بسم الله الرحمن الرحيم

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de La Recherche Scientifique
Université Saad-Dahlab (Blida 1)
Faculté des Sciences
Département d'informatique

Projet de fin d'études en vue d'obtention du Diplôme de Licence en informatique

Fait par :
- Mr. KERKAR Mehdi Nacer
- Mr. KADI Abdelhakim

Davant le jury :
- Mme. Farah
- Mr. Riali

05 - 06 - 2017

Plan

- Introduction
- Problématique
- Classification
- Clustering
- Prétraitement
- Algorithme K-means
- Application
- Résultat
- Conclusion

INTRODUCTION

- La classification automatique est le processus qui permet d'analyser et d'organiser un ensemble de données, selon leurs caractéristiques, dans des classes de similarité. Elle se base principalement sur des représentations classiques de données dont les limites de traitement sont connues et, qui dans la plupart du temps, demande un temps de calcul énorme.
- C'est au début des années soixante et avec l' évènement de l'informatique que les méthodes de classification ont connu de nouveaux développements méthodologiques qui ont permis l'apparition d'algorithmes d'analyse et de classification automatique de données.

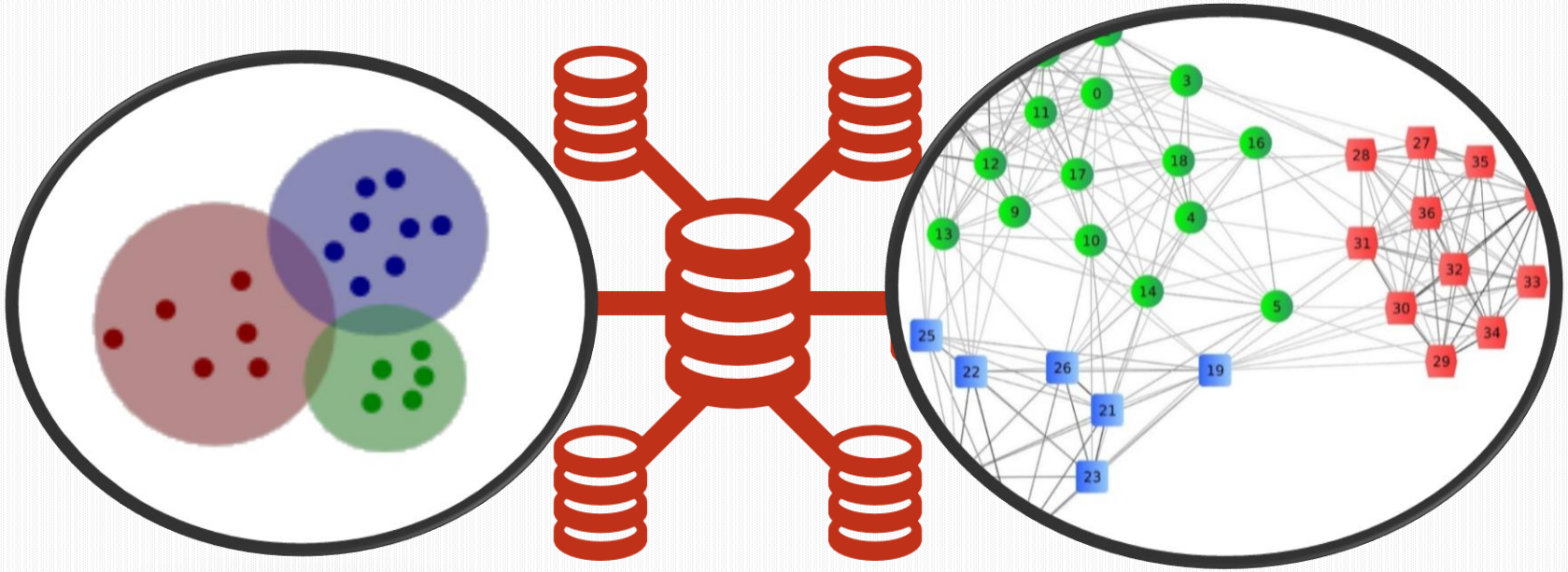
PROBLÉMATIQUE

- La quantité énorme des documents et l'explosion des ressources textuelles non structurées ont suscité dès les années 80, beaucoup d'intérêts pour les différentes techniques de traitement automatique de documents.
- Un besoin urgent de ces techniques ont été suscitées dont les résultats pertinents qu'on peut dégager de ces derniers.
- Donc nous avons proposé de faciliter ce travail de catégorisation avec une application.

Classification

Supervisé

Non supervisé



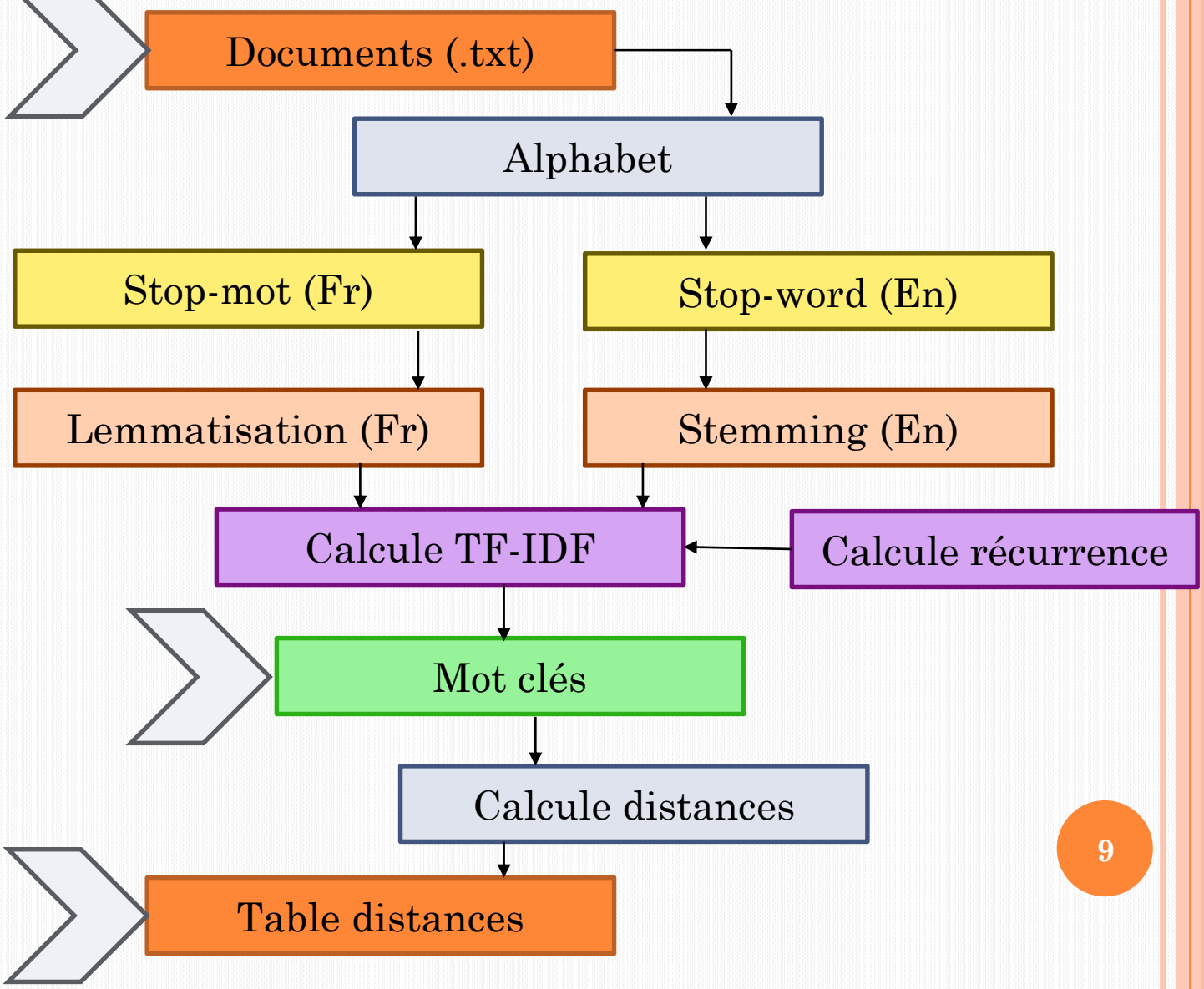
Classification

Supervisé	Non supervisé
Classes connues à priori	Classes fondées sur la structure des objets
Sémantique associée	Sémantique difficile à déterminer
Ensemble d'apprentissage	Pas besoin de base d'apprentissage
	partitionner en sous-ensembles
	Nombre de classes est fixé par l' utilisateur
Les machines à support de vecteurs.	ART

Clustering

- Le **clustering** est un outil statistique utilisé pour obtenir facilement et rapidement une analyse de données.
- Le **clustering** consiste en effet à séparer des données, en constituant différents groupes homogènes.
- Toutes les données placées dans un même groupe doivent alors partager des caractéristiques communes. qui sont difficiles à identifier à l'oeil nu.

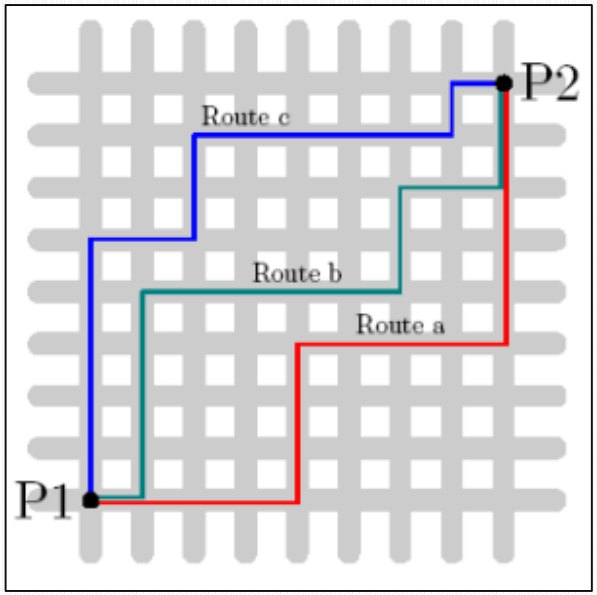
PRÉTRAITEMENT (2)



Distances

Distance de Manhattan :

- appelée aussi taxi-distance.



- $D(A,B) = | x_b - x_a | + | y_b - y_a |$

Distances

Distance Cosinus :

- calcule la similarité entre deux vecteurs à *n* dimensions en déterminant le cosinus de l'angle entre eux.

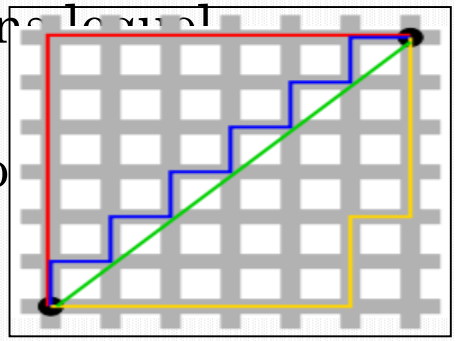
- Soit deux vecteurs A et B , l'angle cos θ s'obtient par le produit scalaire et la norme des vecteurs :

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Distances

Distance Euclidienne :

- On peut imaginer les variables indépendantes comme définissant un espace multidimensionnel dans lequel chaque observation peut être tracée.
- La *distance Euclidienne* est une distance géométrique dans un espace multidimensionnel.



- Entre deux points A et B, de coordonnées respectives (x_a, y_a) et (x_b, y_b), la distance euclidienne est définie par :

$$D(A,B) = \sqrt{((x_b - x_a)^2 + (y_b - y_a)^2}$$

Tf-idf

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

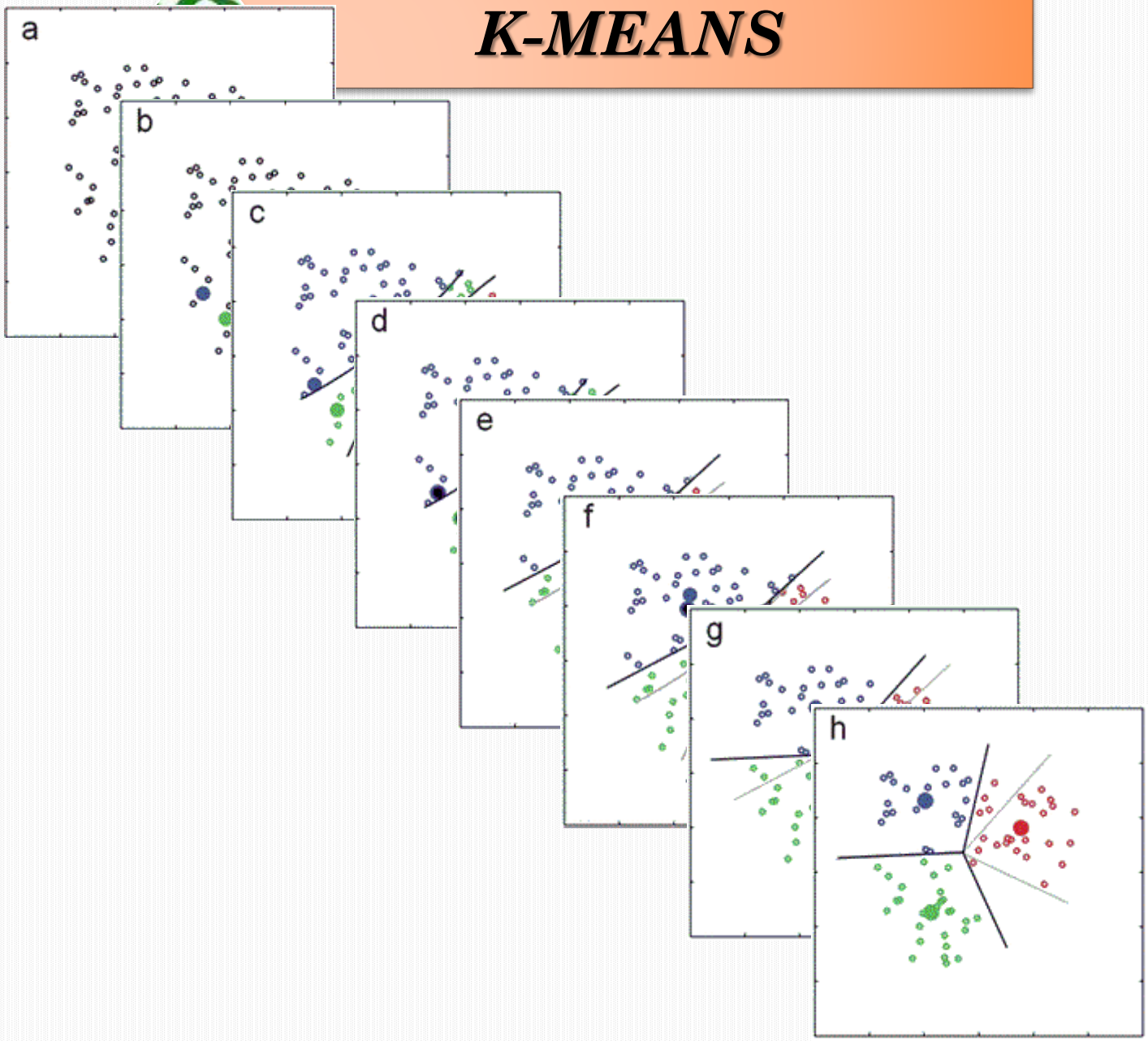
Term x avec document y

$tf_{x,y}$ = frequence de x dans y

df_x = nombre de documents contenon x

N = nombre total de documents

K-MEANS



Définition K-means

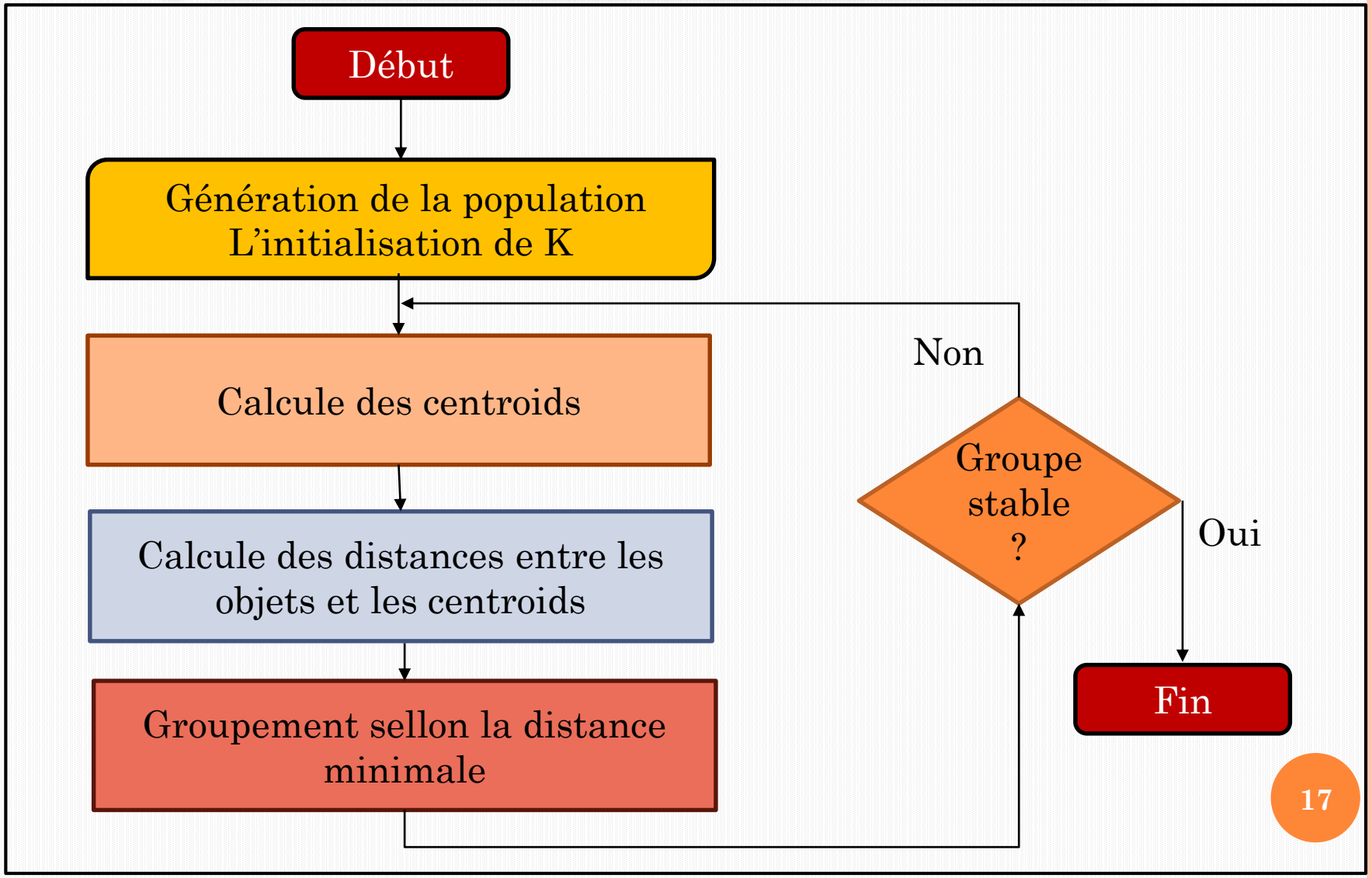
- L'algorithme k-means mis au point par McQueen en 1967, un des plus simples algorithmes d'apprentissage non supervisé , appelée algorithme des centres mobiles, il attribue chaque point dans un cluster dont le centre (centroïde) est le plus proche.

- Le centre est la moyenne de tous les points dans le cluster , ses coordonnées sont la moyenne arithmétique pour chaque dimension séparément de tous les Points dans le cluster c'est à dire chaque cluster est représentée par son centre de gravité.

5.2. Exemples d'applications

Marketing	Bases de données d'achats	Segmentation du marché en découvrant des groupes de clients
Environnement	BD d'observations de la terre	Identification des zones terrestres similaires
Assurance	BD de déclarations	Identification de groupes d'assurés distincts

L'algorithme K-means



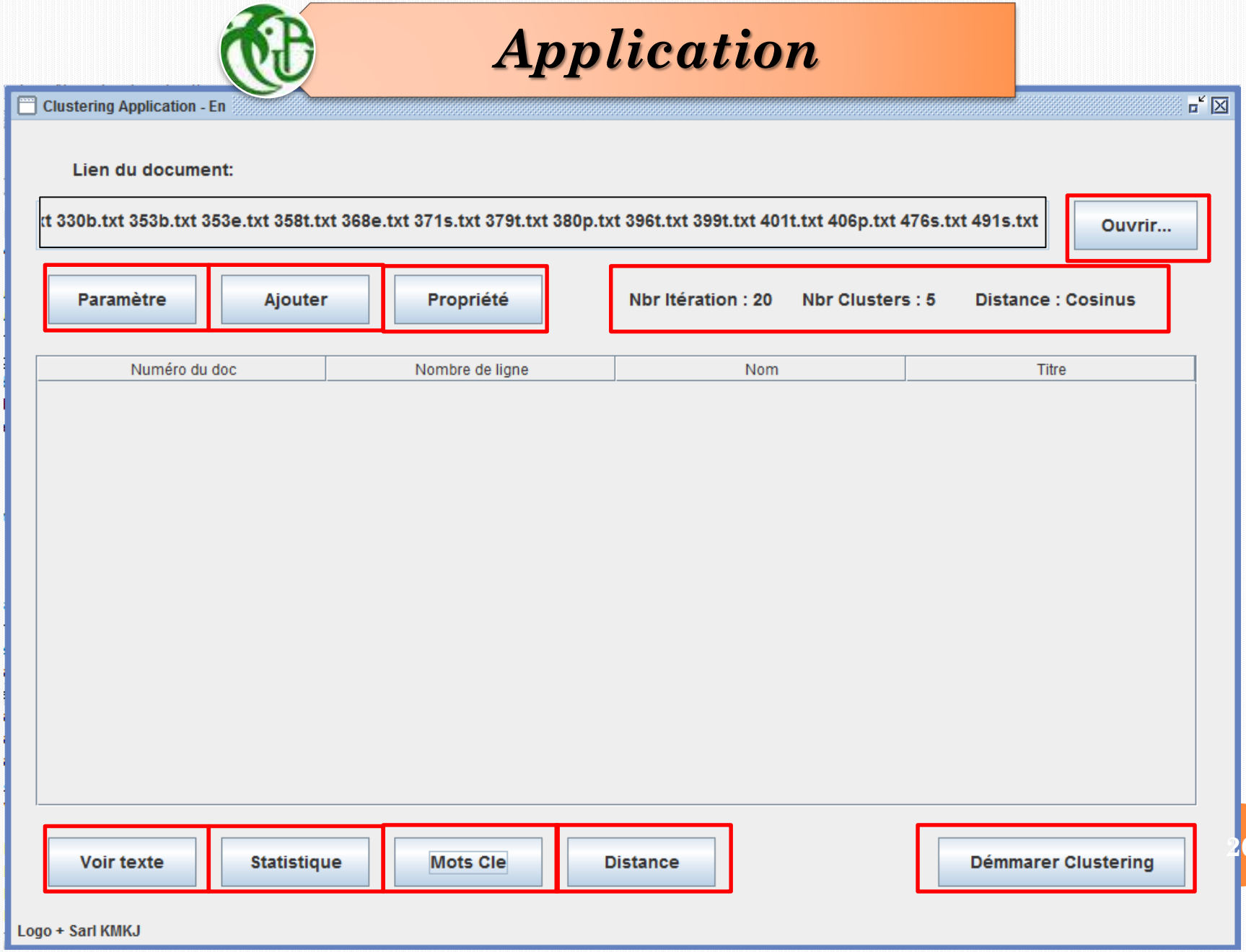
Avantages du K-means

- Sa complexité linière.
- Sa simplicité.
- Sa convergence rapide.
- Son adaptation à de larges bases de données.
- L'ordre d'entrée des objets n'a aucune influence sur les résultats de cette méthode.
- Applicable à des données de grandes tailles.

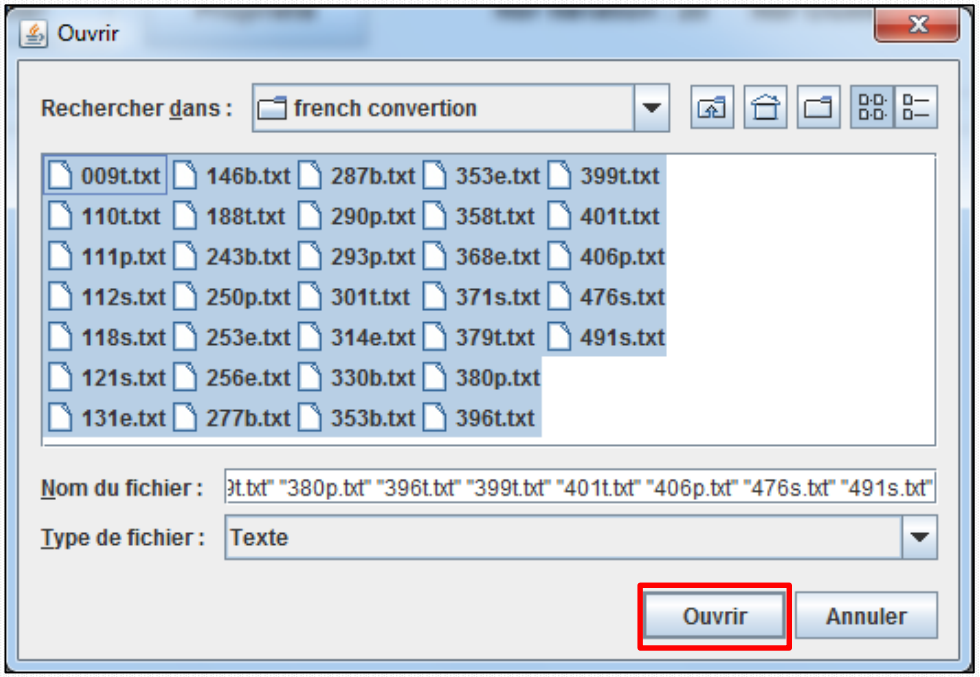
Inconvénients du K-means

- Le nombre d' objets k est fixé au début, ce qui influence les résultats.
- Sa sensibilité aux éléments marginaux.
- Sa mauvaise gestion pour les clusters mal isolés.
- Les clusters sont construits par rapports à des objets inexistants (les milieux) .
- Il converge souvent vers un optimum local .

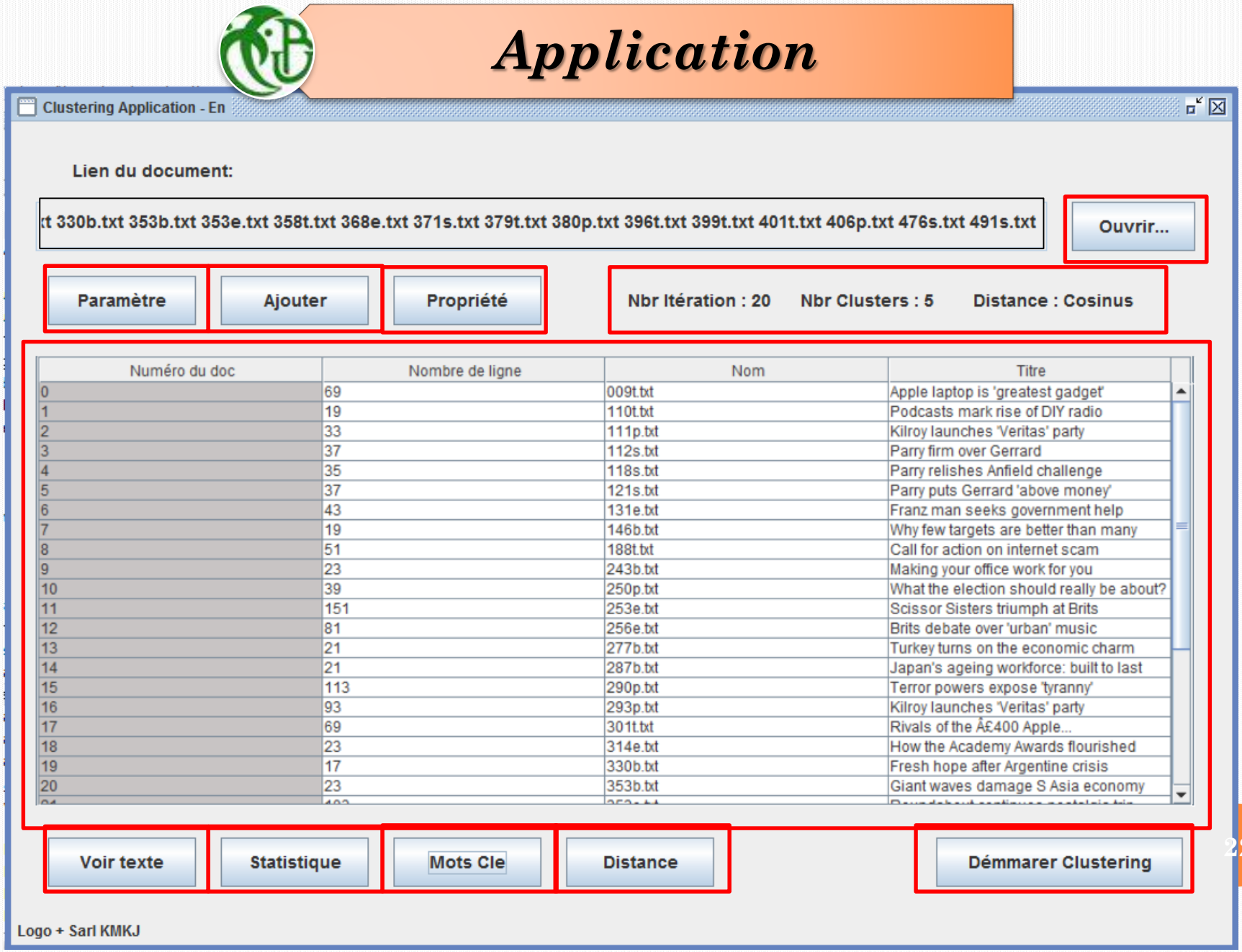
Application



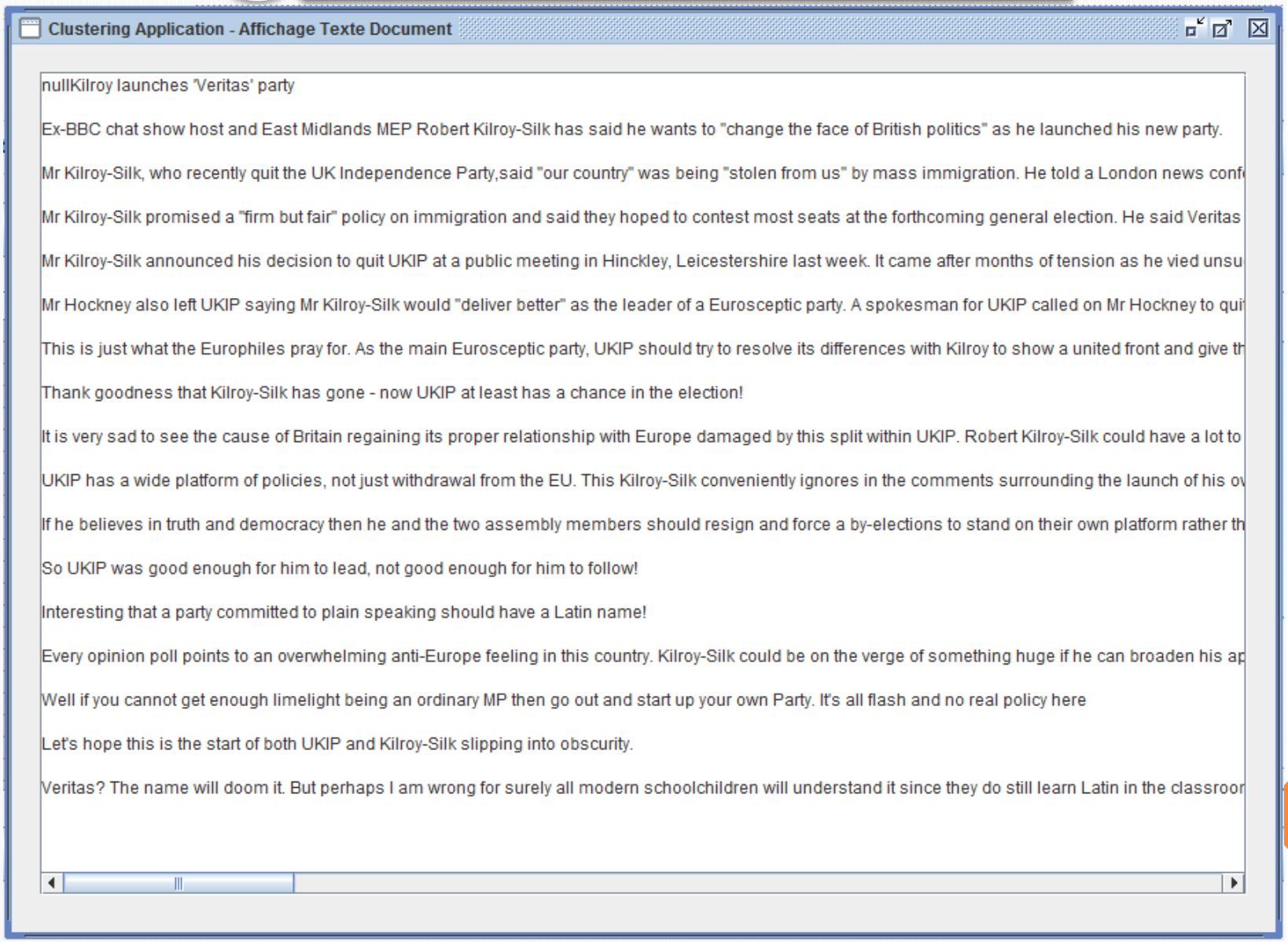
Application



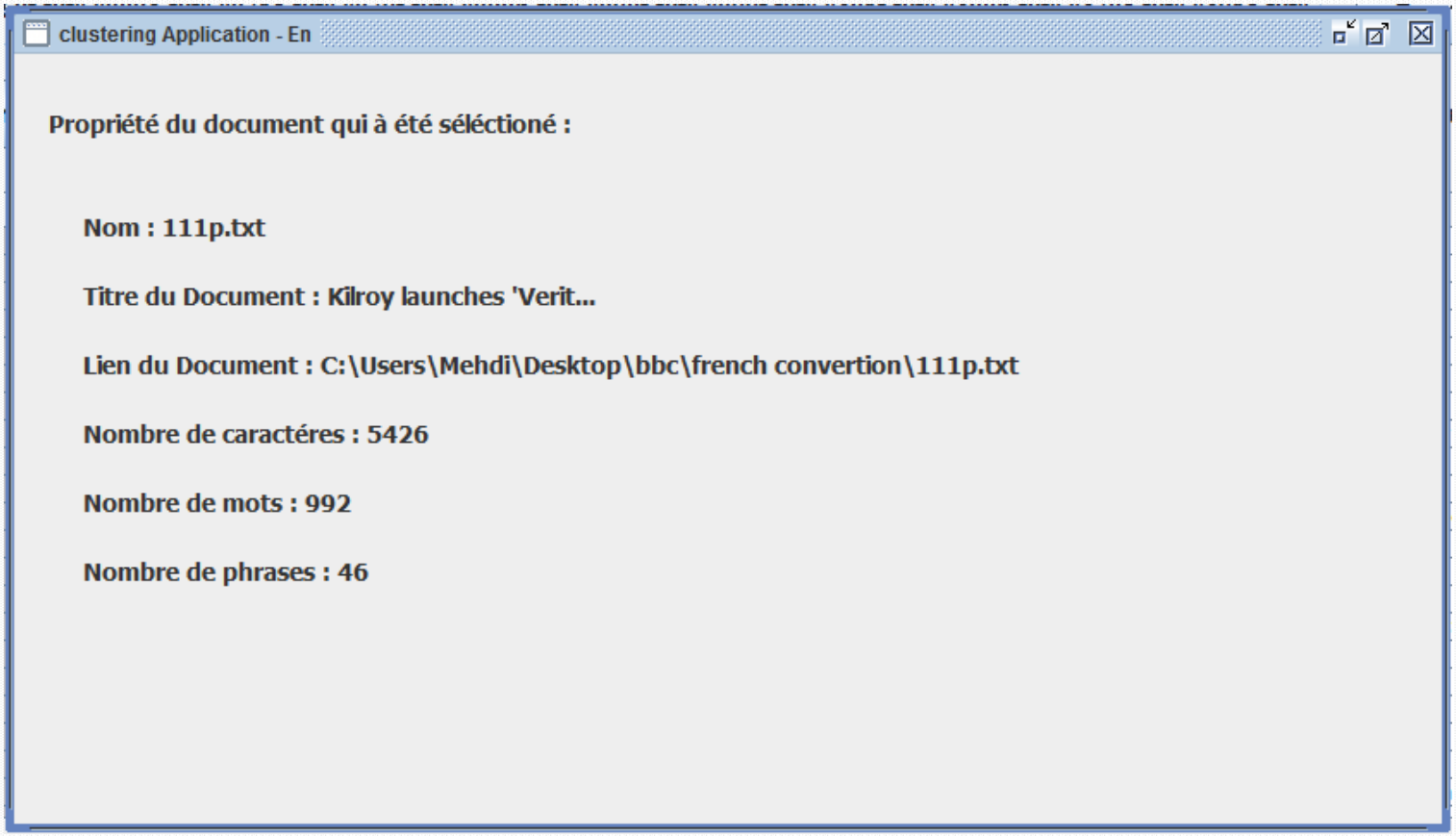
Application



Application



Application

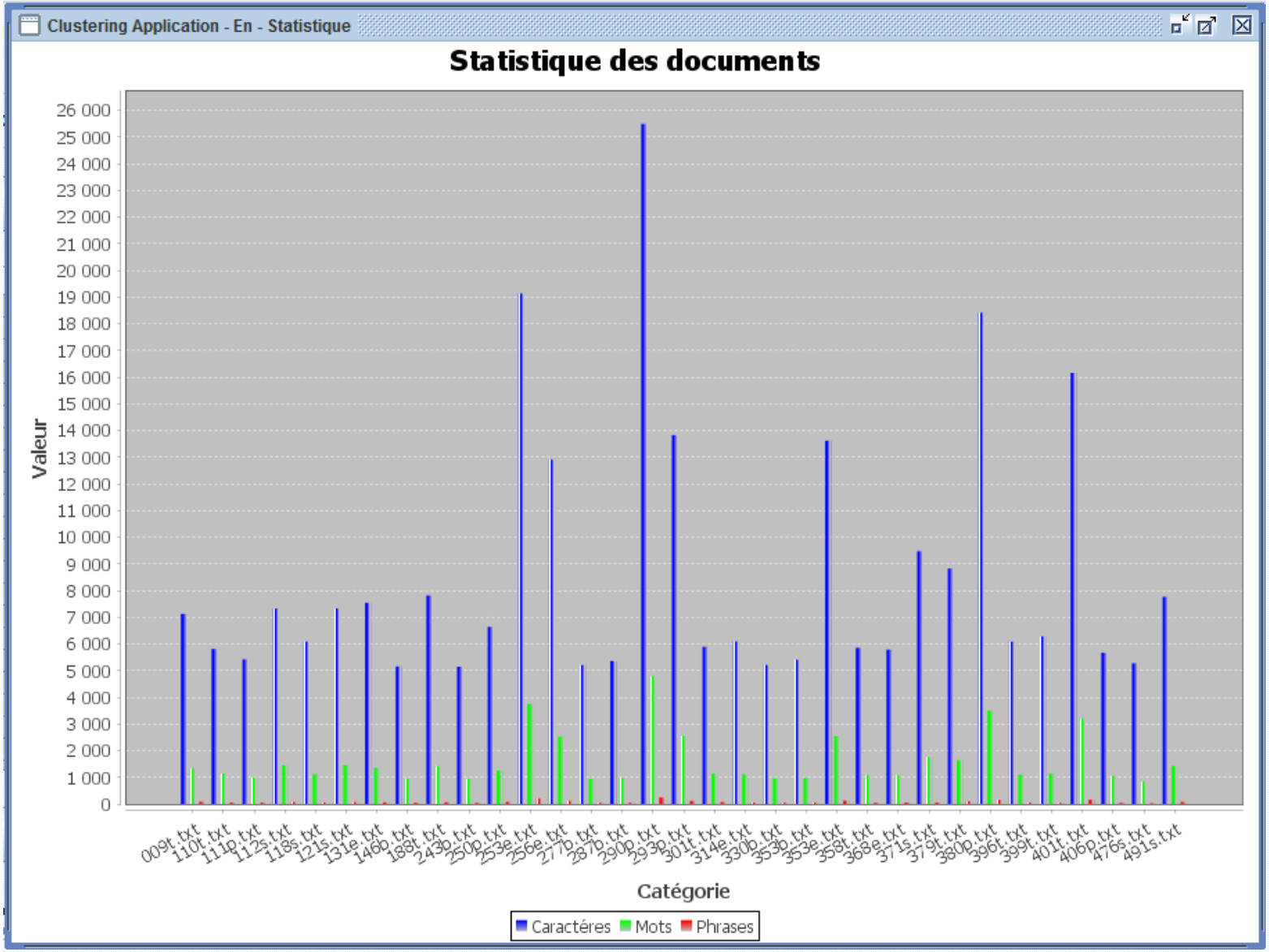


Application

clustering Application - Affichage Mots Cle

Mots cle	Mots Sans Stem	Nombre de répétition	Nom document
make	making,makes,make	2.0	110t.txt
malaysia	malaysia	1.0	110t.txt
mark	mark	1.0	110t.txt
mcintyr	mcintyre	2.0	110t.txt
meant	meant	1.0	110t.txt
microphon	microphone	1.0	110t.txt
mile	mile	1.0	110t.txt
minut	minutes,minute	1.0	110t.txt
mom	mom	2.0	110t.txt
mon	morning	2.0	110t.txt
month	months,month	1.0	110t.txt
more	more	5.0	110t.txt
mother-in-law	mother-in-law	1.0	110t.txt
mp	mps,mp	3.0	110t.txt
mr	mr	7.0	110t.txt
mtv	mtv	3.0	110t.txt
much	much	1.0	110t.txt
music	music,musical,musically	6.0	110t.txt
name	named,names,naming,namely,name	1.0	110t.txt
need	needs,need,needed	3.0	110t.txt
net	nets,net	1.0	110t.txt
netherland	netherlands	1.0	110t.txt
new	news,new	5.0	110t.txt
next	next	2.0	110t.txt
nice	nice	1.0	110t.txt
now	notwithstanding,now	2.0	110t.txt
on	owned,one	6.0	110t.txt
onto	onto	2.0	110t.txt
open	opener,opens,opened,opening,open	2.0	110t.txt
other	others	2.0	110t.txt
out	outing,out	1.0	110t.txt

Application



Résultats (1)

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
379t.txt	243b.txt	293p.txt	491s.txt	371s.txt
253e.txt	110t.txt	380p.txt	131e.txt	277b.txt
476s.txt	188t.txt	111p.txt	368e.txt	406p.txt
009t.txt	358t.txt	250p.txt	256e.txt	118s.txt
401t.txt		287b.txt	353e.txt	396t.txt
301t.txt		290p.txt	112s.txt	330b.txt
399t.txt			121s.txt	353b.txt
			314e.txt	146b.txt

Tableau 1 : Résultat clustering "0,3"

tf-idf (mot clé) < 0.3 :

Cluster 0 : (t, 5) ; Cluster 1 : (t, 3) ; Cluster 2 : (p, 5) ; Cluster 3 : (e, 5) ; Cluster 4 : (b, 4)

purity = (1 / 33) * (5 + 3 + 5 + 5 + 4) = 0.66

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
379t.txt	243b.txt	293p.txt	491s.txt	277b.txt
256e.txt	253e.txt	380p.txt	131e.txt	406p.txt
476s.txt	110t.txt	111p.txt	368e.txt	118s.txt
396t.txt	188t.txt	250p.txt	371s.txt	330b.txt
009t.txt	358t.txt	287b.txt	353e.txt	353b.txt
401t.txt		290p.txt	112s.txt	146b.txt
301t.txt			121s.txt	
399t.txt			314e.txt	

Tableau 2 : RESULTAT CLUSTERING "0,4"

tf-idf (mot clé) < 0.4 :

Cluster 0 : (t, 6) ; Cluster 1 : (t, 3) ; Cluster 2 : (p, 5) ; Cluster 3 : (e, 5) ; Cluster 4 : (b, 4)

purity = (1 / 33) * (6 + 3 + 5 + 5 + 4) = 0.69

Résultats (1)

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
379t.txt	110t.txt	293p.txt	491s.txt	243b.txt
256e.txt	188t.txt	371s.txt	131e.txt	368e.txt
253e.txt	358t.txt	277b.txt	112s.txt	406p.txt
353e.txt		380p.txt	121s.txt	118s.txt
476s.txt		111p.txt	290p.txt	396t.txt
009t.txt		250p.txt	314e.txt	330b.txt
401t.txt		287b.txt		353b.txt
301t.txt				146b.txt
399t.txt				

Tableau 3: RÉSULTAT CLUSTERING "0,5"

tf-idf (mot clé) < 0.5 :

Cluster 0 : (t, 5) ; Cluster 1 : (t, 3) ; Cluster 2 : (p, 4) ; Cluster 3 : (e, 3) ; Cluster 4 : (b, 4)

purity = (1 / 33) * (5 + 3 + 4 + 3 + 4) = 0.57

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
379t.txt	256e.txt	293p.txt	491s.txt	243b.txt
368e.txt	353e.txt	380p.txt	131e.txt	118s.txt
371s.txt	110t.txt	406p.txt	277b.txt	396t.txt
253e.txt	188t.txt	111p.txt	112s.txt	330b.txt
476s.txt	358t.txt	353b.txt	121s.txt	301t.txt
009t.txt	401t.txt	250p.txt	290p.txt	
	399t.txt	287b.txt	314e.txt	
		146b.txt		

Tableau 4 : RÉSULTAT CLUSTERING "0,6"

tf-idf (mot clé) < 0.6 :

Cluster 0 : (t, 2) ; Cluster 1 : (t, 5) ; Cluster 2 : (p, 5) ; Cluster 3 : (s, 3) ; Cluster 4 : (b, 2)

purity = (1 / 33) * (2 + 5 + 5 + 3 + 2) = 0.51

Résultats (2)

- 0/8 le résultat et super a 70% pourcent lorsque tf-idf = 0,4.

distName	Centroid	NbrCluster	tf-idf	resultat	
cos	first	5	0,4	0,63	
cos	last	5	0,4	0,63	
cos	rand	5	0,4	0,63	
cos	rand	5	0,4	0,54	
cos	rand	5	0,4	0,51	
cos	rand	5	0,4	0,63	
cos	rand	5	0,4	0,6	
cos	rand	5	0,4	0,57	
				0,58	Moyen

- 2/8 le résultat et super a 70% pourcent lorsque tf-idf = 0.8.

distName	Centroid	NbrCluster	tf-idf	resultat	
cos	first	5	0,8	0,63	
cos	last	5	0,8	0,63	
cos	rand	5	0,8	0,75	
cos	rand	5	0,8	0,72	
cos	rand	5	0,8	0,51	
cos	rand	5	0,8	0,63	
cos	rand	5	0,8	0,57	
cos	rand	5	0,8	0,66	
				0,64	Moyen

- 1/8 le résultat et super a 70% pourcent lorsque tf-idf = 1,2.

distName	Centroid	NbrCluster	tf-idf	resultat	
cos	first	5	1,2	0,63	
cos	last	5	1,2	0,63	
cos	rand	5	1,2	0,66	
cos	rand	5	1,2	0,54	
cos	rand	5	1,2	0,54	
cos	rand	5	1,2	0,78	
cos	rand	5	1,2	0,54	
cos	rand	5	1,2	0,63	
				0,615	Moyen

Résultats (2)

- REMARQUE : nous avons écarté le problème du nombre de clusters de l'équation, pour tester la qualité du clustering avec un cluster équilibré sur notre datatest.

distName	Centroid	NbrCluster	tf-idf	resultat
Euclidienne	first/last	5	0,8	0,39
Manhatan	first/last	5	0,8	0.36

- La valeur l'optimal est entre les deux valeur de tf-idf 0.8 et 1.2.
- Lorsque la dimension testé augmente, la précision de distance euclidienne et Manhattan diminue.
- Nous avons choisie la meilleur valeur de tf-idf = 0,8, avec les premier et dernier centroid de départ.
- Nous remarquons dans les tableaux vue par la distance de cos, que toujours lorsqu'on commence avec les firsts et lasts, centroides; on a toujours la même valeur pour n'importe quel tf-idf.
- c'est un certain équilibre
- Pour cela nous utilisons c'est valeurs, sans voir la différence; donc il nous faut qu'un jeux de test pour savoir
- la différence entre les 3 distances et les classée.

CONCLUSION

- Dans plusieurs domaines des sciences sociales, nous sommes amenés à constituer des groupes homogènes en leur sein et qui diffèrent suffisamment l'un de l'autre. C'est l'objet des méthodes de classification dont fait partie la méthode des k-means, cet algorithme est une version améliorée et randomisée de la méthode des nuées dynamiques.

- Il est actuellement l'un des plus utilisés et des plus efficaces en analyse des données. Il est utile de noter que l'algorithme k-means est très performant en termes de temps d'exécution, mais il souffre du problème de dépendance des résultats aux choix effectués lors de l'initialisation.

- On peut élargir notre travail, en essayant de comparer nos résultats avec d'autres versions de K-means, travailler sur d'autres algorithmes de classification non supervisé, et même le supervisé.Travailler sur de meilleures distances, un prétraitement accentué et un affinement des interfaces.