INTRODUCTION TO ARTIFICIAL INTELLIGENCE – LECTURE 1 – CLUSTERING

Name and surname: Sofya Aksenyuk   Index no.: 150284

I. Given the representation of five users U1-U5 in terms of two documents D1-D2 (to the left) and the cosine similarity matrix (to the right), use 2-means (K-means with K=2) to group these users into two clusters. Answer the following questions.

	U1	U2	U3	U4	U5
D1	0.2	0.5	0.7	0.7	0.8
D2	0.7	0.2	0.6	0.3	0.6

	U1	U2	U3	U4	U5
U1	1	0.61	0.83	0.63	0.79
U2	0.61	1	0.94	1	0.96
U3	0.83	0.94	1	0.95	0.99
U4	0.63	1	0.95	1	0.97
U5	0.79	0.96	0.99	0.97	1

a) Assume the least similar users are the initial centroids. Which users would be used as centroids?

Answer: The initial cluster centroids C1 and C2 are U1 and U2

b) What would be the clustering obtained in the first iteration? Indicate the groups G1 and G2 with the initial centroids C1 and C2 determined in point a).

G1 (U1)	1	0,61	0,83	0,63	0,79
G2 (U2)	0,61	1	0,94	1	0,96

c) Would the clustering results after the first iteration differ in case U3 and U4 were used as the initial centroids?

G1 (U3)	0,83	0,94	1	0,95	0,99
G2 (U4)	0,63	1	0,95	1	0,97

d) Compute the J measure for the clustering obtained in point b). *Hint*: for our lecture example (see table below): **J** = over all clusters sum the similarities of objects contained in the cluster to the respective cluster centroid = (for G1 compare objects with centroid C1) + (for G2 compare objects with centroid C2) + (for compare objects G3 with centroid C3) = (8 + 6 + 8 + 10) + (10 + 10) + (6 + 9) **= 67**

J = (1)+(1+0,94+1+0,96) =
= 4,9

	U1	U2	U3	U4	U5	U6	U7	U8
G1 (C1)	**8**	**6**	**8**	**10**	3	6	5	7
G2 (C2)	6	5	4	6	5	**10**	6	**10**
G3 (C3)	4	4	0	5	**6**	6	**9**	3

e) Compute the new centroids to be used in the second iteration of 2-means based on the clustering results obtained in point b).

	C1	C2
D1	0.2	0,675
D2	0.7	0,425

(0,5+0,7+0,7+0,8)/4
(0,2+0,6+0,3+0,6)/4

II. Given the below cosine similarity matrix for five users U1-U5, use Agglomerative Hierarchical Clustering (AHC) to cluster these users.

	U1	U2	U3	U4	U5
U1	1	0.61	0.83	0.63	0.79
U2	0.61	1	0.94	1	0.96
U3	0.83	0.94	1	0.95	0.99
U4	0.63	1	0.95	1	0.97
U5	0.79	0.96	0.99	0.97	1

a) Which users would be clustered together first (irrespective of how the similarity between groups is defined)?

Answer: U2, U4

b) Compute the similarity matrix after the first iteration while assuming that the similarity between groups is equal to the maximal (single-link) or average similarity of the objects (users) contained in these clusters?

Single-link

	U1	U24	U3	U5
U1	1	0,63	0.83	0.79
U24	0,63	1	0,95	0,97
U3	0.83	0,95	1	0.99
U5	0.79	0,97	0.99	1

Average

	U1	U24	U3	U5
U1	1	0,62	0.83	0.79
U24	0,62	1	0,945	0,965
U3	0.83	0,945	1	0.99
U5	0.79	0,965	0.99	1

c) Present the process of AHC incorporating the complete-link similarity. Show the similarity matrices in the following iterations (fill the identifiers of groups and similarity values) and draw a dendrogram (fill the identifiers of users, complete the vertical and horizontal lines to reflect the progress in the AHC process).

First iteration (complete-link)

	U1	U24	U3	U5
U1	1	0,61	0.83	0.79
U24	0,61	1	0,94	0,96
U3	0.83	0,94	1	0.99
U5	0.79	0,96	0.99	1

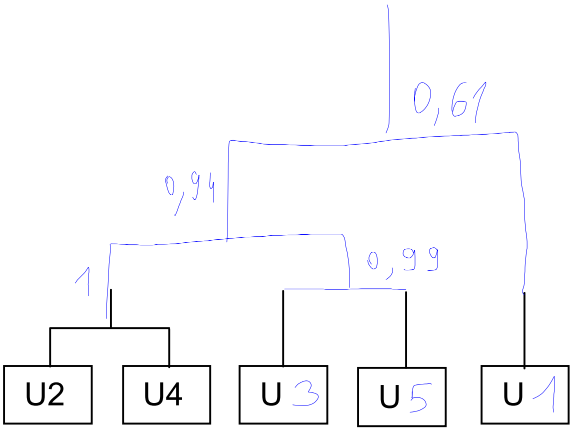
Second iteration

	U1	U24	U35
U1	1	0,61	0,79
U24	0,61	1	0,94
U35	0,79	0,94	1

Third iteration

	U1	U2435
U1	1	0,61
U2435	0,61	1

Dendrogram:



d) How many groups would be obtained in the above clustering process if the similarity threshold for AHC was set to 0.8? Which users would be assigned to which groups (name the groups by G1, G2, …)?

Answer: one group: G1 {U2, U4, U3, U5}. G2{U1} won't be obtained due to the value 0,61<0,8