Problem 1:
Consider the following set of 2-dimensional points in the plane:
\{
(1,0), (0,-1), (-1,0), (0,1), (3,0), (6,0), (0,-3), (0,-6), (-3,0), (-6,0), (0,3), (0,6)
\}
(See Figure 1 below.) Measure the distance between two points using Euclidean distance. Similarity between two points is determined by their distance: more distant is less similar.

Part A: Give the dendrogram for the execution of the hierarchical agglomerative algorithm using complete-link cluster similarity on the given set of points. Stop when four clusters are left. What are the four clusters?

Part B: What is the first merge in your execution of part A that must be different if you use single-link cluster similarity instead of complete-link cluster similarity? That is, if the hierarchical agglomerative algorithm is executed using single-link cluster similarity and ties are always broken to agree with what the complete-link version does, what is the first merge for which the two versions of the algorithm differ? What clusters are merged under single-link cluster similarity?
Figure 1

(0,6)
  
(0,3)
  
(0,1)
  
(-6,0)  (-3,0)  (-1,0)  (1,0)  (3,0)  (6,0)
  
(0,-1)
  
(0,-3)
  
(0,-6)
  
(0,-1)
Problem 2
The algorithm for hierarchical agglomerative clustering giving in Figure 17.8 of *Introduction to Information Retrieval* uses one priority queue for each cluster to efficiently find the most similar pair of clusters to merge. The priority queues are updated for each merge step by deleting the two clusters that have been merged and inserting the new combined cluster. Consider breaking ties when selecting the pair of clusters to merge by choosing the pair that results in the smallest combined cluster. What modifications would be needed in the algorithm of Figure 17.8? Would the running time be affected? Explain.

Problem 3
Slide #17 of the slides for general clustering algorithms posted under March 31 presents an iterative improvement algorithm for divisive partitioning. This problem addresses recalculating the min-max cut cost (slides #13 and #14) incrementally for use with that algorithm.

Let $V$ denote the set of objects to be clustered. Assume that for any objects $v$ and $w$, \( \text{sim}(v,w) = \text{sim}(w,v) \) (we have been assuming this in class). Also assume that for any object $v$, \( \text{sim}(v,v) = 0 \).

The following relationships hold for incremental changes to a cluster by removing or adding an object $x$:

- \[ \text{cutcost}(C_p - \{x\}) = \text{cutcost}(C_p) + 2 \sum_{v_i \in C_p} \text{sim}(v_i, x) - \sum_{v_i \in V} \text{sim}(v_i, x) \]
- \[ \text{cutcost}(C_p \cup \{x\}) = \text{cutcost}(C_p) - 2 \sum_{v_i \in C_p} \text{sim}(v_i, x) + \sum_{v_i \in V} \text{sim}(v_i, x) \]
- \[ \text{intracost}(C_p - \{x\}) = \text{intracost}(C_p) - \sum_{v_i \in C_p} \text{sim}(v_i, x) \]
- \[ \text{intracost}(C_p \cup \{x\}) = \text{intracost}(C_p) + \sum_{v_i \in C_p} \text{sim}(v_i, x) \]

**Question:** Given the relationships above, what is the time complexity of the step: move $v_j$ to that cluster, if any, such that the move gives maximum decrease in cost of the iterative improvement algorithm on slide #17?

You may assume $\sum_{v_i \in V} \text{sim}(v_i, x)$ is precomputed before the initial clustering is chosen. Do not include the cost of this precomputation. Justify your answer.