**Big Data Analytics**

**Homework #2**

**Due Date: April 30th, 2019 (No Extensions)**

1. Consider a 25-by-20 Ratings matrix, Containing rows for 25 users and columns for 20 movies. Each cell of the table contains a rating value between 1(=Poor) and 7(=Best). Perform the following steps with this data structure:
   1. (5) Randomly populate this table with 200 entries (between 1 and 7) and adjust the values to make sure that each row has at least 3 ratings and each column has at least 3 ratings. Leave the other 300 cells as unfilled. Include this table in your submission, shown in table form.
   2. (5) Randomly select 150 of the 200 ratings as the training dataset and the remaining 50 as the test dataset. Show both sets of data in the form of tables.
   3. (10) Use any toolbox to find a matrix factorization of the training data and include only two latent factors. (Use ALS based method and not an SVD method). Show the latent factor matrices, properly labeled.
   4. (15) Use the factorized Ratings matrix to predict the ratings for the 50 ratings of the test-dataset. Measure the error in prediction as the Mean-Square-Error between the actual and the predicted ratings. Submit the test data (in matrix form) with each cell showing the actual and the predicted ratings. Also, report the MSE value computed for all the predictions.
2. Repeat Q#1 with the following difference:
   1. (15) Populate the Ratings matrix in step 1.a in such a way (you are free to choose the rating values) that the MSE obtained in 1.d is minimized. Repeat all steps of Q#1 and submit the answers.
3. Repeat Q#2 with the following difference:
   1. (15) Now populate the Ratings matrix in such a way that the MSE obtained in 1.d is maximized.
   2. (10) Comment on the characteristics of the data distribution that increase or decrease the MSE. Show how the data in your tables in #2 and #3 above reflect those characteristics.
4. (20) Repeat #3 with the difference that now you use four latent factors. Use the same rating values as used in #3 above. Compare the predicted outcomes and the MSE values with those obtained in #3. Comment on any significant differences observed.
5. Perform the following steps with the “SVD-Data” matrix attached with this assignment. This table contains documents as rows and words as columns. Each cell entry is the number of times a word occurs in a document. Use any available toolbox to perform the SVD of this matrix. Then answer all the following questions.
   1. (8) Show all the Eigen-values and Eigen-vectors obtained after the decomposition.
   2. (12) How many, and which Eigen values (and vectors) would you select if you should preserve 85% of the information?
   3. (12) Take one of the dropped Eigen-vectors and interpret it in terms of documents and words to show its relevance/irrelevance to the original data table.
   4. (15) How many concept vectors are chosen by you in step-b above? Write the intuitive interpretation of each of these vectors. What is the intuitive interpretation of the Eigen-values associated with each of these vectors?
   5. (12) Compute the MSE between the original matrix and the predicted matrix using the truncated number of Eigen vectors.
   6. (16) The power-point slides set (Faloutsos-p1-SVD) shows four interpretations of an SVD. Briefly explain the results obtained for SVD of your data matrix, as per each of the four interpretations.
   7. (10) Consider using only the top two Eigen-values and Eigen vectors. Show and interpret how doc-1, doc-8, and doc-16 are summarized in this case?
   8. (10) Compute the MSE if we use only the top two Eigen-vectors and compare it to the value obtained in #5d.
6. (10) Good presentation of all results and answers.