Table des matières

Liste des tableaux

Liste des figures

Chapitre 1 - Introduction

1.1 Introduction

La croissance de l'utilisation des documents électroniques pousse le domaine informatique à évoluer très rapidement. En effet, les difficultés liées à la lecture, à la compréhension et à l'évaluation de certains textes créent le besoin d'avoir des applications plus sophistiquées qui permettent d'améliorer l'expérience des usagers à la lecture des livres, des articles, des notes, etc. D'après les recherches et les travaux antérieurs, on a noté l'insatisfaction des utilisateurs par rapport aux outils existants. D'autres ont rapporté des difficultés de compréhension, d'analyse, et de manipulation des textes relativement longs.

Dans certaines situations, un utilisateur est obligé d'analyser et de comprendre un long texte dans un temps assez réduit ce qui lui complique la tâche. La taille d'un texte représente une contrainte à la lecture. Avec la contrainte additionnelle du temps, le lecteur est ainsi obligé à effectuer une lecture plus rapide. Elle pourrait être assurée par l'extraction des passages les plus importants dans le texte pour avoir une première idée du contexte du contenu. Cette approche peut être une solution pour ce genre de problèmes.

Dans certaines situations, un utilisateur sera obligé d'analyser et comprendre un texte de très grande taille dans un temps très réduit ce qui lui cause différents problèmes. Aussi, la taille d'un texte représente une difficulté, car souvent l'utilisateur se trouve dans des situations urgentes. Ceci l'oblige d'effectuer une lecture rapide et une extraction des

passages importants dans le texte pour avoir une première idée du contexte du contenu. Ce qui nous amène à chercher pour trouver une solution pour ces différents problèmes.

Dans ce travail, nous nous proposons une méthode qui permet de faciliter la lecture et l'analyse d'un texte. En effet, face à un texte complexe, un simple lecteur peut avoir besoin de différents outils pour en simplifier la compréhension.

D'après les recherches et les travaux antérieurs, on a noté une certaine insatisfaction des utilisateurs par rapport aux outils existants. D'autres ont rapporté des difficultés de compréhension, d'analyse, et de manipulation des textes relativement longs [11].

Nous avons développé et implémenté une application qui permet d'assurer les options suivantes:

- La possibilité de lire différents formats de fichiers (.doc .docx, .txt ...) tout en tenant compte des difficultés liées aux caractères spéciaux.
- La possibilité de faire des modifications de police et de mise en page du texte. L'utilisateur peut changer l'affichage de son texte selon ses besoins.
- La possibilité d'enregistrer les documents "fichier" sous un autre nom ou dans un autre répertoire.
- La possibilité de recourir à plusieurs tâches d'édition du texte.

Du côté de l'analyse et de la compréhension du texte, notre application comporte différentes fonctions très utiles pour le lecteur.

- La recherche dans un texte.

- L'enregistrement des mots clés du texte de même que l'écriture et l'enregistrement des notes et des commentaires pour garantir la compréhension du texte.
- L'extraction des notes enregistrées dans une autre interface de l'application ou dans un fichier texte afin d'assister à la rédaction d'un résumé du texte par exemple, d'en extraire les idées principales.

Ces fonctions offrent à l'utilisateur plusieurs options d'analyse et de compréhension. Nous retrouverons les résultats de l'évaluation de notre application complétée par quinze utilisateurs. La majorité des utilisateurs a apprécié les différentes fonctionnalités offertes et ils n'ont pas trouvé des difficultés pour la manipuler. Certains usagers ont souligné l'importance de l'option de la prise des notes.

Dans le chapitre 2, nous présenterons les analyseurs de texte et décrirons leurs types. Nous présenterons la recherche des textes et l'extraction des données et nous détaillerons leurs domaines d'utilisation. Par la suite, nous définirons et catégoriserons les règles de classifications et nous décrirons le rôle de ces règles dans le fonctionnement de l'application projetée.

Dans le chapitre 3, nous allons expliquer en détails le fonctionnement de notre application à l'aide de différents exemples et illustrations. Nous y présentons les étapes à suivre pour l'analyse d'un texte et la rédaction d'un résumé par exemple. Nous allons décrire aussi les fonctions existantes et les résultats qu'ils donnent.

Le chapitre 4 présente les résultats sur l'analyse des textes. Nous avons utilisé un texte pour effectuer différentes tests sur les fonctionnalités existantes et nous avons critiqué les

résultats obtenus. Nous y trouvons également les résultats sur le sondage sur le fonctionnement de cette application.

Chapitre 2 - État de l'art

2.1 Introduction

Les travaux dans le domaine de l'analyse de texte répondent à différents besoins des utilisateurs. Depuis l'apparition des interfaces qui facilitent la lecture par la mise en forme et le changement de police selon l'exigence de l'utilisateur, les outils de lecture et d'analyse de textes ne cessent d'évoluer. En effet, les options recherchées et celles offertes sont en constante progression. Ce chapitre présente les développements dans ce domaine ainsi que les innovations telles que :

- Le surlignage des paragraphes dans le texte.
- L'ajout de commentaires dans le texte.
- L'insertion des signets.
- L'obtention des synonymes et des antonymes des mots.
- etc.

Ce chapitre décrit les méthodes utilisées, leurs types et les problèmes rencontrés. Nous y traitons également de quelques exemples des anciens travaux et de leurs limites.

2.2 Les analyseurs

2.2.1 Historique

L'histoire du domaine des analyseurs de texte se base essentiellement sur la catégorie, sur l'adaptation des modèles linguistiques et logiques à des contextes informatiques, et sur la mise au point des techniques d'ingénierie du langage [11].

Coulomb et Kayser ont créé deux modèles basés sur la langue naturelle. Le premier est nommé le modèle philosophique. Il facilite l'utilisation de la langue avec des programmes qui servent à automatiser la compréhension des langues naturelles. L'autre modèle est appelé « modèle ergonomique ». Il est basé sur la création et l'emploi des nouvelles méthodes qui facilitent la compréhension des langages [23].

Les recherches dans ce domaine ont commencé depuis 1963 avec la programmation de modèles logico-sémantiques. À partir de 1974, les recherches ont rencontré des problèmes de représentation et d'organisation des connaissances avec des modèles cognitifs.

L'analyse des données textuelles permet de considérer les textes comme des données pouvant être manipulées par des programmes et méthodes informatiques.

Les analyses et les tests informatiques, inspirés par la linguistique structurelle et l'analyse du discours, cherchent à qualifier les données textuelles à l'aide de catégories et à les quantifier par des analyses statistiques des éléments de texte [44].

2.2.2 Analyse de discours

L'analyse du discours est un processus de recherche qui permet d'examiner le contenu du discours oral ou écrit. Elle est née dans les années 50, avec les travaux de Zellig Harris parus dans son article « Discourse Analysis » [46]. Cette approche est développée et améliorée en France, Grande-Bretagne et aux États-Unis à partir des années 1960.

L'analyse du discours utilise des méthodes de classifications, d'interprétation et de lecture pour traiter le texte du discours [45].

2.2.3 Analyse syntaxique

L'analyse syntaxique est une technique d'étude des langues naturelles [4]. Généralement elle est réalisée à partir de l'analyse lexicale qui permet le découpage du texte en des flux [47]. Cette méthode permet de découper le texte en phrase; petite unité formée d'un sujet, verbe et des compléments. La méthode de l'analyse syntaxique est très importante dans le traitement automatique de la langue naturelle. En effet un analyseur syntaxique permet d'identifier la fonction syntaxique de chaque mot dans la phrase d'un texte [41]. En informatique, il permet d'expliquer et de reformuler un programme sous la forme d'un arbre appelé arbre de syntaxe abstrait. L'analyseur syntaxique est basé sur deux approches :

- Les analyseurs symboliques permettent de créer une base très puissante d'informations linguistiques. En effet, ils décrivent la nature syntaxique des mots.
- L'analyseur probabiliste se base sur un modèle obtenu. Il reçoit ses résultats des grammaires probabilistes.

2.2.4 Analyse sémantique

L'analyse sémantique permet d'extraire le sens et la signification des phrases du texte. Elle permet d'accomplir des tâches relatives à la sémantique lexicale ou à la compréhension du sens du texte. L'analyse sémantique permet d'obtenir une structure à partir du texte et ainsi d'offrir des solutions à certains problèmes rencontrés. Par exemple, dans les cas où les mots sont identiques, mais utilisés dans des contextes différents.

- L'avocat a une seule graine.
- L'avocat a mangé une pomme dans la cour.

Dans le premier cas, les mots avocat signifient le fruit. Dans le deuxième cas, l'avocat signifie l'homme de loi. Ainsi un avocat, au sens d'humain, peut manger, mais avec une signification de fruit, un avocat peut-être mangé [9].

2.3 Fouille de texte

2.3.1 Historique

La fouille de texte, l'extraction de connaissance dans les textes ou le « Text mining » est incluse dans le domaine de l'intelligence artificielle. Elle est apparue durant les années 1990 aux États-Unis comme une nouvelle technique de l'intelligence artificielle [48]. Les premiers algorithmes de fouille de données ont permis de réduire les volumes importants des données qui ont été utilisées dans le domaine du marketing [49].

2.3.2 Domaine d'utilisation

La technique de fouille de texte est caractérisée par deux traitements principaux :

- L'analyse : Elle consiste à connaitre les unités de la langue et à distinguer leurs fonctions grammaticales, leurs relations et leurs sens.
- Interprétation de l'analyse : Ce traitement permet de sélectionner un texte parmi d'autres. Par exemple, il facilite l'identification des spam dans les courriels [36].

Les méthodes de fouille de textes permettent de résoudre les problèmes liés aux traitements automatiques des données textuelles de grande taille en s'appuyant sur des méthodes linguistiques et statistiques [39].

2.4 La lecture assistée par ordinateur

2.4.1 Présentation

Le développement de l'intelligence artificielle et du traitement et de l'analyse des textes ont permis de résoudre de nombreuses difficultés liées à la recherche scientifique. Ces technologies facilitent le travail en permettant de réduire le temps consacré aux traitements des textes [14].

La lecture assistée par ordinateur a évolué grâce aux nouvelles technologies et en s'adaptant aux besoins de l'utilisateur. L'ordinateur ne peut pas remplacer le raisonnement et la pensée humaine, mais avec le temps, la capacité atteinte et sa flexibilité permettent d'offrir de plus en plus d'options à l'utilisateur. Les avancées de la recherche se manifestent par l'amélioration des lecteurs, par leurs fonctionnalités et par leurs ergonomies.

Il arrive parfois que nous soyons forcés d'imprimer un texte au lieu de le lire à l'écran de l'ordinateur afin de souligner les mots clés et d'y effectuer des opérations manuelles qu'on ne peut pas exécuter automatiquement. Ces inconvénients sont causés par le manque

de fonctionnalités offertes par certains lecteurs. Pour pallier à ces lacunes, il est important d'améliorer les systèmes existants [29].

2.4.2 Travaux récents

Meunier [30] a présenté l'importance de l'interface du lecteur du texte. Il y mentionne les techniques du soulignement et de la mise en forme des paragraphes qui simplifient la lecture et la compréhension du texte. Il nous indique également d'autres fonctionnalités permettant de sauvegarder les mots clés, d'exporter un texte sous d'autres formats afin de le modifier ou de le réorganiser. Aussi, il existe des fonctions qui permettent de segmenter le texte afin de faciliter son analyse basée sur des méthodes de classification. Ce travail permet de produire des lexiques et offre des méthodes qui facilitent la production de résumés. L'affichage des courbes ou graphes statistiques permet également une interprétation générale du texte. D'autre part, l'utilisateur peut modifier son texte en se basant sur des notes enregistrées. Meunier a développé un lecteur de texte. Il emploie des annotations qui facilitent la compréhension et l'analyse du texte. Ce lecteur est dédié aux utilisateurs experts. Ainsi, il facilite la manipulation du texte, la production de résumé, l'extraction des mots clés, la modification ou la suppression de passages du texte, etc.

D'autre part, Lucas et al. [40] décrivent la façon de compter le nombre d'occurrences des mots dans un texte et expliquent comment la réduction des occurrences facilite l'analyse du texte. L'analyse du texte est traitée selon une des deux méthodes suivantes ; la méthode supervisée ou la méthode non supervisée.

La méthode supervisée représente un outil qui permet de reproduire une tâche faite sur un petit échantillon, et répétée à la grandeur du document. La classification représente un

moyen de l'apprentissage supervisé. Cette méthode facilite la vérification de la performance du fonctionnement. Par exemple, elle vérifie l'existence d'un spam ou non dans un courriel, c'est une tâche difficile, mais donne 70% de bons résultats.

La méthode non supervisée organise le texte d'une autre manière. Pour structurer le texte, cette méthode se base sur l'analyse des résultats obtenus. La méthode non supervisée exige de l'utilisateur un ajustement de la structure du travail.

Biskri et al. [2] présentent GRAMEXCO qui est une application d'analyse textuelle. Elle est basée sur les n-grammes, une unité d'information possible pour toutes les langues. Cette application nécessite l'intervention de l'utilisateur pour ajuster certains paramètres lors du fonctionnement.

Dans une autre étude, Brier et al. [6] présentent les façons et les techniques d'analyse du texte assistée par ordinateur en se basant sur l'analyse de contenus quantitatifs. Cette technique utilise des dictionnaires et permet de donner le nombre d'apparitions des mots du texte. Les chercheurs ont utilisé des techniques hiérarchiques et non hiérarchiques de classification pour indiquer et identifier les liens entre la linguistique computationnelle et l'analyse des données relationnelles.

Cette application contient une interface qui permet d'afficher graphiquement les groupes dérivés du texte sous trois catégories. Il y a d'abord les mots rares qui ne se produisent pas assez fréquemment dans le texte. Viennent ensuite les mots les plus communs avec un taux d'occurrence de 500 fois et finalement le troisième groupe comprend les mots dont la fréquence est supérieure à celle de la moyenne de tous les mots contenus dans le texte. Cette technique facilite la compréhension du texte, car les mots les plus présents dans le texte portent des informations sur le sens général du contenu.

Les mots avec une faible occurrence dans le texte présentent un niveau d'association faible par rapport à ceux qui ont un très grand nombre d'apparitions. Cette méthode montre le choix des mots dans un discours ainsi que leur équilibre et leur utilisation.

Reich et al. [38] décrivent les difficultés des étudiants dans la recherche des commentaires sur un texte à évaluer. En effet, les étudiants cherchent toujours à analyser leurs textes, à mettre des commentaires et à sauvegarder les mots clés. Les chercheurs ont proposé une méthode d'analyse de texte qui facilite le traitement des documents avec un temps d'exécution très rapide. L'analyse assistée par ordinateur est disponible avec des méthodes supervisées qui sont basées sur l'intervention de l'utilisateur. Le rôle de l'utilisateur est de faire la lecture puis d'étiqueter un ensemble de données. Le projet de Reich et al. se base sur une méthode d'analyse non supervisée appelée « sujet de modélisation ». Elle permet d'extraire les relations sémantiques entre les cooccurrences des mots. S'il y a deux mots qui se répètent dans plusieurs documents, alors ils font référence à un concept ou un thème similaire.

Les méthodes non supervisées permettent de faciliter l'analyse avec des hypothèses a priori sur le contenu du texte. Mais l'utilisateur a quand même la responsabilité des différentes interprétations et du jugement des résultats obtenus.

A.Hearst et al. [20] présentent les façons d'analyser un texte littéraire. Il s'agit d'un système qui offre des outils de traitement du texte selon des algorithmes qui permettent l'interprétation et l'exploration des données textuelles. Ce système est basé sur la recherche grammaticale, la recherche de similarité contextuelle, la visualisation de modèles de contexte des mots, l'examen et l'organisation du texte pour effectuer une comparaison et construire des hypothèses.

Il est orienté vers la littérature et l'étude de la langue. En effet, il facilite la collecte d'information et assure son organisation. Il explore les mots liés entre eux et il donne l'option d'écrire des notes. Ce système est utilisé actuellement par des groupes de chercheurs pour analyser le récit «The North American Slave», les écrits de Stephen Crane et de William Shakespeare.

Le travail de Hearst et al. permet d'afficher le nombre d'apparitions des mots dans la liste des phrases où le mot cible s'affiche au centre et les mots de contexte à sa gauche ou à sa droite. Il offre aussi des méthodes pour extraire des relations entre les mots en permettant à l'utilisateur de choisir deux mots clés de même que la relation entre-eux à l'aide d'un menu déroulant.

L'utilisateur a une interface qui permet de voir les résultats de la recherche grammaticale et de les évaluer. Ce système simplifie la détection automatique du lien et du rapport entre les phrases grâce à l'extraction des relations entre les mots.

2.4.2.1 Google text editor

L'éditeur de texte Google représente une solution en ligne pour les utilisateurs. Il facilite la création, la lecture et l'analyse du texte. Il est basé sur les n-grammes. Généralement Google utilise les Trigrammes dans son éditeur de texte en se basant sur le modèle de Markov et la probabilité bayésienne.

Le moment où l'utilisateur commence à taper son texte, l'éditeur lui fournit des suggestions des mots selon les deux derniers mots écrits.

Par exemple : Si l'utilisateur commence son texte en tapant une première lettre alors l'éditeur lui recommande des sujets selon la première lettre écrite. Sinon, l'éditeur teste les

mots déjà écrits et donne des recommandations sur les mots qui manquent (complément, verbe...).

L'éditeur de Google est très efficace pour réduire le temps d'analyse du texte, mais il dépend aussi de la vitesse à laquelle l'utilisateur tape au clavier [18].

2.4.2.2 Les liseuses électroniques

Un livre électronique ne prend pas beaucoup d'espace physique comparé à un livre papier. En fait un seul lecteur (liseuse, tablette, etc.) peut contenir plusieurs centaines de livres.

La liseuse est une nouvelle technologie qui facilite aux lecteurs l'achat des livres en ligne à un prix plus économique.

La lecture du texte contient toujours des fonctionnalités standards pour faciliter différentes tâches pour l'utilisateur (mise en forme, recherche, affichage ...). La Liseuse offre aussi d'autres fonctionnalités pour simplifier l'analyse du texte [50] :

- Le surlignage de passages de texte : cette fonction permet de marquer quelques phrases pour que l'utilisateur puisse les identifier après ;
- la prise de notes ou l'ajout de commentaires ;
- la possibilité de partager des phrases ou d'émettre des commentaires sur les médias sociaux ;
- l'ajout d'annotations : l'utilisateur peut écrire des commentaires sur un passage ;
- l'ajout de signets pour retrouver une page plus facilement ;

- l'utilisation de dictionnaires qui est très utile pour connaître une définition ou le sens donné à un mot ;
- la traduction puisque plusieurs liseuses comportent un traducteur.

2.4.2.3 Intelligent text editor pour IOS

L'application Prmac est destinée aux utilisateurs des systèmes iOS et Mac OS X. Il s'agit d'un logiciel éditeur de textes multiplateformes. Prmac est caractérisé par sa rapidité de travail, son efficacité et sa simplicité.

Pour assurer une synchronisation entre un Mac, un iPhone, iPod ou bien un iPad, cette application utilise Dropbox. Il suffit juste de se connecter sur Dropbox à partir de cette application et sélectionner le fichier désiré pour l'utiliser.

Cette application peut être utilisée hors connexion. En effet, on peut avoir accès à tous les documents. L'utilisateur peut travailler sur son document n'importe où, et dès qu'il est en ligne, il lui suffit de faire une synchronisation pour remettre à jour son application.

Avant de faire des modifications sur le document, cette application offre une option qui permet de prendre une image instantanée du document. Cette fonctionnalité permet d'assurer une sauvegarde du fichier. Au cas où les changements ne sont pas désirés, l'utilisateur peut revenir à son document initial grâce à l'image prise.

Cette application donne à l'utilisateur l'option de modifier son texte en plein écran pour assurer un bon affichage dans les tablettes iPad ou bien dans l'iPhone vu la taille réduite de son écran.

Si l'utilisateur utilise une machine qui ne contient pas cette application, il peut exporter son fichier sous format (.txt). Il peut alors importer le fichier de cette application et le mettre à jour [51].

2.4.2.4 Google sheets add-on

Pour améliorer le fonctionnement du logiciel d'analyse de texte API. Google a créé l'add-on (un module complémentaire) pour corriger certains défauts. L'API a été développée pour résoudre les problèmes des développeurs en rendant les documents web plus significatifs. En effet, il facilite l'analyse du texte à tous les usagers. Ce qui donne plus de clarté et plus d'expressivité aux données du Web. Cette application est simple et ne nécessite pas une intervention technique ou bien une installation locale. Elle est incluse dans le navigateur Web.

Les fichiers Google add-on permettent d'effectuer des analyses des sentiments. Ils facilitent l'analyse des textes longs. Ils assurent également l'extraction des données. Ils permettent de détecter la langue du texte et créer les marqueurs de métadonnées (les hashtags). Ils facilitent la classification de données textuelles selon des catégories et donnent la possibilité de les extraire [13].

2.5 Les règles de classification

2.5.1 Introduction

Le chercheur suédois Linné est le plus connu depuis le $18^{ième}$ siècle dans le domaine de la classification du monde vivant en général. Mais avec l'évolution du domaine informatique au $20^{ième}$ siècle, plusieurs algorithmes de classifications sont apparus.

Il s'agit de regrouper des objets similaires entre eux selon des critères et des stratégies bien définis. La classification des données sert à affecter des individus à un groupe ou à une classe. Elle est représentée sous deux catégories [13]:

- La classification non hiérarchique ou partitionnement : Permet d'analyser le groupe des individus en m groupes disjoints. Le nombre m est fixé.
- La classification hiérarchique : Lorsque deux individus se trouvent confondus dans le même groupe, alors avec une précision élevée, ils seront dissociés vers deux groupes distincts [32].

Les méthodes de classification s'articulent en plusieurs phases. On s'intéresse généralement à la phase qui élabore des règles de classification des connaissances disponibles et qui s'appelle la phase d'apprentissage. Cette méthode s'appuie sur l'apprentissage déductif ou inductif. Les algorithmes d'apprentissage inductifs permettent de produire des règles, de nouvelles normes à la suite des nouveaux cas rencontrés. Ce type d'algorithmes utilise plusieurs méthodes de classification. Citons la méthode des K moyenne ou méthode des plus proches voisins, la méthode bayésienne, la méthode d'analyse discriminante, l'approche des réseaux de neurones et la méthode de l'arbre de décision.

La résolution des problèmes de classification nécessite la combinaison des deux types d'apprentissage dans les cas des problèmes. Le besoin de combiner les deux types d'apprentissage permet de créer de nouvelles méthodes de classification [25].

2.5.2 Définition

La classification est l'action d'affecter dans diverses catégories, les objets ayant des caractéristiques ou des attributs communs. Une variété de techniques de classification a été proposée dans différents domaines [19].

Pour vérifier que notre utilisation des méthodes de classification est correcte, il faut prouver que notre méthode suit les conditions nécessaires [37] :

- Il faut bien choisir les mesures d'éloignement (dissimilarité, distance) entre les individus ;
- il faut bien adopter les critères d'homogénéité des classes à améliorer ;
- il faut bien mesurer la qualité de la classification.

2.5.3 Principe

La classification est la méthode qui simplifie la recherche des classes dans des groupes différents. Elle représente un travail préparatoire au classement qui représente la dernière tâche d'affectation des objets sous des classes différentes [28].

2.5.4 Méthodes de classification

2.5.4.1 L'arbre J48

Cette méthode est basée essentiellement sur l'arbre de décision. Il s'agit d'une fonction similaire à un arbre qui commence par les racines et aboutit aux feuilles. Le principe de ce type de méthode est de différencier des objets selon des classes.

L'arbre J48 est une fonction open source de l'algorithme C4.5 [14].

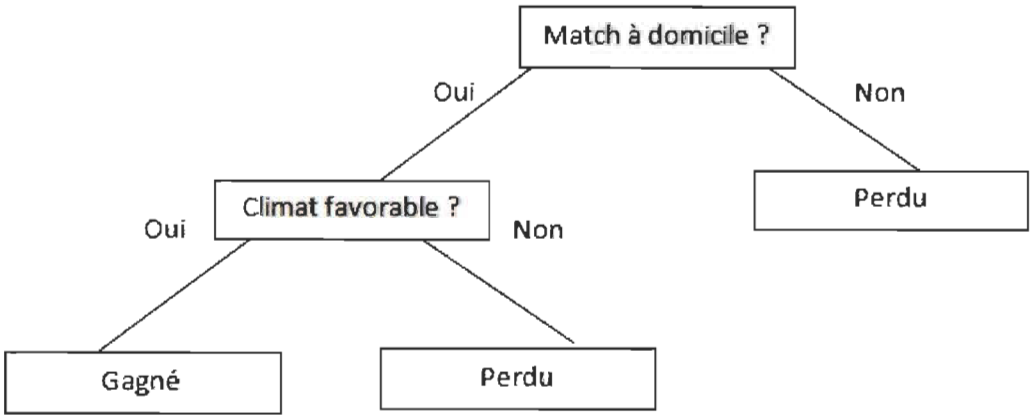

Figure 1: Exemple d'arbre de décision

Algorithme

L'algorithme J48 est basé sur une méthode récursive :

- Il faut initialiser l'arbre vide, les racines et les nœuds.
- Il faut tester si le nœud courant est terminal alors il faut lui affecter une classe. Sinon il faut commencer à créer le sous-arbre.
- Il faut passer au nœud suivant maintenant et lui affecter les mêmes tests jusqu'à obtenir un arbre final.

2.5.4.2 Méthode neuronale

La méthode neuronale est formée de plusieurs entrées qui sont formées par des variables de neurones. Les informations passent toujours par la couche d'entrée vers la couche de sortie qui fait partie de la sortie système [14].

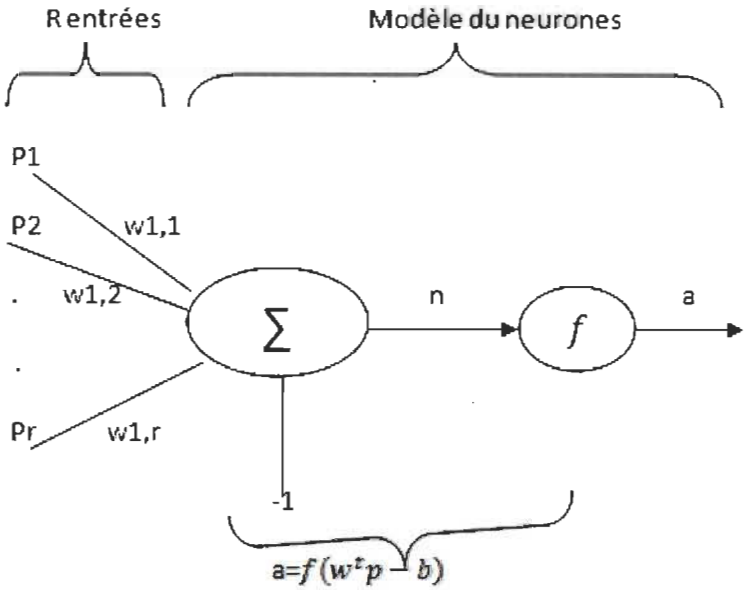

Figure 2 : Perceptron de neurone

R : représente les entrées des neurones.

Suite à l'intégration de la valeur de n, on obtient le résultat a du neurone.

L'équation de n est définie comme suit [19]:

$$n = \sum_{j=1}^{R} w_{1,j}\, p_j - b = w_{1,1}\, p_1 + \cdots + w_{1,R}\, p_R - b$$

Algorithme :

- Il faut initialiser le neurone de départ.
- L'algorithme fait les calculs de la valeur de n décrite dans l'équation précédente.
  - Chaque fois que le calcul est fait, l'algorithme donne un neurone et ses paramètres [52].

2.5.4.3 SVM

SVM (support, vector, machine) est une méthode de classification binaire avec un apprentissage binaire qui est inspiré des formules statistiques de Vladimir Vapnik. Elle consiste à utiliser des fonctions appelées noyaux qui facilitent le partage des données selon leurs critères.

Le SVM est déterminé par la méthode SMO (Sequential minimal optimisation) [53].

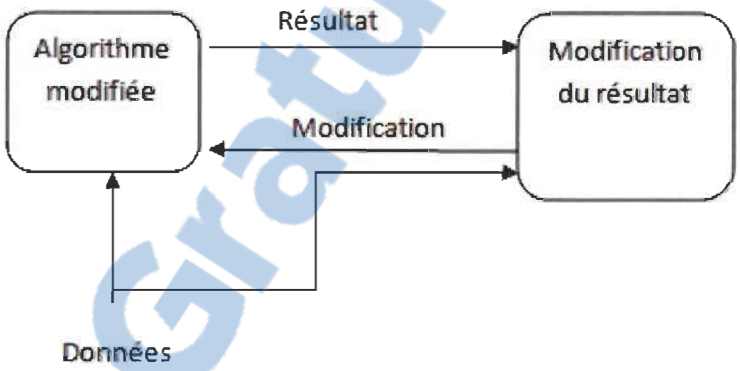

Figure 3 : Méthode SVM

Le principe de cette méthode se présente dans les résultats obtenus de l'algorithme principal, et qui se transforme ensuite en entrée pour un autre algorithme avec les données initiales. Tout ce traitement peut engendrer la modification du programme principal s'il y a des changements dans les résultats obtenus [31].

2.5.1 Calcul de distance

Pour calculer le taux de similarité entre les objets, il faut utiliser une notion de mesure selon le type de données existants. La majorité des algorithmes de classification s'articulent

sur une mesure de distance entre les objets. Cependant, il existe plusieurs méthodes pour calculer ces distances et la méthode choisie aura une influence sur le résultat.

Généralement les algorithmes des méthodes de classification utilisent des mesures euclidiennes pour calculer la distance. Ce choix est la conséquence de l'utilisation des valeurs numériques (exemple : les nombres, le diamètre, le calcul d'aire, etc. Mais dans notre cas, on va calculer les similarités des mots avec les méthodes de Hamming et la distance de Jaccard qui se base sur le calcul de la similarité et la fréquence des objets [25].

2.5.1.1 Distance de Hamming

La distance de Hamming est définie par Richard Hamming. Elle est utilisée généralement en informatique et précisément en traitement de signal, théorie de l'information, théorie des codes et la cryptographie. Elle permet de mesurer la différence entre deux vecteurs binaires. Pour chaque suite de vecteurs, elle permet de calculer sa distance au sens mathématique et regroupe le nombre de situations où les deux objets diffèrent.

La distance de Hamming et le poids (le poids d'un mot est le nombre de symboles non nuls qu'il contient) ont une importance sur les corrections des erreurs.

La distance entre deux objets a et b est le nombre d'éléments de a qui diffère de celle de b [54].

$$\forall a,b \in F\ a = (a_i)i \in [0, n-1] et\ b = (b_i)i \in [0, n-1]$$

$$H(A,B) = |A \cup B| - |A \cap B|$$

Soit les deux vecteurs a = (1, 1, 0, 1, 0) et b = (0, 0, 1, 1, 0)

H (a, b)=|a ∪ b|-|a ∩ b|= 4-1=3.

Dans l'exemple suivant, la distance de Hamming entre les deux vecteurs (101010) et (111010) est égale à 1, car ces deux mots diffèrent seulement en une seule position. La distance de Hamming entre les deux vecteurs (1010) et (0101) est égale à 4, car ces deux mots diffèrent dans toutes les positions. La distance de Hamming entre les deux vecteurs (1010) et (111010) n'est pas définie et ne sera pas calculée, car les deux vecteurs ne sont pas de même taille [8].

2.5.1.2 Distance de Jaccard

La distance de Jaccard est utilisée généralement en statistique pour comparer la similarité des mots. Elle a été établie par le Suisse Paul Jaccard.

Elle consiste à calculer la valeur absolue de l'union des deux vecteurs mots moins leur intersection, puis le résultat divisé par l'union des deux vecteurs.

J (a, b)=|a∪b|-|a∩b| / | (a∪b)|

soient les deux vecteurs a=(1,1,0,1,0) et le vecteur b=(0,0,1,1,0), la mesure de distance calculée avec la méthode de Jaccard serait [55]:

J (a, b)=|a∪b|-|a∩b| / |(a∪b)|=4-1/4=0.75

2.5.1.3 Distance euclidienne

La distance euclidienne représente une distance géométrique dans un espace multidimensionnel. Généralement c'est le type le plus utilisé pour le calcul des distances et elle se calcule de la façon [56] :

Distance d (a, b)= $\{\sum_i (X_i - Y_i)^2\}^{1/2}$

Soient les deux vecteurs a = (1, 1, 0, 1, 0) et b = (0, 0, 1, 1, 0). La distance euclidienne est :

$$d = \sqrt{|1-0|^2 + |1-0|^2 + |0-1|^2 + |1-1|^2 + |0-0|^2} = 1.7321$$

2.5.1.4 Distance du City-block(Manhattan)

Cette distance représente la somme des différences entre les dimensions. Souvent les résultats sont très proches de celui de la distance euclidienne. Cette mesure représente l'addition de la valeur absolue de la soustraction de chaque élément du premier vecteur par ceux du deuxième [57].

La distance du city block se calcule de cette façon :

Distance d (a, b)=$\sum_i (X_i - Y_i)^2$

Soient les deux vecteurs a = (1, 1, 0, 1, 0) et b = (0, 0, 1, 1, 0). La distance du City block est :

d= $|1-0|^2+|1-0|^2+|0-1|^2+|1-1|^2+|0-0|^2=3$

2.5.1.5 Distance de Tchebychev

Cette distance est représentée sur un espace vectoriel où la distance entre les deux vecteurs est très grande. La distance de Tchebychev est donnée comme suit [38]:

$$d = Max(\sum_{i=1}^{n} |a_i - c_i|)$$

Et l'équivalent en limite est de la forme : $d = \lim_{p \to \infty} \sum_{i=1}^{n} \sqrt[p]{(a_i - c_i)^p}$

Soient les deux vecteurs a=(1,1,0,1,0) et b=(0,0,1,1,0). La distance est :

$$d = Max(|1-0|, |1-0|, |0-1|, |1-1|, |0-0|) = 1$$

2.5.2 Méthode de classification automatique

La classification automatique est utilisée comme une méthode statistique dans les recherches des données. Cette méthode permet d'acquérir des informations sans avoir des connaissances préalables, en effet c'est une méthode de type non supervisé. Elle simplifie l'analyse des données et réduit la quantité d'information avant de l'envoyer à une autre méthode.

Généralement la classification automatique est utilisée pour affecter des mécanismes et des principes à une machine. Il faut utiliser des méthodes qui s'articulent sur des données automatiques et sans connaissance préalable.

La classification automatique est utilisée dans plusieurs domaines : En économie, en biologie, en identification des zones dans une image. Il est possible aussi de l'utiliser en

recherche sur les données. En effet, elle permet de vérifier la répartition des données, de vérifier la différence entre les groupes.

Cette méthode de classification contient deux groupes : les méthodes de partitionnement simple et les méthodes hiérarchiques. Les premières méthodes se basent essentiellement sur le classement des objets en différentes catégories et elles doivent répondre aux critères :

- Il faut avoir au minimum un objet dans chaque groupe ;
- un objet doit être dans un seul groupe ;

Ces critères sont utilisés pour vérifier et pour valider les classes obtenues.

Les autres méthodes de classification permettent de traiter les données d'une façon hiérarchique. En effet, ces méthodes peuvent être divisées en deux catégories. La catégorie ascendante qui utilise chaque objet pour former une classe et qui fusionne par la suite les deux classes les plus similaires. L'autre catégorie est descendante. Elle commence son traitement en mettant tous les objets dans une seule classe. Chaque classe sera divisée en classe plus petite jusqu'à obtenir un seul objet dans chaque classe.

Le problème de la classification automatique apparait après chaque itération erronée, car on ne peut pas l'annuler ou la corriger. Aussi, cette méthode est connue pour la complexité de son algorithme [21].

2.5.2.1 Méthodes non hiérarchiques

Parmi les méthodes non hiérarchiques, on peut citer [25] :

2.5.2.1.1 Méthode de leader

Cette méthode affecte au premier objet la première classe et devient son leader. Lorsqu'un nouvel objet apparait, il faut calculer sa distance par rapport au leader de chacune des classes existantes. Puis, on fait la comparaison de cette distance par rapport à un seuil qui sera défini par l'utilisateur. Si cette distance est inférieure au seuil, alors le nouvel objet sera ajouté à cette classe. Sinon cet objet sera le leader d'une nouvelle classe créé.

2.5.2.1.2 Méthode de K-means

En 1967, cette méthode a été élaborée par MacQueen. Elle consiste à choisir un point «K» comme nombre de classe attribué à un objet, puis on calcule la différence entre le nombre des classes «K» et les objets. Cette méthode constitue un mécanisme permettant de partager un ensemble K en différentes classes homogènes. Elle est basée sur le principe de la méthode des plus proches voisins. Cette méthode est la plus utilisée, car elle est très utile et très rapide, mais elle utilise toutes les ressources de la machine et souvent elle est inefficace pour de grands fichiers. Un mauvais choix de «K» donne également de mauvais résultats [15].

Algorithme

Dans la méthode K-means, la sélection du K est aléatoire et il s'effectue de la façon suivante :

- On choisit un nombre aléatoire de K.
- On doit regrouper les observations en classes différentes selon un choix bien défini.

- On doit calculer les centres de gravité des K classes obtenus. Alors on obtient de nouveaux centres. Et à partir de ces centres, on recherche des nouvelles classes.
- On calcule les K nouveaux classes en calculant les centres de gravité des classes précédents.

L'arrêt se fait lorsque deux itérations consécutives donnent le même résultat ou lorsqu'on détermine une condition du notre choix [25].

2.5.2.1.3 Méthode de nuées dynamiques

Cette méthode est créée par Diday en 1972. Elle consiste à faire un tirage au hasard de K noyaux parmi une famille de noyaux. Après il faut chercher les points les plus analogues avec un noyau et les lui affecter. Donc on obtient une partition en K classes. Cette procédure se répète avec tous les objets jusqu'à obtenir une bonne qualité de classification. Ce mécanisme de comparaison est réalisé par le calcul des distances.

Cette méthode donne un temps de traitement très rapide et offre à l'utilisateur le droit de choisir le nombre de classes avant le traitement [25].

Algorithme

L'algorithme de la méthode de nuée dynamique est basé sur un traitement itératif. Il se base essentiellement à la recherche d'une bonne partition d'un ensemble donnée en sous-classes.

- Soit un ensemble E de n points.
- Il faut chercher à partitionner l'ensemble E en sous-classes K tout en gardant les éléments les plus proches dans la même classe.

- Pour avoir une bonne classification dans chaque classe, il faut avoir une distance minimale entre le noyau et chaque point.

2.5.2.2 Méthode hiérarchique

Cette méthode permet de regrouper des objets dans des classes en les ajustant selon une distance bien définie. On affecte un objet à une classe s'il est le plus proche de cette classe par rapport aux autres objets. Cette méthode ne permet pas de donner un critère pour définir la distance entre les classes [25].

Algorithme

- Initialiser les classes de départ et la matrice de distance entre les objets.
- Choisir un critère pour regrouper les éléments.

Mise à jour au niveau de la matrice de distance en calculant de nouveau les distances entre les éléments regroupés ensemble.

2.5.3 Méthodes d'affectation

2.5.3.1 Méthode des K plus proches voisins KNN

La méthode KNN représente une méthode simple et efficace. Elle consiste à trouver les K plus proches voisins les plus similaires dans l'ensemble de l'échantillon. C'est une méthode à base de voisinage.

Pour le calcul de la similarité avec un nouveau cas, l'algorithme KNN utilise tous les attributs des cas déjà calculés. Pour trouver deux mots parmi 2000 mots, c'est presque impossible, et ça cause aussi la perte du temps dans l'algorithme [25].

Algorithme

Pour classer un objet x il faut :

- Chercher l'ensemble des K plus proches voisins de x suivant les méthodes euclidiennes, Manhattan, etc. Il faut chercher les points les proches dans l'objet x.
- choisir la classe de x selon l'ensemble trouvé.

2.5.3.2 Les réseaux de neurones

Les réseaux de neurones sont à la base d'une modélisation du cerveau humain qui a commencé depuis les années 1940 par McCulloch et Pitts. Cette méthode consiste à utiliser des objets appelés neurones qui permettent d'effectuer des calculs sur des données.

La sortie des neurones peut être l'entrée d'un ou plusieurs autres neurones [25].

Les réseaux de neurones sont présentés par cet algorithme [59] :

- On doit initialiser tous les entrées et les poids des réseaux;
- pour effectuer le calcul, on doit créer une condition d'arrêt;
- on calcule la différence entre la sortie attendue est le résultat;
- on utilise la méthode du gradient pour minimiser l'erreur entre le résultat de sortie et la valeur d'entrée.

Les réseaux de neurones sont composés de différentes méthodes :

- ART (adaptative Résonnance théorie) : Le modèle ART est créé par Stephen Grossberg et Gail Carpenter. Ce modèle est basé à la classification non

supervisée et à l'unicité. Il est basé en deux sous modèles qui sont l'ART-1 qui est connu de son classification binaire et ART-2 en représentation analogique. Il se base sur une analyse descendante du problème du niveau abstrait vers la couche d'entrées.

ART permet aussi de reconnaitre les formes et l'anticipation.

Algorithme

L'algorithme ART prend en entrée un vecteur, puis il le compare aux autres nœuds dans sa mémoire. Un calcul sera mis en place pour reconnaitre la similitude avec les autres nœuds. Lorsqu'il y a une ressemblance avec le nœud d'entrée alors il y aura une mise à jour entre les nœuds. Sinon un nœud mémoire sera créé [60].

- SOM (Self Organizing Maps): SOM est un algorithme d'apprentissage non supervisé. Il s'agit d'un réseau de neurones artificiels utilisé pour diminuer les espaces de grandes dimensions. Il est capable de représenter des données de grandes tailles. Cet algorithme d'apprentissage représente une méthode très populaire pour la diminution des dimensions [7].

Algorithme

- Initialiser les vecteurs de neurones d'une manière aléatoire ou selon la méthode de k-moyenne. Généralement, l'initialisation aléatoire est la plus utilisée.
- Pour chaque neurone, il faut chercher les autres les plus similaires.
- Mettre à jour tous les neurones similaires.
- Il faut répéter ce traitement tant qu'on n'a pas atteint une condition d'arrêt.

2.6 Les règles d'association

2.6.1 Présentation

Le concept de règles d'association a été présenté dans un article de Rakesh Agrawal en 1993 [43]. Elle représente un outil des plus utilisées en datamining et qui permet d'extraire des nouvelles connaissances le plus utilisé en Data «mining». À partir d'un ensemble de données, elle permet d'analyser les relations entre les variables ou détecter des associations. Ces associations sont interprétées sous la forme d'implications par exemple « X –> Y » exprimant le fait que les attributs dans X tendent à apparaitre avec ceux dans Y. Dans ce cas, X est appelé l'antécédent de la règle et Y son conséquent.

Par exemple, on parle des achats d'un client dans un magasin : «70% des clients qui achètent du lait et du thé achètent aussi du pain», alors une règle d'association regroupe les attributs lait et thé à l'attribut pain. Ceci permet de bien organiser des produits dans un magasin. Ça veut dire mettre le lait, le thé et le pain l'un à côté de l'autre pour faciliter le processus d'achat à un client. Ceci favorise les bénéfices du magasin puisqu'il va augmenter, de ce fait, le nombre de ses ventes.

2.6.1 Définition

Une règle d'association peut être définie selon plusieurs termes [25]:

- Transactions et Items : Soient I = {i1, i2, ...in} l'ensemble de tous les Items et T = {t1, t2, .., tn}, l'ensemble de toutes les transactions. Chaque transaction contient un sous-ensemble d'items. Le volume de la transaction est le nombre d'items qu'il contient. Pour chaque ensemble d'items, on a un support qui fait

référence au nombre de transactions qu'il contien. Le support d'un ensemble d'items (X) est défini par : $\sigma(X) = Card(\{t_i | X \subseteq t_i, t_i \in T\})$

- Où Card(A) représente le cardinal de l'ensemble A.
- Itemset : Ce champ contient le sous-ensemble d'items.
- Itemsets : Ce champ contient un ensemble des sous-ensembles d'items.
- Règle d'association : La règle d'association est l'ensemble des items disjoints et elle est représenté par la forme « item X->item Y ». Item X représente le produit a analyser et item Y représente le produit associé [62].
- Confiance : Le plus commun des règles d'association est appelé support-confiance, où cette règle est mesurée par le support et la confiance. Soit n(X) et n(Y), les nombres de transactions effectuées respectivement sur les items de X et Y. Le support d'une règle est le rapport d'exploitation qui réalise à la fois X et Y. La confiance est mesurée par le rapport que réalise Y, parmi celles que réalise X, c'est-à-dire Y sachant X [33].

Le tableau suivant donne une vision claire sur le support et la confiance [44].

ID	Produit
1	Pain, Crème, eau
2	crème

3	Pain, Crème,huile
4	eau
5	Crème,eau

Tableau 1 : Support et confiance

Support= Problématique (crème et pain).

Support= (nombre des transactions contenant crème et pain) / (nombre total des transactions) = 2/5=0.4

Confiance=Problème (crème et pain/ crème)

Confiance= (nombre des transactions contenant crème et pain) / (nombre total des transactions contenant crème) = 2/4=0.5 = sup (crème et pain)/ (sup crème)

La confiance se déduit du support conf (X− > Y) = sup(X*Y)/sup(X). Donc c'est important de calculer le support d'abord. L'importance du support peut être remarquable lorsqu'une règle a un support faible. Plus la confiance de X− > Y est élevée, plus la probabilité, d'observer Y avec X est forte : De plus la confiance donne une estimation de la probabilité conditionnelle d'Y sachant X [27].

- Le lift : Le lift d'une règle X− > Y permet de mesurer l'amélioration apportée par la règle d'association à un jeu de transactions aléatoire (X et Y sont indépendants). Il est défini par : Lift (X− > Y) = sup(X U Y) / sup(X).sup(Y). En relation avec la probabilité, le lift représente le rapport d'apparition des

items d'Y dans une même transaction avec X, sur la probabilité d'apparition d'Y dans la même transaction de l'ensemble. En effet, le lift représente une bonne mesure de performance des règles d'association.

Le tableau suivant représente la présence d'achats de la pomme et de fraise à la caisse [49].

Pomme	1	1	1	1	1	0	0	0	0	0
fraises	1	1	1	1	0	1	1	1	1	1

Tableau 2 : La présence de la pomme et de la fraise dans les achats

Le calcul de la confiance de la règle « pomme –> fraise » (pomme) = 5/10, (pomme, fraises) = 4/10, Conf = (4/10)/ (5/10) = 4/5 = 80%. Donc, on a 80% des cas lorsqu'il y a un achat de «pomme », il ya l'achat« fraises ».

Le calcul du support : Supp = (pomme) = 5/10 = 50% donc il y a 50% des individus qui sont intéressés par cette règle. La probabilité des fraises vaut : (9/10) ça veut dire 90%. Le lift vaut : 80/90 = 8/9.

2.6.1 Règle d'association positive et négative

Les mesures de règles d'association dans le cadre de support-confiance reçoivent certaines critiques, dont l'une est que seulement les associations positives entre les attributs X et Y peuvent être découverte, alors que les associations négatives ne pourraient pas être trouvées [40]. C'est parce que les exemples positifs sont considérés comme la négation de

tout article ignorée. Toutefois, les associations négatives sont également très importantes dans notre vie quotidienne. Par exemple, dans les applications de supermarché, les règles d'association positives suggèrent de placer des articles associés ensemble. Tandis les règles d'association négatives représentent une règle qui contient une négation. Ils sont souvent utilisés pour séparer les éléments. Son antécédent ou son conséquent peut être formé par une union. Une règle négative se trouve généralement entre deux Itemsets. On appelle règle d'association positive toute règle de cette forme « $X \rightarrow Y$ », alors que les règles écrites sous autres formes sont négative comme « $X \rightarrow \neg Y$, ou $\neg A \rightarrow Y$ » [16].

2.6.2 L'extraction des règles d'association

L'extraction des règles d'association permet de découvrir des relations significatives entre les attributs binaires extraits des bases de données. Il s'agit d'un processus constitué de plusieurs phases allant de la sélection et la préparation des données jusqu'à l'interprétation des résultats, en passant par la phase de recherche des connaissances : le data mining.

La plupart des approches proposées pour l'extraction des Itemsets fréquents reposent sur les quatre phases suivantes [34] :

-Préparation des données : Cette phase consiste à extraire les règles d'association à partir d'une base de données et permet de transformer ces données en un contexte d'extraction. Elle est nécessaire lors de l'application des algorithmes d'extraction des règles d'association sur des données de natures différentes.

-Extraction des ensembles fréquents d'attributs : Cette phase permet d'extraire tous les ensembles d'attributs binaires, appelés itemsets. Un itemset l est fréquent si son support est supérieur ou égal au seuil minimal de support.

– La génération des règles d'association : Au cours de cette phase, les itemsets d'une phase précédente sont utilisés afin de générer les règles d'association qui sont des implications entre deux itemsets fréquents.

– Interprétation des résultats : c'est le rôle de l'utilisateur qui visualise les règles d'association extraites du contexte et les interprète. Les règles d'association extraites en général imposent le développement d'outils de classification des règles, de sélection par l'utilisateur de sous ensembles de règles, et de leur visualisation sous une forme intelligible.

Exemple :

On va représenter un exemple tiré d'une base de données d'un super marché. Il s'agit de prendre un aperçu sur les achats de la céréale et du sucre : « céréale ∧ sucre → lait (support 7%, confiance 50%)». Cette règle nous montre la probabilité de l'achat du lait lors de l'achat du sucre et de la céréale. La valeur du support nous montre le pourcentage des clients qui ont acheté les trois articles. La confiance indique la précision de la règle, c'est-à-dire elle montre le pourcentage des clients qui ont acheté du lait parmi ceux qui ont acheté de la céréale et du lait [51].

2.6.3 L'algorithme Apriori

L'algorithme Apriori (Agrawal et Srikant, 1994) [50] représente le premier algorithme de recherche de règle d'association incluant les étapes d'élagage qui permet de tenir en

compte la croissance des items. L'algorithme commence par déterminer le support de chaque item. Puis, il génère les ensembles d'items de taille K à partir des ensembles précédents de taille (k-1). Il permet de vérifier si l'ensemble d'items est fréquent, alors tous ses sous-ensembles sont aussi fréquents.

Exemple :

L'algorithme se base sur l'union des ensembles qui ont un seul élément différent [26].

Mots\ ensemble	1	2	3	4	5
A	x				
B	x	x		x	x
C			x	x	
D	x				x

Tableau 3: Exemple de fonctionnement de l'algorithme Apriori

Donc les ensembles ({A}, {B}, {C}, {D}) donnent respectivement le support (1, 4, 2, 2).

Donc l'ensemble on aura l'ensemble E= {{B}, {C}, {D}} et

E1= {{B, C}, {B, D}, {C, D}}.

Algorithme

- Définition d'un ensemble des éléments E de cardinal k.

- Soit C= {c=E1 U E2 / (E1, E2) ∈ E*E, card(c)=k+1}
- Si c ∈ C alors : on prend une variable s ⊂ c dont le card(s)=k. Et ont teste si s ∉ E alors
- C=C\ {c}
- La valeur de C représente le résultat final.

2.6.3.1 Avantages

L'étude des règles d'association nous donne la chance de découvrir ses avantages et on peut citer par exemple :

- Un formalisme très général;
- une méthode simple et facile à comprendre;
- des résultats clairs et faciles à interpréter;
- son application dans plusieurs domaines.

2.6.3.2 Inconvénients

D'après la pratique et les résultats obtenus depuis les algorithmes des règles d'association, on rencontre beaucoup d'inconvénients. On peut citer par exemple :

- Très couteux dans le calcul, il est nécessaire d'utiliser des algorithmes plus performants pour les grands volumes de donnée;
- ils génèrent plusieurs règles;

2.6.4 Problème de la pertinence et de l'utilité des règles d'association

Le problème de la pertinence et de l'utilité est associé aux règles d'association (extraite et) redondante. Ce sont des règles qui ont la même information. Si on a remarqué un très grand nombre de règles, alors généralement on utilise un système de suppression des règles d'association redondantes. De plus les règles d'association redondantes représentent la majorité des règles extraites, leurs suppressions permettent de faciliter la visualisation par l'utilisateur [35].

Exemple : On donne neuf règles d'association extraites du jeu de donnée Mushrooms qui décrit les caractéristiques de 8416 champignons. Ils possèdent un support de confiance presque identique de 51% et 54% respectivement.

1) Lamelles libres ⟶ Comestible.

2) Lamelles libres ⟶ Comestible, voile partiel.

3) Lamelles libres ⟶ Comestibles, voile blanc.

4) Lamelles libres ⟶ Comestibles, voile partiel, voile blanc.

5) Lamelles libres, voile partiel ⟶ Comestible.

6) Lamelles libres, voile partiel ⟶ Comestible, voile blanc.

7) Lamelles libres, voile blanc ⟶ Comestible.

8) Lamelles libres, voile blanc ⟶ Comestible, voile partiel.

9) Lamelles libres, voile partiel, voile blanc ⟶ Comestible.

On remarque que les règles 1 à 3 et 5 à 9 sont redondantes par rapport à la 4ème parce que ces 8 règles n'ont pas d'informations supplémentaires par rapport aux règles 4. Donc, il

est conseillé d'utiliser seulement la règle 4 pour l'utilisateur. Donc, on distingue deux types de règles :

- Des règles d'association exactes qui sont vérifiées dans tous les objets et qui ont une confiance égale à 1;
- des règles d'association approximatives qui sont vérifiées dans une proportion d'objets et qui ont une confiance inférieure à 1.

Dans ce cas, le but est d'essayer d'extraire les règles d'association les plus informatives sans redondances avec une grande précision.

2.6.5 Méthodes de réduction

Il existe plusieurs méthodes permettant de réduire le nombre de règles d'association extraites [51].

- Règles d'association généralisées : Ce sont des règles d'association qui peuvent être classées selon différents niveaux. Par exemple, la règle « r1 : lait→ Sucre » est appelée sur-règles des deux règles «r2 : lait entier → sucre » et « r3 : lait écrémé → sucre », car les items « lait écrémé » et « lait entier » sont hérités de l'item « lait » cette méthode permet de classifier les items et non pas de supprimer les règles d'association redondantes.
- Mesures statistiques : Les mesures statistiques permettent de déterminer la précision des règles d'association selon différentes mesures. Parmi ces mesures, on cite les deux mesures qui donnent les résultats les plus intéressants. La mesure de conviction permet de calculer pour chaque règle le décalage entre la probabilité d'occurrence de l'antécédent et la probabilité de non-occurrence

dans l'objet. La mesure de X^2 présente le degré de dépendance entre les items d'un itemset en comparant la répartition réelle de leur occurrence avec celle attendue. Ces mesures dépendent d'un temps de calcul assez important, ce qui peut causer des problèmes.

- Mesure de déviation : Ce sont des mesures de distance entre les règles d'association définies en fonction de leurs supports et leurs confiances. Ces mesures peuvent être utilisées afin de déterminer les règles d'association semblables, avec une faible distance entre elles. De cette méthode résulte une perte d'informations qui nécessite la comparaison des règles deux à deux. Mais elle permet de déterminer de nouvelles règles d'association qui sont inattendues de la part de l'utilisateur et fournissant de nouvelles informations. En général, la déviation d'une règle, c'est la différence entre une valeur attendue de la règle dans le modèle probabiliste et la valeur réelle pour la règle. Ce qui dans de nombreux cas s'avère très complexe et leurs temps de calcul sont très importants.

- Template : Elle permet de sélectionner un sous-ensemble de l'ensemble des règles d'association valides. Le sous-ensemble est construit en conservant les règles d'association qui vérifient le critère spécifié par les "Template". Cette méthode de traitement ne permet pas de supprimer les règles redondantes, mais elle facilite la visualisation par groupe où chaque groupe correspond à un ensemble de «templates».

- Contraintes sur les items : Ce sont des expressions portant sur les conséquences des précédentes règles définies par l'utilisateur. Ces contraintes sont utilisées

lors de la phase d'extraction des itemsets fréquents. Le candidat peut générer ses propres règles satisfaisant aux contraintes après la génération des itemset candidats. Cette méthode n'élimine pas les règles redondantes et donne seulement un résultat partiel.

- Base pour les règles d'implication : Cette approche permet d'extraire des règles d'association exactes et approximatives. Les résultats obtenus sont des réductions de l'ensemble des règles d'association qui minimisent le nombre des règles générées. Cela exprime que les antécédents et les conséquences de toutes les règles d'association peuvent être obtenus de l'union de ces bases.

- Base pour les règles d'association : Une règle d'association est la conséquence des deux itemsets à laquelle est associée la mesure de support, qui définit la règle, la mesure de confiance et la précision de règle dans le contexte d'extraction. Le support de confiance donne l'utilité de la règle et doit donc être pris en considération lors de la définition des règles d'association redondantes. Une règle d'association r : $l_1 \rightarrow l_2$ de support S et de confiance C est noté

  r : $l_1 \rightarrow^{(s,c)} l_2$ une règle d'association r $\in$ E est redondante si la règle r peut être déduite ainsi que son support S et sa confiance C.

2.6.6 Règles d'association maximale

Les règles d'association régulières permettent aux analystes et aux chercheurs de mettre en évidence des structures cachées dans des fichiers de données volumineux (base de données), par exemple "les clients qui ont acheté le produit A ont également acheté en grande majorité le produit B ou le produit C ". Suite à l'utilisation de l'algorithme, dit a

priori, vous pouvez rechercher et trouver ces associations rapidement dans de gros volumes de données en utilisant des règles de détection prédéfinies.

Mais dans les cas où on a des associations moins évidentes à l'intérieur des corpus, l'utilisation de ces algorithmes ne permet pas de trouver certaines d'entre elles. Par exemple, dans un ensemble de produits, il y a des mots qui sont étroitement liés alors ils apparaissent fréquemment ensemble. Le mot produit laitier a une relation avec yaourt et yaourt a une relation avec fruit. Dans ce cas les règles d'association permettant d'identifier ces articles. Mais dans un autre cas, le produit laitier et le fruit ont une corrélation très basse. Donc, on ne peut pas trouver des associations dans de pareil cas.

Les règles d'associations maximales fournissent une solution efficace pour confronter les problèmes définis auparavant. Elles permettent de résoudre ces cas et d'extraire les relations les moins intéressantes qui sont négligées par les règles d'associations ordinaires.

Les règles d'associations maximales ont pour but de dégager des relations intéressantes dans un objet et en voici les principes de bases [17] :

- Pour une transaction t et un ensemble d'éléments X du même groupe, on dit que t supporte X si $X \subset t$ et on écrit $S_D(X)$, qui représente le nombre des transactions $t \in D$ qui supportent X.
- Le support de $X \rightarrow Y$, où X et Y sont deux sous-ensembles d'éléments disjoints, est le support de $X \cup Y$.
- Dans une règle d'association maximale $X \rightarrow Y$, alors chaque fois que X apparaît seul, Y apparaît aussi, mais avec une certaine confiance.

- Le M-support : Le M-support de l'association maximale X →Y est : S_D(X → Y) avec t : t M–supporte X et t supporte Y, et S_D(X → Y) représente le nombre d'itération dans D qui M-supporte X et supporte Y.

Exemple

Soit les ensembles : E1= {A, B, 1}, E2= {K, C, 2}, E3= {A, 1, 2} qui groupe des chiffres et des lettres.

Chiffre= {1,2}, Lettres= {A, B, C, K}.

Soit la règle d'association A→1. Donc le M support de cette règle est défini par le nombre de transactions qui M-supporte A et 1 en même temps :

(E3 ∩ Lettre = A), car c'est la seule lettre dans l'ensemble.

(E1 ∩ Lettre = A, B).

Dans l'ensemble E2 on n'a pas la lettre A.

(E1 ∩ chiffre= 1), car c'est la seule lettre dans l'ensemble.

(E2 ∩ chiffre = 2).

(E3 ∩ chiffre = 1, 2).

Donc dans cet exemple on a juste E3 qui répond à la règle A→1.

2.7 Conclusion

Dans ce chapitre, nous avons présenté les fonctionnalités utilisés pour l'analyse du texte comme par exemple l'utilisation de commentaires, le surlignage ou bien l'ajout des signets dans les documents. Des travaux récents ont également révélé l'utilisation de dictionnaires permettant l'affichage de définitions, de synonymes et d'antonymes.

Mais malgré les différentes méthodes utilisées et les applications développées, il n'en existe pas une qui répond à tous les besoins de l'utilisateur. En effet, plusieurs problèmes et difficultés sont rencontrés avec les approches utilisées.

- Le temps d'affichage des documents augmente avec la taille des documents.
- La difficulté d'afficher les caractères spéciaux et les documents en différentes langues.
- La difficulté d'extraire les mots clés du texte et de rédiger des remarques et l'absence d'interface qui affiche les données enregistrées par l'utilisateur.
- La difficulté de l'analyse du texte. En effet il n'y a pas des fonctions qui simplifient la tâche d'analyse.

Dans le chapitre suivant nous allons présenter différentes solutions pour faciliter le traitement aux utilisateurs. Nous allons introduire des méthodes qui permettent d'enregistrer des phrases et des mots clés du le texte. Aussi nous allons développer une méthode qui fait compare les phrases enregistrées et supprime ceux qui se répètent pour facilité la compréhension du texte par l'utilisateur.

Nous allons décrire des autres fonctions qui sont basés sur l'extraction des données et l'affichage du nombre des mots qui se répètent dans texte.

Chapitre 3 - Description du fonctionnement de notre application

3.1 Introduction

Dans notre travail, l'analyseur de texte permet d'assister l'utilisateur à la lecture d'un texte. Il donne à l'utilisateur la possibilité de rédiger des notes et des commentaires. Il permet également les enregistrements dans une base de données. Son avantage est de mieux gérer le texte. En fait, faire une analyse complète et équilibrée est souvent une tâche assez compliquée. D'un autre côté, l'utilisateur peut modifier les notes enregistrées, les extraire dans un autre document texte ou dans une autre interface de l'application, rédiger des idées et de les imprimer. Les notes enregistrées peuvent être analysées grâce à une fonction qui permet de supprimer les phrases semblables pour éviter toute ambiguïté. Cet analyseur permet de calculer les occurrences des mots dans le texte et donner pour chaque mot son synonyme et son antonyme. Il améliore également, la recherche dans un texte en un temps de réponse assez réduit ainsi qu'une meilleure mise en forme pour représenter les résultats trouvés.

3.1 Principe

L'utilisateur assure la manipulation et le fonctionnement de cette application. Il peut utiliser les différentes options pour simplifier la lecture et l'analyse du texte.

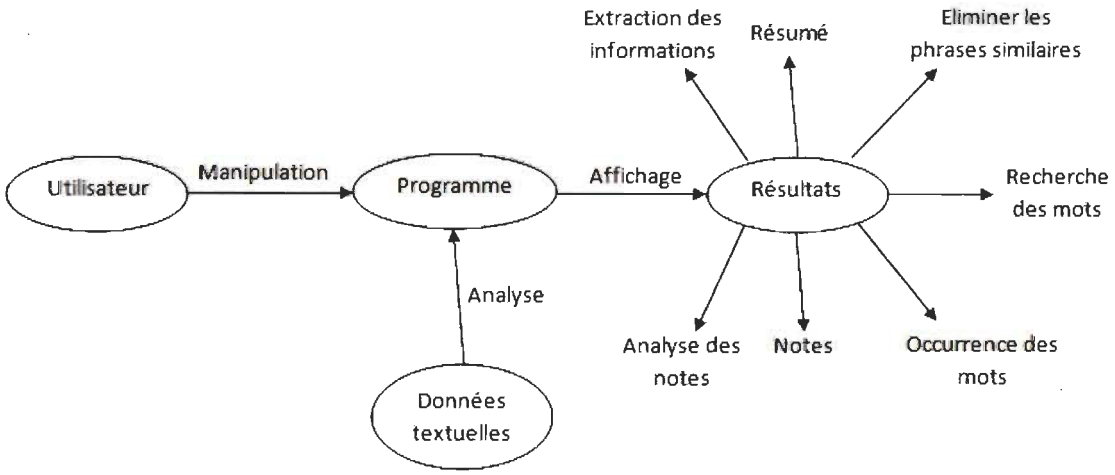

Figure 4 : Principe de fonctionnement

3.2 Création des interfaces graphiques

Une interface graphique est un lien qui facilite le dialogue homme-machine [1]. Il s'agit de pictogrammes (représentations graphiques qui ont une fonction de symbole), à l'écran qui permet à l'usager d'exercer une manipulation physique de ces objets avec un dispositif de pointage. L'utilisation de ces interfaces facilite la manipulation par l'utilisateur. Elle est adaptée à ses besoins. L'utilisateur peut à tout moment sélectionner un objet, changer de fenêtre, changer la forme d'affichage, bref, gérer son application.

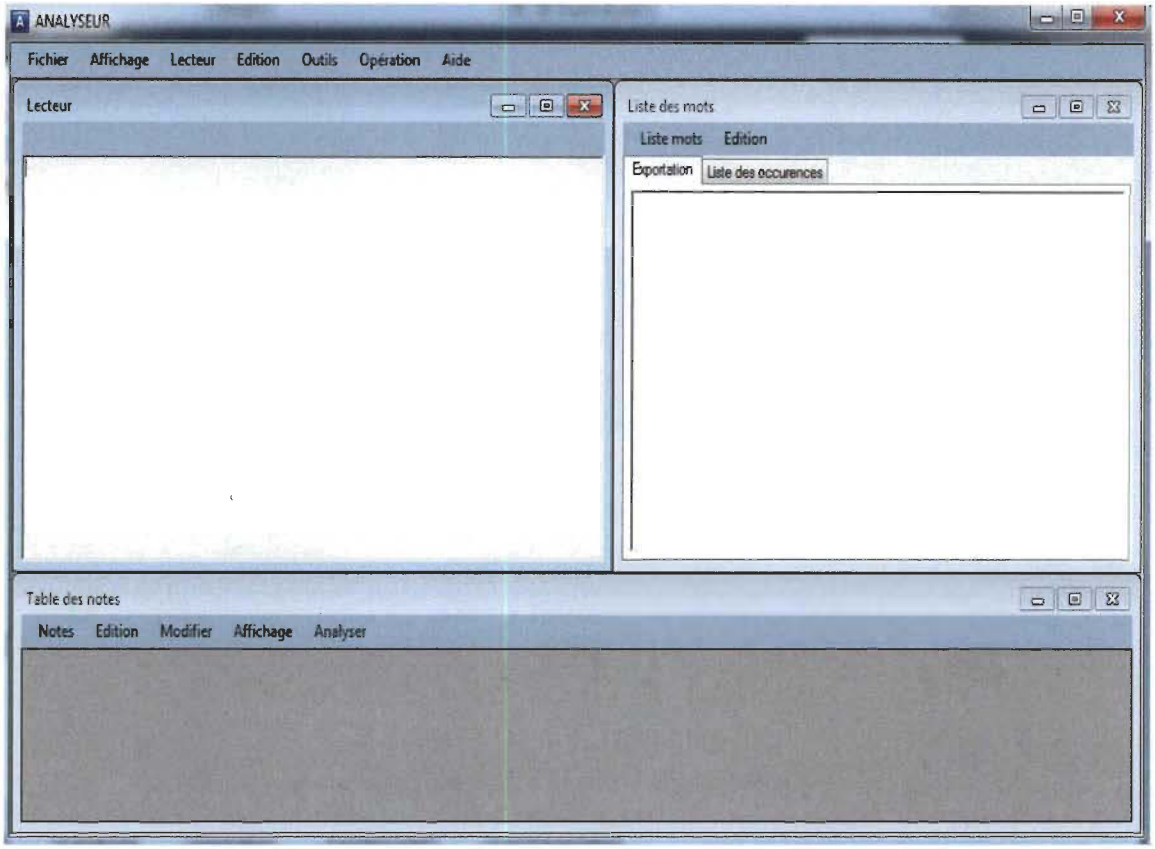

Figure 5 : Interface de l'application

3.3 Fonctionnalité attendu

Parmi les fonctionnalités de l'application, il y a celles qui sont des standards (copier/coller, affichage, lecture, etc.). Ces fonctionnalités sont présentes dans la majorité des applications. Elles permettent de simplifier différentes tâches que nous énumérons ici.

3.3.1 L'affichage du texte

Cette fonction offre à l'utilisateur la possibilité d'ouvrir et d'afficher le texte d'un fichier quelque soit son type (doc, docx, txt, etc.). Elle permet d'éviter les problèmes liés aux caractères spéciaux. Ceci permet d'éviter l'affichage des symboles indésirables qui sont

généralement causés par l'incompatibilité des normes comme les caractères ASCII 7, 8 bits ou Unicode. Il est à noter que l'affichage inclut les lettres grecques et latines.

3.3.2 Mise en forme

L'interface du lecteur de texte donne à l'utilisateur l'option de changer la police, la taille et l'alignement du texte. Ces options permettent d'améliorer l'affichage du texte et facilitent sa compréhension. Cette façon permet d'afficher le contenu textuel selon le besoin de l'utilisateur. En effet la mise en forme donne une ergonomie à l'affichage du texte par une représentation à plusieurs niveaux (les marges, l'alignement, la police, …).

3.3.3 Enregistrement

L'utilisateur a la possibilité d'enregistrer le document sous un autre nom de son choix et dans le répertoire qu'il aura sélectionné.

3.3.4 La recherche

La recherche textuelle permet de sélectionner des mots dans le texte selon une requête de l'utilisateur. Ce type de recherche contient des mots clés ou des phrases. La recherche la plus simple représente une requête qui produit une suite de mots dont on va rechercher les mots similaires et identiques dans le texte [8].

Dans cette méthode, on utilise une fonction qui reçoit une requête formulée par l'utilisateur. Cette recherche commence au début du texte et elle compare les mots avec la requête. Lorsque la recherche est terminée, tous les mots identiques changent de couleur d'arrière-plan afin de faciliter leur identification [43].

Algorithme

Il s'agit de faire une comparaison entre le mot ou la phrase entrée par l'utilisateur et le contenu du texte. Donc il faut :

- Initialiser une variable qui va contenir le mot entré par l'utilisateur.
- Lorsque l'utilisateur lance la recherche, il faut commencer par le début et faire un parcours complet du texte.

Si on trouve un mot identique alors on change la couleur de son arrière-plan et on contenu jusqu'à la fin du texte.

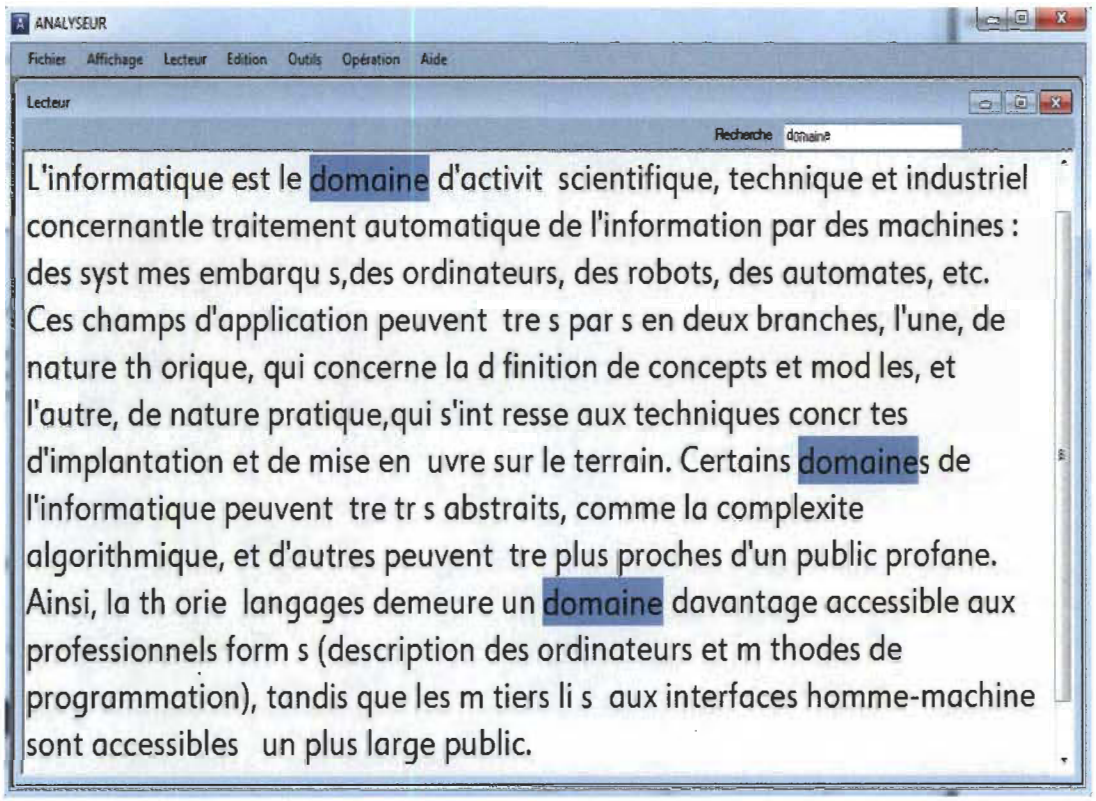

Figure 6 : La recherche d'un mot

Lors de la recherche d'un mot ou d'une séquence de mots, toutes les occurrences dans le texte changent de couleur d'arrière-plan.

3.3.5 La sélection

Pour modifier une zone de texte ou tout le texte, il suffit de sélectionner le paragraphe souhaité puis appliquer les différentes mises en forme (police, alignement, couleur, etc.). À la sélection d'un texte, la zone sélectionnée change de couleur d'arrière-plan. La sélection d'un mot permet de colorer tous les mots qui lui sont similaires [64].

3.3.1 Sommaire

Pour simplifier la manipulation de l'application, l'utilisateur peut accéder au sommaire sous le menu « Aide» qui affiche une structure d'aide concernant l'application. Le sommaire représente le manuel d'utilisation de l'application. Il donne la façon de manipuler toutes les fonctionnalités de l'application avec des images significatives.

3.3.2 Impression

L'impression permet d'extraire un document numérique sous format papier. Cette façon assure que le document imprimé soit conforme au document numérique. La version imprimée du document doit répondre à toutes les attentes de l'utilisateur afin qu'elle donne des informations exactes, les mêmes mises en forme que le document initial. Avant de faire l'impression, on peut passer par la fonction « Aperçu avant impression » qui donne un aperçu sur le document [12].

3.4 Fonctionnalités non standards

3.4.1 Nombre d'occurrences des mots

Pour avoir une idée générale sur le texte, sa décomposition, le nombre et la nature des mots qui existent, on a utilisé une fonction pour simplifier cette tâche à l'utilisateur [57].

Cette fonction permet de donner le nombre d'occurrences de chaque mot dans le texte ce qui aide à connaitre les mots les plus fréquents [65].

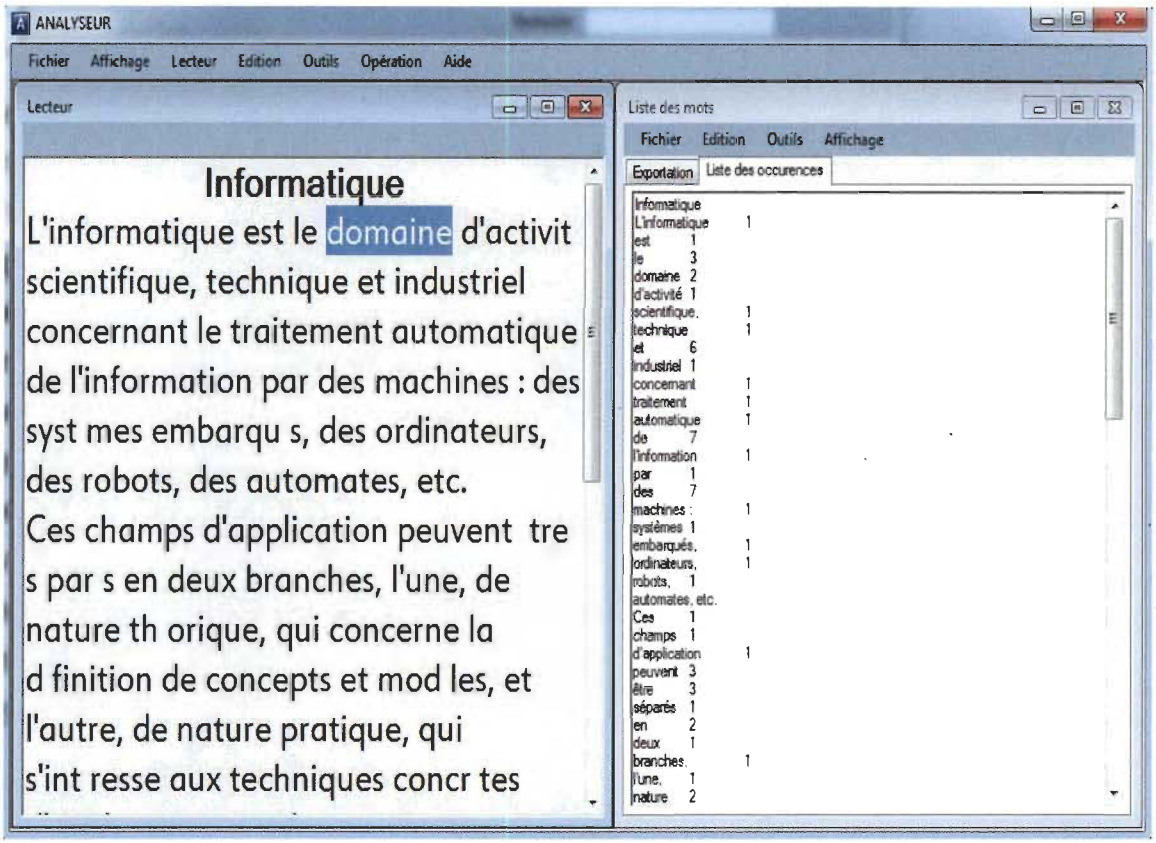

Figure 7 : Nombre d'occurrences

3.4.2 Ajouter des notes

L'ajout d'une note est une fonctionnalité qui sert à rédiger des commentaires ou des remarques qui ne figurent pas dans le texte. Il consiste à enregistrer des phrases pour faciliter la compréhension du texte. Chaque note enregistrée à partir du texte ou rédigée par l'utilisateur est ajoutée automatiquement dans une base de données et affichée dans une autre interface qui sera utile à la fin de la lecture du document.

3.4.1 Tester la similarité des paragraphes

Cette fonctionnalité consiste à comparer les notes enregistrées dans la base de données. En effet c'est une technique qui permet de présenter un rapport de similarité ou de différence entre deux phrases. Dans notre travail ce traitement permet de supprimer les phrases qui se répètent afin de faciliter l'analyse du texte. Cette technique de comparaison est basée sur le traitement par n-gramme.

Exemple de traitement : On va prendre comme exemple cette phrase : «Traitement des langues naturelles.» et on va prendre n=3 (n représente le nombre des caractères à prendre dans l'exemple). Et on va représenter les espaces par : «_»

[Tra=1, rai=1, ait=1, ite=1, tem=1, eme=1, men=1, ent=1, nt_=1, t_d=1, _de=1, des=1, es_=1, s_l=1, _la=1, lan=1, ang=1, ngu=1, gue=1, ues=1, es_=1, s_n=1, _na=1, nat=1, atu=1, tur=1, ure=1, rel=1, ell=1, lle=1, les=1].

En effet le traitement par n-grammes consiste d'abord à choisir le nombre « n », qui représente le nombre des éléments qu'on doit fixer avant de commencer le traitement, puis on passe vers la tâche d'analyse en prenant en compte les espaces. Il faut répéter ce traitement jusqu'au dernier caractère. En comptant les fréquences des n-grammes trouvés [60].

Pour la comparaison de deux mots ou deux paragraphes, cette méthode se base sur un seuil choisi par l'utilisateur et qui représente le taux de différence entre les phrases testées.

La technique par n-gramme facilite la capture des racines des mots les plus fréquents. En effet, la nature de la langue du texte n'a pas d'influence sur le résultat obtenu par ce traitement. Aussi avec cette méthode, il n'est pas obligatoire de segmenter le texte en mot.

Lorsqu'il y a des fautes d'orthographe ou bien de ponctuation, il n'y aura pas d'effet sur le résultat obtenu [66].

Parmi les méthodes utilisées, on peut citer [60]:

- Distance euclidienne : $dist(S_i, S_j) = \|\vec{S}_i - \vec{S}_j\|_2 = \sqrt{\sum_{k=1}^{n}(s_{ik} - s_{jk})^2}$
- Distance de Manhattan : $dist(S_i, S_j) = \|\vec{S}_i - \vec{S}_j\|_1 = \sum_{k=1}^{n}|s_{ik} - s_{jk}|$
- Distance de Minkowski : $dist(S_i, S_j) = \|\vec{S}_i - \vec{S}_j\|_p = \sqrt[p]{\sum_{k=1}^{n}|s_{ik} - s_{jk}|^p}$

Exemple :

On va calculer le taux de similarité des trois ensembles de phrases E1, E2, E3. On considère que le nombre n=2-grams est [61].

E1: [I am], [am sam]

E2: [Sam I], [I am]

E3: [I do], [do not], [not like], [like green], [green eggs]

Distance euclidienne (E1, E2) = 1/3

Distance euclidienne (E1, E3) = 0

Distance euclidienne (E2, E2) =0

3.4.2 Gestion des notes

- Rédiger des annotations : Dans la base de données, l'utilisateur peut modifier ou écrire des notes et des remarques. Le principe de cette tâche est d'extraire les idées clés. En effet il s'agit d'un commentaire ou une explication d'un extrait du texte [58].

- Affichage des notes : Il simplifie la manipulation de l'application et facilite la gestion des notes selon la date, selon l'utilisateur et selon le contenu aussi. Un utilisateur peut afficher juste les notes dont il aura besoin et il peut les modifier. L'utilisateur peut masquer les notes qu'il n'utilise pas. Il peut également les supprimer. La gestion de l'affichage peut être faite aussi selon le nom de la colonne d'affichage ou la ligne, l'utilisateur peut afficher seulement les informations nécessaires dans son interface selon son besoin [20].
- Édition : L'édition des données comporte plusieurs actions [22] :
    - La suppression des données : un utilisateur a le droit de supprimer des notes et des commentaires qui ne lui semblent pas utiles.
    - Annuler : Cette fonction permet d'annuler une action de modification ou de suppression déjà réalisée.
- Impression : L'utilisateur a le droit d'imprimer les notes enregistrées. Il peut ainsi les analyser sur papier. Cette fonction peut être facile pour une catégorie d'utilisateurs qui ne passent pas beaucoup de temps devant un ordinateur.
- Extraction des notes : L'extraction des données peut être faite vers un fichier texte ou vers l'interface de modification dans l'application. Cette technique est très importante et rapide pour des textes de très grand volume qui seront difficiles à les traiter à la main [42].

3.5 Conclusion

Dans ce chapitre nous avons présenté les interfaces de notre application. Elles peuvent être résumées comme suit :

La première interface permet la lecture du texte et donne la possibilité d'enregistrer des notes et des mots clés du texte.

La table des notes permet de parcourir toutes les notes enregistrées puis les classifier et tester les similarités selon des règles pour faire des résumés. La troisième interface contient la liste des mots dans le texte et comporte les phrases exportées de la table des notes et donne la possibilité de les modifier. Aussi, nous avons expliqué le fonctionnement de chaque interface ainsi que les algorithmes sur lesquels sont basées ces fonctionnalités.

Dans le chapitre suivant, nous allons faire tester notre application par plusieurs types d'utilisateurs. Nous allons par la suite présenter les résultats obtenus et revues des évaluateurs.

Chapitre 4 - Expérimentation

4.1 Introduction

Dans ce chapitre, nous allons décrire et évaluer le fonctionnement de l'application. Ce que nous présentons dans un premier temps est une utilisation partielle de l'application. En effet, l'utilisation de cette application peut varier d'un utilisateur à l'autre. En effet les utilisateurs peuvent se servir de ces fonctionnalités :

- La lecture du texte
- La recherche des mots
- L'utilisation des mises en forme et l'impression
- L'extraction des données textuelles

Aussi cette application facilite la lecture du texte tout en réduisant le temps d'accès à l'information. Elle permet en outre le stockage de l'information que l'utilisateur veut conserver pour des applications ultérieures.

Pour faire des tests et effectuer des expérimentations sur l'analyse du texte sur l'application, on a utilisé plusieurs textes, par exemple, le texte de la définition du mot informatique dans Wikipédia, car c'est un texte riche et contient plusieurs mots qui se répètent [1].

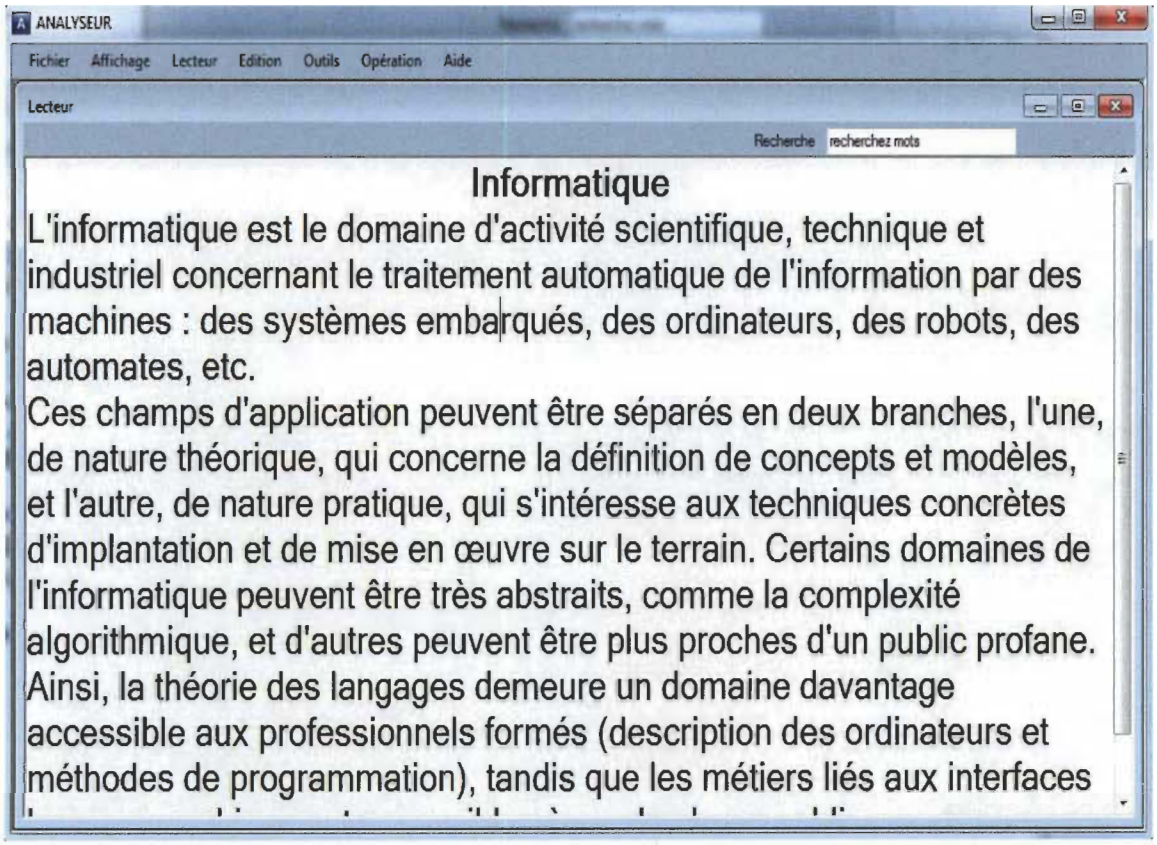

Figure 8 : Page Principale

Calculer le nombre d'occurrence des mots

Avant de passer à une lecture approfondie, le lecteur peut effectuer une analyse générale de son texte. En effet il peut effectuer la tâche de calcul du nombre d'occurrences des mots et les afficher [67].

Cette tache permet de transformer les textes sous la forme de vecteurs numériques pour faciliter l'utilisation de l'apprentissage numérique. La présence de certains mots en attribuant leurs nombres d'apparitions est le seul moyen d'avoir des informations sur le texte. Cette représentation s'appelle « les sacs des mots ». Elle consiste à regrouper les mots en éliminant les signes de ponctuation comme les points et les virgules qui séparent les mots dans plusieurs langues. Cette méthode permet de garder les majuscules pour que l'utilisateur puisse distinguer les noms propres des autres mots. La représentation des occurrences des mots n'introduit pas des analyses ou les calculs de distances entre les mots.

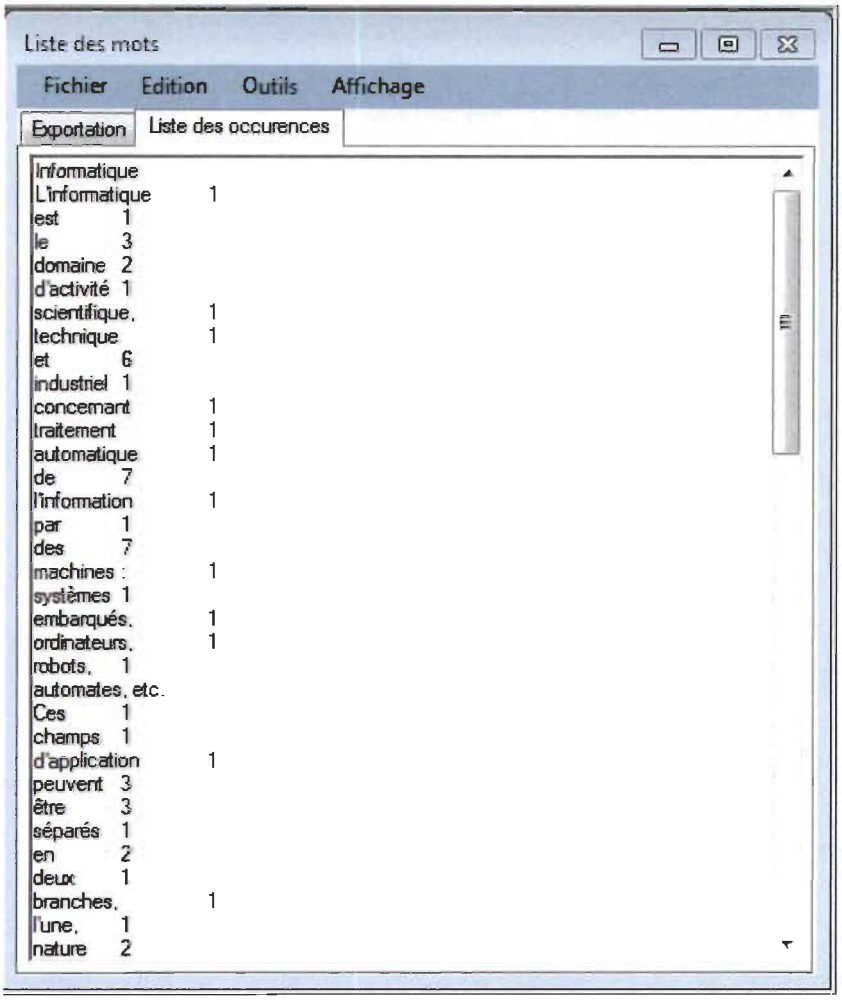

Figure 9 : Liste des mots

4.2 Ajouter des notes

L'option d'ajout des notes et des commentaires permet de faciliter la sauvegarde et compréhension du texte courant par l'utilisateur.

Figure 10 : Liste des notes

4.3 Similarité des notes

Le test de similarité est basé essentiellement sur un seuil de similarité pour éviter une certaine ambiguïté en supprimant les phrases qui se ressemblent. Ces tests sont basés sur la comparaison des n-grammes [68].

Il s'agit de comparer deux phrases pour supprimer les informations redondantes. Cette métrique permet de vérifier la similarité ou la dissimilarité entre les paragraphes.

Après le test de similarité, il aura un affichage de toutes les phrases qui ne sont pas similaires. L'utilisateur peut ainsi les extraire sous un fichier texte ou dans l'interface de l'application où il peut les modifier et les traiter [69].

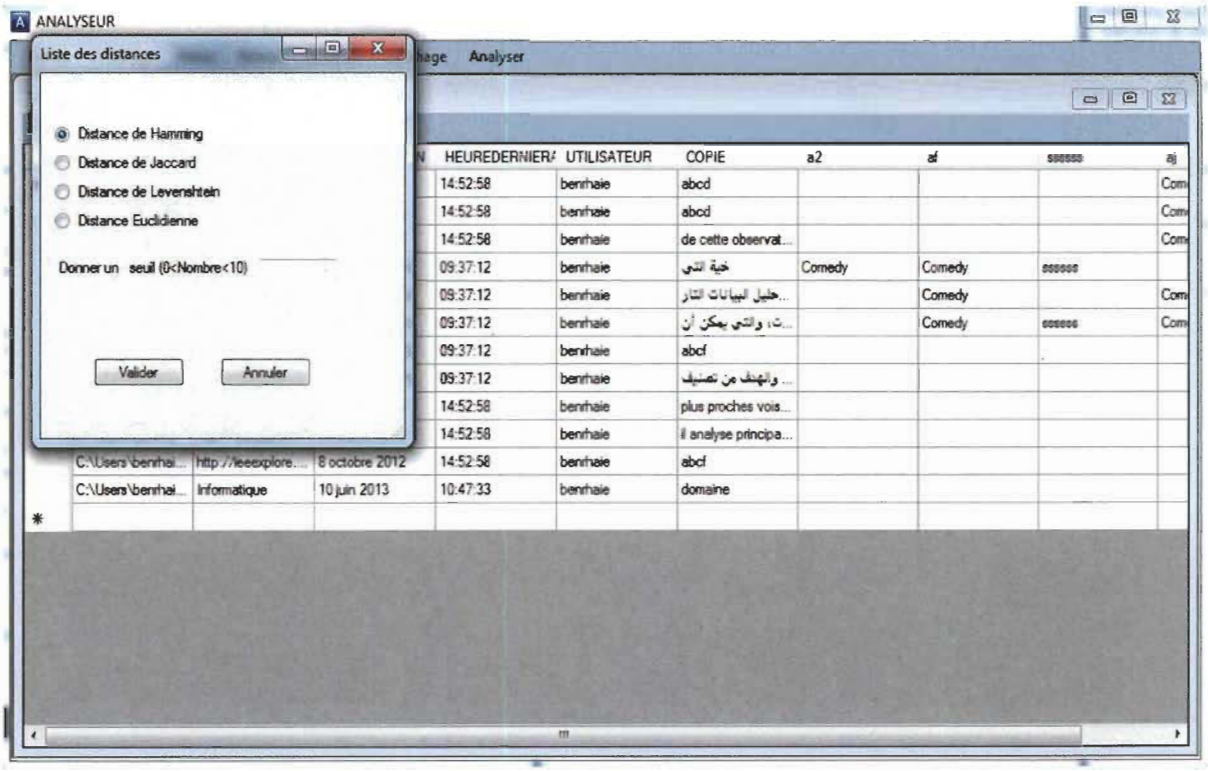

Figure 11 : calcul distance

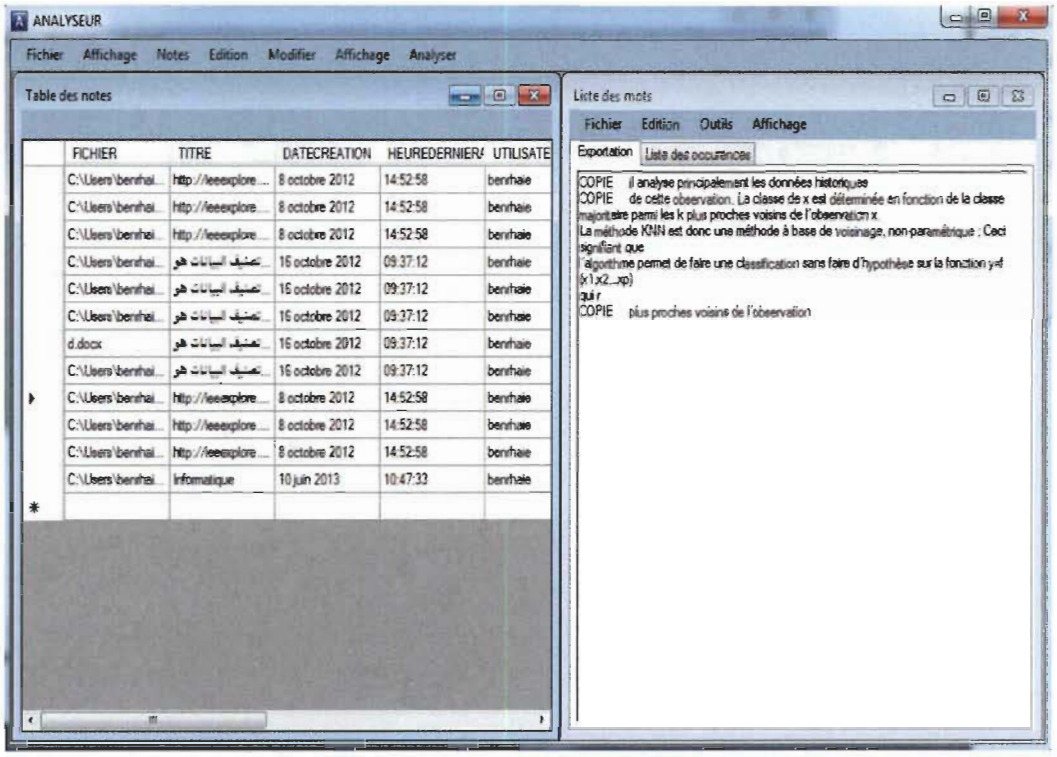

Figure 12 : Affichage des mots

Pour effectuer les tests de comparaison des phrases, on a choisi des exemples de texte qui contiennent des termes qui se répètent fréquemment. Comme paramètres de comparaison, nous avons choisi les bigrammes de caractères. Ils consistent au fait que deux phrases sont liées sémantiquement, si elles comportent une sous-séquence commune des caractères non vides [10]. Nous avons fixé un seuil de ressemblance égal á 30%. Ce seuil a été choisi dans plusieurs travaux tels que [5]. D.Bouscadi, R.tournier, N.aussenac et J. Mothe [69] ont montré qu'avec ce seuil, ils ont obtenus des bons résultats.

Les notes peuvent êtres écrites par les lecteurs comme elles peuvent être des passages copiés directement dans le texte à lire.

Pour les besoins de l'explication nous avons testé ce module par rapport à 64 phrases extraites du mémoire «Modélisation informatique de structures dynamiques de segments textuels pour l'analyse de corpus [10]» Nous avons également utilisé le texte «La fouille de texte » sur wikipédia [23].

Les phrases utilisées :

- On peut distinguer deux étapes principales dans les traitements.
- La première étape.
- Cette première étape est commune à tous les traitements.
- Une analyse sans interprétation n'a que peu d'intérêt.
- l'interprétation des textes.
- C'est le rôle de la seconde étape d'interpréter cette analyse.
- La seconde étape, l'interprétation.
- Des exemples d'applications sont la classification de courriers.
- l'application de requêtes dans un moteur de recherche.
- Le critère de sélection peut être d'au moins deux types.
- la similarité ou contradiction par rapport à un autre texte.
- La fouille de textes se distingue du traitement.
- La fouille de textes recoupe la recherche.
- Les méthodes de fouilles de texte.
- Les techniques de la fouille de texte.

- domaine de la fouille de textes.
- D'un point de vue matériel, le texte se donne d'abord comme une suite de caractères
- Du point de vue du lecteur, le texte se présente d'emblée comme une suite de mots
- Les langages de programmation sont constitués de classes lexicales définies en extension ou en intention.
- Certains algorithmes de calcul peuvent se contenter d'un découpage en unités très primitif.
- Ce découpage peut ne pas aller au-delà du caractère lui-même.
- la nature du corpus en termes de langue utilisée.
- Le lexique du corpus pourrait n'être que l'inventaire des caractères Unicode.
- Ces suites de n-grammes sont souvent utilisées dans le domaine du traitement automatique de la langue
- Les n-grammes sont généralement utilisés avec des méthodes statistiques.
- On pourrait disposer d'un dispositif de segmentation qui ferait appel à un ensemble de ressources linguistiques et terminologiques.
- permet de marquer par un symbole le dépistage d'une unité cognitive ou sémiotique.
- la propriété gramr pourrait servir à définir l'ensemble des fonctions grammaticales possibles du lexème.

- La propriété gramr est une propriété symbolique.
- L'opération de linéarisation du texte.
- La linéarisation du texte représenté dans le plan SATO peut factoriser les valeurs de propriétés.
- permettre d'annoter chacun des mots de chacune des lignes du corpus de texte.
- Le mot en contexte est une instance.
- Toutes les instances partagent les propriétés de la classe.
- À ce titre, la chaine de caractères normalisée qui permet de visualiser la forme lexicale, peut-être aussi considérée comme une propriété de la classe.
- Ajoute un dispositif d'héritage statique entre les propriétés.
- L'héritage statique signifie qu'on peut créer une nouvelle propriété dont les valeurs de départ seront héritées d'une propriété mère.
- L'ajout d'une propriété lexicale aura pour effet d'enrichir la description des classes lexicales existantes.
- L'examen du journal permet également un retour critique sur une démarche analytique effectuée.
- à partir de l'examen du journal que l'on composera les scénarios.
- Le scénario est donc issu d'opérations de repiquage.
- Il est possible d'exporter ses résultats sur un fichier externe et d'agir très finement sur les divers formats de présentation.
- Cette exportation peut être commandée au besoin.

- Il est aussi possible d'extraire divers résultats sur un fichier.
- Il correspond tout à fait à l'idée que l'on se fait d'un laboratoire.
- C'est un outil d'exploration et de découverte.
- Il est aussi possible de filtrer les caractères.
- Tous les mots débutant par a dans l'alphabet.
- Tous les mots débutant par ab.
- L'exportation consiste à écrire sur un fichier en format texte.
- Des valeurs de la propriété sur un ensemble données de formes lexicales ou d'occurrences.
- On peut attribuer des valeurs de propriétés aux formes lexicales ou aux occurrences.
- Les bornes HOMOGÈNES définissent des contextes constitués des mots.
- Les bornes NUMÉRIQUES définissent des contextes constitués d'un nombre maximal.
- On peut aussi se servir de la concordance pour réaliser une catégorisation automatique des mots repérés.
- Le sous-texte est une restriction sur l'axe des occurrences.
- Le sous-texte est une restriction de l'étendue du corpus.
- L'analyseur peut calculer divers indices de répartition et de dispersion des objets.

- L'option DOCUMENT permet de créer un sous-texte constitué des documents.
- L'option PARAGRAPHE permet de créer un sous-texte constitué des paragraphes.
- ce type de logiciel qui permet à l'utilisateur de définir diverses fonctions de calcul associées à une matrice de données.
- Des algorithmes simples de comparaison statistique entre lexiques associés à des sous-textes.
- Le corpus comprend neuf entrevues sur le tabagisme chez les jeunes.
- Le corpus est constitué de transcriptions d'entrevues de groupe sur l'usage du tabac.

Après le test de similarité, le programme a affiché les phrases suivantes :

- On peut distinguer deux étapes principales dans les traitements.
- Cette première étape est commune à tous les traitements.
- Une analyse sans interprétation n'a que peu d'intérêt.
- C'est le rôle de la seconde étape d'interpréter cette analyse.
- La seconde étape, l'interprétation.
- Des exemples d'applications sont la classification de courriers.
- l'application de requêtes dans un moteur de recherche.
- Le critère de sélection peut être d'au moins deux types.
- la similarité ou contradiction par rapport à un autre texte.

- La fouille de textes se distingue du traitement.
- Les techniques de la fouille de texte.
- domaine de la fouille de textes.
- D'un point de vue matériel, le texte se donne d'abord comme une suite de caractères
- Du point de vue du lecteur, le texte se présente d'emblée comme une suite de mots
- Les langages de programmation sont constitués de classes lexicales définies en extension ou en intention.
- Certains algorithmes de calcul peuvent se contenter d'un découpage en unités très primitif.
- Ce découpage peut ne pas aller au-delà du caractère lui-même.
- la nature du corpus en termes de langue utilisée.
- Le lexique du corpus pourrait n'être que l'inventaire des caractères Unicode.
- Ces suites de n-grammes sont souvent utilisées dans le domaine du traitement automatique de la langue
- Les n-grammes sont généralement utilisés avec des méthodes statistiques.
- On pourrait disposer d'un dispositif de segmentation qui ferait appel à un ensemble de ressources linguistiques et terminologiques.
- permet de marquer par un symbole le dépistage d'une unité cognitive ou sémiotique.

- la propriété gramr pourrait servir à définir l'ensemble des fonctions grammaticales possibles du lexème.
- La propriété gramr est une propriété symbolique.
- La linéarisation du texte représenté dans le plan SATO peut factoriser les valeurs de propriétés.
- permettre d'annoter chacun des mots de chacune des lignes du corpus de texte.
- Le mot en contexte est une instance.
- Toutes les instances partagent les propriétés de la classe.
- À ce titre, la chaine de caractères normalisée qui permet de visualiser la forme lexicale, peut-être aussi considérée comme une propriété de la classe.
- L'héritage statique signifie qu'on peut créer une nouvelle propriété dont les valeurs de départ seront héritées d'une propriété mère.
- L'ajout d'une propriété lexicale aura pour effet d'enrichir la description des classes lexicales existantes.
- L'examen du journal permet également un retour critique sur une démarche analytique effectuée.
- Le scénario est donc issu d'opérations de repiquage.
- Il est possible d'exporter ses résultats sur un fichier externe et d'agir très finement sur les divers formats de présentation.
- Cette exportation peut être commandée au besoin.
- Il est aussi possible d'extraire divers résultats sur un fichier.

- Il correspond tout à fait à l'idée que l'on se fait d'un laboratoire.
- C'est un outil d'exploration et de découverte.
- Il est aussi possible de filtrer les caractères.
- Tous les mots débutant par a dans l'alphabet.
- L'exportation consiste à écrire sur un fichier en format texte.
- Des valeurs de la propriété sur un ensemble données de formes lexicales ou d'occurrences.
- On peut attribuer des valeurs de propriétés aux formes lexicales ou aux occurrences.
- Les bornes NUMÉRIQUES définissent des contextes constitués d'un nombre maximal.
- On peut aussi se servir de la concordance pour réaliser une catégorisation automatique des mots repérés.
- Le sous-texte est une restriction de l'étendue du corpus.
- L'analyseur peut calculer divers indices de répartition et de dispersion des objets.
- L'option PARAGRAPHE permet de créer un sous-texte constitué des paragraphes.
- Ce type de logiciel qui permet à l'utilisateur de définir diverses fonctions de calcul associées à une matrice de données.

- Des algorithmes simples de comparaison statistique entre lexiques associés à des sous-textes.
- Le corpus comprend neuf entrevues sur le tabagisme chez les jeunes.
- Le corpus est constitué de transcriptions d'entrevues de groupe sur l'usage du tabac.

Dans cet exemple, nous avons donné à notre programme en entrée 64 phrases et nous avons obtenu seulement 53 phrases. L'application a enlevé tous les termes redondants. Comme par exemple, les termes « fouilles de texte », «La seconde étape » et « La première étape » ont disparu après le filtrage de notre application. La suppression de certaines phrases ou termes n'a pas changé le sens global du paragraphe (ou généralement d'un texte). Ce programme garde les phrases qu'il juge importantes et indispensables à la compréhension du texte. Ceci est en effet, défini par le seuil initial. Le seuil que nous avons choisi dans notre exemple (30%) est tiré de certaines références comme [3]. Le seuil initial est assez important car il est responsable entre autres du résultat final. Si on augmente le seuil, on obtiendra moins de phrases à la fin. De même dans le cas inverse, si on diminue le seuil, le programme enlèvera moins de phrases redondantes.

Le test de ressemblance des phrases ne vérifie pas simplement si deux phrases sont identiques ou pas. Il permet principalement d'identifier les phrases qui sont lexicalement proches. Si le nombre de termes dans les deux phrases à comparer n'est pas le même, notre algorithme est capable de trouver si elles ont un sens proche. Quand l'algorithme détecte que deux phrases sont similaires, il garde celle qui est beaucoup plus riche et plus

informative afin de maintenir le sens général du texte. Comme par exemple, après le test des deux expressions « La première étape » et « Cette première étape est commune à tous les traitements» le programme a retourné la deuxième phrase seulement.

4.3.1 Algorithme

L'algorithme utilisé dans notre application est basé sur les n-grammes. Il prend comme entrée deux phrases. Il commence à calculer la distance entre les deux phrases bigramme par bigramme. Il retourne 1 si les bigrammes sont identiques et 0 s'ils sont différents. La décision finale dépend de la distance calculée et du seuil initialement fixé au début de l'algorithme.

Exemple :

Dans cet exemple, nous avons utilisé les bigrammes, et nous avons utilisé un seuil égal à 30%.

Phrases 1: la première étape est commune (la, a_,_p, pr, re, em, mi, iè, èr; re, e_, _é, ét, ta, ap, pe, _e, es, st, t_, _c, co, om, mm, mu, un, ne).

Phrases 2 : Une première étape (Un, ne, e_,_p, pr, re, em, mi, iè, èr; re, e_, _é, ét, ta, ap, pe).

Donc le premier test entre les phrases donne 0 car les deux premiers bigrammes ne sont pas identiques (la, un). Mais par contre dans les deux bigrammes (_p, _p) donne 1.

La distance est égale à 8> seuil égal à 3.donc les deux phrases sont identiques.

La phrase sélectionnée à la fin, est la plus longue car elles sont plus riche en informations.

La distance est calculée comme suit :

$$Dist\ (p1,p2) = \frac{\sum h(x,P) \frac{1}{d(x,xmax)}}{\sum wi}$$

- p1, p2 représentent les deux phrases à tester.
- P représente le maximum entre la longueur de p1 et p2.
- $\frac{1}{d(x,xmax)}$ Représente un facteur de distance qui réduit le poids de n-gramme.
- w_i représente le poids des termes dans les phrases.

Algorithme

Cet algorithme permet de calculer la distance entre deux phrases [69].

L1 et L2 représente respectivement la longueur des deux phrases mot1 et mot2.

L1 ← length (mot1)

L2 ← length (mot2)

for i ← 0 to L1 do

    Dist[i, 0] ← i

for j ← 1 to L2 do

    Dist[0, j] ← j

for i ← 1 to L1 do

    for j ← 1 to L2 do

Dist[i, j] ← min(Dist[i − 1, j]+1,

Dist[i, j − 1] + 1,

Dist[i − 1, j − 1] + dN(Γ N i−1,j−1))

return Dist[L1, L2]/ max(L1,L2)

Le résultat final est obtenu par le calcul de la distance entre les deux phrases. Par la suite, il suffit de diviser la valeur de la distance de similarité trouvée par la longueur maximale entre les deux phrases (L1 et L2). Ce calcul retourne 1 si les deux phrases sont similaires et 0 si elles ne le sont pas.

Dans ce qui suit, nous allons évaluer les résultats obtenus par notre application sur l'exemple présenté précédemment. Nous allons introduire la sensibilité et la spécificité de notre algorithme. La sensibilité permet de donner un résultat positif lorsqu'il est réellement positif. Elle est définie par le rapport « sensibilité = Vrai Positif/ (Vrai Positif +Faux Négatif) ». La spécificité permet au contraire de donner un résultat négatif lorsqu'il est réellement négatif. Elle est définie par le rapport «spécificité = Vrai Négatif/ (Vrai Négatif + Faux Positif) » [70].

Le vrai positif (VP) est un test qui permet de mesurer le nombre de phrases qui sont réellement identiques et que l'algorithme les a trouvées identiques après le test de similarité.

Le vrai négatif (VN) est un test qui permet de mesurer le nombre de phrases qui sont non identiques et que l'algorithme les a trouvées non identiques après le test de similarité.

Le faux positif (FP) est un test qui permet de mesurer le nombre de phrases qui sont non identiques et que l'algorithme les a trouvées identiques après le test de similarité.

Le faux négatif (FN) est un test qui permet de mesurer le nombre de phrases qui sont réellement identiques et que l'algorithme les a trouvées non identiques après le test de similarité.

La Liste des phrases qu'on le considère comme non similaire et qui sont ressortie comme non similaire (Vrai négatif):

- On peut distinguer deux étapes principales dans les traitements.
- Cette première étape est commune à tous les traitements.
- Une analyse sans interprétation n'a que peu d'intérêt.
- C'est le rôle de la seconde étape d'interpréter cette analyse.
- La seconde étape, l'interprétation.
- Des exemples d'applications sont la classification de courriers.
- l'application de requêtes dans un moteur de recherche.
- Le critère de sélection peut être d'au moins deux types.
- la similarité ou contradiction par rapport à un autre texte.
- D'un point de vue matériel, le texte se donne d'abord comme une suite de caractères
- Du point de vue du lecteur, le texte se présente d'emblée comme une suite de mots
- Les langages de programmation sont constitués de classes lexicales définies en extension ou en intention.

- Certains algorithmes de calcul peuvent se contenter d'un découpage en unités très primitif.

La Liste des phrases qu'on le considère comme similaire et qui sont ressortie comme non similaire (Faux négatif) :

- Ces suites de n-grammes sont souvent utilisées dans le domaine du traitement automatique de la langue
- Les n-grammes sont généralement utilisés avec des méthodes statistiques.
- Le corpus comprend neuf entrevues sur le tabagisme chez les jeunes.
- Le corpus est constitué de transcriptions d'entrevues de groupe sur l'usage du tabac
- Des valeurs de la propriété sur un ensemble données de formes lexicales ou d'occurrences.
- On peut attribuer des valeurs de propriétés aux formes lexicales ou aux occurrences.

La Liste des phrases qu'on le considère comme non similaire et qui sont ressortie comme similaire (Faux positif):

- La fouille de textes se distingue du traitement.
- La fouille de textes recoupe la recherche.
- Ajoute un dispositif d'héritage statique entre les propriétés.

- L'héritage statique signifie qu'on peut créer une nouvelle propriété dont les valeurs de départ seront héritées d'une propriété mère.
- L'examen du journal permet également un retour critique sur une démarche analytique effectuée.
- à partir de l'examen du journal que l'on composera les scénarios.

La liste des phrases qu'on le considère comme similaires et qui sont ressorties comme similaires :

- La première étape.
- Cette première étape est commune à tous les traitements.
- l'interprétation des textes.
- C'est le rôle de la seconde étape d'interpréter cette analyse
- La fouille de textes se distingue du traitement.
- La fouille de textes recoupe la recherche.
- Les méthodes de fouilles de texte.
- Les techniques de la fouille de texte.
- L'opération de linéarisation du texte.
- La linéarisation du texte représenté dans le plan SATO peut factoriser les valeurs de propriétés.
- Ajoute un dispositif d'héritage statique entre les propriétés.

- L'héritage statique signifie qu'on peut créer une nouvelle propriété dont les valeurs de départ seront héritées d'une propriété mère.
- L'examen du journal permet également un retour critique sur une démarche analytique effectuée.
- à partir de l'examen du journal que l'on composera les scénarios.
- Tous les mots débutant par a dans l'alphabet.
- Tous les mots débutant par ab.
- Les bornes HOMOGÈNES définissent des contextes constitués des mots.
- Les bornes NUMÉRIQUES définissent des contextes constitués d'un nombre maximal.
- Le sous-texte est une restriction sur l'axe des occurrences.
- Le sous-texte est une restriction de l'étendue du corpus.
- L'option DOCUMENT permet de créer un sous-texte constitué des documents.
- L'option PARAGRAPHE permet de créer un sous-texte constitué des paragraphes.

Le tableau suivant donne une vision claire sure le vrais positif, le vrais négatif, le faux positif et le faux négatif.

	Positif	Négatif
Vrais	22	13
Faux	6	8

Tableau 4 : Mesure de sensibilité et spécificité

- Sensibilité = VP/ (VP+FN)= 22/22+8=0,7
- La spécificité =VN/ (VN+FP)=13/13+6=0.6

Le calcul de la sensibilité nous donne (0,7). Ceci indique que notre application a donné des résultats assez conformes à la vérité.

Le résultat du calcul de la spécificité a donné (0,6). Ceci indique que l'application a réussi à identifier les phrases qui sont similaires parmi la liste enregistrée par l'utilisateur.

Le ratio des phrases similaires représente le nombre de phrases déclarées identiques par rapport au nombre de phrases qui sont réellement identiques. Il est égal à 38/51.

Le ratio des phrases non similaires représente le nombre de phrases déclarées non identiques par rapport au nombre de phrases qui sont réellement non identiques. Il est égal à 8/13.

4.4 Exportation

L'exportation textuelle permet d'enregistrer le texte sous un autre format et dans un autre emplacement. Dans notre cas l'utilisateur peut exporter son texte dans un fichier «.txt». C'est l'utilisateur qui choisit le nom du fichier et son emplacement. Aussi, il peut l'exporter vers une autre interface de l'application où il peut le modifier selon son besoin pour faciliter son analyse.

L'exportation sous le format (.txt) facilite la manipulation de l'utilisateur, car on peut ouvrir ce fichier sous n'importe quelle machine. Aussi on peut appliquer les modifications sous n'importe qu'elle machine quelque soit son système d'exploitation.

La manipulation du format text est très rapide et facile. Aussi elle nous permet d'enregistrer un document sous plusieurs formats. [67].

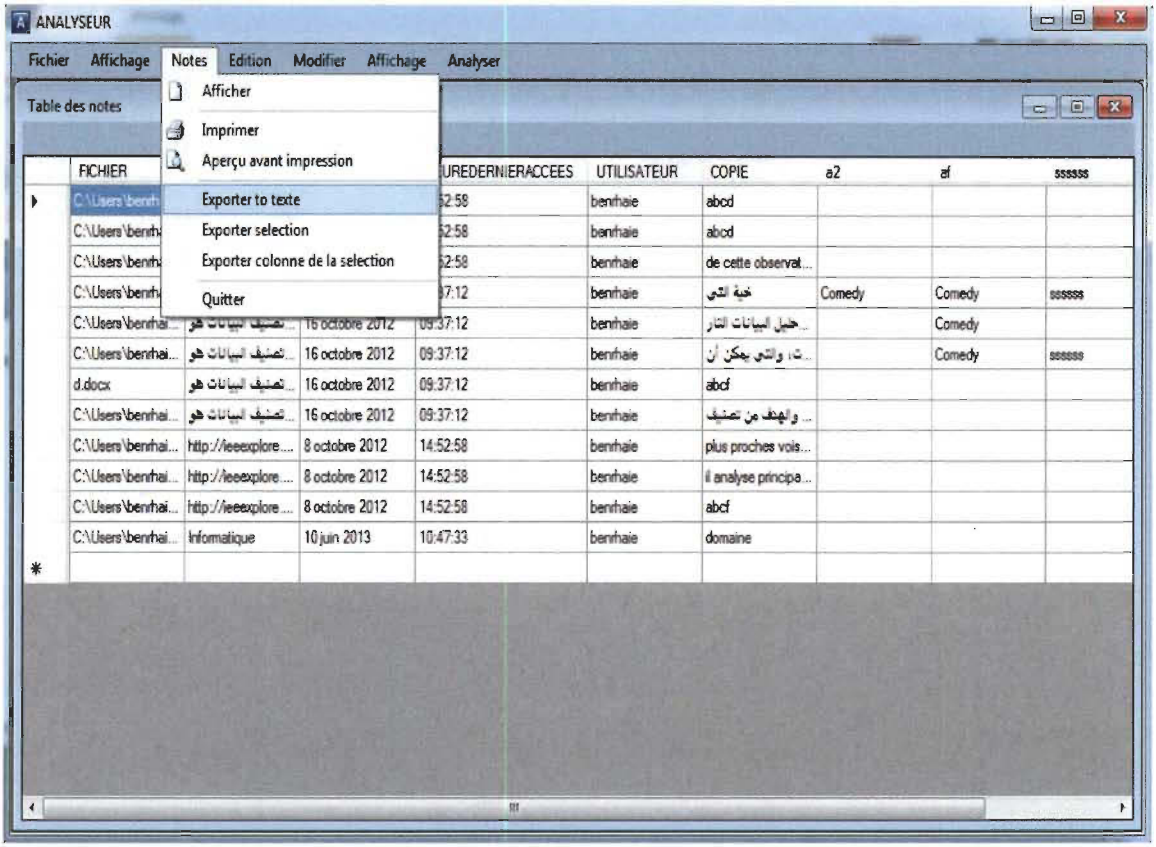

Figure 13 : Exportation

4.5 Questionnaire

En vue d'une expérimentation qui tient compte de l'opinion des utilisateurs. Nous avons réalisé un questionnaire pour savoir les points de vue de 15 utilisateurs et leurs avis sur le fonctionnement de notre application. Nous voulons savoir les fonctionnalités qui représentent les tâches les plus importantes pour eux et celles qui sont facultatifs.

Nous avons mentionné aussi le but de chaque question et son intérêt.

Question 1 : Quel est votre type d'utilisateur ?

Justification 1 : On a posé cette question pour savoir le type de l'utilisateur de cette application (lecteur professionnel, lecteur simple, étudiant).

Question 2 : Quelle est la tache la plus importante d'après vous ?

Justification 2 : Cette question indique la tâche la plus importante et la plus utilisée pour chaque type d'utilisateur dans cette application.

Question 3 : Quel est votre avis sur ce programme ?

Justification 3 : Cette question permet de donner le point de vue des utilisateurs sur cette application et leur niveau de satisfaction.

Question 4 : Que pensez-vous de la lecture que vous venez de faire ? Est-ce que cette application facilite votre lecture et votre compréhension?

Justification 4 : Cette question permet de savoir si les utilisateurs ont utilisé les fonctionnalités existantes, et s'ils aident à la compréhension du texte.

Question 5 : Quel est l'avantage de l'affichage de la liste des mots et le nombre de leurs apparitions dans le texte ?

Justification 5 : Nous posons cette question pour savoir l'utilité de cette tâche pour chaque type d'utilisation.

Question 6 : Trouvez-vous que l'exportation du texte est utile dans votre lecture ?

Justification 6 : Cette question indique si les utilisateurs ont utilisé l'exportation. Et si elle est utile ou non.

Question 7 : Est-ce que la rédaction des commentaires et l'enregistrement des notes sont pratiques d'après vous ?

Justification 7 : Cette question permet de donner la valeur et l'importance de la rédaction des commentaires et l'enregistrement des notes.

Question 8 : La comparaison des phrases enregistrées est-elle efficace pour la compréhension de votre texte ? Comment l'avez-vous utilisée ?

Justification 8 : Cette question permet de monter l'importance de l'outil de comparaison des phrases enregistrées.

4.6 Réponse aux questions

On a soumis le questionnaire à 15 personnes. Cinq personnes sont étudiantes au département de mathématiques et informatique, ils sont diplômés à la maitrise. Cinq autres personnes ont un diplôme en administration des affaires. Cinq personnes sont étudiantes en littérature.

Les réponses des personnes sondées sont mises ci-dessous :

- **Question1 :** Quel est votre type d'utilisation de ce logiciel ?
  - **Réponse 1 :** J'ai utilisé ce logiciel dans le cadre de ma maîtrise pour des lectures bibliographiques des articles et des revues. Et je vais donner comme exemple, l'article intitulé « Reconnaissance faciale automatisée dans les secteurs public et privé ».

- **Réponse2 :** Ce logiciel était très utile pour moi pour la rédaction du chapitre « état de l'art » de mon mémoire. En effet, il m'a facilité la tâche de résumer les documents et de garder les informations les plus pertinentes
- **Réponse 3 :** J'ai utilisé ce logiciel pour la lecture d'un chapitre de mon cours « Droit fiscal ». Il est très efficace pour faciliter ma lecture, et de prendre des notes.
- **Réponse 4 :** J'ai utilisé cette application pour analyser un mémoire intitulé « Modélisation informatique de structures dynamiques de segments textuels pour l'analyse de corpus [10]» que je viens de le lire. C'est très pratique avec l'enregistrement des définitions ainsi que pour des différentes notions. Puis on est capable de supprimer les termes qui se répètent.
- **Réponse 5 :** Pour la lecture et l'analyse, cette application est très utile. Suite à ma lecture de l'article « Une seule intelligence », j'ai utilisé aussi l'exportation des données enregistrées. Ce qui est assez flexible dans ce logiciel.
- **Réponse 6 :** J'ai utilisé cette application pour la lecture d'un article scientifique intitulé « Les spidrons, pliables à l'infini».

- **Réponse 7 :** Cette application est très efficace pour la compréhension de mon livre intitulé « علم الحاسوب». En effet, j'ai rédigé des commentaires et j'ai enregistré des phrases importantes afin de faciliter ma lecture.

- **Réponse 8 :** J'ai utilisé cette application pour une simple lecture du texte « La science ». Le plus important dans cette application est qu'on n'est pas obligé d'utiliser un logiciel spécifique pour chaque type de fichier.

- **Réponse 9 :** J'utilise ce logiciel pour lire mes documents texte enregistrés afin de pouvoir déceler les parties importantes qui me concernent dans le texte. Il m'a aidé beaucoup pour la compréhension du texte « L'image thermique une technologie prête à conquérir le reste du monde ».

- **Réponse 10 :** J'utilise ce logiciel pour résumer mes textes et les simplifier pour plus de clarté. Je trouve l'enregistrement des mots clés intéressants ainsi que l'exportation de mes notes déjà enregistrées.

- **Réponse 11 :** Mon utilisation de ce logiciel consiste principalement à corriger mon texte en éliminant les répétitions afin d'avoir un texte plus clair et plus cohérent. Ceci m'aide à mieux le comprendre.

- **Réponse 12 :** Ce logiciel me permet de faire la mise en forme de mes textes. Il me permet aussi de pouvoir enregistrer des commentaires et des notes sur mes rédactions. Ceci qui facilite l'échange d'information entre les différents utilisateurs de ce document.

- **Réponse 13 : C**e logiciel me permet de choisir la mise en forme, d'enregistrer des notes pour chaque paragraphe qui m'intéresse et de sauvegarder les mots clés qui me seront utiles dans mes travaux. Je

trouve que ce logiciel est vraiment utile pour un traitement clair et approfondi du texte.

- **Réponse 14 :** J'ai utilisé ce logiciel pour des comptes rendus. Il me permet de synthétiser l'essentiel de mon travail.
- **Réponse 15 :** J'ai bien aimé l'aspect intuitif du logiciel. Enfin ce qui est très pratique dans ce logiciel, c'est que l'on puisse utiliser n'importe quel format de document, il n'y a plus de contraintes.

- **Question 2 :** Quelle est la tâche la plus importante d'après vous ?
  - **Réponse 1 :** La gestion et la sauvegarde des notes et des remarques intéressantes.
  - **Réponse 2 :** La compression des données et la sélection des informations les plus pertinentes.
  - **Réponse 3 :** D'après moi la rédaction des notes et des commentaires m'ont bien été utiles pour la compréhension de mes textes. Cela me permet d'avoir directement une idée sur le contenu d'un paragraphe déjà lu, sans être forcé à faire une deuxième lecture.
  - **Réponse 4** : Je pense que l'enregistrement des mots clés et la surpression des phrases répétitives sont une fonction intéressante de ce logiciel.
  - **Réponse 5 :** La fonction qui permet l'affichage du nombre de mots dans le texte et leurs nombres d'apparitions est vraiment essentielle à une rédaction approfondie. Cela permet de choisir précisément les termes exacts dans chaque phrase et de les corriger.

- **Réponse 6 :** L'affichage de la liste des mots est la tâche la plus importante. Car elle donne une vue sur le contenu du document avant de le lire.
- **Réponse 7 :** L'enregistrement des notes est d'après moi le plus utile. Car simplement en me basant sur les notes déjà écrites lors de ma première lecture ou traitement, je peux comprendre mon texte.
- **Réponse 8 :** La rédaction des commentaires sur le texte est la tache la plus importante d'après moi.
- **Réponse 9 :** L'enregistrement des phrases importantes et la vérification et la suppression des répétitions.
- **Réponse 10 :** La recherche des mots est une tâche très importante dans cette application.
- **Réponse 11 :** L'affichage de la liste des mots et la recherche sont deux fonctions complémentaires et importantes.
- **Réponse 12 :** d'après moi la tache la plus importante est la sauvegarde des mots, car elle me permet de revenir sur l'essentiel.
- **Réponse 13 :** La recherche des données dans le texte.
- **Réponse 14 :** La comparaison des phrases enregistrées.
- **Réponse 15 :** Lorsque je veux sauvegarder des parties de mon travail, je peux y revenir facilement grâce à ce logiciel.

o **Question 3** : Quel est votre avis sur ce programme ?

- **Réponse 1 :** Ce programme est efficace et pertinent.
- **Réponse 2 :** Ce programme est utile.
- **Réponse 3 :** ce programme m'a été d'une grande utilité pour mes révisions, il est vraiment pratique.
- **Réponse 4 :** Ce programme est une belle découverte grâce à sa facilité d'utilisation ainsi que son efficacité.
- **Réponse 5 :** ce logiciel m'a permis de bien étudier en profondeur mes textes.
- **Réponse 6 :** J'ai adopté ce programme depuis ma première utilisation. La différence se fait ressentir, je maitrise mieux mes textes. Ce logiciel est vraiment pratique.
- **Réponse 7 :** Ce programme est très utile.
- **Réponse 8 :** Cette application facilite nos travaux.
- **Réponse 9 :** Grâce aux options offertes, cette application nous donne plusieurs solutions pour comprendre notre texte.
- **Réponse 10 :** Programme facile à utiliser.
- **Réponse 11 :** Pour un document volumineux, cette application est nécessaire.
- **Réponse 12** : je dirais que ce programme est utile dans mes travaux lorsque je dois résumer un livre.
- **Réponse 13 :** C'est pratique pour extraire des parties de mon cours.

- **Réponse 14 :** Facilite l'extraction des données textuelles.
- **Réponse 15 :** Permet d'extraire les formules de mon cours.

○ **Question 4**. Que pensez-vous de la lecture que vous venez de faire ? Est-ce que cette application facilite votre lecture et votre compréhension?

- **Réponse 1 :** Ma lecture est plus facile. En fait, le logiciel m'a permis d'enregistrer les phrases et les mots clés dans le texte.
- **Réponse 2 :** La lecture était très pratique. Ce logiciel m'a permis d'afficher la liste des mots dans le texte et le nombre de leurs apparitions.
- **Réponse 3 :** oui en effet, ce programme m'a permis de ressortir les différents termes utilisés lors de mes rédactions. Ceci qui m'a permis de revérifier le contenu de mes textes ainsi que le réajuster et de savoir exactement le nombre des mots utilisés ainsi que leurs nombres d'apparitions. Ceci m'a aidé à faire un texte plus clair.
- **Réponse 4 :** Ce programme m'a permis de résumer mes lectures d'une manière efficace. J'ai pu prendre des notes et enregistrer les mots clés.
- **Réponse 5 :** Oui ce logiciel m'a été vraiment utile pour mes lectures de textes et mes rédactions. J'ai pu utiliser la fonction qui me permet de supprimer les phrases redondantes.
- **Réponse 6 :** C'est facile avec la recherche des mots et la rédaction des notes.

- **Réponse 7 :** Ce programme facilite ma lecture avec les différentes options existantes.
- **Réponse 8 :** Les options de mise en forme m'ont aidé à structurer mon document avant la lecture.
- **Réponse 9 :** Je suis satisfait de cette application.
- **Réponse 10 :** Ce programme fournit beaucoup d'options importantes dans notre lecture.
- **Réponse 11 :** Cela est très pratique pour moi, car ça facilite ma lecture.
- **Réponse 12 :** Il permet de faire un bon compte rendu de la lecture.
- **Réponse 13 :** Grâce à la mise en forme, mon texte est plus lisible.
- **Réponse 14 :** Ce logiciel m'a beaucoup aidé sur la lecture de mon texte.
- **Réponse 15 :** grâce à ce logiciel, je peux comprendre un texte à partir de la première lecture.

- **Question 5**. Quel avantage donne l'affichage de la liste des mots et le nombre de leurs apparitions dans le texte ?
  - **Réponse 1 :** Donner un degré d'importance pour chaque mot de la liste.
  - **Réponse 2 :** Déduire le contenu du document plus rapidement.
  - **Réponse 3 :** cette fonction me permet de savoir sur quels objets mes lectures se basent. Je peux ainsi mieux préciser mon sujet.

- **Réponse 4 :** cet outil m'a permis d'éviter des répétitions afin d'avoir un texte clair et sans répétitions ce qui facilite sa compréhension.
- **Réponse 5 :** cette fonction m'a permis de cerner les mots clés qui sont reliés aux parties qui m'intéressaient.
- **Réponse 6 :** Je n'ai pas trouvé une importance dans cette tache.
- **Réponse 7 :** Cette méthode permet de donner une vue générale sur le texte.
- **Réponse 8 :** l'apparition des mots donne une idée globale sur le texte.
- **Réponse 9 :** Je n'ai pas utilisé cette fonction dans mon travail. Mais elle peut être utile dans un autre travail.
- **Réponse 10 :** Méthode utile pour certains travaux.
- **Réponse 11 :** Cette tâche permet d'avoir une idée rapide sur le document avant de le lire.
- **Réponse 12 :** Je n'ai pas utilisé cette fonction du logiciel.
- **Réponse 13 :** Elle peut être importante si je veux faire ressortir les mots les plus utilisés dans le texte.
- **Réponse 14 :** Facilite la compréhension du texte.
- **Réponse 15 :** Cette fonction n'est pas importante.

o **Question 6**. Vous trouvez que l'exportation du texte est utile dans votre lecture ?

- **Réponse 1 :** Oui, c'est très utile pour rédiger un résumé
- **Réponse 2 :** je pense que cette tâche est secondaire.
- **Réponse 3 :** C'est indispensable pour rédiger un résumé.
- **Réponse 4 :** Cette tâche n'est pas importante dans l'analyse et la compréhension du texte
- **Réponse 5 :** Pas utile.
- **Réponse 6 :** Cette tâche n'est pas nécessaire dans la compréhension du texte.
- **Réponse 7 :** Oui, je l'ai utilisé pour rédiger mon résumé.
- **Réponse 8 :** je ne l'ai pas trouvé utile dans ma lecture.
- **Réponse 9 :** C'est très nécessaire pour qu'on puisse rédiger un résumé sur le texte.
- **Réponse 10 :** Je l'ai utilisé pour extraire les notes dans l'interface de l'application afin d'appliquer des modifications sur les notes écrites.
- **Réponse 11 :** fonction pas utile.
- **Réponse 12 :** selon moi ce n'est pas encore utile.
- **Réponse 13 :** Je pense que pour une grande lecture, cette fonction peut être nécessaire.
- **Réponse 14 :** Cela facilite la modification de mes notes enregistrées.
- **Réponse 15 :** Cette méthode facilite la rédaction de mon résumé.

- **Question 7**. La rédaction des commentaires et l'enregistrement des notes sont pratiques d'après vous ?

  - **Réponse 1 :** Oui, c'est pratique; surtout, quand je prends de temps pour revoir le document. Il me rafraichit la mémoire, sans être obligé de relire tout le document.
  - **Réponse 2 :** Oui, c'est utile. Il permet de gagner du temps.
  - **Réponse 3 :** Oui, en effet, j'ai pu créer des petits mémos à chaque partie de mon texte ce qui m'aidé à comprendre le sujet plus facilement.
  - **Réponse 4 :** en effet, cet outil m'a permis de mettre certaine idée sous forme de note ou commentaires afin de bien pouvoir les pétitionner dans mon texte dans une prochaine utilisation.
  - **Réponse 5 :** Cet outil m'a permis de prendre des notes qui m'aident à clarifier le sens de certaines phrases, alors oui, cet outil est pratique.
  - **Réponse 6 :** Pour un texte volumineux, cette tâche sera très importante.
  - **Réponse 7 :** La rédaction des notes est très importante pour la compréhension du texte.
  - **Réponse 8 :** Cette méthode est très utile dans la lecture du texte.
  - **Réponse 9 :** Oui cet outil est très efficace pour extraire les informations nécessaires du texte.
  - **Réponse 10 :** Cette tâche permet de simplifier le texte.

- **Réponse 11 :** Cette méthode réduit le nombre d'informations dans le texte.
- **Réponse 12 :** Oui c'est pratique, lorsque je veux revenir sur une partie incomplète, je peux revenir sur les idées et les commentaires pour la finaliser.
- **Réponse 13 :** Si je veux commenter une partie de mon cours.
- **Réponse 14 :** Lorsque je veux résumer mon cours.
- **Réponse 15 :** Permet de simplifier la compréhension du texte.

- **Question 8**.Est-ce que la comparaison des phrases enregistrées est efficace pour la compréhension de votre texte ?

  - **Réponse 1 :** La Comparaison des phrases enregistrées et la suppression de celles qui se répètes me permet de garder seulement les informations les plus pertinentes et de réduire la taille du fichier à relire
  - **Réponse 2 :** Cette tâche permet de réduire le nombre d'informations inutiles.
  - **Réponse 3 :** Cette tâche facilite la compréhension du document.
  - **Réponse 4 :** Avec l'ajustement du seuil de ressemblance, on peut contrôler la comparaison.
  - **Réponse 5 :** C'est l'étape la plus utile pour la compréhension du texte.
  - **Réponse 6 :** ça dépend des notes et des phrases enregistrées.

- **Réponse 7 :** Si on a un texte volumineux oui, sinon une lecture est suffisante.
- **Réponse 8 :** Cette fonction est utile pour gérer les phrases enregistrées.
- **Réponse 9 :** C'est la tâche la plus importante dans l'application.
- **Réponse 10 :** Pour comprendre notre texte, et pour rédiger un résumé, il faut vraiment passer par cette étape.
- **Réponse 11 :** Chaque utilisateur a besoin de cette technique pour comprendre son texte.
- **Réponse 12 :** La comparaison entre les commentaires est importante pour supprimer des éléments qui se répètent.
- **Réponse 13 :** Grâce au seuil de ressemblance quand je veux éviter les ambiguïtés.
- **Réponse 14 :** Cette tâche est complémentaire à la tâche d'enregistrement.
- **Réponse 15 :** La compréhension du texte est basée sur cette fonction.

4.6.1 Analyse du questionnaire

On peut résumer les réponses des répondants de cette façon :

En effet on a remarqué d'après les avis des utilisateurs que les taches les plus importantes dans cette application sont :

- L'analyse du texte à travers l'enregistrement des phrases importantes.
- L'ajout des commentaires et des remarques.
- La suppression des notes qui se répètent et son influence sur la compréhension du texte.
- L'exportation des données.

On remarque aussi qu'il y a un grand nombre d'utilisateurs qui sont satisfaits de cette application. Les autres, ça se peut qu'ils aient une utilisation juste basique.

D'après la lecture des réponses, les étudiants en informatique essayent toutes les fonctionnalités du logiciel et donnent leurs avis. Les autres étudiants ont utilisé juste les options nécessaires. Ceci nous montre qu'il y a des utilisateurs qui n'ont pas consulté le manuel d'utilisation de cette application.

La majorité des utilisateurs trouve que les fonctions d'enregistrement des notes et des commentaires ainsi que les tests de ressemblance entre les phrases sont très importants dans la lecture d'un texte.

En général, il n'y a pas de commentaires négatifs sur l'application. Chaque utilisateur s'intéresse à une tâche particulière.

Selon quelques utilisateurs, il y a des fonctions qui ne sont pas importantes. Comme l'extraction des données et des notes enregistrées. En effet, cette tâche est plutôt utilisée par les utilisateurs qui font des lectures ou des rédactions approfondies.

4.7 Conclusion

Dans ce chapitre, nous avons effectué plusieurs tests sur les différentes fonctionnalités de notre application. Nous avons testé notre application en utilisant différents textes en deux langues: français et arabe. Notre application a été testée par différentes catégories d'utilisateurs. On peut distinguer deux principales classes d'utilisateurs: simples et spécialisés. L'évaluation de notre logiciel est assurée par un questionnaire à remplir par chaque utilisateur à la fin de son test. L'objectif du questionnaire est d'évaluer les différentes fonctionnalités de l'application, ressortir celles qui sont les plus importantes et savoir les degrés de satisfaction des utilisateurs.

Les résultats obtenus ont montré une satisfaction parmi les utilisateurs. Selon les résultats obtenus, les fonctions d'enregistrement des notes et des commentaires ainsi que les tests de ressemblance entre les phrases sont très utiles dans la lecture d'un texte.

Conclusion

Dans ce mémoire, nous présentons une nouvelle application d'édition et d'analyse des données textuelles. Comme plusieurs autres applications existantes, notre application offre un grand nombre de fonctionnalités de base de traitement du texte. Elle ajoute également de nouvelles fonctionnalités sur le texte. La contribution majeure de notre travail est l'affichage de tous les mots du texte avec le nombre de leurs apparitions, la sauvegarde des mots clés dans le texte, la suppression de ceux qui se ressemblent et l'affichage de tous les mots dans le texte. La fonctionnalité qui permet d'extraire les données dans un autre fichier texte ou dans une autre interface de l'application facilite à l'utilisateur la rédaction de son résumé. En effet un utilisateur peut tout faire dans cette application sans avoir besoin d'utiliser un autre logiciel.

Notre application a été testée par plusieurs utilisateurs. Nous pouvons les classer en deux grands types d'utilisateurs: utilisateurs simples et utilisateurs spécialisés. Les utilisateurs simples sont ceux qui se contentent de quelques fonctions de base comme la lecture simple. Les utilisateurs spécialisés sont ceux qui exploitent la grande majorité des fonctionnalités de l'application. Les utilisateurs ont évalué notre application sur plusieurs aspects comme leurs avis sur le logiciel, leurs niveaux de satisfaction, leurs types d'utilisation en fonction de chaque fonctionnalité. L'évaluation de notre logiciel a été assurée par un questionnaire remis à chaque utilisateur. Il est demandé de le remplir à la

suite de l'utilisation de l'application. Comme résumé dans le chapitre 4, nous avons obtenu une majorité de satisfactions sur différentes fonctions : la rédaction, l'enregistrement et la comparaison des notes qui aide à mieux comprendre un texte. La majorité des utilisateurs ont trouvé que les fonctions d'enregistrement des notes et des commentaires ainsi que les tests de ressemblance entre les phrases sont très importants dans la lecture d'un texte.

Pour résumer, dans le cadre de ce mémoire, nous avons présenté un logiciel de lecture de texte simple à utiliser, adéquat à différents types d'utilisateurs présentent plusieurs nouvelles fonctionnalités très intéressantes et utiles aux usagers.

L'utilisation d'un dictionnaire dans cette application sera très intéressante pour les utilisateurs. Ce qui facilite l'affichage des synonymes et des antonymes pour les utilisateurs, créer des méthodes qui permettent de classifier les mots qui ont le même sens et de segmenter le texte.

Bibliographies

[1] Abdelghani, A., *Extraction de relations d'associations maximales dans les textes : Présentation graphique*. 2012.

[2] Biskri, I., Achouri, A., Rompré, L., Descoteaux, S. et Bensaber, B., *Computer-Assisted Reading: Getting Help from Text Classification and Maximal Association Rules.* Journal of Advances in Information Technology, 2013.

[3] Biskri, I., Meunier, J. and Joyal, S., *The extraction of Complex term: a semi-automatic modular approach.* 2004.

[4] Blanc, O., *Algorithme d'analyse syntaxique par grammaires lexicalisées : optimisation et traitement de l'ambigüité, thèse de doctorat*. 2006.

[5] Bouscadi, D., tournier, R., aussenac, N., Mothe, J., *IRIT : Textual similarity combining conceptual similarity with n-gram comparison method.* 2012.

[6] Brier, A. et Hopp, B., *Computer assisted text analysis in the social sciences.* Journal Quality & Quantity, 2011.

[7] Cabanes, G., *Classification non supervisée à deux niveaux guidée par le voisinage et la densité*. 2010.

[8] Chappelier, J. et Kolas, J., *Distance de Hamming et poids d'un mot de code*.

[9] Coulomb, D. et Kayser, D., *Informatique et langage naturel : présentation générale des méthodes d'interprétation des textes écrits*. 1986.

[10] Daoust, F., *Modélisation informatique de strectures dynamiques de segments textuels pour l'analyse de corpus*. 2012.

[11] Duchastel, J., Dupuy, L., Paquin, L., Beauchemin, J. et Daoust, F., *Système d'analyse de contenu assistée par ordinateur (SACAO)*. 1989.

[12] *Groupe provincial de soutien pour une approche orientant à l'école.* 2009.

[13] Guess, A., *Aylien Announces Intelligent Text Analysis Google Sheets Add-on.* 2014.

[14] Haccoun, A., *Comparaison de méthodes de classifications, Laboratoire de Recherche Informatique*. 2012.

[15] Hadd, M., *Classification de la population en categories socio-économiques: méthodologie et application pratique*. 1999.

[16] Han, J., *Learning Fuzzy Association Rules and Associative Classificatio Rules.* 2006.

[17] Hilali, H., *Application de la classification textuelle pour l'extraction des règles d'association maximales*. 2009.

[18] Hwang, M., Choi, J., Kim, P., Choi, D. and Lee, H., *Text editor based on google Trigram and its usability.* in IEEE Computer Modeling and Simulation (EMS), Fourth UKSim European Symposium on, 2010.

[19] Jabeur, K. et Guitouni, A., *A generalized framework for concordance/ Disconcordance-based multi-criteria classification methods.* IEEE conference publications, 2007.

[20] Jalam, R. and Chauchat, J., *Pourquoi les n-grammes permettent de classer des textes? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques, Journée internationales d'analyse statistique des données textuelles*. 2002.
[21] Jollois, F., *Contribution de la classification automatique à la fouille de Données*. 2003.
[22] Kober-Smith, A., T. Whitton, *La technique du commentaire de texte*. 2012.
[23] Kondrak, G., *N-gram similarity and distance*. 2005.
[24] Lafourcade, M., *Lexique et analyse sémantique de texte – structure, acquisition, calculs, et jeux de mots*. 2011.
[25] Lallich, S., Teytaud, O., *Evaluation et validation de l'intérêt des règles d'association*. 2004.
[26] Lecroq, T., *Extraction des règles d'association*. 2010.
[27] Liaudet, B., *Cours de data mining 7 : Modelisation non-supervisee recherche d'associations, 3ème édition revue et corrigée*. 2008.
[28] Lucas, C., Nielsen, R., Robert, M., Stewart, B., Storer, A. et Tingley, D., *Computer assisted text analysis for comparative politics*. 2013.
[29] Maitre, A., *La base logiciel du Poste de Lecture Assistée par ordinateur (PLAO)* 2002.
[30] Meunier, J., *CARAT-Computer-Assisted Reading and Analysis of texts : The Appropriation of a technologiy*. 2009.
[31] Muralidharan, A. et Hearst, M., *Supporting exploratory text analysis in Literature study*. Oxford journals, 2012.
[32] Nabil, B., *Méthodes de classification Multicritère Méthodologie et applications à l'aide au Diagnostic Médical*. 2000.
[33] Pascal, P., *Extraction de motifs : Règles d'association et motifs séquentiels*. 2009.
[34] Pasquier, N., *Extraction de Bases pour les règles d'association à partir des Itemsets Fermés fréquents*. Publié en Inforsid 2000 congress, Lyon France, 2000.
[35] Pasquier, N., *Extraction de base pour les règles d'association à partir des itemsets Fermés fréquents*. 2006.
[36] Preux, F., *Fouille de données*. 2011.
[37] Reboul, L., *Classification*. 2003.
[38] Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. and Stewart, B., *Computer Assisted Reading and Discovery for Student Generated Text*. 2014.
[39] Roche, R. , *Fouille de texte : de l'extraction des descripteurs linguistiques à leur induction*. 2011.
[40] Roux, M., *Algorithmes de classification*. 2006.
[41] Tolonne, E., *Analyse syntaxique à l'aide des tables du lexique-grammaire du français*. 2011.
[42] Vandepu, E., *Le traitement du contenu des textes*. 1995.
[43] Weddingen, S., *Le traitement avancé des occurrences : le projet ABEL Analyse du texte fantastique*. 2004.

Webographie

[44] *Analyse des données textuelles* Consulté le 18 dec 2014, http://fr.wikipedia.org/wiki/Analyse_de_donn%C3%A9es_textuelles, 2014.

[45] *L'analyse du discours.* Consulté le 10 sep 2014, http://www.analyse-du-discours.com/l-analyse-du-discours.

[46] *L'analyse du discours.* Consulté le 11 fev 2015, http://fr.wikipedia.org/wiki/Analyse_du_discours, 2014.

[47] *L'analyse syntaxique.* Consulté le 8 dec 2014, http://fr.wikipedia.org/wiki/Analyse_syntaxique, 2014.

[48] *Fouille de texte. Consulté le 22 sep 2015,* http://fr.wikipedia.org/wiki/Fouille_de_textes, 2013.

[49] *http://labh-curien.univ-st-etienne.fr/~jacquemont/Documents/these.pdf.*

[50] *Kobo Arc guide de l'utilisation. Consulté le 11 nov 2014, http://download.kobobooks.com/magento/userguides/downloads/KoboArc/koboarc_user guide_fr.pdf*

[51] *Composition, an intelligent Text Editor For iOS.* http://prmac.com/release-id-59448.htm.

[52] *Réseaux de neurones formels.* Consulté le 5 nov 2014, http://informatique.coursgratuits.net/methodes-numeriques/reseaux-de-neurones-formels.php, 2009.

[53] *Supervised Machine learning with a support vector machine* http://www.neos-guide.org/content/supervised-machine-learning-support-vector-machine#applet, 2012.

[54] *Distance de Hamming. Consulté le 5 jan 2015,* http://en.wikipedia.org/wiki/Hamming_distance.

[55] *J.Chappelier and J.Kolas, Distance de Hamming et poids d'un mot de code.*

[56] *Distance de Jaccard. Consulté le 22 nov 2014,* http://en.wikipedia.org/wiki/Jaccard_distance, 2014.

[57] *Distance euclidienne. Consulté le 7 mar 2015,* http://fr.wikipedia.org/wiki/Espace_euclidien, 2014.

[58] *Distance Manhattan. Consulté le 2 mar 2015,* http://fr.wikipedia.org/wiki/Distance_de_Manhattan, 2014.

[59] *Distance de Tchebychev.* Consulté *le 24 jan 2014,* http://fr.wikipedia.org/wiki/Distance_de_Tchebychev, 2014.

[60] *Fouille de données.* http://www.grappa.univ-lille3.fr/polys/fouille/sortie005.html.

[61] *Adaptative resonance theory (ART).* http://adaptiveresonancetheory.blogspot.ca/.

[62] *Data Mining : Règles d'association ou la problématique du panier de la Ménagère.* http://www.statsoft.fr/concepts-statistiques/regles-d-association/regles-d-association.htm, 2013.

[63] *Règles associatives.* http://georges.gardarin.free.fr/Cours_Total/DM4-Association.ppt.

[64] *Recherche en plein texte.* http://docs.postgresqlfr.org/8.3/textsearch.html.

[65] *Traitement de texte Atlantis- Aide en ligne* http://www.atlantiswordprocessor.com/fr/help/selecting_text.htm.
[66] *Compteur de mots.* http://www.compteurdelettres.com/mots.html.
[67] *Jaccard similarity and k-gram.* http://www.cs.utah.edu/~jeffp/teaching/cs5140/L4-Jaccard+nGram.pdf.
[68] *Représentation des textes.* http://www.neurones.espci.fr/Theses_PS/Stricker_M/CHAP5.pdf.
[69] *Similarité sémantique des mots.* http://theses.ulaval.ca/archimede/fichiers/24972/ch05.html.
[70] *Sensibilité, spécificité, valeurs prédictives et rapports de vraisemblance.* http://www.med.uottawa.ca/sim/data/Sensitivity_f.htm.

Annexe A – Textes utilisés pour l'expérimentation

- **Reconnaissance faciale automatisée dans les secteurs public et privé**

Le Commissariat à la protection de la vie privée du Canada (CPVP) suit l'évolution de la technologie de reconnaissance faciale depuis de nombreuses années dans le cadre de son intérêt pour la biométrie en général. Il y a près de dix ans, nous avons déterminé que la reconnaissance faciale pourrait devenir la plus envahissante des technologies d'identification biométrique populaires modernes, car le sujet n'a pas à donner son consentement ou même à participer sciemment.

La reconnaissance faciale automatisée consiste à identifier un individu à partir de la géométrie de son visage. Pour que la technologie soit efficace, il faut disposer d'une image numérique de qualité du visage de l'individu en question, d'une base de données d'images numériques d'individus identifiés et d'un logiciel de reconnaissance faciale capable d'établir une correspondance exacte entre l'image d'un individu et une image d'un individu identifié qui est enregistrée dans la base de données.

Parmi toutes les technologies biométriques, la reconnaissance faciale est celle qui imite le plus la façon dont les gens s'y prennent pour identifier les autres, c'est-à-dire en examinant leur visage. Il est extrêmement difficile et coûteux de doter une machine de cette aptitude qui ne nécessite aucun effort de la part des humains. Cela dit, grâce à une convergence de facteurs au cours des dernières années, la reconnaissance faciale est devenue une technologie viable et de plus en plus exacte.

Les images numériques sont désormais omniprésentes en raison de la prolifération des caméras de surveillance, des téléphones intelligents équipés d'un appareil photo et des appareils photos numériques de qualité bon marché. Les dispositifs de stockage à prix modique ont donné lieu à la création de vastes bases de données en ligne renfermant des images d'individus identifiés, par exemple les titulaires de permis de conduire ou de passeport, les personnes possédant une carte d'identité d'employé et celles ayant un casier judiciaire. Les individus ont adopté l'affichage et l'étiquetage des photos en ligne sur des plateformes comme Facebook, Instagram, Picasa et Flickr. En outre, la technologie de reconnaissance faciale a fait l'objet de perfectionnements considérables, notamment au chapitre de l'analyse des images et de l'extraction des données.

Les visages ont été transformés en données électroniques qu'il est désormais possible de regrouper, d'analyser et de classer de façons inédites. Les données d'images du visage sont d'autant plus précieuses et sensibles qu'il s'agit d'une caractéristique de notre corps mesurable de façon unique et d'un élément clé de notre identité.

Certaines applications de cette technologie à des fins de sécurité sont incontestablement bénéfiques, par exemple l'authentification des employés autorisés à avoir accès à une centrale nucléaire. La reconnaissance faciale a des répercussions telles au chapitre de la protection de la vie privée et des valeurs de la société en général que certains observateurs1 estiment que cette technologie pourrait sonner le glas de l'anonymat.

Le présent rapport de recherche a pour ambition d'expliquer en termes simples le mode de fonctionnement de la technologie de reconnaissance faciale, d'examiner certaines applications de cette technologie dans les secteurs public et privé, et d'analyser ses répercussions sur la protection de la vie privée.

La technologie de reconnaissance faciale vise à identifier des individus ou à authentifier leur identité en comparant leur visage avec des visages connus stockés dans une base de données pour trouver une correspondance. Le procédé comprend trois grandes étapes. Premièrement, l'ordinateur trouve le visage dans l'image. Il crée ensuite une représentation numérique du visage d'après la position relative, la taille et la forme des traits faciaux. Enfin, cette « carte » numérique du visage représenté sur l'image est comparée avec les images de visages identifiés qui sont enregistrées dans la base de données, par exemple celle des titulaires de permis de conduire.

On peut avoir recours à la reconnaissance faciale pour confirmer ou découvrir l'identité d'une personne. Des systèmes d'authentification sont utilisés pour contrôler l'accès à des installations ou à des équipements. Au nombre des autres usages, mentionnons la lutte contre la fraude, par exemple pour vérifier si un individu a présenté des demande de passeport sous différents noms. D'autres types de technologies biométriques servent à l'heure actuelle aux fins d'authentification, par exemple la lecture des empreintes digitales et le balayage de l'iris.

L'identification est souvent la finalité des applications de sécurité publique et nationale. Il peut s'agir, par exemple, d'identifier des individus au cours d'une émeute ou d'assurer la sécurité dans des lieux publics à fort achalandage, par exemple un aéroport ou un centre sportif. La reconnaissance faciale convient très bien pour les applications d'identification, car les images du visage peuvent être captées à distance et à l'insu de l'individu. On peut aussi avoir recours à d'autres technologies biométriques, comme la reconnaissance de la démarche ou de la voix, pour identifier les individus à distance et sans leur consentement, mais elles comportent des limites évidentes qui les rendent moins utiles.

b) Exactitude

Un certain nombre de facteurs influent sur l'exactitude de la technologie de reconnaissance faciale :

le système peut reconnaître uniquement les individus dont l'image est enregistrée dans la base de données;

les images doivent être de qualité suffisante pour assurer la fiabilité;

le seuil de sensibilité du système doit être réglé de manière à éviter un nombre excessif de faux positifs (erreur sur la personne identifiée) ou de faux négatifs (non détection d'une personne qui aurait dû être identifiée);

l'éclairage, le port de lunettes, une moustache ou une barbe, le maquillage et l'angle sous lequel les photos sont prises.

L'utilisation d'images 3D, qui permettent de capter l'information sur la morphologie du crâne du sujet, représente une innovation récente dans le domaine de la reconnaissance faciale. Elle rend le système moins vulnérable aux problèmes d'éclairage et permet d'établir une correspondance entre des images prises sous des angles différents.

En 2010, après avoir mis à l'essai2 divers systèmes de reconnaissance faciale, le National Institute of Standards and Technology des États-Unis a constaté que le meilleur algorithme permettait de reconnaître avec exactitude 92 % des inconnus au moyen d'une base de données de 1,6 million de dossiers criminels.

Selon une étude menée en 20113 à l'Université Carnegie Mellon, la technologie de reconnaissance faciale pourrait être utilisée pour identifier des individus dans le monde réel à partir d'images personnelles en ligne. Les chercheurs ont été en mesure d'identifier des

inconnus et de trouver leurs renseignements personnels au moyen d'un logiciel de reconnaissance faciale et des profils affichés sur les médias sociaux.

Dans le cadre d'une expérience, des chercheurs ont utilisé des images accessibles au public apparaissant dans des profils affichés en ligne sur des sites de réseaux sociaux pour identifier des individus dont la photographie apparaissait sur un site de rencontre en ligne populaire où les membres utilisent un pseudonyme. Ils ont réussi à identifier un membre sur dix.

Dans une deuxième expérience, les chercheurs ont réussi à identifier 31 % des étudiants qui marchaient sur le campus en utilisant les photos de leur profil Facebook.

Dans une troisième expérience, les chercheurs ont prédit les intérêts personnels des individus et, dans 27 % des cas, les cinq premiers chiffres de leur numéro d'assurance sociale à partir d'une photo de leur visage.

L'étude a montré qu'il est possible d'établir un lien avec l'identité en ligne et hors ligne d'un individu à partir de son visage sans avoir accès à une base de données spéciale. Les chercheurs considèrent que la reconnaissance faciale de tous les individus en tout lieu et en tout temps n'est pas réaliste pour l'heure en raison des contraintes technologiques (exactitude) et des coûts du traitement informatique des données, mais ils estiment que ces contraintes disparaîtront au fil du temps.

- **Droit fiscal**

Les autorités fiscales, tant au niveau provincial que fédéral, disposent d'un certain nombre de pouvoirs et de recours afin d'astreindre un contribuable à collaborer dans le

cadre d'un processus de vérification et d'enquête. Parmi les plus utilisés, nous retrouvons la demande péremptoire, qui permet aux autorités d'exiger qu'une personne produise un document ou renseignement dans un délai raisonnable afin de leur permettre d'appliquer et d'exécuter toute loi fiscale.

Le défaut d'obtempérer à une demande péremptoire peut engendrer des conséquences dommageables importantes pour la personne en défaut. D'abord, le fait de ne pas fournir une information ou de ne pas produire un document dans le délai imparti constitue une infraction pénale, en vertu de laquelle la personne visée peut encourir non seulement une amende mais également une période d'emprisonnement.

En plus de cette infraction pénale, la Loi sur le Ministère du revenu (« LMR »)1 prévoit que la personne qui ne s'est pas conformée à une demande péremptoire peut se faire refuser, par le tribunal, le droit de déposer ce renseignement ou document en preuve dans le cadre d'un litige l'opposant au Sous-ministre du Revenu du Québec.

En dernier lieu, l'article 39.2 LMR permet au ministre de faire une demande auprès d'un juge de la Cour du Québec afin que celui-ci ordonne à la personne visée par une demande péremptoire de fournir les documents et renseignements sollicités.

C'est dans le contexte de ce dernier recours que la Cour d'appel a examiné l'étendue des pouvoirs du ministre du Revenu du Québec relativement aux demandes péremptoires dans la cause Sous-ministre du revenu du Québec c. 6217125 Canada Inc 2.

Dans cette affaire, le Ministère du revenu du Québec (« MRQ ») détient une créance résultant d'un jugement contre 9037-1832 Québec Inc. (« Québec Inc. »), une société exploitant un complexe immobilier. Avant que le MRQ puisse percevoir sa créance,

Québec Inc. vend le complexe immobilier à l'intimée, 6217125 Canada Inc. (« Canada Inc. »).

À la suite de ce refus, le MRQ dépose une requête devant la Cour du Québec pour ordonnance de production en vertu de l'article 39.2 LMR. La Cour du Québec interprète d'abord les articles régissant la demande péremptoire et arrive à la conclusion que ceux-ci doivent recevoir une interprétation stricte, vu l'importance des pouvoirs en jeu. Par conséquent, le recours peut être exercé uniquement aux fins énoncées explicitement dans la loi. Or, en l'espèce, la Cour remarque que la demande péremptoire du MRQ a pour objectif d'étayer sa preuve pour les fins du recours civil, et non pas pour recouvrir un montant recevable en vertu d'une loi fiscale, ni pour appliquer une loi fiscale, tel que requis par les dispositions pertinentes de la LMR. Étendre l'application de la demande péremptoire à tout litige civil dans lequel le MRQ est impliqué accorderait à celui-ci des pouvoirs trop étendus. Ainsi, la Cour conclut que la demande péremptoire est mal fondée et rejette la requête pour ordonnance de production3.

Le jugement du tribunal formule quelques observations. D'abord, la Cour précise qu'il n'y a aucun contentieux fiscal opposant les deux parties et que l'intimée ne fait l'objet d'aucune vérification ou enquête. Elle explique ensuite que l'équité procédurale permet au MRQ d'invoquer les droits procéduraux prévus au Code de procédure civile afin d'obtenir la production des documents et renseignements recherchés. De plus, elle rappelle que les 3278 documents visés par la requête sont publiés et sont donc accessibles. Finalement, la Cour ne reconnaît aucun principe de droit qui permettrait au MRQ de faire constituer son dossier civil par la partie adverse. Saisie du pourvoi du sous-ministre, la Cour d'appel du Québec rejette l'appel mais pour des motifs distincts de ceux établis par le tribunal de

première instance. L'analyse de la Cour débute par une revue des conditions d'application établies par la Cour suprême du Canada relatives à toute demande péremptoire. Parmi celles-ci, la Cour mentionne qu'il n'est pas nécessaire que la personne faisant l'objet de la demande soit celle dont l'assujettissement à l'impôt fait l'objet d'une enquête.

Ainsi, la Cour reconnaît le pouvoir du MRQ d'obtenir d'un tiers des documents relatifs à une vérification sur un autre contribuable, de même que la possibilité qu'un tel recours soit exercé lors de procédures en exécution.

- **Une seule intelligence**

Fabriquer de l'intelligence est un défi que l'informatique veut relever. Quand elle réussit, c'est toujours de manière limitée et en évitant d'aborder de front l'intelligence humaine, qui reste mystérieuse.

L'idée qu'il existe plusieurs types d'intelligence séduit le grand public, car elle évite à chacun de se trouver en un point précis d'une échelle absolue et parce que chacun espère bien exceller dans l'une des formes d'intelligence dont la liste tend à s'allonger. Cette pluralité d'intelligences a été proposée par le psychologue américain Howard Gardner : dans son livre Frame of Mind de 1983, il énumère huit types d'intelligence. Très critiquée, par exemple par Perry Klein de l'Université d'Ontario qui la considère tautologique et non réfutable, cette théorie est à l'opposé d'une autre voie de recherche affirmant qu'il n'existe qu'une sorte d'intelligence à concevoir mathématiquement avec l'aide de l'informatique et de la théorie du calcul.

Dames, échecs, go

Évoquons d'abord l'intelligence des machines et la discipline informatique dénommée « intelligence artificielle ». Il faut l'admettre, aujourd'hui, les machines réussissent des prouesses qu'autrefois tout le monde aurait qualifié d'intelligentes. Nous ne reviendrons pas sur la victoire définitive de l'ordinateur sur les meilleurs joueurs d'échecs, consacrée en 1997 par la défaite de Garry Kasparov (champion du monde) face à l'ordinateur Deep Blue, unanimement saluée comme un événement majeur de l'histoire de l'humanité.

À cette époque, pour se consoler peut-être, certains ont remarqué que les meilleurs programmes pour jouer au jeu de go étaient d'une affligeante médiocrité. Or, depuis quelques années, des progrès spectaculaires ont été réalisés et aujourd'hui les machines, sans égaler les champions et les professionnels, ont atteint le niveau des très bons joueurs amateurs. En mars 2013, le programme Crazy Stone du chercheur français Rémi Coulom, de l'Université de Lille, a battu le joueur professionnel japonais Yoshio Ishida qui, au début de la partie, avait laissé un avantage de quatre pierres au programme. Le programme est considéré avoir un niveau de « sixième dan » (classement KGS) : il y a environ une dizaine de joueurs français à ce niveau, et moins de 500 dans le monde.

Le succès de l'intelligence artificielle au jeu de dames anglaises est absolu. Depuis 1994, aucun humain n'a battu le programme canadien Chinook et, depuis 2007, on sait que le programme joue une stratégie optimale, impossible à améliorer. Pour le jeu d'échecs, on sait qu'il existe aussi des stratégies optimales, mais leur calcul semble hors d'atteinte pour plusieurs décennies encore.

L'intelligence des machines ne se limite plus aux problèmes bien clairs de nature mathématique ou se ramenant à l'exploration d'un grand nombre de combinaisons. Cependant, les chercheurs en intelligence artificielle ont découvert, même avec les jeux de

plateau cités, combien il est difficile d'imiter le fonctionnement intellectuel humain : aux jeux de dames, d'échecs ou de go et bien d'autres, les programmes ont des capacités équivalentes aux meilleurs humains, mais ils fonctionnent très différemment. Cela ne doit pas nous interdire d'affirmer que nous avons mis un peu d'intelligence dans les machines : ce ne serait pas *fair-play,* face à une tâche donnée, d'obliger les machines à l'affronter en imitant servilement nos méthodes et modes de raisonnement.

Véhicules intelligents

Le cas des véhicules autonomes est remarquable aussi de ce point de vue. Il illustre d'une autre façon que lorsque l'on conçoit des systèmes nous imitant à peu près pour les résultats, on le fait en utilisant des techniques le plus souvent totalement étrangères à celles mises en œuvre en nous par la nature, et que d'ailleurs nous ne comprenons que très partiellement : pour le jeu d'échecs par exemple, personne ne sait décrire les algorithmes qui déterminent le jeu des champions.

La conduite de véhicules motorisés demande aux êtres humains des capacités qui vont bien au-delà de la simple mémorisation d'une quantité massive d'informations et de l'exploitation d'algorithmes traitant rapidement et...

- **Les spidrons, pliables à l'infini**

Il est étonnant que l'on invente encore de nouvelles formes géométriques simples. Pourtant, les Spidrons conçus par le designer hongrois Dániel Erdély sont faciles à construire et s'emboîtent miraculeusement.

Au cours des années 1970, Dániel Erdély, un artiste hongrois spécialisé dans le design industriel, a imaginé, perfectionné et étudié une catégorie nouvelle de formes géométriques.

Celles-ci ont depuis suscité une multitude de travaux et de créations artistiques où se mêlent de délicates considérations de géométrie et la volonté de créer de belles figures et de beaux objets.

D. Erdély les a nommés Spidrons, terme qui résulte d'un mélange des mots « spider » (araignée en anglais) et spirale. La figure d'où est issue cette forme est en effet un hexagone où une infinité de tracés ont été dessinés, le tout ressemblant à une toile d'araignée. Soigneusement colorié, on y voit un crochet, décomposé à l'infini, répété six fois et ressemblant à la queue d'un hippocampe : le demi-Spidron.

Cette structure est composée d'une succession illimitée de triangles de deux types : le triangle équilatéral et le triangle isocèle d'angles égaux à 120°, 30° et 30°. Ces triangles, reproduits selon des tailles décroissantes, sont accolés les uns aux autres de manière à ce que leur jonction fasse coïncider deux côtés exactement. Le crochet alterne les deux types de triangles qui s'enroulent en une spirale logarithmique. Accolée à elle-même, cette spirale donne le Spidron à deux branches, ou Spidron complet.

La première propriété remarquable du Spidron complet est qu'il pave le plan parfaitement, ce qui donne de jolis tableaux.

Comme c'est le cas pour de nombreuses molécules du monde vivant, on remarque qu'il existe deux versions du double crochet, selon que l'on fait tourner ses spirales dans un sens ou dans un autre : on passe de l'une à l'autre en regardant à travers une glace.

D. Erdély remarqua que si l'on considère un triangle équilatéral E dans l'une des spirales, son aire est égale à la somme des aires des triangles qui le suivent dans

l'enroulement. Autrement dit, tous les triangles succédant à un triangle équilatéral peuvent s'entasser en lui (pour réaliser ce remplissage, un découpage adéquat est nécessaire).

Voici le petit raisonnement qui établit la propriété des aires :

– Le triangle isocèle I d'angles 120°, 30°, et 30° accolé à un triangle équilatéral E a une aire égale au tiers de l'aire de E, car le triangle I s'obtient par un découpage en trois de E.

– Ce triangle isocèle I a la même aire que le triangle équilatéral E' suivant de la spirale décroissante. En effet, en le coupant en deux par la hauteur issue du sommet d'angle 120°, on obtient deux triangles qui, accolés, reconstituent E'.

– Il en résulte que d'un triangle équilatéral E au suivant, E', l'aire est divisée par 3 (le côté est donc divisé par 3).

– La somme des aires des triangles qui suivent un triangle équilatéral (supposé d'aire unité) est donc 2/3 + 2/9 + 2/27 + 2/81+ ..., qui vaut 1. En effet, si l'on note s cette somme infinie, elle vérifie s – s /3 = 2/3 (tous les termes se simplifient, sauf le premier), ce qui donne 2s /3 = 2/3, donc s =.

- **La science**

La science (latin scientia, « connaissance ») est « ce que l'on sait pour l'avoir appris, ce que l'on tient pour vrai au sens large, l'ensemble de connaissances, d'études d'une valeur universelle, caractérisées par un objet (domaine) et une méthode déterminés, et fondés sur des relations objectives vérifiables [sens restreint] »1.

La volonté de la communauté savante, garante des sciences, est de produire des « connaissances scientifiques » à partir de méthodes d'investigation rigoureuses, vérifiables et reproductibles. Quant aux « méthodes scientifiques » et aux « valeurs scientifiques »,

elles sont à la fois le produit et l'outil de production de ces connaissances et se caractérisent par leur but, qui consiste à permettre de comprendre et d'expliquer le monde et ses phénomènes de la manière la plus élémentaire possible — c'est-à-dire de produire des connaissances se rapprochant le plus possible des faits observables. À la différence des dogmes, qui prétendent également dire le vrai, la science est ouverte à la critique et les connaissances scientifiques, ainsi que les méthodes, sont toujours ouvertes à la révision. De plus, les sciences ont pour but de comprendre les phénomènes, et d'en tirer des prévisions justes et des applications fonctionnelles ; leurs résultats sont sans cesse confrontés à la réalité. Ces connaissances sont à la base de nombreux développements techniques ayant de forts impacts sur la société.

La science est historiquement liée à la philosophie. Dominique Lecourt écrit ainsi qu'il existe « un lien constitutif [unissant] aux sciences ce mode particulier de penser qu'est la philosophie. C'est bien en effet parce que quelques penseurs en Ionie dès le VIIe siècle av. J.-C. eurent l'idée que l'on pouvait expliquer les phénomènes naturels par des causes naturelles qu'ont été produites les premières connaissances scientifiques »2. Dominique Lecourt explique ainsi que les premiers philosophes ont été amenés à faire de la science (sans que les deux soient confondues).

La science se compose d'un ensemble de disciplines particulières dont chacune porte sur un domaine particulier du savoir scientifique. Il s'agit par exemple des mathématiques3, de la chimie, de la physique, de la biologie, de la mécanique, de l'optique, de la pharmacie, de l'astronomie, de l'archéologie, de l'économie, de la sociologie, etc. Cette catégorisation n'est ni fixe, ni unique, et les disciplines scientifiques peuvent elles-mêmes être découpées

en sous-disciplines, également de manière plus ou moins conventionnelle. Chacune de ces disciplines constitue une science particulière.

- **L'image thermique une technologie prête à conquérir le reste du monde**

De nombreuses technologies qui nous semblent aujourd'hui usuelles étaient au départ destinées aux applications militaires. Le radar et les moteurs à réaction ont été mis au point par l'armée et pour elle-même. Le GPS (Global Positioning System) a été créé au début des années 1970 par le ministère étasunien de la Défense pour guider les missiles. Lorsqu'il a été intégré dans des voitures pour son utilisation civile, le GPS était encore très onéreux. Aujourd'hui, presque toutes les nouvelles voitures en sont équipées.

À l'origine de toutes les inventions militaires, il y avait ce que le Pentagone appelait le « réseau intergalactique ». Cette idée révolutionnaire est devenue Internet. C'est sans aucun doute l'invention militaire qui a changé le plus radicalement notre monde. Aujourd'hui, il existe une autre technologie d'origine militaire qui fait son chemin vers de nombreuses applications civiles utiles. Elle va non seulement changer notre vie, mais sauver des vies. C'est l'imagerie thermique.

Alors qu'une caméra normale a besoin de lumière pour produire une image, une caméra thermique (ou infrarouge) peut capter de très faibles différences de température et les convertir en une excellente image thermique sur laquelle les plus petits détails sont visibles. Contrairement à d'autres technologies, comme l'amplification de lumière qui nécessite une petite quantité de lumière pour produire une image, l'imagerie thermique permet de voir dans l'obscurité totale. Elle ne nécessite aucune source de lumière.

La première caméra thermique militaire a été mise au point en Suède en 1958 par une société nommée AGA, devenue aujourd'hui FLIR Systems. Depuis qu'il est possible de produire une image lisible dans l'obscurité totale, la technologie de l'imagerie thermique permet de voir et de cibler les forces ennemies dans la nuit la plus noire. Les caméras thermiques voient à travers la brume, la pluie et la neige. Elles voient aussi à travers la fumée, ce qui était particulièrement intéressant pour l'armée dans le cas d'opérations sur un champ de bataille enfumé.

La première caméra thermique pour applications civiles a été mise au point en 1965. Elle était utilisée pour l'inspection des lignes électriques. Il a fallu attendre 1973 pour voir la première caméra infrarouge « portable » sur batterie.

Bien que qualifiée de portable, elle était très encombrante. La technologie utilisée à cette époque nécessitait de remplir la caméra d'azote liquide pour refroidir son détecteur infrarouge. Cela a duré jusqu'en 1985, année où FLIR Systems a présenté le premier système sans azote liquide. Le détecteur était refroidi par un système cryogénique. C'est en 1997 qu'est apparue sur le marché la première caméra thermique avec détecteur non refroidi, appelé microbolomètre. L'un des principaux avantages du microbolomètre est l'absence de pièce mobile. Il est donc moins exposé aux pannes. Il est également moins coûteux à produire, ce qui a permis aux fabricants de caméras thermiques de baisser leurs prix.

Les caméras thermiques ont toujours été utilisées par l'armée. Il a fallu attendre l'invention du microbolomètre pour qu'elles commencent à être commercialisées, et utilisées par l'industrie.

Les premiers clients industriels à en bénéficier furent les grandes sociétés de production. Non seulement les caméras thermiques produisent une image basée sur les différences de température, mais ces différences peuvent être mesurées. Grâce à des algorithmes complexes intégrés à la caméra thermique, il est possible de calculer les valeurs de température absolue.

L'industrie a rapidement découvert que l'imagerie thermique peut fournir des informations précieuses sur l'équipement électrique. Les fusibles, les connexions, les câbles, le matériel haute tension comme les transformateurs, les lignes et de nombreux autres équipements peuvent tous être inspectés facilement et sans contact au moyen d'une caméra thermique. L'avantage est qu'elle permet aux responsables de la maintenance de voir les anomalies avant qu'elles provoquent un problème. Des pannes coûteuses sont ainsi évitées, ce qui fait gagner du temps et de l'argent.

Dans ces sociétés, les départements Recherche et Développement se sont enthousiasmés pour l'imagerie thermique. C'est au début du cycle de développement d'un produit que son utilisation est la plus avantageuse. Lors de la mise au point, avant la production en série, les appareils sont soumis à des essais complets, car les consommateurs attendent un produit parfait à un prix abordable. L'imagerie thermique permet aux entreprises de raccourcir les phases de mise au point et de bénéficier d'un retour rapide sur leur investissement en développement.

Grâce au nombre croissant de sociétés industrielles adoptant cette technologie, il est devenu possible d'envisager une production en série. Mais même à ce moment-là, les caméras thermiques étaient encore un outil très onéreux, d'au moins 20 000 €. L'imagerie thermique restait une technologie marginale, peu connue des consommateurs. Les seules

images thermiques qu'ils étaient amenés à voir étaient celles utilisées dans les films américains comme Predator I et II.

- **عالم الحاسوب**

كما أيسلفنا فان فكرة ا لحاسب قديمه قدم الإنسان فقد كان أول وسيله حساب استخدمه الإنسان هو أصابع يديه ثم الحصى ثم مع تطور حياته و تعقدها كان غاية أن يطور أساليبه وتقنياته لتلائم متطلبات العصر الذي يعيش فيه وإزاء هذا التسارع المذهل في حياة البشر كان لابد أن يقابلة تسارع مواز في تطور فكرة الحاسوب والاتجاه إلى إبرازه كواحد من أهم العوامل التي تساعد على تقدم ورقى البشرية 0

والتي تستخدم خرزات على أسلاك (abacus) وفى أولى محاولات الإنسان إلى مكنه الحاسوب باستخدام اله العداد 0وقد ساعده العداد في إجراء بعض العمليات الحسابية البسيطة وكن ذلك قبل اكثر من 2000سنه قبل الميلاد ومرت السنوات وبدأت حياه الحاسوب تقاس بالأجيال لا بالسنوات وكما هي السنوات ما هي لحظات في حياة الأمم تعيشها أياما وساعات وتسجلها في سطور التاريخ لحظات . وفي العام 1642 م اختر ع عالم فرنسي يدعى باسكال (وقد سميت لغة البرمجة باسكال باسمه تقديرا وتخليدا لجهوده في هذا المجال) اله ميكانيكية تستطيع إجراء عمليات حسابيه بسيطة في الجمع والطرح وأتم هذه الجهود ليبينز بعد حوالي ثلاثين عاما باختراع آلته والتي سميت بآلة ليبينز وهي ميكانيكي العمل أيضا وتستطيع إجراء عمليات القسمة و الضرب وفي باية القرن التاسع عشر وتحديدا في العام 1804م قام عالم فرنسي يدعى جوزيف كاكوارد باختراع اله تستخدم في عملها البطاقات المثقبة وقد بدا مع اختراع هذه الآلة نشوء فكره البرمجة باستخدام الحاسوب وقد قام بعده تشارلزبابادج بتطوير اله تستطيع استقبال الأوامر عن طريق البطاقات المثقبة0

في منصف الأربعينات وفي حوالي 1945 م قام نيومان بتطوير عمل الحاسوب حيث اصبح الحاسوب يقوم بالتخزين الداخلي للبيانات واستخدام النظام الثنائي كقاعدة لبناء الحاسوب حيث إن النظام الثنائي في العد يشابه أحد حالتي التيار الكهربائي تشغيل إيقاف وهذا يذكرنا بتعريف العمليات المنطقية أي إن الحاسوب قائم في عمله وبنيته على المنطق ويعتبر النظام الثنائي أساس لغة الآلة وهي اللغة التي يفهمها الحاسوب ومنذ ذلك الحين بدا الظهور الفعلي للحاسوب وبدأت أهمية كجزء مهم في حياة البشر وضرورة من حياتهم وتقدمها فعملوا على تطويره وتحديثه ليلائم التسارع الحياتي الذي يعيشون وبدأت أجيال الحاسوب باضهور0

حاسبات الجيل الأول تطلق هذه التسمية حاسبات الجيل الأول على الفترة من حياة وتطور الحاسوب من عام 1945 م إلى العام 1951م 0في هذه الفترة تم استخدام الصمامات المفرغة في صنع حاسبات هذا الجيل استخدامات هذا الجيل لغة الآلة أي لغة الصغر والواحد للتعامل مع الجهاز مما سبب صعوبة في التعامل مع الحاسوب واحتياج الحاسوب إلى إنسان متخصص للتعامل معه كما أن الحاسبات في هذا الجيل كانت كبيرة الحجم وبطيئة نسبيا إضافة إلى أنها تحوي ذاكرة محدودة جدا بالإضافة إلى إنها تولد حرارة عالية جدا