COS 435, Spring 2012 - Problem Set 6

Due at 1:30pm, Wednesday, April 18, 2012.

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.

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:

- No penalty if in Prof. LaPaugh's office or inbox by 5pm Wednesday (4/18/12).
- Penalized 10% of the earned score if submitted by 11:59 pm Wed (4/18/12).
- Penalized 25% of the earned score if submitted by 5pm Friday (4/20/12).
- Penalized 50% if submitted later than 5pm Friday (4/20/12).

Problem 1: Clustering -- iterative improvement for divisive partitioning

Slide #53 of the slides for clustering posted under April 2 presents an iterative improvement algorithm for divisive partitioning. This problem addresses recalculating the total relative cut cost (slides #46 and #47) 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:

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

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 #53? Specify the data structures you are using and how they are used to achieve the time complexity. You may assume

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

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

Problem 2: Detecting near-duplicate documents

Part a: Let D denote a document that is 500 words long and contains each of the words "philanthrepist" , "pendantic" and "androgenous" exactly once each, with "philanthrepist" occurring in word position 100, "pendantic" in position 205, and "androgenous" in position 320. Each of these words is misspelled. Let D_{cor} be the document with these spelling errors corrected ("philanthropist" , "pedantic" and "androgynous"). What is the value of the resemblance $r(D, D_{cor})$ for a 5-shingling of each document if, for each document, 25% of all possible shingles are repeated shingles?

Part b: Let E denote a document that is 500 words long and contains each of the words "philanthrepist" , "pendantic" and "androgenous" exactly once each but as the phrase "pendantic androgenous philanthrepist" starting at word position 200. Let E_{cor} be the document with the spelling errors in this phrase corrected ("pedantic androgynous philanthropist"). What is the value of the resemblance $r(E, E_{cor})$ for a 5-shingling of each document if, for each document, 25% of all possible shingles are repeated shingles?

Part c: For what threshold would one of the pairs (D, D_{cor}) and (E, E_{cor}) be considered near-duplicates and the other not? Which is which? In your opinion, is this a desirable outcome?