CS 6350 – Big Data Management and Analytics
Fall 2015
Assignment #3 Part B
Due: 11/10/2015

Problem 1 Given the following utility matrix which contains ratings (1 to 5 star(s) scale) on eight items (i_1 to i_8) by three users (U_1, U_2 and U_3):

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
U_1	⋆⋆⋆⋆	⋆⋆⋆⋆⋆		⋆⋆⋆⋆⋆	⋆		⋆⋆⋆	⋆⋆
U_2		⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆	⋆	⋆⋆	⋆	
U_3	⋆⋆		⋆	⋆⋆⋆		⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆

Compute the following from the data of this matrix. Please show all your work.

(a) Compute the Jaccard distance between each pair of users.

(b) Compute the cosine distance between each pair of users.

(c) Compute the cosine distance between each pair of users by treating the utility matrix as boolean.

(d) Change the ratings with 3 or more stars to just 1 star and ratings with 2 or less stars to no rating. Compute the Jaccard distance between each pair of users.

(e) Repeat part (d) using cosine distance.

(f) Normalize the matrix by subtracting from each nonblank entry the average value for its user. Show the normalized matrix.

(g) Using the normalized matrix from part (f), compute the cosine distance between each pair of users.

Problem 2 Cluster the items in the matrix of Problem 1 by using the following steps:

(a) Cluster the eight items hierarchically into four clusters. The following method should be used to cluster. Change the ratings with 3 or more stars to just 1 star and ratings with 2 or less stars to no rating. Use Jaccard distance to measure the distance between the resulting column vectors. For clusters of more than one element, take the distance between clusters to be the minimum distance between pairs of elements, one from each cluster.

(b) Then construct from the original utility matrix a new matrix whose rows correspond to users, as before, and whose columns correspond to clusters. Compute the entry for a user and cluster of items by averaging the nonblank entries for that user and all the items in the cluster.

(c) Compute the cosine distance between each pair of users, according to the matrix from part (b).

Problem 3 Given the following eight points in 2-dimensional space with corresponding xy-coordinates:

Point	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
x	2	3	4	5	6	8	8	9
y	2	4	7	3	7	7	1	3

Compute the following.

(a) Using k-means and Euclidean distance, calculate the coordinates of the two final cluster centers. The two cluster centers are initialized at points P_5 and P_6. For each iteration designate which cluster center each point is assigned to. Please show all Euclidean distance calculations.

(b) Repeat part (a) using points P_3 and P_7 as the two new initialized cluster centers instead. Did the clustering assignments change?

Problem 4 Given the following eight points in 1-dimensional space with corresponding x-coordinates:

Point	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
x	11	4	6.5	12	2.2	8	6	1

(a) For each of the following agglomerative clustering techniques, draw the resulting dendrogram using Euclidean distance. The dendrogram should clearly show the order in which the points are merged.

(i) Single link

(ii) Complete link

(b) Assuming you have the following two clusters: $\{P_2, P_3, P_5, P_7, P_8\}$ and $\{P_1, P_4, P_6\}$. Calculate the distance between the two clusters according to:

(i) Single link

(ii) Complete link

(iii) Average link