Steps followed to complete the assignment.

Tools and configurations

1. Installed virtual box and setup Ubuntu on the virtual machine.
2. Then installed Solr 5.3.1 on ubuntu virtual machine for indexing and for searching. [Home directory]
3. After solr installed successfully and started on port 8983, created a new core called myexample in solr by using the command

   ```
   bin/solr create –c myexample
   ```

   This created a core in the solr-3.5/ server/ solr/ with the name myexample.
4. Changes made to the schema files in the solr core as mentioned in the HW document description to copy all the elements to a destination field called _text_.
5. The meta data extracted by TIKA are redirected to this destination _text_ by making changes in solr_config.xml.
6. The crawled data folder consisting of downloaded pages is mounted on ubuntu machine by creating a shared folder MyCrawler

   ```
   sudo mount – t vboxsf  MyCrawler  crawler_data/
   ```
7. After mounting the crawled folder, created an index in the solr core my example by using the crawled data

   ```
   bin/post –c myexample  crawl_data/
   ```

8. This creates the index on the crawled data in solr that can be queried using the Solr interface.

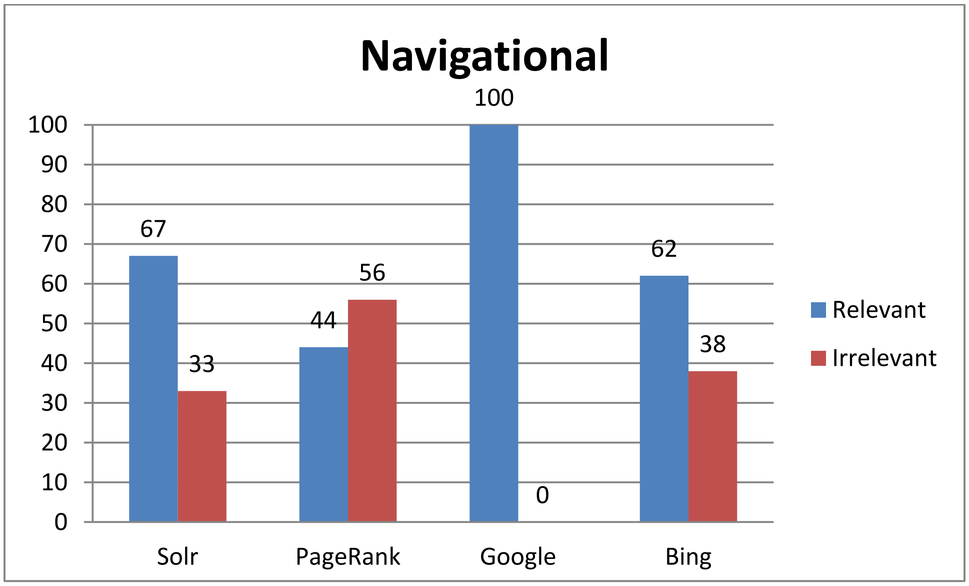
Solr – PHP Client Interface

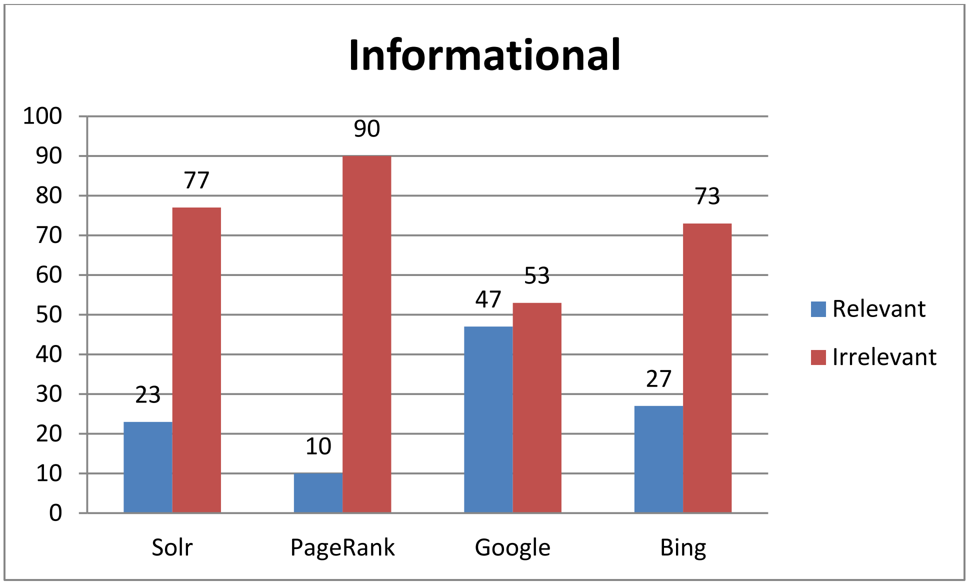
1. To run PHP on ubuntu installed Apache2 server and PHP on ubuntu.
2. Used the php file provided on github for the client interface to Solr server to submit queries to solr index.
3. Modified the php file to incorporate changes mentioned in the HW description
   Changes done :
   a. Added a checkbox to switch between solr search and page rank search.
   b. Set additional parameters to set the sort field to external page rank file.
   c. Extracted meta data from the fetched result like title, stream_size , creation_date and author. Displayed this meta data as part of the search result.
   d. Used a mapping fie to map between the doc id and the URL link that should be displayed in the result page.

Page rank external_page (external_pageRankFile.txt)

1. Copy the pagerank.csv file containing the downloaded pages with their outgoing url’s from each page.
2. Used NetworkX library in python to create the external page rank file. The python script reads downloaded page ids from the pagerank.csv file one by one. The column 1 value from each line represents the downloaded page. This column1 value matches with the id in the solr index.
3. To create the digraph make the column1 value [downloaded page] as the node in the graph. The outgoing links that are there in other columns will form a link with the node that is the downloaded page value.
4. Once the digraph is created, it is passed to the networkx page rank API to compute the pagerank values for downloaded pages.
   Call to api = nx.pagerank(G)
   Alpha value = default (0.85)
5. The values are stored in the external_pageRankFile.txt file in the format
   Docid_ url : page_rank_value.
6. External_pageRankFile.txt is added to the data folder of the myexample core and changes are made in xml file to incorporate it in solr.

Graph Comparison





Analysis of graphs

a. Comparing the performance of Solr 's Lucene and page rank searches on both informational and navigational queries shows that solr lucene out perform the page rank. This is mainly has to do with the kind of models used in each of these approaches. 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. While page rank only considers the number of incoming and outgoing links while computing the page rank. It does not take content of the page into account. Hence pages with higher page rank values are shown top even though they don't have relevant content.

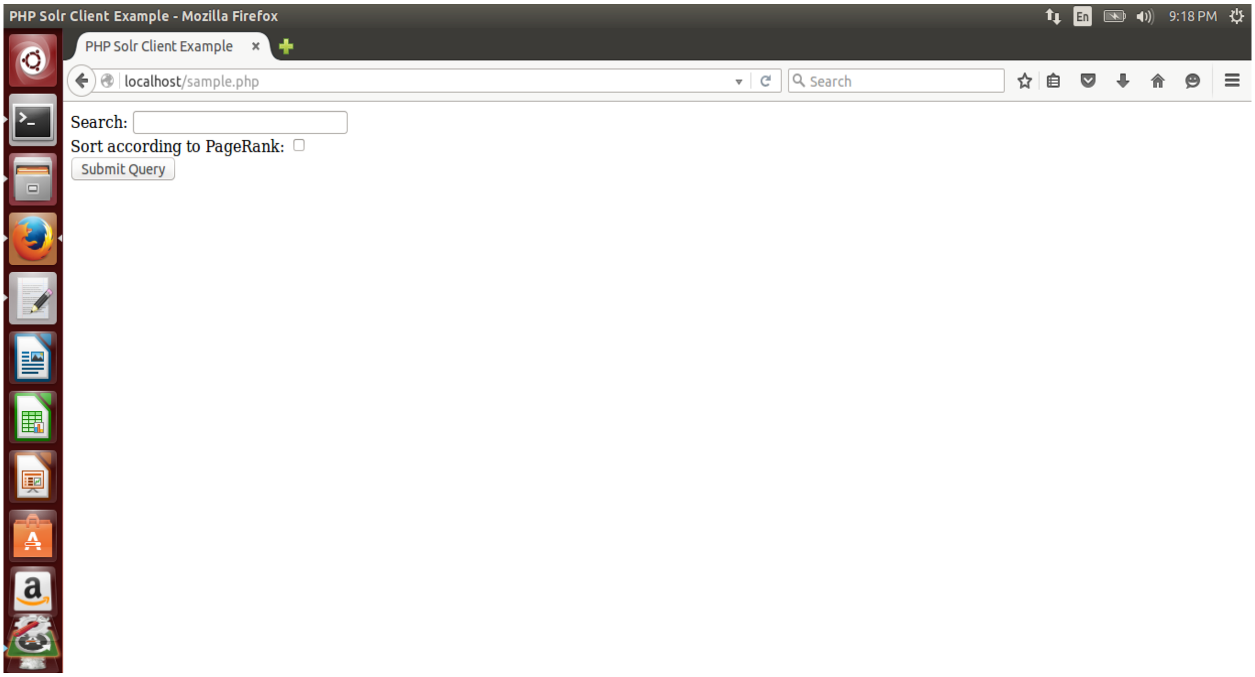
b. When we compare the graph of solr, page rank with search engines results, it show that solr's lucene performance on both navigational and informational queries are almost close to results of Bing Search Engine. Comparing the performance of the two search engines by considering the relevance values for both informational and navigational queries, Google performs better than Bing when top ten results are taken into consideration.

4. Explanation regarding why some pages have higher page rank values

The page rank values of a page depend on the number of links pointing to the page , their page rank values and the number of outgoing links in each of these pages. It does not consider the relevant content or any content for that matter while computing the page rank.

That's why some of the pages have higher page rank values because of the large number of incoming links pointing to this page. Each incoming link pointing to the page will act as a vote and signify the importance of that page. Hence some pages like homepages, department pages that have many other pages pointing to it will always have higher page rank compared to other pages.

5. Snapshots of Queries from Solr and page rank search results.

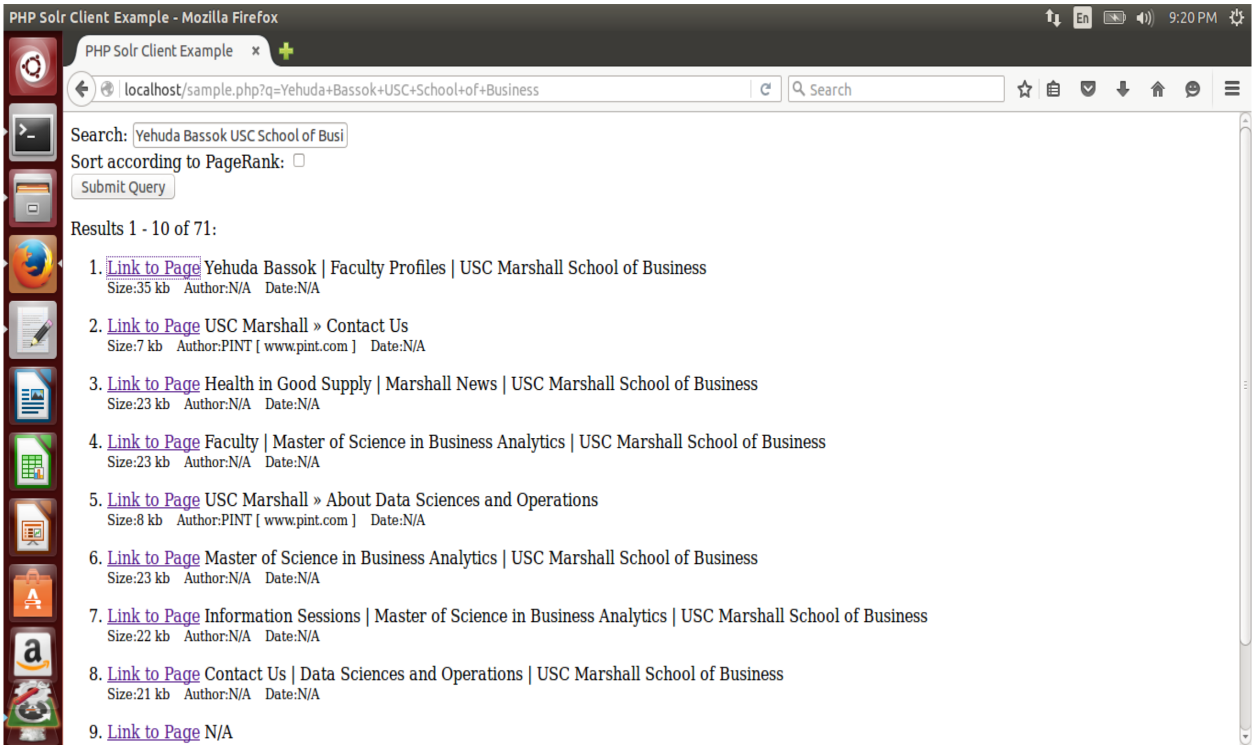


Search:

Sort according to PageRank: □

Submit Query

Query Page


PHP Solr Client Example - Mozilla Firefox
9:20 PM
PHP Solr Client Example
localhost/sample.php?q=Yehuda+Bassok+USC+School+of+Business
Search

Search: Yehuda Bassok USC School of Busi

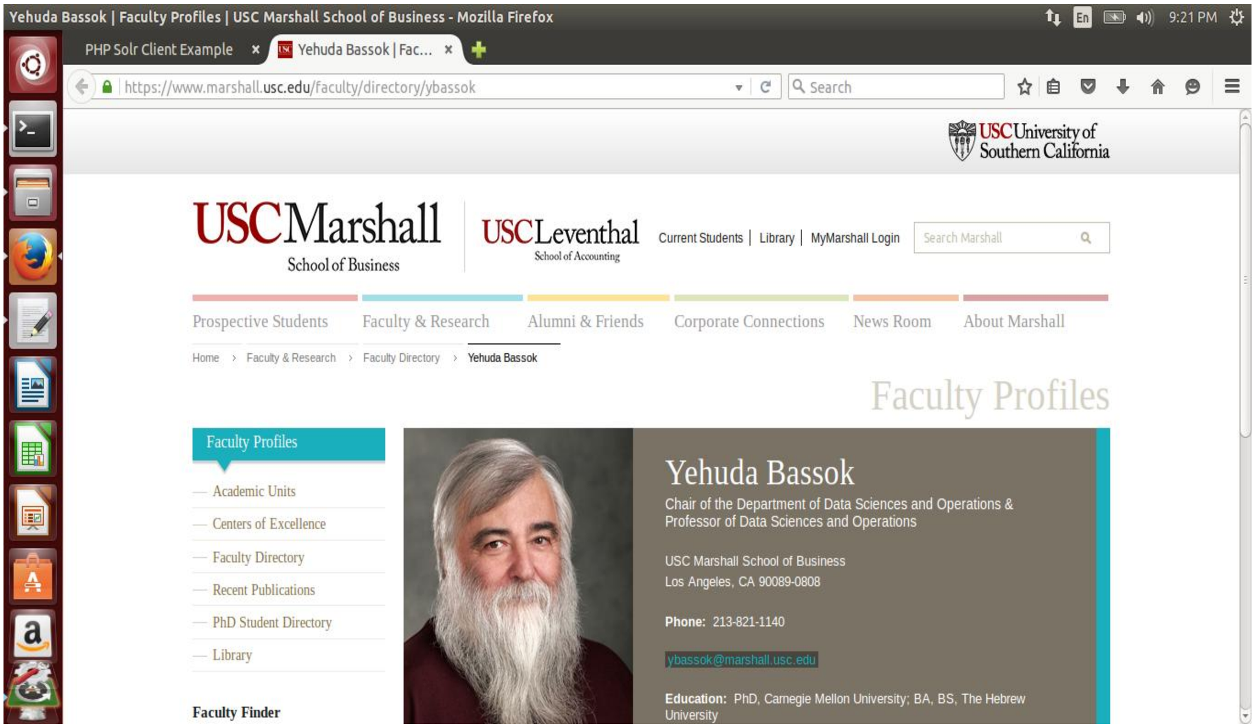
Sort according to PageRank: □

Submit Query

Results 1 - 10 of 71:

1. Link to Page Yehuda Bassok | Faculty Profiles | USC Marshall School of Business
   Size:35 kb Author:N/A Date:N/A
2. Link to Page USC Marshall » Contact Us
   Size:7 kb Author:PINT [www.pint.com] Date:N/A
3. Link to Page Health in Good Supply | Marshall News | USC Marshall School of Business
   Size:23 kb Author:N/A Date:N/A
4. Link to Page Faculty | Master of Science in Business Analytics | USC Marshall School of Business
   Size:23 kb Author:N/A Date:N/A
5. Link to Page USC Marshall » About Data Sciences and Operations
   Size:8 kb Author:PINT [www.pint.com] Date:N/A
6. Link to Page Master of Science in Business Analytics | USC Marshall School of Business
   Size:23 kb Author:N/A Date:N/A
7. Link to Page Information Sessions | Master of Science in Business Analytics | USC Marshall School of Business
   Size:22 kb Author:N/A Date:N/A
8. Link to Page Contact Us | Data Sciences and Operations | USC Marshall School of Business
   Size:21 kb Author:N/A Date:N/A
9. Link to Page N/A

Solr Page Results


Yehuda Bassok | Faculty Profiles | USC Marshall School of Business - Mozilla Firefox
9:21 PM
PHP Solr Client Example
Yehuda Bassok | Fac...
https://www.marshall.usc.edu/faculty/directory/ybassok
Search
USC University of Southern California
USC Marshall School of Business
USC Leventhal School of Accounting
Current Students | Library | MyMarshall Login
Search Marshall
Prospective Students
Faculty & Research
Alumni & Friends
Corporate Connections
News Room
About Marshall
Home › Faculty & Research › Faculty Directory › Yehuda Bassok
Faculty Profiles
Faculty Profiles
Academic Units
Centers of Excellence
Faculty Directory
Recent Publications
PhD Student Directory
Library
Faculty Finder

Yehuda Bassok

Chair of the Department of Data Sciences and Operations & Professor of Data Sciences and Operations

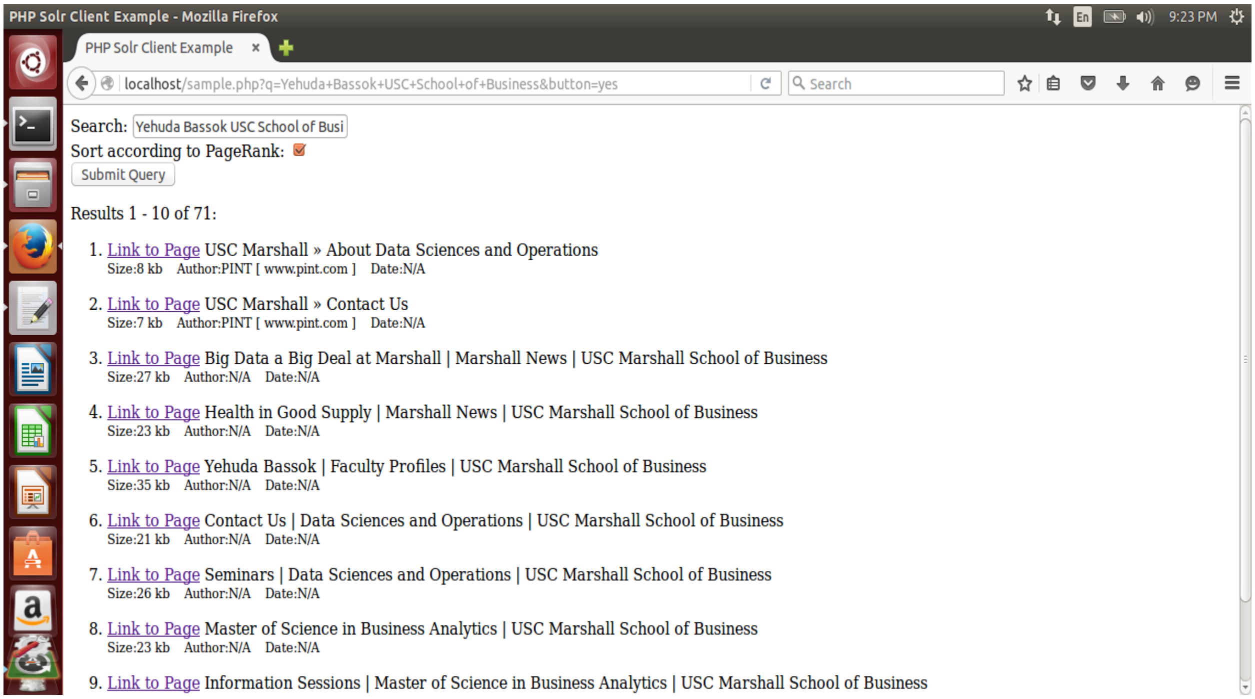
USC Marshall School of Business
Los Angeles, CA 90089-0808

Phone: 213-821-1140

ybassok@marshall.usc.edu

Education: PhD, Carnegie Mellon University; BA, BS, The Hebrew University

Solr Page Link Result


PHP Solr Client Example - Mozilla Firefox
9:23 PM
PHP Solr Client Example
localhost/sample.php?q=Yehuda+Bassok+USC+School+of+Business&button=yes
Search
Submit Query

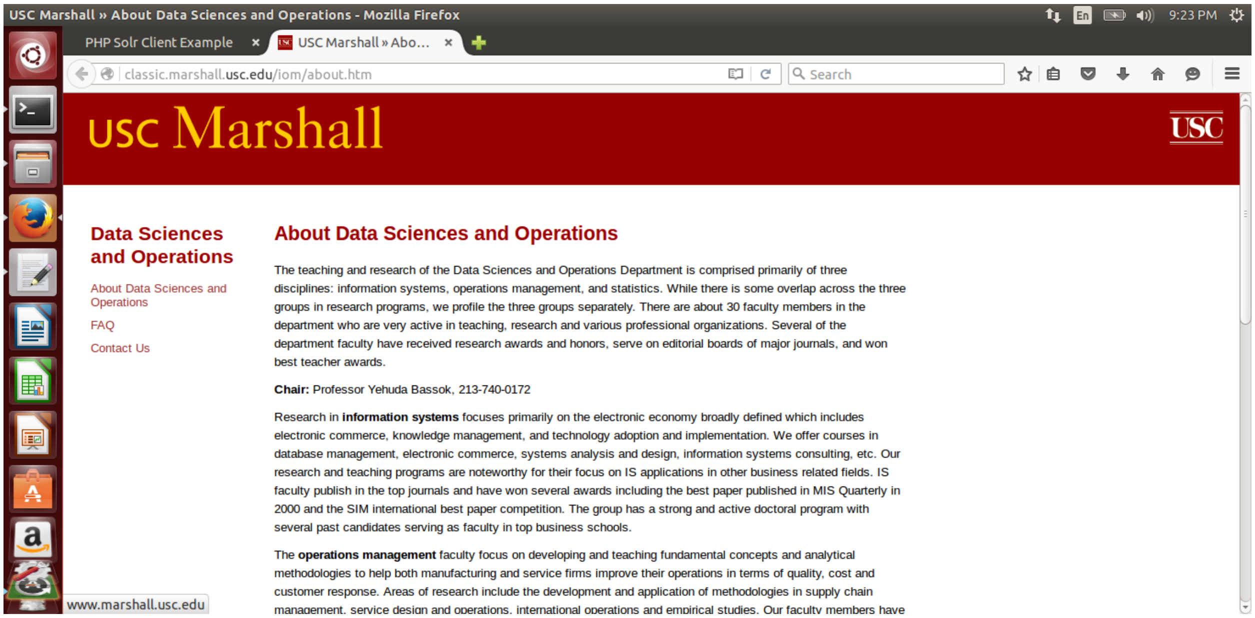
Search: Yehuda Bassok USC School of Busi

Sort according to PageRank: ☑

Results 1 - 10 of 71:

1. Link to Page USC Marshall » About Data Sciences and Operations
   Size:8 kb Author:PINT [www.pint.com] Date:N/A
2. Link to Page USC Marshall » Contact Us
   Size:7 kb Author:PINT [www.pint.com] Date:N/A
3. Link to Page Big Data a Big Deal at Marshall | Marshall News | USC Marshall School of Business
   Size:27 kb Author:N/A Date:N/A
4. Link to Page Health in Good Supply | Marshall News | USC Marshall School of Business
   Size:23 kb Author:N/A Date:N/A
5. Link to Page Yehuda Bassok | Faculty Profiles | USC Marshall School of Business
   Size:35 kb Author:N/A Date:N/A
6. Link to Page Contact Us | Data Sciences and Operations | USC Marshall School of Business
   Size:21 kb Author:N/A Date:N/A
7. Link to Page Seminars | Data Sciences and Operations | USC Marshall School of Business
   Size:26 kb Author:N/A Date:N/A
8. Link to Page Master of Science in Business Analytics | USC Marshall School of Business
   Size:23 kb Author:N/A Date:N/A
9. Link to Page Information Sessions | Master of Science in Business Analytics | USC Marshall School of Business

PageRank Search Results


USC Marshall » About Data Sciences and Operations - Mozilla Firefox
En
9:23 PM
PHP Solr Client Example
USC Marshall » Abo…
classic.marshall.usc.edu/iom/about.htm
Search
usc Marshall
USC
www.marshall.usc.edu

Data Sciences and Operations

About Data Sciences and Operations

FAQ

Contact Us

About Data Sciences and Operations

The teaching and research of the Data Sciences and Operations Department is comprised primarily of three disciplines: information systems, operations management, and statistics. While there is some overlap across the three groups in research programs, we profile the three groups separately. There are about 30 faculty members in the department who are very active in teaching, research and various professional organizations. Several of the department faculty have received research awards and honors, serve on editorial boards of major journals, and won best teacher awards.

Chair: Professor Yehuda Bassok, 213-740-0172

Research in **information systems** focuses primarily on the electronic economy broadly defined which includes electronic commerce, knowledge management, and technology adoption and implementation. We offer courses in database management, electronic commerce, systems analysis and design, information systems consulting, etc. Our research and teaching programs are noteworthy for their focus on IS applications in other business related fields. IS faculty publish in the top journals and have won several awards including the best paper published in MIS Quarterly in 2000 and the SIM international best paper competition. The group has a strong and active doctoral program with several past candidates serving as faculty in top business schools.

The **operations management** faculty focus on developing and teaching fundamental concepts and analytical methodologies to help both manufacturing and service firms improve their operations in terms of quality, cost and customer response. Areas of research include the development and application of methodologies in supply chain management, service design and operations, international operations and empirical studies. Our faculty members have

PageRank Url Page