UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE
APPLIQUÉES

PAR
STEVE DESCÔTEAUX

LES RÈGLES D'ASSOCIATION MAXIMALE AU SERVICE DE L'INTERPRÉTATION DES RÉSULTATS DE LA CLASSIFICATION

JUILLET 2014

SOMMAIRE

Le forage de données textuelles (Text Mining) est devenu un enjeu majeur depuis l'invention du WEB. Plusieurs méthodes de forage de données dont la classification et les règles d'association ont vu le jour afin de soutenir les utilisateurs dans leur recherche d'information.

La force de la classification réside dans sa capacité à répartir les objets textuels (mots, phrases, paragraphes, livres, etc.) en classes distinctes dont les éléments ont les mêmes propriétés. Malheureusement, les classes formées prennent souvent la forme de listes beaucoup trop volumineuses et souvent incompréhensibles pour l'utilisateur. De plus, plusieurs aspects propres aux langues viennent ajouter des difficultés supplémentaires à la tâche.

Les règles d'associations maximales ont plusieurs avantages comme la capacité à faire ressortir les corrélations existantes entre les données textuelles. Cependant, quand vient le moment de traiter un texte, la variété riche du vocabulaire le composant devient un obstacle de taille. Le nombre d'associations est alors beaucoup trop important et les examiner au complet devient alors une tâche trop ardue et trop coûteuse en temps.

Dans ce mémoire, nous proposons une approche combinant à la fois la classification et les règles d'associations maximales afin d'obtenir un accès aux données textuelles plus efficace. Nous démontrerons que les règles d'associations maximales sont d'un grand apport quant à la compréhension des résultats issus de la classification. Nous démonterons que, grâce aux règles d'associations maximales, nous pouvons interpréter les résultats issus des classifieurs même si ces résultats semblent parfois dénués de sens.

REMERCIEMENTS

Tout d'abord, je tiens à exprimer toute ma gratitude envers mon directeur de recherche, Ismaïl Biskri, professeur au département de Mathématiques et d'Informatique de l'Université du Québec à Trois-Rivières. En plus de m'avoir proposé ce projet passionnant, il a su m'encourager et me soutenir tout au long de ma maitrise. La confiance qu'il m'a témoignée, sa grande disponibilité et ses judicieux conseils ont assuré la réussite de ce projet. Un remerciement spécial aussi pour avoir lu attentivement mon mémoire et avoir suggéré des modifications qui ont amélioré grandement la version finale.

Je remercie les membres du jury ayant participés à l'évaluation de ce mémoire. Leurs judicieux commentaires ont permis d'améliorer la qualité de mon travail.

Je tiens également à remercier mon oncle, Gilbert Descôteaux, pour son soutien et son aide.

Merci également à mon employeur pour m'avoir libéré du temps en guise de soutien ainsi que m'avoir soutenu financièrement.

TABLE DES MATIERES

LISTE DES ÉQUATIONS

LISTE DES FIGURES

LISTE DES SIGLES

KNN : K-Nearest-Neighbors.

K-means : Algorithme des K-moyennes

SOM : Self Organizing Map

ART : Adaptative Resonance Technique

LISTE DES TABLEAUX

CHAPITRE 1

INTRODUCTION

Depuis que l'homme a pris conscience du monde l'entourant, il a commencé à chercher à catégoriser l'ensemble de ses connaissances et de son savoir. Le premier but de la catégorisation est de classer pour comprendre le pourquoi. En effet, comment comprendre ce qui nous entoure si nous ne pouvons avoir de références sur lesquelles nous baser et nous comparer ? C'est en répartissant des objets ayant des propriétés similaires en groupes homogènes que nous sommes à même de comprendre les singularités spécifiques à chacun qui les unissent.

L'utilité de la classification est vaste et touche beaucoup de domaines tels que la biologie, la médecine, l'astronomie, la chimie et plusieurs autres. Classer ce fouillis d'informations afin de mieux le comprendre et de mieux s'y retrouver, voilà un défi de taille.

La classification n'est pas un domaine d'application nouveau. En fait, il remonte à plus de 2000 ans. Aristote, aux alentours du troisième siècle avant JC inventa le principe de la classification naturelle des êtres vivants. Il catégorisa chaque être en genres et en espèces en se basant sur leurs propriétés intrinsèques et le mode de production comme divergence. Son modèle de classification perdura jusqu'aux scientifiques du 18^{e} siècle.

Les domaines et les exemples dans lesquels la classification est utilisée sont multiples. Il existe un grand nombre de domaines. En voici quelques-uns [14].

- La biologie

  Dans ce domaine, nous retrouvons la classification que nous appelons classification scientifiques des espèces ou systématique.

- La classification classique (du 18^{e} siècle à nos jours).
- La classification phylogénétique (remplace graduellement la classification classique depuis la moitié du 20^{e} siècle).

➢ La médecine

La médecine comprend la famille des classifications internationales (FCI).

- La classification internationale des maladies,
- La classification internationale du fonctionnement, du handicap et de la santé
- La classification internationale des soins primaires.

➢ L'astronomie.

- La classification des astres.
- La classification des météores.

➢ La chimie

- Le tableau périodique des éléments

➢ La minéralogie

- La classification des minéraux (reposant essentiellement sur leur composition chimique ainsi que de leur structure cristalline.

Avec la facilité d'accès d'internet au grand public, le monde a assisté à la naissance d'une nouvelle source de documentation. La toile est considérée par plusieurs comme la plus grande bibliothèque du monde. Nous pouvons y retrouver une quantité incommensurable de livres, articles, journaux ou autres sous forme de documentation électronique. Un des avantages du document électronique réside dans son immatérialité. Une bibliothèque complète peut être enregistrée sur un support pouvant reposer dans le creux d'une main ou tout simplement être enregistré dans un espace virtuel. Cette ère nouvelle a, cependant, amenée une nouvelle problématique. Avec l'ascension grandissante de l'internet et la prolifération des sites le composant, la toile a été envahie par une multitude d'informations sur tous les sujets. Cette expansion est telle, qu'il est

maintenant humainement impossible de rapatrier toutes les informations sur un sujet précis sans avoir recours à des outils de recherches.
Dans ce mémoire nous proposons une nouvelle approche combinant à la fois deux méthode de forage soit la classification et les règles d'associations maximales.

Le reste se présente comme suit. Le chapitre deux décrit quelques méthodes de classifications existantes et leur fonctionnement. Le chapitre trois permet de passer en revue les règles d'association et les règles d'associations maximales. Nous poursuivons avec le chapitre quatre où nous expliquons comment nous combinons les deux méthodes. Au chapitre cinq, nous décrivons dans un premier temps, les différentes parties du logiciel qui a été développé ainsi que leurs fonctionnalités. Dans un deuxième temps, nous procédons à l'expérimentation du logiciel développé avec l'analyse d'un texte concret.

Enfin, nous terminons par une conclusion dans laquelle nous présentons une synthèse des résultats obtenus.

CHAPITRE 2

LES MÉTHODES DE CLASSIFICATION

INTRODUCTION

Le processus de classification a pour but de regrouper les éléments possédants des caractéristiques communes en sous-groupes distincts. Dans la littérature scientifique nous retrouvons plusieurs classifieurs. Nous nous attarderons à quatre d'entre eux qui ont l'avantage d'être relativement simple :

- K-NN.
- K-means.
- SOM.
- ART.

Ces algorithmes bien qu'ils diffèrent, s'appuient tous sur la notion de distance afin de calculer la proximité (similarité) entre 2 informations à traiter. Dépendant des stratégies et des résultats recherchés, les objets peuvent se présenter sous différentes formes : neurones, vecteurs, chiffres, etc. Dans le contexte de ce mémoire, les objets seront présentés sous la forme de vecteur.

2.1 Représentation vectorielle

Un espace vectoriel est un groupe de vecteurs dans un espace de dimension N. Un vecteur représente un objet pour lequel on attribue une pondération spécifique à chacune des dimensions le définissant (tableau 1).

	C1	C2	C3	C4	C5	C6	C7	C8	C9	...
Objet 1	1	2	3	1	3	2	1	1	1	...
Objet 2	0	2	1	1	0	0	1	1	2	...
...	...	...	...	...	...	...	...	...	...	...

Tableau 1 - représentation vectorielle

2.2 Calcul de distance

Pour mesurer la similarité entre deux objets, une notion de distance, dépendante du type de données utilisées, est nécessaire. Plusieurs méthodes de calcul ont été développées (et continuent encore à être à l'étude) afin de calculer la distance entre 2 éléments à classifier.

Dans la littérature, nous avons pu observer deux types de données existantes. Les éléments dichotomiques et les éléments numériques.

Les éléments dichotomiques (couleurs, qualités, etc.) sont représentés par des éléments binaires. Zéro correspondant à l'absence de l'élément et un, à sa présence. Pour les comparer, il existe des mesures de distance qui basent leur calcul sur la fréquence et la similarité des objets à comparer. Deux des plus connues sont la mesure de distance de Hamming et la mesure de distance de Jaccard.

Pour les éléments numériques tels que le poids, l'âge, un nombre, etc., la mesure de distance la plus connue est la distance euclidienne. Il existe 3 variantes de celle-ci que nous présenterons : 1) la distance de Manhattan (city block), 2) la distance de Tchebychev et 3) la distance de Minkowski.

2.2.1 Calcul de distance

2.2.1.1 Distance de Hamming

La première mesure que nous présentons est la distance de Hamming [15]. Définie par Richard Hamming, elle calcule le nombre de bits différents entre 2 vecteurs binaires. Prenons comme exemple deux groupes, (jaune, bleu) et (rouge, bleu) basés sur le vecteur caractéristique suivant : (bleu, jaune, rouge). Leur représentation est donc A : (1, 1, 0) pour le premier et B : (1, 0, 1) pour le deuxième (équation 1).

$$H(A,B) = |A \cup B| - |A \cap B| = |(1,1,0) \cup (1,0,1)| - |(1,1,0) \cap (1,0,1)| = |3 - 1| = 2$$

Équation 1- distance de Hamming

2.2.1.2 La distance de Jaccard.

Elle doit son nom au botaniste suisse Paul Jaccard [16]. Cette mesure calcule la dissimilarité entre deux vecteurs, en calculant la valeur absolue de l'union des deux vecteurs de laquelle on soustrait la valeur absolue de l'intersection de ceux-ci. Le résultat est ensuite divisé par la fonction cardinale de l'union des deux vecteurs. Si nous prenons l'exemple énoncé au point 2.1.1 où nous avons deux groupes, (jaune, bleu) et (rouge, bleu) basés sur l'ensemble suivant : (bleu, jaune, rouge), nous obtenons une distance de 0.66 (équation 2).

$$J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} = \frac{|(1,1,0) \cup (1,0,1)| - |(1,1,0) \cap (1,0,1)|}{|(1,1,0) \cup (1,0,1)|} = \frac{|3| - |1|}{|3|} = 0.66$$

Équation 2 - distance de Jaccard.

2.2.1.3 Distance Euclidienne.

La distance Euclidienne permet de définir la distance entre 2 vecteurs représentés comme deux coordonnées sur le plan cartésien [17]. La distance représente le calcul de

l'hypoténuse d'un triangle rectangle dont la longueur des côtés correspond à la projection de la distance de chaque dimension des vecteurs sur leurs axes respectifs. Si nous avons deux vecteurs V1 et V2 à deux dimensions soit V1 = (A1, B1) et V2 = (A2, B2), la longueur de chacun des côtés correspond donc à la valeur absolue de la différence des composantes de chaque dimension (figure 1).

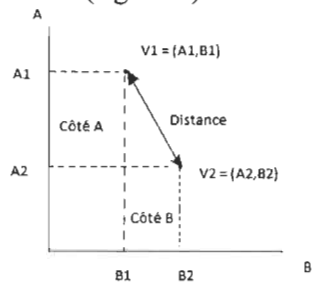


Figure 1 - Représentation de la distance Euclidienne.

Pour un vecteur V1 (3,4) et un vecteur V2(4,5), la distance est de 1.4 (équation 3).

$$d = \sum_{i=1}^{n} \sqrt{|a_i - b_i|^2} = \sqrt{|a1 - b1|^2 + |a2 - b2|^2} = \sqrt{|4 - 3|^2 + |5 - 4|^2} = \sqrt{2} = 1.4$$

Équation 3 - distance Euclidienne.

2.2.1.4 Distance de Manhattan (City-Block).

La distance de Manhattan permet aussi de définir la distance entre 2 vecteurs sur le plan cartésien [17]. La longueur des côtés correspond à la projection de la distance de chaque dimension des vecteurs sur leurs axes respectifs. On obtient cette mesure en additionnant la longueur des côtés A et B (figure 2).

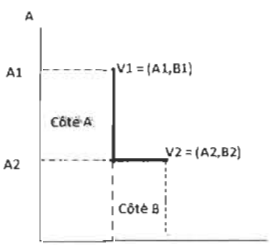


Figure 2 - Représentation de la distance de Manhattan.

Si nous utilisons les vecteurs V1 et V2 de l'exemple précédent soit : V1 (3,4) et V2(4,5). La distance est de 2 (équation 4).

$$d = |a1 - b1| + |a2 - b2| = (4-3) + (5-4) = 2$$

Équation 4 - distance de Manhattan.

2.2.1.5 Distance de Minkowski.

La distance de Minkowski (4)est une généralisation de la distance euclidienne et de la distance de Manhattan à p dimensions (équation 5). Cette distance est rarement utilisée.

$$d = \sum_{i=1}^{n} \sqrt[p]{|a_i - c_i|^p}$$

Équation 5 - distance de Minkowski.

2.2.1.6 Distance de Chebyshev.

La distance de Chebyshev [17] est un cas particulier de la distance de Minkowski. Elle représente la distance entre les deux vecteurs à l'endroit le plus éloigné (équation 6).

$$d = \lim_{p \to \infty} \sum_{i=1}^{n} \sqrt[p]{Max(\sum_{i=1}^{n} |a_i - c_i|)}$$

Équation 6 - distance de Chebyshev.

2.3 Classifieurs

La classification a pour but de regrouper des objets ayant des similitudes en des groupes homogènes. Nous présentons trois classifieurs, soit K-means, SOM et ART

2.3.1 La méthode K-NN

2.3.1.1 Principes fondamentaux

K-NN (k plus proches voisins ou k nearest Neighbors) fut inventée par Weiss and Kulikowski dans les années 1990 [18]. Il s'agit d'un algorithme de classification supervisé, c'est-à-dire qu'il doit obligatoirement posséder un ensemble d'apprentissages auquel se référer durant tout le classement. Le principe général de cette méthode est le suivant : pour chaque nouvelle entrée i, il faut trouver k vecteurs (voisins) qui lui sont le plus similaire et ensuite, il faut retourner la classe C associée à la majorité de ceux-ci.

2.3.1.2 Entrée et sortie

Lors de l'initialisation, on doit spécifier au programme le nombre de classes K que nous voulons obtenir de même qu'un ensemble d'apprentissages de départ. Cet ensemble est composé d'un nombre prédéterminé de vecteurs choisis au hasard ou par l'utilisateur parmi les vecteurs devant être classés. Cette étape complétée, chaque vecteur (vecteur i) de l'ensemble est alors présenté au classifieur qui l'associe à une classe spécifique (classe j) (figure 3).

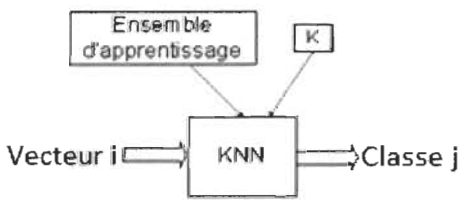


Figure 3 - Représentation des paramètres nécessaires à la méthode K-NN.

2.3.1.3 Algorithme

Voici un algorithme simple de la méthode K-NN[18].

```
Assigner un nombre de voisins K
Introduire les vecteurs de la base d'apprentissage
Pour chaque nouveau vecteur
        Trouver les K plus proches voisins.
        Si une classe de voisins est prédominante
                Assigner le vecteur à la classe.
        Sinon
                Assigner le vecteur à la classe dont le vecteur est le plus proche.
        Fin si
Fin pour
```

2.3.1.4 Explication

Pour expliquer cet algorithme, prenons l'exemple suivant : classifier un vecteur X dans un système possédant 2 classes distinctes A, B avec un K égal à trois. À l'ajout du vecteur, K-NN suivra la démarche suivante :

1. Calcul de la distance entre le vecteur X et tous les vecteurs déjà classifiés.

2. Sélection des K vecteurs (trois dans le cas qui nous concerne) dont la distance entre le vecteur à classifier et les vecteurs du classifieur est la plus petite.

3. Association du nouveau vecteur à la classe possédant le plus grand nombre de voisins se trouvant à proximité. Exemple : Pour un nouveau vecteur X, si nous

avons deux classes comportant chacune trois membres soit A (A1, A2, A3) et B (B1, B2, B3) et si les vecteurs les moins distants sont le vecteur A1, le vecteur A2 et le vecteur B3, alors X sera associé à la classe A.

2.3.1.5 Particularité de l'algorithme

Plusieurs facteurs influent sur les résultats probants de cet algorithme de classification. Nous observons les éléments suivants :

1) Le nombre K

   Si nous prenons un K égal à trois, nous aurons alors une comparaison avec ses 3 plus proches voisins. Il est conseillé d'avoir un nombre impair afin qu'il y ait toujours une classe prédominante lors du traitement. Plus le nombre K est élevé, plus le nombre de comparaisons est grand. Si aucune classe ne prédomine, celle dont les distances entre le vecteur à classer et les vecteurs la composant sont les plus petites est choisie.

2) Le choix des vecteurs des éléments initiaux.

   Les vecteurs initiaux doivent être choisis de manière à représenter l'ensemble des données et non au hasard. Choisir des vecteurs initiaux trop rapprochés ou trop éloignés aura pour effet d'associer les vecteurs à classifier une classe en particulier plutôt que de les répartir sur l'ensemble des vecteurs servant de point central aux classes.

3) L'ordre dans lequel les vecteurs sont présentés à l'algorithme K-NN.

   Une trop faible convergence entraînera une plus grande difformité des classes. Si les vecteurs présentés en entrée sont trop rapprochés les uns des autres, tous les

vecteurs s'agglutineront à la même classe, celle-ci devenant la classe ayant le plus de voisins respectifs.

2.3.2 La méthode K-means

2.3.2.1 Principes fondamentaux

La méthode K-means (K-moyennes) fut inventée en 1967 par McQueen [19]. K-means est un algorithme de classification semi-supervisé, c'est-à-dire qu'elle doit posséder un ensemble d'apprentissages minimal afin de pouvoir être utilisée. Celui-ci sera appelé à évoluer de façon automatique tout au long du processus de classification. Cette méthode classifie les vecteurs en les associant à la classe dont le vecteur central (centroïde) est le plus proche du vecteur à classifier. K-means recentre ensuite le centroïde de la classe sélectionnée.

2.3.2.2 Entrée et sortie

Comme pour la méthode K-NN, lors de l'initialisation, on doit spécifier au programme le nombre K de classes. Le programme choisi ensuite K vecteurs qui formeront l'ensemble des centroïdes de départ (centres des classes de départ). Ces centroïdes sont des vecteurs virtuels à partir desquels la distance sera évaluée et donc ne sont pas inclus dans les classes. Cette étape complétée, chaque vecteur (vecteur i) de l'ensemble est alors présenté au classifieur qui l'associe à une classe spécifique (classe j) (figure 4).

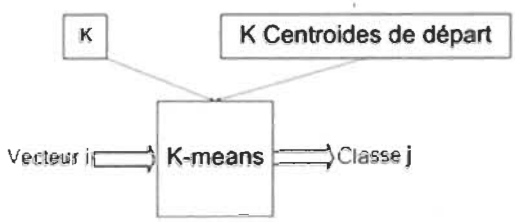


Figure 4 - Représentation des paramètres nécessaire à la méthode K-Means.

2.3.2.3 Algorithme

Voici l'algorithme de la méthode K-Means [20].

> Assigner un nombre de classes K et initialiser leurs centroïdes d'origine à 0.
> Désigner K centroïdes parmi les vecteurs à classifier.
> Remplacer les centroïdes d'origine par les centroïdes désignés.
> **Pour** chaque vecteur à classifier
> Trouver le centroïde le plus proche.
> Assigner le vecteur à la classe du centroïde désigné.
> Recalculer le centre du centroïde désigné.
> **Fin pour**

2.3.2.4 Explication

La méthode K-means suit la même stratégie que la méthode K-NN à l'exception que les centroïdes ne sont pas des vecteurs réels mais plutôt des vecteurs insérés dans l'algorithme.

Afin d'expliquer cet algorithme, prenons l'exemple suivant : classifier un vecteur X dans un système possédant 3 classes distinctes A, B, C (K étant égal à trois).
À l'ajout du vecteur X, K- Means suivra la démarche suivante :

1. Calcul de la distance entre le vecteur X et les centroïdes des classes A, B et C.
2. Association avec la classe donc le centroïde est le plus proche. exemple A.
3. Recalcule du centroïde de A afin qu'il soit représentatif du centre de la classe.

2.3.2.5 Spécificité de l'algorithme

Les facteurs influant sur les résultats probants de cet algorithme de classification sont les suivants :

1) Le nombre K

   Comme il représente le nombre de classes, alors, plus le nombre est élevé, plus le nombre de classes le sera aussi.

2) Le choix de la méthode de calcul de la distance

   La performance de cette technique est proportionnelle à la qualité de la fonction de mesure de distance utilisée.

3) L'ordre dans lequel les vecteurs sont présentés à l'algorithme K-Means

   Comme pour K-NN, une trop faible convergence entraînera une plus grande difformité des classes. Exemple : si les vecteurs présentés en entrée sont ordonnés, tous les vecteurs s'agglutineront à la même classe.

2.3.3 La méthode SOM (carte auto-adaptative)

2.3.3.1 Principes fondamentaux

La méthode SOM (Self Organizing Map) fut inventée par Teuvo Kohonen en 1984 [21]. SOM est un algorithme de classification basé sur des méthodes non supervisées, c'est-à-dire sans ensemble d'apprentissage. Il s'agit en fait d'un réseau de neurones artificiels

(représentant les classes du système de classification) utilisé pour minimiser des espaces de grandes dimensions.

Comme nous le montre la figure 5, la carte de Kohonen a la forme d'une matrice à deux dimensions dont les vecteurs référents (centres des classes) en sont les coefficients. (Dans la figure 5, une carte de largeur trois et de hauteur trois a été préalablement déterminé formant ainsi une matrice composée de neuf intersections représentant les classes. Les valeurs des vecteurs référents (centre) de chacune de celles-ci sont choisies aléatoirement entre zéro et un. Chaque vecteur à classifier est, par la suite, associé à la classe la plus proche. À chaque fois qu'un vecteur est classifié dans une classe, le vecteur référent est recentré afin de refléter la nouvelle réalité ainsi que ses voisins immédiats.

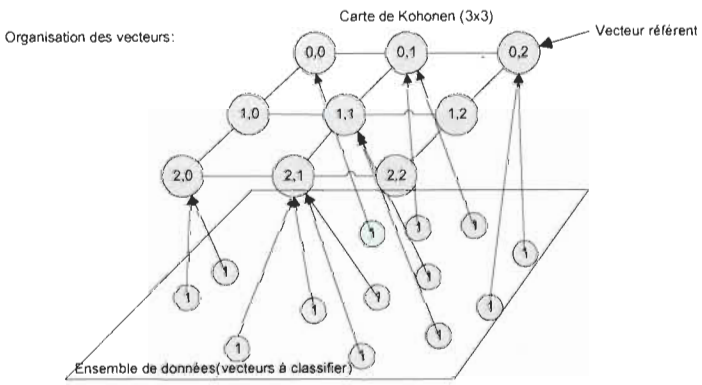


Figure 5 - Représentation de la méthode SOM

2.3.3.2 Entrée et sortie

Lors de l'initialisation, on doit spécifier seulement la grandeur de la carte au programme, soit sa largeur et sa hauteur. Cette étape complétée, chaque vecteur (vecteur i) de l'ensemble est alors présenté au classifieur qui l'associe à une classe spécifique (classe j) (figure 6).

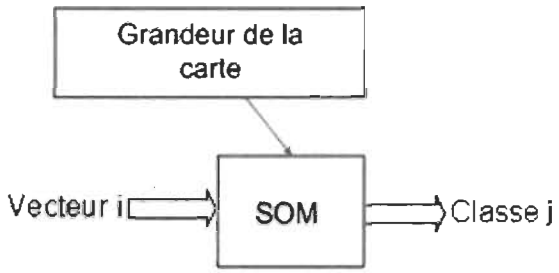


Figure 6 - Représentation des paramètres nécessaire à la méthode SOM.

2.3.3.3 Algorithme

Voici l'algorithme de la méthode SOM [22].

Initialisation du réseau

Création de la table de Kohonen selon les dimensions demandées.

Pour chaque neurone de la matrice de Kohonen

Affecter une valeur vectorielle au hasard (construction des vecteurs référents).

Fin Pour

Classification des vecteurs de l'espace de données (formation de la carte)

Pour chaque vecteur de l'espace de données

Recherche de la classe (nœud) dont le vecteur référent est le plus près grâce à la mesure de distance choisie.

Assignation du vecteur i à la classe désignée.

Pour chaque classe (nœud) de la carte //processus de diffusion.

Réajustement de son vecteur référent grâce à une fonction de type gaussienne

Fin Pour

Fin Pour

2.3.3.4 Explication

Prenons par exemple la classification du vecteur à quatre dimensions : V1 (2, 3, 4, 5). L'algorithme SOM fonctionne de la manière suivante :

1) Phase d'initialisation

Pour la création de la carte, on choisit la forme de celle-ci en premier lieu. (Dans notre exemple (figure 7), nous déterminons une carte de dimension 2x2 mais nous aurions pu choisir une autre dimension). On détermine ensuite la valeur du vecteur référent de chaque classe. Comme chaque vecteur possède 4 dimensions, le processus choisira au hasard 4 nombres entre 0 et 1 pour chacun des quatre vecteurs référents grâce à une fonction d'attribution aléatoire. Par exemple, le vecteur référent de la classe 1 pourrait être le suivant : (0.1, 0.3, 0.12, 0.2).

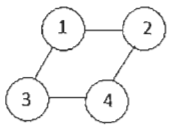


Figure 7 - Carte de Kohonen.

2) Classification des vecteurs de l'espace de données (formation de la carte)

Pour chaque vecteur X à classifier, l'algorithme recherche la classe dont le vecteur référent (centre de la classe) est le plus proche et associe le vecteur X à cette classe en lui attribuant son identifiant (numéro de classe).

Ensuite, l'algorithme réajuste le vecteur référent (centre de la classe) de la classe désignée ainsi que celui des autres classes (processus de diffusion) (figure 8) en leurs additionnant une valeur de correction obtenue grâce à une fonction de type gaussienne (équation 7). Le résultat obtenu est le suivant : plus on s'éloigne du

vecteur référent de la classe désignée, moins la valeur de correction est grande. Les voisins respectifs subissent donc une plus grande modification (figure 9).

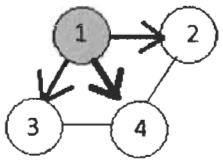


Figure 8 - Modification de la classe choisie ainsi que ses voisins respectifs.

$$e^{-\frac{d^2}{(e^{\overline{(1000/Log(poids)}}*poids)^2\,1}}$$

Équation 7 - Fonction gaussienne de l'algorithme SOM.

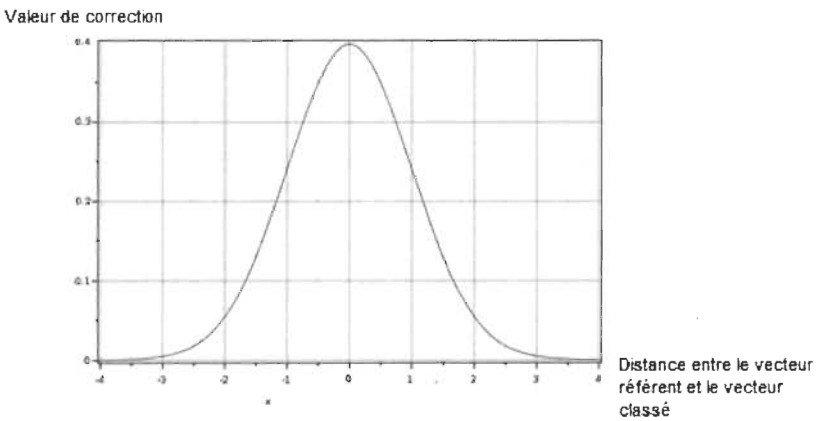


Figure 9 - .Représentation de la fonction de type gaussienne [23].

[1] La distance (d) est égale à la distance entre le vecteur référent et le vecteur classé. Le poids (poids) est égal à la racine carrée de la hauteur de la carte multiplié par la largeur de carte.

2.3.3.5 Particularité de l'algorithme

1) Grandeur de la carte

   La grandeur de la carte détermine le nombre de classes qui sera créée.

2) Aucun ensemble d'apprentissage requis.

   Seule, la dimension de la carte à besoin d'être connue.

3) L'ordre dans lequel les vecteurs sont présentés

   Comme pour les méthodes K-NN et K-means, l'ordre de présentation des vecteurs au processus de classification est intimement lié à la formation des classes.

2.3.4 ART

2.3.4.1 Principes fondamentaux

ART (Adaptative Resonance Technique) est une théorie développée par Stephen Grossberg et Gail Carpenter sur les aspects de la façon dont le cerveau traite l'information [24]. C'est une technique d'apprentissage non supervisé : il est capable d'apprendre à reconnaître un vecteur qu'on lui présente en fonction des catégories qu'il construit. Il auto-organise les catégories et en crée quand cela lui semble nécessaire selon un critère d'autocréation.

Le principe de cette méthode est simple : chaque vecteur est présenté à une base d'apprentissage pour ensuite être comparé aux choix de celle-ci. Si un vecteur de l'ensemble d'apprentissage s'y apparente (c'est-à-dire que le calcul de la distance le

séparant de celui de l'ensemble est inférieur à un certain taux d'apprentissage (ce taux, que l'on nomme Rho est prédéterminé par l'utilisateur), il est choisi, sinon, on ajoute celui-ci comme choix futur dans la base d'apprentissage.

2.3.4.2 Entrée et sortie

Lors de l'initialisation, on doit spécifier au programme le facteur d'apprentissage Rho permettant de comparer les vecteurs à ceux de l'ensemble d'apprentissage. Cette étape complétée, chaque vecteur (vecteur i) de l'ensemble est alors présenté au classifieur qui l'associe à une classe spécifique (classe j) (figure 10).

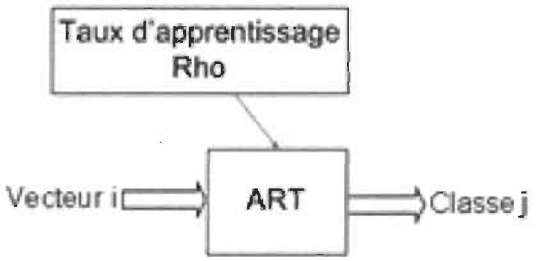


Figure 10 - Représentation des paramètres nécessaire à la méthode ART.

2.3.4.3 Algorithme

Voici l'algorithme de la méthode ART [25].

Présentation d'un vecteur.
Tant qu'il y a des vecteurs disponibles dans la base d'apprentissage.
    Sélection d'un vecteur ressemblant le plus au vecteur d'entrée.
    Essaie d'association du vecteur avec le vecteur d'entrée.
    Si association (distance < rho)
        On quitte et on retourne la classe du vecteur gagnant

Sinon

Élimination du vecteur des choix possibles

Fin si

FinTant que

Si un vecteur a été sélectionné

Ajustement du poids du vecteur gagnant

Sinon

Ajout du vecteur à la base de connaissance

Fin si

2.3.4.4 Explication

Dans cet exemple, nous supposons que la base d'apprentissage comporte quatre vecteurs et que la valeur du taux d'apprentissage (Rho) a été déterminée.

1) Présentation d'un vecteur (V) à la base d'apprentissage E composée de quatre vecteurs (figure 11).

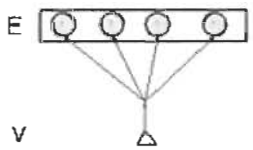


Figure 11 – Présentation d'un vecteur V.

2) On choisit un vecteur présent dans la base d'apprentissages E afin de le comparer à notre vecteur à classifier V. On effectue la comparaison en calculant la distance entre E et V. Si la distance est plus petite que le taux d'apprentissage Rho, alors on les considère semblables. Si le résultat est plus élevé, on écarte le vecteur d'apprentissage des choix possible et on ré exécute

le processus jusqu'à ce qu'un vecteur soit choisi ou que tous les choix aient été écartés.

3) Si un vecteur de la base d'apprentissages est sélectionné, on lui associe le vecteur à classifier et on ajuste le poids du vecteur sélectionné. L'ajustement du poids est nécessaire afin que le vecteur de la base d'apprentissage continue à représenter l'ensemble des vecteurs auxquels il est associé. Si aucun vecteur de la base d'apprentissage n'est sélectionné, on ajoute une copie du vecteur à classifier à la base d'apprentissage (création d'une classe) et on lui associe le vecteur à classifier (figure 12).

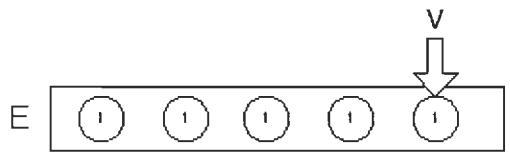


Figure 12 - Ajout d'un vecteur à la base d'apprentissage.

2.3.4.5 Particularité de l'algorithme

Certains facteurs affectent les résultats de cet algorithme de classification. Nous avons :

1) Le facteur Rho(ρ).

   Plus le facteur Rho est élevé, plus le nombre de classes formées sera élevé.

2) L'ordre dans lequel les vecteurs sont présentés à l'algorithme ART.

   Dépendant de l'ordre dans lequel les vecteurs sont présentés à ART, le nombre de classes formées ainsi que les membres composants chacune de celle-ci variera.

2.4 Ordre d'entrée des vecteurs

Les algorithmes de base, tels que nous les connaissons et comme nous les voyons dans la littérature, classifient les vecteurs selon l'ordre dans lequel ils sont présentés à ceux-ci.

Cette particularité est celle de tous les algorithmes de classification que nous avons présentés. La conséquence de cette manière de faire réside au niveau de la phase d'apprentissage. Quand un vecteur s'ajoute à une classe et que son environnement est modifié, sa modification résulte des vecteurs le composant.

Comme nous le voyons dans les figures 13 et 14, si un vecteur est introduit au début de la mise en place de la hiérarchie de classes (figure 13), il risque fort de se retrouver dans une classe différente s'il est introduit en dernier (figure 14).

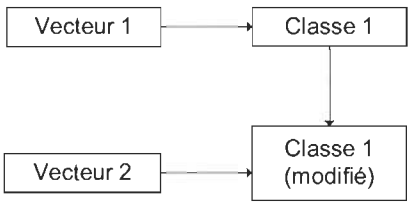


Figure 13 - Vecteur 2 associé à la classe 1.

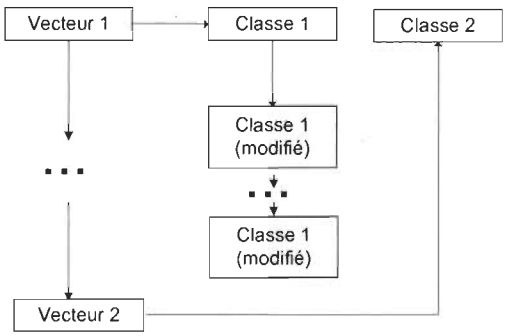


Figure 14 - Vecteur 2 associé à la classe 2.

2.4.1 Stabilisation des classes

Pour pallier à ce problème, il convient d'instaurer des routines pour que les classes se stabilisent d'elles-mêmes. Ceci consiste à modifier les algorithmes de classification pour faire en sorte de recommencer le processus de classification de tous les vecteurs jusqu'à ce que tous les vecteurs soient classés correctement. (Comme nous le voyons dans la représentation de l'algorithme au point 2.3.2)

On parvient à cette finalité en modifiant les algorithmes existants. La méthode de calcul du taux d'erreur diffère d'un classifieur à l'autre mais le principe reste le même. Vérifier si la distance entre les vecteurs à classifier et les vecteurs de références (selon le classifieur choisi) ne dépasse pas un certain seuil d'erreur. Si cette valeur est plus petite que celui-ci pour tous les vecteurs, on suppose alors que la position des vecteurs de référence est optimale et que les classes sont donc stabilisées. Sinon, on considère que des vecteurs se retrouvent au sein de mauvaises classes et on ré exécute tout le processus de classification (exception faite du processus d'apprentissage, celui-ci étant le groupe de vecteurs de référence trouvé lors de la précédente itération). Il est à noter que plus le seuil du taux d'erreur est minime, plus la classification est précise. Mais en contrepartie, le temps d'exécution sera élevé, dû au nombre d'exécutions exhaustif qui sera nécessaire afin d'atteindre le seuil demandé.

2.4.2 Algorithme de stabilisation

Initialisation du seuil de taux d'erreur
Exécuter la phase d'apprentissage du classifieur choisi
Répéter
    Exécuter l'algorithme de classification du classifieur choisi.
    Calculer le taux d'erreur.
Tant que le taux d'erreur est plus grand que le seuil de taux d'erreur

2.5 Conclusion

Dans ce chapitre, Nous avons présenté en premier lieu, différentes méthodes de calcul de distance utilisées par les classifieurs et avons ensuite présenté quatre classifieurs soit K-NN, K-Means, SOM et ART.

Chaque classifieur donne un résultat différent. La variation entre les résultats ont déjà fait l'objet de travaux de recherche comme ceux présentés dans Turenne[13].

Dans le cadre d'un processus de recherche d'information, l'utilisation des méthodes de classification demeure une approche très intéressante.

Une autre approche utilisée consiste en l'utilisation des règles d'associations maximales. C'est ce sujet qui est présenté dans le prochain chapitre.

CHAPITRE 3

LES RÈGLES D'ASSOCIATIONS MAXIMALES

3.1 Introduction

Introduites par Agrawal durant les années 90 [2] [3], les règles d'associations représentent un domaine relativement récent. Elles visent de prime abord à élaborer un modèle basé sur la prérogative suivante : si une condition existe, alors forcément, un résultat issu de celle-ci existe aussi.

Antécédent → conséquent

Plus récemment, des travaux ont été menés pour juger de la qualité de leur interprétation [9] [13].

3.2 Règles d'associations.

Avant de parler des règles d'association maximales, il est de mise d'expliquer en quoi consistent les règles d'associations. Présentées dans les années 60 [8], les règles d'associations sont un outil très pertinent pour extraire des relations existantes dans un ensemble de données, par exemple, un texte. Elles se définissent comme suit : « À chaque fois que des éléments antécédents sont rencontrés dans une transaction, les éléments conséquents le sont aussi. »

Bien sûr, afin de bien cibler les règles d'association pertinentes et éliminer les données non pertinentes, des mesures de qualité existent. Nous avons la mesure du support, (fréquence minimum où les éléments antécédents (X) et les éléments conséquents (Y)

apparaissent ensemble par rapport au total des transactions) et la mesure de confiance (taux minimum où X et Y apparaissent par rapport à toutes les transactions contenant X).

La règle prend donc la forme suivante : X →Y (%support,% confiance) ou X est l'antécédent, Y : le conséquent, %support : le pourcentage de support minimum et % confiance : le pourcentage de la confiance minimum requis.

Exemple :

Pour un ensemble E= $\{PC\ de\ bureau, Portable, Ipad, Clavier, Souris\}$ dont nous tirons 2 sous-ensembles : E1 = {PC de bureau, Portable, Ipad} et E2= {Clavier, Souris} et les transactions T1 = {PC de bureau, Portable, Souris}, T2 = {PC de bureau, souris, clavier} et T3 = {PC de bureau, clavier} (figure 15). Nous voulons étudier la règle d'association : PC de bureau→Souris dont nous établissons le pourcentage du support minimal à 60% et la confiance minimale à 50%.

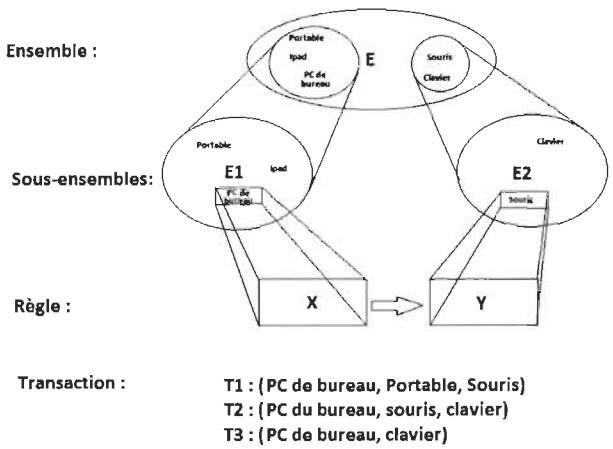


Figure 15 - Représentation du système de règles d'associations.

Nous constatons que le support de la règle est de 66% (PC de bureau et Souris apparaissent ensemble dans 66% de toutes les transactions (T1 et T2)) (équation 8). Le taux de confiance est quant à lui de 66% (PC de bureau et Souris apparaissent ensemble dans 66% de toutes les transactions ou PC de bureau est présent (T1 et T2)) (équation 9). La règle d'association est donc valide.

Support (PC de bureau→Souris) : 2/3 = 66%

{PC de bureau, Souris} ⊆ T1 = VRAI

{PC de bureau, Souris} ⊆ T2 = VRAI

{PC de bureau, Souris} ⊆ T3 = FAUX

Équation 8 - Calcul du support de la règle PC de bureau→Souris (50%,60%).

Confiance (PC de bureau→Souris) : 2/3 = 66%

{PC de bureau, Souris} ⊆ T1 & {PC de bureau} ⊆ T1 = VRAI

{PC de bureau, Souris} ⊆ T2 & {PC de bureau} ⊆ T2 = VRAI

{PC de bureau, Souris} ⊆ T3 & {PC de bureau} ⊆ T3 = FAUX

Équation 9 - Calcul de la confiance de la règle PC de bureau→Souris(50%,60%).

3.2.1 Propriétés des règles d'associations

Ces règles d'associations possèdent plusieurs propriétés générales [26]. (Notons que nous ne sommes pas nécessairement en accord avec certaines de ces propriétés) :

1) Les règles ne peuvent être combinées.

   Même si X →Z et Y →Z sont vraies, X U Y →Z n'est pas nécessairement vrai.

2) Une règle ne peut être décomposée.

Même s'il est juste de dire que X U Y $\rightarrow$Z, X $\rightarrow$Z et Y $\rightarrow$Z ne sont pas nécessairement vraies.

3) Les règles ne sont pas transitives.

Si X $\rightarrow$Y et Y $\rightarrow$Z, on ne peut en déduire que X $\rightarrow$Z.

4) Les items de chaque transaction sont uniques

On ne peut retrouver deux fois le même item.

5) Les items sont inclus dans l'ensemble de départ (X $\subseteq$ Ensemble de départ).

Tous les items d'une transaction doivent être inclus dans leurs ensembles de départ respectifs. (Ex. X $\subseteq$ E1), alors on dit que cette transaction supporte l'ensemble.

6) Les sous-ensembles sont disjoints

L'intersection des sous-ensembles X et Y est vide (X $\rightarrow$Y = $\varnothing$.), c'est-à-dire que nous ne retrouvons aucun des éléments de X dans Y et inversement.

3.2.2 Opérations des règles d'associations

Plusieurs opérations peuvent être effectuées sur les transactions.

3.2.2.1 Calcul du Support de la règle (S (X→Y) = fréquence (X→Y))

Le calcul du support de la règle représente le nombre de transactions contenant les éléments de X et Y($X \subseteq Ti$ ET $Y \subseteq Ti$) [4]. Si on considère la règle de l'exemple du point 3.2 soit ({X→ Y}): {PC de bureau} → {Souris} et les transactions T1 = {PC de bureau, Portable, Souris}, T2 = {PC de bureau, souris, clavier} et T3 = {PC de bureau, clavier}, le support de la règle S(X → Y) est égal à 2 (X et Y sont inclus dans T1 et T2).

3.2.2.2 Calcul de la confiance de la règle (C(X→Y) = S (X → Y) / S(X))

La confiance de la règle est le taux de présence de Y quand X est présent au sein de la transaction [4]. La confiance de la règle de l'exemple au point 3.2.2.1 est donc : C(X→Y) = S(X→Y) / S(X) = 2/3, donc 66% (Y est T1 et T2 tandis X est inclus dans T1, T2 et T3).

3.2.2.3 Avantages des règles d'associations

Les règles d'associations possèdent plusieurs avantages :

- Leurs applications dans de nombreux domaines.
- La découverte de connaissances et relations utiles entre divers mots d'un texte.
- Leur facilité d'utilisation et leur simplicité.
- La facilité d'interprétation des résultats.

3.2.3 Inconvénient des règles d'associations

Malgré tout, les règles d'associations accusent plusieurs faiblesses :

- Le temps considérable consacré à la recherche des ensembles de Y.
- La quantité exorbitante de règles qui découle d'une analyse.
- La difficulté d'évaluer la qualité et la pertinence des règles.

- La probabilité de créer des règles n'apportant rien de plus au modèle recherché.

3.3 Règles d'associations maximales

La plupart du temps, le modèle des règles d'association nous fait perdre de vue les associations qui sont moins fréquentes car elles sont laissées de côté par celui-ci. Les règles d'associations maximales nous permettent de palier à ce problème.

Elles se définissent comme suit : « Chaque fois que X apparait seul dans une transaction, c'est-à-dire que la transaction M-supporte X, Y apparaît également ». On peut affirmer que X apparaît seul si et seulement si pour une transaction donnée (Ti) et un ensemble donné Ej (incluant X), l'ensemble résultant de l'intersection de Ti et Ej est toujours égale à X ($Ti \cap Ej = X$).

Considérons comme exemple la règle ($\{X \rightarrow Y\}$): {PC de bureau} $\rightarrow$ {Souris}, la transaction T2 = {PC de bureau, souris, clavier} et l'ensemble E1 {PC de bureau, Portable, IPad}. La définition de la règle d'association maximale s'applique à T2 car elle M-supporte X ($T2 \cap E1 = X$) et l'ensemble Y est inclus dans la transaction T2 ($Y \subseteq T2$).

3.3.1 Propriétés et opérations des règles d'associations maximales

Les règles d'associations maximales possèdent les mêmes propriétés que les règles d'associations ordinaires. La différence résulte dans la manière de calculer le support de la règle ainsi que la confiance.

3.3.1.1 Opérations

3.3.1.1.1 Support maximal de la règle $X \xrightarrow{max} Y$

Le support maximal ($Smax(X \xrightarrow{max} Y)$) est le calcul du nombre de transactions satisfaisant à la définition de la règle d'association maximale ($X \xrightarrow{max} Y$) [4].

Considérons toujours l'exemple au point 3.2, soit la règle ({X→ Y}): {PC de bureau} → {Souris}, les transactions T1 = {PC de bureau, Portable, Souris}, T2 = {PC de bureau, souris, clavier}, T3 = {PC de bureau, clavier} et l'ensemble E1 (PC de bureau, Portable, IPad). Seule la transaction T2 satisfait la règle d'association maximale (voir l'exemple du point 3.3). T3 M-supporte X mais ne contient pas Y et T1 ne M-supporte pas X. $Smax(X \xrightarrow{max} Y)$ est donc égal à 1. Comme nous le remarquons, le support maximal est plus petit que le support de la règle qui était de 2 faisant du même coup ressortir la transaction T2 du lot de transactions.

3.3.1.1.2 La confiance maximale de la règle $X \xrightarrow{max} Y$

La confiance maximal ($Cmax(X \xrightarrow{max} Y) = Smax(X \xrightarrow{max} Y)/ Smax(X \xrightarrow{max} E2)$) est le pourcentage de transactions satisfaisant à la définition de la règle d'association maximale ($X \xrightarrow{max} Y$) par rapport au nombre de transactions contenant au moins un élément de l'ensemble E2 (ensemble d'où provient Y) [4].

Si on considère toujours l'exemple au point 3.2, la règle ({X→ Y}): {PC de bureau} → {Souris} et les transactions T1 = {PC de bureau, Portable, Souris}, T2 = {PC de bureau, souris, clavier} et T3 = {PC de bureau, clavier}, l'ensemble E1 (PC de bureau, Portable, IPad) et l'ensemble E2 (clavier, souris). Le Msupport de la règle étant de 1 (voir point 3.3.1.1.1 et le Msupport de E2 étant de 2 (T1 et T2 contiennent toutes les deux un

élément de Y et M-supportent X). La confiance maximale est donc égale à 50%. On retrouve donc Souris lorsque PC de bureau est présent et lorsque PC de bureau est seul dans 50 % des cas.

3.4 Relation entre les méthodes de classifications et les règles d'associations maximales

Les classes formées par les différentes méthodes de classification varient beaucoup dépendant de l'algorithme de classification utilisé. Elles prennent souvent la forme de liste beaucoup trop volumineuse et souvent difficile à interpréter pour l'utilisateur.

Les règles d'associations maximales sont, quant à elle, un outil très intéressant pour trouver la relation existante entre les mots d'un texte. Cependant, comme nous l'avons déjà mentionné, la richesse du vocabulaire le composant est un obstacle de taille. Cette méthode devient vite impraticable car le nombre d'association devient beaucoup trop important et les examiner au complet devient alors une tâche trop ardue et trop coûteuse en temps. Elles ont d'ailleurs fait l'objet de plusieurs mémoires de maitrise [1] [9].

Une solution envisageable est de combiner les méthodes de classification avec l'extraction des règles d'associations maximales pour faciliter l'exploitation des résultats issus de la classification.

3.5 Conclusion

Dans ce chapitre, Nous avons introduit les règles d'associations et avons ensuite présenté les règles d'association maximales. Pour chacune des deux approches, nous avons décrit leur fonctionnement ainsi que leurs propriétés et leurs opérations. Finalement, nous avons établi la relation existante entre les méthodes de classification et les règles d'associations maximales.

Dans le chapitre 4 nous présentons la méthodologie de recherche permettant de combiner les méthodes de classifications et les règles d'association maximales.

.

CHAPITRE 4

PROJET

4.1 Introduction

Le principal objectif de ce projet est de développer un outil capable d'analyser un texte ou plusieurs textes en combinant les méthodes de classifications et les règles d'association maximales.

Ce projet nous permettra de démontrer comment nous pouvons utiliser les classes issues de la classification avec les règles d'associations maximales afin d'en faire une interprétation. Nous montrerons que même si le contenu des classes résultantes des classifieurs semble parfois incohérent, il est quand même pertinent. Ce projet a fait l'objet de quelques publications [4] [5] [6] [7].

Dans ce chapitre, nous verrons les étapes que nous devons considérer afin de pouvoir utiliser les règles d'associations maximales.

Dans un premier temps, nous commencerons par voir un aperçu global des étapes du projet (figure 16) ainsi que leurs différentes fonctionnalités. Par la suite, nous approfondirons chacune des étapes en explicitant les algorithmes.

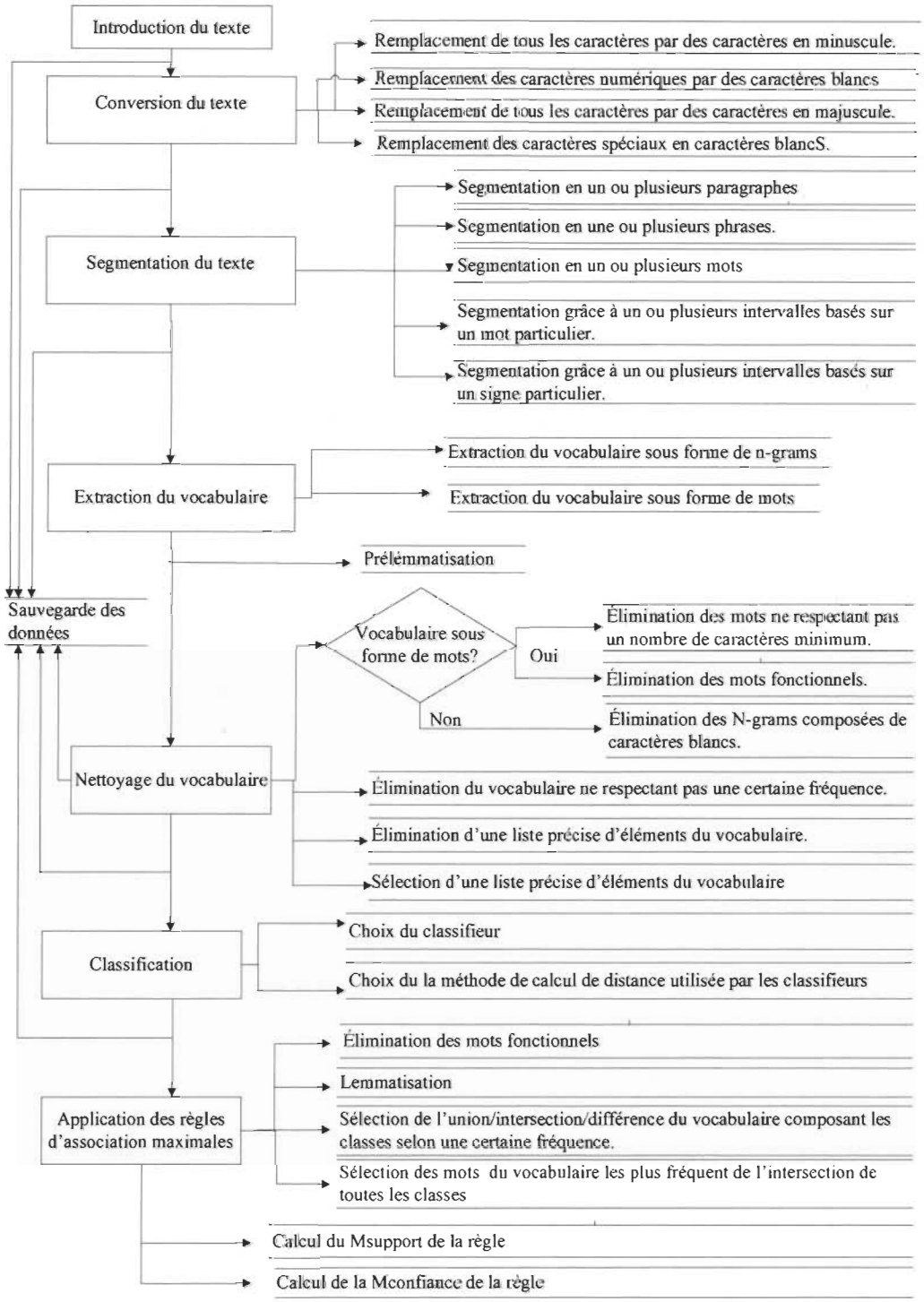


Figure 16 - Aperçu global des étapes du projet.

4.2 Introduction du texte

Le texte doit se trouver préalablement dans un fichier texte afin d'être pris en entrée par le logiciel. Il est ensuite lu, ligne par ligne et emmagasiné dans une liste. Mais le texte tel qu'il est présenté dans ce fichier peut être écrit dans plusieurs langues. C'est pourquoi, au moment d'être lu, celui-ci subit une conversion en langage unicode.

4.2.1 Algorithme

Voici l'algorithme permettant d'introduire le texte.

```
Ouverture en lecture du fichier en format unicode.
Tant que vrai
        Lecture d'une ligne
        Si la ligne est nulle
                On sort de la boucle
        Sinon
                On ajoute la ligne à la liste contenant les lignes du fichier.
        Fin si
Fin tant que
Fermeture du fichier
```

4.3 Conversion du texte

Le texte, tel que lu, comporte souvent des caractères qui, dépendamment de nos attentes, peuvent se révéler non significatifs tels la ponctuation, les caractères spéciaux, numériques ou autres. Afin d'éliminer ceux-ci, si nécessaire, différentes manipulations peuvent être effectuées.

4.3.1 Fonctionnalités

Voici les fonctionnalités pour la conversion du texte :

- Remplacement possible de tous les caractères par des caractères en majuscule.
- Remplacement possible de tous les caractères par des caractères en minuscule.
- Remplacement possible des caractères spéciaux par des caractères blanc.
- Remplacement possible des caractères numériques par des caractères blancs.

4.3.2 Algorithme

Voici l'algorithme permettant de convertir le texte.

Choix (option Majuscule/Minuscule)
    Option : Majuscule
        Transformation du texte en majuscule.*
    Option : Minuscule
        Transformation du texte en minuscule.*
Fin choix
Si l'option de caractères numériques par des caractères blancs est cochée
    Remplacement des caractères numériques en blanc.*
Fin si
Si l'option de caractères spéciaux en caractères blancs est cochée
    Remplacement des caractères spéciaux en blanc.*
Fin si
Retour de texte transformé
* Les transformations sont faites à partir de fonctions déjà existantes.

4.4 Segmentation du texte

La segmentation détermine le contenu ainsi que la taille des vecteurs.

4.4.1 Fonctionnalités

Le choix du type de la segmentation ainsi que le nombre d'éléments contenu dans chacun des segments relèvent de l'utilisateur. Dépendant des objectifs astreints, nous pouvons segmenter celui-ci d'une des façons suivantes (chacun de ces choix étant un paramètre saisi par l'utilisateur):

- Un ou plusieurs paragraphes.
- Une ou plusieurs phrases.
- Un ou plusieurs mots.
- Grâce à un ou plusieurs intervalles basés sur un mot particulier.
- Grâce à un ou plusieurs intervalles basés sur un signe particulier.

Le nombre d'éléments pour chacune des façons est un paramètre saisi par l'utilisateur.

4.4.2 Algorithme

Voici l'algorithme permettant de segmenter le texte.

Choix (option type de segmentation)
    Option: paragraphe
        Segmentation en paragraphe.
    Option: Phrase
        Segmentation en Phrase.

Option: mot

Segmentation en Mot.

Option: mot particulier

Segmentation lors de l'atteinte du mot particulier.

Option: signe particulier

Segmentation lors de l'atteinte du signe particulier.

Fin choix

Déclaration de la liste de vecteurs (segments)

Segment = texte vide

J = 1

Pour i = 0 jusqu'au nombre de segments de texteASegmenter

Si i ++ <= Nombre d'occurrences souhaitées

Segment = Segment + texteASegmenter[i]

Si J>Nombre d'occurrences souhaitées

On ajoute Segment à la liste de vecteur et initialise J à 1

Fin si

Fin si

Fin pour

4.5 Extraction du vocabulaire

Afin de pouvoir utiliser les différents classifieurs, on doit établir une matrice d'entrée contenant les vecteurs à classifier (figure 17).

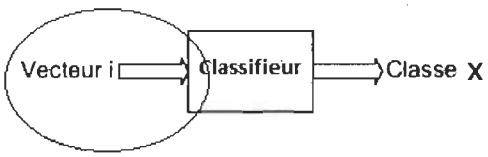


Figure 17 - Système de classification.

La matrice d'entrée (figure 18) se présente sous la forme d'un tableau de fréquence à deux dimensions où les colonnes représentent les différents éléments composant le texte à classifier (dimensions de la matrice) et les lignes, chaque vecteur d'occurrences produit lors de l'étape de segmentation.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	...
Vecteur 1	1	2	3	1	3	2	1	1	1	...
Vecteur 2	0	2	1	1	0	0	1	1	2	...
Vecteur 3	1	1	1	0	1	2	3	2	1	...
...	...	...	...	...	...	...	...	...	...	...

Figure 18 – Matrice d'entrée.

Dans l'application développée, les dimensions de la matrice (représentées ci-dessus par D1, D2, D3,...) peuvent contenir deux types d'informations différentes, soit des mots tels que nous les connaissons ou des n-grams. Un n-gram est un découpage de l'information en n caractères successifs. Nous optons pour ce deuxième type d'unité d'information car le mot ne convient pas à tous les types de langage tels que l'arabe, le chinois ou autre contrairement aux n-grams. (Par exemple, O.N.U. peut être considéré comme trois mots alors que dans les faits, il devrait être considéré comme un seul.)

Pour que cette adaptation soit réalisable, tous les vecteurs être représentés en fonction des mêmes données. Nous tiendrons donc compte du vocabulaire du texte dans sa totalité et calculerons par la suite la fréquence de chaque donnée au sein de chaque vecteur. La solution préconisée tiendra donc compte de l'ensemble du vocabulaire de tous les vecteurs.

4.6 Tables de distribution de fréquences relative et totale.

Afin de pouvoir bâtir la matrice d'entrées des classifieurs, nous devons construire au préalable deux tables de distribution des données de fréquences afin de connaitre la totalité du vocabulaire ainsi que la fréquence de chaque unité d'information décrivant chaque vecteur.

4.6.1 Table de distribution de fréquences relatives

Cette table nous permet de connaître la fréquence de chaque mot au sein de chaque vecteur. Elle contient le vocabulaire pour chacun des vecteurs et la fréquence de chaque unité d'information s'y retrouvant. La figure 19 nous montre un exemple de table de distribution de fréquences relatives.

Vecteur	Unité d'information	Fréquence
1	Jean	1
1	aime	1
...	...	...

Figure 19 - Table de distribution de fréquences relatives.

4.6.2 Table de distribution de fréquences totales

Cette table quant à elle, nous permet de connaître le vocabulaire dans son intégralité. Elle contient le vocabulaire du texte entier et la fréquence de chacune de ses unités d'information. La figure 20 nous montre un exemple de table de distribution de fréquences

Unité d'information	Fréquence
Jean	3
aime	1
…	…

Figure 20 - Table de distribution de fréquences totales.

4.6.3 Algorithme

Voici l'algorithme permettant de bâtir les tables de distributions de fréquences relative et totale.

```
Pour Chaque segment J
    Liste des unités d'information = vide
    Si Option mot/N-gram = N-gram
        Pour n = 0 jusqu'à dernier caractère du segment J
            Pour M = 1 jusqu'à nombre de caractères du N-gram – 1
                Ajouter  N-gram à la liste des unités d'information
            Fin pour
        Fin pour
    Sinon
        Liste des unités d'information = Segmentation en Mot.
    Fin si
    Pour chaque  mot/N-gram i de liste des unités d'information
        Si le mot/N-gram i pour le segment J n'existe pas dans la table de
        fréquences relatives
            On ajoute mot/N-gram i à la table de fréquences relatives
        Sinon
```

On incrémente son compteur

Fin si

Si le mot/N-gram i existe dans la table de fréquences totales n'existe pas

On ajoute mot/N-gram i à la table de fréquences totale

Sinon

On incrémente son compteur

Fin si

Fin pour

Fin pour

4.7 Nettoyage du vocabulaire

Appliquer une classification en considérant la totalité du vocabulaire donne souvent lieu à une matrice dont la taille est gigantesque, ce qui a pour conséquence, d'alourdir le processus de classification. Il est donc recommandé de procéder à un nettoyage du vocabulaire en éliminant plusieurs mots qui, dépendant de nos préférences, peuvent ne pas être pertinentes. On réduit ainsi la taille du vocabulaire.

4.7.1 Fonctionnalités

Les fonctionnalités diffèrent selon le type de vocabulaire.

1. Fonctionnalités pour un vocabulaire composé de mots, que nous avons:

   - Élimination des mots dont la taille en nombres de caractères est inférieure à un seuil établi par l'utilisateur.
   - Élimination des mots dont la fréquence est en dehors d'un intervalle établi par l'utilisateur qui ne sont pas porteurs de significations.

- Élimination des mots fonctionnels. Pour rappel, les mots fonctionnels sont des mots que l'on souhaite ne pas considérer lors de la comparaison. Il peut s'agir d'articles (la, le, les, ...) de pronoms (me, moi,..) ou de certains verbes (ait, aient,...).
- Élimination d'une liste précise de mots jugés non-pertinents.
- Garder une liste précise de mots jugés pertinents.
- Lemmatisation du texte. La lemmatisation permet de remplacer les mots fléchis par leur forme canonique. Par exemple, un verbe conjugué est remplacé par le verbe à l'infinitif, le féminin est remplacé par le masculin, etc.

2. Fonctionnalités pour un vocabulaire composé de n-grams :

- Élimination possible des n-grams composées de caractères blancs.
- Élimination possible des n-grams dont la fréquence est en dehors d'un intervalle établi.
- Élimination d'une liste précise de n-grams jugés non-pertinents.
- Garder une liste précise de n-grams jugés pertinents

4.7.2 Algorithme

Voici l'algorithme permettant de nettoyer le texte.

```
Si « lemmatisation »
    Pour chaque élément de la liste
        S'il existe dans la table de lemmatisation qui n'existe pas dans la liste
            Ajout du mot
        Fin Si
    Fin pour
```

Fin si

Si « Éliminer les mots plus petits »

Éliminer les mots ne respectant pas le nombre de caractères voulu *

Fin si

Si « Conserver l'intervalle »

Éliminer les mots plus petits et plus grands ne correspondant pas aux fréquences voulues *

Fin si

Si « Éliminer les mots fonctionnels»

Éliminer les mots fonctionnels selon le dictionnaire de mots fonctionnels.*

Fin si

Si « Éliminer la sélection »

Éliminer chaque mot sélectionné dans la liste.*

Fin si

Si « Éliminer les n-grams composées de caractères blancs»

Éliminer chaque mot qui n'est pas sélectionné dans la liste.*

Fin si

Si « Ne garder que la sélection »

Éliminer les n-grams composées de caractères blancs. *

Fin si * Les transformations sont faites à partir de fonctions déjà existantes.

4.8 Classification

Pour rappel, le but de la classification est de pouvoir regrouper les vecteurs ayant des similitudes en des groupes homogènes. Le logiciel offre un choix de trois classifieurs dont nous avons parlé dans le chapitre précédent, soit : K-means, SOM et ART.

4.8.1 Création de la matrice d'entrée

La matrice d'entrée, comme mentionnée au point 4.1.4, se présente sous la forme d'un tableau dont les dimensions correspondent à la taille du vocabulaire. On détermine les dimensions grâce à la colonne « unités d'informations » du tableau de distribution de fréquences totales (figure 21).

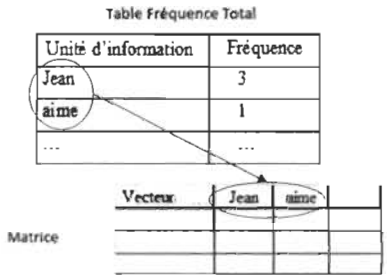


Figure 21 – Création de la matrice d'entrée.

Une fois la matrice créée, nous utilisons les informations fournies par la table de fréquences relatives afin de la compléter. Chaque ligne correspond à un vecteur et chaque donnée correspond au nombre d'occurrences de l'unité d'information au sein de ce vecteur (figure 22).

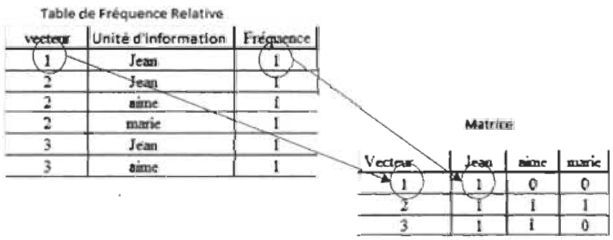


Figure 22 - Remplissage de la matrice d'entrée.

4.8.2 Algorithme

Voici l'algorithme permettant de créer la matrice d'entrée.

mi = nombre de vecteurs

mj = nb de mots de la matrice de fréquences totales

Matrice = tableau [mi X mj]

Pour chaque élément i du tableau de fréquences relatives

On va placer la fréquence de l'élément i dans le tableau Matrice selon le no de vecteur et le mot/N-gram correspondant.

Fin Pour

4.9 Normalisation de la matrice

Comme le poids des unités d'information d'un vecteur dépend de la taille du vecteur en termes d'unités d'information et non de la matrice d'entrée, une normalisation est nécessaire.

Pour effectuer la normalisation, nous commençons par uniformiser le poids de chaque vecteur (recalcul du poids de chaque unité d'information pour un total de un pour chaque ligne et ensuite, nous uniformisons le poids de chaque unité d'information (recalcul du poids de chaque unité d'information pour un total de un pour chaque colonne). Dans la figure 23, pour le mot « Jean » du vecteur 1, nous obtenons premièrement un poids égal à 1 (équation 10) et deuxièmement un poids égal à 0.55 (équation 11).

Matrice non normalisée

Vecteur/Mot	Jean	aime	marie
1	1	0	0
2	1	1	1
3	1	1	0

Vecteur/Mot	Jean	aime	marie
1	1	0	0
2	1/3	1/3	1/3
3	1/2	1/2	0

Matrice normalisée

Vecteur/Mot	Jean	aime	marie
1	0,55	0	0
2	0,18	0,4	1
3	0,27	0,6	0

Figure 23 - Normalisation de la matrice.

$$Poids = \frac{\text{nombre d'occurrences de l'unité d'information au sein du vecteur}}{\text{nombre d'occurrences du vecteur}} = \frac{1}{1} = 1$$

Équation 10 - Calcul du poids de l'occurrence du mot Jean au sein du vecteur 1

$$Poids = \frac{\text{Poids}}{\text{Somme des poids de la colonne}} = \frac{1}{1 + 0.66 + 0.5} = \frac{1}{1.83} = 0.55$$

Équation 11- Calcul du poids de l'occurrence du mot Jean au sein de la colonne « Jean »

4.9.1 Algorithme

Voici l'algorithme permettant de normaliser le texte.

Pour chaque vecteur i de la matrice,
    Calculer la somme des données de toutes les valeurs de toutes les colonnes (unités d'information).
    Pour chaque colonne j de la matrice,
        Matrice [i, j] = Matrice [i, j]/ la somme (si différente de 0)
    Fin Pour
Fin Pour
Pour chaque colonne j de la matrice,
    Calculer la somme de toutes les valeurs de l'unité d'information.
    Pour chaque vecteur i de la matrice,
        Matrice [i, j] = Matrice [i, j]/ la somme (si différente de 0)
    Fin Pour
Fin Pour

4.10 Règle d'association maximales

Afin de pouvoir utiliser les règles d'association maximales, il importe de bien définir l'ensemble de départ E ainsi que les sous-ensembles E1 et E2. Ensuite nous produisons les règles d'associations maximales.

4.10.1 Composition de l'ensemble E

L'ensemble E est l'union des mots présents dans toutes les classes issues de la classification.

Pour un vocabulaire composé de n-grams, une opération supplémentaire est nécessaire afin de bâtir l'ensemble E, car nous devons au préalable rebâtir les classes afin que leur contenu soit composé de mots. Pour chaque classe de n-grams, nous devons donc trouver quels sont les vecteurs associés à la classe lors de classification et ensuite extraire tous les mots contenant le n-gram.

4.10.1.1 Algorithme

Voici l'algorithme permettant de composer l'ensemble E.

Ensemble E = vide
Si l'unité d'information est le N-Gram
Pour chaque n-gram
Liste d'unités d'information = Mots contenant le n-gram au sein du segment dont il fait partie.
Pour chaque mot de la liste d'unités d'information
Ajouter le mot à l'ensemble E.
Fin pour

```
        Fin Pour
Sinon
        Tableau = tableau de distribution relatif
Fin si
Pour chaque vecteur dans Tableau
        Pour chaque élément dans vecteur,
                Si le mot n'existe pas dans l'ensemble E,
                        Ajouter le mot à l'ensemble E.
                Fin Si
        Fin pour
Fin pour
```

4.10.2 Composition des sous-ensembles E1 et E2

Les sous-ensembles E1 et E2 sont quant à eux des sous-ensembles de E. Pour que les règles d'association soient applicables à tout le vocabulaire d'un même texte et du fait que E2 et E1 doivent être disjoints, E1 doit être composé seulement du ou des mots dont nous cherchons à établir les relations. E1 sera donc toujours égal à X. Par conséquent, E2 sera le complémentaire de E1 dans E et sera déterminé lors du choix de X. Exemple, si nous supposons un ensemble E {et, Jean, Marie, se, regarde, tendrement, aime, déteste} formé par les mots des transactions {et, Jean, Marie, se, regarde, tendrement} et {Jean, Marie, aime, déteste} et que E1 est égal à {aime}, E2 sera donc égal à {Jean, Marie, et, se, regarde, tendrement, déteste}.

4.10.2.1 Algorithme pour déterminer E1 et X.

Voici l'algorithme permettant de composer l'ensemble E1 et X.

```
Si un choix a été effectué dans la liste
        Pour x = 0 jusqu'à la fin de la liste
                Si c'est un élément choisi
                        Ajouter l'élément à E1(X)
                Fin si
        Fin Pour
Fin si
```

4.10.2.2 Algorithme pour déterminer E2

Voici l'algorithme permettant de composer l'ensemble E2.

```
Pour chaque transaction
        Pour chaque mot de la transaction
                Si le mot existe dans E et n'existe pas dans E1
                        On le rajoute à E2 s'il n'existe pas.
                Fin Si
        Fin Pour
Fin Pour
```

4.10.3 Ensemble des règles d'associations maximales $X \xrightarrow{max} Y$:

4.10.3.1 Trouver les mots susceptibles de former X.

Comment nous l'avons vu dans la section précédente, X est toujours égal à l'ensemble E1. L'ensemble X ne doit pas être choisi de façon arbitraire. Afin de pouvoir cibler une valeur de X qui soit pertinente, il y a plusieurs possibilités, dont le choix des mots les plus fréquents dans le texte.

4.10.3.1.1 Algorithme

Voici l'algorithme permettant de trouver les mots susceptibles de former X.

```
Pour chaque élément dans E jusqu'au nombre requis
        Trouver le mot ayant la plus grande fréquence
        Ajouter mot
Fin pour
```

Si nous reprenons l'exemple 4.3.4.2.3 et définissons un nombre requis de un, vu la petitesse de l'exemple, le mot le plus fréquent est donc : aime. Donc dans ce cas, l'ensemble X serait : {aime}.

4.10.3.2 Ensemble Y

Y correspond à un ou plusieurs éléments de l'ensemble E2 répondant à la règle d'association maximale. Pour trouver toutes les combinaisons possibles, nous utilisons donc un algorithme récursif. La cardinalité de Y est déterminée par l'utilisateur.

4.10.3.2.1 Algorithme

Voici l'algorithme permettant de composer l'ensemble Y.

```
Pour niveau = 1 jusqu'au nombre de niveaux souhaité
        Pour  mot = 1 jusqu'au nombre de mots dans E2
                Y = trouverMotY (mot, 1)
        Fin pour
        Ajouter Y à la liste de Y
Fin pour

Procédure trouverMotY (mot, niveau)
        Si (niveau < niveau souhaité)
                Pour  mot = mot +1jusqu'au nombre de mot dans E2
                        trouverMotY (mot, niveau+1)
                Fin pour
        Fin si
        Ajouter mot à Y
```

Si nous reprenons l'ensemble X de l'exemple 4.3.4.2.3 avec un niveau de degré 2, il y aurait trois possibilités de Y soit {Jean}, {Marie}, {Jean, Marie}.

4.10.3.3 Notion de Msupport

4.10.3.3.1 Algorithme Msupport

Voici l'algorithme du calcul du Msupport.

```
Pour chaque groupe de mots dans transaction *
        Pour chaque mot de X
                Si le mot n'existe pas dans X
                        ContientTousX = faux
                Fin Si
        Fin pour
        Si ContientTousX
                Pour chaque mot de Y
                        Si le mot n'existe pas dans le groupe de mots
                                ContientTousY = faux
                                On quitte la boucle Pour
                        Fin si
                Fin pour
        Sinon
                ContientTousY = faux
        Fin si
        Si ContientTousY
                On incrémente Msupport
        Fin Si
Fin pour
* On suppose que la transaction contient X et Y
```

4.10.3.4 Notion de Mconfiance

4.10.3.4.1 Algorithme Mconfiance

Voici l'algorithme du calcul de la Mconfiance.

```
Pour chaque groupe de mots dans transaction *
        Pour chaque mot de X
                Si le mot n'existe pas dans X
                        ContientTousX = faux
                Fin Si
        Fin pour
        Si ContientTousX
                Pour chaque mot de E2
                        Si le mot existe dans le groupe de mots.
                                On incrémente motDeYContenuDansE2
                        Fin si
                Fin pour
        Fin si
        Si motDeYContenuDansE2 > 0
                Mconfiance = Msupport/ motDeYContenuDansE2 * 100
        Fin Si
Fin pour
* On suppose que Msupport est connu et qu'au départ, aucun mot de
  Y n'est contenu dans E2
```

4.10.3.5 Fonctionnalité

4.10.3.5.1 Sélection de l'union/intersection/différence du vocabulaire composant les classes selon une certaine fréquence.

Cette opération permet de sélectionner certains mots du vocabulaire afin d'assister au choix lors de la sélection de l'ensemble X. Par exemple, dans ce mémoire, nous avons axé notre choix sur l'intersection des mots les plus fréquents de toutes les classes afin de sélectionner nos ensembles X.

4.10.3.5.2 Algorithme

Voici l'algorithme du traitement de la sélection.

```
Pour chaque élément dans E
    Choix
        Cas : union de tout le vocabulaire
            Ajout du mot dans une nouvelle liste E
        Cas : union de l'intersection du vocabulaire
            S'il est présent dans tous les vecteurs
                Ajout du mot dans une nouvelle liste E
            Fin si
        Cas : union de différence de l'intersection du vocabulaire
            S'il n'est pas présent dans tous les vecteurs
                Ajout du mot dans une nouvelle liste E
            Fin si
    Fin choix
Fin pour
```

4.10.4 Fonctionnalité général du projet

4.10.4.1 Persistance des données.

Afin de pouvoir être en mesure d'apprécier les différentes analyses et de les modifier et/ou les réutiliser dans un temps futur, il s'est avéré nécessaire de développer un outil qui puisse sauvegarder les différents textes et les différentes analyses et ce, à n'importe quelle étape du processus. Des fonctions ont donc été développées afin de pouvoir ajouter un nouveau texte à la base de données, le modifier, le supprimer ou faire de nouvelles analyses (ou analyses complémentaires) sur des textes déjà traités.

4.10.4.1.1 Algorithme

Voici les algorithmes s'occupant de la persistance des données.

```
Ajout et enregistrement
    Capture du nom de sauvegarde
    Recherche dans la base de données Access si le nom existe
    Si le nom existe
        Envoie d'en message d'avertissement pour confirmer
        Si oui
            Effacement des enregistrements des tables.
            Enregistrement ()
        Fin Si
    Sinon
            Enregistrement ()
    Fin si
```

Effacer

Envoi d'en message d'avertissement pour savoir si on efface le texte existant

Si oui

Effacement des enregistrements des tables

Fin Si

4.11 Conclusion

Dans ce chapitre, Nous avons présenté la façon dont nous avons procédé pour développer notre méthode expérimentale.

L'application développé dans un environnement Microsoft Visual C#, permet de soumettre un texte entier (ou plusieurs) aux classifieurs K-Means, SOM et ART et d'interpréter les résultats obtenues grâce aux règles d'association maximales.

Le prochain chapitre sera consacré à l'expérimentation de la méthode expérimentale développée.

CHAPITRE 5

EXPÉRIMENTATION

5.1 Introduction

Le but de l'expérimentation est de démontrer la pertinence de la combinaison de la classification et des associations générées par les règles d'associations maximales. Pour ce faire, nous ferons non seulement ressortir le thème de chaque partie du livre, mais aussi les idées développées dans ce thème. Afin d'avoir des données rigoureuses et concluantes, l'expérimentation a été effectuée sur un livre complet de 1980 pages, soit « La civilisation des Arabes» [11].

5.2 Stratégie

Avant de pouvoir utiliser la classification (et les règles d'associations maximales), nous devons au préalable créer les différentes matrices d'entrées nécessaires à l'utilisation des classifieurs. Pour cela, nous devons effectuer plusieurs traitements sur le texte. La stratégie adoptée est la suivante.

Dans un premier temps, nous effectuons un traitement manuel sur le texte. Ce traitement consiste à diviser le livre en parties distinctes et en épurer le vocabulaire.

Dans un deuxième temps, nous créons une matrice d'entrées (requise lors de l'utilisation des classifieurs) pour chaque partie du livre. Pour ce faire, nous utilisons la démarche suivante :

- Segmentation de chaque partie.

- Extraction du vocabulaire selon deux types d'unités d'information soit les mots usuels, soit les tri-grams.
- Nettoyage du vocabulaire extrait pour chaque type d'unités d'information.
- Création des matrices d'entrées pour chaque type d'unités d'information.

Dans un troisième temps et une fois les matrices d'entrées créées, nous les classifions et en extrayons les règles d'associations maximales. Pour chaque matrice d'entrées, nous procédons de la manière suivante :

1. Classification de chacune d'entre elles grâce aux classifieurs K-Means, SOM et ART.
2. Analyse des systèmes de classes résultantes.
3. Extraction de quelques règles d'associations maximales.
4. Analyse des règles d'associations maximales.

Finalement, nous classifions les classes résultantes des méthodes de classification K-Means, SOM et ART de l'étape précédente et nous soumettons leurs résultats aux règles d'associations maximales.

5.3 Évaluation du livre « La civilisation des Arabes » [11].

5.3.1 Sommaire des opérations

Sous les sections suivantes, nous décrivons chacune des opérations.

5.3.2 Introduction du texte

Afin de pouvoir bien analyser le texte, certaines opérations sont faites au préalable soit :

- Division du livre en six parties afin de réduire le temps de traitement. Le choix de le diviser en six parties résulte du fait que le livre est déjà subdivisé en six livres distincts.
- Pour chacune des parties :
  - Suppression de la mention « Retour à la table des matières ».
  - Suppression des tableaux faisant référence à la figure du livre en édition papier de 1980 pages.
  - Transformation du texte en code Unicode utf-8
  - Élimination de la redondance, des numéros de page, chapitre, etc.

5.3.3 Conversion du texte

Nous faisons subir un prétraitement au texte de chaque partie :

- Conversion des lettres minuscules en lettres majuscules afin d'éliminer la redondance des mots.
- Remplacement des caractères spéciaux par des caractères blancs.
- Remplacement des caractères numériques par des caractères blancs.

5.3.4 Segmentation du texte

Le texte des six parties est segmenté en paragraphe distinct. Donc un vecteur est, dans notre cas, un paragraphe du texte.

5.3.5 Extraction du vocabulaire

Nous procédons à l'extraction du vocabulaire sous forme de mots dans un premier temps et sous forme de tri-grams dans un deuxième temps. Dans le cas des n-grams, nous avons opté pour un n égal à trois car nous pensons que des blocs de trois

caractères n'est ni trop évasif ni trop restreignant. En outre ce choix est fréquemment utilisé dans la littérature.

5.3.6 Nettoyage du vocabulaire

Le vocabulaire obtenu lors de l'extraction est souvent trop large. Il est donc important de le réduire afin d'améliorer l'efficacité des classifieurs. Il est important cependant de mentionner encore une fois que les choix établis ici sont propres à cette expérience. Voici les opérations de nettoyage effectuées pour les deux types de vocabulaire :

- Nettoyage lorsque l'unité d'information est le mot :
  - Supprimer les mots ayant moins de 3 caractères.
  - Élimination des mots fonctionnels. Les mots fonctionnels sont des mots présents dans une base de données annexée au logiciel (mf.mb).
  - Élimination des mots ayant une fréquence totale inférieure à 3.

- Nettoyage lorsque l'unité d'information est le tri-gram:
  - Suppression des tri-grams comportant des espaces blancs
  - Élimination des tri-grams ayant une fréquence totale inférieure à 3.

5.3.7 Classification

Une fois l'opération de nettoyage du vocabulaire effectué, les matrices d'entrées des classifieurs sont formées (voir chapitre précèdent). Ces matrices sont ensuite soumises aux classifieurs dans l'ordre suivant : ART, SOM, K-Means. La raison de cet ordre est très simple. Comme nous l'avons vu dans le chapitre deux, (les méthodes K-means et SOM, demandent en entrée, afin de pouvoir être utilisés, d'établir le

nombre de classes résultantes (la valeur K pour K-means et la grandeur de la carte pour SOM). L'algorithme ART n'ayant pas besoin de cette donnée est donc désigné afin de trouver le nombre de classes. ART requiert néanmoins de connaitre son taux d'apprentissage Rho que nous établissons à 0.001 afin de restreindre le nombre de classes (voir la section 2.3.4.5).

5.3.7.1 Analyse des résultats issus de la classification

Chaque partie du livre est analysée à partir des classes résultantes obtenues lors de la classification selon les méthodes de classification ART, SOM et K-Means. Les groupes de classes considérées sont les douze suivantes :

1. Les classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les mots avec K-Means.
2. Les classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les tri-grams avec K-Means.
3. L'union des classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les tri-grams et celles obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les mots avec K-Means.
4. Les classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les mots avec SOM.
5. Les classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les tri-grams avec SOM.
6. L'union des classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les tri-grams et celles obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les mots avec SOM.

7. Les classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les mots avec ART.
8. Les classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les tri-grams avec ART.
9. L'union des classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les tri-grams et celles obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les mots avec ART.
10. L'union des classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les mots avec K-Means, SOM et ART.
11. L'union des classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les tri-grams avec K-Means, SOM et ART.
12. L'union des classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les tri-grams et celles obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les mots avec K-Means, SOM et ART.

L'intersection des classes n'a pas été retenue, car dépendant des classifieurs utilisés et de l'unité d'information, les groupes de mots sont parfois vide.

Pour chacun de ces groupes, nous calculons les éléments suivants :

- Le nombre de segments trouvés.
- Le nombre de classes (transactions) issues de la classification.
- Le nombre de mots issus de l'union des classes résultantes.

L'identification du thème de chaque partie du livre est maintenant possible. Pour ce faire, nous utilisons les vingt mots les plus fréquents présents dans tous les groupes des classes résultantes énumérées précédemment.

5.3.8 Règle d'associations maximales.

Pour chaque partie du livre, nous extrayons un total de douze groupes de règles d'associations maximales.
Comme sous-ensemble X, on a choisi de garder les quatre mots les plus fréquents présents dans tous les groupes de classes résultantes que l'on retrouve au point 5.2.7. Nous avons opté pour ce choix afin de diminuer le nombre de règles d'associations maximales à traiter et afin de prendre en compte, les résultats de tous les classifieurs, ainsi que les deux types d'analyse avec mots et avec tri-grams. Nous optons pour un Msupport égal à 1 afin d'avoir au minimum un segment du texte contenant les mots de la règle d'association maximale et une Mconfiance de 50% que nous considérons comme un seuil minimum acceptable.

Les ensembles E utilisés résultent de l'union des mots issus de chaque groupe de classes résultantes que l'on retrouve au point 5.2.7.

5.3.8.1 Analyse des résultats issus du processus d'extraction des règles d'associations maximales.

Nous commençons par comptabiliser le nombre de règles pour tous les ensembles E que l'on retrouve à la section 5.2.8. Ensuite, nous calculons le nombre de règles d'associations maximales communes aux deux types d'analyse (avec mots et avec tri-grams) et le nombre de règles d'associations maximales communes aux deux types d'analyse (avec mots et avec tri-grams) et aux règles extraites lors de l'union des résultats des deux types d'analyse.

Nous procédons par la suite à l'analyse des règles extraites. Pour ce faire, Nous utilisons les règles d'association maximales formées à partir de l'ensemble E issu de l'union des classes résultantes obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les tri-grams et celles obtenues lors de la classification de la matrice composée de l'unité d'information basée sur les mots avec K-Means, SOM et ART. Nous utilisons ce choix afin de prendre en compte tous les classifieurs d'une part et minimiser l'impact des choix arbitraires pris lors du nettoyage du vocabulaire.

5.4 Résultats obtenus lors l'analyse du livre « La civilisation des Arabes » [11].

5.4.1 Partie 1

Dans cette section nous présentons les résultats de la partie 1.

5.4.1.1 Résultats des analyses lors du processus de classification

Dans le Tableau 2 nous présentons une synthèse des résultats. Les colonnes « Mots » et « Tri-Grams » représentent les résultats obtenus à la suite des classifications des matrices d'entrées basées sur les unités d'information mot et tri-gram. La colonne « Mots U Tri-Grams » représente les résultats obtenues lorsque nous unissons les résultats des deux types d'unités d'information (mots et tri-grams). Les unités d'information sont toutes classées en utilisant ART, SOM et K-Means. Pour chaque unité d'information, la colonne « Classes » représente le nombre de classes issu des résultats de la classification. La colonne « Mots » représente le nombre de mots obtenus lorsque nous procédons à l'unification de toutes les classes résultantes. La colonne « segments » représente le nombre de segments obtenu lors de l'opération de segmentation.

		Unité d'information					
Classifieurs	Segments	Mots		Tri-Grams		Mots U Tri-Grams	
		Classes	Mots	Classes	Mots	Classes	Mots
ART	335	48	1025	33	4424	81	4424
SOM	335	48	1025	32	4424	80	4424
K-means	335	48	1025	33	4424	81	4424
Total	1005	144	1025	98	4424	242	4424

Tableau 2- Résultats des analyses lors du processus de classification de la partie 1.

Le nombre de segments à classifier (en l'occurrence, en ce qui nous concerne, des paragraphes) nous donne un aperçu de la taille du texte. Nous pouvons aussi constater que le nombre de mots obtenu lors de la classification lorsque l'unité d'information est le tri-gram est quatre fois supérieur à celui obtenu lors de la classification lorsque l'unité d'information est le mot. C'est une conséquence des choix que nous avons pris afin de réduire la taille de la matrice lors de l'étape de nettoyage. Le nombre de mots total lorsque nous faisons l'union des résultats de tous les classifieurs est identique au nombre de mots le plus élevé. Ce qui est normal car nous ne créons pas de mots comme c'est le cas pour les classes.

Pour identifier le thème de la partie du livre, nous utilisons l'ensemble des vingt mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). (242 classes).

Ce groupe de mots est le suivant : (ARABES, ARABIE, FAIT, HISTOIRE, PAYS, ARABE, RACE, FAUT, NOMADES, CIVILISATION, PEUPLE, PEUPLES, VILLES, RESTE, MONDE, YÉMEN, GRANDE, MAHOMET, DIT, DIRE)
Si nous regardons ce groupe de mots, nous pouvons dire qu'il s'agit de l'histoire des arabes mais sans toutefois interpréter spécifiquement le thème. Pour vérifier notre interprétation, regardons la table des matières du livre [11] que voici :

Livre premier : Le milieu et la race

Chapitre I L'Arabie

Chapitre II Les Arabes

Chapitre III Les Arabes avant Mahomet

Bien que les mots « ARABES », « ARABIE », « RACE » et « MAHOMET » apparaissent dans le groupe de mots, la relation entre eux n'est pas évidente quand pris ils sont séparément.

5.4.1.2 Résultats obtenus lors de l'extraction des règles d'associations maximales

Comme ensemble E, nous utilisons un ensemble de 4424 mots. Cet ensemble correspond à la classe la plus large obtenue à la suite de l'unification des classes résultantes des classifications avec ART, K-means et SOM (voir tableau 1 page 68). Nous avons choisi cet ensemble car nous croyons qu'un ensemble E plus large permet d'extraire un plus grand nombre de règles d'associations maximales.

Nous choisissons pour les besoins d'extraction des règles d'association maximales, les quatre mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). Nous établissons donc le sous-ensemble X qui est {ARABES, ARABIE, FAIT, HISTOIRE} duquel nous enlevons le verbe FAIT. Nous avons choisi volontairement d'enlever le verbe FAIT car nous croyons qu'il est un mot fonctionnel n'ayant pas été enlevé lors du nettoyage du vocabulaire. Ceci nous donne donc : X = {ARABES, ARABIE, HISTOIRE}. Bien entendu, d'autres façons de construire l'ensemble X sont possibles.

Dans le tableau 3 nous présentons une synthèse quant au nombre de règles d'association maximales extraites selon la règle {ARABES, ARABIE, HISTOIRE} $\xrightarrow{max}$ Y avec un

seuil minimum de Mconfiance de 50%. Les colonnes « Mots », « Tri-Grams », « Mots U tri-grams », « Mot ∩ Tri-grams » et « Mot ∩ Tri-grams ∩ (Mot U Tri-grams)" représentent le nombre de règles d'association maximales extraites lors de l'utilisation du groupe de classes du classifieur correspondant selon le type d'unité d'information voulu.

Classifieurs	Unité d'information				
	Mot	Tri-gram	Mot U Tri-grams	Mot ∩ Tri-grams	Mot ∩ Tri-grams ∩ (Mot U Tri-grams)
ART	6	10	7	5	5
SOM	7	17	9	7	7
K-MEANS	5	7	5	5	5
A+S+K	6	11	7	6	6

Tableau 3 - Résultats obtenus lors de l'extraction des règles d'associations maximales de la partie 1.

Nous pouvons observer que le nombre de règles d'associations maximales extraites est plus élevé lorsque nous utilisons les tri-grams. Nous pouvons aussi constater que plus de 86% des ensembles Y se retrouvent à la fois dans l'analyse basée sur les mots et l'analyse basée sur les Tri-grams. Finalement dans 100% des cas, lorsqu'un ensemble extrait est présent à la fois dans l'analyse basée sur l'unité d'information « mot » et dans l'analyse basée sur l'unité d'information « tri-gram », il est aussi présent lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ».

Les règles choisies sont les règles extraites lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ». En effet, cette colonne est la plus représentative car elle contient à la fois tous les ensembles présents à la fois lors de l'analyse basée sur les mots et l'analyse basée sur les tri-grams mais aussi des ensembles qui pourraient être pertinents une fois les analyses combinées. Nous avons donc sept règles d'associations maximales que nous présentons dans le tableau 4.

Règles d'association maximales	Mconfiance
{ARABES, ARABIE, HISTOIRE} $\xrightarrow{max}$ {ARABE}	70%
{ARABES, ARABIE, HISTOIRE} $\xrightarrow{max}$ {NOMADE}	67%
{ARABES, ARABIE, HISTOIRE} $\xrightarrow{max}$ {PEUPLES}	65%
{ARABES, ARABIE, HISTOIRE} $\xrightarrow{max}$ {FAIT}	60%
{ARABES, ARABIE, HISTOIRE} $\xrightarrow{max}$ {SIÈCLES}	56%
{ARABES, ARABIE, HISTOIRE} $\xrightarrow{max}$ {FAUT}	53%
{ARABES, ARABIE, HISTOIRE} $\xrightarrow{max}$ {TRIBUS}	52%

Tableau 4 - Règles d'associations maximales extraites de la partie 1.

Si nous éliminons les mots « fait » et « faut » qui sont, à notre avis, des mots fonctionnels, le système de règles d'associations maximales associe fortement {ARABES, ARABIE, HISTOIRE} à {ARABE, NOMADE, PEUPLES, SIÈCLES et TRIBUS}. Ce qui nous donne une idée beaucoup plus précise sur le thème.

Afin de vérifier l'exactitude de cette affirmation, voici en exemple, un extrait de la première partie du livre « La civilisation des arabes » [11] : « Les Arabes nomades ont toujours vécu, comme aujourd'hui, en petites tribus placées sous l'autorité patriarcale d'un chef, nomme cheik ou seigneur, qui est un des chefs de famille de la tribu». Les règles d'associations maximales nous aident donc à l'identification du thème de cette partie.

5.4.2 Partie 2

Dans cette section nous présentons les résultats de la partie 2.

5.4.2.1 Résultats des analyses lors du processus de classification

Voici la synthèse des opérations des trois types d'analyses obtenue suite à l'opération de classification avec ART, SOM, K-Means de la partie 2 (tableau 5).

		Unité d'information					
Classifieurs	Segments	Mots		Tri-Grams		Mots U Tri-Grams	
		Classes	Mots	Classes	Mots	Classes	Mots
ART	321	62	572	27	3075	89	3075
SOM	321	62	572	28	3075	90	3075
K-means	321	62	572	28	3075	90	3075
Total	963	186	572	83	3075	269	3075

Tableau 5- Résultats des analyses lors du processus de classification de la partie 2.

Comme lors de l'analyse de la première partie, le nombre de segments à classer nous donne un aperçu de la taille du texte. Nous pouvons aussi constater que le nombre de mots obtenu lors de la classification des tri-grams est approximativement six fois supérieur à celui obtenu lors de la classification des mots. Si nous comparons avec la première partie où le nombre est quatre fois supérieur, nous pouvons expliquer ce fait comme étant la cause des choix faits lors du nettoyage.

Pour identifier le thème de la partie du livre, nous utilisons l'ensemble des vingt mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). (269 classes).

Ce groupe de mots est le suivant : (MAHOMET, DIEU, PROPHÈTE, ARABES, CORAN, MONDE, HOMMES, JOUR, HOMME, FAIRE, VIE, ANS, RELIGION, EMPIRE, FAIT, ARABE, ISLAMISME, CHRÉTIENS, PUISSANCE, HÉGIRE)

Si nous regardons ce groupe de mots, nous pouvons dire qu'il est question de Dieu, du prophète Mahomet et du Coran sans toutefois pouvoir interpréter spécifiquement le thème. Pour vérifier notre interprétation, regardons la table des matières du livre [11] que voici :

Livre deuxième : Les origines de la civilisation arabe
Chapitre I Mahomet. Naissance de l'empire arabe.
Chapitre II Le Coran
Chapitre III Les conquêtes des Arabes.

On peut observer que l'interprétation n'est pas évidente même si les mots « MAHOMET », « EMPIRE », « ARABE » et « CORAN » sont présents.

5.4.2.2 Résultats obtenus lors de l'extraction des règles d'associations maximales

Comme ensemble E, nous utilisons un ensemble de 3075 mots. Cet ensemble correspond à la classe la plus large obtenue à la suite de l'unification résultantes des classifications avec ART, K-means et SOM (voir tableau 4 page 73).

Nous choisissons pour les besoins d'extractions des règles d'association maximales les quatre mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). Ceci nous donne donc : X = {MAHOMET, DIEU, PROPHÈTE, ARABES}. Bien entendu, d'autres façons de construire X sont possibles.

Dans le tableau 6 nous présentons une synthèse quant au nombre de règles d'association maximales extraites selon la règle {MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ Y avec un seuil minimal de Mconfiance est de 50%.

Classifieurs	Unité d'information				
	Mot	Tri-gram	Mot U Tri-grams	Mot ∩ Tri-grams	Mot ∩ Tri-grams ∩ (Mot U Tri-grams)
ART	42	40	31	23	23
SOM	7	10	10	5	5
K-MEANS	10	19	13	7	7
A+S+K	9	16	11	8	8

Tableau 6- Résultats obtenus lors de l'extraction des règles d'associations maximales de la partie 2.

Nous pouvons observer que le nombre de règles d'associations maximales extraites est généralement plus élevé lorsque nous utilisons les tri-grams. Nous pouvons aussi constater que pour le classifieur SOM, plus de 50% des ensembles Y se retrouve à la fois dans l'analyse basé sur les mots et l'analyse basée sur les Tri-grams. Ceci est dû au fait que nous ne conservons que les résultats dont la Mconfiance est plus de 50%. Finalement dans 100% des cas, lorsqu'un ensemble extrait est présent à la fois dans l'analyse basée sur l'unité d'information « mot » et dans l'analyse basée sur l'unité d'information « tri-gram », il est aussi présent lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ».

Les règles choisies sont les règles extraites lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ». Nous avons donc onze règles d'associations maximales que nous présentons dans le tableau 7.

Règles d'association maximales	Mconfiance
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {HOMME}	71%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {HOMMES}	69%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {FAIRE}	67%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {MONDE}	65%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {VIE}	64%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {JOUR}	62%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {ANS}	60%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {EMPIRE}	55%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {MECQUE}	53%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {PUISSANCE}	53%
{MAHOMET, DIEU, PROPHÈTE, ARABES} $\xrightarrow{max}$ {JUSTICE}	51%

Tableau 7 - Règles d'associations maximales extraites de la partie 2

Si nous éliminons le mot « faire « qui, à notre avis est un mot fonctionnel, le système de règles d'associations maximales associe fortement {MAHOMET, DIEU, PROPHÈTE, ARABES} à {HOMMES, HOMME, MONDE, VIE, SIÈCLES, JOUR, ANS, EMPIRE, MECQUE, PUISSANCE et JUSTICE}. Ce qui nous donne une idée plus clair du thème consacré.

Afin de vérifier l'exactitude de cette affirmation, voici en exemple, un extrait de la deuxième partie du livre « La civilisation des arabes » [11]: « Nous nous sommes surtout occupé dans ce qui précède de la vie publique de Mahomet. Il nous reste maintenant à essayer de reconstituer le caractère et la vie privée du prophète, d'après les documents que

les Arabes nous ont laissés. Dans cet extrait, nous pouvons constater que le texte parle de la vie de Mahomet en tant qu'homme.

Les règles d'associations maximales nous aident donc à l'identification du thème de cette partie.

5.4.3 Partie 3

Dans cette section nous présentons les résultats de la partie 3.

5.4.3.1 Résultats des analyses lors du processus de classification

Voici la synthèse des opérations des trois types d'analyses obtenue suite à l'opération de classification avec ART, SOM, K-Means de la partie 3 (tableau 8).

		Unité d'information					
Classifieurs	Segments	Mots		Tri-Grams		Mots U Tri-Grams	
		Classes	Mots	Classes	Mots	Classes	Mots
ART	695	72	1089	64	7410	136	7410
SOM	695	72	1089	63	7410	135	7410
K-means	695	72	1089	64	7410	136	7410
Total	2085	216	1089	83	7410	407	7410

Tableau 8 - Résultats des analyses lors du processus de classification de la partie 3

Comme lors de l'analyse de la première partie, le nombre de segments à classer nous donne un aperçu de la taille du texte. Nous pouvons aussi constater que le nombre de mots obtenu lors de la classification des tri-grams est approximativement sept fois supérieur à celui obtenu lors de la classification des mots. Nous pouvons expliquer ce fait comme étant la cause des choix faits lors du nettoyage.

Pour identifier le thème de la partie du livre, nous utilisons l'ensemble des vingt mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). (407 classes).

Ce groupe de mots est le suivant : (ARABES, ARABE, FAIT, GRANDE, MONUMENTS, ÉPOQUE, EUROPE, INFLUENCE, VILLE, FAIRE, SIÈCLES, ÉGYPTE, SIÈCLE, HISTOIRE, ESPAGNE, ASSEZ, CIVILISATION, RESTE, MONDE, CHRÉTIENS)

Si nous regardons ce groupe de mots, nous pouvons supposer que le thème traite de différents pays et des arabes. Pour vérifier notre interprétation, regardons la table des matières de cette partie du livre [11] que voici :

Livre troisième : L'empire des Arabes

Chapitre I Les Arabes en Syrie

Chapitre II Les Arabes à Bagdad

Chapitre III Les Arabes en Perse et dans L'Inde

Chapitre IV Les Arabes en Égypte

Chapitre V Les Arabes dans l'Afrique septentrionale

Chapitre VI Les Arabes en Espagne

Chapitre VII Les Arabes en Sicile, en Italie et en France

Chapitre VIII Luttes du christianisme contre l'islamisme. Les croisades.

Les mots « ARABES », « ÉGYPTE » et « ESPAGNE » apparaissent dans le groupe de mots ». Le mot « ARABES » est d'ailleurs omniprésent. Ce que nous ne pouvions savoir en analysant le groupe de mots mais il traite néanmoins de différents pays.

5.4.3.2 Résultats obtenus lors de l'extraction des règles d'associations maximales

Comme ensemble E, nous utilisons un ensemble de 7410 mots. Cet ensemble correspond à la classe la plus large obtenue à la suite de l'unification des classifications avec ART, K-means et SOM (voir tableau 7 page 77).

Nous choisissons pour les besoins d'extractions des règles d'association maximales les quatre mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). Nous établissons donc le sous-ensemble X qui est {ARABES, ARABE, GRANDE} (le mot FAIT a été retiré, étant selon notre avis, un mot fonctionnel). Bien entendu, d'autres façons de construire X sont possibles.

Dans le tableau 9 nous présentons une synthèse quant au nombre de règles d'association maximales extraites selon la règle {ARABES, ARABE, GRANDE} $\xrightarrow{max}$ Y avec un seuil minimal de Mconfiance est de 50%.

Classifieurs	Unité d'information				
	Mot	Tri-gram	Mot ∪ Tri-grams	Mot ∩ Tri-grams	Mot ∩ Tri-grams ∩ (Mot ∪ Tri-grams)
ART	24	37	29	23	23
SOM	14	22	17	14	14
K-MEANS	11	11	10	7	7
A+S+K	16	22	17	16	16

Tableau 9 - Résultats obtenus lors de l'extraction des règles d'associations maximales de la partie 3

Nous pouvons observer que le nombre de règles d'associations maximales extraites est généralement plus élevé lorsque nous utilisons les tri-grams. Nous pouvons aussi

constater que pour le classifieur ART, plus de 62% des ensembles Y se retrouve à la fois dans l'analyse basé sur les mots et l'analyse basée sur les Tri-grams. Finalement dans 100% des cas, lorsqu'un ensemble extrait est présent à la fois dans l'analyse basée sur l'unité d'information « mot » et dans l'analyse basée sur l'unité d'information « tri-gram », il est aussi présent lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ».

Les règles choisies sont les règles extraites lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ». Nous avons donc dix-sept règles d'associations maximales que nous présentons dans le tableau 10.

Règles d'association maximales	Mconfiance
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {FAIT}	71%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {SIÈCLE}	66%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {INFLUENCE}	64%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {ESPAGNE}	61%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {VILLE}	60%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {NOM}	60%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {FAIRE}	58%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {MONDE}	58%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {RESTE}	58%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {PAYS}	57%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {CIVILISATION}	57%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {PARTIE}	55%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {ABORS}	55%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {KHALIFE}	53%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {CONQUÊTE}	52%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {JOUR}	50%
{ARABES, ARABE, GRANDE} $\xrightarrow{max}$ {FIT}	50%

Tableau 10 - Règles d'associations maximales extraites de la partie 3

Si nous éliminons les mots « fait » et « fit » qui, à notre avis sont des mots fonctionnels, le système de règles d'associations maximales associe fortement {ARABES, ARABE, GRANDE} à {SIÈCLE, INFLUENCE, ESPAGNE, VILLE, NOM, MONDE, RESTE, PAYS, CIVILISATION, PARTIE, ABORD, KHALIFE, CONQUÊTE et JOUR}. Nous pouvons donc présager que cet état se reflète dans cette partie du livre.

Afin de vérifier l'exactitude de cette affirmation, voici en exemple, un extrait de la troisième partie du livre « La civilisation des arabes » [11]:

« En consacrant ce chapitre et ceux qui vont suivre à l'étude des Arabes dans les divers pays qu'ils ont occupés, nous voulons donner d'abord une idée générale de leur civilisation, et montrer leur influence sur les peuples avec lesquels ils se sont trouvés en contact ainsi que celle exercée sur eux par ces derniers. Ce sera surtout par l'examen des œuvres laissées dans chaque contrée par les Arabes que nous essaierons d'apprécier leur civilisation. Après cette vue d'ensemble, il sera plus facile d'aborder ensuite un à un dans divers chapitres, les éléments variés dont la réunion constitue une civilisation.

Lorsque les Arabes s'établirent dans les diverses contrées de l'Asie, de l'Afrique et de l'Europe qui contribuèrent à former leur gigantesque empire, ils y rencontrèrent des peuples arrivés à tous les degrés de la civilisation, depuis la demi-barbarie, comme dans certaines parties de l'Afrique, jusqu'à la civilisation grecque et latine la plus avancée, comme en Syrie.»

Dans cet extrait, nous pouvons constater que le texte parle de la civilisation des arabes et de leur influence sur les peuples avec lesquels ils se sont trouvés en contact lors de leur conquête.

Les règles d'associations maximales nous aident donc à l'identification du thème de cette partie.

5.4.4 Partie 4

Dans cette section nous présentons les résultats de la partie 4.

5.4.4.1 Résultats des analyses lors du processus de classification

Voici la synthèse des opérations des trois types d'analyses obtenue suite à l'opération de classification avec ART, SOM, K-Means de la partie 4 (tableau 11).

		Unité d'information					
Classifieurs	Segments	Mots		Tri-Grams		Mots U Tri-Grams	
		Classes	Mots	Classes	Mots	Classes	Mots
ART	351	57	1056	39	4840	96	4840
SOM	351	56	1056	36	4840	92	4840
K-means	351	56	1056	39	4840	95	4840
Total	1053	169	1056	114	4840	283	4840

Tableau 11 - Résultats des analyses lors du processus de classification de la partie 4

Comme lors de l'analyse de la première partie, le nombre de segments à classer nous donne un aperçu de la taille du texte. Nous pouvons aussi constater que le nombre de mots obtenu lors de la classification des tri-grams est approximativement quatre fois supérieur à celui obtenu lors de la classification des mots. Nous pouvons expliquer ce fait comme étant la cause des choix faits lors du nettoyage.

Pour identifier le thème de la partie du livre, nous utilisons l'ensemble des vingt mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). (283 classes).

Ce groupe de mots est le suivant : (ARABES, FEMMES, FAIT, VIE, ORIENT, CORAN, PEUPLES, ORIENTAUX, EUROPE, RESTE, HOMMES, GRANDE, MAHOMET, INFLUENCE, FAIRE, DIT, GRAND, FAMILLE, ENFANTS, HOMME)

Si nous regardons ce groupe de mots, nous pouvons supposer que le thème de la quatrième partie du livre traite de la vie des femmes arabes, leurs enfants et leurs familles en orient. Pour vérifier notre interprétation, regardons la table des matières du livre [11] que voici :

Chapitre quatrième : Les mœurs et les institutions des Arabes.

Chapitre I Les Arabes nomades et Arabes sédentaires des campagnes.

Chapitre II Les Arabes des villes. - Mœurs et coutumes.

Chapitre III Institutions politiques et sociales des Arabes

Chapitre IV Les femmes en Orient.

Chapitre V Religion et morale.

Nous voyons que nous avons cerné une partie du thème. (Les mots « ARABES », « FEMMES » et « ORIENT » apparaissent dans le groupe de mots.)

5.4.4.2 Résultats obtenus lors de l'extraction des règles d'associations maximales

Comme ensemble E, nous utilisons un ensemble de 4840 mots. Cet ensemble correspond à la classe la plus large obtenue à la suite de l'unification des classifications avec ART, K-means et SOM (voir tableau 10 page 82).

Nous choisissons pour les besoins d'extractions des règles d'association maximales les quatre mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). Nous établissons donc le sous-ensemble X qui est {ARABES, FEMMES, VIE} (le mot FAIT a été retiré, étant

selon notre avis, un mot fonctionnel). Bien entendu, d'autres façons de construire X sont possibles.

Dans le tableau 12 nous présentons une synthèse du nombre de règles d'association maximales extraites selon la règle {ARABES, FEMMES, VIE} $\xrightarrow{max}$ Y avec un seuil minimal de Mconfiance est de 50%

Classifieurs	Unité d'information				
	Mot	Tri-gram	Mot U Tri-grams	Mot ∩ Tri-grams	Mot ∩ Tri-grams ∩ (Mot U Tri-grams)
ART	3	5	3	3	3
SOM	2	5	2	2	2
K-MEANS	2	5	2	2	2
A+S+K	3	4	3	3	3

Tableau 12 - Résultats obtenus lors de l'extraction des règles d'associations maximales de la partie 4.

Nous pouvons observer que le nombre de règles d'associations maximales extraites est généralement plus élevé lorsque nous utilisons les tri-grams. Nous pouvons aussi constater que 100% des ensembles Y se retrouve à la fois dans l'analyse basé sur les mots et l'analyse basée sur les Tri-grams. Finalement dans 100% des cas, lorsqu'un ensemble extrait est présent à la fois dans l'analyse basée sur l'unité d'information « mot » et dans l'analyse basée sur l'unité d'information « tri-gram », il est aussi présent lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ».

5.4.4.2.1 Règles d'association maximales extraites.

Les règles choisies sont les règles extraites lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ». Nous avons donc trois règles d'associations maximales que nous présentons dans le tableau 13.

Règles d'association maximales	Mconfiance
{ARABES, FEMMES, VIE} $\xrightarrow{max}$ {ORIENT}	72%
{ARABES, FEMMES, VIE} $\xrightarrow{max}$ {FAMILLE}	63%
{ARABES, FEMMES, VIE} $\xrightarrow{max}$ {HOMMES}	57%

Tableau 13 - Règles d'associations maximales extraites de la partie 4.

Le système de règles d'associations maximales associe fortement {ARABES, FEMMES, VIE} à {ORIENT, FAMILLE, HOMMES}. Nous pouvons donc présager que le sujet de la partie du livre traite de la vie des femmes arabes en rapport avec l'Orient, la famille et les hommes.

Afin de vérifier l'exactitude de cette affirmation, voici en exemple, un extrait de la quatrième partie du livre La civilisation des arabes [11] : « L'influence du climat et de la race est trop évidente pour qu'il soit besoin d'insister. La constitution physiologique de la femme, la nécessité de la maternité, ses maladies, etc., l'obligeant à rester souvent éloignée de son mari, et ce veuvage momentané étant impossible sous le climat de l'Orient et avec le tempérament des Orientaux, la polygamie était absolument nécessaire. Dans cet extrait, nous pouvons constater que le texte parle de la vie de de la femme en Orient.

Les règles d'associations maximales nous aident donc à l'identification du thème de cette partie.

5.4.5 Partie 5

Dans cette section nous présentons les résultats de la partie 5.

5.4.5.1 Résultats des analyses lors du processus de classification

Voici la synthèse des opérations des trois types d'analyses obtenue suite à l'opération de classification avec ART, SOM, K-Means de la partie 5 (tableau 14).

		Unité d'information					
Classifieurs	Segments	Mots		Tri-Grams		Mots U Tri-Grams	
		Classes	Mots	Classes	Mots	Classes	Mots
ART	588	89	1472	54	5715	143	5715
SOM	588	86	1472	54	5715	140	5715
K-means	588	89	1472	54	5715	143	5715
Total	1764	264	1472	162	5715	426	5715

Tableau 14 - Résultats des analyses lors du processus de classification de la partie 5

Comme lors de l'analyse de la première partie, le nombre de segments à classer nous donne un aperçu de la taille du texte. Nous pouvons aussi constater que le nombre de mots obtenu lors de la classification des tri-grams est approximativement quatre fois supérieur à celui obtenu lors de la classification des mots. Nous pouvons expliquer ce fait comme étant la cause des choix faits lors du nettoyage.

Pour identifier le thème de la partie du livre, nous utilisons l'ensemble des vingt mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). (426 classes).

Ce groupe de mots est le suivant : (ARABES, ARABE, SIÈCLE, FAIT, ESPAGNE, ÉPOQUE, EUROPE, GRANDE, OEUVRES, ŒUVRES, ANCIENS, INFLUENCE, DIT, MONUMENTS, DIRE, ART, ÉGYPTE, OUVRAGE, SIÈCLES, SCIENCE)

Si nous regardons ce groupe de mots, nous pouvons supposer que le thème de la cinquième partie du livre traite des œuvres, des monuments, de l'art et de la science. Il faut aussi noter que notre interprétation peut être biaisée vue que nous connaissons le livre. Pour vérifier notre interprétation, regardons la table des matières du livre [11] que voici :

Livre cinquième : La civilisation des Arabes.

Chapitre I	Origine des connaissances des Arabes. Leur enseignement et leurs méthodes.
Chapitre II	Langue, philosophie, littérature et histoire.
Chapitre III	Mathématiques et astronomie.
Chapitre IV	Sciences géographiques.
Chapitre V	Sciences physiques et leurs applications
Chapitre VI	Science naturelles et médicales
Chapitre VII	Les arts Arabes. Peinture, sculpture, arts industriels.
Chapitre VIII	L'architecture des Arabes.
Chapitre IX	Commerce des Arabes. - Leur relation avec divers pays.
Chapitre X	Civilisation de l'Europe par les Arabes. Leur influence en Occident et en Orient

Nous voyons que nous avons cerné approximativement une partie du thème. En effet, l'ensemble ne laisse pas supposer qu'il s'agit de la civilisation arabe.

5.4.5.2 Résultats obtenus lors de l'extraction des règles d'associations maximales

Comme ensemble E, nous utilisons un ensemble de 5715 mots. Cet ensemble correspond à la classe la plus large obtenue à la suite de l'unification des classifications avec ART, K-means et SOM (voir tableau 13 page 86).

Nous choisissons pour les besoins d'extractions des règles d'association maximales les quatre mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). Nous établissons donc le sous-ensemble X qui est {ARABES, SIÈCLE, ARABE} (le mot FAIT a été retiré, étant selon notre avis, un mot fonctionnel). Bien entendu, d'autres façons de construire X sont possibles.

Dans le tableau 15 nous présentons une synthèse quant au nombre de règles d'association maximales extraites selon la règle {ARABES, SIÈCLE, ARABE} $\xrightarrow{\max}$ Y avec un seuil minimal de Mconfiance est de 50%.

Classifieurs	Unité d'information				
	Mot	Tri-gram	Mot U Tri-grams	Mot ∩ Tri-grams	Mot ∩ Tri-grams ∩ (Mot U Tri-grams)
ART	15	43	20	15	15
SOM	7	14	10	5	5
K-MEANS	7	10	10	5	5
A+S+K	9	20	9	9	9

Tableau 15 - Résultats obtenus lors de l'extraction des règles d'associations maximales de la partie 5

Nous pouvons observer que le nombre de règles d'associations maximales extraites est généralement plus élevé lorsque nous utilisons les tri-grams. Nous pouvons aussi constater que vu le peu de règles extraites lors de l'analyse basée sur les mots, le nombre de règles se retrouvant à la fois dans l'analyse basé sur les mots et l'analyse basée sur les Tri-grams est relativement bas. Finalement dans 100% des cas, lorsqu'un ensemble extrait est présent à la fois dans l'analyse basée sur l'unité d'information « mot » et dans l'analyse basée sur l'unité d'information « tri-gram », il est aussi présent lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ».

Les règles choisies sont les règles extraites lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ». Nous avons donc neuf règles d'associations maximales que nous présentons dans le tableau 15.

Règles d'association maximales	Mconfiance
{ARABES, SIÈCLE, ARABE} $\xrightarrow{max}$ {FAIT}	64%
{ARABES, SIÈCLE, ARABE} $\xrightarrow{max}$ {ESPAGNE}	63%
{ARABES, SIÈCLE, ARABE} $\xrightarrow{max}$ {ANCIENS}	62%
{ARABES, SIÈCLE, ARABE} $\xrightarrow{max}$ {ÉPOQUE}	60%
{ARABES, SIÈCLE, ARABE} $\xrightarrow{max}$ {GRANDE}	59%
{ARABES, SIÈCLE, ARABE} $\xrightarrow{max}$ {EUROPE}	59%
{ARABES, SIÈCLE, ARABE} $\xrightarrow{max}$ {INFLUENCE}	56%
{ARABES, SIÈCLE, ARABE} $\xrightarrow{max}$ {DIT}	56%
{ARABES, SIÈCLE, ARABE} $\xrightarrow{max}$ {MONUMENTS}	56%

Tableau 16 - Règles d'associations maximales extraites de la partie 5

Si nous supprimons le mot « fait » qui, à notre avis est un mot fonctionnel, le système de règles d'associations maximales associe fortement {ARABES, SIÈCLE, ARABE} à {ESPAGNE, ANCIENS, ÉPOQUE, GRANDE, EUROPE, INFLUENCE, DIT, MONUMENTS}.

Nous pouvons donc présage qu'une partie spécifique de cette partie du livre traitre des arabes en rapport avec les monuments.

Afin de vérifier l'exactitude de cette affirmation, voici en exemple, un extrait de la cinquième partie du livre « La civilisation des arabes » [11] : « Il suffit de jeter un coup d'œil sur un monument appartenant à une époque avancée de la civilisation arabe, tel qu'un palais ou mosquée, ou simplement sur un objet quelconque, un encrier, un poignard, la reliure d'un coran, pour constater que ces œuvres d'art sont tellement caractéristiques qu'il n'y a jamais d'erreur possible sur leur origine. Grands ou petits, les produits divers du travail arabe n'ont aucune parenté visible avec les productions d'aucun autre peuple. Leur originalité est évidente et complète.» Dans cet extrait, nous pouvons constater que le texte parle d'un monument.

Les règles d'associations maximales nous aident donc à l'identification du thème de cette partie.

5.4.6 Partie 6

Dans cette section nous présentons les résultats de la partie 6.

5.4.6.1 Résultats des analyses lors du processus de classification

Voici la synthèse des opérations des trois types d'analyses obtenue suite à l'opération de classification avec ART, SOM, K-Means de la partie 6 (tableau 17).

		Unité d'information					
Classifieurs	Segments	Mots		Tri-Grams		Mots U Tri-Grams	
		Classes	Mots	Classes	Mots	Classes	Mots
ART	139	23	439	15	2470	38	2470
SOM	139	23	439	15	2470	38	2470
K-means	139	23	439	15	2470	38	2470
Total	417	69	439	45	2470	114	2470

Tableau 17 - Résultats des analyses lors du processus de classification de la partie 6

Comme lors de l'analyse de la première partie, le nombre de segments à classer nous donne un aperçu de la taille du texte. Nous pouvons aussi constater que le nombre de mots obtenu lors de la classification des tri-grams est approximativement cinq fois supérieur à celui obtenu lors de la classification des mots. Nous pouvons expliquer ce fait comme étant la cause des choix faits lors du nettoyage.

Pour identifier le thème de la partie du livre, nous utilisons l'ensemble des vingt mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). (114 classes).

Ce groupe de mots est le suivant : (ARABES, MONDE, PEUPLES, CIVILISATION, HISTOIRE, PUISSANCE, EMPIRE, FAIRE, HOMMES, FAIT, DIRE, DÉCADENCE, DERNIERS, GRANDEUR, CAUSES, INDE, PEUPLE, EUROPE, SIÈCLES, ÉPOQUE)

Si nous regardons ce groupe de mots, nous pouvons supposer que le thème de la sixième partie du livre traite de l'histoire de la civilisation de sa grandeur, sa puissance et sa décadence. Pour vérifier notre interprétation, regardons la table des matières du livre [11] que voici :

Livre sixième : La décadence de la civilisation arabe.

Chapitre I Les successeurs des arabes. – Influence des européens en Orient.

Chapitre II Causes de la grandeur et de la décadence des Arabes. État actuel de l'islamisme.

Nous voyons que nous avons cerné approximativement le thème. En effet, on retrouve plusieurs mots de l'ensemble dans les titres tels que : « DÉCADENCE », « CIVILISATION », ARABES », « INFLUENCE », « EUROPE », « CAUSES » et « GRANDEUR ».

5.4.6.2 Résultats obtenus lors de l'extraction des règles d'associations maximales

Comme ensemble E, nous utilisons un ensemble de 2470 mots. Cet ensemble correspond à la classe la plus large obtenue à la suite de l'unification des classifications avec ART, K-means et SOM (voir tableau 16 page 90).

Nous choisissons pour les besoins d'extractions des règles d'association maximales les quatre mots les plus fréquents présents dans toutes les classes issues de l'union des résultats des deux types d'unités d'information (mots et tri-grams). Nous établissons donc que le sous-ensemble X est {ARABES, MONDE, PEUPLES, CIVILISATION}. Bien entendu, d'autres façons de construire X sont possibles.

Dans le tableau 18 nous présentons une synthèse quant au nombre de règles d'association maximales extraites selon la règle {ARABES, SIÈCLE, ARABE} $\xrightarrow{\max}$ Y avec un seuil minimal de Mconfiance est de 50%.

Classifieurs	Unité d'information				
	Mot	Tri-gram	Mot U Tri-grams	Mot ∩ Tri-grams	Mot ∩ Tri-grams ∩ (Mot U Tri-grams)
ART	29	79	36	28	28
SOM	15	29	25	12	12
K-MEANS	30	27	21	18	18
A+S+K	26	38	29	23	23

Tableau 18 - Résultats obtenus lors de l'extraction des règles d'associations maximales de la partie 6

Nous pouvons observer que le nombre de règles d'associations maximales extraites est généralement plus élevé lorsque nous utilisons les tri-grams. Nous pouvons aussi constater que vu le peu de règles extraites lors de l'analyse basée sur les mots, le nombre de règles se retrouvant à la fois dans l'analyse basé sur les mots et l'analyse basée sur les Tri-grams est relativement bas. Finalement dans 100% des cas, lorsqu'un ensemble extrait est présent à la fois dans l'analyse basée sur l'unité d'information « mot » et dans l'analyse basée sur l'unité d'information « tri-gram », il est aussi présent lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ».

5.4.6.2.1 Règles d'association maximales extraites.

Les règles choisies sont les règles extraites lors de l'analyse basée sur les deux types d'unités d'information « mot » et « tri-gram ». Nous avons donc 29 règles d'associations maximales que nous présentons dans le tableau 19.

Règles d'association maximales	Mconfiance
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {PUISSANCE}	81%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {HISTOIRE}	80%

{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {EMPIRE}	80%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {DECADENCE}	72%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {CAUSES}	72%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {FAIRE}	69%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {GRANDEUR}	67%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {QUALITES}	67%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {HOMMES}	65%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {EPOQUE}	61%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {VUES}	61%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {RACES}	61%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {PEUPLE}	59%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {GRANDE}	59%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {JOUR}	57%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {FAIT}	57%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {EUROPE}	57%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {EGYPTE}	56%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {INFLUENCE}	56%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {MENT}	54%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {SIECLES}	54%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {INSTITUTIONS}	54%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {FACILEMENT}	52%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {MAHOMET}	52%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {ORIENT}	52%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {BESOIN}	50%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {ESPAGNE}	50%
{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {POLITIQUE}	50%

{ARABES, MONDE, PEUPLES, CIVILISATION} $\xrightarrow{max}$ {NIVEAU}	50%

Tableau 19 - Règles d'associations maximales extraites de la partie 6

Si nous supprimons les mots « fait » et « faire » qui, à notre avis sont des mots fonctionnels, le système de règles d'associations maximales associe fortement {ARABES, MONDE, PEUPLES, CIVILISATION} à {PUISSANCE, HISTOIRE, EMPIRE, DECADENCE, CAUSES, GRANDEUR, QUALITÉS, HOMMES, EPOQUE, VUE, RACES, PEUPLE, GRANDE, JOUR, EUROPE, EGYPTE, INFLUENCE, MENT, SIECLES, INSTITUTION, FACILEMENT, MAHOMET, ORIENT, BESOIN, ESPAGNE, POLITIQUE, NIVEAU}.
Nous pouvons présager qu'une bonne partie de cette partie du livre traite de la civilisation arabe en rapport avec sa puissance, son histoire, son empire, sa grandeur et sa décadence. (La Mconfiance se situe en haut de 70%).

Afin de vérifier l'exactitude de cette affirmation, voici en exemple, un extrait du livre « La civilisation des arabes » [11] : « Nous terminerons notre histoire de la civilisation des Arabes en résumant dans une vue d'ensemble les causes de leur grandeur et de leur décadence.» Dans cet extrait, nous pouvons constater que le texte parle de la civilisation arabe, des causes de sa grandeur et de sa décadence.

Les règles d'associations maximales nous aident donc à l'identification du thème de cette partie.

5.5 Évaluation des transactions issues de la combinaison des trois classifieurs

Comme nous l'avons déjà spécifié, le fait que chaque classifieur utilise sa propre stratégie lors de la classification soulève l'interrogation suivante. Est-ce que la stratégie de classification des classifieurs a une influence sur leur capacité à regrouper les segments ayant un vocabulaire commun dans les mêmes classes? Pour répondre à cette question nous soumettons donc en dernier lieu l'union des classes résultantes des méthodes de classification K-Means, SOM et ART aux règles d'associations maximales. Nous serons à même de déterminer si chaque classifieur a associé les mêmes segments aux mêmes classes.

5.5.1 Règles d'associations maximales.

Les transactions, dans ce cas sont formées des segments de texte et non du vocabulaire. Pour l'expérience qui suit, nous avons déterminé un M support de 3 et M confiance à 100%. Une Mconfiance de 100% signifie que les 3 classifieurs ont classifié les segments de texte se retrouvant dans le sous-ensemble X et le sous-ensemble Y dans la même classe. Une confiance moindre n'est pas pertinente car nous voulons évaluer dans notre cas, la capacité des classifieurs à regrouper les mêmes éléments. C'est-à-dire que pour un segment X1 et un segment X2, à chaque fois que l'on retrouvera X1, on retrouvera X2 également. On ajuste également le nombre de niveaux à 3 indiquant que l'ensemble Y peut contenir jusqu'à trois éléments. Ce choix est arbitraire, nous aurions pu le laisser à 1 mais nous voulons vérifier si les classifieurs ont classifié plus de deux segments ensembles. Pour cette dernière expérience, afin d'alléger le processus, seul les résultats issus de la combinaison des classifieurs basés sur les mots de la première partie du livre seront utilisés.

5.5.2 Analyse des résultats issus des règles d'association maximales.

Après avoir passé en revue toutes les règles d'associations maximales, il s'est avéré que 21 règles correspondaient au critère : avoir une Mconfiance de 100%. Sur ce nombre, une seule avait plus d'un niveau, soit la règle $\{136\} \xrightarrow{max} \{137, 138\}$. Cela signifie donc que seuls les segments 136, 137 et 138 se retrouvent toujours ensemble lors de leur classification par la méthode ART, SOM et K-Means dans le cadre de cette analyse. Afin d'alléger le texte, nous n'avons retenu que dix d'entre elles pour l'analyse. Afin de mieux visualiser les mots communs jugés pertinents, nous les surlignons.

5.5.2.1 Règle d'association maximales

Les dix énoncés suivant ont un taux de Mconfiance de 100%. Ceci signifie que les trois classifieurs ont classifiés ART, SOM et K-Means dans la même classe les segments contenu dans l'ensemble X et l'ensemble Y.

5.5.2.2 Analyse de la règle : $\{2\} \xrightarrow{max} \{3\}$

Si nous examinons les segments 2 et 3:

Segment 2: LA CIVILISATION DES ARABES

Segment 3: LA CIVILISATION DES ARABES

Nous retrouvons dans les deux segments l'énoncé « la civilisation des arabes ». Cela signifie que ces deux segments contenant le même énoncé ont été classifiés dans la même classe.

5.5.2.3 Règle :{6} $\xrightarrow{max}$ {8}

Si nous examinons les segments 6 et 8 :

Segment 6 : LES LECTEURS DE NOS PRÉCÉDENTS OUVRAGES CONNAISSENT LA GENÈSE DE CE NOUVEAU LIVRE. ILS SAVENT QU APRÈS AVOIR ÉTUDIÉ L HOMME ET LES SOCIÉTÉS NOUS DEVIONS ABORDER L HISTOIRE DES CIVILISATIONS.
Segment 8 : APRÈS AVOIR ÉTUDIÉ L HOMME ISOLÉ ET L ÉVOLUTION DES SOCIÉTÉS IL NOUS RESTE POUR COMPLÉTER NOTRE PLAN À APPLIQUER À L ÉTUDE DES GRANDES CIVILISATIONS LES MÉTHODES QUE NOUS AVONS EXPOSÉES.

Nous retrouvons dans les deux segments l'énoncé, « après avoir étudié l'homme » ainsi que les mots clés : sociétés et civilisations. Cela signifie que ces deux segments contenant les mêmes énoncés ont été classifiés dans la même classe.

5.5.2.4 Règle :{14} $\xrightarrow{max}$ {20}

Si nous examinons les segments 14 et 20 :

Segment 14 : L ACTION DES ARABES DÉJÀ SI GRANDE EN OCCIDENT FUT PLUS CONSIDÉRABLE ENCORE EN ORIENT. AUCUNE RACE N Y A JAMAIS EXERCÉ UNE INFLUENCE SEMBLABLE. LES PEUPLES QUI ONT JADIS RÉGNÉ SUR LE MONDE ASSYRIENS PERSES ÉGYPTIENS GRECS ET ROMAINS ONT DISPARU SOUS LA POUSSIÈRE DES SIÈCLES ET N ONT LAISSÉ QUE D INFORMES DÉBRIS ; LEURS RELIGIONS LEURS LANGUES ET LEURS ARTS NE SONT PLUS QUE DES SOUVENIRS. LES ARABES ONT DISPARU À LEUR TOUR; MAIS LES ÉLÉMENTS LES PLUS ESSENTIELS DE LEUR CIVILISATION LA RELIGION LA LANGUE ET LES ARTS SONT VIVANTS ENCORE ET DU MAROC JUSQU À L INDE PLUS DE CENT MILLIONS D HOMMES OBÉISSENT AUX INSTITUTIONS DU PROPHÈTE.

Segment 20 : C EST DE L ORIENT QUE L OCCIDENT EST NÉ ET C EST ENCORE À L ORIENT QU IL FAUT ALLER DEMANDER LA CLEF DES ÉVÉNEMENTS PASSÉS. SUR CETTE TERRE MERVEILLEUSE LES ARTS LES LANGUES ET LA PLUPART DES GRANDES RELIGIONS SE SONT MANIFESTÉS LES HOMMES N Y SONT PAS CE QU ILS SONT AILLEURS. IDÉES PENSÉES ET SENTIMENTS SONT AUTRES. LES TRANSFORMA TIONS Y SONT MAINTENANT SI LENTES

QU ON PEUT EN LE PARCOURANT REMONTER TOUTE LA CHAÎNE DES ÂGES. ARTISTES SAVANTS ET POÈTES Y REVIENDRONT TOUJOURS. QUE DE FOIS ASSIS À L OMBRE D UN PALMIER OU DU PYLÔNE DE QUELQUE TEMPLE ME SUIS JE PLONGÉ DANS DE LONGUES RÊVERIES PLEINES DE CLAIRES VISIONS DES ÂGES DISPARUS. ON S ASSOUPIT LÉGÈRE MENT ; ET SUR UN FOND LUMINEUX S ÉLÈVENT BIENTÔT DES VILLES ÉTRANGES DONT LES TOURS CRÉNELÉES LES PALAIS FÉERIQUES LES TEMPLES LES MINARETS SCINTILLENT SOUS UN SOLEIL D OR ET QUE PARCOURENT DES CARAVANES DE NOMADES DES FOULES D ASIATIQUES VÊTUS DE COULEURS ÉCLATANTES DES TROUPES D ESCLAVES À LA PEAU BRONZÉE DES FEMMES VOILÉES. ELLES SONT MORTES AUJOURD HUI POUR LA PLUPART CES GRANDES CITÉS DU PASSÉ NINIVE DAMAS JÉRUSALEM ATHÈNES GRENADE MEMPHIS ET LA THÈBES AUX CENT PORTES. LES PALAIS DE L ASIE LES TEMPLES DE L ÉGYPTE SONT MAINTENANT EN RUINES. LES DIEUX DE LA BABYLONIE DE LA SYRIE DE LA CHALDÉE DES RIVES DU NIL NE SONT PLUS QUE DES SOUVENIRS. MAIS QUE DE CHOSES DANS CES RUINES QUEL MONDE D IDÉES DANS CES SOUVENIRS. QUE DE SECRETS À DEMANDER À TOUTES CES RACES DIVERSES QUI SE SUCCÈDENT DES COLONNES D HERCULE AUX PLATEAUX FERTILES DE LA VIEILLE ASIE DES PLAGES VERDOYANTES DE LA MER ÉGÉE AUX SABLES BRILLANTS DE L ÉTHIOPIE.

Nous retrouvons dans les deux segments l'énoncé «ne sont plus que des souvenirs.» ainsi que les mots clés : plus, grandes, occident, encore, plus, orient, race, religion, art, langues et hommes. Cela signifie que ces deux segments contenant plusieurs éléments similaires ont été classifiés dans la même classe.

5.5.2.5 Règle :{26} $\xrightarrow{max}$ {27}

Si nous examinons les segments 26 et 27 :

Segment 26: NOUS AVONS SUFFISAMMENT EXPOSÉ AILLEURS LES MÉTHODES D INVESTIGATION QUI NOUS SEMBLENT APPLICABLES À L ÉTUDE DES PHÉNOMÈNES HISTORIQUES IL SUFFIRA DE RAPPELER LES PLUS ESSENTIELS

Segment 27: LA NOTION DE CAUSE QUI DOMINE AUJOURD HUI L ÉTUDE DES FAITS SCIENTIFIQUES DOMINE ÉGALEMENT CELLE DES FAITS

HISTORIQUES. LES MÉTHODES D INVESTIGATION APPLICABLES AUX UNS LE SONT ÉGALEMENT AUX AUTRES.

Nous retrouvons dans les deux segments les énoncés «les méthodes d'investigations» et « l'étude des » ainsi que les mots clés : applicables et historiques. Cela signifie que ses deux segments contenant plusieurs éléments similaires ont été classifiés dans la même classe.

5.5.2.6 Règle :{136} $\xrightarrow{max}$ {137}

Si nous examinons les segments 136 et 137 :

Segment 136 : LA GRANDE MOSQUÉE DE LA MECQUE A LA FORME D UN QUADRILATÈRE RÉGULIER. LORSQU ON A PÉNÉTRÉ DANS L INTÉRIEUR DU MONUMENT PAR UNE DES PORTES QUI Y DONNENT ACCÈS ON SE TROUVE DANS UNE VASTE COUR ENTOURÉE D ARCADES SOUTENUES PAR UNE VÉRITABLE FORÊT DE COLONNES AU DESSUS DESQUELLES S ÉLÈVENT UN NOMBRE CONSIDÉRABLE DE PETITES COUPOLES. DES MINARETS DISPOSÉS SUR DIVERSES PARTIES DU QUADRILATÈRE LE SURMONTENT

Segment 137: LE TEMPLE DE LA MECQUE A SERVI DE MODÈLE NOTAMMENT EN SYRIE À UN GRAND NOM BRE D AUTRES MOSQUÉES. J EN AI TROUVÉ PLUSIEURS CONSTRUITES SUR LE MÊME TYPE À DAMAS. CELLES DU CAIRE SONT AU CONTRAIRE ASSEZ DIFFÉRENTES PAR LA FORME DES MINARETS ET LES DÉTAILS DE LEUR ORNEMENTATION.

Nous retrouvons dans les deux segments, l'énoncé «de la Mecque» ainsi que le mot clé mosquée. Cela signifie que ces deux segments contenant plusieurs éléments similaires ont été classifiés dans la même classe.

5.5.2.7 Règle :{136} $\xrightarrow{max}$ {138}

Si nous examinons les segments 136 et 138 :

Segment 136 : LA GRANDE MOSQUÉE DE LA MECQUE A LA FORME D UN QUADRILATÈRE RÉGULIER. LORSQU ON A PÉNÉTRÉ DANS L INTÉRIEUR DU MONUMENT PAR UNE DES PORTES QUI Y DONNENT ACCÈS ON SE TROUVE DANS UNE VASTE COUR ENTOURÉE D ARCADES SOUTENUES PAR UNE VÉRITABLE FORÊT DE COLONNES AU DESSUS DESQUELLES S ÉLÈVENT UN NOMBRE CONSIDÉRABLE DE PETITES COUPOLES. DES MINARETS DISPOSÉS SUR DIVERSES PARTIES DU QUADRILATÈRE LE SURMONTENT

Segment 138 : LE PETIT TEMPLE DE LA KAABA SE TROUVE DANS LA COUR MÊME DE LA GRANDE MOSQUÉE DE LA MECQUE. C EST UN CUBE DE PIERRE GRISE AYANT SUIVANT BURCKHART PIEDS DE HAUTEUR PAS DE LONGUEUR ET DE LARGEUR. ELLE N A D AUTRE OUVERTURE QU UNE PETITE PORTE PLACÉE À PIEDS DU SOL À LAQUELLE ON NE PEUT ARRIVER QUE PAR UN ESCALIER MOBILE QU ON N APPLIQUE QUE PENDANT LA PÉRIODE DES PÈLERINAGES. SON INTÉRIEUR EST UNE SALLE PAVÉE DE MARBRE ÉCLAIRÉE PAR DES LAMPES D OR MASSIF ET RECOUVERTE D INSCRIPTIONS

Nous retrouvons dans les deux segments, l'énoncé «la grande mosquée de la Mecque» ainsi que le mot clé mosquée. Cela signifie que ces deux segments contenant plusieurs éléments similaires ont été classifiés dans la même classe.

5.5.2.8 Règle :{136} $\xrightarrow{max}$ {137, 138}

Si nous examinons les segments 136, 137 et 138 :

Segment 136 : LA GRANDE MOSQUÉE DE LA MECQUE A LA FORME D UN QUADRILATÈRE RÉGULIER. LORSQU ON A PÉNÉTRÉ DANS L INTÉRIEUR DU MONUMENT PAR UNE DES PORTES QUI Y DONNENT ACCÈS ON SE TROUVE DANS UNE VASTE COUR ENTOURÉE D ARCADES SOUTENUES PAR UNE VÉRITABLE FORÊT DE COLONNES AU DESSUS DESQUELLES S ÉLÈVENT UN NOMBRE CONSIDÉRABLE DE PETITES COUPOLES. DES MINARETS DISPOSÉS SUR DIVERSES PARTIES DU QUADRILATÈRE LE SURMONTENT

Segment 137: LE TEMPLE DE LA MECQUE A SERVI DE MODÈLE NOTAMMENT EN SYRIE À UN GRAND NOM BRE D AUTRES MOSQUÉES. J EN AI TROUVÉ PLUSIEURS CONSTRUITES SUR LE MÊME TYPE À DAMAS. CELLES DU CAIRE SONT AU CONTRAIRE ASSEZ DIFFÉRENTES PAR LA FORME DES MINARETS ET LES DÉTAILS DE LEUR ORNEMENTATION.

Segment 138 : LE PETIT TEMPLE DE LA KAABA SE TROUVE DANS LA COUR MÊME DE LA GRANDE MOSQUÉE DE LA MECQUE. C EST UN CUBE DE PIERRE GRISE AYANT SUIVANT BURCKHART PIEDS DE HAUTEUR PAS DE LONGUEUR ET DE LARGEUR. ELLE N A D AUTRE OUVERTURE QU UNE PETITE PORTE PLACÉE À PIEDS DU SOL À LAQUELLE ON NE PEUT ARRIVER QUE PAR UN ESCALIER MOBILE QU ON N APPLIQUE QUE PENDANT LA PÉRIODE DES PÈLERINAGES. SON INTÉRIEUR EST UNE SALLE PAVÉE DE MARBRE ÉCLAIRÉE PAR DES LAMPES D OR MASSIF ET RECOUVERTE D INSCRIPTIONS

Nous retrouvons dans les trois segments, l'énoncé «de la Mecque» ainsi que le mot clé mosquée. Cela signifie que ces deux segments contenant plusieurs éléments similaires ont été classifiés dans la même classe.

5.5.2.9 Règle :{273} $\xrightarrow{max}$ {275}

Si nous examinons les segments 273 et 275 :

Segment 273: LES ARABES AVANT MAHOMET

Le segment 275 soit : DES ARABES AVANT MAHOMET

Nous retrouvons dans les deux segments, l'énoncé «arabes avant Mahomet » Cela signifie que ces deux segments contenant plusieurs éléments similaires ont été classifiés dans la même classe.

5.5.2.10 Règle :{283} $\xrightarrow{max}$ {284}

Si nous examinons les segments 283 et 284 :

Segment 283 : L HISTOIRE N EST PAS RESTÉE AUSSI MUETTE SUR L ANCIENNE CULTURE DES ARABES QU ELLE L A ÉTÉ SUR D AUTRES CIVILISATIONS QUE LA SCIENCE MODERNE VOIT SORTIR AVEC ÉTONNEMENT DE LA POUSSIÈRE ; MAIS EÛT ELLE GARDE UN SILENCE COMPLET NOUS AURIONS PU ASSURER QUE LA CIVILISATION ARABE FUT

BIEN ANTÉRIEURE À MAHOMET. IL NOUS AURAIT SUFFI DE RAPPELER QU A L ÉPOQUE DU PROPHÈTE LES ARABES POSSÉDAIENT DÉJÀ UNE LITTÉRATURE ET UNE LANGUE TRÈS DÉVELOPPÉES ET SE TROUVAIENT DEPUIS PLUS DE ANS EN RELATIONS COMMERCIALES AVEC LES PEUPLES LES PLUS CIVILISÉS DU MONDE ET RÉUSSIRENT EN MOINS DE CENT ANS À CRÉER UNE DES PLUS BRILLANTES CIVILISATIONS DONT LES SIÈCLES ONT GARDÉ LA MÉMOIRE.

Segment 284 : OR UNE LITTÉRATURE ET UNE LANGUE NE S IMPROVISENT PAS ET LEUR EXISTENCE EST DÉJÀ LA PREUVE D UN LONG PASSÉ. LES RELATIONS SÉCULAIRES AVEC LES NATIONS LES PLUS CIVILISÉES FINISSENT TOUJOURS PAR CONDUIRE À LA CIVILISATION LES PEUPLES QUI EN SONT SUSCEPTIBLES ; ET LES ARABES ONT SUFFISAMMENT PROUVÉ QUE TEL ÉTAIT LEUR CAS. POUR AVOIR RÉUSSI ENFIN À CRÉER EN MOINS D UN SIÈCLE UN VASTE EMPIRE ET UNE CIVILISATION NOUVELLE IL FALLAIT DES APTITUDES QUI SONT TOUJOURS LE FRUIT DE LENTES ACCUMULATIONS HÉRÉDITAIRES ET PAR CONSÉQUENT D UNE LONGUE CULTURE ANTÉRIEURE. CE N EST PAS AVEC DES PEAUX ROUGES OU DES AUSTRALIENS QUE LES SUCCESSEURS DE MAHOMET EUSSENT CRÉÉ CES CITÉS BRILLANTES QUI PENDANT HUIT SIÈCLES FURENT LES SEULS FOYERS DES SCIENCES DES LETTRES ET DES ARTS EN ASIE ET EN EUROPE. BIEN D AUTRES PEUPLES QUE LES ARABES ONT RENVERSÉ DE GRANDS EMPIRES MAIS ILS N ONT PAS FONDÉ DE CIVILISATION ET FAUTE DE CULTURE ANTÉRIEURE SUFFISANTE ILS N ONT PROFITÉ QUE BIEN TARD DE LA CIVILISATION DES PEUPLES QU ILS AVAIENT VAINCUS. IL A FALLU DE LONGS SIÈCLES D EFFORTS AUX BARBARES QUI S EMPARÈRENT DE L EMPIRE ROMAIN POUR SE CRÉER UNE CIVILISATION AVEC LES DÉBRIS DE LA CIVILISATION LATINE ET SORTIR DE LA NUIT DU MOYEN ÂGE

Nous retrouvons dans les deux segments les énoncés «une littérature et une langue» et « les plus civilisées » ainsi que le mot clé : culture, arabes, civilisation, science, Mahomet, relations et siècles. Cela signifie que ces deux segments contenant plusieurs éléments similaires ont été classifiés dans la même classe.

5.5.2.11 Règle:{285} $\xrightarrow{\max}$ {286}

Si nous examinons les segments 285 et 286 :

Segment 285 : AVANT D ESSAYER DE DÉCOUVRIR AU MOYEN DES FAIBLES DOCUMENTS QUE NOUS POSSÉ DONS CE QUE FUT LA CIVILISATION DES ARABES AVANT MAHOMET NOUS ALLONS RÉSUMER RAPI DEMENT CE QUE NOUS SAVONS DE LEUR HISTOIRE

Segment 286 : HISTOIRE DES ARABES AVANT MAHOMET

Nous retrouvons dans les trois segments, l'énoncé «des arabes avant Mahomet» ainsi que le mot clé histoire. Cela signifie que ces deux segments contenant plusieurs éléments similaires ont été classifiés dans la même classe.

CHAPITRE 6

CONCLUSION

Tout d'abord, nous tenons à préciser que le présent travail de recherche a déjà fait l'objet de quatre publications [4] [5] [6] [7]

Le principal objectif de ce projet a été de démontrer la pertinence de l'utilisation des règles d'association maximales dans un processus d'assistance à l'interprétation des résultats issus des méthodes de classification. Pour ce faire, nous avons développé un prototype informatique pour analyser un ou plusieurs textes en entrée au moyen d'un processus d'extraction de règles d'association maximales combiné en aval à des processus de classification.

Dans ce mémoire, nous avons présenté, en premier lieu, différentes méthodes de calcul de distance utilisées par les classifieurs et avons ensuite présenté quatre classifieurs régulièrement cités dans la littérature, en l'occurrence K-NN, K-Means, SOM et ART.

Ensuite, nous avons introduit la notion de règles d'associations et de règles d'association maximales. Pour chacune des deux, nous avons décrit leur fonctionnement ainsi que leurs propriétés et leurs opérations. Pour finir, nous avons établi la relation existante entre les méthodes de classification et les règles d'associations maximales.

Nous avons, également, décrit chacune des parties de l'outil développé en décrivant leur rôle, leur fonction et leur algorithme.

Finalement, nous avons procédé à l'expérimentation en analysant un livre complet de 1980 pages soit : « La civilisation des Arabes de Gustave Le Bon (1884) ».

Cette expérimentation nous a permis de valider plusieurs aspects de notre hypothèse de recherche initiale.

Bien que chaque classifieur utilise sa propre stratégie lors de la classification, certains segments du texte ayant un vocabulaire commun sont classifiés de la même façon par tous les classifieurs (pour un segment X1 et un segment X2, à chaque fois que l'on retrouve X1, on retrouve X2 également). Les associations dont il est question ici sont considérées comme très fortes.

Par ailleurs, nous utilisons les classes résultantes d'une méthode de classification pour former l'ensemble nécessaire à la formation des sous-ensembles des règles d'associations maximales ($X \xrightarrow{max} Y$). Nous pouvons, également, utiliser la combinaison des classes résultantes de plusieurs méthodes de classification et/ou des classes résultantes issus de la classification d'unités d'information différentes pour former les ensembles nécessaire à la composition des règles d'associations maximales. Nous permettons ici de tenir compte des points de vue sous-jacents aux choix faits par l'utilisateur et ainsi proposer la possibilité d'une métaclassification.

En conclusion, un processus d'extraction de règles d'associations maximales dans des transactions (classes dans notre cas) issues de l'application de méthodes de classification, s'avère utile pour assister un ingénieur de la langue dans un processus de lecture et d'analyse de texte.

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Achouri, A., « Extraction de relations d'associations maximales dans les textes : représentation graphique », Mémoire de maitrise en informatique, Université du Québec à Trois-Rivières, Trois-Rivières, 2012.
2. Agrawal, R., Srikant, R., « Fast algorithms for mining association rules in large databases », In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proceedings of the 20th International Conference on Very Large Data Bases.* Santiago, Chile, 1994.
3. Agrawal, R., Imielinski, T., Swami, A., « Mining association rules between sets of items in large databases », In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Washington, 1993.
4. Biskri I., Achouri A., Rompré L., Descôteaux S, Bensaber Boucif A., «Computer-Assisted Reading: Getting Help from Text Classification and Maximal Association Rules », Journal of Advances in Information Technology 2013
5. Biskri I., Achouri A., Rompré L., Jouis C. Descôteaux S, Bensaber Boucif A., «Seeking for High Level Lexical Association in Texts», CIIA 2013, Spinger Publishing
6. Biskri I., Achouri A., Rompré L., Descôteaux S, Bensaber Boucif A., «Extraction of Strong Associations in Classes of Similarities», IEEE/ICMLA 2012, IEEE Press
7. Biskri I., Hallalli H., Rompré L., Descôteaux S., « Aide à l'interprétation des classes de similarité », Livre « L'organisation des sciences dynamisme et stabilité », chapitre 8, Presse Hermès Lavoisier, 2012.
8. Hajek, P., Havel, I., Chytil, M., The GUHA method of automatic hypotheses determination. Computing, 1966.
9. Hilali, H, « Application de la classification textuelle pour l'extraction des règles d'association maximales », Thèse de maitrise en informatique, Université du Québec à Trois-Rivières, Trois-Rivières, 2009.

10. Lallich, S., Teytaud O., « Évaluation et validation de l'intérêt des règles d'association », Revue des nouvelles Technologies de l'information, 2003.
11. Le Bon, Gustave, « La civilisation des Arabes ». Paris : Firmin-Didot, 1884.
12. Turenne, N., Apprentissage statistique pour l'extraction de concepts à partir de textes (Application au filtrage d'informations textuelles), Thèse de doctorat en informatique, Université Louis-Pasteur, Strasbourg, 2000.
13. Vaillant, B., Meyer P., « Mesurer l'intérêt des règles d'association », Revue des Nouvelles Technologies de l'Information (Extraction et gestion des connaissances: État et perspectives), 2006.

RÉFÉRENCES WEBOGRAPHIQUES

14. http://fr.wikipedia.org/wiki/Classification
15. http://fr.wikipedia.org/wiki/Distance_de_Hamming
16. http:/ http://www.uqtr.uquebec.ca/~biskri/Personnel/mol/RD.doc
17. http://docs.happycoders.org/orgadoc/artifical_intelligence/classification_documents/classification.pdf
18. http://www.codeproject.com/KB/recipes/K_nearest_neighbor.aspx
19. http://www.tsi.telecom-paristech.fr/pages/enseignement/ressources/beti/hyste-dyn/node2.html
20. http://www.codeproject.com/KB/recipes/K-Mean_Clustering.aspx
21. http://fr.wikipedia.org/wiki/Carte_auto_adaptative
22. http://dynamicnotions.blogspot.com/2008/11/c-self-organising-map-som.html
23. http://fr.wikipedia.org/wiki/Fonction_gaussienne
24. http://www.scholarpedia.org/article/Adaptive_resonance_theory
25. http://www.tralvex.com/pub/nap/nn-src/art1.txt
26. http://www.lirmm.fr/~teisseir/SUPPORTCOURS/RAMS07.pdf

ANNEXE 1

RÉSULTATS OBTENUS LORS DE L'EXTRACTION DES RÉGLES D'ASSOCIATIONS MAXIMALES

Voici les règles d'associations maximales obtenues suite à l'opération de leurs extractions avec ART, SOM, K-Means de chacune des parties du livre « La civilisation des Arabes de Gustave Le Bon (1884) ». Les tableaux « Mots » et « Tri-Grams » représentent les résultats obtenus à la suite des classifications des matrices d'entrées basées sur les unités d'information mot et tri-gram. Le tableau « Mots U Tri-Grams » représente les résultats obtenues lors de l'union des résultats des deux types d'unités d'information (mots et tri-grams). La colonne Y de chacun des tableaux représente l'ensemble Y de la règle d'association maximale $X \xrightarrow{max} Y$ avec un seuil minimum de Mconfiance de 50%. La colonne « MS » représente le Msupport de la règle et la colonne « MC » représente la Mconfiance.

Partie 1

X =: ARABES, ARABIE, HISTOIRE

2.1 ART

Mots

Y	MS	MC
ARABE	13	87
NOMADES	13	87
SIÈCLES	10	67
POPULATIONS	8	53
PEUPLES	8	53
VIE	8	53

Tri-grams

Y	MS	MC
PEUPLES	13	72
FAIT	12	67
NOMADES	12	67
ARABE	12	67
POPULATIONS	12	67
ÉTUDE	11	61
SIÈCLES	10	56
TRIBUS	10	56

Mots U Tri-Grams

Y	MS	MC
NOMADES	25	76
ARABE	25	76
PEUPLES	21	64
SIÈCLES	20	61
POPULATIONS	20	61
FAIT	18	55
TRIBUS	17	52

HOMME	9	50
EUROPE	9	50

2.2 SOM

Mots

Y	MS	MC
FAUT	13	62
ARABE	13	62
NOMADES	13	62
FAIT	12	57
ÉPOQUE	12	57
PEUPLES	11	52
SIÈCLES	11	52

Tri-grams

Y	MS	MC
PEUPLES	10	100
FAUT	9	90
ÉPOQUE	8	80
NOMADES	7	70
ÉTUDE	7	70
ARABE	7	70
SIÈCLES	7	70
EUROPE	6	60
DÉBRIS	6	60
PREMIERS	6	60
PIERRE	6	60
FAIT	6	60
HOMME	5	50
NOM	5	50
SOL	5	50
ROIS	5	50
TRIBUS	5	50

Mots U Tri-Grams

Y	MS	MC
FAUT	22	71
PEUPLES	21	68
ÉPOQUE	20	65
ARABE	20	65
NOMADES	20	65
SIÈCLES	18	58
FAIT	18	58
ÉTUDE	17	55
EUROPE	16	52

2.3 K-means

Mots

Y	MS	MC
ARABE	9	69
TRIBUS	9	69
PEUPLES	8	62
FAIT	8	62
NOMADES	7	54

Tri-grams

Y	MS	MC
FAIT	9	82
ARABE	8	73
TRIBUS	8	73
NOMADES	7	64
PEUPLES	7	64
SIÈCLES	6	55
ÉPOQUE	6	55

Mots U Tri-Grams

Y	MS	MC
FAIT	17	71
ARABE	17	71
TRIBUS	17	71
PEUPLES	15	63
NOMADES	14	58

2.4 ART U SOM U K-means

Mots

Y	MS	MC
ARABE	36	71
NOMADES	33	67
PEUPLES	27	55
SIÈCLE	26	53
FAIT	26	53
FAUT	25	51

Tri-grams

Y	MS	MC
PEUPLES	30	77
ARABE	27	69
FAIT	27	69
NOMADES	26	67
SIÈCLES	23	59
ÉTUDE	23	59
TRIBUS	23	59
FAUT	22	56
ÉPOQUE	21	54
EUROPE	20	51
POPULATIONS	20	51

Mots U Tri-Grams

Y	MS	MC
ARABE	62	70
NOMADES	59	67
PEUPLES	57	65
FAIT	53	60
SIÈCLES	49	56
FAUT	47	53
TRIBUS	46	52

Partie 2

X =: MAHOMET, PROPHÈTE, DIEU, ARABES

3.1 ART

Mots

Y	MS	MC
HOMME	5	100
HOMMES	5	100
VIE	5	100
ARABIE	4	80
GRANDE	4	80
EMPIRE	4	80
MONDE	4	80
PRINCIPAUX	4	80
DERNIER	4	80
COMPAGNONS	4	80
JOUR	4	80
DIT	4	80

Tri-grams

Y	MS	MC
EMPIRE	7	87,5
FAIRE	6	75
PUISSANCE	6	75
VIE	6	75
JOUR	6	75
HOMME	6	75
ABOU	6	75
HOMMES	6	75
JUSTICE	6	75
MONDE	6	75
HISTORIENS	5	62,5
MISSION	5	62,5

Mots U Tri-Grams

Y	MS	MC
HOMMES	11	85
EMPIRE	11	85
VIE	11	85
HOMME	11	85
MONDE	10	77
JOUR	10	77
JUSTICE	9	69
DIT	9	69
FAIRE	9	69
ARABIE	9	69
PUISSANCE	9	69
DERNIER	8	62

RECONNAÎTRE	3	60
SIMPLICITÉ	3	60
ANNÉES	3	60
ÈRE	3	60
FAMILLE	3	60
PARENTS	3	60
PUISSANCE	3	60
JUSTICE	3	60
CARACTÈRE	3	60
HISTORIENS	3	60
FAIRE	3	60
TROUVER	3	60
LOI	3	60
ALI	3	60
PERSONNE	3	60
MAIN	3	60
UNS	3	60
AMOUR	3	60
VOIX	3	60
FAIM	3	60
PIERRE	3	60
SUCCESSEUR	3	60
CHRÉTIENNE	3	60
RELIGIEUSE	3	60
DISCOURS	3	60
ANS	3	60
PASSAIT	3	60
NUIT	3	60
MOIS	3	60
TRIBUS	3	60

PREMIÈRE	5	62,5
ANS	5	62,5
SUCCESSEURS	5	62,5
ARABIE	5	62,5
MAÎTRE	5	62,5
DIT	5	62,5
SUCCESSEUR	4	50
LOI	4	50
TROUVER	4	50
MOURUT	4	50
POLITIQUE	4	50
RECONNAÎTRE	4	50
ÂGE	4	50
DERNIER	4	50
GABRIEL	4	50
PREMIER	4	50
MÉDINE	4	50
ÈRE	4	50
ENVOYÉ	4	50
NOMBRE	4	50
MAIN	4	50
ANNÉES	4	50
TERRE	4	50
CARACTÈRE	4	50
NOM	4	50
GRANDE	4	50
FAMILLE	4	50
PARENTS	4	50

HISTORIENS	8	62
GRANDE	8	62
ABOU	8	62
ANS	8	62
MAIN	7	54
ÈRE	7	54
LOI	7	54
ANNÉES	7	54
FAMILLE	7	54
PARENTS	7	54
PREMIÈRE	7	54
CARACTÈRE	7	54
PRINCIPAUX	7	54
SUCCESSEUR	7	54
COMPAGNONS	7	54
RECONNAÎTRE	7	54
TROUVER	7	54
SUCCESSEURS	7	54
MAÎTRE	7	54

3.2 SOM

Mots

Y	MS	MC
FAIRE	8	73
HOMMES	7	64
MECQUE	7	64
ANS	7	64
HOMME	7	64
JOUR	6	55
VIE	6	55

Tri-grams

Y	MS	MC
FAIRE	9	69
MONDE	9	69
HOMME	9	69
ANS	9	69
VIE	9	69
PUISSANCE	9	69
EMPIRE	8	62
SUCCESSEUR	8	62
HOMMES	7	54
MISSION	7	54

Mots U Tri-Grams

Y	MS	MC
FAIRE	17	71
HOMME	16	67
ANS	16	67
VIE	15	63
HOMMES	14	58
EMPIRE	14	58
MONDE	14	58
PUISSANCE	13	54
MECQUE	13	54
JOUR	12	50

3.3 K-means

Mots

Y	MS	MC
FAIRE	9	69
MONDE	9	69
HOMME	9	69
ANS	9	69
VIE	9	69
PUISSANCE	9	69
EMPIRE	8	62
SUCCESSEUR	8	62
HOMMES	7	54
MISSION	7	54

Tri-grams

Y	MS	MC
HOMMES	7	78
HOMME	7	78
MONDE	7	78
FAIRE	6	67
JOUR	6	67
MECQUE	6	67
DIT	6	67
PREMIER	5	56
PREMIÈRE	5	56
ANS	5	56
FEMME	5	56
MOIS	5	56
ENVOYÉ	5	56
COMPAGNONS	5	56
PEINE	5	56
MAÎTRE	5	56
MISSION	5	56
VIE	5	56

Combiné

Y	MS	MC
HOMMES	13	72
HOMME	12	67
JOUR	12	67
MONDE	12	67
MISSION	11	61
MECQUE	11	61
FAIRE	11	61
MAÎTRE	10	56
DIT	10	56
ABOU	10	56
ANS	9	50
VIE	9	50
ENVOYÉ	9	50

ABOU	5	56

2. 4 ART U SOM U K-means

Mots

Y	MS	MC
HOMMES	18	72
HOMME	17	68
JOUR	16	64
FAIRE	16	64
VIE	15	60
MONDE	14	56
MECQUE	14	56
ANS	14	56
HISTORIENS	13	52

Tri-grams

Y	MS	MC
MONDE	22	73
HOMME	22	73
FAIRE	21	70
VIE	20	67
HOMMES	20	67
PUISSANCE	19	63
ANS	19	63
EMPIRE	18	60
JOUR	18	60
MISSION	17	57
PREMIÈRE	16	53
JUSTICE	16	53
MAÎTRE	15	50
SUCCESSEUR	15	50
DIT	15	50
MECQUE	15	50

Mots U Tri-Grams

Y	MS	MC
HOMME	39	71
HOMMES	38	69
FAIRE	37	67
MONDE	36	65
VIE	35	64
JOUR	34	62
ANS	33	60
EMPIRE	30	55
MECQUE	29	53
PUISSANCE	29	53
JUSTICE	28	51

Partie 3

X = ARABES, ARABE, GRANDE

4.1 ART

Mots

Y	MS	MC
FAIT	20	80
ESPAGNE	20	80
SIÈCLE	18	72
VILLE	18	72
RESTE	17	68
PAYS	17	68

Tri-grams

Y	MS	MC
INFLUENCE	18	82
VILLE	18	82
SIÈCLE	17	77
PAYS	17	77
FAIT	17	77
CIVILISATION	17	77

Mots U Tri-Grams

Y	MS	MC
FAIT	37	79
VILLE	36	77
ESPAGNE	36	77
SIÈCLE	35	74
INFLUENCE	34	72
PAYS	34	72

JOUR	17	68
MONDE	17	68
NOM	17	68
CONQUÊTE	17	68
AFRIQUE	16	64
VILLES	16	64
KHALIFE	16	64
INFLUENCE	16	64
PREMIERS	15	60
HOMMES	15	60
MUSULMANS	15	60
FAIRE	15	60
ABORD	14	56
FIT	14	56
CIVILISATION	13	52
PEUPLES	13	52
PALAIS	13	52
FAUT	13	52

ESPAGNE	16	73
PARTIE	16	73
PREMIERS	15	68
MONUMENT	15	68
FAIRE	15	68
DAMAS	15	68
BERBÈRES	15	68
ABORD	15	68
HOMMES	14	64
MONDE	14	64
PEUPLES	14	64
VILLES	14	64
FIT	14	64
FORME	14	64
PALAIS	14	64
GRECS	14	64
JOUR	14	64
NOM	14	64
CONQUÊTE	14	64
RESTE	14	64
INDE	13	59
KHALIFE	13	59
TERRE	12	55
DIT	12	55
RELIGION	12	55
MUSULMANS	12	55
FAUT	12	55
CHRÉTIENNE	11	50
NOMBREUX	11	50
ARCHITECTURE	11	50
POPULATION	11	50

MONDE	31	66
NOM	31	66
RESTE	31	66
CONQUÊTE	31	66
JOUR	31	66
VILLES	30	64
FAIRE	30	64
PREMIERS	30	64
CIVILISATION	30	64
KHALIFE	29	62
HOMMES	29	62
ABORD	29	62
FIT	28	60
DAMAS	27	57
PALAIS	27	57
MUSULMANS	27	57
PEUPLES	27	57
PARTIE	26	55
AFRIQUE	26	55
FAUT	25	53
BERBÈRES	24	51
MONUMENT	24	51
FORME	24	51

4.2 SOM

Mots

Y	MS	MC
ESPAGNE	21	72

Tri-grams

Y	MS	MC
SIÈCLE	16	73

Mots U Tri-Grams

Y	MS	MC
ESPAGNE	36	71

FAIT	19	66
RESTE	19	66
MONDE	19	66
NOM	18	62
INFLUENCE	17	59
KHALIFE	17	59
FAIRE	16	55
JOUR	16	55
MUSULMANS	16	55
FAUT	16	55
SIÈCLE	16	55
CIVILISATION	15	52
DIT	15	52

ESPAGNE	15	68
FAIT	15	68
CIVILISATION	14	64
PAYS	14	64
INFLUENCE	14	64
FAIRE	14	64
VILLE	14	64
FAUT	13	59
DIT	13	59
RESTE	13	59
PEUPLES	12	55
MONDE	12	55
MUSULMANS	12	55
PARTIE	12	55
JOUR	12	55
NOM	12	55
DEVAIT	12	55
PALAIS	11	50
FIT	11	50
FORME	11	50
MONUMENT	11	50

FAIT	34	67
SIÈCLE	32	63
RESTE	32	63
MONDE	31	61
INFLUENCE	31	61
NOM	30	59
FAIRE	30	59
CIVILISATION	29	57
FAUT	29	57
JOUR	28	55
MUSULMANS	28	55
DIT	28	55
VILLE	27	53
KHALIFE	27	53
PARTIE	26	51
PAYS	26	51

4.3 K-means

Mots

Y	MS	MC
FAIT	25	76
NOM	22	67
SIÈCLE	20	61
PARTIE	20	61
FAIRE	20	61
ABORD	19	58
INFLUENCE	19	58
VILLE	19	58
CIVILISATION	17	52
FORME	17	52
PAYS	17	52

Tri-grams

Y	MS	MC
SIÈCLE	20	67
FAIT	19	63
INFLUENCE	19	63
PARTIE	17	57
ABORD	17	57
RESTE	16	53
MONDE	16	53
VILLES	16	53
CIVILISATION	15	50
CONQUÊTE	15	50
VILLE	15	50

Mots U Tri-Grams

Y	MS	MC
FAIT	44	70
SIÈCLE	40	63
INFLUENCE	38	60
PARTIE	37	59
ABORD	36	57
NOM	35	56
VILLE	34	54
FAIRE	34	54
CIVILISATION	32	51
MONDE	32	51

4.4 ART U SOM U K-means

Mots

Y	MS	MC
FAIT	64	74
NOM	57	66
Espagne	54	2
SIÈCLE	54	62
MONDE	52	60
INFLUENCE	52	60
FAIRE	51	59
VILLES	50	57
RESTE	50	57
KHALIFE	48	55
ABORD	46	53
PAYS	46	53
CONQUÊTE	45	52
CIVILISATION	45	52
PARTIE	44	51
JOUR	44	51

Tri-grams

Y	MS	MC
SIÈCLE	53	71.6
INFLUENCE	51	69
FAIT	51	69
VILLE	47	64
CIVILISATION	46	62
PARTIE	45	61
PAYS	45	61
Espagne	44	59
FAIRE	43	58
RESTE	43	58
MONDE	42	57
ABORD	42	57
VILLES	39	53
NOM	39	53
FORME	39	53
CONQUÊTE	38	51
FIT	38	51
PALAIS	38	51
JOUR	37	50
KHALIFE	37	50
FAUT	37	50
DIT	37	50

Mots U Tri-Grams

Y	MS	MC
FAIT	115	71
SIÈCLE	107	66
INFLUENCE	103	64
ESPAGNE	98	61
VILLE	97	60
NOM	96	60
FAIRE	94	58
MONDE	94	58
RESTE	93	58
PAYS	91	57
CIVILISATION	91	57
PARTIE	89	55
ABORD	88	55
KHALIFE	85	53
CONQUÊTE	83	52
JOUR	81	50
FIT	81	50

Partie 4

X = ARABES, FEMME, VIE

5.1 ART

Mots

Y	MS	MC
FAMILLE	9	64
HOMMES	8	57

Tri-grams

Y	MS	MC
FAMILLE	12	80
ORIENT	12	80

Mots U Tri-Grams

Y	MS	MC
FAMILLE	21	72
ORIENT	20	69

ORIENT	8	57

HOMMES	11	73
ARABE	9	60
MUSULMANS	9	60

HOMMES	19	66

5.2 SOM

Mots

Y	MS	MC
ORIENT	11	85
HOMMES	7	54

Tri-grams

Y	MS	MC
ORIENT	10	83
ARABE	8	67
HOMMES	7	58
FAMILLE	7	58
MUSULMANS	6	50

Mots U Tri-Grams

Y	MS	MC
ORIENT	21	84
HOMMES	14	56

5.3 K-means

Mots

Y	MS	MC
ORIENT	8	62
FAMILLE	8	62

Tri-grams

Y	MS	MC
FAMILLE	9	75
ORIENT	8	67
HOMMES	6	50
BESOIN	6	50
PÈRE	6	50

Mots U Tri-Grams

Y	MS	MC
FAMILLE	17	68
ORIENT	16	64

5.4 ART U SOM U K-means

Mots

Y	MS	MC
ORIENT	27	68
FAMILLE	22	55
HOMMES	21	53

Tri-grams

Y	MS	MC
ORIENT	30	77
FAMILLE	28	72
HOMMES	24	62

Mots U Tri-Grams

Y	MS	MC
ORIENT	57	72
FAMILLE	50	63
HOMMES	45	57

Partie 5

X =: ARABES, ARABE, SIÈCLE

6.1 ART

Mots

Y	MS	MC
EUROPE	20	71
FAIT	19	68
GRANDE	18	64
INFLUENCE	17	61
Espagne	17	61
ÉPOQUE	16	57
MONUMENTS	16	57
ANCIENS	16	57
DIT	15	54
SCIENCE	15	54
ARCHITECTURE	15	54
DIXIÈME	14	50
MODERNES	14	50
INDE	14	50
OEUVRES	14	50

Tri-grams

Y	MS	MC
FAIT	21	81
ESPAGNE	21	81
EUROPE	21	81
ÉPOQUE	19	73
GRANDE	19	73
DIT	19	73
MONUMENTS	19	73
ANCIENS	18	69
OEUVRES	18	69
USAGE	17	65
INFLUENCE	17	65
HISTOIRE	17	65
ART	17	65
SIÈCLES	17	65
MODERNES	16	62
AFRIQUE	16	62
MUSULMANS	16	62
MOSQUÉE	16	62
BAGDAD	15	58
DIXIÈME	15	58
ARCHITECTURE	15	58
PLUPART	15	58
AUTEURS	15	58
ARTS	14	54
OUVRAGES	14	54
FORME	14	54
AUTEUR	14	54
DIRE	14	54
ORIENT	14	54

Mots U Tri-Grams

Y	MS	MC
EUROPE	41	76
FAIT	40	74
ESPAGNE	38	70
GRANDE	37	69
ÉPOQUE	35	65
MONUMENTS	35	65
DIT	34	63
INFLUENCE	34	63
ANCIENS	34	63
OEUVRES	31	57
MUSULMANS	30	56
MODERNES	30	56
ART	30	56
ARCHITECTURE	30	56
DIXIÈME	29	54
USAGE	29	54
SCIENCE	28	52
SIÈCLES	28	52
HISTOIRE	28	52
INDE	28	52

INDE	14	54
CONNAISSANCES	14	54
MONDE	13	50
NOTAMMENT	13	50
ÉTUDE	13	50
TREIZIÈME	13	50
GRAND	13	50
NOM	13	50
FAIRE	13	50
SEULEMENT	13	50
SCIENCE	13	50
DOIT	13	50
STYLE	13	50
TRAVAUX	13	50

6.2 SOM

Mots

Y	MS	MC
DIT	18	60
ÉPOQUE	17	57
FAIT	16	53
MUSULMANS	15	50
ANCIENS	15	50
BAGDAD	15	50
Espagne	15	50

Tri-grams

Y	MS	MC
FAIT	20	74
Espagne	18	67
FAIRE	17	63
MONUMENTS	17	63
ART	17	63
MODERNES	16	59
ÉPOQUE	16	59
EUROPE	16	59
INFLUENCE	15	56
ARCHITECTURE	15	56
DIT	15	56
GRANDE	15	56
ANCIENS	14	52
DIRE	14	52

Mots U Tri-Grams

Y	MS	MC
FAIT	36	63
ÉPOQUE	33	58
ESPAGNE	33	58
DIT	33	58
MODERNES	30	53
EUROPE	30	53
MONUMENTS	30	53
FAIRE	29	51
INFLUENCE	29	51
ANCIENS	29	51

6.3 K-means

Mots

Y	MS	MC
ANCIENS	19	66
GRANDE	18	62
Espagne	17	59
ÉPOQUE	16	55
DIXIÈME	15	52
FAIT	15	52
MONUMENTS	15	52

Tri-grams

Y	MS	MC
ANCIENS	22	76
Espagne	19	66
GRANDE	18	62
ÉPOQUE	18	62
INFLUENCE	18	59
FAIT	17	59
DIRE	17	59
NOTAMMENT	16	55
FORME	16	55
DIT	16	55

Mots U Tri-Grams

Y	MS	MC
ANCIENS	41	71
GRANDE	36	62
ESPAGNE	36	62
ÉPOQUE	34	59
FAIT	32	55
INFLUENCE	31	53
DIRE	29	50
MONUMENTS	29	50
NOTAMMENT	29	50
DIXIÈME	29	50

6.4 ART U SOM U K-means

Mots

Y	MS	MC
ANCIENS	50	57
FAIT	50	57
ÉPOQUE	49	56
Espagne	49	56
GRANDE	48	55
EUROPE	48	55
DIT	44	51
MONUMENTS	44	51
INFLUENCE	44	51

Tri-grams

Y	MS	MC
Espagne	58	71
FAIT	58	71
ANCIENS	54	66
ÉPOQUE	53	65
GRANDE	52	63
EUROPE	51	62
DIT	50	61
INFLUENCE	50	61
MONUMENTS	50	61
ART	46	56
DIRE	45	55
MODERNES	44	54
SIÈCLE	43	52
FAIRE	43	52
ARCHITECTURE	43	52
FORME	42	51
HISTOIRE	42	51
NOTAMENT	41	50
MUSULMANS	41	50

Mots U Tri-Grams

Y	MS	MC
FAIT	108	64
ESPAGNE	107	63
ANCIENS	104	62
ÉPOQUE	102	60
GRANDE	100	59
EUROPE	99	59
INFLUENCE	94	56
DIT	94	56
MONUMENTS	94	56

MOSQUÉE	41	50

Partie 6

X =: ARABES, MONDE, PEUPLES, CIVILISATION

7.1 ART

Mots

Y	MS	MC
PUISSANCE	9	90
FAIRE	8	80
PEUPLE	8	80
HISTOIRE	8	80
CAUSES	8	80
EMPIRE	8	80
GRANDE	7	70
DÉCADENCE	7	70
DIVERSES	7	70
RACES	7	70
HOMMES	7	70
INSTITUTIONS	6	60
JOUR	6	60
BESOINS	6	60
GRANDEUR	6	60
EUROPE	6	60
QUALITÉS	6	60
ÉPOQUE	6	60
NIVEAU	6	60
FACILEMENT	5	50
ESPAGNE	5	50
GRANDS	5	50
POPULATIONS	5	50
FACTEURS	5	50
RAISON	5	50

Tri-grams

Y	MS	MC
DÉCADENCE	5	100
EMPIRE	5	100
ÉGYPTE	5	100
EUROPE	5	100
GRANDEUR	5	100
RACES	5	100
CAUSES	5	100
FAIRE	5	100
INSTITUTIONS	5	100
PUISSANCE	5	100
MENT	5	100
POPULATIONS	5	100
JOUR	4	80
PROPHÈTE	4	80
DIVERSES	4	80
RELIGION	4	80
RELIGIEUX	4	80
ORIENTAUX	4	80
SURENT	4	80
RAISON	4	80
SUBIR	4	80
POSSIBLE	4	80
POLITIQUE	4	80
HISTOIRE	4	80
GRANDE	4	80

Mots U Tri-Grams

Y	MS	MC
PUISSANCE	14	93
EMPIRE	13	87
CAUSES	13	87
FAIRE	13	87
HISTOIRE	12	80
DÉCADENCE	12	80
RACES	12	80
HOMMES	11	73
GRANDE	11	73
INSTITUTIONS	11	73
GRANDEUR	11	73
EUROPE	11	73
PEUPLE	11	73
DIVERSES	11	73
ÉPOQUE	10	67
POPULATIONS	10	67
NIVEAU	10	67
JOUR	10	67
BESOINS	9	60
MENT	9	60
ORIENTAUX	9	60
RAISON	9	60
QUALITÉS	9	60
ROMAINS	9	60
ESPAGNE	9	60

ROMAINS	5	50
PEUVENT	5	50
ORIENTAUX	5	50
VUE	5	50

HEUREUX	4	80
ORIENT	4	80
NIVEAU	4	80
ESPAGNE	4	80
MAHOMET	4	80
SIÈCLES	4	80
FAIT	4	80
HOMMES	4	80
ÉPOQUE	4	80
GRECS	4	80
ROMAINS	4	80
SUPÉRIEURS	4	80
INFLUENCE	4	80
POLITIQUES	4	80
SUPÉRIORITÉ	4	80
FACILEMENT	4	80
CONSÉQUENCES	3	60
NOUVELLE	3	60
RÉUSSI	3	60
CIVILISÉS	3	60
PEUPLE	3	60
CHINOIS	3	60
AMÉRIQUE	3	60
PRÉSENTE	3	60
PLACER	3	60
VUE	3	60
FIRENT	3	60
FRAPPANT	3	60
BESOINS	3	60
COUCHES	3	60
VRAI	3	60
INFÉRIEURES	3	60
EMPIRES	3	60
FONDÉ	3	60
HABILETÉ	3	60
PROSPÉRITÉ	3	60
SENTIR	3	60

FACILEMENT	9	60
ÉGYPTE	9	60
SUPÉRIORITÉ	8	53
POLITIQUE	8	53
PEUVENT	8	53
FAIT	8	53
SUPÉRIEURS	8	53
ORIENT	8	53
VUE	8	53
SIÈCLES	8	53
GRANDS	8	53

CORAN	3	60
ISLAMISME	3	60
CROISEMENTS	3	60
ÉPOQUES	3	60
PROGRÈS	3	60
ARABIE	3	60
DURENT	3	60
CONQUÊTES	3	60
SOCIALES	3	60
MODIFICATIONS	3	60
DOUCEUR	3	60
GRANDS	3	60
FACILITÉ	3	60
ÉGAUX	3	60
PEUVENT	3	60
QUALITÉS	3	60
POSSÈDENT	3	60

7.2 SOM

Mots

Y	MS	MC
PUISSANCE	9	82
ORIENT	8	73
RACES	8	73
HISTOIRE	8	73
CAUSES	8	73
ÉGYPTE	7	64
EMPIRE	7	64
DÉCADENCE	7	64
INFLUENCE	6	55
INTELLIGENCE	6	55
ORIENTAUX	6	55
NIVEAU	6	55
PEUPLE	6	55
JOUR	6	55
QUALITÉS	6	55

Tri-grams

Y	MS	MC
FAIRE	8	89
PUISSANCE	8	89
EMPIRE	8	89
GRANDEUR	7	78
DÉCADENCE	7	78
QUALITÉS	7	78
CAUSES	7	78
HOMMES	7	78
HISTOIRE	7	78
ÉPOQUE	7	78
SIÈCLES	6	67
INFLUENCE	6	67
EUROPE	6	67
POLITIQUE	6	67
SUPÉRIORITÉ	5	56

Mots U Tri-Grams

Y	MS	MC
PUISSANCE	17	85
EMPIRE	15	75
CAUSES	15	75
HISTOIRE	15	75
DÉCADENCE	14	70
RACES	13	65
QUALITÉS	13	65
INFLUENCE	12	60
HOMMES	12	60
ORIENT	12	60
ÉPOQUE	12	60
GRANDEUR	12	60
FAIRE	12	60
PEUPLE	11	55
INTELLIGENCE	11	55

PEUPLE	5	56
TÊTE	5	56
ESPAGNE	5	56
ORIENTAUX	5	56
INTELLIGENCE	5	56
VUE	5	56
MAHOMET	5	56
JOUR	5	56
GRANDE	5	56
RELIGION	5	56
RACES	5	56
INSTITUTIONS	5	56
BESOINS	5	56
ARABIE	5	56

JOUR	11	55
ORIENTAUX	11	55
EUROPE	11	55
ÉGYPTE	11	55
SIÈCLES	10	50
POLITIQUE	10	50
ESPAGNE	10	50
GRANDE	10	50
VUE	10	50
NIVEAU	10	50

7.3 K-means

Mots

Y	MS	MC
GRANDEUR	6	86
QUALITÉS	6	86
MAHOMET	6	86
VUE	6	86
EMPIRE	6	86
HISTOIRE	6	86
FAIT	6	86
PUISSANCE	5	71
FAIRE	5	71
DÉCADENCE	5	71
CAUSES	5	71
MENT	5	71
FACILEMENT	5	71
ÉPOQUE	5	71
ÉGYPTE	4	57
SOCIALES	4	57
BESOINS	4	57
SIÈCLES	4	57

Tri-grams

Y	MS	MC
HISTOIRE	10	83
FAIT	9	75
VUE	9	75
EMPIRE	9	75
HOMMES	8	67
PUISSANCE	8	67
INFLUENCE	8	67
DÉCADENCE	8	67
QUALITÉS	8	67
FAIRE	7	58
JOUR	7	58
SIÈCLES	7	58
MENT	7	58
EUROPE	7	58
GRANDE	7	58
GRANDEUR	7	58
PEUPLE	7	58
INTELLIGENCE	6	50

Mots U Tri-Grams

Y	MS	MC
HISTOIRE	16	84
FAIT	15	79
EMPIRE	15	79
VUE	15	79
QUALITÉS	14	74
GRANDEUR	13	68
DÉCADENCE	13	68
PUISSANCE	13	68
MENT	12	63
FAIRE	12	63
HOMMES	12	63
MAHOMET	12	63
FACILEMENT	11	58
CAUSES	11	58
INFLUENCE	11	58
SIÈCLES	11	58
GRANDE	11	58
ÉPOQUE	11	58

INSTITUTIONS	4	57
HOMMES	4	57
ARABIE	4	57
FACILITÉ	4	57
POLITIQUES	4	57
ESPAGNE	4	57
FIRENT	4	57
RACES	4	57
FACTEURS	4	57
PROPHÈTE	4	57
PEUVENT	4	57
GRANDE	4	57

ORIENT	6	50
ISLAMISME	6	50
ÉGYPTE	6	50
ÉPOQUE	6	50
MAHOMET	6	50
FACILEMENT	6	50
NOUVELLE	6	50
POLITIQUE	6	50
CAUSES	6	50

PEUPLE	10	53
JOUR	10	53
ÉGYPTE	10	53

7.4 ART U SOM U K-means

Mots

Y	MS	MC
PUISSANCE	23	82
HISTOIRE	22	79
CAUSES	21	75
EMPIRE	21	75
RACES	19	68
DÉCADENCE	19	68
QUALITÉS	18	64
PEUPLE	17	61
FAIRE	17	61
GRANDEUR	17	61
VUE	16	57
ÉPOQUE	16	57
GRANDE	16	57
HOMMES	16	57
DIVERSES	15	54
JOUR	15	54
ÉGYPTE	15	54
FAIT	15	54
NIVEAU	14	50
ESPAGNE	14	50
ORIENT	14	50

Tri-grams

Y	MS	MC
EMPIRE	22	85
PUISSANCE	21	81
HISTOIRE	21	81
DÉCADENCE	20	77
FAIRE	20	77
HOMMES	19	73
GRANDEUR	19	73
QUALITÉS	18	69
CAUSES	18	69
INFLUENCE	18	69
EUROPE	18	69
ÉPOQUE	17	65
VUE	17	65
SIÈCLES	17	65
JOUR	16	62
FAIT	16	62
MENT	16	62
GRANDE	16	62
POLITIQUE	16	62
INSTITUTIONS	15	58
MAHOMET	15	58

Mots U Tri-Grams

Y	MS	MC
PUISSANCE	44	81
HISTOIRE	43	80
EMPIRE	43	80
DÉCADENCE	39	72
CAUSES	39	72
FAIRE	37	69
GRANDEUR	36	67
QUALITÉS	36	67
HOMMES	35	65
ÉPOQUE	33	61
VUE	33	61
RACES	33	61
PEUPLE	32	59
GRANDE	32	59
JOUR	31	57
FAIT	31	57
EUROPE	31	57
ÉGYPTE	30	56
INFLUENCE	30	56
MENT	29	54
SIÈCLES	29	54

PEUVENT	14	50
FACTEURS	14	50
BESOINS	14	50
INSTITUTIONS	14	50
FACILEMENT	14	50

PEUPLE	15	58
ÉGYPTE	15	58
FACILEMENT	14	54
SUPÉRIORITÉ	14	54
RACES	14	54
ORIENT	14	54
POPULATIONS	14	54
RELIGION	14	54
NIVEAU	13	50
RAISON	13	50

INSTITUTIONS	29	54
FACILEMENT	28	52
MAHOMET	28	52
ORIENT	28	52
BESOINS	27	50
ESPAGNE	27	50
POLITIQUE	27	50
NIVEAU	27	50

ANNEXE 2

RÉSULTATS OBTENUS LORS DE L'EXTRACTION DES RÈGLES D'ASSOCIATIONS MAXIMALES DANS LES TRANSACTIONS FORMÉES DES SEGMENTS DE TEXTE DE LA PREMIÈRE PARTIE DU LIVRE « LA CIVILISATION DES ARABES ».

Voici les six premiers résultats des transactions formées des segments de texte de la première partie dans le cadre de cet expérience au point 5.4.2.1. La colonne Y de chacun des tableaux représente les segments de texte contenus dans le sous-ensemble Y de la règle d'association maximale $X \xrightarrow{max} Y$. La colonne «Msupport » représente le Msupport de la règle (le Msupport correspond au nombre de classifieurs ayant classifiés les segments de texte se retrouvant dans le sous-ensemble X et le sous-ensemble Y, dans la même classe). La colonne « Mconfiance » représente la Mconfiance. Finalement, la colonne « texte des segments » représente le segment du texte de l'élément de l'ensemble Y correspondant.

Règle : $\{2\} \xrightarrow{max} \{?\}$

Le segment 2 soit : LA CIVILISATION DES ARABES

Msupport	Mconfiance	Y	Texte des segments
3	100	3	LA CIVILISATION DES ARABES ( )
2	66.67	27	LA NOTION DE CAUSE QUI DOMINE AUJOURD'HUI L ÉTUDE DES FAITS SCIENTIFIQUES DOMINE ÉGALEMENT CELLE DES FAITS HISTORIQUES. LES MÉTHODES D'INVESTIGATION APPLICABLES AUX UNS LE SONT ÉGALEMENT AUX AUTRES
		67	GUSTAVE LE BON LA CIVILISATION DES ARABES ( )
2	66.67	67	GUSTAVE LE BON LA CIVILISATION DES ARABES ( )
		215	LES QUALITÉS ET LES DÉFAUTS DES ARABES NOMADES SONT

			NATURELLEMENT LES QUALITÉS ET LES DÉFAUTS ENGENDRÉS PAR LEURS CONDITIONS D EXISTENCE.
2	66.67	165	GUSTAVE LE BON LA CIVILISATION DES ARABES ( )
		188	ORIGINE DES ARABES
2	66.67	2	DR GUSTAVE LE BON ( )
1	33.33	62	NOUS TERMINERONS CETTE INTRODUCTION EN DÉGAGEANT DE CE QUI PRÉCÈDE LA MÉTHODE QUE NOUS AVONS SUIVIE DANS CE VOLUME ET QUE NOUS SUIVRONS DANS TOUS CEUX CONSACRÉS À NOTRE HISTOIRE DES CIVILISATIONS.
		178	IMPORTANCE DE L ÉTUDE
		309	MAIS AU MOMENT OÙ PARUT MAHOMET ELLE ÉTAIT MENACÉE D INVASIONS REDOUTABLES. L AN DE J. C. L YÉMEN QUI N'AVAIT JUSQU ALORS OBÉI QU'À DES SOUVERAINS ARABES AVAIT ÉTÉ ENVAHI PAR LES ABYSSINS QUI ESSAYÈRENT D'Y PROPAGER LA RELIGION CHRÉTIENNE ET RÉUSSIRENT À CONVERTIR PLUSIEURS TRIBUS. EN C' EST À DIRE FORT PEU DE TEMPS AVANT MAHOMET, ILS FURENT CHASSÉS PAR LES PERSES QUI Y ÉTABLIRENT DES VICE ROIS. CES DERNIERS RÉGNÈRENT SUR LE YÉMEN, L' HADRAMAUT ET L' OMAN JUSQU À L' ARRIVÉE DU PROPHÈTE.

**Le tableau possède 44492 données (3 niveaux)*

Règle : {6} $\xrightarrow{max}$ {?}

Le segment 6 soit : LES LECTEURS DE NOS PRÉCÉDENTS OUVRAGES CONNAISSENT LA GENÈSE DE CE NOUVEAU LIVRE. ILS SAVENT
QU'APRÈS AVOIR ÉTUDIÉ L'HOMME ET LES SOCIÉTÉS, NOUS DEVIONS ABORDER L'HISTOIRE DES CIVILISATIONS.

Msupport	Mconfiance	Y	Texte des segments
3	100	8	APRÈS AVOIR ÉTUDIÉ L HOMME ISOLÉ ET L ÉVOLUTION DES SOCIÉTÉS IL NOUS RESTE POUR COMPLÉTER NOTRE PLAN À APPLIQUER À L ÉTUDE DES GRANDES CIVILISATIONS LES

			MÉTHODES QUE NOUS AVONS EXPOSÉES
2	66.67	7	NOTRE DERNIER TRAVAIL AVAIT ÉTÉ CONSACRÉ À DÉCRIRE LES FORMES SUCCESSIVES DE L ÉVOLUTION PHYSIQUE ET INTELLECTUELLE DE L HOMME LES ÉLÉMENTS DIVERS DONT LES SOCIÉTÉS SE COMPOSENT. REMONTANT AUX PLUS LOINTAINES PÉRIODES DE NOTRE PASSÉ NOUS AVONS FAIT VOIR COMMENT SE FORMÈRENT LES PREMIÈRES AGGLOMÉRATIONS HUMAINES COMMENT NAQUI RENT LA FAMILLE ET LES SOCIÉTÉS L INDUSTRIE ET LES ARTS LES INSTITUTIONS ET LES CROYANCES ; COMMENT CES ÉLÉMENTS SE TRANSFORMÈRENT À TRAVERS LES ÂGES ET QUELS FURENT LES FACTEURS DE CES TRANSFORMATIONS
2	66.67	3	LA CIVILISATION DES ARABES ( )
2	66.67	2	LA CIVILISATION DES ARABES
1	33.33	4	INTRODUCTION
1	33.33	9	L ENTREPRISE EST VASTE SES DIFFICULTÉS SONT GRANDES. IGNORANT JUSQU OÙ NOUS POUR RONS LA CONDUIRE NOUS AVONS VOULU QUE CHACUN DES VOLUMES QUI COMPOSERONT CET OUVRAGE FÛT COMPLET ET INDÉPENDANT. S IL NOUS EST DONNÉ DE TERMINER LES HUIT À DIX VOLUMES QUE NOTRE PLAN COMPREND RIEN NE SERA PLUS SIMPLE QUE DE CLASSER ENSUITE DANS UN ORDRE MÉTHODIQUE L HISTOIRE DES DIVERSES CIVILISATIONS À L ÉTUDE DESQUELLES CHACUN D EUX AURA ÉTÉ CONSACRÉ.
...	...	...	

**Le tableau possède 25 données (1 niveau)*

Règle : $\{14\} \xrightarrow{max} \{?\}$

Le segment 14 : L ACTION DES ARABES DÉJÀ SI GRANDE EN OCCIDENT FUT PLUS CONSIDÉRABLE ENCORE EN ORIENT. AUCUNE RACE N Y A JAMAIS EXERCÉ UNE INFLUENCE SEMBLABLE. LES PEUPLES QUI ONT JADIS RÉGNÉ SUR LE MONDE ASSYRIENS PERSES ÉGYPTIENS GRECS ET ROMAINS ONT DISPARU SOUS LA POUSSIÈRE DES SIÈCLES ET N ONT LAISSÉ QUE D INFORMES DÉBRIS ; LEURS RELIGIONS LEURS LANGUES ET LEURS ARTS NE SONT PLUS QUE DES SOUVENIRS. LES ARABES ONT DISPARU À LEUR TOUR; MAIS LES ÉLÉMENTS LES PLUS ESSENTIELS DE LEUR CIVILISATION LA RELIGION LA LANGUE ET LES ARTS SONT VIVANTS ENCORE ET DU MAROC JUSQU À L INDE PLUS DE CENT MILLIONS D HOMMES OBÉISSENT AUX INSTITUTIONS DU PROPHÈTE.

Msupport	Mconfiance	Y	Texte des segments
3	100	20	C EST DE L ORIENT QUE L OCCIDENT EST NÉ ET C EST ENCORE À L ORIENT QU IL FAUT ALLER DEMANDER LA CLEF DES ÉVÉNEMENTS PASSÉS. SUR CETTE TERRE MERVEILLEUSE LES ARTS LES LANGUES ET LA PLUPART DES GRANDES RELIGIONS SE SONT MANIFESTÉS LES HOMMES N Y SONT PAS CE QU ILS SONT AILLEURS. IDÉES PENSÉES ET SENTIMENTS SONT AUTRES. LES TRANSFORMA TIONS Y SONT MAINTENANT SI LENTES QU ON PEUT EN LE PARCOURANT REMONTER TOUTE LA CHAÎNE DES ÂGES. ARTISTES SAVANTS ET POÈTES Y REVIENDRONT TOUJOURS. QUE DE FOIS ASSIS À L OMBRE D UN PALMIER OU DU PYLÔNE DE QUELQUE TEMPLE ME SUIS JE PLONGÉ DANS DE LONGUES RÊVERIES PLEINES DE CLAIRES VISIONS DES ÂGES DISPARUS. ON S ASSOUPIT LÉGÈRE MENT ; ET SUR UN FOND LUMINEUX S ÉLÈVENT BIENTÔT DES VILLES ÉTRANGES DONT LES TOURS CRÉNELÉES LES PALAIS FÉERIQUES LES TEMPLES LES MINARETS SCINTILLENT SOUS UN SOLEIL D OR ET QUE PARCOURENT DES CARAVANES DE NOMADES DES FOULES D ASIATIQUES VÊTUS DE COULEURS ÉCLATANTES DES TROUPES D ESCLAVES À LA PEAU BRONZÉE DES FEMMES VOILÉES. ELLES SONT MORTES AUJOURD HUI POUR LA PLUPART CES GRANDES CITÉS DU PASSÉ NINIVE DAMAS JÉRUSALEM ATHÈNES GRENADE MEMPHIS ET LA THÈBES AUX CENT PORTES. LES PALAIS DE L ASIE LES TEMPLES DE L ÉGYPTE SONT MAINTENANT EN RUINES. LES DIEUX DE LA BABYLONIE

			DE LA SYRIE DE LA CHALDÉE DES RIVES DU NIL NE SONT PLUS QUE DES SOUVENIRS. MAIS QUE DE CHOSES DANS CES RUINES QUEL MONDE D IDÉES DANS CES SOUVENIRS. QUE DE SECRETS À DEMANDER À TOUTES CES RACES DIVERSES QUI SE SUCCÈDENT DES COLONNES D HERCULE AUX PLATEAUX FERTILES DE LA VIEILLE ASIE DES PLAGES VERDOYANTES DE LA MER ÉGÉE AUX SABLES BRILLANTS DE L ÉTHIOPIE.
2	66.67	21	ON RAPPORTE BIEN DES ENSEIGNEMENTS DE CES CONTRÉES LOINTAINES ; ON Y PERD AUSSI BIEN DES CROYANCES. LEUR ÉTUDE NOUS MONTRE COMBIEN EST PROFOND L ABÎME QUI SÉPARE LES HOMMES ET À QUEL POINT SONT CHIMÉRIQUES NOS IDÉES DE CIVILISATION ET DE FRATERNITÉ UNIVERSELLE COMBIEN LES VÉRITÉS ET LES PRINCIPES QUI SEMBLENT LES PLUS ABSOLUS PEUVENT CHANGER EN RÉALITÉ D UN PAYS À L AUTRE
2	66.67	16	C EST UNE MERVEILLEUSE HISTOIRE QUE CELLE DE CET HALLUCINÉ ILLUSTRE DONT LA VOIX SOUMIT CE PEUPLE INDOCILE QU AUCUN CONQUÉRANT N AVAIT PU DOMPTER AU NOM DUQUEL FURENT RENVERSÉS LES PLUS PUISSANTS EMPIRES ET QUI DU FOND DE SON TOMBEAU TIENT ENCORE DES MILLIONS D HOMMES SOUS SA LOI.
1	33.33	227	IL EN EST AINSI DE TOUTES LES RACES OU DE TOUTES LES NATIONS PRIMITIVES ET A EN EST AINSI DES FEMMES ET DES ENFANTS PARCE QU ILS REPRÉSENTENT ÉGALEMENT DES FORMES INFÉRIEURES DE L ÉVOLUTION HUMAINE. LE NOMADE N EST EN RÉALITÉ QU UN DEMI SAUVAGE. DEMI SAUVAGE INTELLIGENT ASSURÉMENT MAIS QUI DEPUIS DES MILLIERS D ANNÉES N A PAS FAIT UN PAS VERS LA CIVILISATION ET PAR CONSÉQUENT N A SUBI AUCUNE DES TRANSFORMATIONS ACCUMULÉES PAR L HÉRÉDITÉ CHEZ L HOMME CIVILISÉ. SI COMME NOUS LE CROYONS LES CARACTÈRES PSYCHOLOGIQUES SUFFISENT À ÉTABLIR DES DIFFÉRENCES PROFONDES ENTRE LES HOMMES ON PEUT DIRE QUE L ARABE SÉDENTAIRE ET L ARABE NOMADE FORMENT DEUX RACES VÉRITABLEMENT SÉPARÉES PAR UN ABÎME
1	33.33	301	CONNUES DES GRECS PLUS DE QUATRE SIÈCLES AVANT J. C. LES RICHESSES DES

			ARABES AVAIENT DÉTERMINÉ ALEXANDRE À TENTER LA CONQUÊTE DE L ARABIE ET L EXPÉDITION DE NÉARQUE AUTOUR DE LA PÉNINSULE AURAIT ÉTÉ LE PRÉSAGE DE L EXÉCUTION PROCHAINE D UN DESSEIN QUE LA MORT VINT INTERROMPRE. LORS DU PARTAGE DE L EMPIRE D ALEXANDRE LES RÉGIONS VOISINES DES FRONTIÈRES DE L ÉGYPTE ET DE LA PALESTINE HABITÉES PAR LES ARABES TOMBÈRENT AU POUVOIR DE PTOLÉMÉE. LES NABATHÉENS PRIRENT PARTI POUR PTOLÉMÉE CONTRE ANTIGONE. QUAND CELUI CI FUT MAÎTRE DE LA SYRIE ET DE LA PHÉNICIE IL ENVOYA CONTRE EUX UN DE SES MEILLEURS GÉNÉRAUX QUI APRÈS S ÊTRE EMPARÉ DE PÉTRA PAR SURPRISE EUT SON ARMÉE DE HOMMES ENTIÈREMENT DÉTRUITE. ANTIGONE ENVOYA ALORS CONTRE EUX SON FILS DÉMÉTRIUS. LORSQUE CE DERNIER ARRIVA À PETRA LES ARABES AU DIRE DE DIODORE LUI TINRENT CE LANGAGE ROI DÉMÉTRIUS POURQUOI NOUS FAIS TU LA GUERRE À NOUS QUI HABITONS DES DÉSERTS OÙ L ON NE TROUVE RIEN DE CE QUI EST NÉCESSAIRE À LA VIE PAISIBLE DES HABITANTS D UNE CITÉ C EST PARCE QUE NOUS SOMMES DÉTERMINÉS À FUIR L ESCLAVAGE QUE NOUS AVONS CHERCHÉ UN REFUGE AU MILIEU D UNE CONTRÉE PRIVÉE DE TOUTES RESSOURCES. CONSENS DONC À ACCEPTER LES PRÉSENTS QUE NOUS T OFFRONS POUR FAIRE RETIRER TON ARMÉE ET SOIS SÛR QUE TU AURAS DORÉNAVANT DANS LES NABATHÉENS DE FIDÈLES AMIS. QUE SI TU VOULAIS PROLONGER LE SIÈGE TU ÉPROUVERAIS BIENTÔT DES PRIVATIONS DE TOUTE ESPÈCE ET TU NE POURRAIS JAMAIS NOUS CONTRAINDRE À MENER UN GENRE DE VIE DIFFÉRENT DE CELUI AUQUEL NOUS SOMMES HABITUÉS DES NOTRE ENFANCE. SI TOUT AU PLUS TU PARVENAIS À FAIRE PARMI NOUS QUELQUES PRISONNIERS TU NE TROUVERAIS EN EUX QUE DES ESCLAVES DÉCOURAGÉS ET INCAPABLES DE VIVRE SOUS D AUTRES INSTITUTIONS QUE LES NÔTRES
1	33.33	316	LES ANCIENNES CHRONIQUES ARABES SONT CONFORMES AUX RENSEIGNEMENTS FOURNIS PAR LES AUTEURS CLASSIQUES. TOUS SONT UNANIMES À VANTER LA RICHESSE DE L YÉMEN ON Y VOYAIT DIT

			MASOUDI À PROPOS DU PAYS DE MAREB DE BEAUX ÉDIFICES DES ARBRES MAGNI FIQUES DES CANAUX EN GRAND NOMBRE DES RIVIÈRES QUI LE PARCOURAIENT EN TOUS SENS. TEL ÉTAIT L ÉTAT DE CE PAYS QUI AVAIT EN LONGUEUR ET EN LARGEUR L ÉTENDUE QUE POURRAIT PAR COURIR EN UN MOIS DE TEMPS UN BON CAVALIER. UN VOYAGEUR SOIT À PIED SOIT À CHEVAL POUVAIT SUIVRE TOUTE CETTE ROUTE D UNE EXTRÉMITÉ À L AUTRE SANS RESSENTIR LES ARDEURS DU SOLEIL ; IL Y TROUVAIT PARTOUT UN OMBRAGE TOUFFU QUI NE LE QUITTAIT PAS ; CAR LES ARBRES DONT LA CULTURE FAISAIT LA RICHESSE DE CE PAYS COUVRAIENT TOUTE CETTE TERRE ET LUI FAISAIENT UN ABRI CONTINUEL. LES HABITANTS JOUISSAIENT DE TOUTES LES AISANCES DE LA VIE ILS AVAIENT EN ABONDANCE TOUS LES MOYENS DE SUBSISTANCE ; UNE TERRE FERTILE UN AIR PUR UN CIEL SEREIN DES SOURCES D EAU NOMBREUSES UNE GRANDE PUISSANCE UNE DOMINATION BIEN AFFERMIE UN EMPIRE AU PLUS HAUT POINT DE PROSPÉRITÉ TOUT CONTRIBUAIT À FAIRE DE LEUR PAYS UN SÉJOUR DONT LES AVANTAGES ÉTAIENT PASSÉS EN PROVERBE. ILS SE DISTINGUAIENT AUSSI PAR LA NOBLESSE DE LEUR CONDUITE ET PAR L EMPRESSEMENT AVEC LEQUEL ILS ACCUEILLAIENT DE TOUT LEUR POUVOIR ET SUIVANT LEURS FACULTÉS TOUS LES ÉTRANGERS QUI VENAIENT DANS LEUR PAYS ET TOUS LES VOYAGEURS. CET ÉTAT DE PROSPÉRITÉ DURA AUSSI LONGTEMPS QU IL PLUT A DIEU ; AUCUN ROI NE LEUR RÉSISTA QUI NE FÛT DÉFAIT ; AUCUN TYRAN NE MARCHA CONTRE EUX AVEC SES ARMÉES QUI NE FÛT MIS EN DÉROUTE ; TOUTES LES RÉGIONS LEUR ÉTAIENT SOUMISES TOUS LES HOMMES RECONNAISSAIENT LEURS LOIS ; ILS ÉTAIENT COMME LE DIADÈME SUR LE FRONT DE L UNIVERS.
...	...	...	

**Le tableau possède 19 données (1 niveau)*

Règle : {136} $\xrightarrow{max}$ {?}

Le segment 136 soit : LA GRANDE MOSQUÉE DE LA MECQUE A LA FORME D UN QUADRILATÈRE RÉGULIER. LORSQU ON A PÉNÉTRÉ DANS L INTÉRIEUR DU MONUMENT PAR UNE DES PORTES QUI Y DONNENT ACCÈS ON SE TROUVE DANS UNE VASTE COUR ENTOURÉE D ARCADES SOUTENUES PAR UNE VÉRITABLE FORÊT DE COLONNES AU DESSUS DESQUELLES S ÉLÈVENT UN NOMBRE CONSIDÉRABLE DE PETITES COUPOLES. DES MINARETS DISPOSÉS SUR DIVERSES PARTIES DU QUADRILATÈRE LE SURMONTENT

Msupport	Mconfiance	Y	Texte des segments
3	100	138	LE PETIT TEMPLE DE LA KAABA SE TROUVE DANS LA COUR MÊME DE LA GRANDE MOSQUÉE DE LA MECQUE. C EST UN CUBE DE PIERRE GRISE AYANT SUIVANT BURCKHART PIEDS DE HAUTEUR PAS DE LONGUEUR ET DE LARGEUR. ELLE N A D AUTRE OUVERTURE QU UNE PETITE PORTE PLACÉE À PIEDS DU SOL À LAQUELLE ON NE PEUT ARRIVER QUE PAR UN ESCALIER MOBILE QU ON N APPLIQUE QUE PENDANT LA PÉRIODE DES PÈLERINAGES. SON INTÉRIEUR EST UNE SALLE PAVÉE DE MARBRE ÉCLAIRÉE PAR DES LAMPES D OR MASSIF ET RECOUVERTE D INSCRIPTIONS
3	100	137	LE TEMPLE DE LA MECQUE A SERVI DE MODÈLE NOTAMMENT EN SYRIE À UN GRAND NOM BRE D AUTRES MOSQUÉES. J EN AI TROUVÉ PLUSIEURS CONSTRUITES SUR LE MÊME TYPE À DAMAS. CELLES DU CAIRE SONT AU CONTRAIRE ASSEZ DIFFÉRENTES PAR LA FORME DES MINARETS ET LES DÉTAILS DE LEUR ORNEMENTATION.
		138	LE PETIT TEMPLE DE LA KAABA SE TROUVE DANS LA COUR MÊME DE LA GRANDE MOSQUÉE DE LA MECQUE. C EST UN CUBE DE PIERRE GRISE AYANT SUIVANT BURCKHART PIEDS DE HAUTEUR PAS DE LONGUEUR ET DE LARGEUR. ELLE N A D AUTRE OUVERTURE QU UNE PETITE PORTE PLACÉE À PIEDS DU SOL À LAQUELLE ON NE PEUT ARRIVER QUE PAR UN ESCALIER MOBILE QU ON N APPLIQUE QUE PENDANT LA PÉRIODE DES PÈLERINAGES. SON INTÉRIEUR EST UNE SALLE PAVÉE DE MARBRE ÉCLAIRÉE PAR DES LAMPES D OR MASSIF ET RECOUVERTE D INSCRIPTIONS

3	100	137	LE TEMPLE DE LA MECQUE A SERVI DE MODÈLE NOTAMMENT EN SYRIE À UN GRAND NOM BRE D AUTRES MOSQUÉES. J EN AI TROUVÉ PLUSIEURS CONSTRUITES SUR LE MÊME TYPE À DAMAS. CELLES DU CAIRE SONT AU CONTRAIRE ASSEZ DIFFÉRENTES PAR LA FORME DES MINARETS ET LES DÉTAILS DE LEUR ORNEMENTATION.
2	66.67	140	LES MURS DE LA KAABA SONT TOUS REVÊTUS DE MARBRE DE DIVERSES COULEURS. DU COTÉ DE L OCCIDENT IL Y A SIX MIRAHBS EN ARGENT FIXÉS À LA MURAILLE PAR DES CLOUS CHACUN D EUX A LA HAUTEUR D UN HOMME ; ILS SONT ORNÉS D INCRUSTATIONS EN OR ET EN ARGENT NIELLÉ D UNE TEINTE NOIRE FONCÉE. LES MURAILLES SONT JUSQU À LA HAUTEUR DE QUATRE ARECH AU DESSUS DE LA TERRE DANS LEUR ÉTAT PRIMITIF ; À PARTIR DE CETTE HAUTEUR ELLES SONT JUSQU AU PLAFOND RECOUVERTES DE PLAQUES DE MARBRE ORNÉES D ARABESQUES ET DE SCULPTURES DONT LA PLUS GRANDE PARTIE EST DORÉE.
2	66.67	142	LA KAABA EST TOUJOURS RECOUVERTE D UN IMMENSE VOILE NOIR SAUF À L ENDROIT OÙ SE TROUVE LA PIERRE SACRÉE. CE VOILE EST RELEVÉ À QUELQUES PIEDS DU SOL. PENDANT LES PRE MIERS JOURS DU PÈLERINAGE IL EST ENTOURÉ VERS LE MILIEU DE SA HAUTEUR D UNE BANDE PORTANT EN LETTRES D OR DES INSCRIPTIONS DU CORAN. UNE FOIS PAR AN CE VOILE EST RENOUVELÉ.
2	66.67	141	DANS UNE DES MURAILLES EXTÉRIEURES DE LA KAABA SE TROUVE ENCHÂSSÉE LA CÉLÈBRE PIERRE NOIRE APPORTÉE SUIVANT LES ARABES DU PARADIS PAR LES ANGES POUR SERVIR DE MARCHEPIED À ABRAHAM LORSQU IL CONSTRUISAIT LE TEMPLE. CETTE RELIQUE N A GUÈRE QUE POUCES DE DIAMÈTRE. AUCUN AUTRE OBJET N A ÉTÉ ENTOURÉ D UNE AUSSI LONGUE VÉNÉRATION DE LA PART DES HOMMES CAR BIEN DES SIÈCLES AVANT MAHOMET LA PIERRE NOIRE ÉTAIT DÉJÀ VÉNÉRÉE.
2	66.67	135	C EST AU MILIEU DE LA MECQUE QUE S ÉLÈVE LA MOSQUÉE À LAQUELLE LA MÈRE DES CITÉS DOIT SA CÉLÉBRITÉ. DANS SON INTÉRIEUR SE TROUVE LA KAABA TEMPLE CÉLÈBRE DONT LA FONDATION SUIVANT LES HISTORIENS ORIENTAUX REMONTE À

			ABRAHAM. KHALIFES SULTANS CONQUÉRANTS ONT DEPUIS MAHOMET TENU À TÉMOIGNER LEUR PIÉTÉ EN ORNANT LA CÉLÈBRE MOSQUÉE ; ET AUJOURD HUI IL NE RESTE RIEN DE SON ORNEMENTATION PRIMITIVE.
...	...		

**Le tableau possède 11 données (3 niveaux)*

Règle : {26} $\xrightarrow{\max}$ {?}

Le segment 26 soit : NOUS AVONS SUFFISAMMENT EXPOSÉ AILLEURS LES MÉTHODES D INVESTIGATION QUI NOUS SEMBLENT APPLICABLES À L ÉTUDE DES PHÉNOMÈNES HISTORIQUES IL SUFFIRA DE RAPPELER LES PLUS ESSENTIELS

Msupport	Mconfiance	Y	Texte des segments
3	100	27	LA NOTION DE CAUSE QUI DOMINE AUJOURD HUI L ÉTUDE DES FAITS SCIENTIFIQUES DOMINE ÉGALEMENT CELLE DES FAITS HISTORIQUES. LES MÉTHODES D INVESTIGATION APPLICABLES AUX UNS LE SONT ÉGALEMENT AUX AUTRES.
1	33.33	187.	J AI EXAMINÉ DANS UN RÉCENT OUVRAGE L INFLUENCE PROFONDE QUE PEUVENT AVOIR SUR LES DESTINÉES D UN PEUPLE LES ÉLÉMENTS QUI ENTRENT DANS SON SEIN SURTOUT LORSQUE CES ÉLÉMENTS ONT DES TENDANCES DIFFÉRENTES ET CHERCHÉ À MONTRER QUE C EST DANS CETTE ÉTUDE BEAUCOUP PLUS QUE DANS CELLE DES INSTITUTIONS POLITIQUES INSTITUTIONS QUI SONT DES CONSÉQUENCES ET BIEN RAREMENT DES CAUSES QU ON PEUT TROUVER LA CLEF DU RÔLE QUE LES NATIONS ONT JOUÉ OU JOUERONT DANS L HISTOIRE. JE NE SAURAIS INSISTER DAVANTAGE SUR UN SUJET QUE JE NE POUVAIS QU EFFLEURER ICI. LE PEU QUE J EN AI DIT A DÛ SUFFIRE POUR MONTRER AU LECTEUR L IMPORTANCE DE L ÉTUDE DE LA PSYCHOLOGIE DES PEUPLES SCIENCE À PEINE ÉBAUCHÉE ENCORE . EN CE QUI CONCERNE LES ARABES NOTAMMENT NOUS VERRONS QUE C EST DANS L ÉTUDE DE LEUR CARACTÈRE QUE NOUS TROUVERONS EN GRANDE PARTIE L

			EXPLICATION DES CAUSES QUI ONT DÉTERMINÉ LEUR GRANDEUR ET LEUR DÉCADENCE.
1	33.33	178	IMPORTANCE DE L ÉTUDE
1	33.33	269	C EST DANS LES TEMPS PASSÉS QU ONT ÉTÉ ÉLABORÉS LES MOTIFS DE NOS ACTIONS ET DANS LE TEMPS PRÉSENT QUE SE PRÉPARENT CEUX DES GÉNÉRATIONS QUI NOUS SUCCÉDERONT. ESCLAVE DU PASSÉ LE PRÉSENT EST MAÎTRE DE L AVENIR. L ÉTUDE DE L UN SERA TOUJOURS INDISPENSABLE POUR LA CONNAISSANCE DE L AUTRE.
1	33.33	23	LE CONTRASTE ENTRE L ORIENT ET L OCCIDENT EST AUJOURD HUI TROP GRAND POUR QU ON PUISSE JAMAIS ESPÉRER DE FAIRE ACCEPTER À L UN LES IDÉES ET LES FAÇONS DE PENSER DE L AUTRE. NOS VIEILLES SOCIÉTÉS SUBISSENT DES TRANSFORMATIONS PROFONDES ; LES RAPIDES PROGRÈS DES SCIENCES ET DE L INDUSTRIE ONT BOULEVERSÉ TOUTES NOS CONDITIONS PHYSIQUES ET MORALES D EXISTENCE. ANTAGONISME VIOLENT DANS LE CORPS SOCIAL ; MALAISE GÉNÉRAL QUI NOUS CONDUIT SANS CESSE À CHANGER NOS INSTITUTIONS POUR REMÉDIER AUX MAUX QUE CES CHANGE MENTS MÊMES ENGENDRENT ; DÉFAUT DE CONCORDANCE ENTRE LES SENTIMENTS ANCIENS ET LES CROYANCES NOUVELLES ; DESTRUCTION DES IDÉES SUR LESQUELLES AVAIENT VÉCU LES ANCIENS ÂGES. TEL AUJOURD HUI EST L OCCIDENT. FAMILLE PROPRIÉTÉ RELIGION MORALE CROYANCES TOUT CHANGE OU VA CHANGER. LES PRINCIPES DONT NOUS AVIONS VÉCU JUSQU ICI LES RECHER CHES MODERNES LES REMETTENT EN QUESTION. CE QUI SORTIRA DE LA SCIENCE NOUVELLE NUL NE POURRAIT LE DIRE. LES FOULES S ENTHOUSIASMENT MAINTENANT POUR QUELQUES THÉORIES TRÈS SIMPLES CONSTITUÉES SURTOUT PAR UN ENSEMBLE DE NÉGATIONS RADICALES ; MAIS LES CONSÉ QUENCES DE CES NÉGATIONS ELLES NE LES ENTREVOIENT PAS ENCORE. DES DIVINITÉS NOUVELLES ONT REMPLACÉ LES ANCIENS DIEUX. LA SCIENCE ACTUELLE ESSAIE DE LES DÉFENDRE QUI POURRAIT DIRE QU ELLE LES DÉFENDRA DEMAIN
1	33.33	22	IL Y A DONC BIEN DES QUESTIONS À RÉSOUDRE DANS L HISTOIRE DES ARABES

			ET PLUS D UNE LEÇON À RETENIR. CE PEUPLE EST UN DE CEUX QUI PERSONNIFIENT LE MIEUX CES RACES DE L ORIENT SI DIFFÉRENTES DE CELLES DE L OCCIDENT. L EUROPE LES CONNAÎT BIEN PEU ENCORE ; ELLE DOIT APPRENDRE À LES CONNAÎTRE CAR L HEURE APPROCHE OÙ SES DESTINÉES DÉPENDRONT BEAUCOUP DES LEURS..
...	...	...	

**Le tableau possède 23 données (1 niveau)*

Règle : {273} $\xrightarrow{max}$ {?}

Le segment 273 soit : LES ARABES AVANT MAHOMET

Msupport	Mconfiance	Y	Texte des segments
3	100	275	DES ARABES AVANT MAHOMET
2	66.67	309	MAIS AU MOMENT OÙ PARUT MAHOMET ELLE ÉTAIT MENACÉE D INVASIONS REDOUTABLES. L AN DE J. C. L YÉMEN QUI N AVAIT JUSQU ALORS OBÉI QU À DES SOUVERAINS ARABES AVAIT ÉTÉ ENVAHI PAR LES ABYSSINS QUI ESSAYÈRENT D Y PROPAGER LA RELIGION CHRÉTIENNE ET RÉUSSIRENT À CONVERTIR PLUSIEURS TRIBUS. EN C EST À DIRE FORT PEU DE TEMPS AVANT MAHOMET ILS FURENT CHASSÉS PAR LES PERSES QUI Y ÉTABLIRENT DES VICE ROIS. CES DERNIERS RÉGNÈRENT SUR L YÉMEN L HADRAMAUT ET L OMAN JUSQU À L ARRIVÉE DU PROPHÈTE.
2	66.67	312	AVANT MAHOMET
2	66.67	274	PRÉTENDUE BARBARIE
2	66.67	286	HISTOIRE DES ARABES AVANT MAHOMET
2	66.67	280	IL EN EST AINSI DE LA CIVILISATION DES ARABES AVANT MAHOMET. DIRE EXACTEMENT AUJOURD HUI CE QUE FUT CETTE CIVILISATION SERAIT DIFFICILE MAIS LES DOCUMENTS QUE NOUS POSSÉDONS SUFFISENT À MONTRER QU ELLE A EXISTÉ ET QU ELLE NE FUT PAS INFÉRIEURE PEUT

			ÊTRE À CES ANTIQUES CIVILISATIONS DE L ASSYRIE ET DE LA BABYLONIE IGNORÉES PENDANT SI LONGTEMPS MAIS QUE L ARCHÉOLOGIE MODERNE RECONSTITUE MAINTENANT.
...	...	...	

**Le tableau possède 86 données (1 niveau)*

Règle : {283} $\xrightarrow{\text{max}}$ {?}

Le segment 283 soit : L HISTOIRE N EST PAS RESTÉE AUSSI MUETTE SUR L ANCIENNE CULTURE DES ARABES QU ELLE L A ÉTÉ SUR D AUTRES CIVILISATIONS QUE LA SCIENCE MODERNE VOIT SORTIR AVEC ÉTONNEMENT DE LA POUSSIÈRE ; MAIS EÛT ELLE GARDE UN SILENCE COMPLET NOUS AURIONS PU ASSURER QUE LA CIVILISATION ARABE FUT BIEN ANTÉRIEURE À MAHOMET. IL NOUS AURAIT SUFFI DE RAPPELER QU A L ÉPOQUE DU PROPHÈTE LES ARABES POSSÉDAIENT DÉJÀ UNE LITTÉRATURE ET UNE LANGUE TRÈS DÉVELOPPÉES ET SE TROUVAIENT DEPUIS PLUS DE ANS EN RELATIONS COMMERCIALES AVEC LES PEUPLES LES PLUS CIVILISÉS DU MONDE ET RÉUSSIRENT EN MOINS DE CENT ANS À CRÉER UNE DES PLUS BRILLANTES CIVILISATIONS DONT LES SIÈCLES ONT GARDÉ LA MÉMOIRE.

Msupport	Mconfiance	Y	Texte des segments
3	100	284	OR UNE LITTÉRATURE ET UNE LANGUE NE S IMPROVISENT PAS ET LEUR EXISTENCE EST DÉJÀ LA PREUVE D UN LONG PASSÉ. LES RELATIONS SÉCULAIRES AVEC LES NATIONS LES PLUS CIVILISÉES FINISSENT TOUJOURS PAR CONDUIRE À LA CIVILISATION LES PEUPLES QUI EN SONT SUSCEPTIBLES ; ET LES ARABES ONT SUFFISAMMENT PROUVÉ QUE TEL ÉTAIT LEUR CAS. POUR AVOIR RÉUSSI ENFIN À CRÉER EN MOINS D UN SIÈCLE UN VASTE EMPIRE ET UNE CIVILISATION NOUVELLE IL FALLAIT DES APTITUDES QUI SONT TOUJOURS LE FRUIT DE LENTES ACCUMULATIONS HÉRÉDITAIRES ET PAR CONSÉQUENT D UNE LONGUE CULTURE ANTÉRIEURE. CE N EST PAS AVEC DES PEAUX ROUGES OU DES AUSTRALIENS QUE LES SUCCESSEURS

			DE MAHOMET EUSSENT CRÉÉ CES CITÉS BRILLANTES QUI PENDANT HUIT SIÈCLES FURENT LES SEULS FOYERS DES SCIENCES DES LETTRES ET DES ARTS EN ASIE ET EN EUROPE. BIEN D AUTRES PEUPLES QUE LES ARABES ONT RENVERSÉ DE GRANDS EMPIRES MAIS ILS N ONT PAS FONDÉ DE CIVILISATION ET FAUTE DE CULTURE ANTÉRIEURE SUFFISANTE ILS N ONT PROFITÉ QUE BIEN TARD DE LA CIVILISATION DES PEUPLES QU ILS AVAIENT VAINCUS. IL A FALLU DE LONGS SIÈCLES D EFFORTS AUX BARBARES QUI S EMPARÈRENT DE L EMPIRE ROMAIN POUR SE CRÉER UNE CIVILISATION AVEC LES DÉBRIS DE LA CIVILISATION LATINE ET SORTIR DE LA NUIT DU MOYEN ÂGE
2	66.67	290	NOUS IGNORONS QUELLES FURENT LES INFLUENCES DE MILIEUX ET DE CONDITIONS D EXIS TENCE QUI DÉTERMINÈRENT LA DIFFÉRENCIATION DES PEUPLES ISSUS DE LA RACE PRIMITIVE DONT NOUS VENONS DE PARLER ET NOUS NE POUVONS PAR CONSÉQUENT QU INDIQUER LEUR PARENTÉ AVEC LES ARABES LES SEULS DONT NOUS ALLONS NOUS OCCUPER MAINTENANT.
2	66.67	288	LES DÉBRIS RETROUVÉS DANS LES COUCHES GÉOLOGIQUES DU SOL ONT PROUVÉ QUE CET ÂGE DE LA PIERRE PRÉSENTE LES PLUS GRANDES ANALOGIES CHEZ LES DIVERS PEUPLES. AVEC CES ÉLÉMENTS IL A ÉTÉ FACILE DE RECONSTITUER LES CONDITIONS D EXISTENCE ET MÊME L ÉTAT INTELLECTUEL DE NOS PLUS LOINTAINS ANCÊTRES. C EST UN TRAVAIL QUE NOUS AVONS FAIT DANS NOTRE PRÉCÉDENT OUVRAGE ET SUR LEQUEL IL SERAIT INUTILE DE REVENIR ICI.
2	66.67	287	LES ARABES ONT EU COMME TOUS LES PEUPLES UNE PÉRIODE PRÉHISTORIQUE. L ÉTUDE DES DÉBRIS D ARMES D INSTRUMENTS DE DEMEURES LAISSÉS DANS LES COUCHES GÉOLOGIQUES DU GLOBE PAR NOS PRIMITIFS ANCÊTRES

			PROUVE QUE BIEN DES SIÈCLES AVANT LA COURTE DURÉE DES TEMPS DONT S OCCUPE L HISTOIRE ET PENDANT UNE PÉRIODE QUI NE PEUT SE CHIFFRER QUE PAR MILLIONS D ANNÉES L HOMME IGNORA LES MÉTAUX L AGRICULTURE L ART DE RENDRE LES ANIMAUX DOMESTIQUES ET N EUT QUE DES FRAGMENTS DE SILEX POUR ARMES. ON A DONNÉ À CETTE PRIMITIVE PÉRIODE LE NOM D ÂGE DE LA PIERRE TAILLÉE ET PARTOUT OÙ L ARCHÉOLOGIE PRÉHISTORIQUE A PORTÉ SES RECHERCHES EN ARABIE COMME EN EUROPE ET EN AMÉRIQUE ELLE A RETROUVÉ DES TRACES DE CETTE LOINTAINE ÉPOQUE.
2	66.67	289	LES PLUS ANCIENNES TRADITIONS DES ARABES NE REMONTENT PAS AU DELÀ D ABRAHAM MAIS LA LINGUISTIQUE NOUS PROUVE QU À UNE ÉPOQUE BEAUCOUP PLUS RECULÉE TOUTES CES VASTES RÉGIONS COMPRISES ENTRE LE CAUCASE ET LE SUD DE L ARABIE ÉTAIENT HABITÉES SINON PAR UNE MÊME RACE AU MOINS PAR DES PEUPLES PARLANT LA MÊME LANGUE. L ÉTUDE DES LANGUES DITES SÉMITIQUES DÉMONTRE EN EFFET QUE L HÉBREU LE PHÉNICIEN LE SYRIAQUE L ASSYRIEN LE CHALDÉEN ET L ARABE ONT UNE ÉTROITE PARENTÉ ET PAR CONSÉQUENT UNE COM MUNE ORIGINE.
1	33.33	259	L ARABE DE L ALGÉRIE N EST EN RÉALITÉ QU UN VÉRITABLE MÉTIS ET NOUS DEVONS NOUS ATTENDRE À TROUVER EN LUI TOUTES LES QUALITÉS INFÉRIEURES DES MÉTIS. LES HABITANTS SÉDENTAIRES DES VILLES SONT LES PRODUITS DU MÉLANGE DE TOUS LES PEUPLES CITÉS PLUS HAUT PRODUITS DÉGÉNÉRÉS PAR TOUTES LES DOMINATIONS QUI ONT PESÉ SUR EUX. MOINS MÉLANGÉS ET PARTANT BIEN MOINS DÉGÉNÉRÉS LES NOMADES SE RAPPROCHENT DES VÉRITABLES ARABES NOMADES DES AUTRES CONTRÉES ET COMME EUX SONT

			RÉFRACTAIRES À TOUTE CIVILISATION.
...	...	...	

**Le tableau possède 21 données (1 niveau)*

$\{285\} \xrightarrow{max} \{?\}$

Le segment 285 soit : AVANT D ESSAYER DE DÉCOUVRIR AU MOYEN DES FAIBLES DOCUMENTS QUE NOUS POSSÉ DONS CE QUE FUT LA CIVILISATION DES ARABES AVANT MAHOMET NOUS ALLONS RÉSUMER RAPI DEMENT CE QUE NOUS SAVONS DE LEUR HISTOIRE

Msupport	Mconfiance	Y	Texte des segments
3	100	286	HISTOIRE DES ARABES AVANT MAHOMET
2	66.67	288	LES DÉBRIS RETROUVÉS DANS LES COUCHES GÉOLOGIQUES DU SOL ONT PROUVÉ QUE CET ÂGE DE LA PIERRE PRÉSENTE LES PLUS GRANDES ANALOGIES CHEZ LES DIVERS PEUPLES. AVEC CES ÉLÉMENTS IL A ÉTÉ FACILE DE RECONSTITUER LES CONDITIONS D EXISTENCE ET MÊME L ÉTAT INTELLECTUEL DE NOS PLUS LOINTAINS ANCÊTRES. C EST UN TRAVAIL QUE NOUS AVONS FAIT DANS NOTRE PRÉCÉDENT OUVRAGE ET SUR LEQUEL IL SERAIT INUTILE DE REVENIR ICI.
2	66.67	292	LES LIVRES DES HÉBREUX RECONNAISSENT LEUR PARENTÉ AVEC LES ARABES ET CONSIDÈRENT CES DERNIERS COMME UN PEUPLE PLUS ANCIEN QU EUX MÊMES.
2	66.67	309	MAIS AU MOMENT OÙ PARUT MAHOMET ELLE ÉTAIT MENACÉE D INVASIONS REDOUTABLES. L AN DE J. C. L YÉMEN QUI N AVAIT JUSQU ALORS OBÉI QU À DES SOUVERAINS ARABES AVAIT ÉTÉ ENVAHI PAR LES ABYSSINS QUI ESSAYÈRENT D Y PROPAGER LA RELIGION CHRÉTIENNE ET RÉUSSIRENT À CONVERTIR PLUSIEURS TRIBUS. EN C EST À DIRE FORT PEU DE TEMPS AVANT MAHOMET ILS FURENT CHASSÉS PAR LES PERSES QUI Y ÉTABLIRENT DES VICE ROIS. CES DERNIERS RÉGNÈRENT SUR L YÉMEN L HADRAMAUT ET L OMAN JUSQU À L ARRIVÉE DU PROPHÈTE.

2	66.67	273	LES ARABES AVANT MAHOMET
2	66.67	275	DES ARABES AVANT MAHOMET
...	...	...	

**Le tableau possède 87 données (1 niveau)*

ANNEXE 3

TABLE DES MATIÈRES DU LIVRE « LA CIVILISATION DES ARABES »

La table des matières, que nous retrouvons dans « Gustave Le Bon (1884), La civilisation des Arabes : livres I et II » de la page 3 à 13, a servi à assister à l'identification du thème de chaque partie du livre.

Partie de la table utilisée pour la première partie du livre

Chapitre premier : Le milieu et la race

Chapitre I L'Arabie

Chapitre II Les Arabes

Chapitre III Les Arabes avant Mahomet

Partie de la table utilisée pour la première partie du livre

Chapitre deuxième : Les origines de la civilisation arabe

Chapitre I Mahomet. Naissance de l'empire arabe.

Chapitre II Le Coran

Chapitre III Les conquêtes des Arabes.

Partie de la table utilisée pour la première partie du livre

Chapitre troisième : L'empire des Arabes

Chapitre I Les Arabes en Syrie

Chapitre II Les Arabes à Bagdad

Chapitre III Les Arabes en Perse et dans L'Inde

Chapitre IV Les Arabes en Égypte

Chapitre V Les Arabes dans l'Afrique septentrionale

Chapitre VI Les Arabes en Espagne

Chapitre VII Les Arabes en Sicile, en Italie et en France

Chapitre VIII Luttes du christianisme contre l'islamisme. Les croisades.

Partie de la table utilisée pour la première partie du livre

Chapitre quatrième : Les mœurs et les institutions des Arabes.

Chapitre I Les Arabes nomades et Arabes sédentaires des campagnes.

Chapitre II Les Arabes des villes. - Mœurs et coutumes.

Chapitre III Institutions politiques et sociales des Arabes

Chapitre IV Les femmes en Orient.

Chapitre V Religion et morale.

Chapitre cinquième : La civilisation des Arabes.

Chapitre I Origine des connaissances des Arabes. Leur enseignement et leurs méthodes.

Chapitre II Langue, philosophie, littérature et histoire.

Chapitre III Mathématiques et astronomie.

Chapitre IV Sciences géographiques.

Chapitre V Sciences physiques et leurs applications

Chapitre VI Science naturelles et médicales

Chapitre VII Les arts Arabes. Peinture, sculpture, arts industriels.

Chapitre VIII L'architecture des Arabes.

Chapitre IX Commerce des Arabes. - Leur relation avec divers pays.

Chapitre X Civilisation de l'Europe par les Arabes. Leur influence en Occident et en Orient

Partie de la table utilisée pour la première partie du livre

Chapitre sixième : La décadence de la civilisation arabe.

Chapitre I Les successeurs des arabes. – Influence des européens en Orient.

Chapitre II Causes de la grandeur et de la décadence des Arabes. État actuel de l'islamisme.

ANNEXE 4

INTERFACES DE L'OUTIL DÉVELOPPÉ

La recherche théorique a été mise en œuvre dans une application implémentée en C#. Nous présentons ici les principales interfaces utilisateur.

1 Fenêtre principale

Ce formulaire permet de faire la gestion et la manipulation de texte grâce à la fonction d'ajout de modification et de destruction. Elle donne aussi accès à l'outil d'analyse générale par règle d'associations maximales.

Elle permet, en outre, de pouvoir réactualiser les fichiers où sont contenus les mots fonctionnels et ceux servant à la lemmatisation du texte lors d'un changement des bases de données.

Figure 24 - Fenêtre principale lors de l'exécution de logiciel.

Fonctions :

1) Permet de modifier ou poursuivre l'analyse (complété ou incomplète) d'un texte choisi dans la liste de la base de données.
2) Permet de détruire une analyse d'un texte choisi dans la liste de la base de données.
3) Permet d'ajouter un nouveau texte à la base de données et donc une nouvelle analyse de ce texte.
4) Donne accès à l'outil permettant l'analyse des textes de la base de données.
5) Permet de réactualiser les bases servant à la lemmatisation et à l'élimination des mots fonctionnels.
6) Permet de quitter le programme.
7) Permet de modifier les options prédéfinies.

1.1 Analyse de texte

Permet l'analyse d'un nouveau texte ou d'un texte choisi dans la liste de la fenêtre précédente. Chaque traitement est enregistré en mémoire et affiché dans la fenêtre prévue à cet effet. L'utilisateur est donc en mesure de suivre le processus en cour.

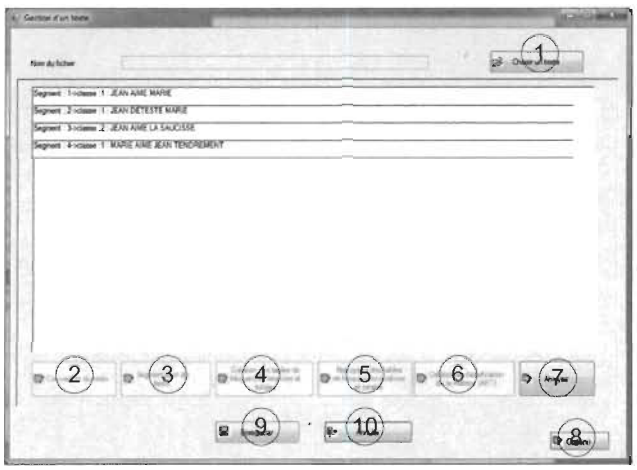


Figure 25 - Fenêtre de gestion du texte.

Fonctions :

1) Permet d'importer un texte situé sur un support matériel (clé USB, disques durs, etc.).
2) Permet de convertir le texte selon les options choisies (voir la gestion des options).
3) Permet de segmenter le texte selon les options choisies (voir la gestion des options).
4) Permet de construire les tables de distribution de fréquences relatives et totales qui serviront à bâtir la matrice utilisés par les classifieurs selon les options choisies (voir la gestion des options).
5) Donne accès à l'outil permettant le nettoyage des tables de distribution de fréquences relatives et totales.
6) Donne accès à l'outil permettant de procéder à la classification des segments.
7) Donne accès à l'outil permettant l'analyse du texte.
8) Permet de modifier les options.
9) Permet d'enregistrer l'analyse de texte en cours (lors de la modification, on peut aussi le sauvegarder sous un autre nom gardant intact le texte original).
10) Permet de quitter la gestion de texte et de revenir à la fenêtre principale.

** La touche CRTL-Z permet d'annuler la dernière opération et la touche CRTL-Y, d'annuler la dernière annulation.

1.2 Option

Voici les onglets dont se compose la fenêtre :

Figure 26 - Onglet de la fenêtre des options.

1) Onglet « Classifieur » : Permet de changer les options des classifieurs.

2) Onglet « Conversion du texte » : Permet de changer les options des opérations de conversion.
3) Onglet « Segmentation » : Permet de changer les options de segmentation.
4) Onglet « Unités d'information » : Permet de changer le type d'unités d'information utilisé.
5) Onglet « Règle d'associations maximales » : Permet de changer les paramètres des règles d'association maximales.

1.2.1 Onglet « Classifieur »

Permet de spécifier les paramètres par défaut des classifieurs

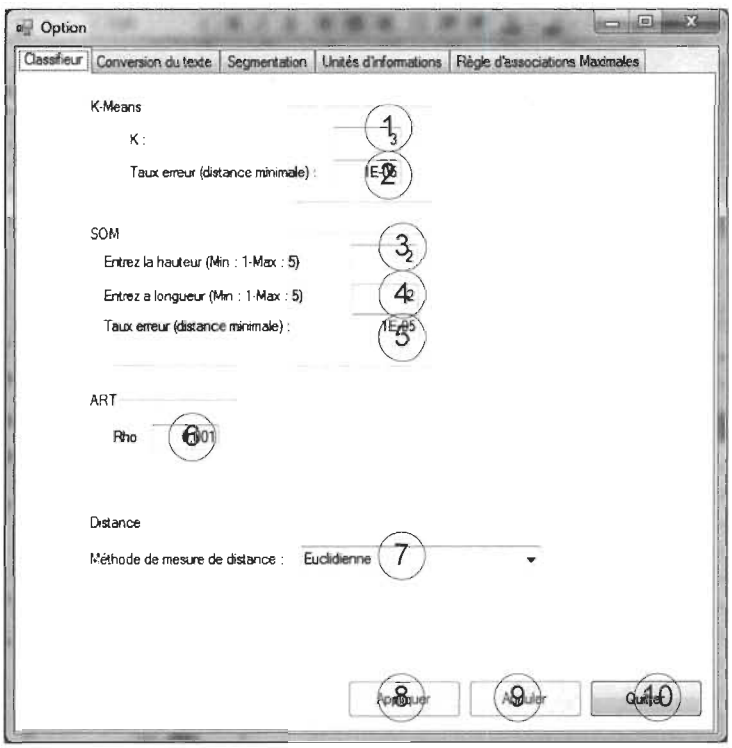


Figure 27 -Onglet « classifieur »

1) Permet de spécifier le nombre de centroïdes (K) du classifieur K-Means.

2) Permet de spécifier le taux d'erreur acceptable entre un vecteur et un centroïde.
3) Permet de spécifier la hauteur de la carte utilisée du classifieur SOM.
4) Permet de spécifier la largeur de la carte utilisée du classifieur de SOM.
5) Permet de spécifier le taux d'erreur acceptable entre un vecteur et un vecteur référent.
6) Permet de spécifier la valeur Rho du classifieur ART.
7) Permet de spécifier la mesure de distance utilisée lors de l'exécution du classifieur K-Means et du classifieur SOM.
8) Permet d'appliquer les modifications effectuées.
9) Permet d'annuler les modifications effectuées.
10) Permet de quitter la fenêtre d'options et de revenir à la gestion de texte.

1.2.2 Onglet « Conversion du texte »

Permet de changer les options des opérations de conversion.

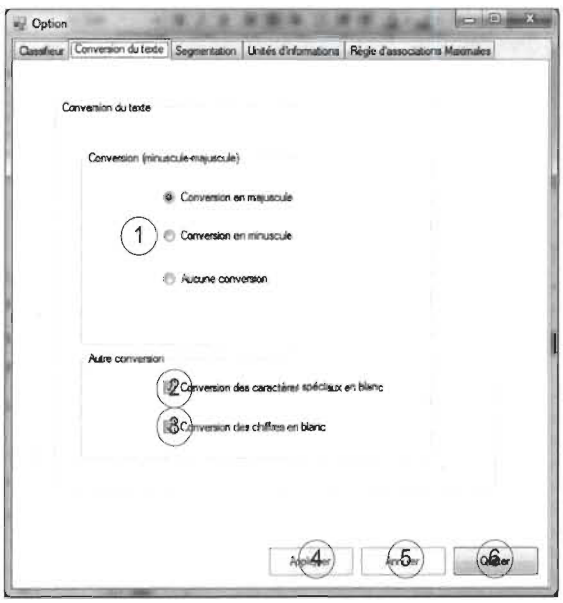


Figure 28 - Onglet « Conversion de texte »

1) Permet de spécifier le type de conversion entre :
    a. Remplacer tous les caractères par des caractères en majuscule.

b. Remplacer tous les caractères par des caractères en minuscule.

c. Ne rien faire.

2) Permet de remplacer les caractères spéciaux par des caractères blancs.

3) Permet de remplacer les caractères numériques par des caractères blancs.

4) Permet d'appliquer les modifications effectuées.

5) Permet d'annuler les modifications effectuées.

6) Permet de quitter la fenêtre d'options et de revenir à la gestion de texte.

1.2.3 Onglet « Segmentation »

Permet de changer les options de segmentation.

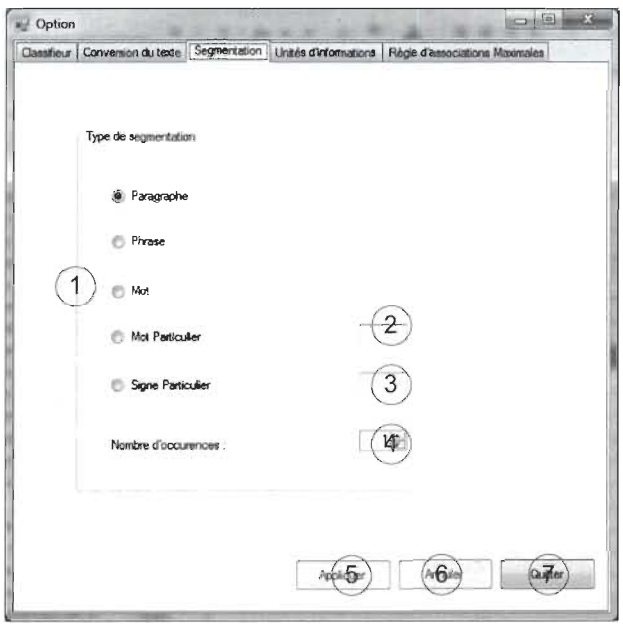


Figure 29 — Onglet « Segmentation »

1) Permet de spécifier que la gestion de la segmentation se fera par :

a. Paragraphe

b. Phrase
c. Mot
d. Grâce à un mot particulier
e. Grâce à un signe particulier

2) Permet de spécifier mot utilisé par l'option « mot particulier ».
3) Permet de spécifier signe utilisé par l'option « signe particulier ».
4) Permet de spécifier nombre d'occurrences observé avant de pouvoir segmenter.
5) Permet d'appliquer les modifications effectuées.
6) Permet d'annuler les modifications effectuées.
7) Permet de quitter la fenêtre d'options et de revenir à la gestion de texte

1.2.4 Onglet « Unités d'information »

Permet de changer le type d'unités d'information utilisé.

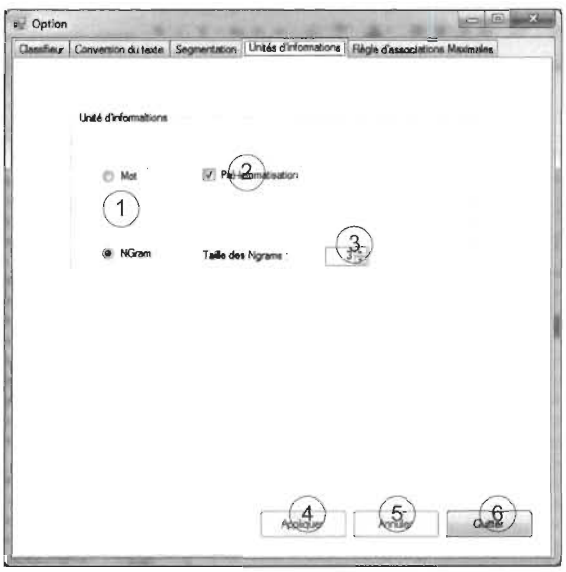


Figure 30 — Onglet « Unités d'information »

1) Permet de spécifier que l'unité d'information est le mot.
2) Permet de spécifier que l'unité d'information est le n-gram.

3) Permet de spécifier si les mots doivent subir une lemmatisation.
4) Permet de spécifier le nombre de caractères formant les n-grams.
5) Permet d'appliquer les modifications effectuées.
6) Permet d'annuler les modifications effectuées.
7) Permet de quitter la fenêtre d'options et de revenir à la gestion de texte.

1.2.5 Onglet « Règle d'associations maximales »

Permet de changer les paramètres des règles d'association maximales.

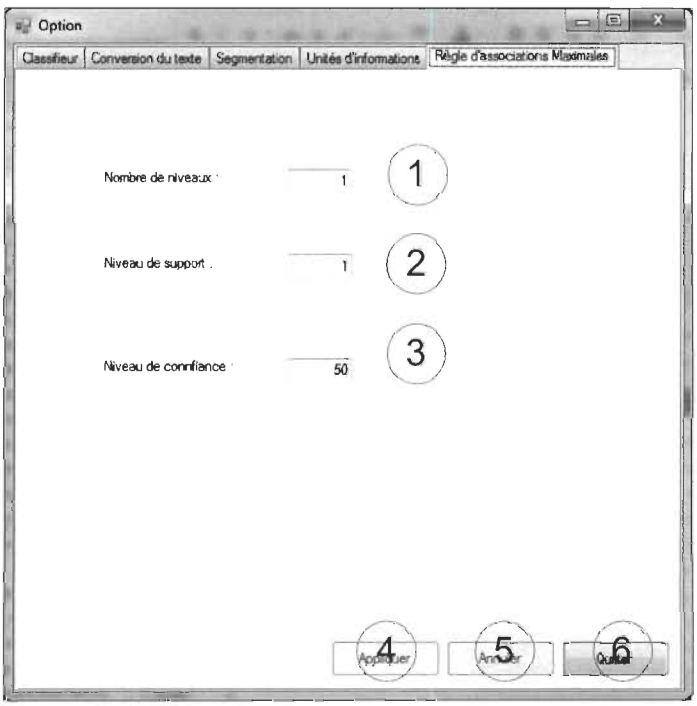


Figure 31 – Onglet « Règles d'association maximales »

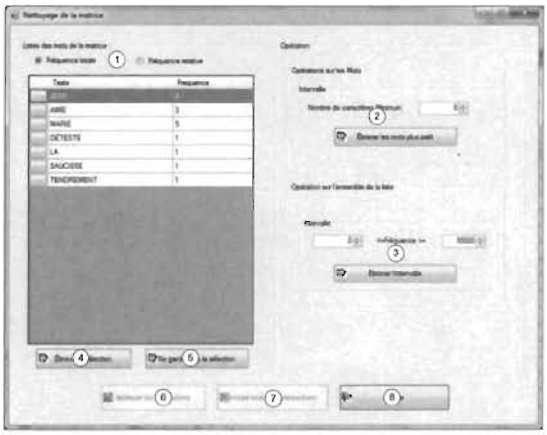
1) Permet de spécifier le nombre de niveaux utilisés lors de la formation des ensembles Y.
2) Permet de spécifier le niveau de support minimum que chaque règle doit respecter.
3) Permet de spécifier le niveau de confiance minimum que chaque règle doit respecter.

4) Permet d'appliquer les modifications effectuées.
5) Permet d'annuler les modifications effectuées.
6) Permet de quitter la fenêtre d'options et de revenir à la gestion de texte.

1.3 Nettoyage de la matrice

Permet de faire un nettoyage des mots (n-grams) dont est composée la matrice.

Par Mots : Par N-grams :



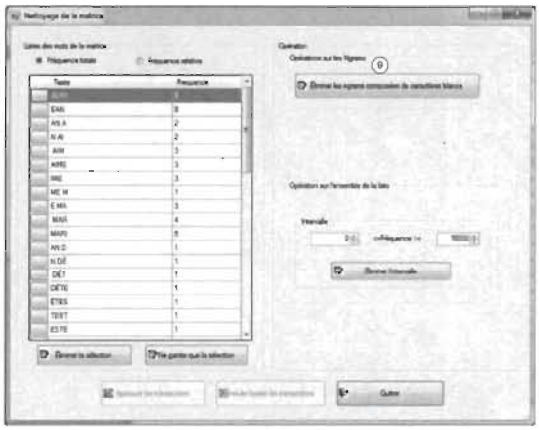


Figure 32 - Nettoyage de la matrice (mots) Figure 33 - Nettoyage de la matrice(Ngrams)

1) Permet d'afficher la liste selon la fréquence relative (par segments) ou la fréquence totale.
2) Permet d'éliminer les mots ne respectant pas un nombre de caractères minimums.
3) Permet d'éliminer les mots ne respectant pas une certaine fréquence.
4) Permet d'éliminer la sélection faite par l'utilisateur dans la liste.
5) Permet de ne conserver que la sélection faite par l'utilisateur dans la liste.
6) Permet de rendre effective les modifications en cours.
7) Permet d'annuler les modifications en cours.
8) Permet de quitter et revenir à la gestion de texte.

9) Permet d'éliminer les n-grams composées de caractères blancs.

1.4 Classification de la matrice

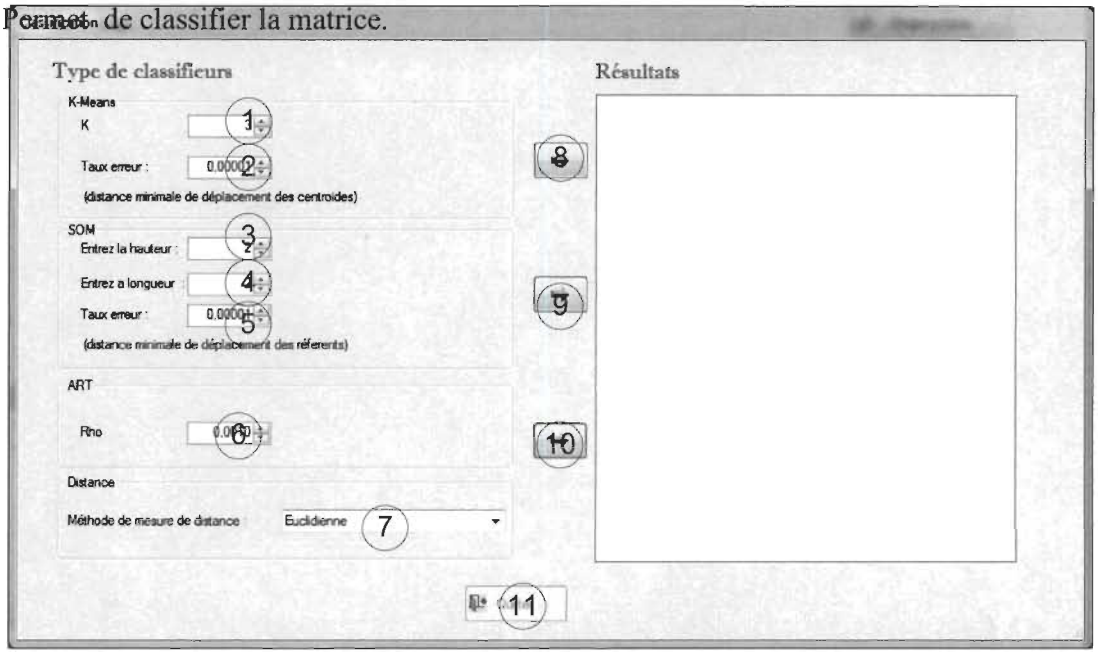


Figure 34 – Classfication des vecteurs afin de les relier à une classe distincte.

1) Permet de spécifier le nombre de centroïdes (K) du classifieur K-Means.
2) Permet de spécifier le taux d'erreurs acceptable entre un vecteur et un centroïde.
3) Permet de spécifier la hauteur de la carte utilisée du classifieur SOM.
4) Permet de spécifier la largeur de la carte utilisée du classifieur de SOM.
5) Permet de spécifier le taux d'erreur acceptable entre un vecteur et un vecteur référent.
6) Permet de spécifier la valeur Rho du classifieur ART.
7) Permet de spécifier la mesure de distance utilisée lors de l'exécution du classifieur K-Means et du classifieur SOM.
8) Permet l'exécution du classifieur K-Means et l'obtention des classes formées.
9) Permet l'exécution du classifieur SOM et l'obtention des classes formées.

10) Permet l'exécution du classifieur ART et l'obtention des classes formées.

11) Permet de quitter et revenir à la gestion de texte lorsque la classification est complétée.

1.5 Analyse de texte

Permet d'analyser le texte grâce aux classes obtenues par le classifieur sous différentes formes.

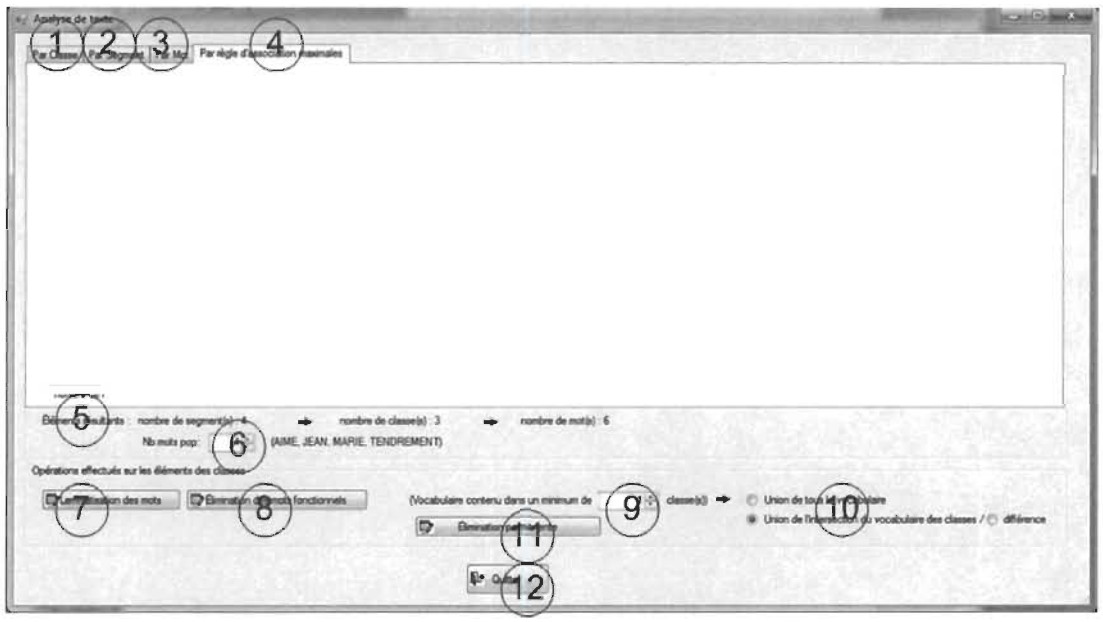


Figure 35- Analyse du texte selon différentes options.

1) Onglet « Par classe » : Permet de visualiser chaque classe, leur contenu ainsi que les segments associés à celles-ci.
2) Onglet « Par segment » : Permet de visualiser chaque segment, sa classe associée ainsi que les mots composant cette classe.
3) Onglet « Par mot» : Permet de visualiser chaque mot ainsi que les classes et les segments dans lesquels il est présent.

4) Onglet « Par règle d'associations maximales » : Permet de visualiser les règles d'association maximales.
5) Commentaire indiquant le nombre de segments obtenus lors de l'opération de segmentation, le nombre de classes issu des résultats de la classification et le nombre de mots obtenu lors de l'unification de toutes les classes résultantes.
6) Permet de choisir le nombre de mots les plus fréquents du texte qui pourront éventuellement former notre ensemble type X.
7) Permet de lemmatiser les éléments des classes.
8) Permet d'éliminer les mots fonctionnels tels que : « la », «le », etc.
9) Permet de choisir le nombre de classes minimal ou une unité d'information est présente.
10) Permet d'effectuer soit l'union de tous les vocabulaires des classes, soit l'intersection de tous les vocabulaires des classes, soit la différence.
11) Permet d'éliminer définitivement les éléments non-choisis par les opérations 9 et 10.
12) Permet de quitter et revenir à la gestion de texte.

1.5.1 Onglet « Par classe »

Permet de visualiser chaque classe, leur contenu ainsi que leurs segments associés.

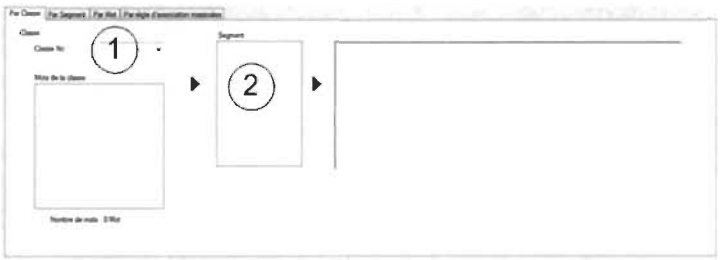


Figure 36 – Analyse du texte par classe.

1) Permet de choisir la classe.
2) Permet de choisir un segment et de le visualiser.

1.5.2 Onglet « Par segment »

Permet de visualiser chaque segment, sa classe associée ainsi que les mots composant cette dite classe.

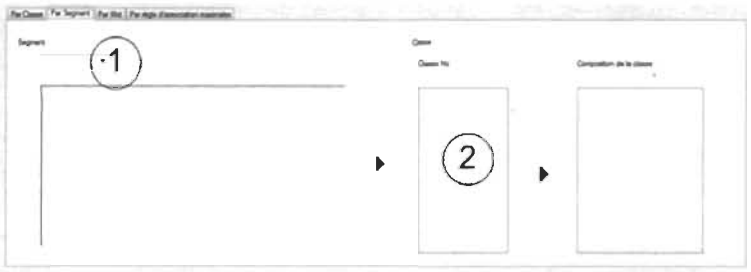


Figure 37 – Analyse du texte par segment.

1) Permet de choisir le segment.
2) Permet de choisir une classe et de visualiser son contenu.

1.5.3 Onglet Par mot»

Permet de visualiser chaque mot.

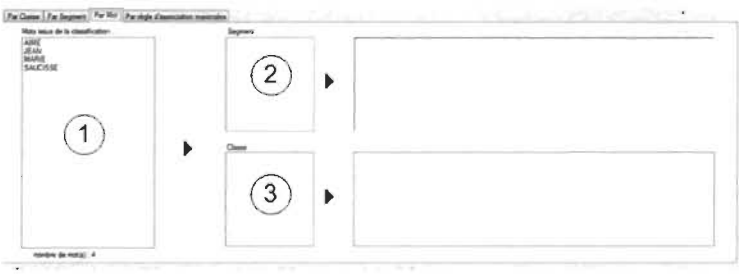


Figure 38 – Analyse du texte par mot.

1) Permet d'afficher les segments et les classes liés à ce mot.
2) Permet de sélectionner un segment et de le visualiser.
3) Permet de choisir une classe et de visualiser son contenu.

1.5.4 Onglet « Par règle d'associations maximales »

Permet d'analyser les règles d'association maximales.

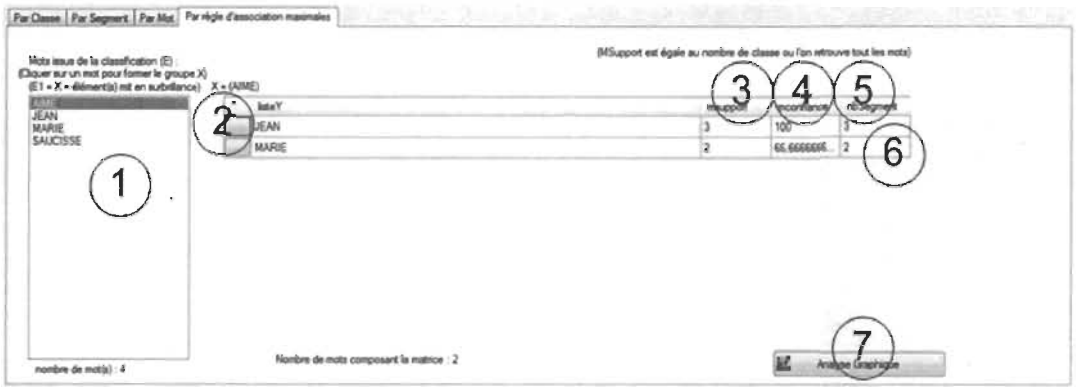


Figure 39 - – Analyse du texte par règles d'associations maximales.

1) Permet de choisir la composition d'E1 et X.
2) Permet d'atteindre la fenêtre d'analyse de la règle.
3) Permet d'afficher les segments correspondants.
4) Permet de trier la liste par ordre de support.
5) Permet de trier la liste par ordre de confiance.
6) Permet de trier la liste par ordre de nombre de segments.
7) Permet d'atteindre la fenêtre d'analyse graphique (Achouri, 2012)

1.5.5 Composition de la règle

Permet de visualiser chaque groupe formant la règle ainsi que la manière dont le support et la confiance ont été calculés.

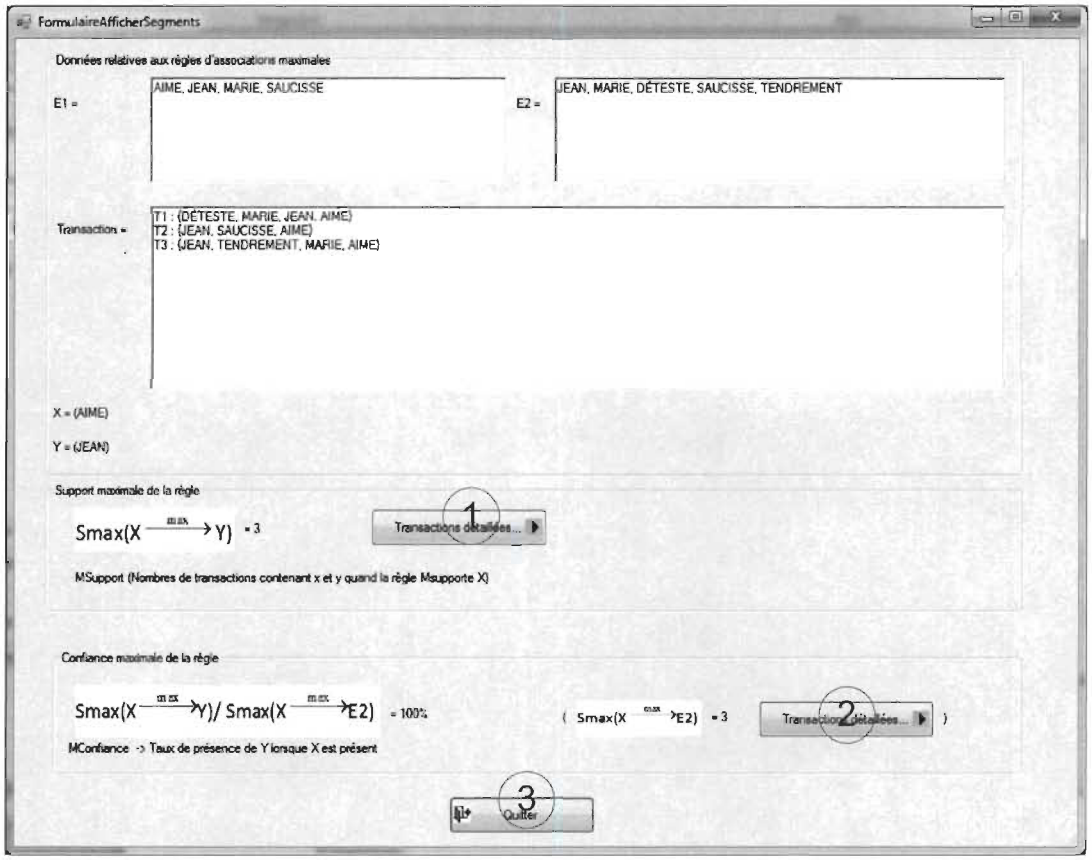


Figure 40 – fenêtre affichant les éléments composant la règle.

1) Visualiser les transactions en rapport avec le support d'une règle.
2) Visualiser les transactions en rapport avec la confiance d'une règle.
3) Permet de quitter et revenir à la gestion du texte.

1.5.6 Affichage

Permet de visualiser des segments spécifiques ou transactions dépendamment de la commande qui l'appelle. Soit :

1) Visualiser les segments correspondant au mot formé par une règle.
2) Visualiser les transactions en rapport avec le support d'une règle.

3) Visualiser les transactions en rapport avec la confiance d'une règle.

Figure 41 – Fenêtre affichant les segments

Figure 42 – Fenêtre affichant les transactions en lien au support de la règle.

Figure 43 - Fenêtre affichant les transactions en lien au confiance de la règle.

1.5.7 Analyse générale par règles d'associations

Permet d'analyser la combinaison de plusieurs textes à la fois

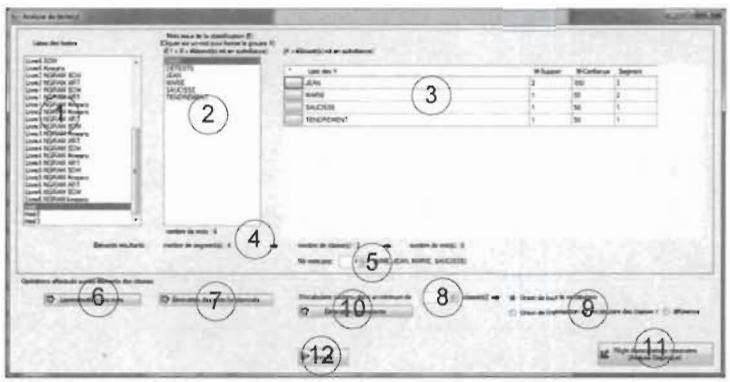


Figure 44 - Analyse de plusieurs textes

1) Permet de choisir la composition de X (E1).
2) Permet d'atteindre la fenêtre d'analyse de la règle.
3) Permet d'afficher les segments correspondants. On peut aussi changer l'ordre en pesant sur le titre voulu (Msupport, Mconfiance,...)
4) Indique le nombre de segments, classes et mots contenus.
5) Permet de choisir le nombre de mots le plus fréquents du texte qui formeront notre ensemble type X.
6) Permet de lemmatiser les éléments des classes.
7) Permet d'éliminer les mots fonctionnels tels que : « la », «le », etc.
8) Permet de choisir le nombre de classes minimal ou une unité d'information est présente.
9) Permet d'effectuer soit l'union de tous les vocabulaires des classes, soit l'intersection de tous les vocabulaires des classes, soit la différence.
10) Permet d'éliminer les éléments non-choisis par les opérations 9 et 10.
11) Permet d'atteindre la fenêtre d'analyse graphique (Achouri, 2012).
12) Permet de quitter et revenir à la gestion de texte.