CS160 - Homework 1 solutions

1. (15 points) Find 5 different IR systems

2. (20 points) Find 2 articles in popular media (e.g. nytimes, cnn, digg, slashdot, ...) discussing IR or issues related to IR (e.g. online advertising).

3. (5 points) Describe a set of term postings lists and a boolean query where the heuristic for merging the postings lists of the query terms with the shortest lists first is sub-optimal. State what the optimal query order should be.

   Given the postings lists:
   $w1 : 1 \rightarrow 2 \rightarrow 3$
   $w2 : 4 \rightarrow 5$
   $w3 : 4 \rightarrow 5$

   and the query $w1$ AND $w2$ AND $w3$, the heurstic merging would be $w2$ and $w3$ first. The optimal solution would merge $w1$ with either $w2$ or $w3$ first.

4. (5 points) 2.5 pg. 35 - 1 to 2 sentences is sufficient

   For OR queries, we must traverse every entry in both postings lists, so skip pointers don't help. All we're really doing for an OR query is removing duplicate entries between the two lists.

5. (6 points) 2.9 pg. 41 - For part (a), also state the position(s) where the query matches.

   (a) 2 at position 1, 4 at position 8, 7 at positions 3, 13
   (b) 4

6. (5 points) What is the potential problem of using stop words in combination with positional indexing? How could you solve this problem? A sentence or two for each question sufficient.

Generally, stop word removal occurs before indexing happens. If we take this approach, the positional indices of our index will be different than the actual indices of the document. For example, we if a document had the phrase “... states of america” and “of” was a stop word, then the tokenized version would be “states america” and a query asking if we had the phrase “states america” would incorrectly return true. This will also cause complications for proximity queries since the positional indices will be off.

One solution is to do stop word list removal ofter indexing on the postings lists. You then need to be careful with your query processing.

7. (5 points) Read all of the “Administrative” handout. What is the URL of that handout?

   http://www.cs.pomona.edu/classes/cs160/handouts/admin.pdf