Qn V.[12pt] A search engine returned 5 documents in response to a query. The documents are completely decribed in terms of the weights and type of keywords they have in the following table:

	Computer	Repair	Science
D1	6	4	0
D2	3	0	7
D3	1	6	3
D4	1	3	6
D5	6	2	2

Suppose you are trying to cluster these results using K-means clustering algorithm into 2 clusters. Assume you start the clusters off with D3 and D4 as the initial cluster centers. Assume that you are using a bag-based similarity measure discussed in the LSH paper:

$$sim(d1, d2) = \frac{|B1 \cap B2|}{|B1 \cup B2|}.$$

Show the operation of the K-means algorithm on this data. For each iteration of the K-means, show the cluster dissimilarity measure (which is defined as the sum of the similarities of docs from their respective cluster centers). You are allowed to use calculators. Show all details of your work if you expect partial credit.

Iteration 1: cluster centers: c1: D3 c2: D4

For D1, sim(D1,D3) = 5/15 = 1/3 = 0.333 (explanation {6,4,0} intersection {1 6 3}= {1 4 0}. The cardinality of the bag is thus 5. {6 4 0} union { 1 6 3} = {6 6 3}. Cardinality is 15.)

Sim(D1,D4) is 4/16=.25

So, D1 goes into cluster c1 because it is more similar to its cluster center.

Similar computation shows that D2 goes into c2 {its similarity is .25 with D3 and .54 with D4}

For D5, the similarity is .33 for each cluster center, so we can put it in either. Suppose we randomly put it in c1

So our clusters are {D1, D3,D5} and {D2,D4}

The aggregate cluster dissimilarity is computed with respect to the CLUSTER CENTROIDS (not the original cluster centers): {Most people did this with respect to the original cluster centers.}

So we find the cluster centers to be the centroids of the current clusters (which are computed as the mean of each of the bag coordinates.

For {D1, D3, D5}, the centroid is the bag {4.3, 4, 1,7} (4.3 = (6+1+6)/3= 13/3)

For {D2, D4} the centroid is {4.5, 1, 4.5}

So the dissimilarity measure will turn out to be:

1/.71+ 1/.60 + 1/.50 +1/.48+ 1/.67 = 8.65 {here we are taking dissimilarity as the inverse of similarity. The 1/.71 comes because , sim(D1,{4.3,4,1.7} } = (4.3+4+0)/(6+4+1.7)= 8.3/11.7 = .71; and inverse of it is 1/.71.

(if you computed it with respect to the original centers, it would come out to 9.85)

Iteration 2:

With respect to the centroids, we again sort each of the elements, *including D3 D4,* into clusters

For D1, sim(D1,{4.3,4,1.7} } = (4.3+4+0)/(6+4+1.7)= 8.3/11.7 = .71
Similary, sim(D1,{4.5,1,4.5}) = .38

So, D1 continues to be in cluster C1

When we do this for the other four elements, we find that each of them continue to stay in their own respective clusters.

Since the clusters didn't change, the iteration stops, and the clusters {D1,D3,D5}; {D2,D4} are output.

The aggregate dissimilarity measure stays at 8.65

==========The END==========