DATA20021 Information Retrieval (Spring 2021)
Exercise Set 1 (due date is Friday, 29 January).

Answer at least 6 out of the 8 questions below *well* to get full marks.
Solutions are to be submitted on Moodle.

1. Draw the inverted index that would be built for the following document collection.

   Doc 1 new pants sales top forecasts
   Doc 2 pants sales rise in july
   Doc 3 increase in pants sales in july
   Doc 4 july new pants sales rise

2. Consider these documents:

   Doc 1 breakthrough drug for peace
   Doc 2 new peace drug
   Doc 3 new approach for peace in our time
   Doc 4 new hopes for peace process

   (a) Draw the term-document incidence matrix for the collection.

   (b) Draw the inverted index representation of the collection.

   (c) What are the returned results for these queries:

      i. peace **AND** drug
      ii. for **AND NOT** (drug **OR** approach)

3. In the lecture we went through an algorithm for intersecting two lists involved in an **AND** query that ran in time $O(x + y)$. For the queries below, can we still run through the intersection in time $O(x + y)$? If not, what can we achieve?

   (a) Brutus **AND NOT** Caesar

   (b) Brutus **OR NOT** Caesar

4. Recommend a query processing order for the following query:

   (tangerine **OR** trees) **AND** (marmalade **OR** skies) **AND** (kaleidoscope **OR** eyes)

   given the following posting list sizes:

   - eyes: 213312
   - kaleidoscope: 87009
   - marmalade: 107913
   - skies: 271658
   - tangerine: 46653
   - trees: 316812

5. Write pseudocode for an algorithm for an x **OR** y query (in the style of the algorithm in the lecture slides for an x **AND** y query).

6. How should the Boolean query x **AND NOT** y be handled? Why is naive evaluation of this query normally very expensive? Write pseudocode for an algorithm that evaluates this query efficiently.

7. Try using the Boolean search features on a couple of major web search engines. For instance, choose a word, such as hacker, and submit the queries (i) hacker, (ii) hacker AND hacker, and (iii) hacker OR hacker. Look at the estimated number of results and top hits. Do they make sense in terms of Boolean logic? Often they do not. Can you make sense of what is going on? What about if you try different words? For example, query for (i) author, (ii) prize, and then author OR prize. What bound should the number of results from the first two queries place on the third query? Is this what you see?

8. Draw the trie for the set of terms $\mathcal{R} = \{\mathsf{manne}, \mathsf{manu}, \mathsf{minna}, \mathsf{salla}, \mathsf{saul}, \mathsf{sauli}, \mathsf{vihtori}\}$.