VeriMadenciliği – Data Mining

HW – 6

1. Asagıda mesafe matrisi verilen veride bir noktada iki demete indirgenmiş. A demetinde 1,2,5 noktaları B demetinde ise 3,4,6 noktaları var. Buna göre A demeti ile B demeti arasındaki mesafe maximum mesafe metriğine göre nedir?
   (ENG) For the given distance matrix the algorithm finds two clusters. Cluster A contains points 1,2 and 5 while, cluster B contains points 3,4 and 6. According to the max distance metric, find the distance between clusters A and B.

	p1	p2	p3	p4	p5	p6
p1	0	0,24	0,22	0,37	0,34	0,23
p2	0,24	0	0,15	0,2	0,14	0,25
p3	0,22	0,15	0	0,15	0,28	0,11
p4	0,37	0,2	0,15	0	0,29	0,22
p5	0,34	0,14	0,28	0,29	0	0,39
p6	0,23	0,25	0,11	0,22	0,39	0

Uzaklık matrisi (Öklid)

A demetindeki noktalar ile B demetindeki noktalardan birbirine en uzak olanların arasındaki mesafe.
d(3,5) =0.28 d(4,1) = 0.37 d(6,5)=0.39
en uzağı 0.39 olduğu için A ile B arasıdaki mesafe 0.39

2. k= 2 için k-means uyguluyorum ve bir noktada aşağıdaki 2 demeti elde ettim. Bu demetlerin centroid değerlerini bulun.
   (eng) We found the two clusters below using k-means with k=2. Find the centroids of the clusters.

   A{ (1,3), (2,2)} B{ (2,3), (3,3), (4,1)}

   $C_A = (\frac{1+2}{2}, \frac{3+2}{2})$ $C_B=\left(\frac{2+3+4}{3}, \frac{3+3+1}{3}\right)$

3. DBSCAN algorihması için tek boyutlu verim şu şekilde: {9, 10, 11, 12, 14, 17}

   Bu algorithma için eps=1.2 minPts = 2 seçilmiş.

   a. 9'dan 11'e erişilebilir mi? (eng) Is 9 reachable from 11?
      Hayır, 9 çekirdek değil

   b. 9 ile 12 bağlantılı mı? (eng) Are 9 and 12 connected?
      Evet aynı demetteler, 10dan ikisine de erişilebilir

   c. veride sapan degerler (outlier) ne? (eng) Find the outliers.
      14, 17

Document/term	T1	T2	T3	T4	T5	T6
d1	0	4	10	8	5	0
d2	5	19	7	16	0	32
d3	15	0	0	4	0	17
d4	22	3	12	0	15	0
d5	0	7	0	9	4	12

4. Verilen tabloyu kullanarak, d4 dokumanı için, T5 teriminin normalize edilmiş TF (terim frekansı) degerini hesaplayın TF(d4,T5) =?
   (Eng) For the given table calculate the normalized term frequency for term T5 in document d4 TF(d4,T5) =?

$$TF(t,d) = 0.5 + \frac{0.5 * f(t,d)}{MaxFreq(d)}$$

TF(d4,t5) = 0.5 + (0.5 * 15) / 22 = 0.84, actual tf was 15.

5. Aynı veriye gore T5 teriminin tum veri kumesindeki IDF (inverse document frequency) degerini hesaplayın ve 1'inci sorudaki sonucu da kullanarak d4 dokumanındaki T5 teriminin TF-IDF degerini bulun.
   (eng) For the same data, calculate the IDF (inverse document frequency) for the term T5. Then calculate the TF-IDF value for term T5 in document d4 using the result in question 1.

$$IDF(t) = 1 + log(\frac{n}{k})$$

idf (t5) = 1 + log(5 / 3) = 1.22

tf-idf(d4,t5) = tf(d4,t5) * idf(t5) = 0.84 * 1.22 = 1.024

6. Aynı veri için d1 ve d2 dokumanları arasındaki kosinus benzerliğini (iki vector arasındaki açının kosinusu) bulun. Terim frekansını normalize etmenize gerek yok.
   For the same data, calculate the cosine similarity between documents d1 and d2, which is the cosine of the angle between the corresponding vectors. You do not need to normalize the term frequency.

$$Sim(D_i, D_j) = \frac{\sum_{t=i}^{N} w_{it} * w_{jt}}{\sqrt{\sum_{t=1}^{N} (w_{it})^2 * \sum_{t=1}^{N} (w_{jt})^2}}$$

IDF	t1	t2	t3	t4	t5	t6
	1.22	1.097	1.22	1.097	1.22	1.22

tf-idf	T1	T2	T3	T4	T5	T6
d1	0*1.22	4*1.097	10*1.22	8*1.097	5*1.22	0*1.22
d2	5*1.22	19*1.097	7*1.22	16*1.097	0*1.22	32*1.22

tf-idf	T1	T2	T3	T4	T5	T6
d1	0	4.38	12.2	8.77	6.1	0
d2	6.1	20.84	8.54	17.55	0	39.04

cos(d1,d2) = $\frac{0+(4.38*20.84)+(12.2*8.54)+(8.77*17.55)+(0)+(0)}{\sqrt{0+4.38^2+12.2^2+8.77^2+6.1^2+0} * \sqrt{6.1^2+20.84^2+8.54^2+17.55^2+0+39.04^2}} = \frac{347}{16.6 * 48} = \frac{347}{797} = 0.43$