Série de TD N°3
Clustering 2 (Clustering Hiérarchique)

Exercice1 :

Le dendrogramme est une représentation graphique d'une classification (clustering) hiérarchique par un arbre.

1- Le dendrogramme d'une classification ascendante (ou descendante) est-il unique ? Si OUI dites Comment ? Si NON dites pourquoi ?

2- Comment déterminer le nombre de classes (clusters) à partir du dendrogramme ? Cette méthode est-elle exacte ou approximative ? Quel est le moyen le plus efficace pour avoir un nombre de classes (clusters) proche de la réalité?

3- Considérer la matrice de similarité suivante de cinq documents d1, d2, d3, d4 et d5. Déterminer le dendrogramme résultant de l'application du text clustering hiérarchique ascendant en utilisant le « **lien maximum** ».

	d1	**d2**	**d3**	**d4**	**d5**
d1	0	0.5	0.5	0.6	0.8
d2	0.5	0	0.7	0.6	0.5
d3	0.5	0.7	0	0.6	0.5
d4	0.6	0.6	0.6	0	0.9
d5	0.8	0.5	0.5	0.9	0

Exercice2 :

Considérer la matrice de dissimilarité suivante P.

Déterminer les dendrogrammes résultants de l'application du « **single link algorithm** », puis du « **complete link algorithm** » sur P et commentez.

a	**b**	**c**	**d**	**e**
0	4	9	6	5
4	0	3	8	7
9	3	0	3	2
6	8	3	0	1
5	7	2	1	0

Exercice3 :

Considérons la matrice de similarité suivante entre les exemples x_1, x_2, x_3, x_4 et x_5

Déterminer le dendrogramme résultant de l'application du « single link algorithm » (lien minimum).

$$\begin{array}{c} \\ x1 \\ x2 \\ x3 \\ x4 \\ x5 \end{array} \begin{pmatrix} 0 & 0.4 & 0.9 & 0.6 & 0.5 \\ 0.4 & 0 & 0.3 & 0.8 & 0.7 \\ 0.9 & 0.3 & 0 & 0.3 & 0.2 \\ 0.6 & 0.8 & 0.3 & 0 & 0.1 \\ 0.5 & 0.7 & 0.2 & 0.1 & 0 \end{pmatrix} \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}$$

Série de TD N°3 (Corrigé)
Clustering 2 (Clustering Hiérarchique)

Exercice1 :

1- Le dendrogramme d'un clustering n'est pas unique ; il dépend de la stratégie de regroupement : lien minimum, maximum ou moyen. En plus, si la distance minimale choisie n'est pas unique, le choix aléatoire diversifie le dendrogramme.

2- On détermine le nombre de clusters à partir d'un dendrogramme, en trouvant le nombre de points d'intersection entre la droite *y=d* et le dendrogramme, *d* étant la distance choisie. Cette méthode est exacte. Le moyen le plus efficace pour avoir un nombre de classes (clusters) proche de la réalité se fait généralement à l'aide de l'avis d'un expert dans le domaine.

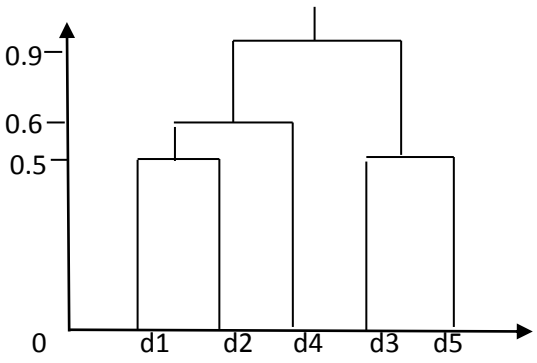
3- Le dendrogramme résultant de l'application du text clustering hiérarchique ascendant de cinq documents d1, d2, d3, d4 et d5, en utilisant le « **lien maximum** ».

	d1	**d2**	**d3**	**d4**	**d5**
d1	0				
d2	**<u>0.5</u>**	0			
d3	0.5	0.7	0		
d4	0.6	0.6	0.6	0	
d5	0.8	0.5	0.5	0.9	0

grouper (d1, d2)

	d1d2	**d3**	**d4**	**d5**
d1d2	0			
d3	0.7	0		
d4	0.6	0.6	0	
d5	0.8	**<u>0.5</u>**	0.9	0

grouper (d3, d5)



Dendrogramme

	d1d2	**d3d5**	**d4**
d1d2	0		
d3d5	0.8	0	
d4	**<u>0.6</u>**	0.9	0

grouper (d1d2, d4)

	d1d2d4	**d3d5**
d1d2d4	0	
d3d5	**<u>0.9</u>**	0

grouper(d1d2d4,d3d5)

Exercice3 :

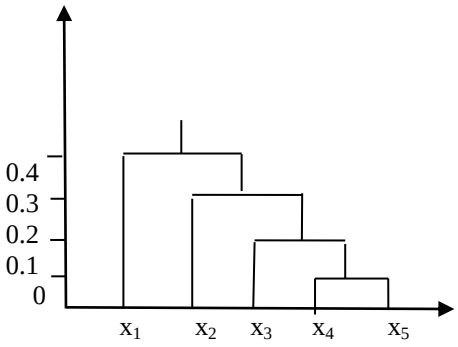
Single Link (Lien mimimum)

	x_1	x_2	x_3	x_4	x_5
x_1	0				
x_2	0.4	0			
x_3	0.9	0.3	0		
x_4	0.6	0.8	0.3	0	
x_5	0.5	0.7	0.2	**0.1**	0

grouper (x_4, x_5)

	x_1	x_2	x_3	x_4x_5
x_1	0			
x_2	0.4	0		
x_3	0.9	0.3	0	
x_4x_5	0.5	0.7	**0.2**	0

grouper(x_3,x_4x_5)



Dendrogramme

	x_1	x_2	$x_3x_4x_5$
x_1	0		
x_2	0.4	0	
$x_3x_4x_5$	0.5	**0.3**	0

grouper (x_2, $x_3x_4x_5$)

	x_1	$x_2x_3x_4x_5$
x_1	0	
$x_2x_3x_4x_5$	**0.4**	0