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.

.

USC ID ends with	School to crawl	Root URL
01~14	Architecture	http://arch.usc.edu/
15~26	Cinematic Arts	http://cinema.usc.edu/
27~38	Dornsife (College)	http://dornsife.usc.edu/
39~49	Gould (Law)	http://lawweb.usc.edu/
50~60	Keck (Medicine)	http://keck.usc.edu/
61~71	Marshall (Business)	http://www.marshall.usc.edu/
72~78	Viterbi (Engineering)	http://viterbi.usc.edu/
79~88	Price (Public Policy)	http://www.usc.edu/schools/price/
89~00	Social Work	http://sowkweb.usc.edu/

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”
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”
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;

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

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.

Output

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

A graph showing the **percentage of result overlap** between the search results of different 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.

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

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.

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

References

Evaluating the retrieval effectiveness of Web search engines using a representative query sample" by Dirk Lewandowski, Hamburg University of Applied Sciences, Journal of the American Society for Information Science and Technology, 2013

http://arxiv.org/ftp/arxiv/papers/1405/1405.2210.pdf

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.