Homework 3.

1. K Means on Documents

$D_3(1,6,3)$ and $D_4(1,3,6)$ are initial cluster centers.

Let K_1 & K_2 be their respective clusters.

Iteration 1

$$\therefore \text{Sim}(D_3, D_1) = \frac{1+4+0}{6+6+3} = \frac{5}{15} = 1/3$$

$$\text{Sim}(D_4, D_1) = \frac{1+3+0}{6+4+6} = \frac{4}{16} = 1/4$$

$$\text{sim}(D_3, D_2) = \frac{1+0+3}{3+6+7} = \frac{4}{16} = 1/4$$

$$\text{sim}(D_4, D_2) = \frac{1+0+6}{3+3+7} = \frac{7}{13}$$

$$\text{Sim}(D_3, D_5) = \frac{1+2+2}{6+6+3} = \frac{5}{15} = \frac{1}{3}$$

$$\text{Sim}(D_4, D_5) = \frac{1+2+2}{6+3+6} = \frac{5}{15} = 1/3$$

So D_1 is more similar to D_3.

D_2 is more similar to D_4.

D_5 can go to either one.

$K_1 = \{D_1, D_3, D_5\}$ $K_2 = \{D_2, D_4\}$

Cluster dissimilarity measure

$K_1 = 1 + \frac{1}{3} + \frac{1}{3} = \frac{5}{3}$

$K_2 = Sim(D_4, D_2) = \frac{7}{13} + Sim(D_4, D_4)$

$= \frac{7}{13} + 1 = \frac{20}{13}$

Iteration 2

Computing new cluster centers for K_1 & K_2 we get

$C_1 = \frac{D_1 + D_3 + D_5}{3} = \frac{6+1+6}{3}, \frac{4+6+2}{3}, \frac{0+3+2}{3}$

$= \left(\frac{13}{3}, 4, \frac{5}{3}\right)$

$C_2 = \frac{D_2 + D_4}{2} = \frac{3+1}{2}, \frac{0+3}{2}, \frac{7+6}{2}$

$= 2, 3/2, 13/2$

Now calculating similarities again.

$Sim(D_1, C_1) = \frac{13/3 + 4 + 0}{6 + 4 + 5/3} = \frac{5}{7}$

$Sim(D_1, C_2) = \frac{2 + 3/2 + 0}{6 + 4 + 13/2} = \frac{7}{33}$

$$\text{Sim}(D_2, C_1) = \frac{3 + 0 + 5/3}{13/3 + 4 + 7} = \frac{7}{23}$$

$$\text{Sim}(D_2, C_2) = \frac{2 + 0 + 13/2}{3 + 3/2 + 7} = \frac{17}{23}$$

$$\text{Sim}(D_3, C_1) = \frac{1 + 4 + 5/3}{13/3 + 6 + 3} = 1/2$$

$$\text{Sim}(D_3, C_2) = \frac{1 + 3/2 + 3}{2 + 6 + 13/2} = \frac{11}{29}$$

$$\text{Sim}(D_4, C_1) = \frac{1 + 3 + 5/3}{13/3 + 4 + 6} = \frac{17}{43}$$

$$\text{Sim}(D_4, C_2) = \frac{1 + 3/2 + 6}{2 + 3 + 13/2} = \frac{17}{23}$$

$$\text{sim}(D_5, C_1) = \frac{13/3 + 2 + 5/3}{6 + 4 + 2} = 2/3$$

$$\text{sim}(D_5, C_2) = \frac{2 + 3/2 + 2}{6 + 2 + 13/2} = \frac{11}{29}$$

Assuming clusters we get

$C_1 \rightarrow D_1, D_3, D_5$

$C_2 \rightarrow D_2, D_4$

Cluster Dissimilarity

$$K_1 = \text{Sim}(D_1, C_1) + \text{Sim}(D_3, C_1) + \text{Sim}(D_5, C_1)$$

$$= \frac{5}{7} + \frac{1}{2} + \frac{2}{3} = 1.88$$

$$K_2 = \text{Sim}(D_2, C_2) + \text{Sim}(D_4, C_2)$$

$$= \frac{17}{23} + \frac{17}{23} = 1.48$$

Q:2 Hierarchical agglomerative clustering.

Similarity matrix.

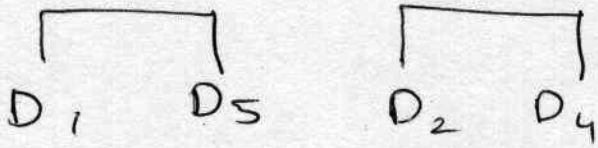
	D_1	D_2	D_3	D_4	D_5
D_1	1	3/17	1/3	1/4	2/3
D_2		1	1/4	7/13	1/3
D_3			1	7/13	1/3
D_4				1	1/3
D_5					1

With single link measure the cluster distance is the distance between closest points.

1. Max similarity 2/3 bet^n D_1 & D_5

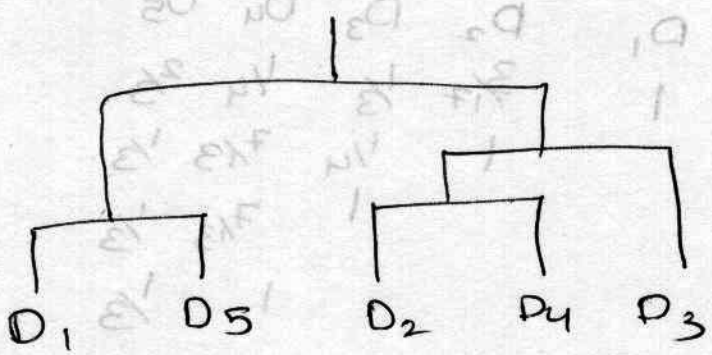
2. Max similarity $\frac{7}{13}$ bet^n D_2 & D_4, D_3 & D_4

taking D_2 & D_4



3. D_3 is left Max similarity of

D_3 for 2nd cluster is $7/13$ b/w D_3 & D_4



So clusters are (D_1, D_5) & (D_2, D_4, D_3)

Q:3

A: The carbon atom is the foundation of life on earth.

$P(\text{Physics} \mid \text{atom, carbon, life, earth})$

$\approx P(\text{atom/physics}) \times P(\text{carbon/physics}) \times P(\text{life/physics}) \times P(\text{earth/physics}) \times P(\text{Physics})$

$= 0.1 \times 0.005 \times 0.001 \times 0.005 \times 0.35$

$= 8.75 \times 10^{-10}$

$P(\text{Biology} \mid \text{atom, carbon, life, earth})$

$\approx P(\text{atom/Bio})\, P(\text{carbon/Bio})\, P(\text{life/Bio})\, P(\text{earth/Bio})\, P(\text{Bio})$

$= 0.01 \times 0.03 \times 0.1 \times 0.006 \times 0.4$

$= 7.2 \times 10^{-8}$

$P(\text{Chemistry} \mid \text{atom, carbon, life, earth})$

$\approx P(\text{atom/Ch})\, P(\text{carbon/Ch})\, P(\text{life/Ch})\, P(\text{earth/Ch})$

$= 0.2 \times 0.05 \times 0.008 \times 0.003 \times 0.25 = 6 \times 10^{-8}$

∴ The most likely classification of sentence A is Biology.

B: the carbon atom contains 12 protons.

$$P(\text{Physics} \mid \text{carbon, atom, proton}) \approx P(\text{atom}/\text{phy})\, P(\text{carbon}/\text{phy})\, P(\text{proton}/\text{phy})\, P(\text{Phy})$$

$$= 0.1 \times 0.005 \times 0.05 \times 0.35$$

$$= 8.75 \times 10^{-6}$$

$$P(\text{Biology} \mid \text{carbon, atom, proton}) \approx P(\text{atom}/\text{Bio})\, P(\text{carbon}/\text{Bio})\, P(\text{proton}/\text{Bio})\, P(\text{Bio})$$

$$\approx 0.01 \times 0.03 \times 0.001 \times 0.4$$

$$= 1.2 \times 10^{-7}$$

$$P(\text{Chemistry} \mid \text{atom, carbon, proton}) \approx 0.2 \times 0.05 \times 0.05 \times 0.25$$

$$= 1.25 \times 10^{-4}$$

∴ The most likely category is Chemistry.

Q:4 Collaborative Filtering:

$$\text{Correlation}(U1, AU) = \frac{(10-6)(9-5)+0+(6-2)(5-1)}{3\sqrt{\frac{1}{3}(4^2+4^2)}\sqrt{\frac{1}{3}(4^2+4^2)}}$$

$$= 1$$

$$\text{correlation}(U_2, AU) = \frac{(5-7)(9-5)+(9-7)(5-5)+(7-7)(5-1)}{3\sqrt{\frac{1}{3}(2^2+2^2)}\sqrt{\frac{1}{3}(4^2+4^2)}}$$

$$= -0.5$$

$$\text{correlation}(U_3, AU) = \frac{(9-6)(9-5)+(3-6)(1-5)}{2\sqrt{\frac{1}{2}(3^2+4^2)}\sqrt{\frac{1}{2}(3^2+4^2)}}$$

$$= 1.$$

Correlation $(U_4, AU) = 1$

(∵ Because only one item is rated by u_4 in common with Au and same rating by both.

Users 1 & 3 can be considered as highest correlated users.

$$P_{aD} = 5 + \frac{(4-6) + (8-9)(3-6)}{2}$$

$$= 2.5$$

$$P_{aE} = 5 + \frac{(8-6) + (9-6)}{2}$$

$$= 7.5$$