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 the class is divided across the set of Schools at USC. Students are pre-assigned according to their USC ID number, as given in the table below.

.

USC ID ends with	School to crawl	Root URL
01~20	Dornsife (College)	http://dornsife.usc.edu/
21~40	Gould (Law)	http://gould.usc.edu/
41~60	Keck (Medicine)	http://keck.usc.edu/
61~70	Marshall (Business)	http://marshall.usc.edu/
71~80	Viterbi (Engineering)	http://viterbi.usc.edu/
81~00	Price (Public Policy)	http://priceschool.usc.edu/

Now that you have been assigned a USC School, below are the queries you will submit. There are a total of nine (3+3+1+1+1) navigational queries, three informational queries, and one final query.

Input Navigational Queries: Devise a set of queries for your USC School as follows;

- Choose 3 Faculty names from your school and enter the following query using the names from your school, e.g. "Ellis Horowitz Viterbi" or "David Cruz Gould" or "Tara Blanc Price" (do NOT use quotes in your query; include only the faculty name and the school name)

Determine relevance for each individual faculty name; do not average over the three names;

- Choose 3 Faculty departments, e.g. “Computer Science Viterbi”, or if there is no department use a division name, e.g. “Director of Admissions, Gould”. If there are no departments or divisions, come up with a suitable categorization on your own.

  Determine relevance for each individual department name; do not average over the three names;
- School Location, a map, e.g. “Viterbi USC map” or “Price USC map”
- The USC School of Engineering is named after Andrew Viterbi, the USC School of Business is named for Gordon S. Marshall; the USC School of Public Policy is named for Sol Price, etc. Issue a query to find a USC web page describing the individual who has named the school, e.g. “Andrew Viterbi”, “Gordon Marshall”, “Sol Price”; the web page can be a Wikipedia entry or a USC web page.
- Alumni News web page, e.g. “USC Viterbi Alumni” or “USC Gould Alumni”.

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

- Requirements for an undergraduate degree in a given department or if there are no departments than simply the requirements for an undergraduate degree, e.g.
  “USC Computer Science Undergraduate degree requirements”
- Requirements for a Masters degree in a given department or if there are no departments than simply the requirements for a Masters degree , e.g.
  “USC Computer Science Masters degree requirements”
- Requirements for a Ph.D. degree in a given department or if there are no departments than simply the requirements for a Ph.D. degree or whatever the most advanced degree that is offered , e.g.
  “USC Computer Science Ph.D. degree requirements”
  If your School does not offer an undergraduate, Masters, or Ph.D. degree, devise a query for whatever degree(s) are offered.

Query 13: Attempt to create a query for your USC school that produces no common matches in Google and Bing within the first five results.

Note: do not alter the above queries so more relevant results are returned; use only the queries as specified above since they are typical of what a casual user might enter.

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[1]; 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;

[1] Notes on special cases: a professor may have more than one home page, perhaps one created by him and one created by his department; both may receive a relevance score of 1; to receive a relevance score of 1, the homepage must have a usc.edu domain; links to external sites such as a LinkedIn entry for a professor is not

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/or directions, otherwise relevance = 0; note that a Google map that provides the exact building location is as relevant as a USC campus map.

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

For alumni news web page relevance = 1 for a result that points to an alumni news page; if one exists and is not returned, then relevance is 0. A returned page that talks about the school's alumni get a relevance of 1. A page describing a specific alum gets a relevance of 0.25.

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;

2.2 For each query a number representing the number of 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[2]. A bar graph where the x-axis represents query 1, 2, 3, etc and the y-axis represents the percentage 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

---

considered a home page, though it can be recorded with relevance 0.5; a resume or CV is not considered a home page, but may get relevance = 0.25

[2] If Google and Bing show different URLs, but they point to the identical page, this should be considered as an overlap

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

Note3: do not reformulate your queries in such a way that the search engine produces more relevant results; the point of the exercise is to examine the results when a "normal" query (as defined above) is entered

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.