Problem 1: (20 points)
The leader algorithm represents each cluster using a point, known as a leader, and assigns each point to the cluster corresponding to the closest leader, unless this distance is above a user-specified threshold. In that case, the point becomes the leader of a new cluster.
(a) What are the advantages and disadvantages of the leader algorithm as compared to K-means?
(b) Suggest ways in which the leader algorithm might be improved.

Problem 2: (10 points)
Traditional agglomerative hierarchical clustering routines merge two clusters at each step. Does it seem likely that such an approach accurately captures the (nested) cluster structure of a set of data points? If not, explain how you might post-process the data to obtain a more accurate view of the cluster structure.

Problem 3: (30 points)
Download and install CLUTO software for clustering high dimensional data (http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview).
Apply this software to cluster Enron Emails dataset available at https://archive.ics.uci.edu/ml/datasets/Bag+of+Words
(a) You are allowed to apply CLUTO on a sample if the data is too large for your computer. In such a case report sample size you used and how consistent the result is if you repeat experiments 3 times on 3 samples of that size.
(b) Report your clustering results when using partitional clustering in CLUTO package
(c) Report results when using agglomerative clustering algorithms in CLUTO package. In agglomerative clustering compare the results of when using complete-link vs. single-link merging schemes. Then, for single-link merging compare the results when using cosine versus Euclidean distance function.

Problem 4: (50 points)
Write a research proposal for the class project that you plan to perform. Some ideas for project topics you can find at https://www-users.cs.umn.edu/~kumar001/dmbook/projects.htm
Teams of two undergraduate students are allowed and also teams of one undergraduate and one graduate student. Teams of two graduate students are not allowed.
Write the proposal using the following format:
(0) Your name(s) and e-mail address (such that the instructor can approve your topic quickly or to ask for a revision/clarification)
(1) Title;
(2) Objective and Significance;
(3) Background;
(4) Proposed Approach (make sure to explain where you will get data and how much preprocessing is needed);
(5) References.
The proposal description may not exceed 2 pages in 12 pt style.