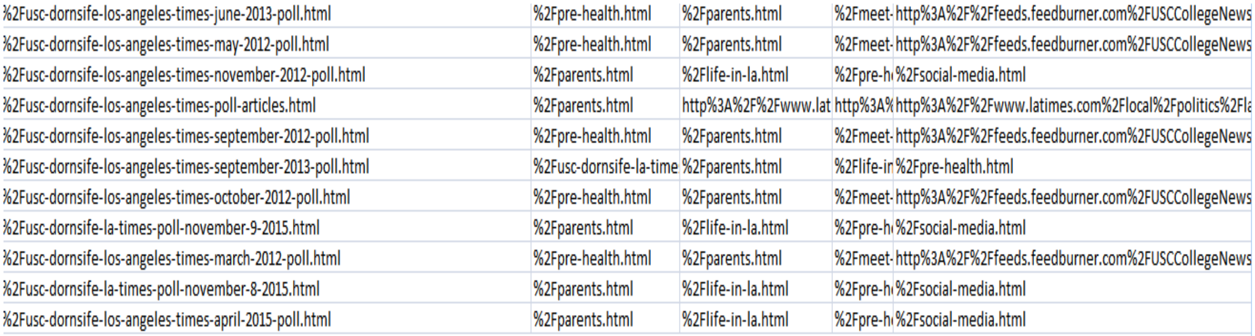
Homework 3-Design and Report Analysis

Website Crawled: USC Dana and David Dornsife College of Letters, Arts and Sciences (http://dornsife.usc.edu)

Crawling and Downloading

The site was crawled and downloaded by crawler4j library in JAVA. The library was customized to fetch , download pages and generate statistics. Maximum pages to fetch were 5000, around 2872 pages were downloaded, the downloaded file types were HTML,PDF and Word Documents. Maximum size was 50 MB to allow download, Pages larger than size of 50MB were not be fetched. For each downloaded file entry ,the link itself and corresponding outgoing links were recorded in pagerank csv. The file downloaded was saved as encoded path (URL-DOMAIN i.e. domain was stripped from the name and path was encoded by URL encoder).Each URL entry was encoded in pagerank csv. Internal URLS were without the domain and other were saved and encoded with the domain.

Pagerank CSV looked Like this:



%2Fusc-dornsife-los-angeles-times-june-2013-poll.html	%2Fpre-health.html	%2Fparents.html	%2Fmeet-	http%3A%2F%2Ffeeds.feedburner.com%2FUSCCollegeNews
%2Fusc-dornsife-los-angeles-times-may-2012-poll.html	%2Fpre-health.html	%2Fparents.html	%2Fmeet-	http%3A%2F%2Ffeeds.feedburner.com%2FUSCCollegeNews
%2Fusc-dornsife-los-angeles-times-november-2012-poll.html	%2Fparents.html	%2Flife-in-la.html	%2Fpre-h	%2Fsocial-media.html
%2Fusc-dornsife-los-angeles-times-poll-articles.html	%2Fparents.html	http%3A%2F%2Fwww.lat	http%3A%	http%3A%2F%2Fwww.latimes.com%2Flocal%2Fpolitics%2Fla
%2Fusc-dornsife-los-angeles-times-september-2012-poll.html	%2Fpre-health.html	%2Fparents.html	%2Fmeet-	http%3A%2F%2Ffeeds.feedburner.com%2FUSCCollegeNews
%2Fusc-dornsife-los-angeles-times-september-2013-poll.html	%2Fusc-dornsife-la-time	%2Fparents.html	%2Flife-in	%2Fpre-health.html
%2Fusc-dornsife-los-angeles-times-october-2012-poll.html	%2Fpre-health.html	%2Fparents.html	%2Fmeet-	http%3A%2F%2Ffeeds.feedburner.com%2FUSCCollegeNews
%2Fusc-dornsife-la-times-poll-november-9-2015.html	%2Fparents.html	%2Flife-in-la.html	%2Fpre-h	%2Fsocial-media.html
%2Fusc-dornsife-los-angeles-times-march-2012-poll.html	%2Fpre-health.html	%2Fparents.html	%2Fmeet-	http%3A%2F%2Ffeeds.feedburner.com%2FUSCCollegeNews
%2Fusc-dornsife-la-times-poll-november-8-2015.html	%2Fparents.html	%2Flife-in-la.html	%2Fpre-h	%2Fsocial-media.html
%2Fusc-dornsife-los-angeles-times-april-2015-poll.html	%2Fparents.html	%2Flife-in-la.html	%2Fpre-h	%2Fsocial-media.html

NetworkX Graph and PageRank File

PYTHON (PY-2.7) library networkx to create a directed graph ,as the pagerank csv has outgoing edges for each internal URL. Each encoded URL(internal ones) were prefixed by the path of folder of indexed data and a map is generated using networkx ,alpha the damping parameter was set 0.9 as default mentioned in documentation for the normalized sum and not let the number of links influence too much in the page rank, used max iterations as default 100 as it was converging around there, didn't make much difference in the search results computed. The key value(float value) pair was thus printed in a text file tor all the urls (both internal with local folder path and external website path) .

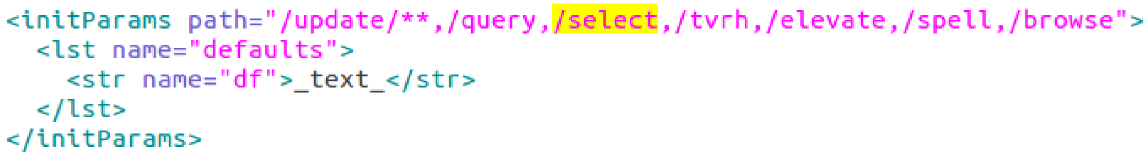
Indexing Website and Fetching Results

Tools used for creating the search engine:

Ubuntu Version: 15.10	VirtualBox:5.0	Solr : 5.5.0
Apache 2 on Linux: 2.4	PHP: 5.2	

Solr : Created a core named dornsife, folder containing downloaded pages was mounted on VM and was indexed by SOLR . Managed-schema.xml was edited for extracting the fields from meta tags of html files title, author, description , content , keywords etc and all are appended to _text_

solarconfig.xml was changed to query for the query parameter in _text_ for all search handlers.



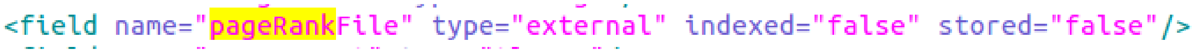
```
<initParams path="/update/**,/query,/select,/tvrh,/elevate,/spell,/browse">
  <lst name="defaults">
    <str name="df">_text_</str>
  </lst>
</initParams>
```

PHP script consisting of both frontend and backend logic connecting to SOLR server was hosted at Apache web server. The request was entered in a input box and top 10 results were displayed with title(as hyperlink),if title is not found it takes default value as "Document", author, date created, size in KB and the result would be the link to be displayed and clicked on. The result link with .Top 10 results were fetched .This gave results for default SOLR algorithm to fetch search results.

PageRankFile Configuration

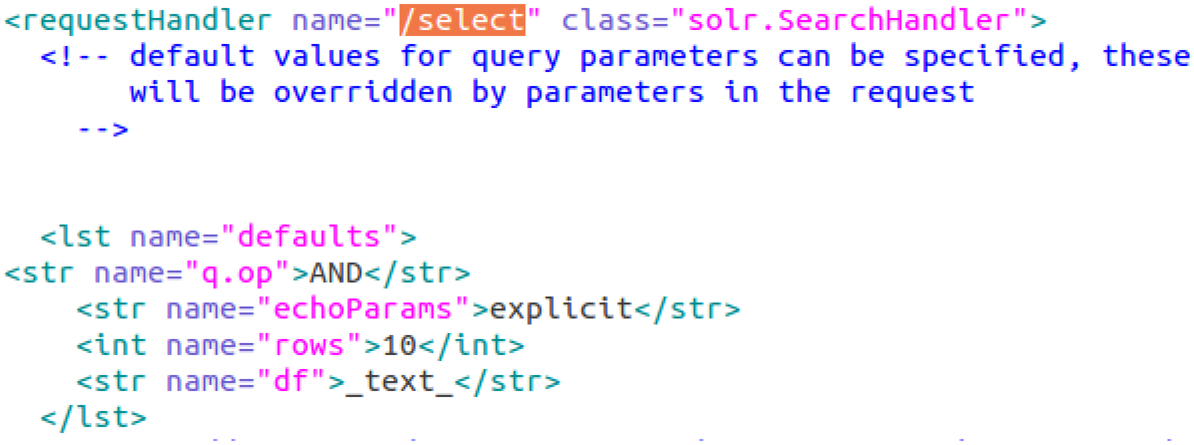
A checkbox was given to enable sorting with Page rank for each file. In the page rank text file, consisting of all links with local URLS(ID for SOLR as well) and external URLS from csv are kept in the text file. The file was stored with the name (externaPageRankFile) in data folder of the respective core.

External field is configured in SOLR for each corresponding document ID in managed-schema:



```
<field name="pageRankFile" type="external" indexed="false" stored="false"/>
```

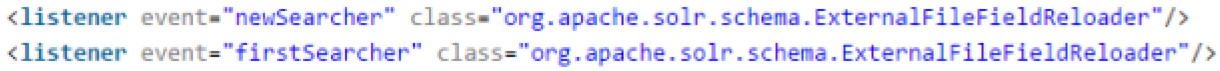
AND query model was configured to use both SOLR and Pagerank algorithm, configured in solrconfig.xml.



```
<requestHandler name="/select" class="solr.SearchHandler">
  <!-- default values for query parameters can be specified, these
       will be overridden by parameters in the request
    -->


  <lst name="defaults">
<str name="q.op">AND</str>
    <str name="echoParams">explicit</str>
    <str name="rows">10</int>
    <str name="df">_text_</str>
  </lst>
```

Added the listener eventListeners to reload the external file, everytime searcher is reloaded or new searcher is started



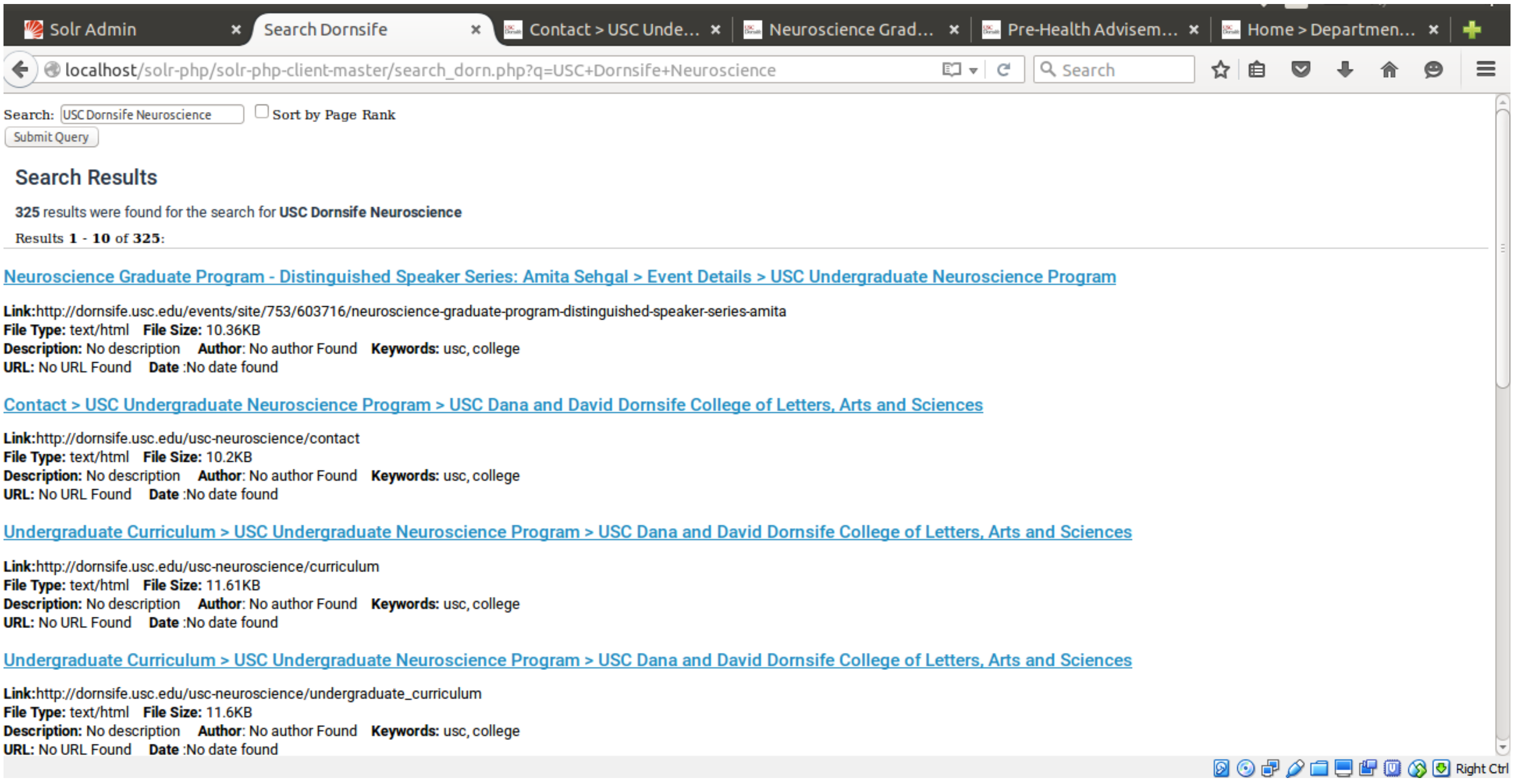
```
<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
<listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
```

In the query ,additional parameters for sort is sent to sort according to pagerank.

i.e. sort: pageRankFile desc;

Queries and Search Resulting Comparison

Example: USC Dornsife Neurocience without pagerank



Search: USC Dornsife Neuroscience □ Sort by Page Rank

Submit Query

Search Results

325 results were found for the search for **USC Dornsife Neuroscience**

Results **1** - **10** of **325**:

Neuroscience Graduate Program - Distinguished Speaker Series: Amita Sehgal > Event Details > USC Undergraduate Neuroscience Program

Link:http://dornsife.usc.edu/events/site/753/603716/neuroscience-graduate-program-distinguished-speaker-series-amita
File Type: text/html **File Size:** 10.36KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

Contact > USC Undergraduate Neuroscience Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

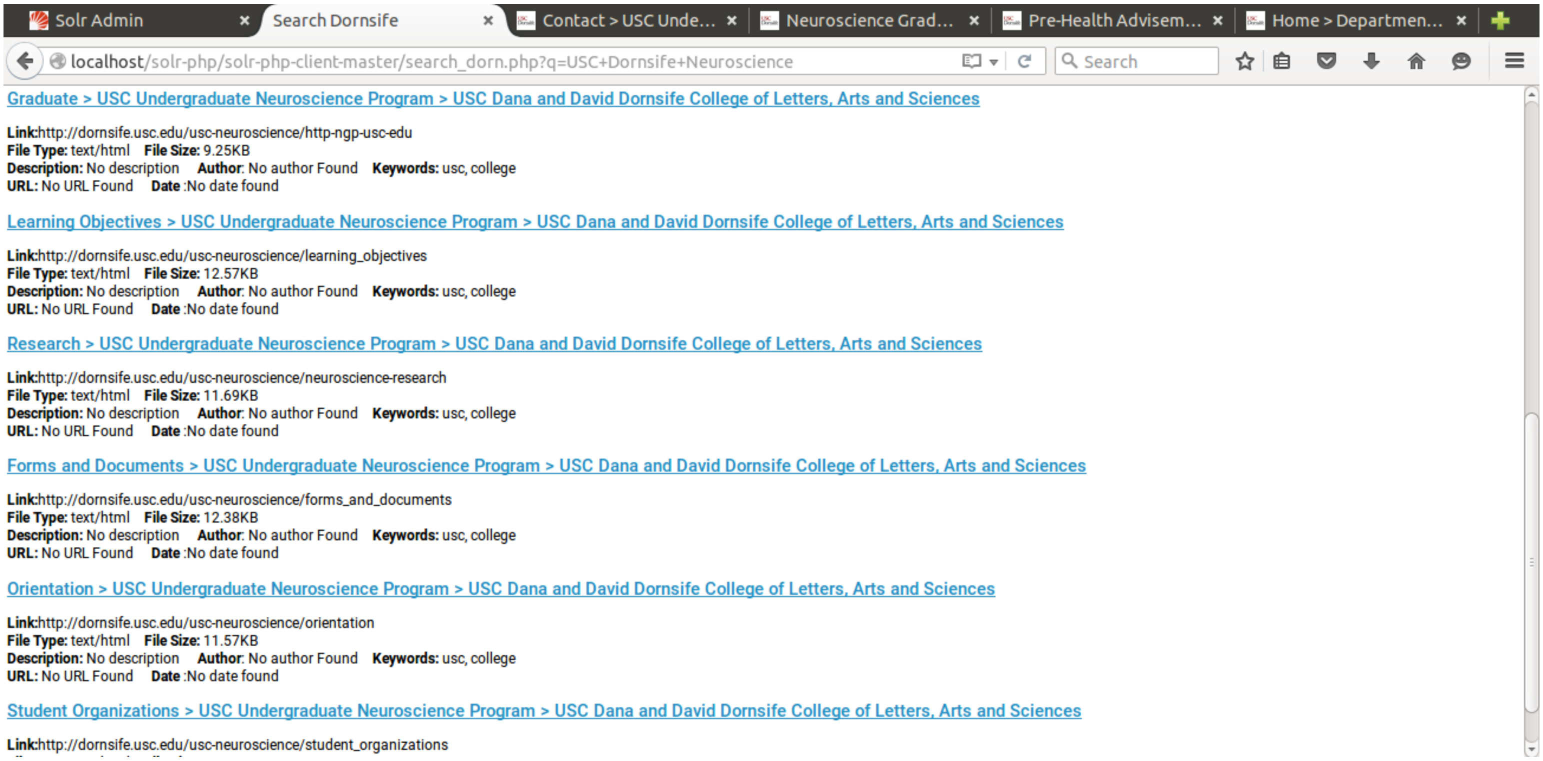
Link:http://dornsife.usc.edu/usc-neuroscience/contact
File Type: text/html **File Size:** 10.2KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

Undergraduate Curriculum > USC Undergraduate Neuroscience Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/usc-neuroscience/curriculum
File Type: text/html **File Size:** 11.61KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

Undergraduate Curriculum > USC Undergraduate Neuroscience Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/usc-neuroscience/undergraduate_curriculum
File Type: text/html **File Size:** 11.6KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found



Graduate > USC Undergraduate Neuroscience Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/usc-neuroscience/http-ngp-usc-edu
File Type: text/html **File Size:** 9.25KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

Learning Objectives > USC Undergraduate Neuroscience Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/usc-neuroscience/learning_objectives
File Type: text/html **File Size:** 12.57KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

Research > USC Undergraduate Neuroscience Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/usc-neuroscience/neuroscience-research
File Type: text/html **File Size:** 11.69KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

Forms and Documents > USC Undergraduate Neuroscience Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/usc-neuroscience/forms_and_documents
File Type: text/html **File Size:** 12.38KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

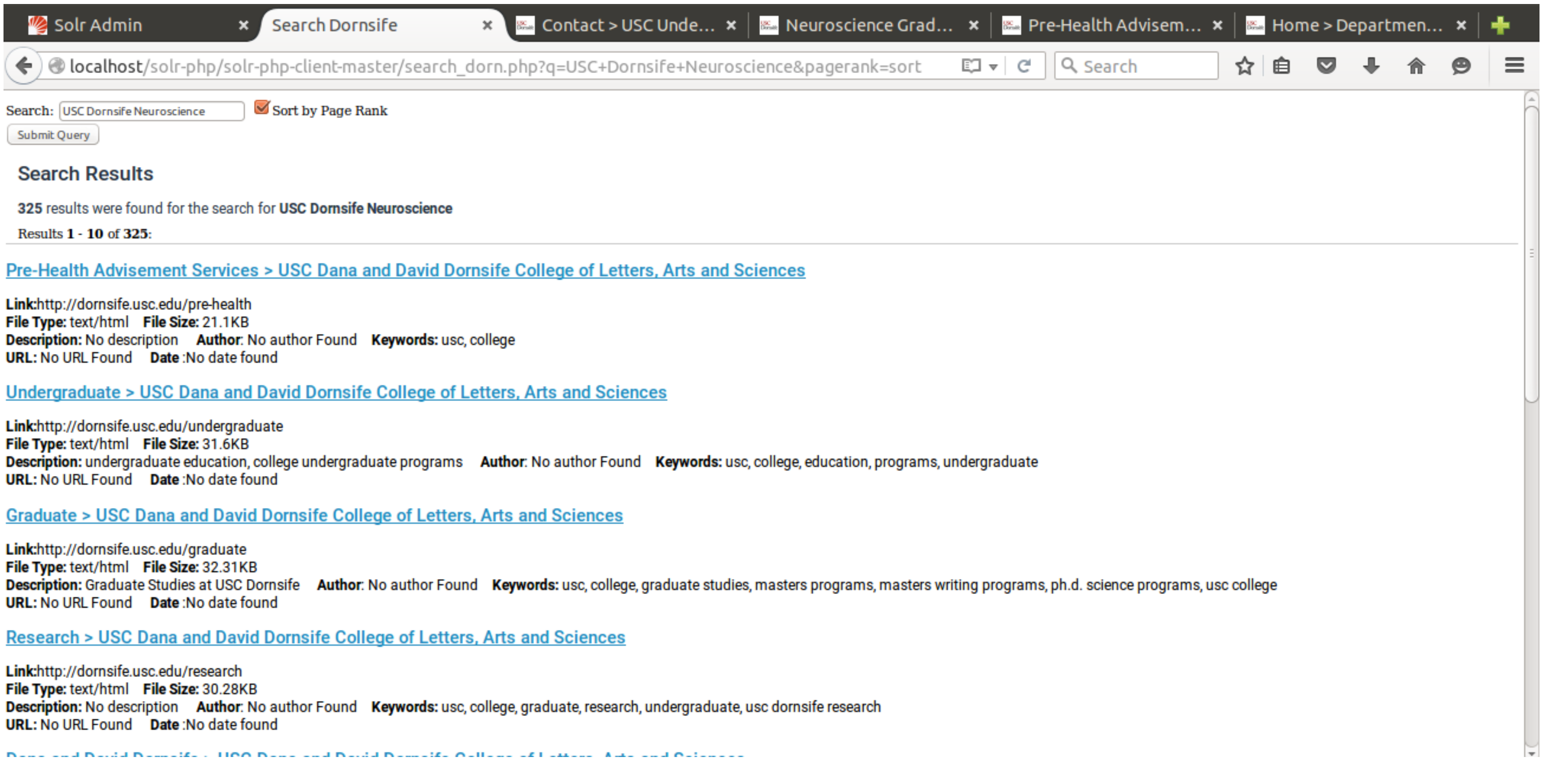
Orientation > USC Undergraduate Neuroscience Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/usc-neuroscience/orientation
File Type: text/html **File Size:** 11.57KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

Student Organizations > USC Undergraduate Neuroscience Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/usc-neuroscience/student_organizations

Sorted Results for the example "USC Dornsife Neuroscience" with Pagerank



Search: USC Dornsife Neuroscience ☑ Sort by Page Rank

Submit Query

Search Results

325 results were found for the search for **USC Dornsife Neuroscience**

Results **1** - **10** of **325**:

[Pre-Health Advisement Services > USC Dana and David Dornsife College of Letters, Arts and Sciences]

Link:http://dornsife.usc.edu/pre-health
File Type: text/html **File Size:** 21.1KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

[Undergraduate > USC Dana and David Dornsife College of Letters, Arts and Sciences]

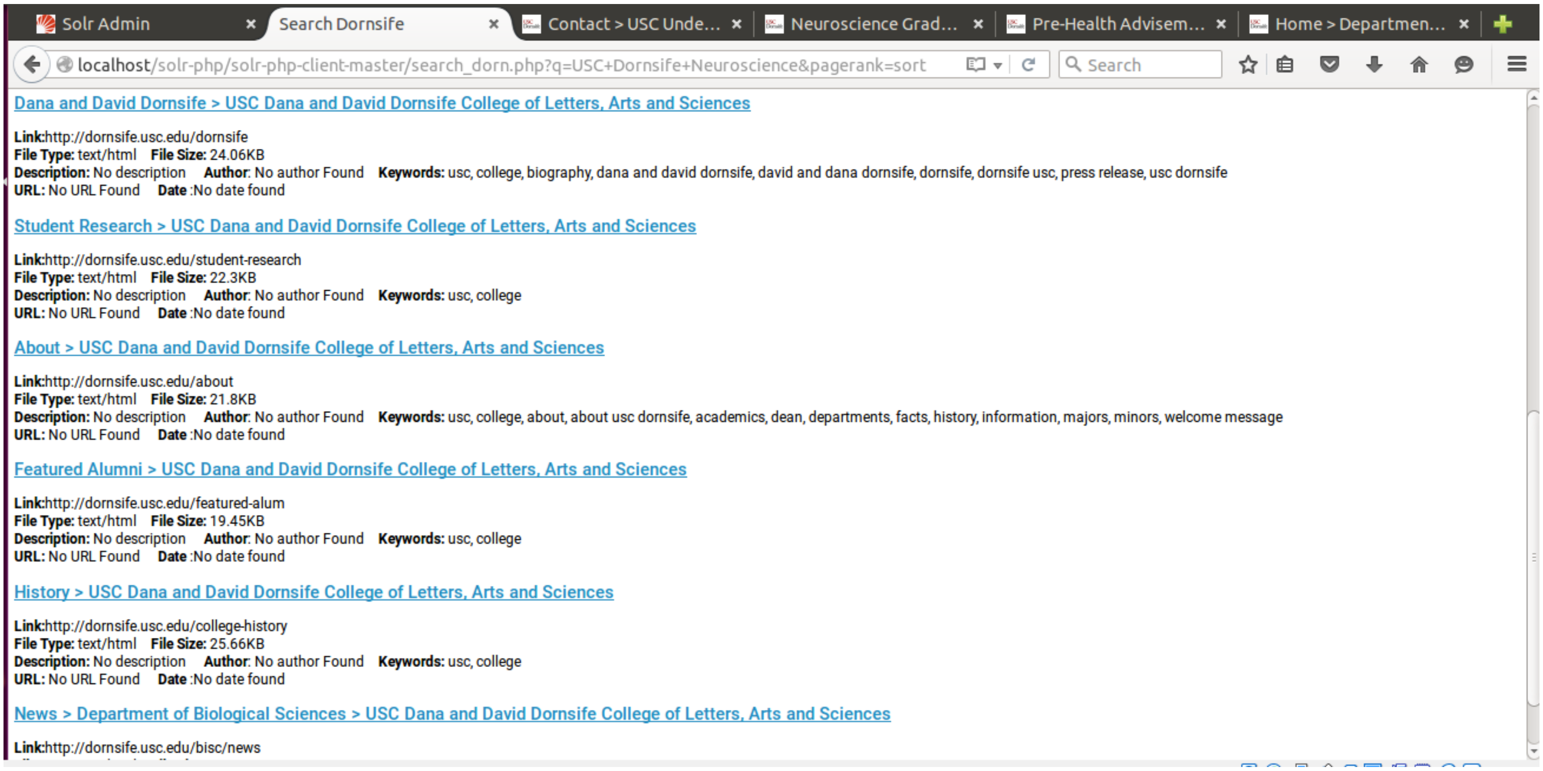
Link:http://dornsife.usc.edu/undergraduate
File Type: text/html **File Size:** 31.6KB
Description: undergraduate education, college undergraduate programs **Author:** No author Found **Keywords:** usc, college, education, programs, undergraduate
URL: No URL Found **Date** :No date found

[Graduate > USC Dana and David Dornsife College of Letters, Arts and Sciences]

Link:http://dornsife.usc.edu/graduate
File Type: text/html **File Size:** 32.31KB
Description: Graduate Studies at USC Dornsife **Author:** No author Found **Keywords:** usc, college, graduate studies, masters programs, masters writing programs, ph.d. science programs, usc college
URL: No URL Found **Date** :No date found

[Research > USC Dana and David Dornsife College of Letters, Arts and Sciences]

Link:http://dornsife.usc.edu/research
File Type: text/html **File Size:** 30.28KB
Description: No description **Author:** No author Found **Keywords:** usc, college, graduate, research, undergraduate, usc dornsife research
URL: No URL Found **Date** :No date found



[Dana and David Dornsife > USC Dana and David Dornsife College of Letters, Arts and Sciences]

Link:http://dornsife.usc.edu/dornsife
File Type: text/html **File Size:** 24.06KB
Description: No description **Author:** No author Found **Keywords:** usc, college, biography, dana and david dornsife, david and dana dornsife, dornsife, dornsife usc, press release, usc dornsife
URL: No URL Found **Date** :No date found

[Student Research > USC Dana and David Dornsife College of Letters, Arts and Sciences]

Link:http://dornsife.usc.edu/student-research
File Type: text/html **File Size:** 22.3KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

[About > USC Dana and David Dornsife College of Letters, Arts and Sciences]

Link:http://dornsife.usc.edu/about
File Type: text/html **File Size:** 21.8KB
Description: No description **Author:** No author Found **Keywords:** usc, college, about, about usc dornsife, academics, dean, departments, facts, history, information, majors, minors, welcome message
URL: No URL Found **Date** :No date found

[Featured Alumni > USC Dana and David Dornsife College of Letters, Arts and Sciences]

Link:http://dornsife.usc.edu/featured-alum
File Type: text/html **File Size:** 19.45KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

[History > USC Dana and David Dornsife College of Letters, Arts and Sciences]

Link:http://dornsife.usc.edu/college-history
File Type: text/html **File Size:** 25.66KB
Description: No description **Author:** No author Found **Keywords:** usc, college
URL: No URL Found **Date** :No date found

[News > Department of Biological Sciences > USC Dana and David Dornsife College of Letters, Arts and Sciences]

Link:http://dornsife.usc.edu/bisc/news

Top result for USC Dornsife Neuroscience : Solr vs PageRank

Relevancy of Solr Vs Pagerank

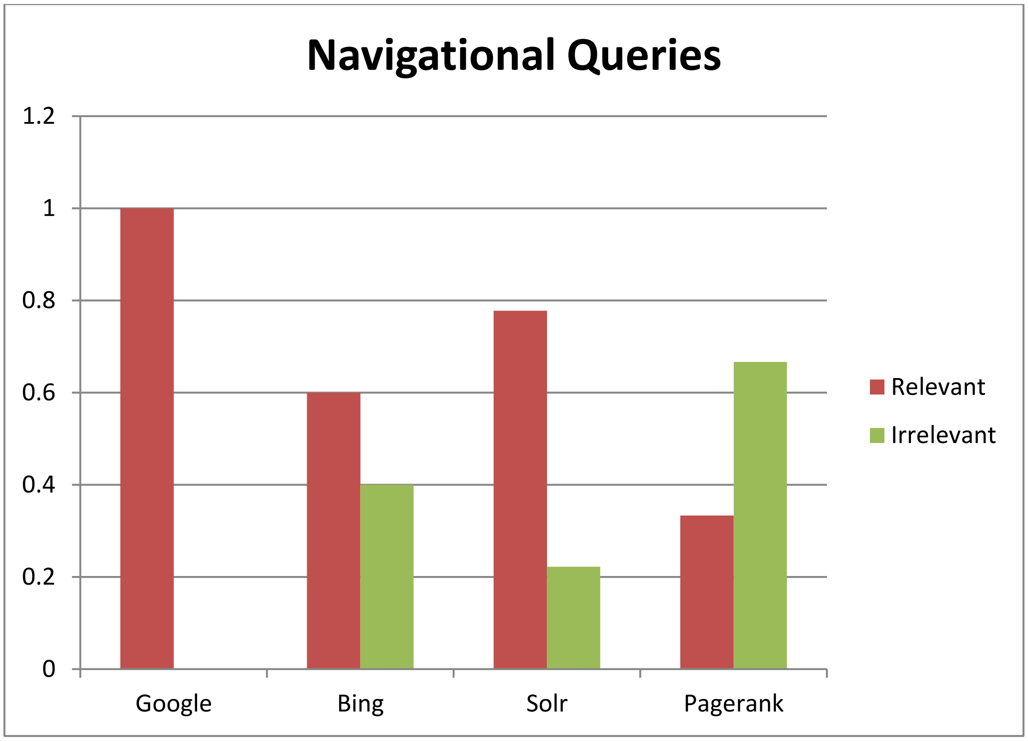
SOLR works on Lucene, this means that it takes all the documents, splits them into words, and then builds an index *for each word*. Since the index is an exact string-match, unordered.

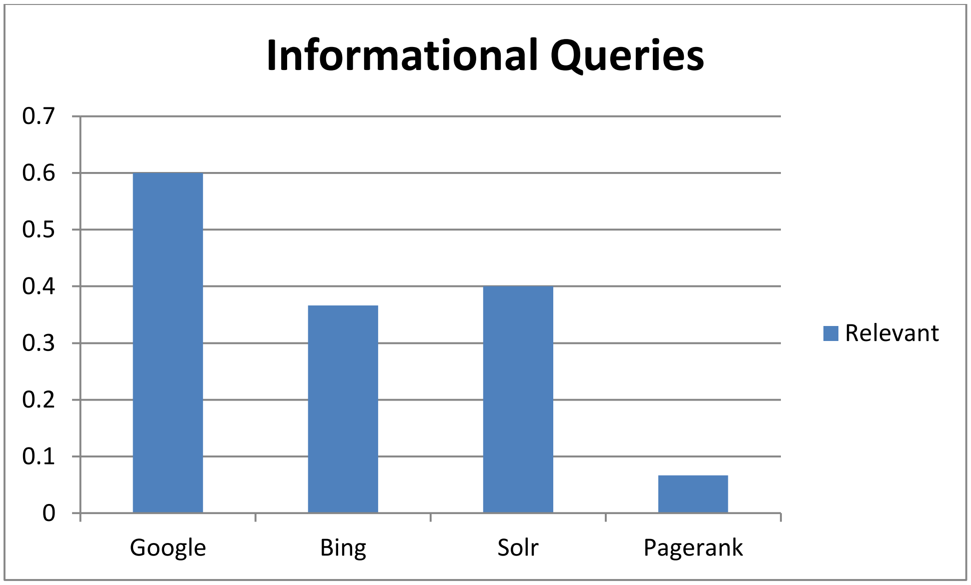
1. As SOLR works on score match with index it resulted around 70 % relevance in navigational queries as it could match the keywords in the downloaded documents. Whereas pagerank algorithm could give relevance of only 30% in navigational queries. Thus as we are querying "USC Dornsife" as prefix in most queries, the search results when matched with indexed words makes the real query "Neuroscience" , the results become more irrelevant. The top result has maximum word match score for USC Dornsife becomes the topmost.As data set is huge and 5000 pages were downloaded , thus pages about certain faculty could not be collected. Some relevant pages could be sparsely connected thus irrelevant results shown in pagerank. Example the pages belonging to philosophy are sparsely connected in the page graph thus, irrelevant results took priority over it and thus reducing the relevance in page rank algorithm results.

2. For informational queries as the words could match several pages undergraduate, graduate, degree, requirements, usc , dornsife, english. These keywords could be present in content of several pages but may not exactly relevant. For example : There could be English requirements of TOEFL exam in Economics Degree. Thus the relevance became even lower in case of informational queries, only 40 % relevance, as economics page is more nested than statistics page , the results order gave even more skewed results thus resulting in very low relevance of 7%in case of pagerank.

Similarly for entire set of queries as USC Dornsife was there in each query it skewed with exact relevance of the pages. As is SOLR relevancy was quite medium , the descending of page score gave rise to more of irrelevant results , as those pages might have decent score and has both incoming and outgoing links.

	Navigational			
	Google	**Bing**	**Solr**	**Pagerank**
Relevant	1	0.6	0.77777778	0.333333
Irrelevant	0	0.4	0.22222222	0.666667
	Informational			
	Google	**Bing**	**Solr**	**Pagerank**
Relevant	0.6	0.37	0.4	0.07
Irrelevant	0.4	0.63	0.6	0.93

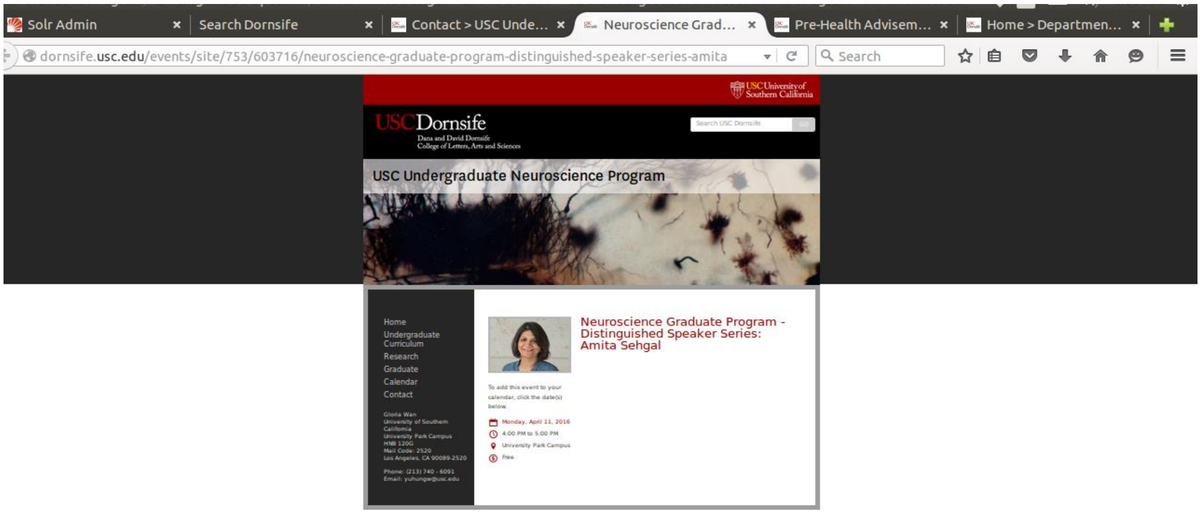




Pages with Higher Page Rank

Home page and pages directly connected with home page have higher page rank value because of their connected, thus a hierarchy concentrates votes and Pagerank into connected page. For the demonstrated results it is just under homepage and it is well structured hierarchy, thus a well structured site will amplify the effect of any contributed to pushing of higher pagerank value because of outgoing links. As there are few internal links repeated like "Home page" "Life in LA","Career Pathway", "Internships" thus pagerank values are not evenly distributed as we can see comparison in the demonstration: neuroscience page pagerank(3.092e-05)) and pre health page (Pagerank=0.00072) .

Demonstrated page :If The link of page comes in the left navigation, breadcrumb of first hierarchy pages of the school's website. As it has a major set of outgoing URLS to other first level pages from left navigation and top navigation links to major hierarchies. As this page lies in left navigation bar thus it is pointed at by majority pages of the same level. As the home page is being referred in the most pages and this has both incoming and outgoing URLs along with homepage. It will be ranked higher for pagerank according to pagerank algorithm. Even though this page is very less relevant to neuroscience has a mention of the neuroscience building , but it has decent mention of USC Dornsife thus the query fetches the pre-health page as the top most result, followed by alumni ,research page, graduate page. All the pages closely connected to home page.


Solr Admin
Search Dornsife
Contact > USC Unde…
Neuroscience Grad…
Pre-Health Advisem…
Home > Departmen…
dornsife.usc.edu/events/site/753/603716/neuroscience-graduate-program-distinguished-speaker-series-amita
Search
USC University of Southern California
USC Dornsife
Dana and David Dornsife College of Letters, Arts and Sciences
USC Undergraduate Neuroscience Program
Home
Undergraduate Curriculum
Research
Graduate
Calendar
Contact

Gloria Wan
University of Southern California
University Park Campus
HNB 120G
Mail Code: 2520
Los Angeles, CA 90089-2520

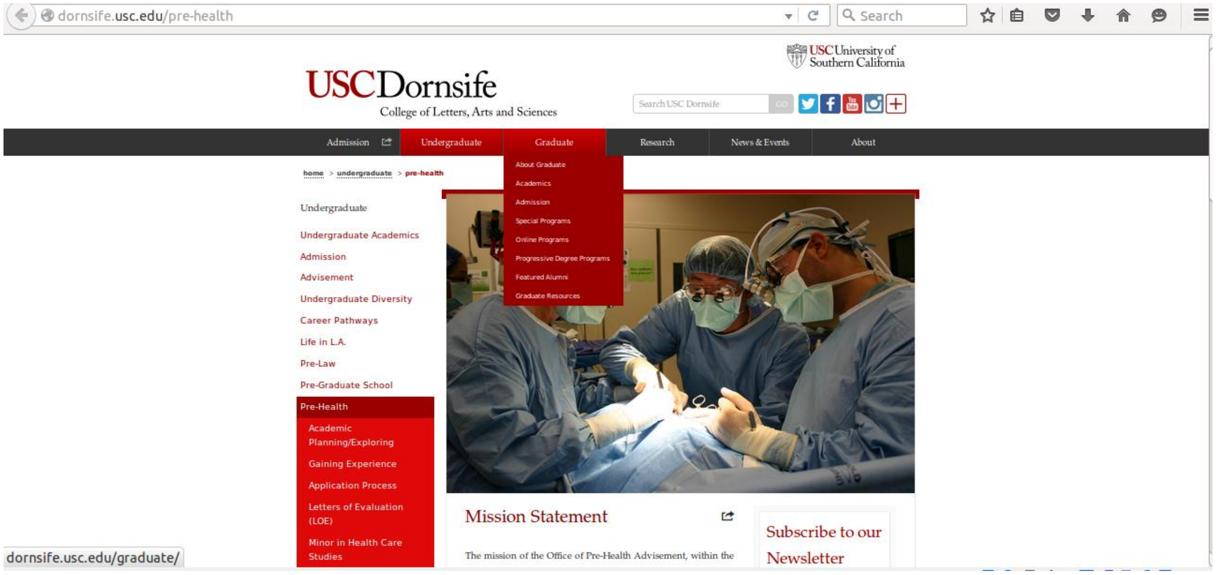
Phone: (213) 740 - 6091
Email: yuhungw@usc.edu

Neuroscience Graduate Program - Distinguished Speaker Series: Amita Sehgal

To add this event to your calendar, click the date(s) below.

Monday, April 11, 2016
4:00 PM to 5:00 PM
University Park Campus
Free

Pagerank(3.092e-05)



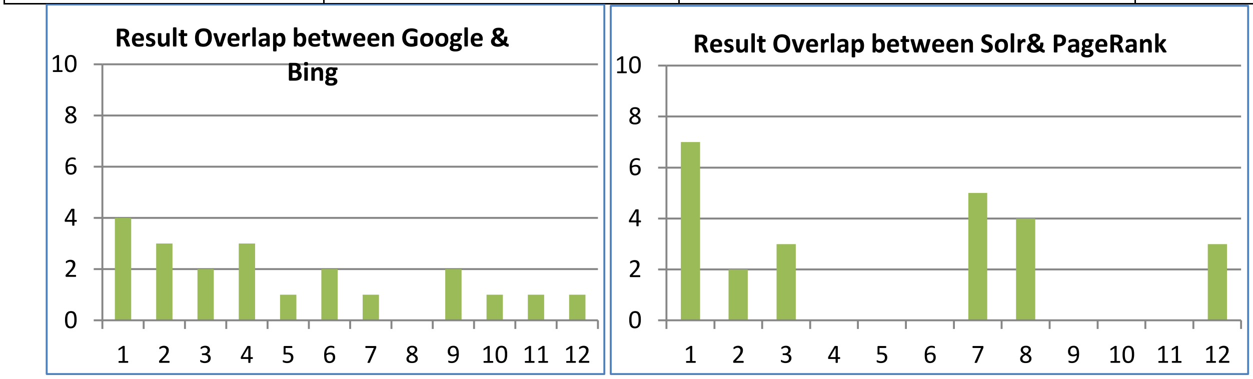
The mission of the Office of Pre-Health Advisement, within the

(Pagerank=0.00072)

Queries :

Faculty:	**Departments:**	**Others:**	**Informational Queries:**
Eric Friedlander USC Dornsife	USC Dornsife Neuroscience	Directions: USC Dana and David Dornsife Map and Directions	USC Dornsife English Undergraduate degree

			requirements
Darren Ruddell USC Dornsife	USC Dornsife Economics	Founders: Founders Dana and David Dornsife USC College of Letters, Arts and Sciences	USC Dornsife Master of Science in Statistics degree requirements
Peter C. Mancall USC Dornsife	USC Dornsife Philosophy	Alumni: USC Dana and David Dornsife Alumni News	USC Dornsife English PHD degree requirements



Overlap Google vs Bing vs Solr vs Pagrank

The overlap between solr and pagerank occurred where number of results was low, thus they could overlap. For example for a faculty "Eric Friedlander USC Dornsife" only 7 results could be fetched, other faculties also had fewer results thus more of a overlap.

	Eric Friedlander USC Dornsife					Darren Ruddell USC Dornsife					Peter C. Mancall USC Dornsife				
	Google	Bing	Solr	PageRank	Overlap	Google	Bing	Solr	PageRank	Overlap	Google	Bing	Solr	PageRank	Overlap
1	1	1	0.5	0.5	4	1	1	1	0.25	3	1	1	0.25	0	2
2	0.25	1	0.5	0.5	7	1	0	0.25	0	2	0.25	0	0.25	0.25	3
3	0.25	0.5	0.25	0.25		0.5	0.5	0	0		0.25	0	0.25	0	
4	0	0.25	0	0		0.25	0	0	0.5		1	0.25	0.25	0	
5	0.25	0.25	0	0.5		0	1	0.25	0		0.25	0	0	0	
6	1	0	0.5	0		0	0.25	0	0.25		0.5	0	0	0	
7	0.5	0.25	0	0		0	0	0	0		0.25	0.25	0	0	
8	0	0				0	0.25	0.5	0		0.5	0	0	0.25	
9	0	0				0.25	0	0	0.25		0	0.25	0	0.25	
10	0.25	0				0	0	0.25	0.25		1	0.5	0	0	

	USC Dornsife Neuroscience					USC Dornsife Economics					USC Dornsife Philosophy				
	Google	Bing	Solr	PageRank	Overlap	Google	Bing	Solr	PageRank	Overlap	Google	Bing	Solr	PageRank	Overlap
1	0.5	1	0.5	0.5	3	1	1	0.5	0	1	1	1	0.5	0	2
2	1	0.5	1	0		0.5	0.5	0.5	0		0	1	1	0	
3	0.5	0.5	0.5	0		0.5	1	1	0.5		0	0.5	0.5	0.5	
4	0.5	0.5	0.5	0		0	0.5	0.5	0.5		0	0.5	0.5	0	
5	0	0.5	0.5	0		0.5	0	0.5	0		0.5	0.5	0.5	0	
6	0	0	0.5	0		0	0	1	0		0.5	0	0.5	0	
7	0.5	0.5	1	0.5		0.5	0	0.5	0.5		0.5	0.5	0.5	0	
8	0.5	0	0.5	0		0	0.5	0.5	0		0.5	0.5	0.5	0	
9	1	0	0.5	0		0	0	0.5	0		0.5	0	0.5	0	
10	1	0	0.5	0		0	0	0.5	0		0.5	0	0.5	0	

	usc dana and david dornsife map and directions					Founders Dana and David Dornsife USC College of Letters, Arts and Sciences					USC Dana and David Dornsife Alumni News				
	Google	Bing	Solr	PageRank	Overlap	Google	Bing	Solr	PageRank	Overlap	Google	Bing	Solr	PageRank	Overlap
1	1	0	1	1	1	1	0	0	0	0	0.5	0.5	0.5	0	2
2	0	1	0	1	5	0	1	0	0	4	0	0	0.5	0	
3	0	0	0	1		0	0	0.25	0.25		1	0	0.5	0	
4	1	0	0	0		0	0	0	0		1	0	0.5	0	
5	1	0	0	0		0	0.5	0	0		0.5	0.5	0.5	0.5	
6	1	0	0	0		0	0	0	0		0	0	1	0	
7	1	0	1	0		0	0	0	0		0	0	0	0	
8	0	0	0	0		0.5	0	0	0		0	0	0.5	0	
9	0	0	1	0		0.5	0.5	0	0		1	0	0	0	
10	1	0	0	0		0	0	0	0		0	0	0.5	0	

	USC Dornsife English Undergraduate degree requirements					USC Dornsife Master of Science in Statistics degree requirements					USC Dornsife English PHD degree requirements				
	Google	Bing	Solr	PageRank	Overlap	Google	Bing	Solr	PageRank	Overlap	Google	Bing	Solr	PageRank	Overlap
1	1	1	1	0	1	0	1	0.5	0	1	1	0	1	0	1
2	0	1	0	0	0	1	0	0.5	0	0	1	1	0.5	0	3
3	0.5	0	0	0		0.5	0	0	0		0.5	0	1	0	
4	0	0	0.5	0		0	0	0	0		0	0.5	0	0	
5	0.5	0	0	0		0	0	0	0		1	0.5	0	0	
6	0	0.5	0.5	0		1	0.5	0	0		1	0	0.5	0	
7	0.5	0	0	0		0.5	0.5	0	0		1	0	0.5	0	
8	1	0	0.5	0		1	0	0	0.5		0	0	0.5	0	
9	0.5	0	0	0		0	0	0	0.5		0	0	0	0	
10	0.5	0.5	0	0		0	0	0	0		0	0.5	0	0	