Informatique décisionnelle -2

M. M Lebbah

Installer SPARK

- http://blog.prabeeshk.com/blog/2014/10/31/install-apache-spark-on-ubuntu-14-dot-04/

Matrice des distances

- Pour un nuage d'individus, on peut résumer l'ensemble des distances entre individus au sein d'une matrice des distances que l'on note D.
- Chaque coefficient d_{ij} représente la distance entre l'individu M_i et l'individu M_j
- Par exemple, si l'on choisit comme critère de ressemblance la distance euclidienne, on a $d_{ij} = d_2(M_i, M_j)$.
- Les propriétés de ce type de matrice :

  Une matrice de distances est :
  - Une matrice carré.
  - Une matrice symétrique ($d_{ij} = d_{ji}$).
  - De coefficients positifs ($d_{ij} >= 0$).
  - De coefficients nuls sur la diagonale ($d_{ii} = d(M_i,M_i) = 0$).

105

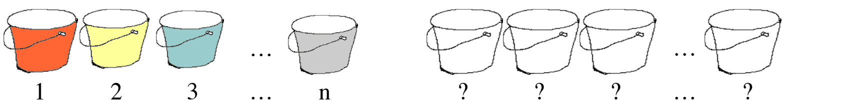
Problématique : clustering

Classification supervisée

Les classes et le nombre des classes sont connus

Classification non supervisée

Les classes et le nombre des classes ne sont pas disponibles



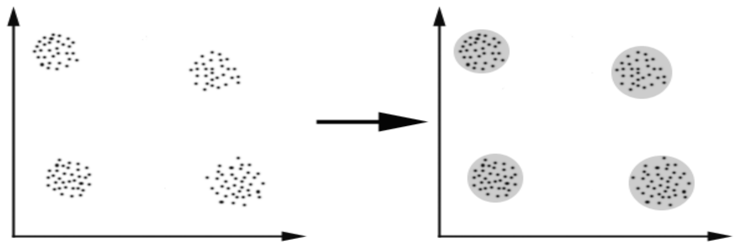
➢ Les difficultés :

- Existence réelle d'une structure
- Choix de similarité
- Choix du nombre de groupes (Combinatoire)

- Validation (absence de labels)
- Nature des données

106

C'est quoi le clustering ?



Trouver K clusters/ groupes/ensemble de données homogènes. (les données appartenant à des clusters différents sont dissimilaires)

construire des classes automatiquement en fonction des exemples disponibles
• L'apprentissage non supervisé est tr ès souvent synonyme de clustering

107

Quelques bonnes raisons de s'intéresser à l'apprentissage non supervisé

➔Constituer des échantillons d'apprentissage étiquetés peut être très couteux

➔Découvertes de la structure et la nature des données à travers l'analyse exploratoire

- Utile pour l'étude des caractéristiques pertinentes
- Prétraitement avant l'application d'une autre technique de fouille de données

108

Approches de Clustering

- Algorithmes de Partitionnement: Construire plusieurs partitions puis les évaluer selon certains critères
- Algorithmes hiérarchiques: Créer une décomposition hiérarchique des objets selon certains critères
- Algorithmes basés sur la densité: basés sur des notions de connectivité et de densité
- À Base de modèle de mélange

Notion de proximité

➔Mesure de dissimilarité : plus la mesure est faible plus les points sont similaires (~ distance)

➔Mesure de similarité : plus la mesure est grande, plus les points sont similaires

K-means

Algorithmes à partionnement

- Construire une partition à ***k*** clusters d'une base ***A*** de ***n*** objets
- Les *k clusters doivent* optimiser le critère choisi
  - *k-means* (MacQueen'67): Chaque cluster est représenté par son centre
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Chaque cluster est représenté par un de ses objets

Quantification vectorielle

D : espace des données $A \subset D \subset \Re^n$

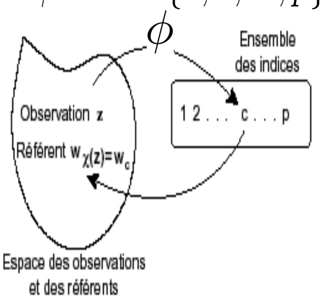
A : ensemble d'apprentissage $\mathcal{A} = \{\mathbf{x}_i, i = 1 : N\}$

$\phi : \mathcal{D} \rightarrow \{1, 2, ..., p\}$

Réduire l'information de D

- En la <u>résumant</u> par un ensemble de **p référents**

$$\mathcal{W} = \{\mathbf{w}_c, c = 1 : p\}$$

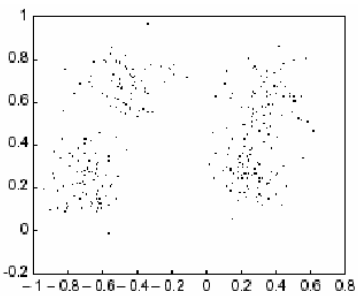


- En <u>réalisant</u> une partition de D en **p** sous-ensembles par l 'intermédiaire d 'une <u>**fonction d'affectation**</u> ϕ

$$P_c = \{\mathbf{x} \in \mathcal{D}, \phi(\mathbf{x}) = c\}$$

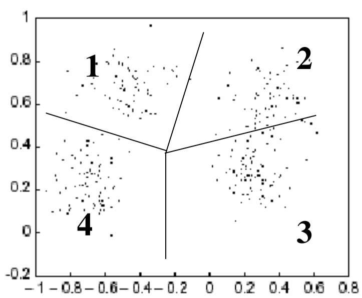
113

Quantification



$\mathcal{A} = \{\mathbf{x}_i, i = 1 : N\}$

$A \subset D \subset \Re^n$



Partition

$P = \{P_1, P_2, P_3, P_4\}$

$\phi : \mathcal{D} \rightarrow \{1, 2, 3, 4\}$

114

K-means
Version nuées dynamiques

(Diday 1972, 1974)

- Chaque cluster est associé à un centre (prototype)
- Chaque donnée est affectée au centre le plus proche
- Nombre de clusters doit être fixé
- L'algorithme est simple

115

Méthode des k-moyennes

- Minimiser la somme des inerties locales par rapport à χ et **W**

$$I(\mathcal{W}, \phi) = \sum_{\mathbf{x}_i} ||\mathbf{x}_i - \mathbf{w}_{\phi(\mathbf{x}_i)}||^2 = \sum_{c} \underbrace{\sum_{\mathbf{x}_i \in P_c} ||\mathbf{x}_i - \mathbf{w}_c||^2}_{I_c}$$

- L'inertie I_c représente l'erreur de quantification obtenue si l'on remplace chaque observations de P_c par son référent $\mathbf{w}_c$

116

•Minimisation itérative qui fixe alternativement la partition (**c**)puis minimise l'inertie

Phase d'affectation:

Pour un ensemble **W** de référents fixe, la minimisation de **I** par rapport à **Φ** s'obtient en affectant chaque observation **x** au référent $\mathbf{w}_c$ selon la nouvelle fonction d'affectation **Φ**

$$\phi(\mathbf{x}) = \arg\min_r ||\mathbf{x} - \mathbf{w}_r||^2$$

Phase de minimisation:

La partition **Φ** est fixée. La fonction $I(\mathcal{W}, \phi)$ est quadratique et convexe par rapport à **W.** Le minimum global est atteint pour

$$\frac{\partial I}{\partial W} = \left[\frac{\partial I}{\partial \mathbf{w}_1}, \frac{\partial I}{\partial \mathbf{w}_2}, \ldots, \frac{\partial I}{\partial \mathbf{w}_p}\right]^p = 0 \qquad \mathbf{w}_c = \frac{\sum_{\mathbf{x}_i \in P_c} \mathbf{x}_i}{|P_c|}$$

117

L'algorithme

L'algorithme de base

1. Sélectionner K centres
2. **Repeat**
3. Affecter chaque données au centre centre le plus proche
4. Mise à jour des centres
5. **Until** non changement des centres

118

Initialisation

◆aléatoirement dans l'intervalle de définition des x_i

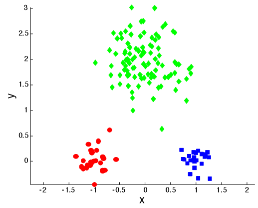
◆aléatoirement dans l'ensemble des x_i

Des initialisations différentes peuvent mener à des clusters différents (problèmes de minima locaux)

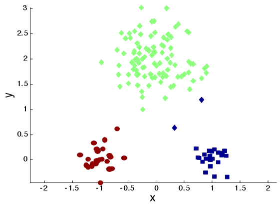
◆méthode **générale pour obtenir des clusters "stables" =** formes fortes, on répète l'algorithme k fois

119

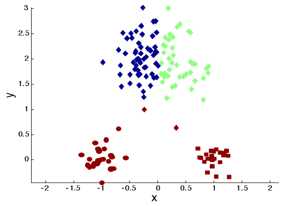
K-Means



Données originales



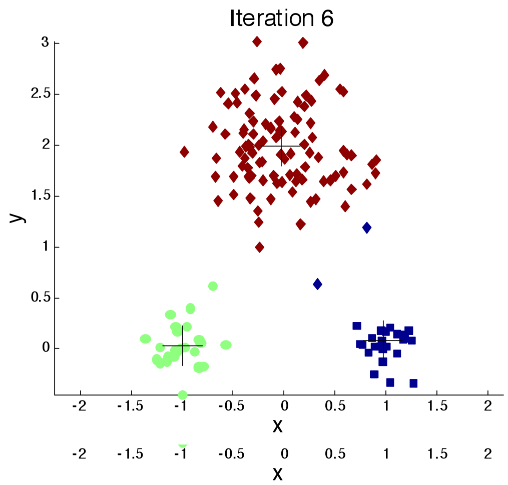
Solution optimale



Autre solution

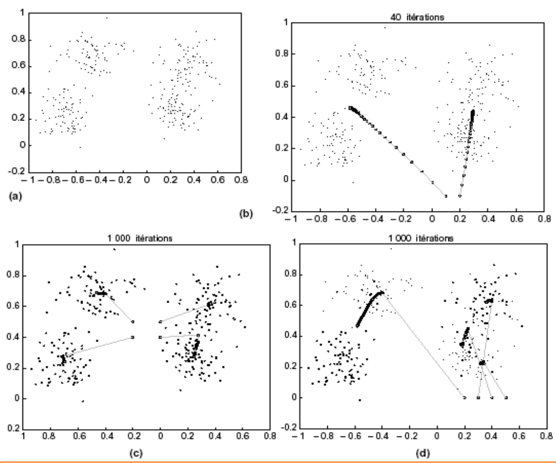
120

Importance de l'initialisation



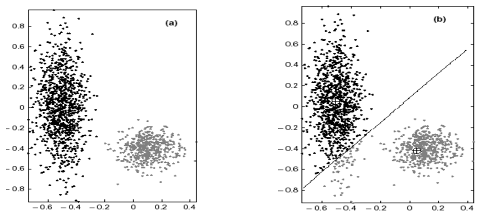
121

Sensibilité aux conditions initiales



12
2

Comportement de l'algorithme des k-moyennes en fonction des densités sous-jacente



(a) (b) (c)

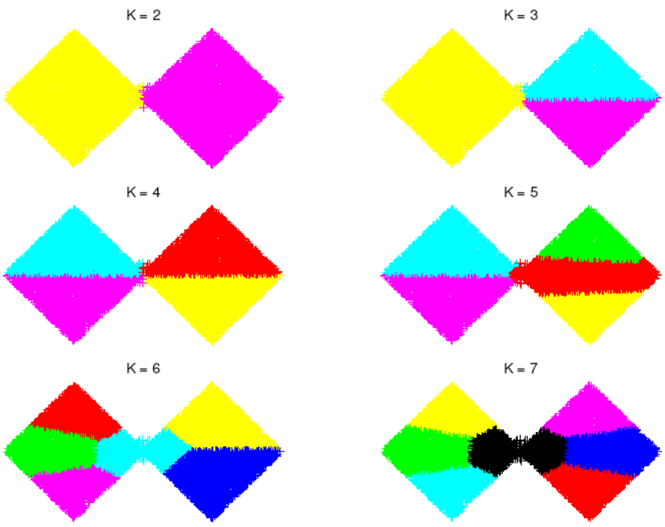
(a) Données simulées selon deux distributions gaussiennes de matrice de variance-covariance différentes

(b) référents et partition obtenue à la convergence avec deux référents

(c) avec cinq référents;

123

K-Means : exemple



124

Compression d'image: Quantification vectorielle

[redacted]

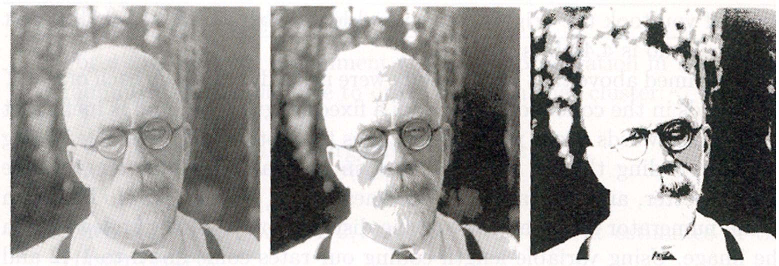


Image à gauche 1024*1024 pixels
256 niveaux de gris
8bits par pixel-
mémoire 1 mégabit

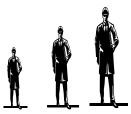
Image au centre 512*512 blocs de 2*2 pixels quantifiés en 200 référents
mémoire 0,239 mégabit

Image à droite 512*512 blocs de 2*2 pixels quantifiés en 4 référents
mémoire 0,063 mégabit

125

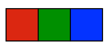
Données qualitatives

Qualitatives / Catégorielles

Taille 

Sexe: 

Diabète: Oui/NON

Couleur : 

Questions:
•Comment partitionner ces données ?
•Quelle distance utilisée ?
•Avoir des prototypes du même type que les données

126

Variables qualitatives et codage

Taille: **Petit**, Moyen, Grand		1	0	0
Petit, **Moyen**, Grand	→ Ordinale	1	1	0
Petit, Moyen, **Grand**		1	1	1

Couleur : **rouge**, vert, bleu		1	0	0
rouge, **vert**, bleu	→ Disjonctif	0	1	0
rouge, vert, **bleu**		0	0	1

127

Données binaires

- une matrice de contingence

		Object j: 1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

$$H(i,j) = b + c$$

$$sim_{Jaccard}(i,j) = \frac{a}{a+b+c}$$

128

Distance de Hamming
w, x ∈ {0,1}^{d}

$$I(\mathcal{W},\phi)=\sum_{i=1}^{N}|\mathbf{x}_i-\mathbf{w}_{\phi(\mathbf{x}_i)}|=\sum_{i=1}^{N}\sum_{j=1}^{n}|x_i^j-w^j_{\phi(\mathbf{x}_i)}|$$

$$\Downarrow$$

$$I(\mathcal{W},\phi)=\sum_{j=1}^{n}\left(\sum_{i=1}^{N}(1-x_i^j)w^j_{\phi(\mathbf{x}_i)}+\sum_{i=1}^{N}x_i^j(1-w^j_{\phi(\mathbf{x}_i)})\right)$$

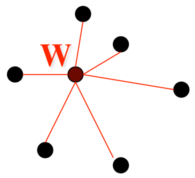
$$\Downarrow$$

$$I(\mathcal{W},\phi)=\sum_{j=1}^{n}\left(w^j_{\phi(\mathbf{x}_i)}\Gamma_0^j+(1-w^j_{\phi(\mathbf{x}_i)})\Gamma_1^j\right)$$

129

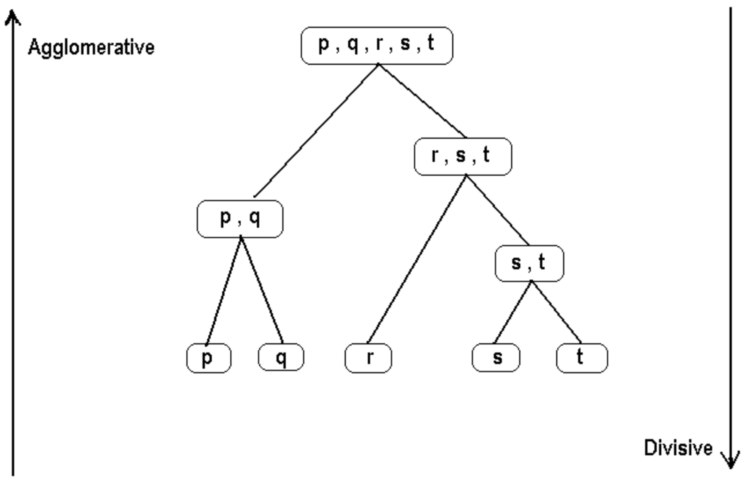
Centre médian

1	1	1	1	1	0	0	1	1	1	0	0
1	1	0	1	1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	1	1	1	1	1
1	1	1	1	1	1	0	1	1	0	0	0
1	1	1	1	1	1	0	1	1	1	1	0
1	0	0	1	1	1	0	1	1	1	0	0
1	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**0**



130

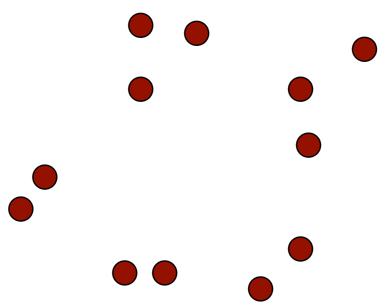
Classification hiérarchique

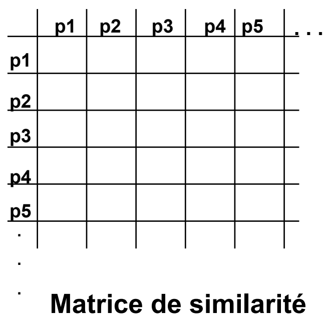


131

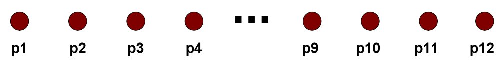
Situation initiale

Un point ==cluster



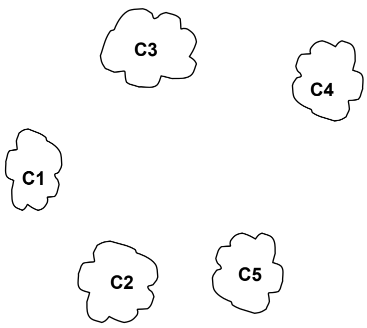


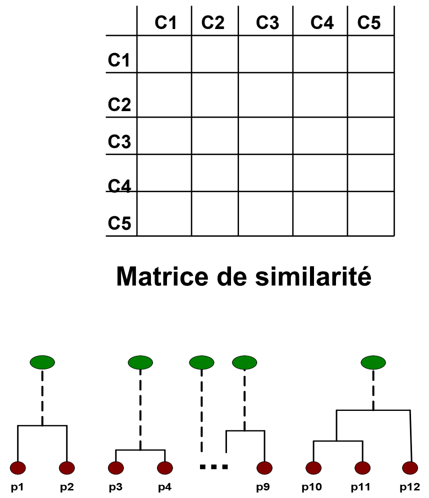
Matrice de similarité



Situation intermédiaire

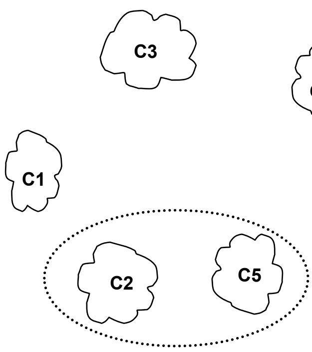
Après quelques itérations

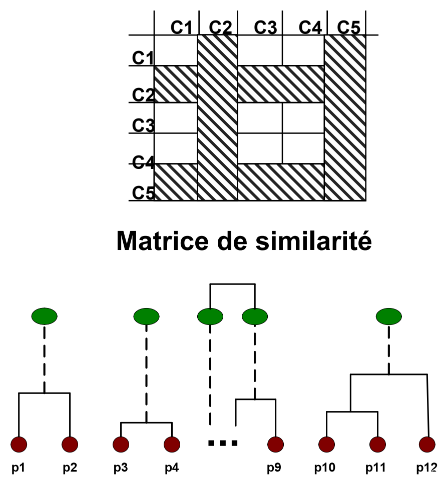




Situation intermédiaire

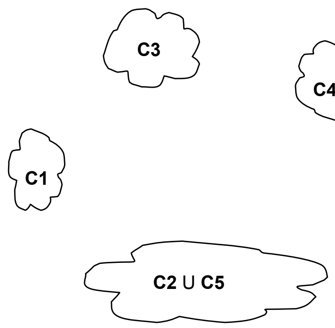
Fusionner C2 et C5 puis mise à jour de la matrice de similarité.





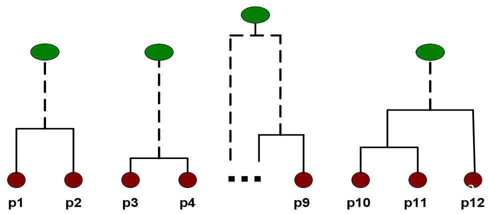
Après fusion

La question: "comment mettre à jour la matrice de similarité?"

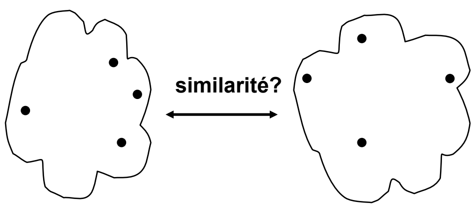


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Matrice de similarité



Similarité inter-classe

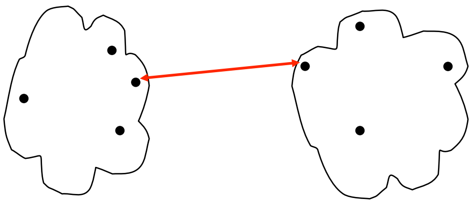


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matrice de similarité

- MIN
- MAX
- moyen
- Distance entre centres
- Ward

Similarité inter-classe

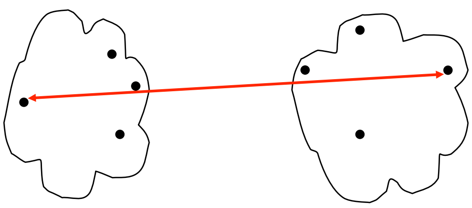


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matrice de similarité

- MIN
- MAX
- moyen
- Distance entre centres
- Ward

Similarité inter-classe

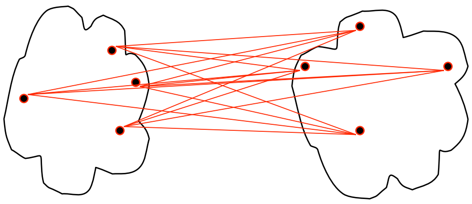


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matrice de similarité

- MIN
- MAX
- Moyen
- Distance entre centres
- Ward

Similarité inter-classe

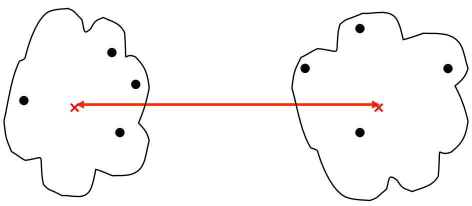


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
. . .						

Matrice de similarité

- MIN
- MAX
- moyen
- Distance entre centres
- Ward

Similarité inter-classe



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
. . .						

Matrice de similarité

- MIN
- MAX
- Average
- Distance entre centres
- Ward

Indice de Ward

- Basé sur la perte d'inertie
- Moins sensible aux outliers
- A chaque itération, on agrè   ge de mani   ère à avoir un gain minimum d'inertie intra-classe : perte d'inertie interclasse due à cette agrégation

$$\frac{n_A n_B}{n_A + n_B} ||g_A - g_B||^2$$

Algorithme agglomératif

L'algorithme de base

1. Calculer la matrice de similarité
2. Affecter chaque donnée à un cluster
3. **Repeat**
4. fusionner les deux clusters les plus proches
5. Mise à jour de la matrice de similarité
6. **Until** trouver un seul cluster

142