CS 6350 – Big Data Management and Analytics
Fall 2015
Assignment #3 Part B
*** Solution ***

Please contact `lkc130030@utdallas.edu` if you find any errors.

Problem 1

(a) $d_J(U_1, U_2) = 1 - J(U_1, U_2) = 1 - |U_1 \cap U_2|/|U_1 \cup U_2| = 1 - 4/8 = 1/2 = .5$
$d_J(U_1, U_3) = 1 - J(U_1, U_3) = 1 - |U_1 \cap U_3|/|U_1 \cup U_3| = 1 - 4/8 = 1/2 = .5$
$d_J(U_2, U_3) = 1 - J(U_2, U_3) = 1 - |U_2 \cap U_3|/|U_2 \cup U_3| = 1 - 4/8 = 1/2 = .5$

(b) $||U_1|| = \sqrt{4^2 + 5^2 + 5^2 + 1^2 + 3^2 + 2^2} = 8.94$
$||U_2|| = \sqrt{3^2 + 4^2 + 3^2 + 1^2 + 2^2 + 1^2} = 6.32$
$||U_3|| = \sqrt{2^2 + 1^2 + 3^2 + 4^2 + 5^2 + 3^2} = 8$

$$d_{\cos}(U_1, U_2) = 1 - \cos_s(U_1, U_2) = 1 - \frac{(5)(3) + (5)(3) + (1)(1) + (3)(1)}{||U_1||\ ||U_2||} = .399$$

$$d_{\cos}(U_1, U_3) = 1 - \cos_s(U_1, U_3) = 1 - \frac{(4)(2) + (5)(3) + (3)(5) + (2)(3)}{||U_1||\ ||U_3||} = .385$$

$$d_{\cos}(U_2, U_3) = 1 - \cos_1(U_2, U_3) = 1 - \frac{(4)(1) + (3)(3) + (2)(4) + (1)(5)}{||U_2||\ ||U_3||} = .486$$

(c) $||U_1|| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = 2.449$
$||U_2|| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = 2.449$
$||U_3|| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = 2.449$

$$d_{\cos}(U_1, U_2) = 1 - \cos_s(U_1, U_2) = 1 - \frac{(1)(1) + (1)(1) + (1)(1) + (1)(1)}{||U_1||\ ||U_2||} = .333$$

$$d_{\cos}(U_1, U_3) = 1 - \cos_s(U_1, U_3) = 1 - \frac{(1)(1) + (1)(1) + (1)(1) + (1)(1)}{||U_1||\ ||U_3||} = .333$$

$$d_{\cos}(U_2, U_3) = 1 - \cos_s(U_2, U_3) = 1 - \frac{(1)(1) + (1)(1) + (1)(1) + (1)(1)}{||U_2||\ ||U_3||} = .333$$

(d)

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
U_1	1	1	0	1	0	0	1	0
U_2	0	1	1	1	0	0	0	0
U_3	0	0	0	1	0	1	1	1

$$d_J(U_1,U_2) = 1 - J(U_1,U_2) = 1 - |U_1 \cap U_2|/|U_1 \cup U_2| = 1 - 2/5 = 3/5 = .6$$
$$d_J(U_1,U_3) = 1 - J(U_1,U_3) = 1 - |U_1 \cap U_3|/|U_1 \cup U_3| = 1 - 2/6 = 2/3 = .667$$
$$d_J(U_2,U_3) = 1 - J(U_2,U_3) = 1 - |U_2 \cap U_3|/|U_2 \cup U_3| = 1 - 1/6 = 5/6 = .833$$

(e) $$d_{\cos}(U_1,U_2) = 1 - \cos_s(U_1,U_2) = 1 - \frac{(1)(1)+(1)(1)}{\sqrt{1^2+1^2+1^2+1^2}\sqrt{1^2+1^2+1^2}} = .423$$

$$d_{\cos}(U_1,U_3) = 1 - \cos_s(U_1,U_3) = 1 - \frac{(1)(1)+(1)(1)}{\sqrt{1^2+1^2+1^2+1^2}\sqrt{1^2+1^2+1^2+1^2}} = .5$$

$$d_{\cos}(U_2,U_3) = 1 - \cos_s(U_2,U_3) = 1 - \frac{(1)(1)}{\sqrt{1^2+1^2+1^2}\sqrt{1^2+1^2+1^2+1^2}} = .711$$

(f)

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
U_1	2/3	5/3		5/3	-7/3		-1/3	-4/3
U_2		2/3	5/3	2/3	-4/3	-1/3	-4/3	
U_3	-1		-2	0		1	2	0

(g) $$||U_1|| = \sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{5}{3}\right)^2 + \left(\frac{5}{3}\right)^2 + \left(-\frac{7}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{4}{3}\right)^2} = 3.651$$
$$||U_2|| = \sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{5}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(-\frac{4}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{4}{3}\right)^2} = 2.708$$
$$||U_3|| = \sqrt{(-1)^2 + (-2)^2 + 0^2 + 1^2 + 2^2 + 0^2} = 3.162$$

$$\begin{aligned} d_{\cos}(U_1,U_2) &= 1 - \cos_s(U_1,U_2) \\ &= 1 - \frac{\left(\frac{5}{3}\right)\left(\frac{2}{3}\right) + \left(\frac{5}{3}\right)\left(\frac{2}{3}\right) + \left(-\frac{7}{3}\right)\left(-\frac{4}{3}\right) + \left(-\frac{1}{3}\right)\left(-\frac{4}{3}\right)}{||U_1||\ ||U_2||} = .416 \end{aligned}$$

$$\begin{aligned} d_{\cos}(U_1,U_3) &= 1 - \cos_s(U_1,U_3) \\ &= 1 - \frac{\left(\frac{2}{3}\right)(-1) + \left(\frac{5}{3}\right)(0) + \left(-\frac{1}{3}\right)(2) + \left(-\frac{4}{3}\right)(0)}{||U_1||\ ||U_3||} = 1.115 \end{aligned}$$

$$\begin{aligned} d_{\cos}(U_2,U_3) &= 1 - \cos_s(U_2,U_3) \\ &= 1 - \frac{\left(\frac{5}{3}\right)(-2) + \left(\frac{2}{3}\right)(0) + \left(-\frac{1}{3}\right)(1) + \left(-\frac{4}{3}\right)(2)}{||U_2||\ ||U_3||} = 1.74 \end{aligned}$$

Problem 2

(a) Initial Jaccard distance matrix

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
i_1	0	1/2	1	2/3	1	1	1/2	1
i_2	1/2	0	1/2	1/3	1	1	2/3	1
i_3	1	1/2	0	2/3	1	1	1	1
i_4	2/3	1/3	2/3	0	1	2/3	1/3	2/3
i_5	1	1	1	1	0	1	1	1
i_6	1	1	1	2/3	1	0	1/2	**0**
i_7	1/2	2/3	1	1/3	1	1/2	0	1/2
i_8	1	1	1	2/3	1	0	1/2	0

1. Merge i_6 and i_8

Jaccard distance matrix after 1.

	i_1	i_2	i_3	i_4	i_5	(i_6, i_8)	i_7
i_1	0	1/2	1	2/3	1	1	1/2
i_2	1/2	0	1/2	**1/3**	1	1	2/3
i_3	1	1/2	0	2/3	1	1	1
i_4	2/3	1/3	2/3	0	1	2/3	**1/3**
i_5	1	1	1	1	0	1	1
(i_6, i_8)	1	1	1	2/3	1	0	1/2
i_7	1/2	2/3	1	1/3	1	1/2	0

2. Can merge either i_2 and i_4 or i_4 and i_7. However, either choice will end up with all three in a cluster resulting in the following Jaccard distance matrix:

	i_1	(i_2, i_4, i_7)	i_3	i_5	(i_6, i_8)
i_1	0	**1/2**	1	1	1
(i_2, i_4, i_7)	1/2	0	**1/2**	1	**1/2**
i_3	1	1/2	0	1	1
i_5	1	1	1	0	1
(i_6, i_8)	1	1/2	1	1	0

3. Can merge either i_1 and (i_2, i_4, i_7) or (i_2, i_4, i_7) and i_3 or (i_2, i_4, i_7) and (i_6, i_8). The following three final clusterings are all valid:

	C_1	C_2	C_3	C_4
1	(i_1, i_2, i_4, i_7)	i_3	i_5	(i_6, i_8)
2	i_1	(i_2, i_3, i_4, i_7)	i_5	(i_6, i_8)
3	i_1	$(i_2, i_4, i_6, i_7, i_8)$	i_3	i_5

(b) Clustering choice 1:

	C_1	C_2	C_3	C_4
U_1	(4+5+5+3)/4 = 4.25		1	2
U_2	(3+3+1)/3 = 2.33	4	1	2
U_3	(2+3+5)/3 = 3.33	1		(4+3)/2 = 3.5

Clustering choice 2:

	C_1	C_2	C_3	C_4
U_1	4	(5+5+3)/3 = 4.33	1	2
U_2		(3+4+3+1)/4 = 2.75	1	2
U_3	2	(1+3+5)/3 = 3		(4+3)/2 = 3.5

Clustering choice 3:

	C_1	C_2	C_3	C_4
U_1	4	(5+5+3+2)/4 = 3.75		1
U_2		(3+3+2+1)/4 = 2.25	4	1
U_3	2	(3+4+5+3)/4 = 3.75	1	

(c) Clustering choice 1:

$$||U_1|| = \sqrt{4.25^2 + 1^2 + 2^2},\ ||U_2|| = \sqrt{2.33^2 + 4^2 + 1^2 + 2^2},\ ||U_3|| = \sqrt{3.33^2 + 1^2 + 3.5^2}$$

$$d_{\cos}(U_1, U_2) = 1 - \frac{(4.25)(2.33) + (1)(1) + (2)(2)}{||U_1||\ ||U_2||} = .396$$

$$d_{\cos}(U_1, U_3) = 1 - \frac{(4.25)(3.33) + (2)(3.5)}{||U_1||\ ||U_3||} = .107$$

$$d_{\cos}(U_2, U_3) = 1 - \frac{(2.33)(3.33) + (4)(1) + (2)(3.5)}{||U_2||\ ||U_3||} = .26$$

Clustering choice 2:

$$||U_1|| = \sqrt{4^2 + 4.33^2 + 1^2 + 2^2},\ ||U_2|| = \sqrt{2.75^2 + 1^2 + 2^2},\ ||U_3|| = \sqrt{2^2 + 3^2 + 3.5^2}$$

$$d_{\cos}(U_1, U_2) = 1 - \frac{(4.33)(2.75) + (1)(1) + (2)(2)}{||U_1||\ ||U_2||} = .243$$

$$d_{\cos}(U_1, U_3) = 1 - \frac{(4)(2) + (4.33)(3) + (2)(3.5)}{||U_1||\ ||U_3||} = .116$$

$$d_{\cos}(U_2, U_3) = 1 - \frac{(2.75)(3) + (2)(3.5)}{||U_2||\ ||U_3||} = .144$$

Clustering choice 3:

$$||U_1|| = \sqrt{4^2 + 3.75^2 + 1^2},\ ||U_2|| = \sqrt{2.25^2 + 4^2 + 1^2},\ ||U_3|| = \sqrt{2^2 + 3.75^2 + 1^2}$$

$$d_{\cos}(U_1, U_2) = 1 - \frac{(3.75)(2.25) + (1)(1)}{||U_1||\ ||U_2||} = .639$$

$$d_{\cos}(U_1, U_3) = 1 - \frac{(4)(2) + (3.75)(3.75)}{||U_1||\ ||U_3||} = .093$$

$$d_{\cos}(U_2, U_3) = 1 - \frac{(2.25)(3.75) + (4)(1)}{||U_2||\ ||U_3||} = .394$$

Problem 3

(a) Iteration 1:

Euclidean distances from each point to each cluster center:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
C_1	6.4	4.2	2	4.1	0	2	6.3	5
C_2	7.8	5.8	4	5	2	0	6	4.1

Cluster assignments:

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
C_1	C_1	C_1	C_1	C_1	C_2	C_2	C_2

Updated cluster center coordinates:

	C_1	C_2
x	4	8.3
y	4.6	3.7

Iteration 2:

Euclidean distances from each point to each cluster center:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
C_1	3.2	1.2	2.4	1.9	3.1	4.7	5.4	5.3
C_2	6.5	5.3	5.5	3.4	4.1	3.4	2.7	0.9

Cluster assignments:

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
C_1	C_1	C_1	C_1	C_1	C_2	C_2	C_2

Converged

(b) Iteration 1:

Euclidean distances from each point to each cluster center:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
C_1	5.4	3.2	0	4.1	2	4	7.2	6.4
C_2	6.1	5.8	7.2	3.6	6.3	6	0	2.2

Cluster assignments:

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
C_1	C_1	C_1	C_2	C_1	C_1	C_2	C_2

Updated cluster center coordinates:

	C_1	C_2
x	4.6	7.3
y	5.4	2.3

Iteration 2:

Euclidean distances from each point to each cluster center:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
C_1	4.3	2.1	1.7	2.4	2.1	3.8	5.6	5
C_2	5.3	4.6	5.7	2.4	4.9	4.7	1.5	1.8

Cluster assignments:

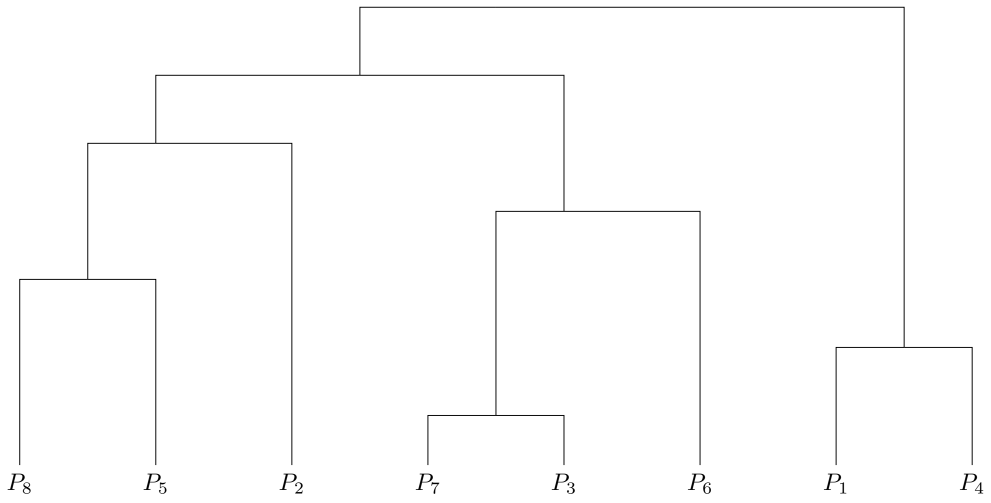
P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
C_1	C_1	C_1	C_2	C_1	C_1	C_2	C_2

Converged

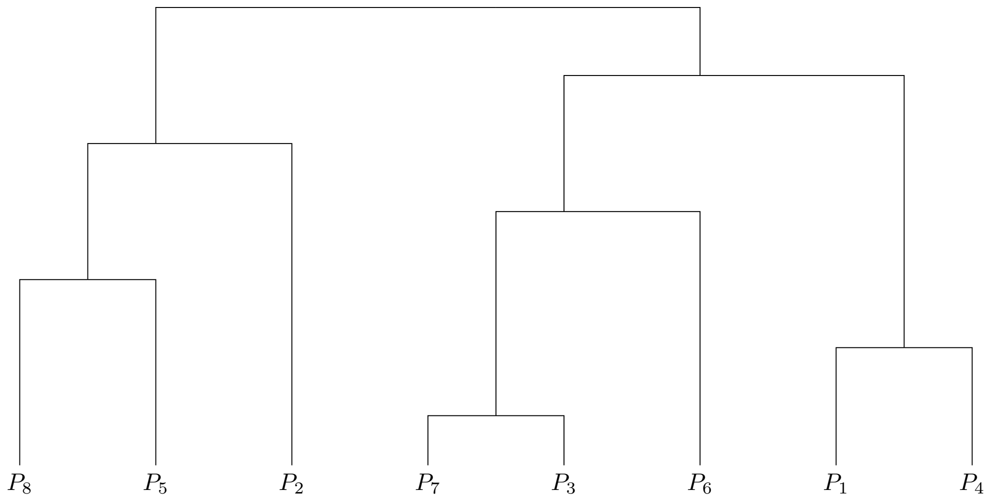
Yes, clusters changed

Problem 4

(a)(i)



(a)(ii)



(b)(i) $8 - 6.5 = 1.5$

(b)(ii) $12 - 1 = 11$

(b)(iii) $(7 + 10 + 11 + 5.8 + 8.8 + 9.8 + 4 + 7 + 8 + 2 + 5 + 6 + 1.5 + 4.5 + 5.5)/15 = 6.39$