COMP 4321 Search Engines for Web and Enterprise Data
Homework 1
Due: Oct 19, 2014

1. [25 points] (Open questions, no fixed answers)

   (a) Type the query “comp4321 project” in Google, Bing, and Yahoo. List the first 10 results returned by the above search engines.
   - ● How do you judge which search engine is better?
   - ● Which search engine has better performance according to your experience? Explain with example.

   (b) Can you give any suggestions of query formulations so that you can get more related results?

2. [15 points]

   Suppose there are only 5 unique terms, t_1 to t_5, in the collection, which contains a total of 100 documents. The term’s term frequencies in a document D and their document frequencies are given below:

   $tf_{D,1} = 2$ $df_1 = 10$
   $tf_{D,2} = 0$ $df_2 = 20$
   $tf_{D,3} = 0$ $df_3 = 30$
   $tf_{D,4} = 5$ $df_4 = 20$
   $tf_{D,5} = 2$ $df_5 = 1$

   - Write down the document vector for D when idf * tf/tf_max weighting method is used, where tf_max is the largest term frequency in a document
   - Given the query vector, $Q = \langle 1, 1, 1, 0, 1 \rangle$, compute the inner product and cosine similarity values between Q and D. Which of D’s terms contributes the most to the similarity scores?

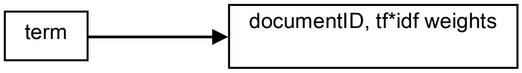
3. [40 points]

   A small document collection contains only the following three short documents:

   (a) Give the document vectors for the three documents using tf *idf weights. All words are indexed.
   D1: to be or not to be
   D2: to live or not to live
   D3: publish or perish

	be	live	not	or	perish	publish	to
D1							
D2							
D3							

   (b) Show the inverted file structure for the documents as in the form below. Sort the terms in ascending order in the inverted file.



   (c) Write the pseudo code for calculating the cosine similarity between a query and each of the documents based on the inverted list constructed in (b).

(d) Calculate the centroid of the document set and calculate the Jaccard coefficient values between the query Q=<0, 1, 1, 0, 1, 1, 0> and the centroid.

4. [20 points]

We have not (and will not) discuss the extended Boolean model slides in lecture. Read the file http://www.suntek.com.hk/031000005/1.pdf (also accessible from the course homepage), the corresponding slides https://home.cse.ust.hk/~dlee/4321/Password_Only/extended-Boolean.ppt, and answer the following question:

- Someone uses the query `apple AND juice` to find information about apple juice.
- Consider the following pages:
    1. http://global.britannica.com/EBchecked/topic/30599/apple
    2. http://nutritiondata.self.com/facts/fruits-and-fruit-juices/1822/2
- What are the tf of apple and juice in the above pages (Use Chrome to load the page and CNTL F to find a word and you will see the count of the word)?
- Given that the most frequent word in the above pages is apple, what are the weights ($w_{x,j}$) of apple and juice in the above pages according to Slide 2, assuming that we use tf only (since there is no DF information)?
- What are the similarity scores of the query to the above pages according to Slide 3?
- Discuss the advantage of the extended Boolean model compared to the vector space model in answering the query, assuming that the vector space model simply uses tf as term weight.