Sheetal Parikh
Problem Set – Module 1
605.744.81 Information Retrieval
Fall 2021

1. *List four English stopwords*

   Stopwords are common words that are generally not helpful in identifying relevant documents. Examples of English stopwords are:
   - The
   - And
   - That
   - To

2. *What is the difference between a disjunctive clause and a conjunctive clause in a Boolean query? Give an example of each.*

   A disjunctive clause in a Boolean query would contain the OR operator meaning that we are searching for the documents that contain either term or both terms (for a simple disjunctive clause). Whereas a conjunctive clause in a Boolean query would contain the AND operator meaning that we are searching for documents that only contain both terms (for a simple conjunctive clause). For more complicated queries, for example, we could have conjunctive clause that is a conjunction of disjunctions.

   Examples:
   - Simple conjunctive clause: Brutus AND Calpurnia
   - Simple disjunctive clause: Brutus OR Calpurnia

3. *Chapter 2 of the text shows how a Boolean AND query can be performed for two terms with postings lists lengths of m and n in less than O(m+n) time. What is the method presented by the authors for sublinear postings list intersection and explain why it can result in a time savings?*

   The authors present a method of using a skip list in which at indexing time, skip pointers would be added to a postings list. The skip pointers are shortcuts that reference terms further along in the list. This method would result in a time savings because they would allow us to avoid processing parts of the postings list that would lead to empty results because they would not be included in the intersection or search results.

   Overall, in this method, the postings intersection can use a skip pointer when the end point is less than the item on the other list. For example, using Figure 2.9 of the textbook, if we matched 8 on the list then we move on to 16 on the top list and 41 on the bottom list. We can check the skip list pointer and see that 16 and the following 19, 23, and 28 (on the top list) are all less than 41. Therefore, we can advance the top pointer to 28, avoiding moving to 19 and 23. A simple method for placing skips would be to use sqrt(N) evenly-spaced skip pointers in which N is the length of the postings list.

4. *Exercise 2.1 in IIR: Are the following statements true or false?*

   a. In a Boolean retrieval system, stemming never lowers precision.
   False – Stemming can lower precision because stemming may increase the number of documents that are retrieved but doesn't increase the number of relevant documents

   b. In a Boolean retrieval system, stemming never lowers recall.
   True – Stemming should not lower recall because it would increase the number of documents that are retrieved which could either increase recall or leave it the same

   c. Stemming increases the size of the vocabulary.
   False – Stemming would decrease the size of the vocabulary

   d. Stemming should be invoked at indexing time but not while processing a query
   False – The same processing should be done to both queries and documents to make sure that we get matches and to avoid errors

5. *Suppose we are using a biword index as described in IIR Section 2.4. Give an example of a short plausible English document that is 1 or 2 sentences in length and that would be retrieved for the query "HEAVY METAL BAND", but which is actually a false positive and does not contain all three words in consecutive order. An example true positive document might be: "Motörhead is my favorite heavy metal band."*

   Using the biword index method, consecutive pairs of words in the text would be indexed. For example, if we had text, "Heavy Metal Band", it would be indexed as *heavy metal* and *metal band*. Since we are searching for these two word phrases, we can have many false positives that don't actually contain the entire phrase in the query. Also, the phrases may be from a different context than what the query intended. For example, for the query "HEAVY METAL BAND", a false positive document might be:

   "This metal band that looks like a common bracelet actually contains a toxic heavy metal".