CSci 5304 HW4 Due: Nov 10 2014

- This assignment is a mix of handwritten questions and Matlab coding questions. You should hand in everything on paper, including a printout of your matlab functions and scripts.
- You are encouraged to do this in pairs: hand in one copy of the entire assignment with both names on it. You should still satisfy yourself that you could answer every question on your own, since similar questions will appear on future exams.

1. Write a 2×2 Givens rotation $\begin{bmatrix} c & s \\ -s & c \end{bmatrix}$ as a product of two Householder reflections. Hint: one way to do this involves making one of the two Householder reflectors represent the reflection across a coordinate axis.
2. Prove or disprove: any square orthogonal matrix can be written as a product of Householder reflectors.
3. Prove or disprove: any square orthogonal matrix can be written as a product of Givens rotations.
4. What are the eigenvalues and eigenvectors for a Householder reflector of the form $I - 2uu^T$ where $u^T u = 1$?
5. Let $P = \begin{bmatrix} c & s \\ s & -c \end{bmatrix}$, with $c^2 + s^2 = 1$. Prove or disprove: P is an orthogonal transformation. Is this a Givens rotation, a Householder transformation, or neither? If a Givens rotation, write it as a Givens rotation. If a Householder transformation, write it in the form $I - 2uu^T$ for some unit vector u.
6. Paul, Jane, and Ann, share information about their likes and dislikes of movies in order to make decisions about selecting films to see. They rates films they see with a scale of 0 to 10, (10 means they liked the movie very much). Here is the status of their table of ratings when Ann was interested in a new film which soon came to a 'theater near her' (titled 'Title 6' in the table):

movie	Paul	Jane	Ann
Title-1	4	8	4
Title-2	9	3	8
Title-3	2	6	1
Title-4	7	4	4
Title-5	8	3	6
Title-6	3	8	x

Ann generally follows a combination of Paul and Jane's ratings. Ann wants to predict how well she will like the movie Title-6, which Paul and Jane have already seen. Ann reasons as follows: she will give her 'similarity' coefficients α and β for Paul and Jane respectively. If the missing rating (call it x) were known then the column of Ann's ratings should be the closest in the least-squares sense to the combination of Paul's and Jane's ratings:

$$\min_{\alpha,\beta} \left\| \begin{pmatrix} \text{Paul's ratings} \end{pmatrix} \cdot \alpha + \begin{pmatrix} \text{Jane's ratings} \end{pmatrix} \cdot \beta - \begin{pmatrix} \text{Ann's ratings} \end{pmatrix} \right\|_2^2 = \min_{\alpha,\beta} \left\| \begin{pmatrix} \text{Paul's ratings} & \text{Jane's ratings} \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \begin{pmatrix} \text{Ann's ratings} \end{pmatrix} \right\|_2^2$$

Determine, α, β using the first 5 movie ratings, and from that the combination of Paul's and Jane's ratings that is closest to Ann's, in a least squares sense. Use the resulting best combination to infer the induced rating for Ann for Title-6. Should Ann see 'Title 6'? Is her taste closer to Paul's or to Jane's?

7. The data in the file lsidata.m* contains the term-frequency matrix A for a collection of 2340 documents, using a dictionary of 21839 words. Load this data into matlab and design a query to retrieve all documents containing the word "bipolar". Apply this query to the original term frequency matrix (with columns scaled to unit length) and then repeat this procedure using the best rank-50 approximation A_{50} to the term-frequency matrix A. Only the document headlines are provided, but the word counts are based on the document contents which can be found at the web site mentioned within the datafile.

   Hints: use `svds` to obtain the singular value decomposition $A = U\Sigma V^T$ instead of `svd`, because (a) the matrix A is too big for `svd` and (b) you can get the rank 50 approximation A_{50} directly from `svds`. Do not try to form the rank-50 approximation explicitly – rather work directly with the factors U, Σ, V obtained from `svds`. Do not forget to normalize the columns to unit length before applying a query or computing the SVD. A skeleton matlab script for this problem is given in LSIpreamble.m*.

   The query vector is a vector with 1's in the positions corresponding to the words in your query and zeros for all other words. So if your query is only one word, then the query vector has only one nonzero element. If n is the number of words, and m the number of documents, then $\mathbf{q}$ is an n-vector with only a few nonzero elements. The term-frequency matrix A is an $n \times m$ with the j-th column corresponding to the j-th document. If each column $\mathbf{a}_j$ (for $j = 1, \ldots, m$) is normalized to length 1 in the usual 2-norm, and the query vector $\mathbf{q}$ is also normalized to unit length, then the inner product $\mathbf{q} \circ \mathbf{a}_j$ is the cosine of the angle between the two vectors. If the vectors are almost the same, then the angle will be small and the cosine will be very close to 1. If all the columns of A are normalized in this fashion, then all the cosines can be computed at once with the formula $\mathbf{q}^T A$. So your task is to compute $\mathbf{q}^T A$ and also $\mathbf{q}^T A_{50}$. Once you have computed these two vectors of similarities, you need to sort each of them to find the positions of the 10 biggest values, and then print out the document headlines corresponding to those positions.

   You will need to use the `sort` function in matlab, which returns two results. The first result is the values sorted, and the second result consists of the indices of the positions of those values. You can use that second result to retrieve the document headlines.

   A sample script which carries out all of this on a toy example is given in LSItoy.m*.

*All external links are located in `http://www-users.cselabs.umn.edu/classes/Fall-2014/csci5304/Notes/MatlabDemos/`.