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:

- Which search engine performs best when considering the first five results for a given query?
- Is there a difference in relevance between the search engines when considering informational queries and navigational queries, respectively?

THE USC SCHOOLS

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)[1]	http://dornsife.usc.edu/
21~40	Gould (Law)[2]	http://gould.usc.edu/
41~60	Keck (Medicine)[3]	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/

[1] Founder of the College is David and Dana Dornsife

[2] Since, there are no departments in Gould, you can consider the different programs offered as divisions, like: Business Law Program; Media Entertainment & Technology Program; Alternative Dispute Resolution Program; Also, the founder of the Law School is James Gould, found here: http://gould.usc.edu/about/history/timeline/

[3] In Keck for faculty departments you can use: Department of Anesthesiology Keck; Department of Dermatology Keck; Department of Emergency Medicine Keck; the founder of the school is the W.M. Keck Foundation or its namesake, William Myron Keck.

THE QUERIES

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 nine 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. Your query should be exactly as shown above.)

  Determine relevance (see below for how to 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. Your query should ONLY contain the department or division name followed by the school name, and no extra keywords. To determine relevance for each individual department/division name, do not average over the three names;
- *Determine School Location*, a map, e.g. “Viterbi USC map” or “Price USC map”. Your query should be exactly as shown, the school name, USC followed by the word “map”.
- *Determine the Founder:* 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 web page describing the individual who has named the school, e.g. “Andrew Viterbi USC”, “Gordon Marshall USC”, “Sol Price USC”; the web page can be a USC page, or if not, a Wikipedia entry. Your query should contain ONLY the name of the founder of the school and USC.
- *Determine School Alumni News* web page, e.g. “USC Viterbi Alumni” or “USC Gould Alumni”. Your query should only contain “USC” followed by the name of the school, followed by the word “Alumni”.

Input Informational Queries: Devise a set of three 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 Google includes in its top five results, but Bing does not include in its top five results, or vice-versa (Bing includes it, but Google does not). It is supposed to be an entirely different query from the earlier ones. It is anything you can make up.

Note1: 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.

Note2: do NOT consider ad results, we are only concerned with the organic (non-ad) search results; ignore ads that are placed at the top of the search results page

DETERMINING RELEVANCE

Each of your thirteen queries should be run on both Google and Bing. You should capture the top five results (the URL) for each query. For each of the top 5 results for each query you should compute a relevance score as follows:

For faculty names relevance = 1 for a search result to the faculty's home page[4]; 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 a 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 founder'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.

For the 13th query, you can use your own best judgement to determine the relevance to assign to the results of this query.

Note3: in the event that your Google account enables personalized search, please turn this off before performing your tests.

[4] 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 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

Note4: For ambiguous/not mentioned cases please use your best judgment when choosing the scores. Make sure to be consistent across search engines. As long as you follow a consistent scoring that makes sense that is considered to be acceptable.

Output

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

1. An Excel or Google docs spreadsheet showing the following:

the list of queries that you used and for each query the top five URLs produced as results, and for each URL the relevance score that you assigned. The data should include the results for both Google and Bing using the following column headings:

QUERY 1	" . . . . . "			
	Google Results	Relevance Score	Bing Results	Relevance Score
1.	URL1		URL1	
2.	URL2		URL2	
3.	URL3		URL3	
4.	URL4		URL4	
5.	URL5		URL5	

2. In addition to the above data you need to provide:

2.1 Thirteen bar graphs, one for each query, with Y-axis from 0 to 1 and X-axis the top five results; the value for each result is the relevance score for Google and Bing; so your bar graph should have ten bars

2.2 A single bar graph whose Y-axis is 0 to 5 and whose X-axis is query 1, query 2, . . . , query 13 and whose value for each query is the number of overlapping search results for that query. Results are assumed to overlap if the identical link is contained in the top 5 results[5].

2.3 Two bar graphs, one for Google and the other for Bing; each bar graph has query 1 through query 9 on the x-axis, and for each query there are two bars, one showing the number of relevant pages for that query and the other showing the number of irrelevant pages for that query. A page is considered relevant if its relevance score is greater than 0. A page is considered irrelevant if its score is 0.

2.4 Two bar graphs, one for Google and the other for Bing; each bar graph has query 10, query 11 and query 12 (the three informational queries) on the x-axis, and for each query there are two bars, one showing the number of relevant pages for that query and the other showing the number of irrelevant pages for that query. A page is considered relevant if its relevance score is greater than 0. A page is considered irrelevant if its score is 0.

Note5: place all of your results on a single sheet of the spreadsheet

[5] If Google and Bing show different URLs, but they point to the identical page, this should be considered as an overlap; if the same URL occurs twice in the top five results, it should be counted twice.

Note6: 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

Finally, provide a one sentence answer to the two questions posed at the beginning of this exercise.

- Which search engine performs best when considering the first five results for a given query?
- Is there a difference in relevance between the search engines when considering informational queries and navigational queries, respectively?

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  MYNAME.XSLX
```

where MYNAME (use your own login name) contains your results.