Web Search Engine Comparison

This 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 a more modest research question which is:

- which search engine performs best when considering the first five results for a given small set of queries?

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.

Note: Please stick with the assigned schools according to your ID number and don’t work on another school and later ask for an exception.

USC ID ends with	School to query	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/

THE QUERIES

Now that you have been assigned a USC School, below are the queries you will submit. There are a total of six queries.

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

[2] The founder of the Law School is James Gould, found here: http://gould.usc.edu/about/history/timeline/

[3] For Keck the founder of the school is the W.M. Keck Foundation or its namesake, William Myron Keck.

Queries:

- *Choose 2 Faculty names from your school and try to find their home page*. 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 any query**; **do NOT include the name USC**; include only the faculty name and the school name. Your query should be exactly as shown above, but without the quotes.) See below for how to determine relevance;
- *Locate the home page for an academic department or concentration area*. For the school you are assigned issue the query below:

School	**Example Query**
Dornsife	USC Dornsife Economics
Gould	USC Gould Dispute Resolution
Keck	USC Keck Preventive Medicine
Marshall	USC Marshall Marketing
Viterbi	USC Viterbi Computer Science
Price	USC Price Public Policy

- *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.

School	**Query**
Dornsife	David Dornsife USC
Gould	James Gould USC
Keck	William Myron Keck USC
Marshall	Gordon Marshall USC
Viterbi	Andrew Viterbi USC
Price	Sol Price USC

- *Determine requirements for degrees:* some schools offer undergraduate degrees, e.g. Dornsife and Viterbi, while others primarily offer graduate degrees, e.g. Marshall offers an MBA and Keck offers an MD degree. Depending upon which school you have been assigned, here is the query to be issued:

School	**Query**
Dornsife	USC economics undergraduate degree requirements
Gould	USC school of law JD degree requirements
Keck	USC school of medicine MD degree requirements
Marshall	USC school of business MBA degree requirements
Viterbi	USC school of engineering computer science undergraduate degree requirements
Price	USC school of public policy master's degree requirements

Note 1: 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.

Note 2: 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 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 pointing to the faculty's home page[4]; relevance = 0.5 for a 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 (or division's) home page, relevance = 0.5 for a page that is internal to the department (or division) and otherwise relevance = 0;

For school location, relevance = 1 for a search result containing a 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 the "requirements" query 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 3: In the event that your Google account enables personalized search, please turn this off before performing your tests.

Note 4: For ambiguous/not mentioned cases please use your best judgment when choosing the relevance 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 the search results for all the queries you should produce the following statistics.

[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

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
Result 1.	URL1		URL1	
Result 2.	URL2		URL2	
Result 3.	URL3		URL3	
Result 4.	URL4		URL4	
Result 5.	URL5		URL5	

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

2.1 Six bar graphs, one for each query, with Y-axis from 0 to 1 and X-axis results 1, 2, 3, 4, and 5; the value for each result is two bars, the relevance score for Google and the relevance score for 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 11 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 You will compute a form of Discounted Cumulative Gain for Google and Bing using the formula:

for each query i, i from 1 to 11, DCG(i) = $\Sigma_{\text{j = 1 to 5}}$ RelevanceScoreOfResult(j) / $\log_2$ (j+1)

and then the final DCG = $\Sigma_{\text{i = 1 to 11}}$ (DCG(i)). Make sure you provide a final DCG for both Google and Bing.

Note 5: Place all your results on a single sheet of the spreadsheet

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

- which search engine performs best when considering the first five results for your set of queries?

Points to note:

1. If the professor doesn't have a home page in usc.edu domain, then you can give a score of 1 to a page in isi.edu (as domain belongs to USC). However, if the professor has home page in usc.edu domain as well as isi.edu, please give score of 1 to USC webpage and 0.75 to ISI webpage.

[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.

2. Provide the actual URLs so that the graders can verify if the URLs are legitimate and the relevance scores have been assigned correctly.
3. For the degree requirements, sometimes we get links to older catalogues (for example year 2014), you may assign a relevance score of less than 1, since they are not relevant today.
    a. Note: Degree requirements here mean the number of credits required to complete, compulsory courses, etc and not the application requirements (ie requirements to be satisfied for applying to the program)
4. If the first link to the map query is Bing images, and if the images are serving the purpose, i.e, if you can locate the school you are searching for, then it can be relevant.
5. Ignore snippets and advertisements returned in your search results.
6. **Since this is an experimental exercise, as long as the relevance scores are consistent across search engines, it is acceptable. You can even mention the relevance rule that you followed in your report i.e., why you have given that score. We'll not be deducting points for the relevance scores assigned, as long as the scoring is consistent.**
7. If some URL is unreachable, Please ignore that and consider next result.
8. If two of the top-5 results returned the same URL, consider both of them .
9. If you get a link to the Rate My Professor website as result, you can assign the relevance score 0.25 as long as it is related to the professor and his courses.
10. If a professor has 2 pages and both have the same content and both are in the usc.edu domain, You can assign relevance score of 1 to both of them.
11. Search Engines may not return the same results for the same query issued multiple times. Search results can change based on your previous search, your search history, location etc. Run the query and take down the results at the same time.
12. If 2 urls only differ in "http" and "https", you can consider them as an overlap when counting overlaps.
13. Google and Bing may offer knowledge graph cards on the right side of results, for the sake of consistency ignore these.

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.xlsx
```

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

You can use either cs-server.usc.edu or aludra.usc.edu to submit your results

You will get a “SUCCEEDED” message after successful submission of the homework.

You can submit your homework as many times as you want as long as you don’t surpass the deadline. In case of multiple submissions, your previous submission will be overwritten.