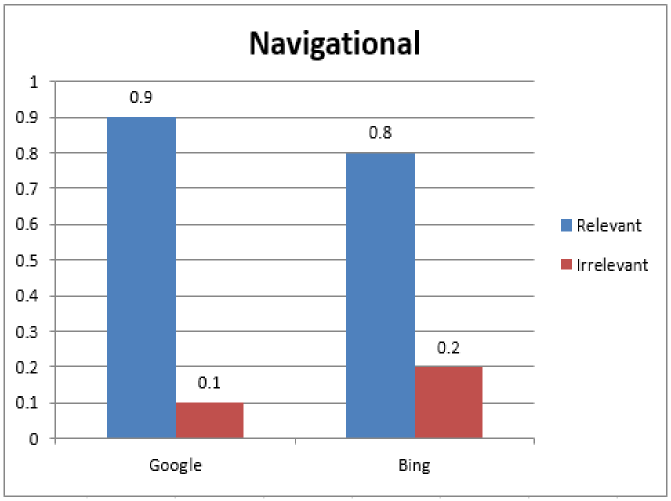
4. Document

Title	USC Cinematic Arts | Media Arts + Practice Faculty
Author	N/A

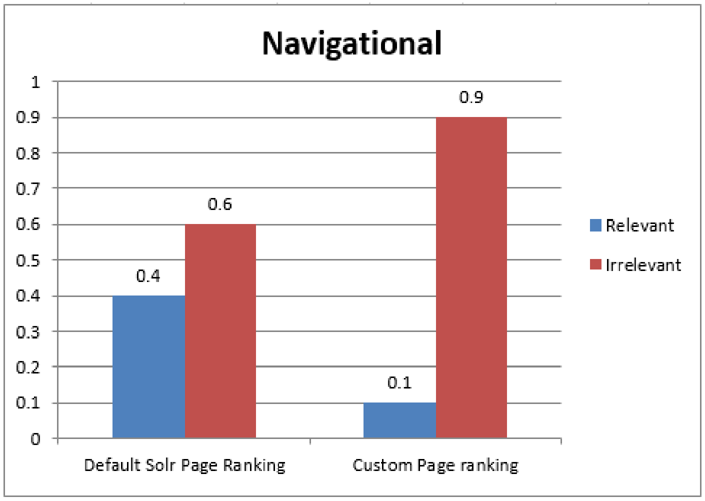
Analysis of the results:

The relevancy graphs for the two ranking algorithms compared against the graphs of HW1 are provided for both the navigational and informational queries.

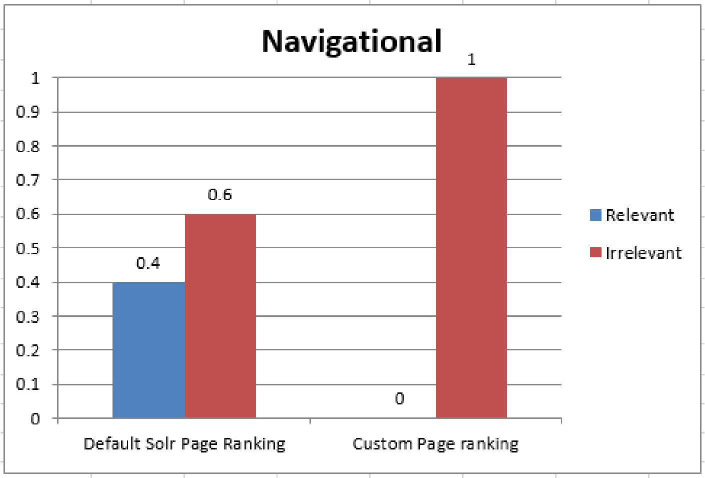
Navigational relevancy result for Google and Bing



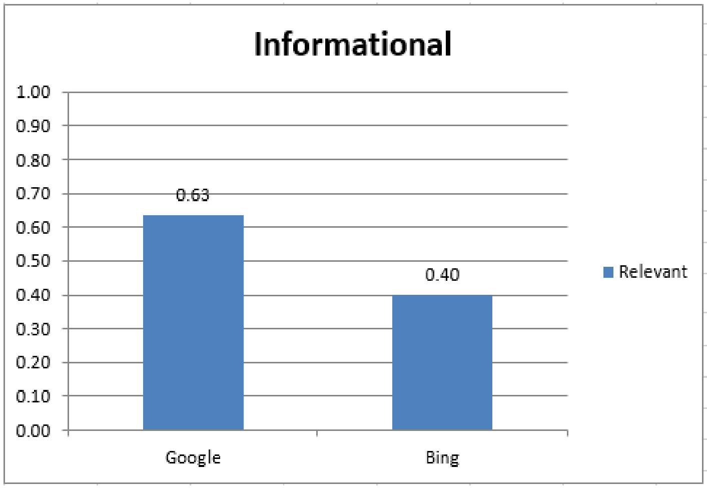
Navigational relevancy result for approach 1



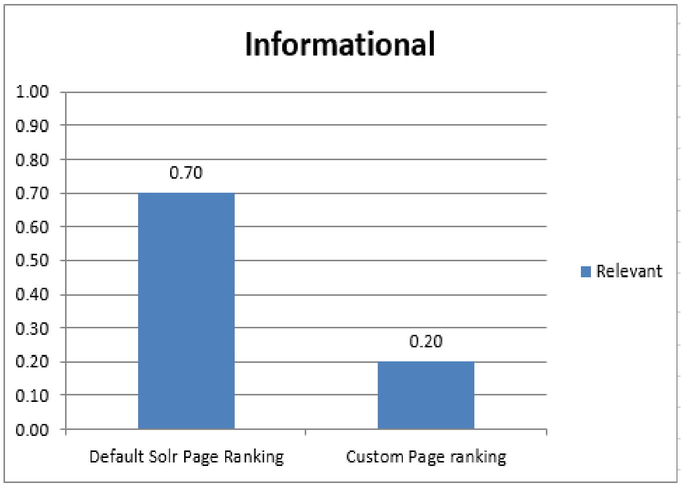
Navigational relevancy result for approach 2



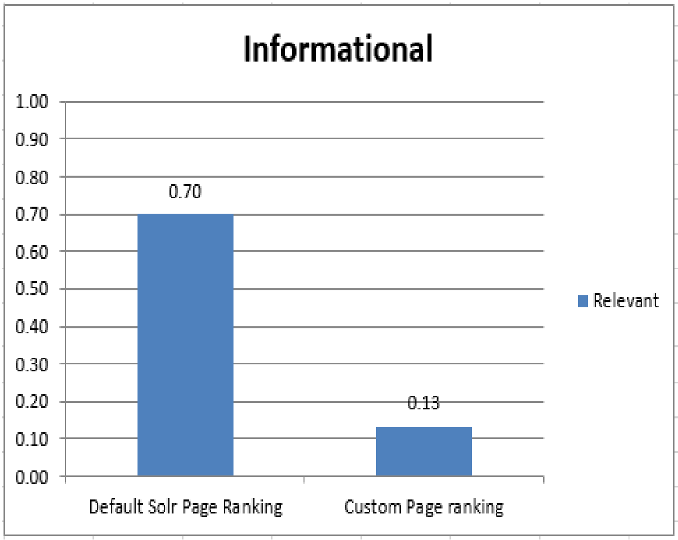
Informational relevancy result for Google and Bing



Information relevancy result for approach 1:



Informational relevancy result for approach 2:



According to the observations,

On the whole, for the navigational queries and informational queries, Google and Bing are found to give results with better relevance when compared with Solr, with and without page ranking approach. This is because, Solr uses Lucene to facilitate ranking. Lucene uses a combination of the Vector Space Model and the Boolean model to determine how relevant a given document is to a user's query. Though, "***AND***" can be included as shown above to increase the relevancy. But, the vector space model is based on term frequency. Hence, the queries would fetch web pages that have better term frequency and there is no existence of an intelligent model that adopts to give out results based on the context and semantics of the query. This behavior is similar to Solr, with the aid of an external page ranking file. The pages which have a good combination of inner and outer links would be fetched for the query, instead of an intelligent approach that considers the context and semantics of the query. This is the reason why some pages have better page rank when compared with the others. But, on the other hand, Google and

Bing is a global search engine that incorporates sophisticated algorithms to give out better search results for the queries.

The relevancy results can better explained with an example. If the query is “Interactive Media and Games USC School of Cinematic Arts”, the user’s primary intent is to find the pages of Interactive Media and Games division within USC School of Cinematic Arts. When this query is provided to Solr, it provides pages with higher term frequency of the query within the indexed files and hence provide a page which contains some information about the topic of Interactive Media and Games within USC School of Cinematic Arts. For the default Solr page rank behavior, the page has a good page rank because of the term frequency. On the other hand, for Solr with external page rank behavior, the page would have a better page rank compared to other relevant pages because of a good combination of inbound and outbound links. Though this page might have a better page rank, it may not be relevant to the query. Therefore, results which are not as good as the ones provided by Google and Bing is provided by Solr.

For informational queries, the relevance of Solr with default page rank behavior is considered to be comparable with Google, since the top pages that were displayed had better term frequencies of Undergraduate, Master’s or PhD requirements when compared with the other pages.