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 more modest research questions which are:

- which search engine performs best when considering the first five results for a given query?

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 other school and later ask for exceptions.

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/

[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 eleven queries, and one final query.

Queries:

- *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 any query**; include only the faculty name and the school name. Your query should be exactly as shown above, but without the quotes.)

  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.
- *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.

Final Query: Attempt to create a query for your USC school where Google's top five results are entirely different from Bing's top five results. It is supposed to be an entirely different query from the earlier queries. It is anything you can make up. For this query there should be no overlap between top 5

results of Google and Bing. You can use your own best judgement to determine the relevance to assign to the results of this query.

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" 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 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 Eleven 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) = $\sum_{j=1 \text{ to } 5}$ RelevanceScoreOfResult(j) / $\log_2$ (j+1)

and then the final DCG = $\sum_{i=1 \text{ to } 11}$ (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

Note 6: 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 question posed at the beginning of this exercise.

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

[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.

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.
2. Put the actual URLs so that the graders can verify if the URLs are legitimate and the relevance scores have been assigned correctly.
3. For Gould school, you can consider areas of concentration as division. e.g. Business Law, Gould
4. When the query is about a professor or a department and we get the results as a map, video or image, you may include the result in top 5 and assign relevance.
5. If it shows the link with the map of the building, not a direction to the building, you may not assign any relevance.
6. Since there are no departments in Gould school you may use either USC School of Law Masters requirements or USC Gould Masters requirements.
7. For the degree requirements, sometimes we get links to older catalogues (for example year 2014), you may assign a relevance score of < 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)
8. 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.
    a. Note: Consider images only for map query and not for other queries like alumni, faculty, etc.
9. For Gould School, Instead of Ph.D. you can use Gould's any other degree program. For example, you can use Gould "Dual degree program".
10. Please ignore snippets and advertisements returned in your search results.
11. **Since this is an experimental exercise, as long as the relevance scores are consistent across search engines, it is acceptable.** **You can even mention about your judgment in the 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.**
12. If some URL is unreachable, Please ignore that and consider next result.
13. If two of the top-5 results returned the same URL, consider both of them .
14. If you get rate my professor website as result, see if you find some useful information regarding the courses taken by professor and rating for those courses and you can assign the relevance score 0.25 if it is related to the professor and his courses.
15. If a prof 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.
16. Search Engines are not consistent. That is expected browser behavior. Search results do change time to time based on your previous search, your search history,loaction etc. Run the query and take down the results at the same time.

17. If 2 urls only differ in "http" and "https", you can consider them as an overlap when counting overlaps.
18. 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.