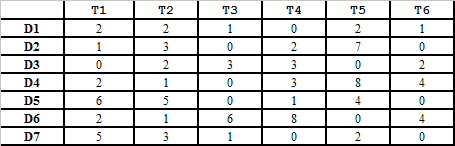
# Homework 7

Your Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

**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 are encouraged to use a spreadsheet program such as Microsoft Excel to facilitate computation in intermediate steps.



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:

**2. (40 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:

**3. (20 points) How to find the best value of K in K-Means clustering**

1). Read the paper X-Means.pdf, and introduce how you can find the best value of K in K-Means clustering [10]

2). Search online, and find at least one extra method to define the best value of K in K-Means clustering [10]