EE412 Foundation of Big Data Analytics, Fall 2021 HW2

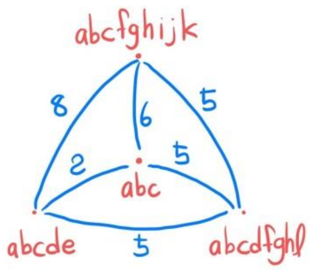
20170616 정희진

※ Because I submitted late, I will use 2 tokens.

1. Clustering

(a) Solve the following problem, which is based on the exercises in the Mining of Massive Datasets 2nd edition (MMDS) textbook.

A set of strings: [abc, abcde, abcfghijk, abcdfghl] 이 있다고 하자. 그렇다면 각 edit distance는 다음과 같다.



각 점의 distance의 합을 구해보면 다음과 같다.

abc	abcde	abcfghijk	abcdfghl
13	15	19	15

각 점의 maximum distance를 구해보면 다음과 같다.

abc	abcde	abcfghijk	abcdfghl
6	8	8	5

따라서 clustroid와 다른 점들과의 distance의 합을 최소화하도록 clustroid를 고르면 abc가 된다.

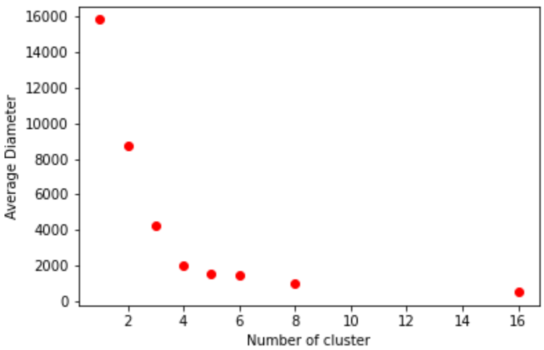
하지만 clustroid와 다른 점들과의 distance 중 maximum distance를 최소화하도록 clustroid를 고르면 abcdfghl이 된다.

(b) Implement the k-Means algorithm using Spark

각 number of cluster에 대한 average diameter은 다음과 같다.

k (Number of cluster)	Average diameter
1	15840.015935162377
2	8712.758000610926
3	4225.773466007263
4	2018.267391109598
5	1526.2137994123182
6	1420.7214130743844
8	1003.3225742993732
16	504.6115133174929

위의 표를 그래프로 그리면 다음과 같다.



처음에는 k = 1, 2, 4, 8, 16일 때의 average diameter를 구했다. 이때, k = 8일 때와 k = 16일 때 많은 변화가 없으므로 k가 8보다 작을 때 변화가 큰 지점이 있다는 것을 알 수 있었다. 따라서 추가적으로 k = 3, 5, 6일 때의 average diameter를 구했다. 그래프를 보면 k = 4일 때 가장 큰 변화가 일어난다. k = 4일 때 변화가 가장 크게 일어나는지 알아보기 위해 다음의 식을 계산해보았다.

$$k = m - 1\text{일때와 } k = m\text{일 때의 } \text{average diameter 차이}$$

m	3	4	5	6
A	4486.985	2207.506	492.0536	105.4924

M= 3, 4일 때는 변화량이 1000 이상이지만 그 이후부터 1000 미만이 된다. 따라서 k = 4일 때 가장 적절하게 군집될 것이다.

2. Dimensionality Reduction

Exercise 11.1.7

문제를 풀기 위한 python 코드는 다음과 같다.

```
import numpy as np

M = np.array([[1,1,1],[1,2,3],[1,3,6]])
x = np.array([[1,1,1]]).T

for i in range(3):
    print("######## Finding %dth eigenpair ########" %(i+1))
    print("Matrix is")
    print(M)
    for j in range(10):
        Mx = np.matmul(M, x)
        x = Mx/ np.linalg.norm(Mx)
#       print("x%d" %(i+1),"is")   # Check the value change
#       print(x)

    eig_vec = x
    print("Eigenvector is")
    print(eig_vec)
    eig_val = np.matmul(np.matmul(eig_vec.T, M), eig_vec)
    print("Eigenvalue is")
    print(eig_val)

    M = M - eig_val * np.matmul(eig_vec, eig_vec.T)
    x = np.array([[1,1,1]]).T
```

(a) Starting with a vector of three 1's, use power iteration to find an approximate value of the principal eigenvector.

```
Eigenvector is
[[0.19382266]
 [0.47224729]
 [0.8598926 ]]
```

(b) Compute an estimate the principal eigenvalue for the matrix.

```
Eigenvalue is
[[7.87298335]]
```

(c) Construct a new matrix by subtracting out the effect of the principal eigenpair, as in Section 11.1.3.

```
Matrix is
[[ 0.70423389  0.27936833 -0.31216389]
 [ 0.27936833  0.24418694 -0.19707639]
 [-0.31216389 -0.19707639  0.17859583]]
```

(d) From your matrix of (c), find the second eigenpair for the original matrix of Exercise 11.1.5.

```
Eigenvector is
[[ 0.81649658]
 [ 0.40824829]
 [-0.40824829]]
Eigenvalue is
[[1.]]
```

(e) Repeat (c) and (d) to find the third eigenpair for the original matrix.

```
Matrix is
[[ 0.03756722 -0.053965    0.02116944]
 [-0.053965    0.07752028 -0.03040973]
 [ 0.02116944 -0.03040973  0.01192916]]
Eigenvector is
[[ 0.54384383]
 [-0.78122713]
 [ 0.30646053]]
Eigenvalue is
[[0.12701665]]
```

Exercise 11.3.1

문제를 풀기 위한 전체 python 코드는 다음과 같다.

```
import numpy as np

M = np.array([[1,2,3],[3,4,5],[5,4,3],[0,2,4],[1,3,5]])

print("(a)#################################################################")

MTM = np.matmul(M.T, M)
MMT = np.matmul(M, M.T)

print("MTM is")
print(MTM)
print("MMT is")
print(MMT)

print("(b)#################################################################")

w1, v1 = np.linalg.eig(MTM)

print("MTM eigenvalues are")
print(w1)
print("MTM eigenvectors are")
print(v1)

w2, v2 = np.linalg.eig(MMT)

print("MMT eigenvalues are")
print(w2)
print("MMT eigenvectors are")
print(v2)

print("(c)#################################################################")

U = v2[:,[0,2]]

S = np.zeros((2,2))
np.fill_diagonal(S, (w1[0:2])**(1/2.0))
```

```
V = v1[:,0:2]

print("M is")
print(M)
print("But USV.T is")
print(np.matmul(np.matmul(U, S),V.T))

U = -1 * U

print("U is")
print(U)

print("S is")
print(S)

print("V is")
print(V)

print("Now USV.T is")
print(np.matmul(np.matmul(U, S), V.T))

print("(d)############################################################")

new_U = U[:,[0]]
new_S = S[0,0]
new_V = V[:,[0]]

print("Reduced U is")
print(new_U)

print("Reduced S is")
print(new_S)

print("Reduced V is")
print(new_V)

new_M = new_S * np.matmul(new_U, new_V.T)
print("Reduced M is")
```

```
print(new_M)


print("(e)###################################################################")
print("The retained energy is")
print("%f%%" %(S[0,0]**2/(S[1,1]**2 + S[0,0]**2)*100))
```

(a) Compute the matrices $M^T M$(MTM) and MM^T(MMT).

```
MTM is
[[36 37 38]
 [37 49 61]
 [38 61 84]]
MMT is
[[14 26 22 16 22]
 [26 50 46 28 40]
 [22 46 50 20 32]
 [16 28 20 20 26]
 [22 40 32 26 35]]
```

(b) Find the eigenpairs (eigenvalues, eigenvectors) for your matrices of part (a) using Python NumPy function (numpy.linalg.eig()).

```
MTM eigenvalues are
[1.53566996e+02 1.54330035e+01 6.69501359e-15]
MTM eigenvectors are
[[-0.40928285 -0.81597848  0.40824829]
 [-0.56345932 -0.12588456 -0.81649658]
 [-0.7176358   0.56420935  0.40824829]]
```

```
MMT eigenvalues are
[ 1.53566996e+02 -6.38239498e-15  1.54330035e+01 -3.86181189e-15
 -1.03191483e-15]
MMT eigenvectors are
[[ 0.29769568  0.94131607 -0.15906393 -0.78673469  0.17238272]
 [ 0.57050856 -0.17481584  0.0332003  -0.09936245  0.08563818]
 [ 0.52074297 -0.04034212  0.73585663  0.21995086 -0.19713125]
 [ 0.32257847 -0.18826321 -0.5103921   0.56795623 -0.78393509]
 [ 0.45898491 -0.21515796 -0.41425998 -0.01493222  0.55635901]]
```

(c) Find the SVD for the original matrix M from parts (b). Note that there are only two nonzero eigenvalues, so your matrix Σ(S) should have only two singular values, while U and V have only two columns.

U, Σ, V를 계산했지만, 실제로 UΣV.T를 계산해보니 -1*M이 나왔다. 이는 위에서 계산한 eigenvector의 (크기만 1일 뿐) 첫번째 element가 음수의 값을 가질 수도, 양수의 값을 가지게 나타날 수 있기 때문이다.

```
M is
[[1 2 3]
 [3 4 5]
 [5 4 3]
 [0 2 4]
 [1 3 5]]
But USV.T is
[[-1.00000000e+00 -2.00000000e+00 -3.00000000e+00]
 [-3.00000000e+00 -4.00000000e+00 -5.00000000e+00]
 [-5.00000000e+00 -4.00000000e+00 -3.00000000e+00]
 [ 1.11022302e-15 -2.00000000e+00 -4.00000000e+00]
 [-1.00000000e+00 -3.00000000e+00 -5.00000000e+00]]
```

따라서 다음과 같은 코드를 넣어 다시 계산해 보았다.

```
U = -1 * U
```

결과는 M과 비슷하게 나왔다.

```
Now USV.T is
[[ 1.00000000e+00  2.00000000e+00  3.00000000e+00]
 [ 3.00000000e+00  4.00000000e+00  5.00000000e+00]
 [ 5.00000000e+00  4.00000000e+00  3.00000000e+00]
 [-1.11022302e-15  2.00000000e+00  4.00000000e+00]
 [ 1.00000000e+00  3.00000000e+00  5.00000000e+00]]
```

따라서 U, Σ(S), V는 다음과 같다.

```
U is
[[-0.29769568  0.15906393]
 [-0.57050856 -0.0332003 ]
 [-0.52074297 -0.73585663]
 [-0.32257847  0.5103921 ]
 [-0.45898491  0.41425998]]
S is
[[12.39221516  0.        ]
 [ 0.          3.92848616]]
V is
[[-0.40928285 -0.81597848]
 [-0.56345932 -0.12588456]
 [-0.7176358   0.56420935]]
```

(d) Set your smaller singular value to 0 and compute the one-dimensional approximation to the matrix M.

Dimension을 줄이기 위해 U, Σ(S), V를 다음처럼 보정했다.

```
Reduced U is
[[-0.29769568]
 [-0.57050856]
 [-0.52074297]
 [-0.32257847]
 [-0.45898491]]
Reduced S is
12.39221515549012
Reduced V is
[[-0.40928285]
 [-0.56345932]
 [-0.7176358 ]]
```

위에서 구한 U, Σ(S), V를 가지고 M을 구하면 다음과 같다.

```
Reduced M is
[[1.509889   2.0786628  2.64743661]
 [2.89357443 3.98358126 5.0735881 ]
 [2.64116728 3.63609257 4.63101787]
 [1.63609257 2.25240715 2.86872172]
 [2.32793529 3.20486638 4.08179747]]
```

(e) How much of the energy of the original singular values is retained by the onedimensional approximation? (Hint: energy = sum of the squares of the singular values)

```
The retained energy is
90.868045%
```

2. Recommendation Systems

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

문제 풀이에서 Jaccard()는 Jaccard similarity
Cosine()은 cosine similarity
Jaccard list의 element의 3번째 값은 Jaccard similarity를 나타낸다

Exercise 9.3.1

(a) $\text{Jaccard}(A,B) = \frac{4}{8} = \frac{1}{2}$ $\therefore \text{Distance}(A,B) = 1 - \frac{1}{2} = 0.5$

$\text{Jaccard}(B,C) = \frac{4}{8} = \frac{1}{2}$ $\therefore \text{Distance}(B,C) = 1 - \frac{1}{2} = 0.5$

$\text{Jaccard}(A,C) = \frac{4}{8} = \frac{1}{2}$ $\therefore \text{Distance}(A,C) = 1 - \frac{1}{2} = 0.5$

(b) $\text{Cosine}(A,B) = \frac{5\times3+5\times3+1\times1+3\times1}{\sqrt{4^2+5^2+5^2+1^2+3^2+2^2}\sqrt{3^2+4^2+3^2+1^2+2^2+1^2}} = \frac{34}{\sqrt{80}\sqrt{40}} = \frac{17}{40}\sqrt{2}$ $\therefore \text{Distance}(A,B) = \cos^{-1}\left(\frac{17}{40}\sqrt{2}\right) = 0.926$

$\text{Cosine}(B,C) = \frac{4\times1+3\times3+2\times4+1\times5}{\sqrt{3^2+4^2+3^2+1^2+2^2+1^2}\sqrt{2^2+1^2+3^2+4^2+5^2+3^2}} = \frac{26}{\sqrt{40}\sqrt{64}} = \frac{13}{80}\sqrt{10}$ $\therefore \text{Distance}(B,C) = \cos^{-1}\left(\frac{13}{80}\sqrt{10}\right) = 1.031$

$\text{Cosine}(A,C) = \frac{4\times2+5\times3+3\times5+2\times3}{\sqrt{4^2+5^2+5^2+1^2+3^2+2^2}\sqrt{2^2+1^2+3^2+4^2+5^2+3^2}} = \frac{44}{\sqrt{80}\sqrt{64}} = \frac{11}{40}\sqrt{5}$ $\therefore \text{Distance}(A,C) = \cos^{-1}\left(\frac{11}{40}\sqrt{5}\right) = 0.909$

(c)

	a	b	c	d	e	f	g	h
A	1	1	0	1	0	0	1	0
B	0	1	1	1	0	0	0	0
C	0	0	0	1	0	1	1	1

$\text{Jaccard}(A,B) = \frac{2}{5}$ $\therefore \text{Distance}(A,B) = 1 - \frac{2}{5} = \frac{3}{5}$

$\text{Jaccard}(B,C) = \frac{1}{6}$ $\therefore \text{Distance}(B,C) = 1 - \frac{1}{6} = \frac{5}{6}$

$\text{Jaccard}(A,C) = \frac{2}{6} = \frac{1}{3}$ $\therefore \text{Distance}(A,C) = 1 - \frac{1}{3} = \frac{2}{3}$

(d) $\text{Cosine}(A,B) = \frac{1\times1+1\times1}{\sqrt{1^2+1^2+1^2+1^2}\sqrt{1^2+1^2+1^2}} = \frac{2}{\sqrt{4}\sqrt{3}} = \frac{\sqrt{3}}{3}$ $\therefore \text{Distance}(A,B) = \cos^{-1}\left(\frac{\sqrt{3}}{3}\right) = 0.955$

$\text{Cosine}(B,C) = \frac{1\times1}{\sqrt{1^2+1^2+1^2}\sqrt{1^2+1^2+1^2+1^2}} = \frac{1}{\sqrt{4}\sqrt{3}} = \frac{\sqrt{3}}{6}$ $\therefore \text{Distance}(B,C) = \cos^{-1}\left(\frac{\sqrt{3}}{6}\right) = 1.278$

$\text{Cosine}(A,C) = \frac{1\times1+1\times1}{\sqrt{1^2+1^2+1^2+1^2}\sqrt{1^2+1^2+1^2+1^2}} = \frac{2}{4} = \frac{1}{2}$ $\therefore \text{Distance}(A,C) = \cos^{-1}\left(\frac{1}{2}\right) = 1.047$

(e) $\text{mean } A = \frac{4+5+5+1+3+2}{6} = \frac{20}{6} = \frac{10}{3}$

$\text{mean } B = \frac{3+4+3+1+2+1}{6} = \frac{14}{6} = \frac{7}{3}$

$\text{mean } C = \frac{2+1+3+4+5+3}{6} = \frac{18}{6} = 3$

	a	b	c	d	e	f	g	h
A	$\frac{2}{3}$	$\frac{5}{3}$		$\frac{5}{3}$	$-\frac{7}{3}$		$-\frac{1}{3}$	$-\frac{4}{3}$
B		$\frac{2}{3}$	$\frac{5}{3}$	$\frac{2}{3}$	$-\frac{4}{3}$	$-\frac{1}{3}$	$-\frac{4}{3}$	
C	-1		-2	0		1	2	0

(f)

$$\text{Cosine}(A,B) = \frac{\frac{5}{3}\times\frac{2}{3}+\frac{5}{3}\times\frac{2}{3}+\left(-\frac{7}{3}\right)\times\left(-\frac{4}{3}\right)+\left(-\frac{1}{3}\right)\times\left(-\frac{4}{3}\right)}{\sqrt{\left(\frac{2}{3}\right)^2+\left(\frac{5}{3}\right)^2+\left(\frac{5}{3}\right)^2+\left(-\frac{7}{3}\right)^2+\left(-\frac{1}{3}\right)^2+\left(-\frac{4}{3}\right)^2}\sqrt{\left(\frac{2}{3}\right)^2+\left(\frac{5}{3}\right)^2+\left(\frac{2}{3}\right)^2+\left(-\frac{4}{3}\right)^2+\left(-\frac{1}{3}\right)^2+\left(-\frac{4}{3}\right)^2}} = \frac{\frac{52}{9}}{\sqrt{\frac{120}{9}}\sqrt{\frac{66}{9}}} = \frac{13}{165}\sqrt{55}$$

$$\text{Cosine}(B,C) = \frac{\frac{5}{3}\times(-2)-\frac{1}{3}\times1-\frac{4}{3}\times2}{\sqrt{\left(\frac{2}{3}\right)^2+\left(\frac{5}{3}\right)^2+\left(\frac{2}{3}\right)^2+\left(-\frac{4}{3}\right)^2+\left(-\frac{1}{3}\right)^2+\left(-\frac{4}{3}\right)^2}\sqrt{(-1)^2+(-2)^2+(-1)^2+(-2)^2}} = \frac{-\frac{19}{3}}{\sqrt{\frac{66}{9}}\sqrt{10}} = \frac{-19}{660}\sqrt{660}$$

$$\text{Cosine}(A,B) = \frac{\frac{2}{3}\times(-1)-\frac{1}{3}\times2}{\sqrt{\left(\frac{2}{3}\right)^2+\left(\frac{5}{3}\right)^2+\left(\frac{5}{3}\right)^2+\left(-\frac{7}{3}\right)^2+\left(-\frac{1}{3}\right)^2+\left(-\frac{4}{3}\right)^2}\sqrt{(-1)^2+(-2)^2+(-1)^2+(-2)^2}} = \frac{-\frac{4}{3}}{\sqrt{\frac{120}{9}}\sqrt{10}} = -\frac{\sqrt{3}}{15}$$

Distance(A,B) = $\cos^{-1}\left(\frac{13}{165}\sqrt{55}\right) = 0.947$

Distance(B,C) = $\cos^{-1}\left(\frac{-19}{660}\sqrt{660}\right) = 2.403$

Distance(A,C) = $\cos^{-1}\left(-\frac{\sqrt{3}}{15}\right) = 1.687$

Exercise 9.3.2

(a)

	a	b	c	d	e	f	g	h
A	1	1	0	1	0	0	1	0
B	0	1	1	1	0	0	0	0
C	0	0	0	1	0	1	1	1

Jaccard list = [$(a,b,\frac{1}{2})$, $(a,c,0)$, $(a,d,\frac{1}{3})$, $(a,e,0)$, $(a,f,0)$, $(a,g,\frac{1}{2})$, $(a,h,0)$, $(b,c,\frac{1}{2})$, $(b,d,\frac{2}{3})$, $(b,e,0)$,
$(b,f,0)$, $(b,g,\frac{1}{3})$, $(b,h,0)$, $(c,d,\frac{1}{3})$, $(c,e,0)$, $(c,f,0)$, $(c,g,0)$, $(c,h,0)$, $(d,e,0)$, $(d,f,\frac{1}{3})$,
$(d,g,\frac{2}{3})$, $(d,h,\frac{1}{3})$, $(e,f,0)$, $(e,g,0)$, $(e,h,0)$, $(f,g,\frac{1}{2})$, $(f,h,1)$, $(g,h,\frac{1}{2})$]

Jaccard similarity가 1일 때가 가장 높으므로 가장 먼저 f,h가 한 cluster가 될 것이다

다음으로 similarity가 $\frac{2}{3}$일 때 가장 높으므로 b,d,g가 한 cluster가 될 것이다

다음으로 similarity가 $\frac{1}{2}$일때 가장 높다 $(a,b,\frac{1}{2})$, $(a,g,\frac{1}{2})$, $(b,c,\frac{1}{2})$, $(f,g,\frac{1}{2})$이 있는데, 이 순서대로 우선 순위를 둔다면($(a,b,\frac{1}{2})$가 우선 순위가 가장 높다면) a,b,d,g가 한 cluster가 된다

cluster1: a,b,d,g, cluster2: c, cluster3: e, cluster4: f,h

(b)

	a,b,d,g	c	e	f,h
A	$\frac{17}{4}$		1	2
B	$\frac{7}{3}$	4	1	2
C	$\frac{10}{3}$	1		$\frac{7}{2}$

(c)

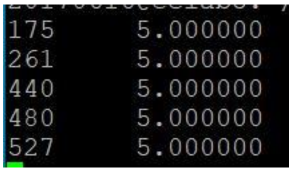
$$\text{Cosine}(A,B) = \frac{\frac{17}{4}\times\frac{7}{3}+1\times1+2\times2}{\sqrt{\left(\frac{17}{4}\right)^2+1^2+2^2}\sqrt{\left(\frac{7}{3}\right)^2+4^2+1^2+2^2}} = 0.604 \quad \text{Distance}(A,B) = \cos^{-1}(0.604) = 0.922$$

$$\text{Cosine}(B,C) = \frac{\frac{7}{3}\times\frac{10}{3}+4\times1+2\times\frac{7}{2}}{\sqrt{\left(\frac{7}{3}\right)^2+4^2+1^2+2^2}\sqrt{\left(\frac{10}{3}\right)^2+1^2+\left(\frac{7}{2}\right)^2}} = 0.740 \quad \text{Distance}(B,C) = \cos^{-1}(0.740) = 0.738$$

$$\text{Cosine}(A,C) = \frac{\frac{17}{4}\times\frac{10}{3}+2\times\frac{7}{2}}{\sqrt{\left(\frac{17}{4}\right)^2+1^2+2^2}\sqrt{\left(\frac{10}{3}\right)^2+1^2+\left(\frac{7}{2}\right)^2}} = 0.893 \quad \text{Distance}(A,C) = \cos^{-1}(0.893) = 0.467$$

(b) Implement collaborative filtering

User-based Result



```
175      5.000000
261      5.000000
440      5.000000
480      5.000000
527      5.000000
```

Item-base Result는 정상적으로 작동하는 것 같으나 시간이 오래 걸려 구하지 못했다.