DATA MINING – CSE 5334

PROJECT 1 - KMEANS

NAME: KARTHIK KUMARASUBRAMANIAN
STUDENT ID: 1001549999

1. The cluster centers for the problem 1 are given by,

k = {3, 5}

```
Enter the number of cluster : 3

 For questions 1 & 2

Clusters labels are
 [0 2 0 ..., 1 0 0]
Centroids are
 [[-0.15405503  0.10437123  0.35863273 ..., -0.12051449  0.27503158
  -0.09178372]
 [ 0.14153213 -0.07229925  0.63375808 ...,  1.30527087  0.82383105
   1.45072623]
 [ 0.11522678 -0.09524852 -0.96993902 ..., -0.78006421 -0.98992167
  -0.92674284]]

In [88]: runfile('C:/Users/Karthik Subramanian/Documents/Classes/FALL_2018/
FALL_2018/Data Mining/Project 1/KMeans/q1')


Enter the number of cluster : 5

 For questions 1 & 2

Clusters labels are
 [3 0 2 ..., 2 2 0]
Centroids are
 [[-0.22762499 -0.13958593 -1.09807852 ..., -0.86978814 -0.98725765
  -0.96940962]
 [ 0.55380839  0.07373267  0.60013805 ...,  0.72368361  0.58158753
   0.82105533]
 [-1.14794491 -0.06607409  0.46813293 ...,  0.2536075   1.02578085
   0.36632208]
 [ 0.32792121  0.12503431  0.29624666 ..., -0.20966344 -0.08298033
  -0.26316386]
 [ 0.46768126 -0.03944423  0.64534976 ...,  2.07562443  0.7953158
   2.24380019]]
```

OBSERVATION:

The cluster centers are initially chosen at random for the given data. With the chosen cluster center, the Euclidean distance for each data point in each attribute is found. The data with least Euclidean distance is assigned the cluster in every iteration. The process is repeated till centroid values are the same for all clusters.

The program results in different cluster center when it is run again. Since it randomly assigns a cluster center for each run.

2. The attributes “Tm” and “Player” are redundant in terms of Linear Algebra. They are redundant because they cannot be mapped into any numerical value. The attribute “Pos” can be mapped to numbers and cannot be dropped as it is important feature that can be used in clustering.

3. The cluster centers for the problem 4 are given by,

   k = {3, 5}

```
Enter the number of cluster : 3

 For questions 4

Clusters labels are
 [0 0 1 ..., 1 1 0]
Centroids are
 [[-0.04409634  0.22315181  0.32376812 ..., -0.20274791 -0.28605242
  -0.38745247]
 [ 0.27355119 -0.09689549  0.04221254 ...,  0.6985572   0.91822329
   0.91541477]
 [-0.51771358 -1.0321652  -2.01610189 ..., -0.794098   -0.92930594
  -0.32742437]]

In [91]: runfile('C:/Users/Karthik Subramanian/Documents/Classes/FALL_
FALL_2018/Data Mining/Project 1/KMeans/q4')


Enter the number of cluster : 5

 For questions 4

Clusters labels are
 [1 1 0 ..., 0 0 2]
Centroids are
 [[ 0.59144472 -1.02609817 -0.34274586 ..., -0.24578247  0.05035302
   1.72916945]
 [ 0.04758892  0.52281955  0.20252643 ...,  0.14540883  0.45161204
  -0.19045539]
 [-0.12030008 -0.02512079  0.34353294 ..., -0.50345851 -0.71917496
  -0.47181411]
 [-0.91467364 -0.81449933 -2.6363511  ..., -0.85236343 -0.95342989
  -0.53652562]
 [ 0.01353791  0.52107434  0.47153497 ...,  2.26377644  1.67253532
   0.04336261]]
```

OBSERVATION:

The cluster centers are initially chosen at random for the given data. With the chosen cluster center, the Euclidean distance for each data point in each attribute is found. The data with least Euclidean distance is assigned the cluster in every iteration. The process is repeated till centroid values are the same for all clusters.

The number of attributes is reduced to 7 chosen attributes and clustering is performed on the attributes listed in the question.

The program results in different cluster center when it is run again. Since it randomly assigns a cluster center for each run.