COS 435, Spring 2008 - Problem Set 5

Due at 5:00pm, Monday April 14, 2008 (end of the business day).

Collaboration 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.

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 5:00 pm Tuesday (4/15/08).
- Penalized 25% of the earned score if submitted by 5:00 pm Saturday (4/19/08).
- Penalized 50% if submitted later than 5:00 pm Saturday (4/19/08).

Problem 1 (divisive clustering):

On slide #11 of the class presentation April 2, there is an iterative improvement algorithm intended to be used for divisive partitioning. This problem addresses recalculating the min-max cut cost (slides #8 and #9) incrementally for use with that algorithm. 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.

Part a: Give an equation for

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

when x is an object in C_p. Here C_p is an arbitrary cluster and $C_p - \{x\}$ is C_p with x removed. Your equation should be in terms of similarities between x and other objects.

Hint: the quantity

$$\sum_{v_i \text{ in } U} \text{sim}(v_i, x)$$ 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, express cutcost($C_p - \{x\}$) as an incremental change to cutcost(C_p). ***Also*** express cutcost($C_q \cup \{x\}$) as an incremental change to cutcost(C_q) when C_q is an arbitrary cluster that does not contain x. ($C_q \cup \{x\}$ is C_q with x added.)

Part c: Give an equation for

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

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

Part d: Using your equation of Part c, express intracost($C_p - \{x\}$) as an incremental change to intracost(C_p) ***and*** express intracost($C_q \cup \{x\}$) as an incremental change to intracost(C_q) when C_q does not contain x.

Part e: Given your equations in Parts b and d, what is the 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 #11. 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. Justify your answer.

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"). For a 5-shingling of each document, give a lower bound and an upper bound on the resemblance *r(D, D_{cor})* as given on slide #7 of the class presentation (April 7 and 9). (Do not expect your lower and upper bounds to be close.)

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"). For a 5-shingling of each document, give a lower bound and an upper bound on the resemblance *r(E, E_{cor})* as given on slide #7 of the class presentation. (Again, do not expect your lower and upper bounds to be close.)

Part c: When computing the 5-shingling of each of D, D_{cor}, E, and E_{cor}, suppose 25% of all possible shingles are repeated shingles. In this case, what are *r(D,D_{cor})* and *r(E,E_{cor})*? For what threshold would one pair be considered near-duplicates and the other not? Which is which? In your opinion, is this a desirable outcome?