Rules. This is to be handed out no later than the start of class on the day it is due. Earlier submission during office hours is also possible. For the questions of part 2 you are explicitly allowed to search the Web to find auxiliary information that will help you formulate answers to the questions asked; however such answers must be justified as asked.

Due Date: No later than start of class on Mon Sep 26, 2011.

Topic: Bing and Google query language comparison

1 Objectives of the assignment

You will use two search engines, `Google` (`www.google.com`) and `Bing` (`www.bing.com`) to perform some search queries, understand the search results obtained, and evaluate the quality of the answers based on your original expectations. You will also estimate the size of the document corpus of each one of the two search engines.

2 Conventions

When we ask you to type in `course` in the query (a.k.a. text) box of a search engine, we will represent this in this document as ⟨`course`⟩. You do not need to type the left and right brackets ⟨ and ⟩, just the text in-between, which for this example is `course`. Case will also be significant. You will be expected to type in `course` in the case indicated rather than arbitrarily picking case such as typing `Course` or `COURSE`. If we ask you to type in ⟨"`course`"⟩, this will mean that you must type the word `course` enclosed in double-quotes, i.e. you type a double-quote followed by the word `course` followed by another double-quote. If a query contains a minus (-) sign, this immediately precedes the following letter with no space inbetween.

3 Requirements

1. Present the queries specified in Part 1 to the two search engines and figure out how to locate in each search engine the number of web-page hits for each query and record your results in tabular form as specified in the relevant Table and in the order indicated. Do not swap rows or columns! Make also sure that figures are presented as requested (in multiples of millions).
2. Answer the questions in Part 2 relevant to the queries of Part 1. You can navigate around the help systems of the two search engines and learn more about them and find out how they work. You are also explicitly allowed to search the Web and try to find answers for the questions of Part 2 (but you are not allowed to obtain past solutions from this instructor that might provide hints to some of these questions).
3. Answer the questions of Part 3.

4 Part 1 (52 points)

In the two search engines `Google` and `Bing`, perform the following searches and record the number of hits for each such search. Tabulate the results in a form similar and order identical to that of the sample table given. It's imperative that you do not swap the order of the queries in the table of your answer sheet. The first column is a reference to a specific query by index, the second column is the query that needs to be performed (per the outlined convention of section 2). You need to report the number of hits observed in columns 3 and 4 in millions

Advice. Collect the results twice (or thrice) one day/one hour apart.

No	Query Term(s)	GOOGLE (number of hits in '000,000)	BING (number of hits in '000,000)
1	⟨`a b or c`⟩		
2	⟨`a b OR c`⟩		
3	⟨`A B OR C`⟩		
4	⟨`a (b OR c)`⟩		
5	⟨`(a b ) OR c)`⟩		
6	⟨`a or b`⟩		
7	⟨`a OR b`⟩		
8	⟨`b or a`⟩		
9	⟨`b OR a`⟩		
10	⟨`a not b`⟩		
11	⟨`a NOT b`⟩		
12	⟨`a NOT a`⟩		
13	⟨`a −a`⟩		

Table 1: Table for Part 1

5 Part 2 (32 points)

Answer the following questions.

Q2.1 With reference to queries 1 and 2, does capitalization matter ? Why? Explain.

Q2.2 With reference to queries 2 and 3, does capitalization matter ? Explain.

Q2.3 With reference to queries 1-5, which queries are equivalent? Explain.

Q2.4 Is there a difference between `OR` and `or` for the two engines? Explain.

Q2.5 Is there a difference between `NOT` and `not` for the two engines? Explain.

Q2.6 What do queries 12 and 13 ask (note that there is no space between the - and the `a`)?

Q2.7 Can you identify clusters (i.e. groups) of queries that return approximately the same number of answers for Google (separately), Bing (separately), and Google and Bing (combined)?

Q2.8 Can you identify clusters of queries in which one engine returns a constant multiple more queries than the other? What is that constant?

6 Part 3 (41 points)

Q3.1 Which of the two search engines returns the highest number of results?

Q3.2 Based on Q3.1 and Part 1 what is an estimate of the web-pages indexed by the two engines?

Q3.3 Formulate a query other than the ones in Part 1/2/3 that would return a number of pages for Google greater than or equal to the maximum of the results reported in Part 1.

Q3.4 Formulate a query other than the ones in Part 1/2/3 that would return a number of pages for Bing greater than or equal to the maximum of the results reported in Part 1.

Q3.5 Based on Q3.1-Q3.4 what is your esitamte for the size of Google's and Bing's corpus?

Q3.6 Perform the following 3 queries in Google and Bing. How many hits did you get? Can you explain any discrepancies? (9 points)

No	Query Term(s)	GOOGLE (number of hits in '000,000)	BING (number of hits in '000,000)
14	⟨`to be or not to be`⟩		
15	⟨`TO BE OR NOT TO BE`⟩		
16	⟨`to be OR not to be`⟩		

Table 2: Table for Q3.6

Q3.7 Perform the query below in Google and Bing. How many hits did you get? Can you explain what it asks? And why are there as many hits? (8 points)

No	Query Term(s)	GOOGLE (number of hits in '000,000)	BING (number of hits in '000,000)
17	⟨ `(a AND b) OR ( −a OR − b )` ⟩		

Table 3: Table for Q3.7

Note. Search engine behavior is sometimes erratic. Repeat your experiments twice or thrice at different times and from different browsers. Get connected to the engines either through the toolbar of or directly through a browser.