COS 435, Spring 2009 - Problem Set 2
Due at 3:00pm, Thursday, Feb. 19, 2009.

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:

* Penalized 10% of the earned score if submitted by noon Friday (2/20/09).
* Penalized 25% of the earned score if submitted by noon Monday (2/23/09).
* Penalized 50% if submitted later than noon Monday (2/23/09).

Problem 1

In class we discussed the LSI example appearing in the original paper *Indexing by Latent Semantic Analysis* by Deerwester, Dumais, et. al. (Journal of the Society for Information Science, 41(6), 1990, 391-407). The term-document matrix and the matrices of the reduced-dimension approximation (k=2) are shown below.

Terms	**Documents**								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

$$\mathbf{V'_2}^{T} = \begin{pmatrix} 0.20 & 0.61 & 0.46 & 0.54 & 0.28 & 0.00 & 0.02 & 0.02 & 0.08 \\ -0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53 \end{pmatrix}$$

$$\mathbf{\Sigma'_2} = \begin{pmatrix} 3.34 & \\ & 2.54 \end{pmatrix}$$

$$\mathbf{U'_2} = \begin{pmatrix} 0.22 & -0.11 \\ 0.20 & -0.07 \\ 0.24 & 0.04 \\ 0.40 & 0.06 \\ 0.64 & -0.17 \\ 0.27 & 0.11 \\ 0.27 & 0.11 \\ 0.30 & -0.14 \\ 0.21 & 0.27 \\ 0.01 & 0.49 \\ 0.04 & 0.62 \\ 0.03 & 0.45 \end{pmatrix}$$

The matrix $C_2 = U'_2\,\Sigma'_2\,{V'_2}^{T}$ is shown below for your information.
C_2 IS NOT NEEDED. IT IS INCORRECT TO USE IT FOR THIS PROBLEM!

$$\mathbf{C_2} = \begin{pmatrix} 0.16 & 0.40 & 0.38 & 0.47 & 0.18 & -0.05 & -0.12 & -0.16 & -0.09 \\ 0.14 & 0.37 & 0.33 & 0.40 & 0.16 & -0.03 & -0.07 & -0.10 & -0.04 \\ 0.15 & 0.51 & 0.36 & 0.41 & 0.24 & 0.02 & 0.06 & 0.09 & 0.12 \\ 0.26 & 0.84 & 0.61 & 0.70 & 0.39 & 0.03 & 0.08 & 0.12 & 0.19 \\ 0.45 & 1.23 & 1.05 & 1.27 & 0.56 & -0.07 & -0.15 & -0.21 & -0.05 \\ 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ 0.22 & 0.55 & 0.51 & 0.63 & 0.24 & -0.07 & -0.14 & -0.20 & -0.11 \\ 0.10 & 0.53 & 0.23 & 0.21 & 0.27 & 0.14 & 0.31 & 0.44 & 0.42 \\ -0.06 & 0.23 & -0.14 & -0.27 & 0.14 & 0.24 & 0.55 & 0.77 & 0.66 \\ -0.06 & 0.34 & -0.15 & -0.30 & 0.20 & 0.31 & 0.69 & 0.98 & 0.85 \\ -0.04 & 0.25 & -0.10 & -0.21 & 0.15 & 0.22 & 0.50 & 0.71 & 0.62 \end{pmatrix}$$

Problem 1 Part a: Consider the query ***trees, minors***. This query is represented by the 0/1 vector **q** = (0,0,0,0,0,0,0,0,0,1,0,1). What is the transformed query $\mathbf{q_k}$ for this query using the dimension 2 (k=2) LSI approximation? Show the equations you are using. You may use a calculation program to actually do the calculation, although hand calculation should not be too burdensome.

Problem 1 Part b: Give the scores of the 9 documents using the dimension 2 (k=2) LSI approximation with $\mathbf{q_k}$ of Part a. Show the equations you are using. You may use a calculation program to actually do the calculation, although hand calculation should not be too burdensome.

Problem 1 Part c: Give the scores of the 9 documents for the query ***trees, minors*** using the standard vector model. For both the documents and the query, start with 0/1 vectors, then normalize them to unit vectors and use the cosine similarity measure. It suffices to calculate the vector components that will actually affect the computation.

Problem 2

The computational cost of comparing a query to all documents by calculating $C^T\mathbf{q}$ is M*N multiplications and (M-1**)**N additions. As in class, **q** denotes the vector representation of the query and C denotes the matrix whose columns are the vector representations of the documents. There are M terms and N documents.

What is the computational cost of doing a comparison of a query to all documents after latent semantic indexing has been used to compute the rank-k approximation in terms of matrices U'_k, Σ'_k, and V'_k? **Do not include** the preprocessing cost to find matrices U'_k, Σ'_k, V'_k, $(V'_k(\Sigma'_k)^2)$, and $((\Sigma'_k)^{-1}(U'_k)^T)$ – all of these can be computed once for the collection of documents and a given value of k. **Do include** all computation steps that must be done after **q** is given. Unit-cost computations are scalar addition, multiplication, comparison and other basic program steps - not vector operations. You should list each step of the computation contributing to the cost. Your analysis should be in terms of M, N, and k.

Problem 3

In class I mentioned that scoring how well a document matches a query based on how close together two query terms appear in the document is not something that can be achieved in the vector model using M-dimensional vectors, where M is the number of terms in the lexicon. Give another example of a property one might like to use in scoring documents that could not be modeled in the M-dimensional vector model. You should justify your answer, but need not give a proof of your claim.