

## 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).

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