Web Search Engine Comparison

The exercise is about comparing the search results from Google versus Bing, the two leading US search engines. Many search engine comparison studies have been done. All of them use samples of data, some small and some large, so no general conclusions can be drawn. But it is always instructive to see how the two search engines match up, even on a small data set.

The process you will follow is to issue a set of queries and to evaluate the returned results for relevance. These studies do not seek to answer the ultimate question of which search engine is “best”. Rather we stick to more modest research questions which are:

RQ1: Which search engine performs best when considering the first 10 results for a given query?

RQ2: Is there a difference in relevance between the search engines when considering informational queries and navigational queries, respectively?

To begin I have divided the class across the set of Schools at USC. I have pre-assigned the school you are assigned to according to your USC ID number, as given in the table below.

<table>
<thead>
<tr>
<th>USC ID ends with</th>
<th>School to crawl</th>
<th>Root URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>01~14</td>
<td>Architecture</td>
<td><a href="http://arch.usc.edu/">http://arch.usc.edu/</a></td>
</tr>
<tr>
<td>15~26</td>
<td>Cinematic Arts</td>
<td><a href="http://cinema.usc.edu/">http://cinema.usc.edu/</a></td>
</tr>
<tr>
<td>27~38</td>
<td>Dornsife (College)</td>
<td><a href="http://dornsife.usc.edu/">http://dornsife.usc.edu/</a></td>
</tr>
<tr>
<td>39~49</td>
<td>Gould (Law)</td>
<td><a href="http://gould.usc.edu/">http://gould.usc.edu/</a></td>
</tr>
<tr>
<td>50~60</td>
<td>Keck (Medicine)</td>
<td><a href="http://keck.usc.edu/">http://keck.usc.edu/</a></td>
</tr>
<tr>
<td>61~71</td>
<td>Marshall (Business)</td>
<td><a href="http://www.marshall.usc.edu/">http://www.marshall.usc.edu/</a></td>
</tr>
<tr>
<td>72~78</td>
<td>Viterbi (Engineering)</td>
<td><a href="http://viterbi.usc.edu/">http://viterbi.usc.edu/</a></td>
</tr>
<tr>
<td>79~88</td>
<td>Price (Public Policy)</td>
<td><a href="http://priceschool.usc.edu/">http://priceschool.usc.edu/</a></td>
</tr>
<tr>
<td>89~00</td>
<td>Social Work</td>
<td><a href="http://sowkweb.usc.edu/">http://sowkweb.usc.edu/</a></td>
</tr>
</tbody>
</table>

Now that you have been assigned a USC School, here are the queries you will submit:
Input **Navigational** Queries: Devise a set of queries for your USC School as follows; in this case I am assuming I have been assigned to the Viterbi School of Engineering; here are five navigational queries and three informational queries

1. Choose 3 Faculty names from your school and enter the following query using the names from your school, e.g. “Ellis Horowitz USC School of Engineering” or “David Cruz USC Gould School of Law” (do NOT use quotes in your query)
2. Choose 3 Faculty departments, e.g. “Computer Science USC School of Engineering”, or if there is no department use a division name, e.g. “Director of Admissions, USC Gould School of Law”. If there are no departments or divisions, come up with a suitable categorization on your own.
3. School Location, map and directions, e.g. “Viterbi School of Engineering USC map and directions”
4. Founder’s Name for the school, if there is one, e.g. “USC School of Engineering founder Viterbi” or “USC School of Business founder Marshall”; if there is no founder you can skip this part.
5. Alumni News web page, e.g. “USC School of Engineering Alumni” or “USC Gould School of Law Alumni”

Input **Informational** Queries: Devise a set of queries for your USC School as follows

1. Requirements for an undergraduate degree in a given department, e.g. “USC Computer Science Undergraduate degree requirements”
2. Requirements for a Masters degree in a given department, e.g. “USC Computer Science Masters degree requirements”
3. Requirements for a Ph.D. degree in a given department, e.g. “USC Computer Science Ph.D. degree requirements”
   If your School does not offer an undergraduate or Masters degree, devise a query for whatever degree(s) are offered.

Place your queries in a separate text file which you will submit with the rest of your assignment.

For each result you should compute a relevance score as follows:

**For faculty names** relevance = 1 for a search result to the faculty’s home page; relevance = 0.5 for course page taught by the faculty member, and relevance = 0.25 for a page with only a little information about the faculty member, and otherwise relevance = 0;

**For faculty departments or divisions** relevance = 1 for a search result to the department’s home page, relevance = 0.5 for an page that is internal to the department and otherwise relevance = 0;

**For school location**, relevance = 1 for a search result containing map and directions, otherwise relevance = 0; note that a Google map that provides the exact building location is as relevant as a USC campus map.

**For founder’s name** relevance = 1 for a search result that describes the founder, relevance = 0.5 for a page that gives the history of the school and mentions the founder, and otherwise relevance = 0;

**For alumni news web page** relevance = 1 for a result that points to an alumni news page; if one does not exist, then relevance is 0.5, but if one exists and is not returned, then relevance is 0.
For the informational queries relevance = 1 if the page describes the requirements, relevance = 0.5 if it contains a link to the actual requirements, and otherwise relevance = 0.

Note: in the event that your Google account enables personalized search, please turn this off before performing your tests.

Output

Once you score all of the search results for all of the queries you should produce the following statistics.

1. A text file containing the list of queries that you used and a general conclusion about the effectiveness of Google and Bing based upon your experiment; you should address items RQ1 and RQ2 defined at the beginning of this exercise.

2. An Excel or Google docs spreadsheet showing the following:
   2.1 For each query the relevance scores for Google and Bing spreadsheets; there will be ten scores each;
   2.2 For each query a number representing the overlapping search results
   2.3 A graph showing the percentage of result overlap between the search results of the two search engines. Results are assumed to overlap if the identical link is contained in the top 10 results. A bar graph where the x-axis represents query 1, 2, 3, etc and the y-axis represents the number of overlapping results, up to a maximum of 10.
   2.4 For navigational queries, a bar graph showing the ratio of relevant vs. irrelevant pages at the top position. See pp. 7, figure 1 of the Lewandowski paper
   2.5 For informational queries, a bar graph showing the ratio of relevant vs. irrelevant pages in the top 10 results. See pp 8, Figure 2 of the Lewandowski paper

The precision is computed as the fraction of retrieved instances that are relevant. For our purposes, any score of 0.5 or higher is deemed to be relevant.

Note1: you should use a spreadsheet program to produce the above data, e.g. Microsoft Excel.

Note2: place all of your results on a single sheet of the spreadsheet

References


Submission
You are required to submit your results electronically to the csci572 account on SCF so that it can be graded. To submit your file electronically, enter the following command from your Unix prompt:

```
submit -user csci572 -tag hw1 MYFILE1 MYFILE2
```

where MYFILE1 contains your queries and MYFILE2 contains your results.