COS 435, Spring 2015 - Problem Set 6

Due at 1:30PM, Wednesday, April 15, 2015.

Collaboration and Reference Policy

You may discuss the general methods of solving the problems with other students in the class. However, each student must work out the details and write up his or her own solution to each problem independently. For each problem, list the students with whom you discussed general methods of solving the problem (excluding very brief casual conversations).

Some problems have been used in previous offerings of COS 435. You are NOT allowed to use any solutions posted for previous offerings of COS 435 or any solutions produced by anyone else for the assigned problems. You may use other reference materials; you must give citations to all reference materials that you use.

Lateness Policy

A late penalty will be applied, unless there are extraordinary circumstances and/or prior arrangements:

- Penalized 10% of the earned score if submitted by 11:59 pm Wed. (4/15/15).
- Penalized 25% of the earned score if submitted by 4:30pm Friday (4/17/15).
- Penalized 50% if submitted later than 4:30 pm Friday (4/17/15).

Problem 1: Clustering – hierarchical agglomerative

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 and data structures of Figure 17.8? Be sure to address all the data structures, not just the priority queues. Would the running time be affected? Explain.

Problem 2: Clustering -- iterative improvement for divisive partitioning

Slide #25 of Part 2 of the slides for clustering, posted under April 6, presents an iterative improvement algorithm for divisive partitioning. This problem addresses recalculating the total relative cut cost (slides #18 and #19) incrementally for use with that algorithm.

Let U denote the set of objects to be clustered. Assume that for any objects v and w, sim(v,w)=sim(w,v) (we have been assuming this in class). Also assume that for any object v, sim(v,v)=0. Let C_p be an arbitrary cluster containing object x, C_q be an arbitrary cluster that does not contain x. (The set notation $C_p - \{x\}$ denotes C_p with x removed, and $C_q \cup \{x\}$ denotes C_q with x added.)

The following relationship holds for incremental changes to the intracost of a cluster when removing or adding an object x.

$$\text{intracost}(C_p) - \text{intracost}(C_p - \{x\}) = \sum_{v_i \text{ in } C_p - \{x\}} \text{sim}(v_i, x)$$

$$= \sum_{v_i \text{ in } C_p} \text{sim}(v_i, x) \qquad \text{since sim}(x,x) = 0$$

From this relationship we derive the incremental cost changes for intracost:

$$\text{intracost}(C_p - \{x\}) = \text{intracost}(C_p) - \sum_{v_i \text{ in } C_p} \text{sim}(v_i, x)$$

$$\text{intracost}(C_q \cup \{x\}) = \text{intracost}(C_q) + \sum_{v_i \text{ in } C_q} \text{sim}(v_i, x)$$

Your task is to derive incremental cost equations for cutcost. The ultimate goal is to minimize the computation time used by the iterative improvement algorithm.

Part a: Give an equation for

$$\text{cutcost}(C_p) - \text{cutcost}(C_p - \{x\})$$

when x is an object in C_p. Your equation should be in terms of similarities between x and other objects.

Hint: the quantity

$$\sum_{v_i \text{ in } U} \text{sim}(v_i, x) \quad \text{where U is the set of all objects}$$

is useful because it is a function of x independent of the clustering and can be precomputed before the clustering construction is begun.

Part b: Using your equation of Part a, derive equations for

i. cutcost($C_p - \{x\}$) as an incremental change to cutcost(C_p);
ii. cutcost($C_q \cup \{x\}$) as an incremental change to cutcost(C_q).

Part c: Given the equations for the incremental changes in intracost and cutcost, what is the computational time complexity of the step:

move v_j to that cluster, if any, such that move gives maximum decrease in cost

of the iterative improvement algorithm on slide #25? Specify the data structures you are using and how they are used to achieve the time complexity. You may assume

$$\sum_{v_i \text{ in } U} \text{sim}(v_i, x)$$ where U is the set of all objects

is precomputed before the initial clustering is chosen; don't include the cost of this precomputation.