CSE 4510/5400 Data Mining
Due 5pm, Apr 6
Submit Server: Class = datamining, Assignment = hw4

1. Written assignment (from textbook) [pdf file or hardcopy in class]:
   (a) Q13, p562
   (b) Q16, p563 [assume distance = 1 - similarity]
   (c) Q23, p565 [assume distance = 1 - similarity]
   (d) CSE 5400 only: Q11, p562

2. Programming assignment:
   (a) Implement K-means (Algorithm 8.1, p497) and Bisecting K-means (Algorithm 8.2, p509)
      i. allow K (number of clusters) as a user-specified parameter
      ii. use Euclidean distance (Eq 2.1, p69)
      iii. Algorithm 8.2, line 3: cluster with the largest SSE (Eq 8.1 but for one cluster, p500)
      iv. Algorithm 8.2, line 5: use 10 for *number of trials*
   (b) CSE 5400 only:
      i. Implement Agglomerative Hierarchical Clustering (Algorithm 8.3, p516)
      ii. allow proximity options of Single Link, Complete Link, and Group Average
      iii. Algorithm 8.3, line 5: until K clusters remain
   (c) Allow the option of normalizing the values x_{ij} of each attribute j to be between 0 and 1 ($(x_{ij} - min_j/(max_j - min_j))$ [e.g. food dataset]
   (d) For each cluster, print IDs of data items in the cluster
   (e) Print performance of an algorithm based on:
      i. Silhouette Coefficient (p541-542)
      ii. Rand Statistic (p551-552) [food dataset only]
   (f) Datasets are on the course website:
      i. table8-3 (Table 8.3, p519)
      ii. food
   (g) Implementation:
      i. The same implementation should be able to handle the two different data sets
      ii. Use C (GNU gcc), C++ (GNU g++), Java (Oracle Java), LISP (CLISP), or Python. If you don't have a preference, use Java since it's more portable.
      iii. Your program preferably runs on code01.fit.edu (linux).
   (h) Submission:
      i. README.txt:
         - what the different files are
         - how to compile and run your program with the two data sets with a different K and with/without data normalization, and for CSE 5900 different proximty options for Agglomerative Hierarchical Clustering (preferably on code01.fit.edu).
      ii. source code files