TRA301/COS401: Homework 4
Due Date: April 9, 2009

1. **Lexical Semantics:** A famous linguist, J.R. Firth has been attributed as saying that the meaning of a word is derived from "the company it keeps". The idea is that a word derives its meaning from the word contexts in which it appears. If the word is polysemous, then there should be different contexts that attest to the different senses of the word.

   For a given word W, an n-word-profile for W is a vector of words that appear in a context window of n words to the left and right of W, not counting the closed-class words. For example, for the word *jumped* in the sentence: *the green frog jumped over the red snake*; the 2-word-context would be {*green, frog, over, red*}.

   (a) (10 points) Build a 2-word-profile vectors for the words in the file "4-1a.txt", using the closed-class word list "4-1b.txt". Show the two most frequent profiles for each word in "4-1a.txt".

   (b) (10 points) Refine the profiles to distinguish the words from the left context and those from the right context and redo 1(a). For our example, this vector would be {*L_green, L_frog, R_over, R_red*}

   (c) (10 points) Refine the profile further to distinguish the word positions relative to the target word from the left context and those from the right context and redo 1(a).

   (d) (10 points) For a word in "4-1a.txt", convert each of its profiles from 1(a) into a bit vector. Use one cell of the vector to represent a word in context. Show the two most frequent bit vector profiles for each word in "4-1a.txt".

   (e) (10 points) For a word in "4-1a.txt", convert each of its profiles from 1(b) into a bit vector. Use one cell of the vector to represent a word in its left/right context. Show the two most frequent bit vector profiles for each word in "4-1a.txt".

   (f) (10 points) For a word in "4-1a.txt", convert each of its profiles from 1(c) into a bit vector. Use one cell of the vector to represent a word in its relative positional context. Show the two most frequent bit vector profiles for each word in "4-1a.txt".

   (g) (40 points) Cluster the vectors using the k-means clustering. Initialize the algorithm by choosing k vectors randomly as centroids. Iterate the following steps until no changes to cluster membership.

      i. For each data vector x, compute its membership in clusters by choosing the nearest centroid

      ii. For each centroid, recompute its location according to members

   For each word in "4-1a.txt", show the word profiles of the centroids of clusters, after partitioning the vectors in 1(d), 1(e), and 1(f) into two clusters using the k-means algorithm. Try to characterize the two clusters and judge if they represent different senses of the word.