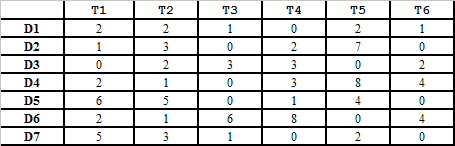
# Homework 6

Your Name: \_\_\_\_Tsucheng Lu\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student ID:\_\_\_\_\_\_A20525040\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**1. (40 points) Clustering by the K-Means approach**

Consider the following document-term matrix, where each entry represents the raw frequency of a term Ti in document Dj. We would like to apply clustering to automatically group these documents into 3 classes (clusters). Note: you can use a spreadsheet program such as Microsoft Excel to facilitate computation in intermediate steps. But you need to show the results in each step, until the clustering is converged



Suppose we initially assign D2 to Cluster 1, D4 and D6 to Cluster 2, and D5 and D7 to Cluster 3. Using the K-means clustering method discussed in class, compute the final contents of the 3 Clusters. Use the Cosine similarity of two vectors (NOT only the dot product) as your similarity measure. Show the details of your computation, including intermediate steps in each iteration of the algorithm.

Note: Recall that the Cosine similarity of two vectors is their dot product divided by the product of their norms. For example, Consider the two vectors X and Y:

X = <3, 0, 1, 2, 0, 3>

Y = <2, 0, 0, 3, 8, 4>

The dot product is given by sum of the coordinate-wise multiples:

dot-product(X, Y) = 3\*2 + 0\*0 + 1\*0 + 2\*3 + 0\*8 + 3\*4

= 6 + 0 + 0 + 6 + 0 + 12

= 24.

The norm of each vector is the square-root of the sum of the squares of its dimension values. So, the norms of X and Y are:

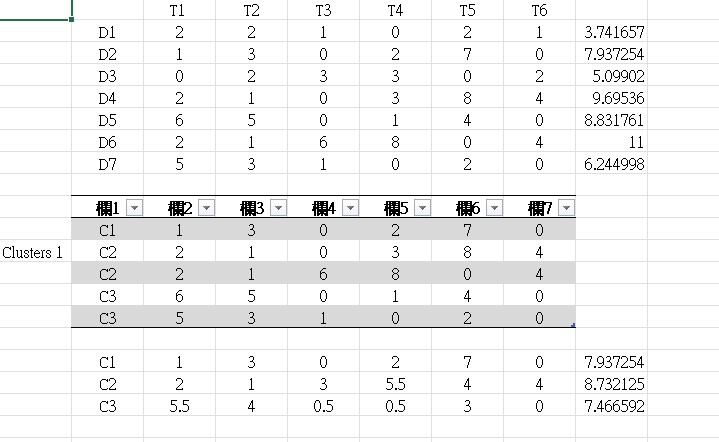
http://facweb.cs.depaul.edu/mobasher/classes/ect584/HW/norm-d1.gif     http://facweb.cs.depaul.edu/mobasher/classes/ect584/HW/norm-d4.gif

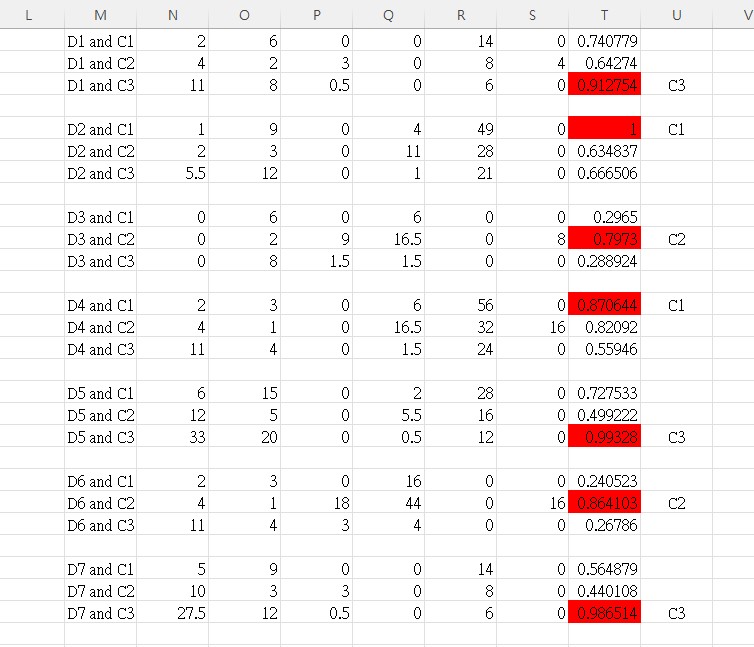
and the Cosine similarity of X and Y is given by:

http://facweb.cs.depaul.edu/mobasher/classes/ect584/HW/sim.gif

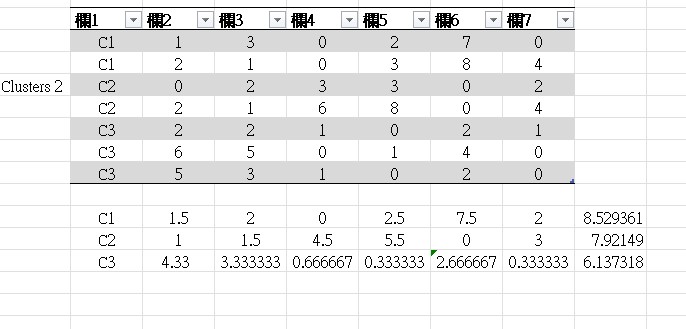
Your answer:

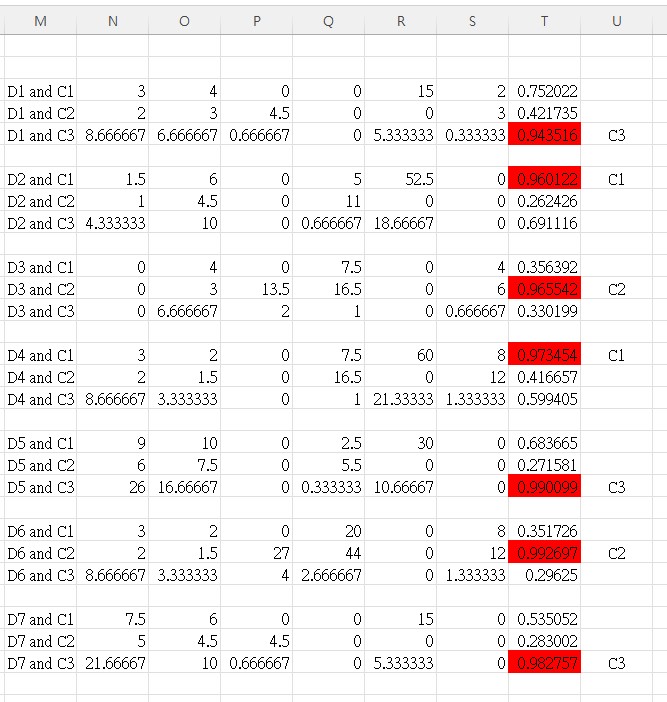
First of all, calculate the cluster centroids and calculate the Cosine distance. After that depending on the highest value and create new cluster centroids.



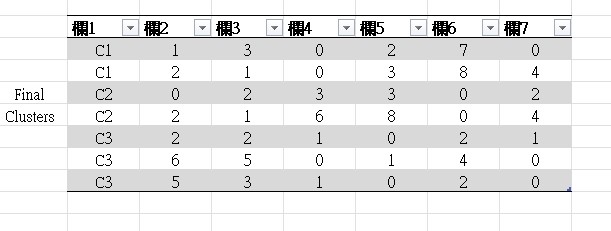


We have to calculate the new cluster centroids and Cosine distance again until the value is the same.





After that we can see the value is the same, so we can stop the calculation.



**2. (30 points) Practice with hierarchical clustering**

Perform a hierarchical clustering of the following data points:

D1: <1, -2>

D2: <-1, 9>

D3: <-6, 4>

D4: <0, -5>

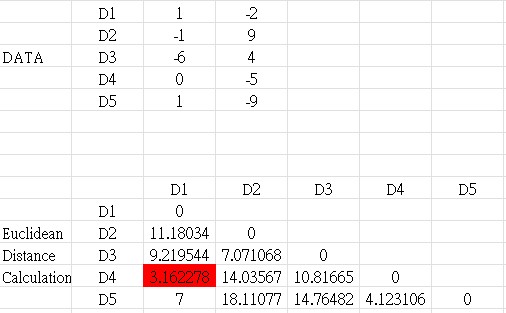
D5: <1, -9>

clusters are represented by their centroid (use means), and at each step the clusters with the closest centroid are merged. And you should use the “Bottom-to-Top” approach to construct your hierarchical tree. Use the Euclidean distance to measure the distance between instances or clusters.

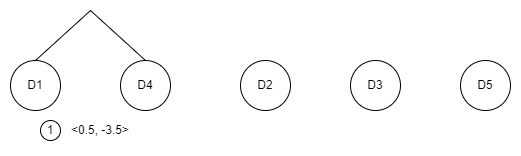
You must draw a hierarchical tree structure and give the sequence (such as, ①②③…) to each merging operation. For example, number ① means this merge is the 1st merge operation. In addition, you must use the centroid value or vector to represent the new value or vector for each group of points. Write down the centroid value or vector to the side of the ① notations.

Your answer:

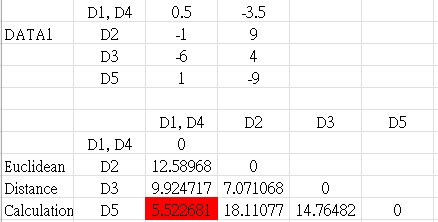
For the following data, we can know D1 and D4 have the closest distance, so they will merge together.



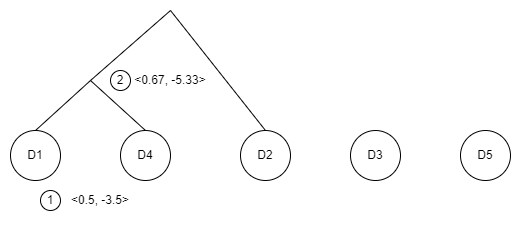
**Tree 1**



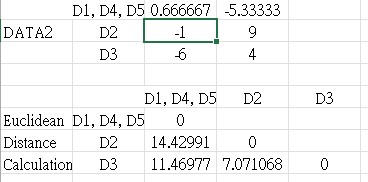
After calculate the new data, we will know D1, D4 have the closest distance with D5, so they will merge together

****

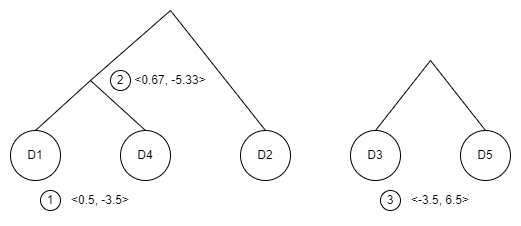
**Tree 2**



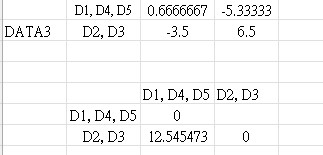
After calculated, D2 and D3 have the closest distance, so they will merge together.

****

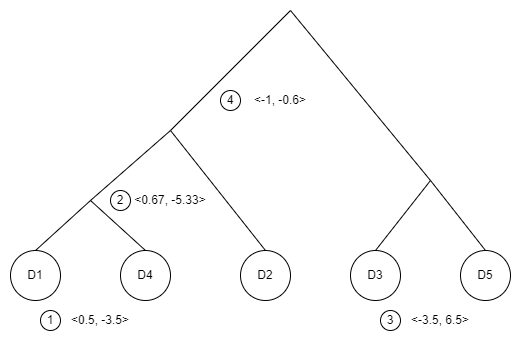
**Tree 3**



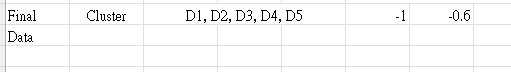
After calculated, D2 and D3 will merge with D1, D4, D5 and become 1 data.



**Tree 4**



And the final data will merge together.



**3. (30 points) Run K-Means on Malware\_MultiClass.csv data**

**Notes**

* **When you run K-Means clustering, you should exclude the label in the classification task**
* **Set maximal learning iteration = 100**
* **Use elbow method to find the optimal K**
* **Evaluate the K-Means results by using at least two methods**