CS160 - Homework 3
Due: Wednesday Sept. 23, Before class

1. (2.5 points) 6.22 (1-2 sentences)
2. (2.5 points) 7.1 (1-2 sentences)
3. (20 points) Calculating tf-idf

   Feel free to use matlab, python or excel to calculate these answers.

   (a) 6.10

   (b) 6.15

   (c) Compute the document similarities with the document vectors above for the query "insurance car insurance for new autos" using boolean term frequncies for the query.

   (d) Compute the document similarities with the document vectors above for the query "insurance car insurance for new autos" using "**lnc**" query weighting. Use your proposal from above to handle out of vocabulary terms in the query. Did your ranking change from the previous question? What type of query would you expect to see a change in answer?

4. (10 points) Let $K = 2$ and $r = 3$. Give an index and a query such that the list of candidate documents generated using the champion list approach does NOT contain the best K documents using a nnn.nnn weighting model (i.e. only tf weighted).
5. (5 points) Cluster pruning - If we let $b_1 = 2$ and $b_2 = 1$, that is, assign followers to the two nearest leaders and then search documents associated with the closest leader to the query, can we still have the case that the closest document to the query is not found? If yes, provide an explaination, if no, a counterexample.

6. (5 points) Play with Google's page query system (found at www.google.com/advance_search, under "Page-specific tools" with the heading "Find pages similar to the page:"). How well does it work? How do you think it works?