République Algérien Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahleb Blida

Faculté des Sciences

Département de l'Informatique

Mémoire Présenté par :

Tiouguiouine Yousra

Samar Mohamed Bilel

En vue de l'obtention du Diplôme de Master

Domaine: Mathématique et Informatique

Filière : Informatique

Spécialité : Traitement Automatique de Langue

Sujet :

Vers l'élaboration d'une Approche Contextuelle Pour l'Indexation Automatique Des Textes Arabes Non Structurés

Soutenue le : 07 décembre 2020, devant le jury composé de :

Mme M. MEZZI	President
Mme H.Ykhlef	Examinatrice
Mme Droua-Hamdani Ghania	Encadreur
Mme Benbldia Nadjia	Promotrice

Année universitaire 2019/2020

ملخص

ضمن مواجهة الزيادة الهائلة في حجم الوثائق العربية على الإنترنت والبريد الإلكتروني والمكتبات الرقمية، يمثل هذا التغيير الثوري تحديات كبيرة، ويصبح الوصول إلى البيانات بطريقة دقيقة وسريعة أمرًا صعبًا للغاية. حيث يعد التصنيف اليدوي للنصوص في هذه الحالة أمرًا صعبًا للغاية. لذلك من الضروري تطوير برامج التصنيف التلقائي لمساعدة المستخدمين على البحث عن المعلومات بطريقة أكثر كفاءة أو حتى استهدافًا أفضل.

في هذا المشروع، نقدم تطوير نهج سياقي للفهرسة التلقائية للنصوص العربية غير المهيكلة باستخدام ترجيح (TF-IDF)

نحن نستخدم مجموعة مستندات من ثلاث فئات ثم نصنف الوثائق إلى هذه الثلاث ونرتبها وفقًا للوثيقة التي نسعى إليها. يتم تقييم كل ذلك باستخدام نماذج "SVM" "KNN" التي تعطينا أفضل النتائج.

الكلمات الدالة:

المعالجة الآلية للغة العربية، الفهرسة الآلية، التعلم الآلي الخاضع للإشراف، نموذج "KNN"، نموذج "SVM".

Résumé

Devant l'augmentation énorme du volume de documents arabe sur Internet, e-mail, les bibliothèques numériques, Ce changement révolutionnaire présente de grands défis, l'accès aux données d'une façon précise et rapide devient très difficile. La catégorisation manuelle des textes dans ce cas est très difficile même s'il est possible elle infect l'efficacité, la rapidité et le coût. Pour cela il est donc nécessaire de développer des programmes de catégorisation automatique pour aider les utilisateurs à rechercher de l'information d'une manière plus efficace voire mieux ciblée.
Dans ce projet, nous présentons le développement d'une approche contextuelle pour l'indexation automatique de textes arabes non structurés en utilisant la pondération TF-IDF.

Nous utilisons un corpus de trois catégories puis catégorisent les documents dans ces trois et les ordonnent en fonction du document recherché. Tout cela est évalué en utilisant les modèles "KNN" "SVM" qui nous donne les meilleurs résultats.

Mots clés:

TALA : traitement automatique de la langue arabe, indexation automatique, apprentissage automatique supervisé, le modèle "KNN", le modèle "SVM".

Abstract:

Faced with the huge increase in the volume of Arabic documents on the Internet, e-mail, digital libraries, this revolutionary change presents great challenges, access to data in a precise and fast way becomes very difficult. Manual categorization of texts in this case is very difficult even if it is possible it infects efficiency, speed and cost. For this, it is necessary to develop automatic categorization programs to help users search for information in a more efficient.

In this project, we present the development of a contextual approach for the automatic indexing of unstructured Arabic texts using the TF-IDF weighting.

We use a corpus of three categories then categorises the documents into these three & order it according to the document we seek. All this is evaluated by using the "KNN" "SVM" models which gives us the best results.

Keywords:

Natural Language Processing, automatic indexing, supervised machine learning, the "KNN" model, the "SVM" model.

Remerciements

En premier lieu et avant tout nous remercions **DIEU « ALLAH »** le tout puissant de nous avoir donné le courage, la patience et la force de réaliser ce projet de fin d'études.

Nous tenons tout d'abord à exprimer nos profonds remerciements à notre encadreur **Mme Droua-Hamdani Ghania** pour avoir encadré et dirigé ce travail avec une grande rigueur scientifique, sa disponibilité, ses conseils et la confiance qu'elle nous accordait nous a permet de réaliser ce travail.

Dans un deuxième temps, Nous remercions notre promotrice de thèse Mme **Benbldia** qui a bien voulu nous encadrer et nous aider par ces conseils pendant ce projet.

Nous remercions tous nos collègues et amis, pour les conseils, les services et plus particulièrement pour l'amitié qu'ils nous ont témoignée. Nous vous souhaitons à tous bonheur, réussite et tout le bien que vous méritez.

En terminant, nous souhaitons démontrer notre grande gratitude à toutes les personnes ayant participé de près ou de loin et plus particulièrement à nos familles à la réalisation de ce projet.

Dédicaces

À ma très chère mère

Quoi que je fasse ou que je dise, je ne saurai point te remercier comme il se doit.

À mon très cher père

Tu as toujours été à mes côtés pour me soutenir et m'encourager.

Que ce travail traduit ma gratitude et mon affection.

À mon très cher frère et ma sœur

À tous mes amis surtout Amel, Amina et Selma

À mon binôme Bilel

À toute ma famille

Puisse Dieu vous donne santé, bonheur, courage et surtout réussite

Yousra

Dédicaces

Je dédie le fruit de ce modeste travail comme un geste de gratitude

A mes chers parents

Pour leur sacrifice, leur amour, leur soutien et encouragements tout au long de mes études

A mes très chères sœurs

A mes amis

A mon binôme Yousra

A tous les étudiants du Master II TAL

A tous les professeurs que ce soit du primaire, du moyen, du secondaire ou de l'enseignement supérieur

BILEL

Table des matières

Table of Contents

Liste des figures

Liste des tableaux

Liste des Équations

Liste des abréviations :

TAL : traitement automatique de la langue

TALA : traitement automatique de la langue arabe

RI : recherche information

SRI : system de recherche information

IA : intelligence artificielle

SVM : machine à vecteur du support (Mupport Vector Machine)

KNN : les k plus proches voisines (k Nearest Neighbors)

TF : la Fréquence d'apparition d'un terme dans un document (term frequency)

IDF : fréquence inverse des documents (inverse document frequency)

RSV : Retrieval Status Value

PDF : Portable Document Forma

Introduction Générale

Contexte globale

Le monde connait une avance technologique considérable dans tous les secteurs et cela grâce à l'informatique qui est une science qui étudie les techniques du traitement automatique de l'information. Elle joue un rôle important dans la société d'information d'aujourd'hui.

Un système de recherche d'information doit représenter, stocker et organiser l'information, puis fournir à l'utilisateur les éléments correspondant au besoin d'information exprimé par sa requête il à un objectif de fournir à un utilisateur un accès facile à l'information qui l'intéresse.

Notre projet est un projet du domaine de traitement automatique de la langue arabe « TALA », il s'agit de l'élaboration d'une approche contextuelle pour l'indexation automatique des textes arabes non structurés utilisant la pondération TF-IDF.

Avant de commencer les calculs dont nous avons besoin, les textes arabes passent par quelques étapes qui sont la tekenisation, l'élimination des mots vides, la normalisation et la dérivation.

Et pour évaluer l'approche, des tests seront effectués en utilisant des modèles K –NN (K plus proches voisins k-Nearest Neighbors) KNN et Support Vecteurs Machine (SVM).

Problématique et objectifs

L'indexation des textes est une étape fondamentale dans le traitement de texte. Elle a pour objectif d'associer ou de marquer des documents avec des informations ou caractéristiques les plus pertinentes afin de pouvoir par la suite les rechercher et les récupérer. Les solutions de gestion des connaissances sont variées, tandis que certaines indexent uniquement le nom de fichier d'autres indexent le contenu complet ainsi que le méta data relatif. Plusieurs approches de recherche d'information sont utilisées à cet

effet. Cependant, extraire des connaissances à partir de données textuelles est un problème important, en particulier lorsque la masse de données est importante.

La plupart des travaux de recherche effectués dans ce domaine ont été menés sur des langues européennes (surtout en anglais) et asiatiques (japonais et chinois), très peu de travaux ont été réalisés sur le plan des langues qui sont morphologiquement riches (comme l'arabe).

L'objectif principal de notre travail est de proposer un système de recherche d'information local qui traitre les documents arabes pour automatiser et faciliter le processus de recherche.

Organisation du mémoire

Ce mémoire, est constitué de cinq chapitres. Dans le chapitre 1 nous présentons les notions de base. Il se focalise particulièrement sur des SRI. Il présente l'architecture générale d'un SRI telle qu'elle est admise actuellement ainsi qu'un aperçu sur les principaux modèles de recherche existant dans la littérature, il aborde aussi les différentes mesures d'évaluation de pertinence. En deuxième partie, dans le chapitre 2, nous présentons les différentes étapes de prétraitement et la richesse et l'ambiguïté de la morphologie Arabe.

Dans le chapitre 3 nous exposons quelques méthodes d'apprentissage automatique. Ces méthodes se divisent en deux catégories. La première catégorie concerne les méthodes supervisées. La deuxième catégorie concerne les méthodes non supervisées.

Le chapitre 4 présentera la conception du notre projet et la méthodologie suivie pour la réalisation de système.

Dans le chapitre 5, nous exposerons les outils utilisés pour la réalisation du projet et la présentation et la discussion des différents résultats obtenus lors de la phase d'expérimentation.

Chapitre 01: Introduction aux Systèmes de Recherche d'Information

1 Chapitre 01: Introduction aux Systèmes de Recherche d'Information.

1.1 Introduction

Les systèmes de recherche d'information ont pour objectif de retrouver l'information demandée par un utilisateur dans un ensemble de documents. La manière la plus simple serait de parcourir l'ensemble des documents à la recherche de l'information souhaitée. Mais ce processus a un inconvénient majeur : il est très coûteux en termes de performances. Pour faire face à cette problématique les systèmes de recherches indexent les documents et les requêtes.

1.2 Recherche d'Information (RI)

La recherche d'information a un but de faciliter à l'utilisateur l'accès à l'information qui l'intéresse, l'objectif est de développer des systèmes capables de retourner tous les documents pertinents et écarter tous les documents non pertinents.

Un Système de Recherche d'Information (SRI) est un système informatique constitué d'un ensemble de programmes, dont l'objectif principal est de sélectionner, dans une collection de documents préalablement enregistrée, les informations (documents) pertinentes répondant à un besoin en information formellement exprimé par un utilisateur sous forme de requête

1.2.1 Processus de recherche information

Les différentes étapes de ce processus de recherche sont schématisées dans la (figure 1) **[1]:**

1. La représentation de la requête.
2. La représentation des documents.
3. La fonction de correspondance

Le document représente l'information élémentaire recherchée par un SRI. Cette information structurée (HTML, XML, ...) ou non structurée (textuelle), peut apparaitre sous plusieurs formes (texte, image, vidéo, son, ...) et dans différents langages (français, anglais, arabe, ...). L'ensemble des documents sur lequel porte une recherche forme une collection de documents. (Nous focalisons dans la suite de ce mémoire sur les documents textuels non structurés)

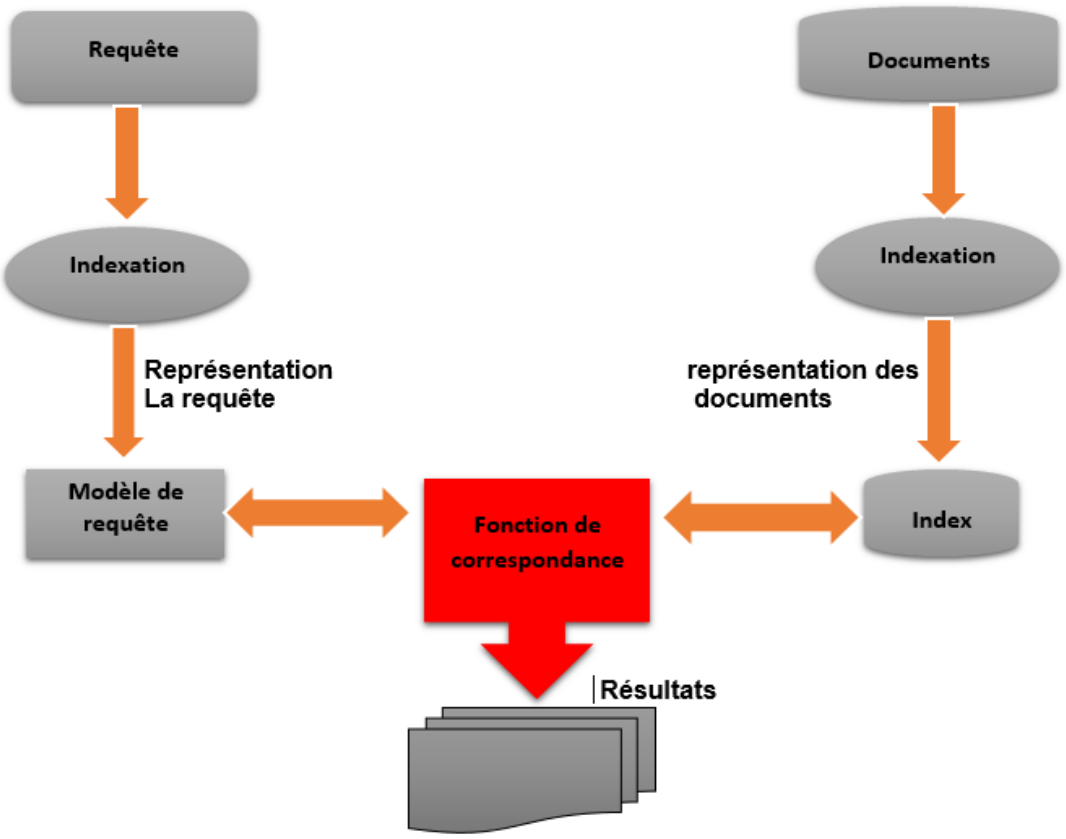


Figure 1:Architecture générale d'un système de Recherche d'Information. [1]

On a deux principales phases dans le déroulement du processus: indexation et recherche.

1.2.1 Indexation

L'indexation est la représentation qui permet de repérer et retrouver facilement l'information dans un ensemble de documents **[2].** C'est une activité de nature cognitive qui consiste à décrire le contenu d'un document à l'aide d'un langage d'indexation pour faciliter sa mémorisation dans un fichier, en vue d'une recherche ultérieure de l'information contenue dans ce document **[3]**. Elle consiste, après une analyse approfondie du contenu d'un document, à repérer, à sélectionner et à exprimer les informations contenues dans les documents primaires. Elle peut également servir à

comparer et classifier des documents, proposer des mots-clés, faire une synthèse automatique de documents, calculer des cooccurrences des termes, etc. **[4]**

La qualité de la recherche dépend en grande partie de la qualité de l'indexation donc cette étape consiste à déterminer et extraire les termes représentatifs du contenu d'un document ou d'une requête afin de créer une représentation textuelle qui soit utilisable par le SRI.

1.2.1.1 Langages d'indexation

1.2.1.1.1 Langage libre

Le langage libre est un langage évolutif, proche de notre Langue Naturelle (LN). Son vocabulaire est l'ensemble des éléments qui compose le langage, il est choisi a posteriori et n'est pas limité par un contrôle. Le vocabulaire est donc composé de tous les descripteurs choisis librement pour indexer les documents **[4]**. Le langage libre n'est pas régi par une syntaxe car aucune contrainte n'est spécifiée a priori. Par conséquent, le vocabulaire évolue rapidement et peut contenir des synonymes, polysémie etc. Ce qui entraîne des incohérences et diminue les performances du système de recherche d'information **[5].**

1.2.1.2 Langage contrôlé

Le langage contrôlé ou langage documentaire est un langage normalisé, il sert à éviter les problèmes d'ambiguïté (dû à l'homonymie et à la polysémie de certains termes) ainsi que les problèmes de redondance (synonymie, etc.) du langage libre **[6]**. Le langage contrôlé est construit a priori et permet de limiter le nombre de représentations d'un document ou d'une requête. Dans les langages contrôlés on peut trouver :

- **Vocabulaire contrôlé** : les vocabulaires contrôlés sont des listes énumérées de termes. Ces listes sont constituées de paires {concept, terme}. Tous les termes du vocabulaire contrôlé doivent être non ambigus et non redondants. Les vocabulaires contrôlés sont donc conçus afin d'éviter toute ambigüité. Ils contiennent une liste fermée des mots représentant le vocabulaire technique du domaine. **[7]**
- **Taxonomie (ou taxinomie)** : la taxinomie est la science qui a pour finalité de décrire des objets et de les regrouper en entités appelées taxons dans

l'intention de les identifier puis les nommer, et enfin les classer. Elle complète la systématique qui est la science qui organise le classement des taxons et leurs relations. **[7]**

- **Thésaurus** : un thésaurus **[8]** est un réseau de vocabulaires contrôlé. Il s'agit d'un niveau supérieur par rapport aux taxonomies. Il s'agit d'une représentation de données incluant les relations associatives en plus des relations hiérarchiques. Les thésaurus contiennent entre autres des relations hiérarchiques, des relations de synonymie, des traductions dans le cas de systèmes multilingues et peuvent contenir des descriptions sur les sujets. Les thésaurus sont très efficaces pour l'indexation. Ils sont à la fois la simplicité et la richesse des relations.
- **Ontologie** : le terme «ontologie» couvre plusieurs champs de la science. En philosophie, l'ontologie est la branche de la métaphysique concernant l'étude de l'être ; en médecine, l'ontologie s'intéresse à la genèse des maladies; en informatique, une ontologie est un système de représentation des connaissances **[8].**

1.2.2 Types d'indexation

L'indexation a pour but d'extraire les descripteurs les plus pertinents d'un document. Plus cette sélection est sophistiquée plus les tâches ultérieures de fouille de textes exploitant le système d'indexation (classification, recherche d'information, etc.) s'avèrent précises. Il existe deux types fondamentaux d'indexation : l'indexation manuelle et l'indexation automatique.

1.2.2.1 Indexation manuelle

L'indexation manuelle se fait par des experts en linguistique. Les experts analysent les textes et choisissent les termes descripteurs qui permettent de mieux représenter le contenu des documents. Le problème de cette indexation c'est que les index ne sont pas toujours les mêmes (deux experts différents peuvent choisir des descripteurs différents selon leur vision). Elle est basée sur un vocabulaire contrôlé.

1.2.2.1.1 Limites de l'indexation manuelle

- Indexation très coûteuse pour construire le vocabulaire et pour affecter les concepts (termes) aux documents.
- Difficile à maintenir puisque la terminologie évolue plusieurs termes sont ajoutés tous les jours.
- Processus humain donc subjectif qui engendre que des termes différents peuvent être effectués à un même document par des indexe différents.
- Les utilisateurs ne connaissent pas forcément le vocabulaire utilisé par les indexeurs.

1.2.2.2 Indexation automatique

Elle extrait des descripteurs automatiquement des textes en se fondant sur des règles d'analyse morphosyntaxique, des méthodes statistiques ou même sur des approches hybrides combinant les deux. Ce type d'indexation elle permet de pallier aux problèmes liés aux interprétations différentes et permet donc de limiter les problèmes de disparité. **[9].**

1.2.2.2.1 Les types d'indexation automatique

- **Indexation orientée document** : L'objectif est de résumer ou de présenter le contenu de chaque document.
- **Indexation orientée requête** : pour chaque document, refléter les requêtes pour lesquelles il est pertinent : l'indexation d'un document doit alors représenter les raisons pour lesquelles un utilisateur consulte ce document (i.e. : confronter chaque document de la base à une liste de requêtes prédéfinie). **[11]**

1.2.2.2.2 Les méthodes d'indexation automatique

- des méthodes **statistiques** et **probabilistes** : pour sélectionner les termes d'index, ces méthodes combinent les critères distributionnels (calcule les poids des attributs des vecteurs représentant les textes utilisant les méthodes de pondération).

- des méthodes **linguistiques** : **lexicographiques** et **morphosyntaxiques**. Ce sont les techniques employées dans le traitement automatique des langues naturelles (lemmatisation, radicalisation, normalisation, élimination des mots vide ...).
- des méthodes **informatiques** (telles que : **algorithmes de recherche, langages évolués spécifiques**, etc.) utilisées aussi bien dans le traitement automatique des langues qu'en documentation automatique **[12]**.

1.2.2.2.3 Avantages de l'indexation automatique

L'indexation automatique présente l'avantage d'une régularité du processus, car elle fournit toujours le même index pour le même document.

Elle permet d'offrir de longues «listes» de mots-clés très spécifiques sans délai et à coût réduit.

1.2.2.3 Fonction de pondération

La pondération des termes est l'une des méthodes de prétraitement; utilisées pour la présentation améliorée des documents texte en tant que vecteur de caractéristiques. La pondération des termes nous aide à localiser les termes importants dans une collection de documents à des fins de classement **[13].** Les schémas populaires pour le poids des termes sont le modèle booléen, la fréquence des termes (TF), la fréquence inverse du document (IDF) et la fréquence du document inversée (TF-IDF).

➢ Fréquence des termes

Cette approche consiste à attribuer à chaque terme d'un document un poids pour ce terme qui dépend du nombre d'occurrences du terme dans le document. Pour obtenir cela, calculez un score entre un terme de requête t et un document d, basé sur le poids de t dans d. L'approche la plus simple consiste à attribuer un poids égal au nombre d'occurrences du terme t dans le document d. Ce schéma de pondération est appelé terme fréquence et est noté$TF_{t,d}$, les indices désignant le terme et le document dans l'ordre **[14].**

$$TF(d, t_i) = \frac{n(d, t_i)}{\sum_i n(d, t_i)}$$

Équation 1: équation de TF

Où $\boldsymbol{n}\ (\boldsymbol{d}, \boldsymbol{t_i})$ est le nombre d'occurrences de t_i dans un document et $\boldsymbol{\Sigma i}\ \boldsymbol{n}\ (\boldsymbol{d}, \boldsymbol{t_i})$ est le nombre total de jetons dans le document.

➢ Fréquence inverse des documents

La fréquence des termes bruts souffre d'un problème critique: tous les termes sont considérés comme également importants lorsqu'il s'agit d'évaluer la pertinence d'une requête. En fait, certains termes ont peu ou pas de pouvoir de discrimination pour déterminer la pertinence. À cette fréquence de document inverse IDF (t) est réduit les termes qui se produisent dans de nombreux documents. Nous introduisons un mécanisme pour atténuer l'effet des termes qui apparaissent trop souvent dans la collection pour être significatifs pour la détermination de la pertinence.

$$IDF(t_i) = \log(\frac{D}{D_i})$$

Équation 2: équation de IDF.

Où $\boldsymbol{d_i}$ est le nombre de documents contenant $\boldsymbol{t_i}$ et $\boldsymbol{D}$ est le nombre total de documents

de la collection.

➢ Fréquence de terme-fréquence de document inverse (TF-IDF)

La fréquence des termes et la fréquence inverse des documents (TF-IDF), sont des méthodes populaires de prétraitement des documents dans la communauté de recherche d'information **[15].**

$$TF - IDF = TF * IDF$$

Équation 3: équation de TF-IDF

1.2.3 Recherche

Un processus qui permet de sélectionner l'information jugée pertinente pour l'utilisateur. Il se base sur un formalisme précis défini par un modèle de RI.

Les documents présentés en résultat à l'utilisateur considérés comme les plus pertinents, sont ceux dont les termes d'indexation sont les plus proches de ceux de la requête. **[16]**

1.2.3.1 Les modèles de RI

L'objectif des modèles de RI est de fournir une formalisation du processus de RI pour identifier et d'ordonner les documents pertinents par rapport à une requête donnée **[16].** Dans ce qui suit, nous décrivons brièvement les modèles de RI, et quelques modèles dérivés :

1.2.3.1.1 Modèles ensemblistes

- **Modèle booléen:**

Ce modèle est basé sur la théorie de l'ensemble. Dans ce modèle, les documents et les requêtes sont représentés par des ensembles de mots clés pondérés de façon binaire (0 ou 1), correspondant à la présence ou l'absence du terme dans le document relié par des connecteurs (le ou ∨, le et ∧, le non ¬). **[17]**

1.2.3.1.2 Modèles algébriques

- **Modèle vectoriel :**

Le modèle vectoriel repose sur les bases mathématiques des espaces vectoriels. Les documents et les requêtes sont représentés dans un espace vectoriel engendré.

La représentation d'un document par un vecteur :

$$D_j = (d_{1j}, d_{1j}, \ldots\ldots, d_{ij}, \ldots\ldots, d_{Tj})$$

La représentation d'une requête par un vecteur :

$$Q = (q_1, q_2, \ldots\ldots, q_i, \ldots\ldots,, q_T)$$

Avec :

$\boldsymbol{d_{ij}}$: Poids de terme $\boldsymbol{t_i}$ dans le document $\boldsymbol{D_j}$,

$\boldsymbol{q_i}$: Poids du terme $\boldsymbol{t_i}$ dans la requête$\boldsymbol{Q}$.

Dans ce modèle, la pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel, et appelée ***Retrieval Status Value*** ou RSV.

Ce coefficient de similarité est calculé sur la base d'une fonction qui mesure la colinéarité des vecteurs documents et requête. On peut citer notamment les fonctions suivantes **[11] :**

Produit scalaire :

$$RSV(Q, D_j) = \sum_{i=1}^{T} q_i * d_{ij}$$

Équation 4: Produit scalaire

Mesure de Jaccard :

$$RSV(Q, D_j) = \frac{\sum_{i=1}^{T} q_i * d_{ij}}{\sum_{i=1}^{T} q_i^2 + \sum_{i=1}^{T} d_{ij}^2 - \sum_{i=1}^{T} q_i * d_{ij}}$$

Équation 5: Mesure de Jaccard

Mesure de cosinus :

$$RSV(Q, D_j) = \frac{\sum_{i=1}^{T} q_i * d_{ij}}{(\sum_{i=1}^{T} q_i^2)^{1/2} * (\sum_{i=1}^{T} d_{ij}^2)^{1/2}}$$

Equation 6: Mesure de cosinus

- **Modèle LSI (latent Semantic indexing)**

Le modèle LSI est basé sur une représentation conceptuelle, le but de LSI est de transformer une représentation par des mots-clés en une autre représentation telle que les documents et les requêtes sémantiquement similaires seront plus proches avec la représentation transformée qu'avec les mots-clés **[18].**

1.2.3.1.3 Modèles probabilistes

Le modèle de recherche probabiliste utilise un modèle mathématique fondé sur la théorie de la probabilité. L'idée est d'implémenter les notions de la théorie des probabilités sur les SRI.

Le modèle probabiliste représente :

La probabilité de la pertinence d'un document D par rapport à une requête R. L'idée de base, dans ce modèle, est de tenter de déterminer les probabilités:

- $P\ (Pert/d, q)$: la probabilité de pertinence de document vis-à-vis de la requête.
- $P\ (Non\_Pert/d, q)$: la probabilité de non-pertinence de document vis-à-vis de la requête.

Le but de cette fonction de similarité dans ce modèle est d'essayer de séparer les documents pertinents des non pertinents au sein d'une collection **[12].**

1.2.4 L'évaluation d'un système de recherche d'Information

L'évaluation permettant de juger l'efficacité des SRI à retrouver des documents pertinents et à mesurer la différence entre un résultat attendu et un résultat obtenu, pour cela l'un des plusieurs métriques sont la précision le rappel et F-mesure leurs valeurs entre 0 et 1 pour faciliter l'interprétation

1.2.4.1 Rappel et précision

- **Le Rappel**

La capacité d'un système à sélectionner tous les documents pertinents de la collection, c'est-à-dire il mesure la capacité du système à retrouver tous les documents pertinents répondants à une requête.

$$\text{Rappel} = \frac{\textbf{nombre total de documents pertinents retrouvés par le système}}{\textbf{nombre total de documents pertinents dans la collection}}$$

Équation 7: formule de Rappel

- **La Précision**

La capacité d'un système a sélectionné que des documents pertinents, c'est-à-dire elle mesure la capacité du système de rejeter tous les documents non pertinents à une requête **[19]**.

$$\text{Précision} = \frac{\textbf{nombre total de documents pertinents retrouvés par le système}}{\boldsymbol{nombre\ total\ de\ documents\ retrouvés\ par\ le\ système}}$$

Équation 8: formule de Précision.

- **F-mesure**

Combine les mesures de rappel et de précision, On peut choisir la mesure F comme valeur synthétique exploitant la précision et le rappel.

Elle est calculée comme suit :

$$F= 2 \times \frac{(\boldsymbol{précision} * \boldsymbol{rappel})}{(\boldsymbol{précision} + \boldsymbol{rappel})}$$

Équation 9 : formule de F-mesure.

Conclusion

Dans ce chapitre, nous avons présenté le SRI, nous avons commencé par parler de l'indexation et de ses types, en nous concentrant sur l'indexation automatique que nous aurons dans notre programme, après cela nous avons parlé du RI, de ses modèles et des systèmes d'évaluation, sans oublier de parler des fonctions de pondération qui auront beaucoup besoin.

La recherche est un terme que nous utilisons fréquemment dans notre vie quotidienne, maintenant l'automatisation de ce concept et le rendre efficace est ce que les chercheurs et les programmeurs essaient de faire. Le chapitre suivant présente les différentes étapes de prétraitement des textes et la richesse et l'ambiguïté de la morphologie Arabe.

Chapitre 2:

Notions sur la langue arabe et son traitement automatique

2 Chapitre 02 : Notions sur la Langue Arabe et son Traitement Automatique

2.1 Introduction

La langue arabe est l'une des langues très riche morphologiquement et fortement flexionnelle, parlée par plus de 400 millions de locuteurs, la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du TALN.

Dans ce chapitre, nous présenterons les bases de la langue arabe et parlerons du «comment et pourquoi» il est difficile à gérer, en même temps, nous parlerons des méthodes de traitement.

2.2 Caractéristiques de la langue arabe

2.2.1 L'alphabet arabe :

L'alphabet arabe (tableau 1) est une *abjad* s'écrit et se lit de droite à gauche, comprend des consonnes et plusieurs signes diacritiques. La plupart des lettres s'attachent entre elles ce qui forme l'agglutination, leur graphie diffère selon leur position dans le mot : au début, au milieu ou à la fin.

Tableau 1:L'Alphabet arabe dans toutes les positions.

lettre	Nom	fin	milieu	début
ا	Alif	ـا	ـا	ا
ب	Ba	ـب	ـبـ	بـ
ت	Ta	ـت	ـتـ	تـ
ث	ṯa (tha)	ـث	ـثـ	ثـ
ج	ğim (jim)	ـج	ـجـ	جـ
ح	Ḥa	ـح	ـحـ	حـ
خ	ẖa (kha)	ـخ	ـخـ	خـ
د	Dal	ـد	ـد	دـ
ذ	ḏal (dhal)	ـذ	ـذ	ذـ
ر	Ra	ـر	ـر	ر-
ز	Zay	ـز	ـز	ز-
س	Sin	ـس	ـسـ	سـ

ش	šin (shin)	ـش	ـشـ	شـ
ص	Ṣad	ـص	ـصـ	صـ
ض	Ḍad	ـض	ـضـ	ضـ
ط	Ṭa	ـط	ـطـ	طـ
ظ	Ẓa	ـظ	ـظـ	ظـ
ع	ʿayn	ـع	ـعـ	عـ
غ	ġayn (ghayn)	ـغ	ـغـ	غـ
ف	fa	ـف	ـفـ	فـ
ق	qaf	ـق	ـقـ	قـ
ك	kaf	ـك	ـكـ	كـ
ل	lam	ـل	ـلـ	لـ
م	mim	ـم	ـمـ	مـ
ن	nun	ـن	ـنـ	نـ
ه	ha	ـه	ـهـ	هـ
و	waw	ـو	ـو-	و-
ي	ya	ـي	ـيـ	يـ
ء	hamza	أ ؤ إ ئ		

2.2.2 Les consonnes

Il existe 28 consonnes arabes fondamentales, mais il y a des auteurs qui traitent la lettre *alif (*ا) comme la vingt-neuvième consonne. L'*alif* se comporte comme une voyelle longue qu'on ne trouve jamais en tant que consonne de la racine **[20]**.

Il y a deux symboles *waw, yah* (و،ي) qui sont des semi-consonnes (glides), autrement dit, ils peuvent être considérés comme des consonnes ou des voyelles longues.

Toutes les consonnes se lient entre elles sauf *waw,reh,zain,dal,thal* (ذ , د ,ز , ر , و) celles qui ne se joignent jamais à gauche. De plus certaines lettres comme ا Alef peut symboliser le *ah* آ, *A* أ ou *I* إ; de même que pour les lettres ي et ه qui symbolise respectivement *yah*ى et *teh marbouta* ة.

2.2.3 Les voyelles

Les voyelles jouent un rôle important dans les mots arabes, non seulement parce qu'elles lèvent l'ambiguïté, mais aussi parce qu'elles donnent la fonction

grammaticale d'un mot indépendamment de sa position dans la phrase. Autrement dit, les voyelles ont une double fonction : l'une est morphologique ou sémantique et l'autre est syntaxique **[21]**.

Les voyelles sont de deux types : les voyelles brèves et les voyelles longues. Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte et permettent de différencier des mots ayant les mêmes consonnes.

- **Les voyelles brèves**

Les voyelles brèves (Fatha ـَ ,Damma ـُ ,Kasra ـِ) sont ajoutées au-dessus ou au-dessous des consonnes. Lorsque la consonne n'a aucune voyelle, on marquera une absence de voyelle représentée en arabe par une voyelle muette (Sukun ـْ).

- **Les voyelles longues**

Les voyelles longues (tableau 2) sont des lettres prolongées, elles sont formées par une des voyelles brèves suivis d'une des lettres correspondantes suivantes :*Alef,waw,yeh (ي،و،أ)* **.**

Tableau 2: Les voyelles longues

Voyelles longues	ـا	ـو	ـي

2.2.4 Les signes diacritiques :

En plus des 3 mouvements qui sont considérés comme voyelles, nous avons aussi Tanwin « التنوين » & aussi des contrôleurs "الضوابط" qu'ils sont "الشد" & "المد " & "الوصل" & "القطع" qui servent à aider à connaître la prononciation du mot et aussi à aider à expliquer davantage le sens du mot.

Šadda : est un signe qui peut être placé au-dessus d'une consonne mais qui ne peut pas être à la position initiale du mot. La consonne surmontée de ce signe est analysée comme une séquence de deux consonnes identiques géminées, la première avec une voyelle

brève : Fatha,Damma ou Kasra , dite motaharik ,et la deuxième sans voyelle avec sukun .par exemple *Mada* ~ مدّ (donner) est analysé comme *Madad* مدَدْ.

Tanwin : ou bien La désinence (ً an, ٌ un , ٍ in) considéré par quelques auteurs comme étant le double de même voyelles brèves(tableau 3), il est ajouté seulement à la fin des mots indéterminés, par conséquent il n'apparaît jamais avec l'article de détermination *AL* (ال). Le signe du *tanwin Fathatan*« » (à l'accusatif) est suivi toujours par ا *alif.*

Tableau 3:Les différents signes diacritiques.

Illustration en arabe	Nom de signe	Prononciation et fonction
Voyelle brève		
ـَـ	Fathatun	a/ signe d'accusatif
ـُـ	Damatun	u/signe de nominatif
ـِـ	Kasratun	i/signe de génétive
Voyelles casuelles (Tanwin)		
ـًـ	FathatAni	An
ـٌـ	DamatAni	Un
ـٍـ	KasratAni	In
Signes de syllabation		
ـْـ	Sukun	/aucune voyelle
ـّـ	shadda	Doublement de consonne

2.3 Lexique arabe

Le lexique de la langue arabe comprend trois catégories grammaticales de mots : verbe, nom et particule.

2.3.1 Verbe

Unité lexicale référant à un état ou une action exprimant un sens dépendant du temps comme : Eamila عَمِلَ (travailler),dahaba ذَهَبَ (partir) **[22]**.Nous pouvons classer les verbes arabes selon plusieurs critères **[23]** :

- Selon le critère de temps, il existe trois types : l'accompli, inaccompli, impératif.
- Selon leur sens et leur transitivité de sujet au complément aux deux types : Intransitive, transitive.
- Selon leurs modes aux deux types : la voix passive et la voix active.
- Selon le nombre des consonnes de la racine, la majorité des verbes a peu prés de 85% sont formés sur 3 lettres et le reste entre les racines de 4 et 5 lettres.

Ces racines peuvent donner plusieurs schèmes avec des transformations morphologiques.

- Selon le schème et le nombre de consonnes qui constituent la structure verbale, nous avons soit des verbes nus (Mojarad مجرد), soit des verbes augmentés (Mazid مزيد)

2.3.2 Nom

Toute unité lexicale référant à un sens indépendant du temps regroupe:
les adjectifs ; féminin et masculin ; les noms démérites, les noms prolongés ainsi que les noms réduits ; les noms communs et les noms propres ; les pronoms et leurs types (connectés et séparés) ; les pronoms relatifs ; les pronoms démonstratifs ; les noms d'interrogations ; les noms déterminés et non déterminés ; les noms de périphrases ; les noms du verbe ; les noms de voix ; les semblables des verbes de noms **[24] .**

2.3.3 Particule

Entité invariable exprimant un sens dépendant de compréhension. La langue arabe contient un nombre limité ne dépasse pas 80 éléments, ils se nommaient en arabe les particules de sens (حروف المعاني), par contre l'alphabet arabe se nommait les particules de construction (حروف المباني) **[23].**

Les particules de sens sont de type : unitaire, binaire, tertiaire, quaternaire ou quintette, Elles jouent un rôle important dans l'articulation et l'interprétation de la phrase ainsi la cohérence et l'enchaînement d'un texte.

Les particules sont classées selon leur sémantique et leur fonction dans la phrase. Il existe deux classes selon leur fonction (active, inactive) et 31 classes de particules selon leur sens, parmi lesquels on peut citer **[23]**:

- Particules de préposition : exemple *MaEa,ILA,Fi,Ka,Bi* (ب،ك،في،إلى،مع)
- Particules de coordination : exemple *Wa,Voma,Fa,Aaw* (أو،ف،ثم،و)
- Particules interrogatives : exemple *Aa,MaA,Hal* (هل،ما،أ)
- Particules d'affirmation : exemple *LaA,NaEam,Bala,Ajal (أجل،بلى،نعم،لا)*
- Particules de négat
- ion : exemple *Lame,LaA,Lane* (لن،لا،لم)
- Particules distinctive : exemple *Aye* (أي)
- Particules relatives : exemple *MaA* (ما)
- Particules de future : exemple *Sa,Sawefa,Lane,Aan* (أن،لن،سوف،س)

- Particules conditionnelles : exemple *Ine,Aaw* (إن،لو)
- Particule d'appel **:***YaA,Aa,AalaA* **(** يا،أ،الا**)**

2.4 Morphologie Arabe

La langue arabe a une morphologie riche et différente, par rapport aux autres langues. L'analyse morphologique d'un mot, consiste principalement à déterminer la structure générale de ce mot, les éléments essentiels utilisés pour construire ce mot sont :

2.4.1 Les racines

Les racines sont des verbes formés souvent de trois consonnes (Mustafa et al. 2008), elles sont à l'origine de la plupart des mots arabes. À partir d'une racine, on peut générer jusqu'à 30 mots, considérons l'exemple de la racine trigramme «كتب » (écrire) où on peut produire plusieurs nominaux et verbaux.

Dans cet exemple (tableau 4), nous remarquons qu'à partir d'une racine trilittérale «كتب », on peut générer plusieurs mots dans lesquels les trois lettres (ك, ت, ب) figurent, ainsi que d'autres lettres représentant les patrons insérés au début, au milieu ou à la fin du mot.

Tableau 4:Dérivation de plusieurs mots à partir de la racine «كتب , écrire »

Ecrire	Ecrivain	Livre	Petit livre	Ecrit
كتب	كاتب	كتاب	كتيب	مكتوب

2.4.2 Les schèmes

Les schèmes (ou modèles) sont des déclinaisons du mot «فعل », qui sont obtenus en ajoutant des affixes ou en utilisant des diacritiques. Par exemple, le modèle «مستفعل » est obtenu en y ajoutant les préfixes, par contre le modèle « فَعَّلَ » est obtenu en utilisant les diacritiques.

Les schèmes servent à extraire la racine d'un mot ou inversement à produire des stems à partir d'une racine **[25].**

2.4.3 Les affixes

Les affixes sont des morphèmes qui s'ajoutent au début ou à la fin des mots arabes. En général, Ils permettent de former, à partir d'une même racine, de nouveaux lemmes. Les affixes peuvent être subdivisés en deux types : préfixes et suffixes.

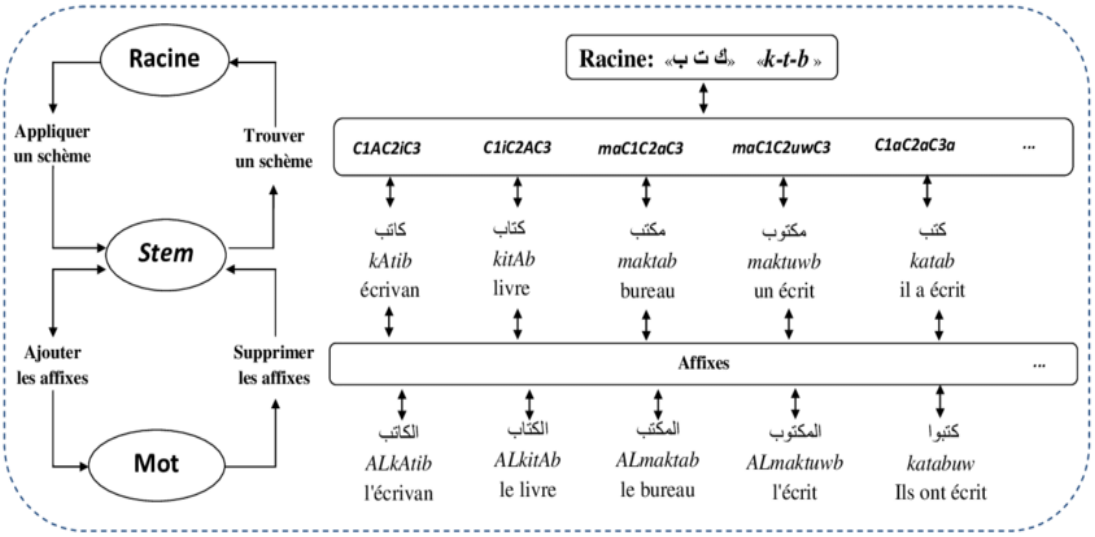
Les préfixes se placent avant le radical, et dépendent des mots auxquels ils s'attachent. Il y a trois types des préfixes: les préfixes nominaux qui sont réservés pour les noms et les adjectifs, les préfixes verbaux qui sont réservés aux verbes et les préfixes généraux qui sont indépendants du type des mots.

Les suffixes sont des morphèmes placés après le radical. Il existe deux types des suffixes: les suffixes verbaux qui dépendent de la transitivité, et les suffixes nominaux indiquant la flexion du nom, du nombre et du genre, etc.

2.4.4 Les stems

Un stem est obtenu par troncature sur les deux extrémités du mot sans modification interne sur le mot. C'est la dérivation obtenue à partir d'une racine donnée selon un patron. Par exemple, le lemme « مدرس, enseignant», il est obtenu à partir de la racine «درس, il a étudié» selon le schème « مفعل ». Le tableau présente un exemple de génération des stems (tableau 5).

Tableau 5:Exemples de génération des stems.



2.5 Mots dérivés

La plupart des mots arabes sont considérés comme des mots dérivés, puisqu'ils sont construits à partir des racines. Ainsi, les mots qui dérivent d'une même racine ont des

sens différents. En effet, les mots dérivés sont construits à partir d'un stem en y ajoutant des affixes comme c'est le cas du nom «أتطلبون, Est-ce que vous demandez ?»(Figure 2)

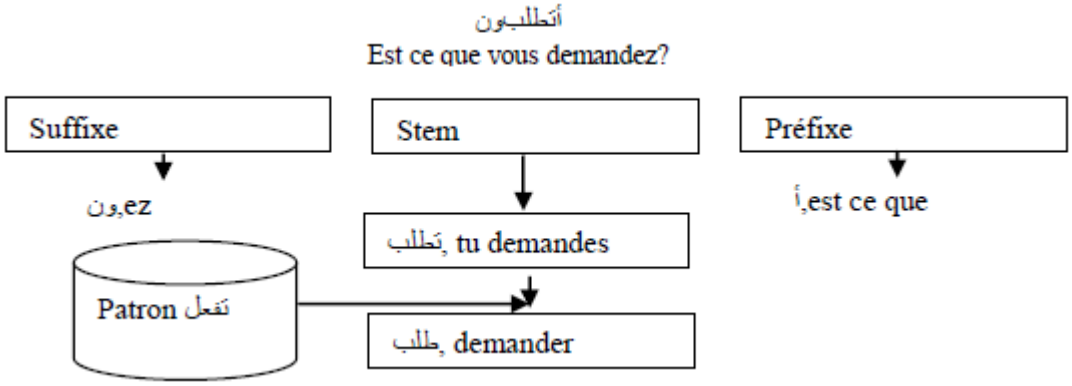


Figure 2:Exemple de mot dérivé « أتطلبون » [26]

2.6 Structure d'un mot arabe

En arabe un mot peut signifier toute une phrase grâce à sa structure composée qui forme une agglutination d'éléments de grammaire (tableau 6), ceci définit le mot graphique arabe ; cette appellation est désignée par David Cohen à un mot décomposable aux proclitiques, forme fléchies, enclitique avec la forme fléchie représente le noyau lexical.

La représentation suivante (figure 3) schématise une structure possible d'un mot. Notons que la lecture et l'écriture d'un mot se font de droite vers la gauche.

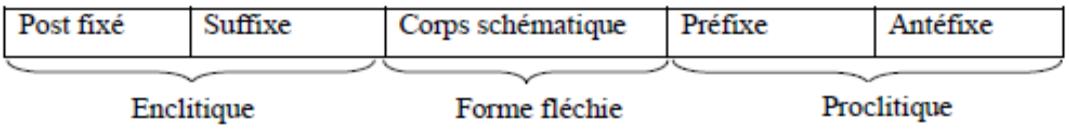


Figure 3:Structure d'un mot arabe [23].

ها	ون	أكل	ت	أف

Tableau 6:exemple du enclitique et proclitique d'un mot arabe .

2.6.1 Les proclitiques :

Les proclitiques sont des antéfixes et des préfixes, les antéfixes sont des prépositions ou des conjonctions et les préfixes sont les traits grammaticaux dépendus de l'aspect

verbal dans le cas des verbes, et de déclinaison dans le cas des noms et déverbaux (Nombre, genre, personne,…) **[27].**Quelques exemples de proclitiques **[28]**:

- Les proclitiques réservés aux noms :

  L'article de définition ' ال ' (*Al*) (préfixes).

  La préposition' بِ' (*bi*) (Antéfixes).

- Les proclitiques réservés aux verbes :

  La particule de subjonctif ' لِ ' (*li*) (Antéfixe).

  La particule de futur ' س' (*sa*) (Antéfixe)

  La particule de l'apocopé ' لْ (*li*)

- Les proclitiques réservés aux verbes et noms :

  L'article d'interrogation ' أ' (*Aa*) (Antéfixe)

  Les conjonctions de coordination' وet ' ف ' (*wa,fa*)(Antéfixe)

  La particule d' affirmation ' لَ' (*la*) (Antéfixe)

On peut indiquer qu'il existe plusieurs ambigüités dans le rôle d'un proclitique, par exemple: la particule و est utilisé dans la majoritaire des cas comme une particule de coordination, dans des moindres cas comme particule d'accompagnement, et rarement une particule de serment.

2.6.2 Les enclitiques

Représentent les suffixes et les post fixés, les suffixes sont des traits grammaticaux par contre les post fixés sont des pronoms personnels. Dans le cas des noms et en mode non déterminé, les noms acceptent toutes les enclitiques, par contre les noms qui se terminent par ى (*Y*) ou par ي (*y*) nécessitent des transformations morphologiques avant leur suffixation comme dans le cas du mot 'مبنى ' (MAbenaY, immeuble) qui est transformé par changement de ي (*Y*) au (*A*) ا et l'ajout d'un suffixe comme ه(*h*), le mot donc devenir :

مبناه (*MabenaAh,* ses immeuble)**[28].**Dans le cas des verbes, les enclitiques sont varis selon leur aspect et pronom personnel.

Le tableau (tableau 7) suivant indique toutes les particules unitaires qui ont exprimé soit des proclitiques soit des enclitiques du verbe ou nom.

Tableau 7:Les rôles des particules unitaires dans un mot arabe.

La particule	Le sens de particule
أ (>)	Antéfixe de Question, appel, égalité
ا (A)	Suffixe de l'exclamation, et de secours
ب (b)	Antéfixe de préposition
ت (t)	Antéfixe de serment, et suffixe de féminin
س (s)	Antéfixe de futur
ف (f)	Antéfixe de conjonction
ك (k)	Antéfixe de préposition
ل (l)	Antéfixe de préposition pour les noms et d'affirmation pour les verbes
م (m)	Suffixe d'indication de pluriel masculin
ن (n)	Préfixes de l'inaccompli et suffixe d'affirmation
ه (h)	Post fixe de l'absence
و (w)	Antéfixe de conjonction
ي (y)	Préfixe de l'inaccompli

- Dans notre travail nous avons considéré que tous les proclitiques et les enclitiques sont des préfixes et des suffixes.

2.7 Les difficultés du traitement automatique de la langue Arabe

Le traitement automatique des langues(TAL) est la conception de logiciels ou programmes, capables de traiter de façon automatique des données linguistiques (textes) exprimées dans une langue dite « naturelle». Le TAL arabe rencontre plusieurs défis dépendant de l'absence fréquente des voyelles courtes dans le texte arabe et d'autres phénomènes morphologiques et syntaxiques cités au-dessous, ce qui risque de générer une certaine ambiguïté.

2.7.1 Absence des voyelles :

Le problème de la voyelle réside dans son absence dans les textes arabes. En effet, les signes des voyelles sont des signes diacritiques placés au-dessus ou au-dessous des lettres, qui apparaissent dans certains ouvrages scolaires pour débutants et dans le Coran. Le non voyelle génère plusieurs cas d'ambigüités et des problèmes lors de l'analyse automatique (tableau 8) **[29].**

Tableau 8:exemple d'un mot arabe avec différents sens à cause d'absence de voyelle.

Le mot : **الجد**	**الجَدّ**	**الجِدّ**	**الجُدّ**
Le mot en alphabet français	aljadu	Aljidu	aljudu
Le sens en arabe	أبو الأب	الاجتهاد	ساحل البحر
Le sens en français	le grand-père	Diligence	côte de la mer

2.7.2 Agglutination :

Le phénomène d'agglutination de mot arabe est lié aux clitiques rattachés aux verbes et noms, ils sont définis comme une liste d'affixes (suffixes, préfixes, postfixes, antéfixes). Ces clitiques génèrent certains problèmes d'ambigüité spécifiques à la segmentation d'un mot, ce qui permet d'avoir plusieurs formes comme dans l'exemple suivant (tableau 9) : أوصلوهم (*AwSlwhm*) :

Tableau 9:Illustration d'un exemple de plusieurs segmentations d'un mot.

Antéfixe	préfixe	Forme fléchie	suffixe	Post fixé
أ:article d'interrogation *(>a)*	وَ:conjonction de coordination (wa)	صَلَّ:verbe a l'accompli prière(Sala~u)	و:suffixe verbal exprime le pluriel (*w*)	هُمْ:pronom complément de nom(*hum*)
	أ:article d'interrogation *(>a)*	وَصَل:verbe a l'accompli arriver (*waSala*)	و:suffixe verbal de pluriel (*w*)	هُمْ:pronom complément de nom (*hum*)
		أَوْصَل:verbe a l'inaccompli faire arriver *(>aweSala)*	و:suffixe verbal du pluriel (*w*)	هُمْ:pronom complément de nom (*hum*)

La bonne représentation du mot est indiquée par une analyse morpho-lexical puissant pour affecter les catégories grammaticales justes suivant les règles d'agglutination des proclitiques et des enclitiques parmi les il existe :

- La relation d'ordre : il faut toujours respecter l'ordre des proclitiques entre eux ainsi les enclitiques selon la catégorie grammaticale de chacun pour former le bon sens d'un mot par exemple l'article d'interrogation أ se précède toujours les proclitiques du verbe de l'inaccompli : أ،ن،ت،ي *(>,n,t,y)* .
- La compatibilité entre les proclitiques et les enclitiques : pour former la bonne expression d'un mot arabe, il faut aussi respecter la compatibilité entre les proclitiques et les enclitiques, pour cela il existe plusieurs contraintes grammaticales pour gérer leurs enchainements, et diriger les analyses morphologiques.
    - Les contraintes grammaticales pour les verbes **[28]**:
- L'article d'interrogation أ ne peut pas être collé avec un verbe conjugué à l'impératif ou subjonctif.
- La particule س ne peut pas joindre qu'a un verbe conjugué à l'inaccompli (active ou passive).
- Les pronoms personnels ne se collent ni aux verbes intransitifs, ni aux verbes conjugués à la voix passive.
- Lorsqu'un verbe est conjugué avec les premiers et les deuxièmes pronoms personnels alors il ne peut pas agglutiner avec un pronom de la même personne.
    - Les contraintes grammaticales pour les noms :
- L'article de définition ' ال '(*Al*) ne peut être compatible avec les enclitiques de pronoms personnels, ni avec *tanwin*.

2.7.3 Ordre des mots dans la phrase

En langue arabe, on met au début de la phrase le mot sur lequel on veut attirer l'attention ou le mot le plus riche en sens. Cet ordre provoque des ambiguïtés syntaxiques artificielles dans la mesure où il faut prévoir dans la grammaire toutes les règles de combinaisons possibles d'inversion de l'ordre des mots dans la phrase **[30].** Ainsi par exemple (tableau 10), on peut changer l'ordre de la phrase (1) pour obtenir une autre phrase (2) ayant le même sens.

Tableau 10:Exemple de changement d'ordre de la phrase.

La phrase en arabe	La phrase en français
لعب الطفل مع اصدقاءه في الساحة	L'enfant a joué avec ses amis dans la cour.
في الساحة، لعب الطفل مع اصدقاءه	Dans la cour, l'enfant a joué avec ses amis.

2.7.4 Mots homographiques

C'est tous les mots qui ont les mêmes formes orthographiques mais la prononciation est déférente **[28]**, ils ont apparait dans la majorité des cas dans les textes non vocalisés et qui ont causé des ambigüités lexicales et syntaxiques. (Sens du mot et la difficulté à identifier sa fonction dans la phrase) **[31].**

Le lexique arabe contient plusieurs mots homographies qui ont des significations et des catégories grammaticales différentes comme (tableau 11) : علم

Tableau 11:Exemple de mot « علم »

عَلَمٌ :drapeau	عِلْمٌ:science	عَلِمَ :savoir
(Ealamun)	*(Eilemun)*	*(Ealima)*

Aussi les verbes défectueux peuvent générer des mots graphiques lors de modifications de la lettre défectueux, ainsi l'existence de chadda en leur conjugaison comme (tableau 12): يعد

Tableau 12: Exemple de mot « يعد»

Il a plusieurs sens :	يُعِدُّ : prépare (le verbe أَعَدَّ)		يَعِدُ : promesse(le verbe وَعَدَ)	
	(yuEidu~)	*(>aEada~)*	*(yaEido)*	*(waEada)*
يَعُدُّ : compte(le verbe عَدَّ)		يَعُدُ : revient(le verbe عَادَ)		يُعِدُ : refait(le verbe أَعَادَ)
(yaEudu~)	*(Eada~)*	*(yaEude)*	*(EaAda)*	*(yuEide)*	*(>aEaAda)*

2.7.5 Système numérique arabe

En observant les écrits arabes, on remarque une double norme dans l'usage des chiffres selon le pays. Ainsi, les pays d'Afrique du Nord utilisent les chiffres arabes dans leurs formes arabes, alors que cet usage est différent dans la plupart des pays arabes du Moyen-Orient, de l'Egypte et de l'Arabie Saoudite où l'usage des anciens chiffres arabes dits indiens est en vigueur (tableau 13) **[32].**

Au niveau de la lecture, le nombre est lu en commençant par la plus petite valeur comme

21 se lit un et vingt. Les nombres sont appartenus à la catégorie des noms.

Tableau 13:Le système numérique arabe.

Type	Exemple
Chiffres arabes standards (Tunisie, Algérie, Maroc).	0 1 2 3 4 5 6 7 8 9
Chiffres arabes *variantes occidentales* (Égypte, Syrie, Palestine.)	٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩

2.8 Les prétraitements des documents textes en langue arabe

Les documents arabes peuvent être écrits en caractères arabes ou/et en caractères latins translittérés. Ces documents subissent un prétraitement afin d'améliorer l'efficacité de la recherche d'information. Les étapes de prétraitement comprennent les techniques suivantes :

2.8.1 Segmentation :

La segmentation de textes en phrases reste une phase préalable pour le traitement automatique des langues. Cette phase de traitement ne fait pas l'objet de beaucoup de recherches. La segmentation de textes est basée sur l'étude linguistique d'une part, et sur une modélisation informatique d'autre part. Ces deux études se complètent. Comme d'autres types de traitement automatique de la langue, la segmentation a ses particularités, que ce soit au niveau linguistique, ou au niveau informatique.

La plupart des segmenteurs (tokenizers) existants sont limités à la simple utilisation des espaces et des marques de ponctuation «. », « ! », « ? », avec une étude de quelques cas d'ambigüités sur des corpus bien déterminés.

Toutefois, les particularités de la langue arabe rendent la segmentation toujours difficile. Par exemple, il n'y a pas de majuscules qui marquent le début d'une nouvelle phrase. De plus, les signes de ponctuation, ne sont pas utilisés de façon régulière. D'après l'étude réalisée par Belguith **[30],** certaines particules comme "et | و ", "donc | ف ", etc. jouent un rôle principal dans la séparation des phrases et peuvent être déterminantes pour guider la segmentation.

2.8.2 Elimination des mots vides :

Les mots vides (Stop Word) sont des mots fréquents qui ne sont pas porteur de sens **[33]**,

La stratégie générale pour déterminer une liste des mots vides est de trier les termes par fréquence de corpus (le nombre total d'apparition de chaque terme dans la collection de documents) et puis prendre les termes les plus fréquents, ces termes sont ensuite rejetés lors de la phase d'indexation.

2.8.3 Normalisation :

La normalisation d'un texte consiste à apporter des modifications sur quelques lettres principalement la lettre «أ ». Par exemple, certaines lettres subissent une simple modification dans l'écriture qui n'influe pas considérablement sur le sens du mot. Mais l'encodage de ces lettres change d'un mot à un autre. Une autre raison pour ce prétraitement est la mal écriture des différentes formes de hamza. Par exemple, le mot « أخذ » est généralement écrit « اخذ ». Aussi la lettre «ة » à la fin des mots peut être écrite de deux façons : « ة » ou « ه ». Par exemple les deux mots « كبيرة » et « كبيره » ont la même signification « grande » malgré que leur dernière lettre sont différente.

Également, les voyelles sont souvent omises dans les textes, néanmoins, on peut parfois trouver quelques voyelles présentes avec les mots. Alors, l'élimination de ces voyelles est nécessaire pour la normalisation des textes.

2.8.4 Racinisation (Stemming)

Le stemming (racinisation en français) vise à garder la Racine du mot c'est-à-dire consiste à extraire la racine d'un mot et à associer les mots lies morphologiquement à

la même racine **[34].** Par exemple, les mots « درسوا »،«يدرس», et « درست » sont considérés comme des descripteurs différents alors qu'il s'agit de trois formes conjuguées du même verbe «درس ».

2.9 Travaux apparentés aux TALA

Plusieurs travaux ont été développés récemment dans plusieurs domaines tels que la traduction automatique et la recherche d'information, les différentes expériences réalisées n'utilisent pas les mêmes types de traitements.

- Liu et al. **[35]** proposer une méthode d'extraction de mots-clés basés sur le regroupement sémantique qui garantit une bonne couverture sémantique du document. La méthode permet d'extraire les termes candidats qui être regroupée en classes après avoir calculé les liens sémantiques entre ces termes. Ce regroupement consiste à développer un ensemble de mots de référence pour chaque classe. Les mots de référence sont utilisés pour extraire les mots-clés après filtrage des termes candidats.
- Raheel et al. **[36]** a combiné la méthode Boosting et l'arbre de décision en tant que classificateur hybride. Ils ont utilisé la lemmatisation comme méthode d'extraction des caractéristiques et le TFIDF pour la pondération. Une comparaison de la méthode a été réalisée avec deux classificateurs, Bayesian Naïve (NB) et SVM (Support Vector Machine). Le résultat montre que SVM et NB surpassent l'approche proposée.
- Thabtah et al**.[37]** mettre en place un système de catégorisation arabe utilisant le classificateur bayésien naïf basé sur les caractéristiques de pondération fournies par le test $\chi 2$(Chi-square Testing) pour classer une base de données étiquetée simple. Les résultats expérimentaux, comparés à l'ensemble de données classifiées, montrent que la sélection des caractéristiques améliore souvent la précision de la classification en supprimant les termes vides ou rares.
- Bawaneh et al. **[38]** ont comparé les deux classificateurs, KNN (K Nearest Neighbor) et NB (Naïve Bayesian). Le stammer a été utilisée comme caractéristique et la mesure TFIDF comme méthode de pondération des caractéristiques. Le classificateur KNN a été jugé plus efficace.

- Kanaan et al. **[39]** classé les documents en arabe avec l'algorithme de maximisation des attentes (Expectation-Maximization). La mesure TFIDF est appliquée comme méthode de pondération des éléments caractéristiques, tandis que l'algorithme bayésien naïf est utilisé pour calculer les étiquettes des documents, et enfin la classification est faite à l'aide de l'algorithme EM
- Jamoussi **[40],** propose une méthode d'extraction de mots clés basée sur la représentation sémantique des termes. Dans son travail, Elle a présente deux méthodes basées sur les distances sémantiques, la distance Kullback-Leibler et les informations mutuelles moyennes pour calculer la quantité d'informations entre deux mots ou deux classes de mots. La nouvelle méthode introduite par Jamoussi est testée par rapport à une représentation vectorielle simple, avec trois classificateurs non supervisés : l'algorithme K-means, les cartes de Kohonen et le réseau bayésien AutoClass.

Conclusion

Dans ce chapitre en a présenté les bases de la langue arabe, aussi pour démontré les difficultés on a pour traiter les documents arabes par rapport à autre langue.

Le TALA reste très difficile mais comme en a prouvé dans la fin de ce chapitre, il existe des prétraitements qui simplifient le travail pour donner des résultats plus pertinents.

Le chapitre prochain expose quelques méthodes d'apprentissage automatique. Ces méthodes se divisent en deux catégories. La première catégorie concerne les méthodes supervisées. La deuxième catégorie concerne les méthodes non supervisées.

Chapitre 3:

Apprentissage automatique (Machine Learning)

3 Chapitre 03 : apprentissage automatique

3.1 Introduction:

À une époque où l'intelligence artificielle est un terme qui est entendu partout, beaucoup de gens ne le lient qu'à des robots et à des choses futuristes, alors que la vérité est que l'AI peut affecter beaucoup de domaines et l'un d'entre eux est le domaine de la recherche d'information et l'indexation. Dans ce chapitre, nous allons parler d'apprentissage automatique pour la lecture et la classification des documents.

3.2 Les données non structurée :

Près de 85% des données sont sous forme non structurée; emails, contrats, rapports de médecins et messages sur les réseaux sociaux : ce sont toutes des données non structurées et elles renferment une véritable mine d'or d'informations.

Les informations non structurées ou données non structurées sont des données représentées ou stockées sans format prédéfini. Ces informations sont toujours destinées à des humains. Elles sont typiquement constituées de documents textes ou multimédias, mais peuvent également contenir des dates, des nombres et des faits **[41]**

Mais comment faire pour extraire rapidement et correctement les informations les plus pertinentes de cette énorme quantité de données non structurées ?

ca nécessite certaines approches, l'une de ces approches est ce qu'on a fait dans notre travail.

3.3 L'apprentissage automatique :

L'apprentissage automatique consiste en l'acquisition de connaissances à partir d'observations de phénomène, inclut des méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune, Nous distinguons deux types d'apprentissage : le supervisé et le non supervisé **[42]**

3.3.1 Types d'apprentissage:

Les algorithmes d'apprentissage peuvent se catégoriser selon le type d'apprentissage qu'ils emploient **[42]:**

3.3.1.1 Supervisé :

L'apprentissage est supervisé, si les classes sont prédéterminées et les exemples connus, c'est-à-dire lorsque les données qui entrent dans le processus sont déjà catégorisées après pour étudier la fiabilité des règles, on utilise souvent un échantillon indépendant, dit de validation ou de test.

3.3.1.2 Non supervisé:

L'apprentissage non supervisé ou **clustering** en anglais, si le système ou l'opérateur ne dispose que d'exemples, mais non d'étiquette, et que le nombre de classes et leur nature n'ont pas été prédéterminés

3.4 Les méthodes d'apprentissage supervisé :

3.4.1 Les k plus proches voisins (K-PPV) :

L'algorithme KNN est l'un des plus simples de tous les algorithmes d'apprentissage automatique, Il a prouvé son efficacité face au traitement de données textuelles. La phase d'apprentissage consiste à stocker les exemples étiquetés. Le classement de nouveaux textes s'opère en calculant la distance entre la représentation vectorielle du document et celle de chaque exemple du corpus, donc pour affecter un nouvel individu à une classe, l'algorithme cherche les k plus proches voisins parmi les individus déjà classés. Ainsi, l'individu est affecté à la classe qui contient le plus d'individus parmi les candidats trouvés.

Pour tester la similarité entre deux vecteurs. Il existe plusieurs types de distance parmi lesquels on trouve **[43]**

La distance Euclidienne :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Équation 10:La distance Euclidienne

La distance de Minkowsky :

$$d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

Équation 11:La distance de Minkowsky

La distance de Manhattan :

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Équation 12: La distance de Manhattan.

- **Le choix du K :**

Pour sélectionner la valeur de **k** qui convient, on doit exécuter plusieurs fois l'algorithme KNN avec différentes valeurs de **k**. Puis on choisit le **k** qui réduit le nombre d'erreurs rencontrées tout en maintenant la capacité de l'algorithme à effectuer des prédictions avec précision lorsqu'il reçoit des données nouvelles (non vues auparavant) **[44]** .

- **L'algorithme de KNN**

L'algorithme ci-dessous montre comment classer un nouvel exemple par la méthode K plus proche voisin KPPV (figure 4):

- Initialisation, choix de :
  - ✓ Nombre de classes
  - ✓ Valeur de k
  - ✓ exemples initiaux

- Pour chaque vecteur d'objet à classer :
  - ✓ Mesurer la distance du vecteur avec tous les autres déjà classés
  - ✓ Déterminer la liste des k vecteurs les plus proches de lui (k-ppv)

- Déterminer la classe la plus représentée dans la liste des k-ppv et affecter notre vecteur à cette classe.

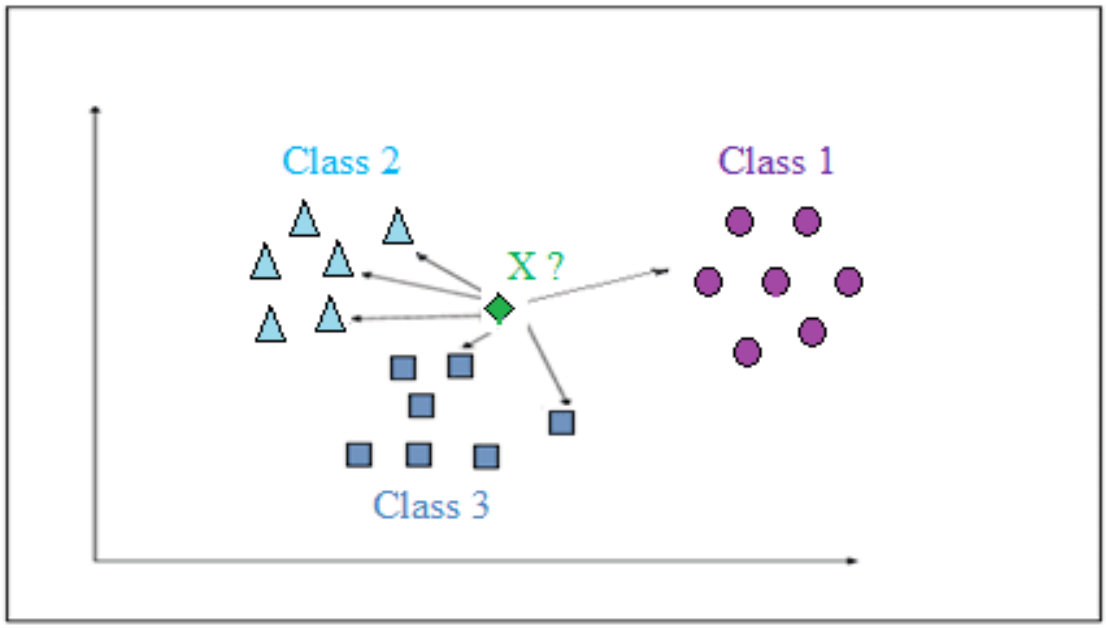


Figure 4:Explication de processus kNN. [45].

3.4.2 Machine à vecteur de support (SVM)

Les Machines à vecteur de support (SVM) sont à l'origine de nouvelles méthodes de catégorisation **[46],** le principe des SVM consiste en une stratégie de minimisation structurelle du risque. En ce qui concerne son application à la problématique de catégorisation de documents, l'approche par SVM permet de définir, par apprentissage, une surface de séparation entre des exemples positifs et négatifs minimisant le risque d'erreur et maximisant la marge entre deux classes. La figure (figure 5) montre une telle séparation dans le cas d'une séparation linéaire par un hyperplan. Il est intéressant de remarquer qu'en réduisant le jeu d'entraînement uniquement aux vecteurs de support, l'algorithme calculerait le même hyperplan que pour le jeu d'entraînement complet. La marge se présente alors comme la plus courte distance entre un vecteur de support et "son" hyperplan.

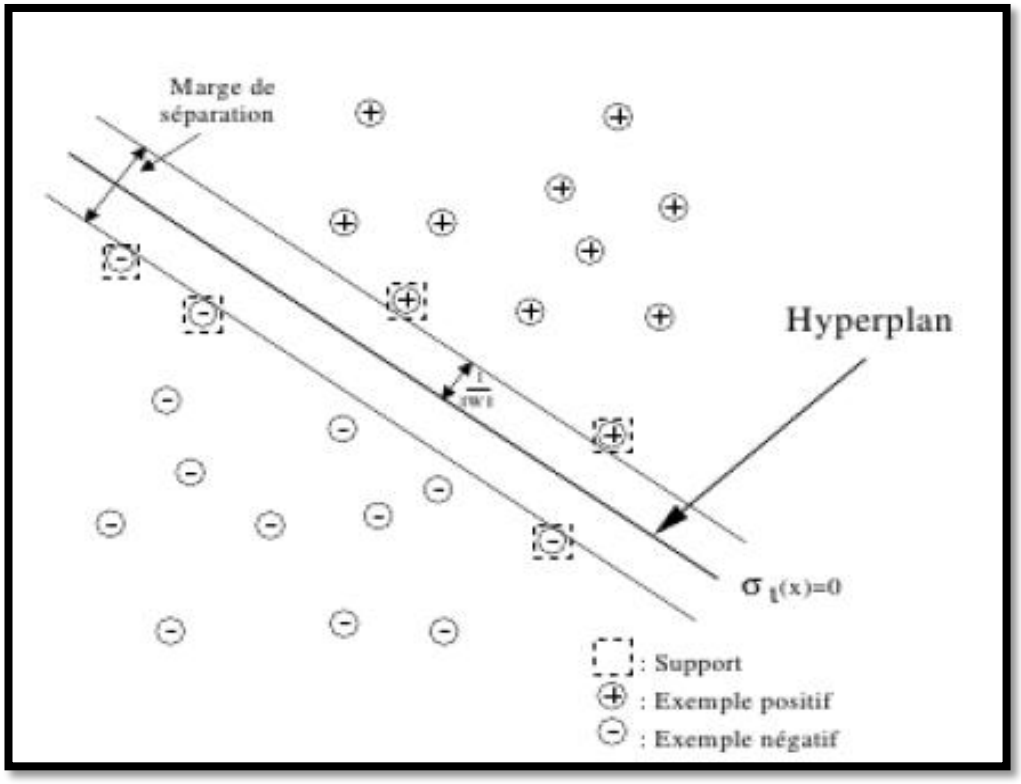


Figure 5:Explication de processus SVM.[46]

De manière formelle, un hyperplan peut être défini par :

$$\vec{w} \cdot \vec{x} + b = 0$$

Avec x un point arbitraire, w un vecteur et b le biais.

Soit $D = \{(xi, yi)\}$ le jeu d'entraînement et yi ∈ {±1} définissant l'état, positif ou négatif, de l'exemple. Trouver l'hyperplan maximisant la marge séparatrice revient à résoudre le problème suivant :

$$\begin{cases} minimiser\ \frac{1}{2}\|w\|^2 \\ sous\ les\ contraintes\ \forall_i\ , y_i(\vec{w}.\vec{x}_i + b) - 1 \geq 0 \end{cases}$$

Équation 13: équation de l'hyperplan.

Grâce à une extension de cet algorithme, il est aussi possible de résoudre des problèmes qui ne sont pas linéairement séparable, mais l'amélioration obtenue pour la catégorisation de documents reste minime. Pour la construction vectorielle

des textes, ce sont en général les stems (radicaux) qui sont utilisés comme termes d'indexation **[46].** SVM est considéré comme un des algorithmes les plus performants en classification textuelle.

- **Implémentation de SVM :**

Les SVMs sont une famille d'algorithmes d'apprentissage automatique, il ont pour but de séparer les données en classes à l'aide d'une frontière (hyperplan), de telle façon que la distance (marge) entre les différents groupes de données et la frontière qui les sépare soit maximale. Les « vecteurs de support » étant les données les plus proches de la frontière.

Pour que le SVM puisse trouver cette frontière, il est nécessaire de lui donner des données d'entraînement. En l'occurrence, on donne au SVM un ensemble de points.

- **Systèmes non linéaires (astuce du noyau):**

Abordons maintenant le problème des données non linéairement séparables. Pour rappel, des données sont non linéairement séparables quand il n'existe pas d'hyperplan capable de séparer correctement les deux catégories.

Le passage à la recherche de surfaces séparatrices non linéaires est obtenu par l'introduction d'une fonction noyau (kernel) dans le produit scalaire induisant implicitement une transformation non linéaire des données vers un espace intermédiaire de plus grandes dimensions, On peut citer des exemples de noyaux suivants **[47]**

1. Linéaire : $K\ (x, x') = (x * x')$

2. Polynomial : $K\ (x, x') = (x * x')^d$

3. Sigmoïde $k(x, y) = tanh(x\,.\,y + 1)$ / $tanh$: La fonction tangente hyperbolique

4. RBF (Radial Basis Function) $k(x, y) = e - \sigma\ (x\_y)^2$

➢ Classification multi-classe

Le classifier SVM ne pouvant résoudre que des problèmes binaires (deux catégories), différentes approches ont été mises en place pour résoudre les problèmes multi-classes. Leur principe est toujours le même : transformer le problème multi-classes en plusieurs problèmes binaires. Il existe deux méthodes différentes : un contre un (one-vs-one) et un contre tous (one-vs-all) **[48]** .

- **Un contre un (one-vs-one) :**

Cela consiste à apprendre chaque couple de classes (la premier classe contre la deuxième classe , la premier contre la 3eme ,la 2eme contre la 3eme ,etc.) puis de procéder à un vote majoritaire pour déterminer l'appartenance d'un exemple à l'une des classes

- **Un contre tous (one-vs-all) :**

Consiste à regrouper dans une classe (négative) toutes les classes sauf une (qui sera la classe positive) et d'apprendre ainsi chaque classe contre les autres. Le choix de l'appartenance à une classe se fait alors selon la valeur max calculée par chacun des SVM binaires.

3.4.3 Naïve Bayes :

Le classificateur Naïve Bayes est un catégoriseur du type probabiliste fondé sur le théorème de Bayes (1763). Considérons $v_j = (v_{j1}, \ldots, v_{jk}, \ldots, v_{jd})$ un vecteur de variables aléatoires représentant un document $d_j$et C un ensemble de classes.

En s'appuyant sur le théorème de Bayes, la probabilité que ce dernier appartienne à la classes $ci \in C$ est définie par **[48]:**

$$P(c_i/v_j) = \frac{P(c_i)P(v_j/c_j)}{P(v_j)}$$

Équation 14: équation de probabilité.

Ce théorème repose sur l'hypothèse que des solutions recherchées peuvent être trouvées à partir de distributions de probabilité dans les hypothèses et dans les données. Un classificateur bayésien naïf, dans le cadre de la classification de textes, permet de

déterminer la classe d'un document spécifié en supposant que les documents sont indépendants.

Cette hypothèse d'indépendance ne reflète pas la réalité d'où l'appellation naïve. La classe la plus probable d'un nouvel objet est déterminée en combinant les prédictions de toutes les hypothèses en les pondérant par leurs probabilités a priori **[49].**

3.4.4 Arbres de décision

Un arbre de décision est un arbre au sens informatique du terme. Les nœuds internes sont appelées nœuds de décision. Chaque nœud de décision est étiqueté par un test qui peut être appliqué à toute description d'un individu de la population. En général, chaque test examine a valeur d'un unique attribut de l'espace des descriptions. Les réponses possibles au test correspondent aux étiquettes des arcs issus de ce nœud. Les feuilles sont étiquetées par une classe appelée classe par défaut. Chaque nœud interne ou feuille est repéré par sa position : la liste des numéros des arcs qui permettent d'y accéder depuis la racine. **[50].**

Conclusion

Dans ce chapitre on a parlé sur l'apprentissage automatique, on a introduit les deux typer & et on a basé sur l'apprentissage superviser et ses méthodes, spécifiquement sur les méthodes de KNN et SVM qui sont la base de notre travail.

Ce chapitre est une porte vers le chapitre suivant où commencera à parler de la façon dont nous avons utilisé ces méthodes dans notre travail et comment nous avons réussi à utiliser l'apprentissage automatique pour nous donner le meilleur résultat possible.

Le chapitre suivant présente la conception du projet et la méthodologie suivie pour la réalisation de système.

Chapitre 4 : Conception du Système

4 Chapitre 4 : Conception du Système

4.1 Introduction

Dans ce chapitre nous présenterons notre approche pour l'élaboration d'une approche contextuelle pour l'indexation automatique des textes arabes non structurés. Qui comporte trois étapes. Premièrement : le processus de prétraitement qui se divise en deux sous étapes : prétraitement linguistique et prétraitement statistique. Deuxièmes : une formalisation des KNN pour l'application à la classification de texte. Troisièmes étapes : une autre formalisation des SVM pour la classification.

4.2 Architecteur du système

1. **Collection des textes arabes**
2. **Prétraitement des textes (prétraitement linguistique, prétraitements statistique)**
3. **Classification des textes**
4. **Evaluation**

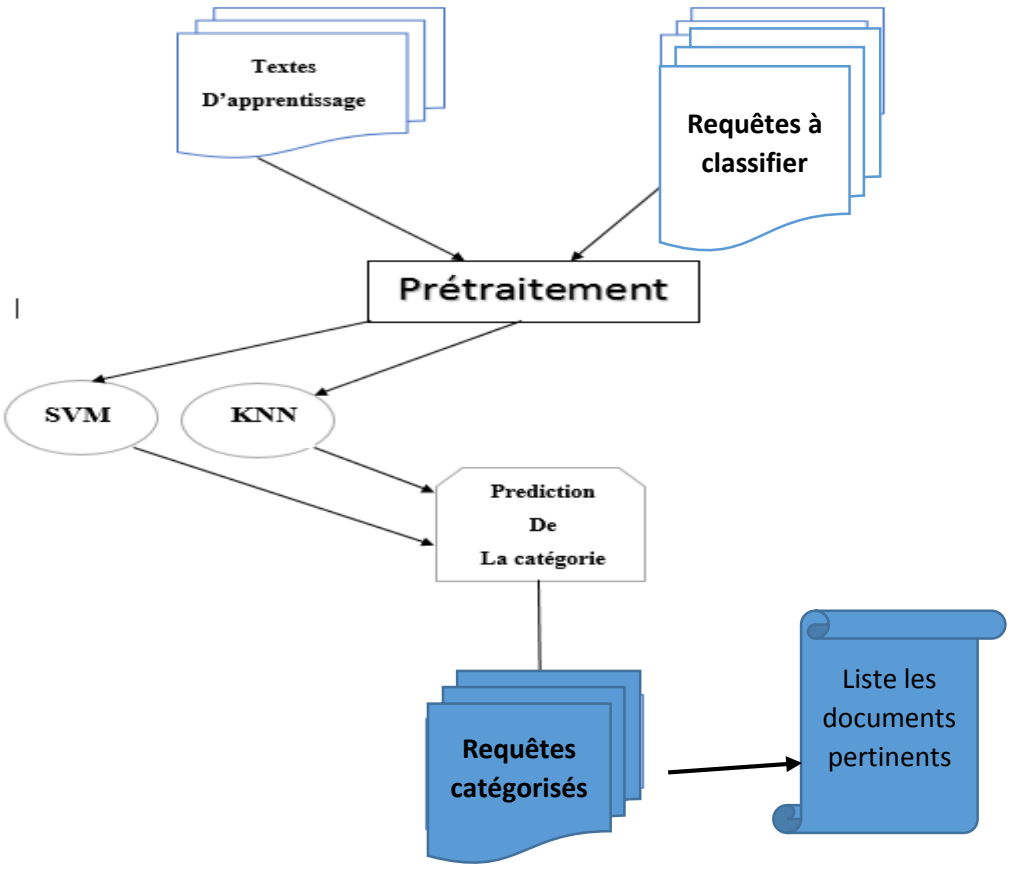


Figure 6:Architecture Globale du Système [51].

On a trois grandes étapes dans notre système :

- **L'apprentissage**

  L'apprentissage automatique s'intéresse à la faculté d'apprendre à effectuer une tache à partir de l'observation d'un environnement donc il commence généralement par un ensemble de données bien défini et une certaine compréhension de la façon dont ces données sont classifiées, dans cette optique, plusieurs algorithmes mis au point pour des problèmes quelconques en apprentissage automatique ont été adaptés et appliqués dans notre domaine de recherche. L'objectif est de trouve une liaison entre les besoins exprimés par l'utilisateur et l'ensemble des données classifiées, que l'on appelle également modelé de prédiction, dans notre travail l'algorithme d'apprentissage utilisé est les SVMs et les KNNs **[47].**

- **La prédiction de la classe d'un nouveau texte (requête)**

  En appliquant le modèle de prédiction générée dans la phase d'apprentissage pour prédire la classe de ce texte. Le modèle, en entrée, reçoit une requête et en sortie, lui associe une étiquette (classe).

- **Afficher les documents les plus pertinentes Trier :**

  Après la prédiction on prend les documents est en fait le trier, du document le plus pertinent au le moins pertinent.

4.2.1 Collection du corpus :

Notre corpus utilisé pour les tests, un ensemble de données « RTAnews » **[51]** est une collection de textes arabes multi-étiquettes, collectés à partir du portail d'informations « Russia Today in Arabic ». Il se compose de 23 837 textes répartis en 40 catégories, et divisés en 15 001 textes pour l'apprentissage et 8 836 textes pour le test. L'ensemble de données sont des textes arabes non prétraités et en format (pdf).

4.2.2 Les prétraitements des textes arabes

Tout texte soumis au système doit être traité dans le but d'en extraire les fréquences des occurrences des termes (figure 7).

L'idée consiste à représenter les textes dans un espace approprié.

Elle comporte des processus de **[47]** :

- **Prétraitement linguistique**: Élimination des mots inutiles (mots vides, caractères spéciaux,…) et extraction des radicaux (lemmatisation).
- **Prétraitement statistique**: calcule les poids des attributs des vecteurs représentant les textes en utilisant la pondération TF-IDF.

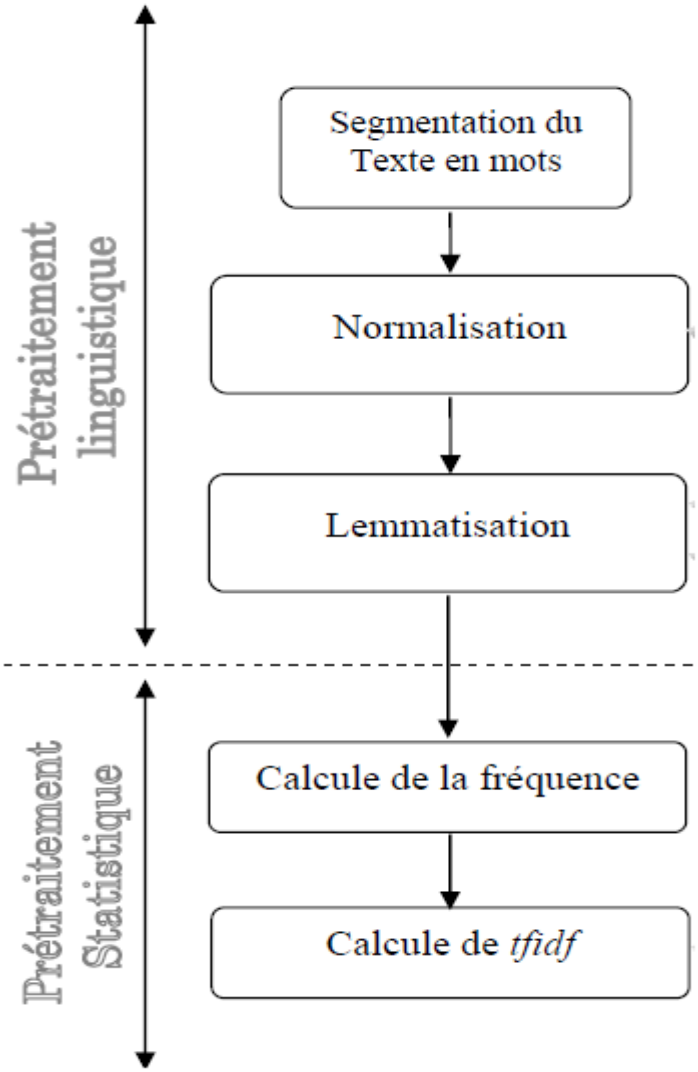


Figure 7:Processus de Prétraitement [47*]*

4.2.2.1 Prétraitement linguistique

4.2.2.1.1 Analyse Lexicale (Tokenisation/Segmentation)

La segmentation que l'on peut l'appeler aussi tokenisation de texte ou analyse lexicale permet de couper chaque phrase du texte en mots. Dans notre approche nous avons

opté à une segmentation en tokens pour que des traitements ultérieurs comme la recherche d'information puissent s'appliquer **[52].**

Exemple:

Texte arabe avant la segmentation:

(أول ما يمكن ملاحظته من خلال سردنا لهذه المصطلحات الخاصة)

Texte arabe après la segmentation:

(أول ،ما ، يمكن ، ملاحظته ، من ، خلال ، سردنا ، لهذه ، المصطلحات ، الخاصة)

4.2.2.1.2 Élimination des mots vides

Les mots vide (stop words) correspondent aux termes non porteurs d'information utile c'est-à-dire ce sont des mots fréquents qui ne sont pas porteur de sens, cette étape consiste à éliminer tous les mots vides. Ces mots sont généralement des pronoms personnels, des articles ou des conjonctions comme **[28]** :
("في"،"كل"،"لم"،"لن"،"من"،"هي"،"كما"،"لها"،"لكن"،"هي"،"يكون").

Exemple (figure 8):

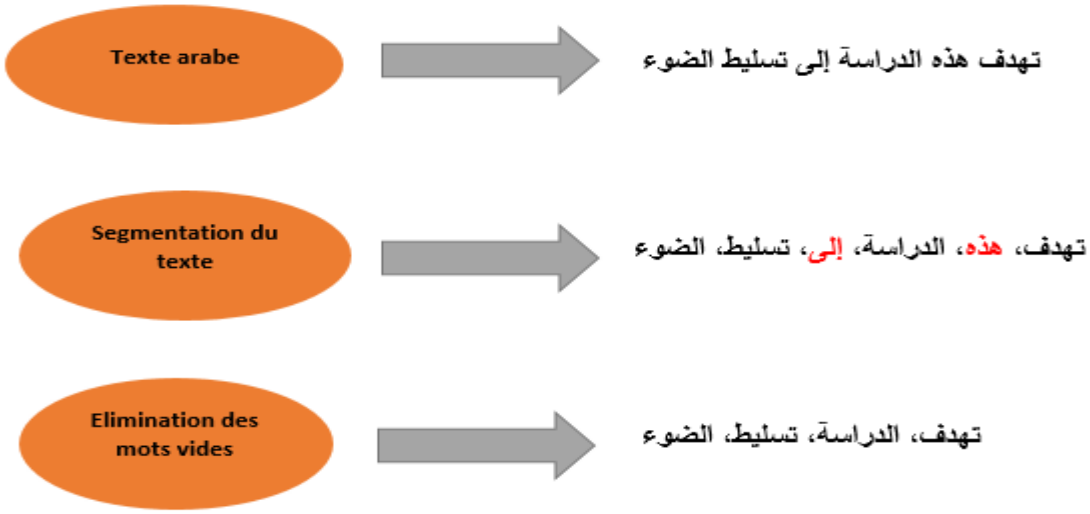


Figure 8:Exemple d'élimination des mots vide.

4.2.2.1.3 Normalisation

La normalisation est un processus morphologique consisté à transformer une copie du document original dans un format standard plus facilement manipulable. Dans l'arabe écrit, certaines lettres subissent une simple modification dans l'écriture qui n'influe pas considérablement sur le sens du mot par exemple, le mot « أمر » est généralement écrit « امر ».Donc on fait la conversion des caractères qui a pour but de normaliser les lettres qui peuvent être écrites sous plusieurs formes :

- « إ » «أ» et « آ» sont remplacés par « ا»
- « ة» est convertie en « ه»
- « ئ » ،«ي» en « ى».

La deuxième conversion est l'élimination des signes diacritiques, les signes diacritiques sont ajoutés au-dessus ou en dessous des lettres arabes afin de spécifier la prononciation du mot, enlever la ponctuation et les chiffres et retirer les non lettrés arabes **[28].**

4.2.2.1.4 Racinisation (Stemming)

Cette technique consiste à détecter la racine d'un mot **[34]**.

Exemple:

Mot (يعملون) ⇨⇨⇨⇨ stem (عمل)

Mot (المراجع) ⇨⇨⇨⇨ stem (مرجع)

4.2.2.2 Prétraitement statistique :

4.2.2.2.1 Modèle vectoriel « calcule de TF-IDF » :

La représentation d'un ensemble de documents en tant que vecteurs dans un espace vectoriel commun est connue sous le nom de modèle d'espace vectoriel et est fondamentale pour une multitude d'opérations de recherche d'informations allant de l'évaluation des documents sur une requête, la classification de documents et le regroupement de documents**[53].** Dans le modèle d'espace vectoriel, le contenu d'un document est représenté par un vecteur d'espace multidimensionnel. La procédure du modèle d'espace vectoriel peut être divisée en trois étapes:

1. La première étape est l'indexation des documents, lorsque les termes les plus pertinents sont extraits.

2. La deuxième étape est basée sur l'introduction de pondérations associées au terme d'index afin d'améliorer la recherche pertinente pour l'utilisateur.

3. La dernière étape classe le document avec une certaine similitude.

La pondération des termes est l'une des méthodes de prétraitement; utilisées pour la présentation améliorée des documents texte en tant que vecteur de caractéristiques. La pondération des termes nous aide à localiser les termes importants dans une collection de documents à des fins de classement **[53].** Pour cela on a utilisé la méthode tf-idf .Celle-ci pour calculer le poids sachant que, pour chaque terme, il est possible de calculer non seulement sa fréquence dans le corpus, mais aussi le nombre de documents contenant ce terme.

- **Représentation du document**

  Les documents sont représentés par des vecteurs de caractéristiques. Un document est représenté sous forme de vecteur - une séquence d'entités et leurs poids. Le modèle de sac de mots le plus courant utilise simplement tous les mots d'un document comme entités **[54].**

  On utilise cette représentation vectorielle et le produit interne pour trouver les documents les plus pertinents après on fait le trier du document le plus pertinent au le moins pertinent

4.2.3 Classification des documents « avec KNN et SVM »

La classification est de classer de façon automatique les documents dans des catégories qui ont été définies soit préalablement par un expert, il s'agit alors de classification supervisée ou catégorisation, soit de façon automatique.

4.2.3.1 Apprentissage avec KNN

La réalisation d'un programme d'apprentissage par KNN, elle est constituée de trois éléments:

1) l'échantillon d'apprentissage(les classes du corpus)

2) la distance (utilisant les fréquents des mots)

3) la méthode de combinaison des voisins (déterminer la liste des k vecteurs les plus proches et déterminer la classe la plus représentée dans la liste des K-ppv)

Avec la distance de manhattan en calculer la distance d(w_i , w_{ij}) entre chaque fréquence de terme w_i de la requête (la fréquence de terme i dans la requête) et la fréquence w_{ij} (la fréquence de terme i dans le document j) , après le choix du k en fait la sélection des k qu'ont les plus petites distances entre les documents indique leur plus grande similitude avec la requête et à la fin en déterminer la classe la plus représentée dans la liste des K-ppv .

4.2.3.2 Apprentissage avec SVM

Pour chaque document d on prendre les fréquences (tf-idf) des termes représentatifs par rapport à la requête.

Pour l'étape d'apprentissage on prend en entrée les fréquences des termes de chaque document de chaque classe qui sont présentées en 1-dimension et avec l'utilisation de l'astuce du noyau (karnel) avec la méthode polynomiale de 2 degrés $\emptyset(X) = (X, X^2)$ parce que les données sont non linéairement séparables donc on peut avoir un couple de chaque fréquence et on fait la même chose pour la requête .

Pour les classes le SVM ne pouvant résoudre que des problèmes de deux classes par contre nous avons multi-classes donc on a utilisé la méthode un contre un (one-vs-one) la stratégie semble plus adaptée à une utilisation pratique **[35]**

Avec l'utilisation la bibliothèque sklearn sur python pour le SVM en peut faire la prédiction de la classe.

4.2.4 Évaluation

Il existe de nombreuses normes d'évaluation dans la recherche d'informations utilisées dans regroupement de documents tels que « Entropie », Cluster Purity ,matrice de confusion et F-mesure **[55].**

- La précision (P) est la fraction des documents récupérés qui sont pertinents
- Rappel (R) est la fraction des documents pertinents qui sont récupérés

Ces notions peuvent être clarifiées en examinant le tableau de contingence suivante (tableau 14): **[56]**

Tableau 14 : matrice de confusion.

	Pertinent	**Non Pertinent**
Récupéré	Vrai positif (VP)	Faux positif (FP)
Non Récupéré	Faux Négatif (FN)	Vrai Négatif (VN)

1. le nombre de textes correctement classé comme appartenant à la classe, noté VP (pour Vrai Positif) ;

2. le nombre de textes incorrectement classé comme appartenant à la classe, noté FP (pour Faux Positif) ;

3. le nombre de textes incorrectement rejetés, noté FN (pour Faux Négatif) ;

4. le nombre de textes correctement rejetés, noté VN (pour Vrai Négatif).

Donc :

$$Précision = VP / (VP + FP)$$

$$Rappel = VP /(VP + FN)$$

Conclusion

Dans ce chapitre, nous avons expliqué tous les mécanismes que nous avons mis dans notre application, en nous concentrant sur la façon dont nous avons réussie à fixer les textes arabes et comment nous avons fait les calculs qui nous donneront les résultats.

Dans le prochain chapitre se concentrera sur la façon dont ces termes se transforment en une véritable application et comment nous avons réussi à obtenir les meilleurs résultats.

Chapitre 5 :

Implémentation et Test

5 Chapitre 05 : Implémentation et Test

5.1 Introduction

Dans ce chapitre, nous allons présenter l'implémentation de notre système de recherche. Premièrement on commence par la présentation de l'environnement de développement, en détaillant les différents outils utilisés, après on explique le déroulement de l'application, et enfin on interprète et on commente les résultats obtenus.

5.2 Les outils de développement

Le choix de l'environnement de programmation convenable est très important pour le développement des projets. Cela se fait suivant plusieurs facteurs : la puissance de compilation, la facilité d'utilisation, la disponibilité de plusieurs fonctionnalités, la communication avec d'autres environnements… etc.

L'outil que nous avons adopté est PYTHON, notre choix s'est porté sur cet outil car il porte beaucoup de bibliothèque de traitement des textes qui nous avons besoins pour faire ce projet.

Pour implémenter notre système nous avons utilisé les outils suivants

5.2.1 Le langage python

Python est un langage de script de haut niveau, structuré et open source. Ila était créé au début des années 1990 par Guido van Rossum de Stichting Mathematisch Centrum aux Pays-Bas pour succéder à un langage appelé' ABC. Guido reste l'auteur principal de Python, bien qu'il inclue de nombreuses contributions d'autres.

Python est un langage orienté objet, il supporte l'héritage multiple et la surcharge des opérateurs. Dans son modèle objets, et en reprenant la terminologie de C++, toutes les méthodes sont virtuelles. Il est réputé par la rapidité de développement et très apprécié pour la clarté et simplicité de sa syntaxe, ce qui oppose à d'autres langages, en prenant exemple le langage Perl. Un programme Python est souvent de 3 à 5 fils plus court qu'un programme C ou C++ (ou mémé Java), ce qui représente en général un temps de développement de 5 2-1 10 fois plus court et une facilité de maintenance largement accrue **[57]**.

La bibliothèque standard de Python, et les paquetages contribués, donnent accès à une grande variété de services : chaines de caractères et expressions régulières, services UNIX standard (fichiers, pipes, signaux, sockets, threads...), protocoles Internet (Web, News,…), persistance et bases de données, interfaces graphiques.

Python est un langage qui continuée d'évoluer, soutenu par une communauté d'utilisateurs enthousiastes et responsables, dont la plupart sont des supporteurs du logiciel libre. Parallèlement a |'interpréteur principal, écrit en C et maintenu par le créateur du langage, un deuxième interpréteur, Python écrit en Java, est en cours de développement **[58]**.

5.2.2 Anaconda

C'est une distribution des langages de programmation Python et R pour la science des données et l'apprentissage automatique (figure 9).

Maintenant, si vos faites principalement du travail de science des données, Anaconda est également une excellente option. Anaconda est créé par continuum Analytiques, et il s'agit d'une distribution Python préinstallée avec de nombreuses bibliothèques Python utiles pour la science des données.

Anaconda est populaire car il intègre de nombreux outils utilisés dans la science des données et l'apprentissage automatique en une seule installation. Il est donc idéal pour une configuration courte et simple.

Comme Virtualenv, Anaconda utilise également le concept de création d'environnements afin d'isoler différentes bibliothèques et versions. Anaconda introduit également son propre gestionnaire de paquets, appelé conda, à partir duquel vous pouvez installer des bibliothèques.

De plus, Anaconda a toujours l'interaction utile avec Pip qui vous permet d'installer toutes les bibliothèques supplémentaires qui ne sont pas disponibles dans le gestionnaire de packages Anaconda **[59].**

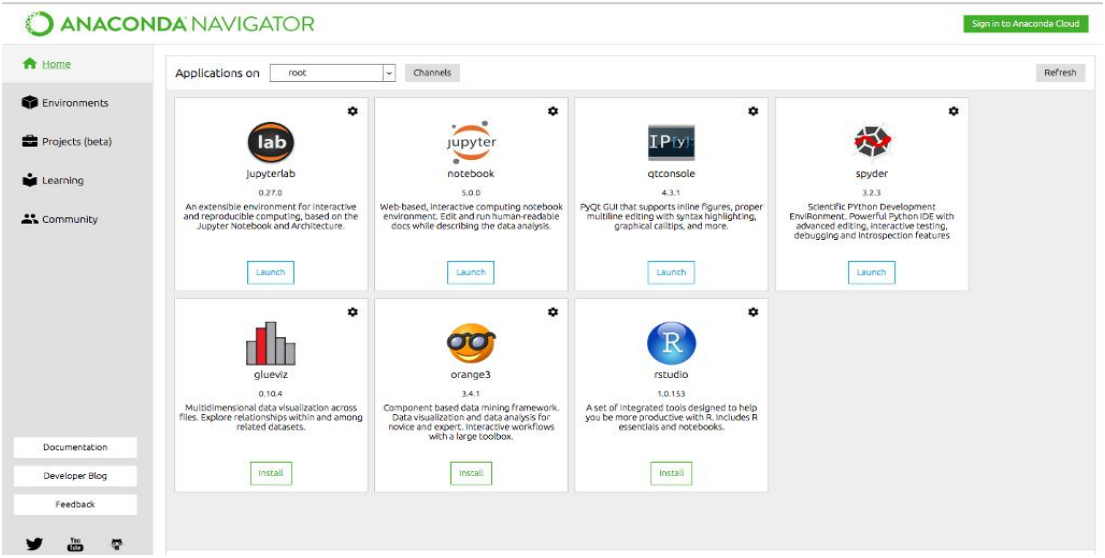


Figure 9:Le navigateur anaconda de python [59].

5.2.3 Spyder

Spyder est un environnement scientifique puissant écrit en Python, pour Python conçu par et pour les scientifiques, les ingénieurs et les analystes de données. Il offre une combinaison unique des fonctionnalités avancées d'édition, d'analyse, de débogage et de profilage d'un outil de développement complet avec |‘exploration de données, |'exécution interactive, |'inspection approfondie et les superbes capacités de visualisation d'un progiciel scientifique. En outre, Spyder offre une intégration intégrée a de nombreux logiciels scientifiques populaires, notamment NumPy, SciPy, Pandas, lPython, QtConso|e, Matplotlib, SymPy, etc. Au-delà de ses nombreuses fonctionnalités intégrées, les capacités de Spyder peuvent être étendues encore davantage via son système de plug-in et son API. Spyder peut également être utilisé en tant que bibliothèque d'extensions PyQt5, vous permettant de développer ses fonctionnalités et d’incorporer ses composants, tels que la console interactive, dans votre propre Logiciel (figure10) **[60].**

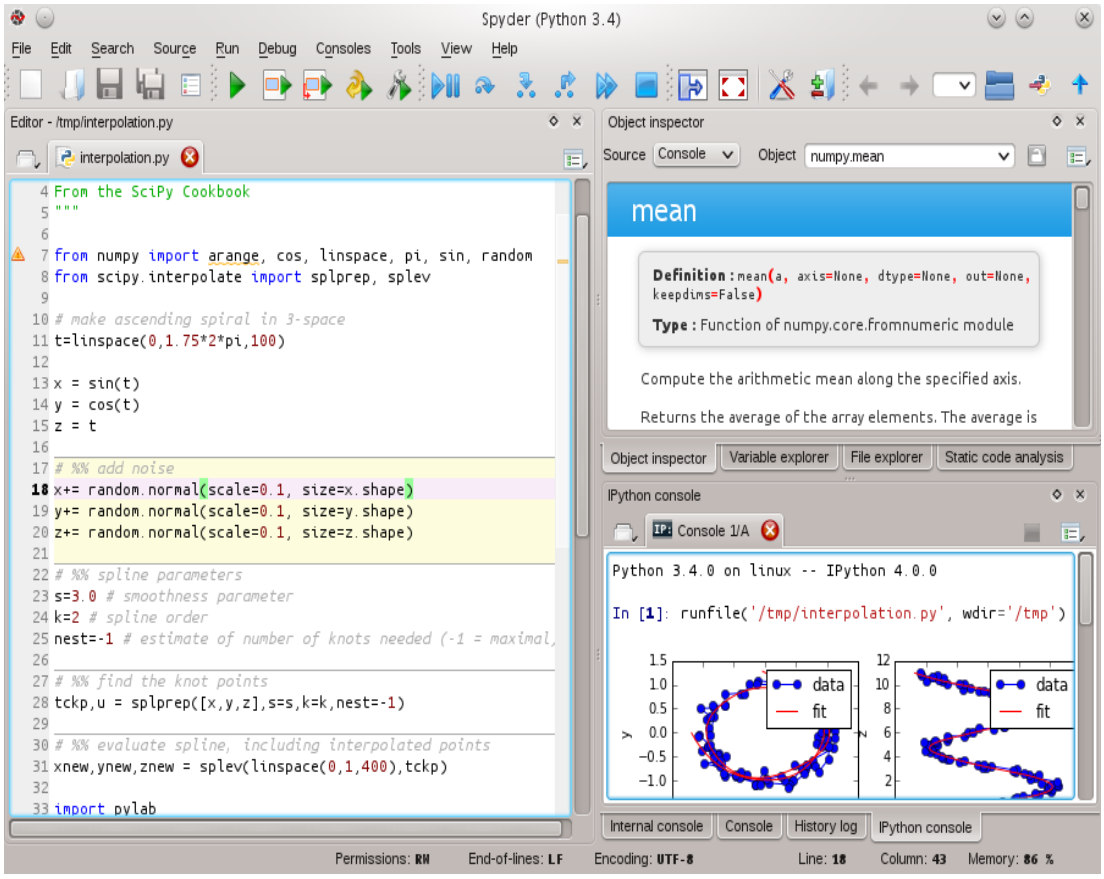


Figure 10:La plateforme spyder de python [60].

5.2.4 PYQT5 :

Qt est un ensemble de bibliothèques C ++ multiplateformes qui implémentent des API de haut niveau pour accéder à de nombreux aspects des systèmes de bureau et mobiles modernes. Il s'agit notamment des services de localisation et de positionnement, du multimédia, de la connectivité NFC et Bluetooth, d'un navigateur Web basé sur Chromium, ainsi que du développement d'interface utilisateur traditionnelle.

PyQt5 est un ensemble complet de liaisons Python pour Qt v5. Il est implémenté sous la forme de plus de 35 modules d'extension et permet d'utiliser Python comme langage de développement d'applications alternatives au C ++ sur toutes les plates-formes prises en charge, y compris iOS et Android.

PyQt5 peut également être intégré dans des applications basées sur C ++ pour permettre aux utilisateurs de ces applications de configurer ou d'améliorer les fonctionnalités de ces applications.

5.3 Description d’application

Notre application est développée en python à l’aide de l’environnement spyder, avec une interface graphique créée à travers PYQT5. Et pour réaliser ce projet on a utilisé quelque bibliothèque comme ; NLTK, Snowball stemmer …etc.

Tout cela pour que l'utilisateur puisse écrire la requête qu'il veut, afin qu'il puisse connaître la classe de son requête et les documents pertinents en fonction de cette requête , avec la possibilité de cliquer sur ces documents et de choisir celui dont il a le plus besoin.

5.3.1 Le Déroulement :

Dans cette partie, nous parlerons des étapes que nous avons franchies pour obtenir les résultats recherchés par l'utilisateur, de la saisie au tableau des textes pertinents.

Avant de sauter dans les étapes, voici l'apparence de l'interface de l'application que nous avons créée (figure 11)

Figure 11:L’interface de notre application.

5.3.2 La recherche :

Pour commencer à travailler sur cette application, vous devez cliquer sur le bouton d'accueil pour accéder à la zone de travail principal.

Tout d'abord, vous entrez votre propre requête (mot clés) dans la zone de recherche, puis vous cliquez sur Rechercher pour permettre à l'application de faire le travail (figure 12)

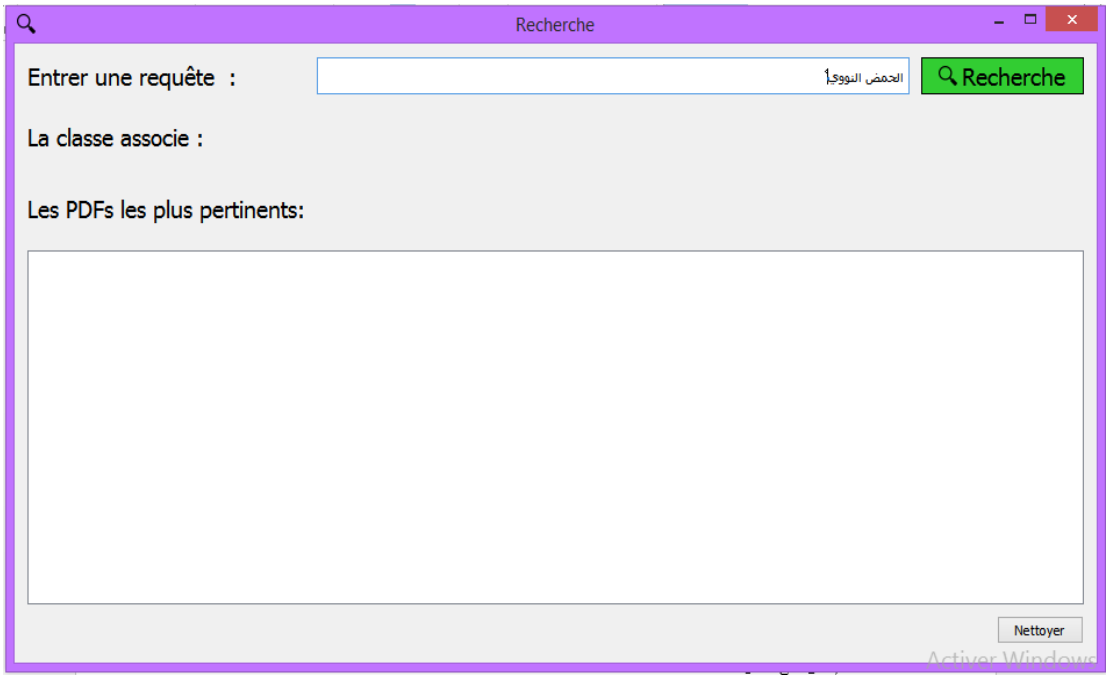


Figure 12:L'insertion de requête

5.3.3 Le Traitement

Avant d'obtenir les résultats, le texte passe par de nombreuses étapes pour être prêt pour le processus (nous avons mentionné celles au chapitre 4), après cela, il passe par les calculs de fréquences, et se termine en utilisant les algorithmes de "KNN" et "SVM" sur eux (plus de détails au chapitre 4).

Dans ces captures d'écran, nous voyons certains des codes que nous avons utilisés pour cette étape (figure13).

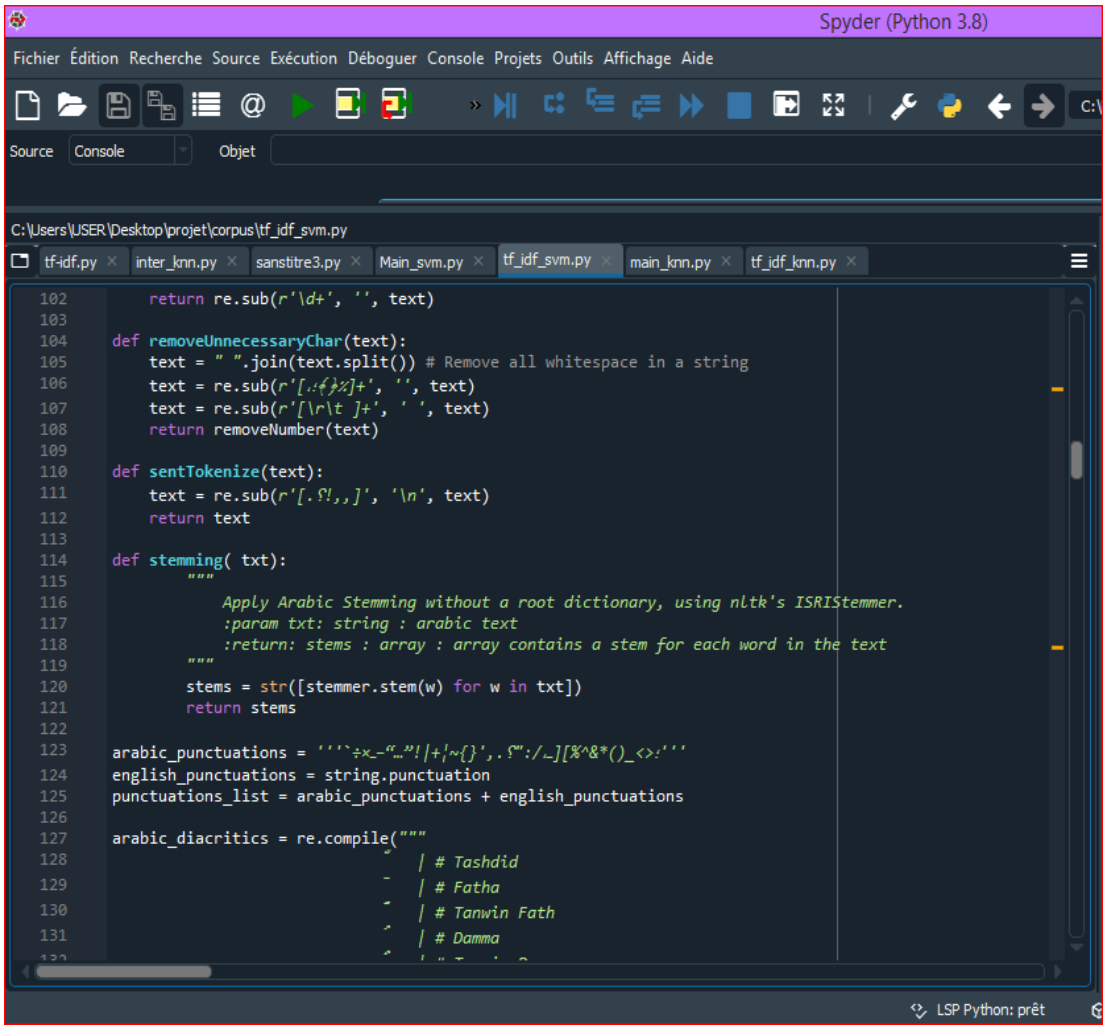


Figure 13:Code source de notre application.

5.3.4 Les résultats

Après avoir cliqué sur le bouton de recherche, l'application prend quelques secondes pour traiter tous les documents (format PDF) que vous avez sur votre ordinateur portable, après cela, elle vous donne 2 résultats (figure 14):

1- c'est la classe à laquelle votre requête est associée.

2- les documents pertinents à votre requête, qui est la partie la plus importante du résultat, ici nous trouvons la liste des documents (format PDF) classés par le plus au moins pertinent.

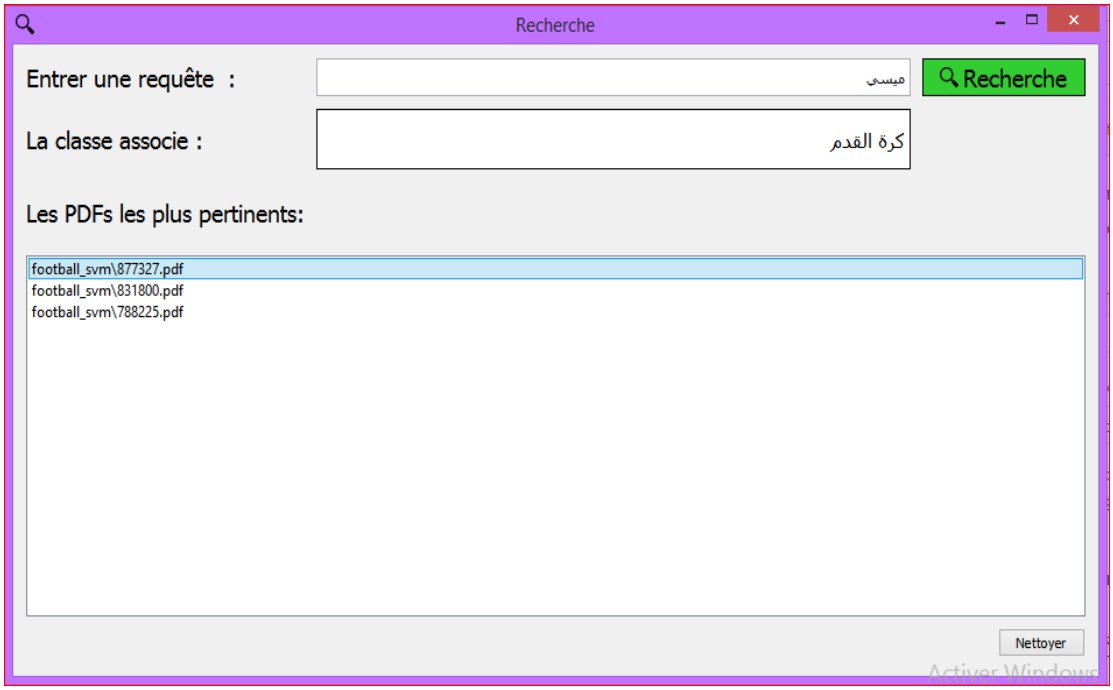


Figure 14:Les résultats de la recherche.

Lorsque vous cliquez sur le document il sera automatiquement ouvert sans besoin de parcourir le répertoire de votre ordinateur (figure 15).

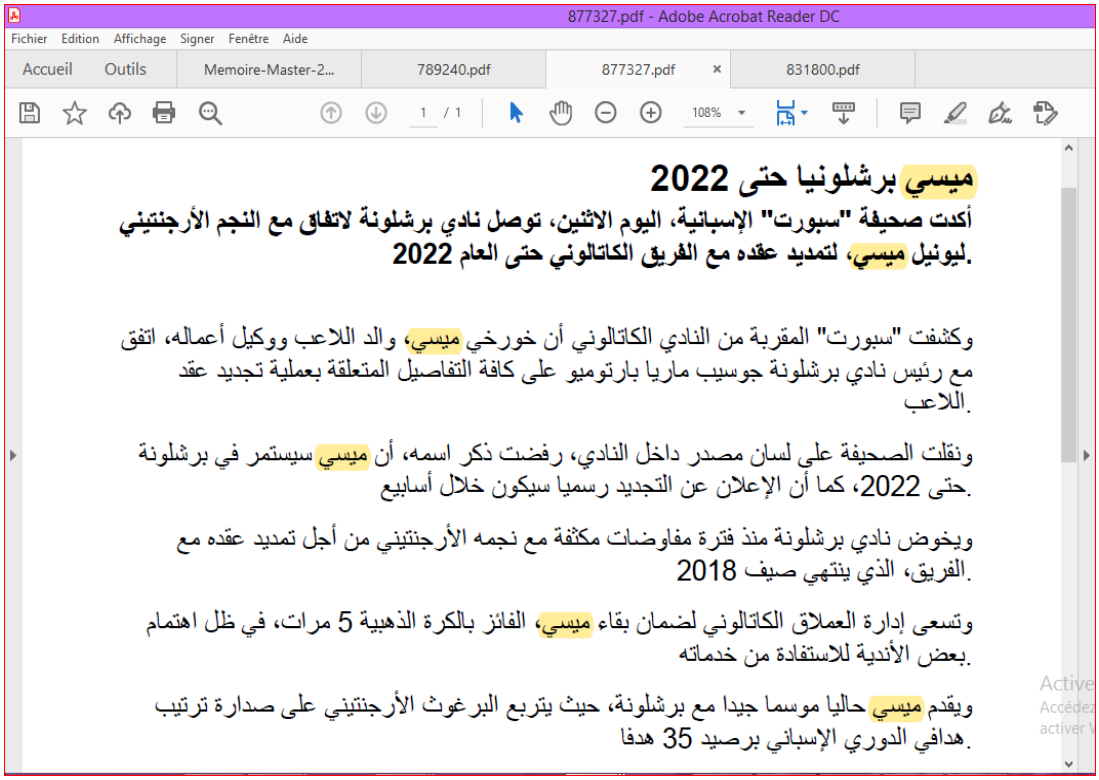


Figure 15:l'ouverture de document.

Mais il y a 2 cas particuliers:

- Quand il n'y a pas de classe et pas de documents pertinents (figure 16).

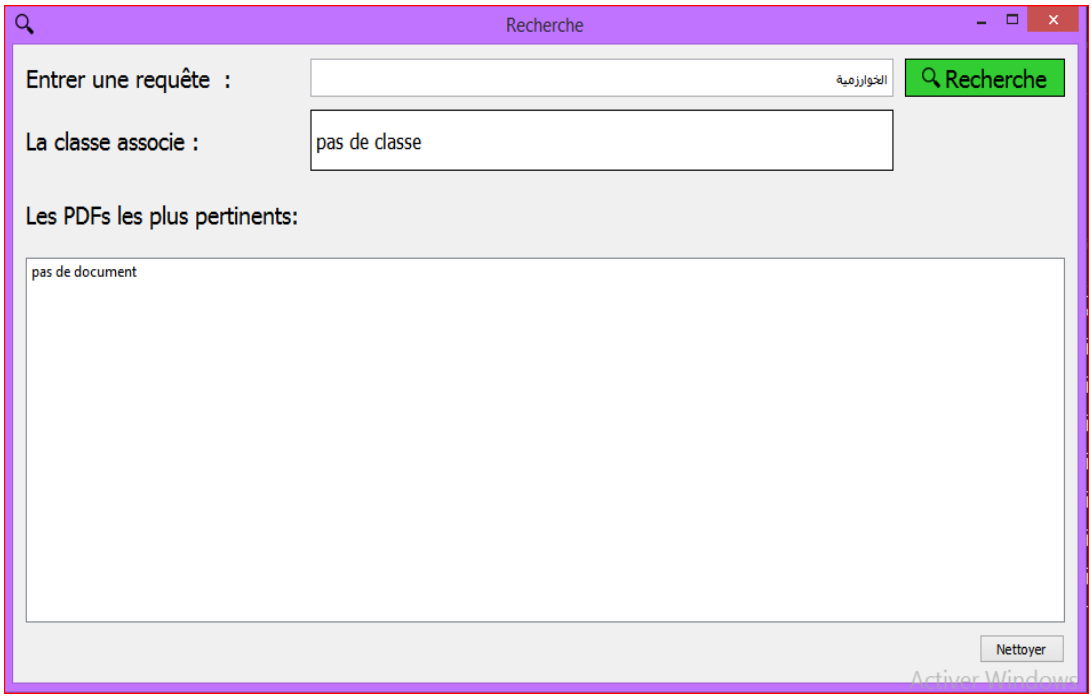


Figure 16:le cas d'aucune classe et d'aucun document.

- Lorsqu'il n'y a pas de classe mais qu'il y a des documents pertinents (figure 17).

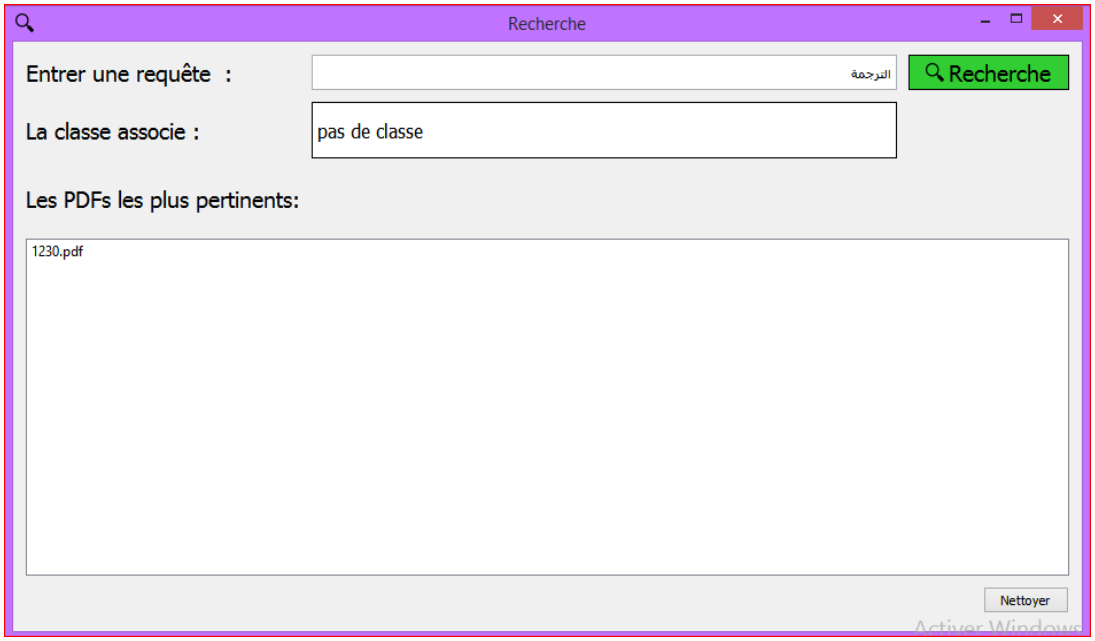


Figure 17:le cas d'aucune classe mais il y a un document pertinent.

5.4 Test et Résultats :

5.4.1 Les données de test:

Pour la réalisation des expériences, nous avons utilisé une partie du corpus « RTAnews » qui a une grande taille du 110 Mo.

Nous avons utilisé des textes arabes du corpus dans les domaines (tableau 15) : football (**كرة القدم**), Musique (**موسيقى**), Recherche médicale (**البحوث الطبية**), qui contient 120 documents comme un jeu d'entrainement(format pdf sans avoir converti en une autre format) et la construction du jeu de tests a été un travail long et difficile parce que notre objective est de construire un système de recherche d'information interne donc nous avons utilisé 60 documents du corpus des tests pour l'extraction des requêtes (des mots-clés). Pour permettre une bonne comparaison avec k-NN et SVM.

Tableau 15:nombre des documents dans chaque catégorie

Catégories	Nombre des textes d'Entrainent	Nombre des requêtes pour le test
كرة القدم	40	20
البحوث الطبية	40	20
موسيقى	40	20
Total	120	60

5.4.2 Résultats

L'expérimentation consiste à comparer les deux méthodes de classification. Nous avons subdivisé la collection en deux bases : le premier pour l'apprentissage (120 documents) et la deuxième pour le test (60 requêtes).

Le tableau (tableau 16) Montre les performances de la classification des deux méthodes Machine à vecteur de support (SVM) les k plus proches voisins (K-PPV), en termes de précision, rappel et f-Mesure :

Tableau 16:résultats du Rappel, Précision et F-mesure de multi-classes.

Nom de la catégorie	SVM			KNN		
	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure
كرة القدم	0.900	0.947	0.922	0.750	0.789	0.769
البحوث الطبية	0.950	1.00	0.974	0.850	0.850	0.850
موسيقى	0.950	0.863	0.904	0.800	0.761	0.780

Pour comparer les deux méthodes, on a calculé Rappel, Précision et F-mesure total, le tableau 17 montre les résultats obtenus.

À partir du tableau (tableau 17) on peut dire que la méthode de SVM donne les meilleures performances dans les trois catégories par rapport à la méthode de KNN.

Tableau 17:La moyenne de performance du KNN et SVM.

	Rappel	Précision	F-mesure
SVM	93.33 %	93.66 %	93.33 %
KNN	80 %	80 %	82.2 %

Ces résultats sont présentés dans la figure (figure 18).

À partir de figure et du tableau on remarque les meilleurs résultats sont donnés par la méthode SVM avec un rappel, une précision et une F-mesure supérieure à 90%.

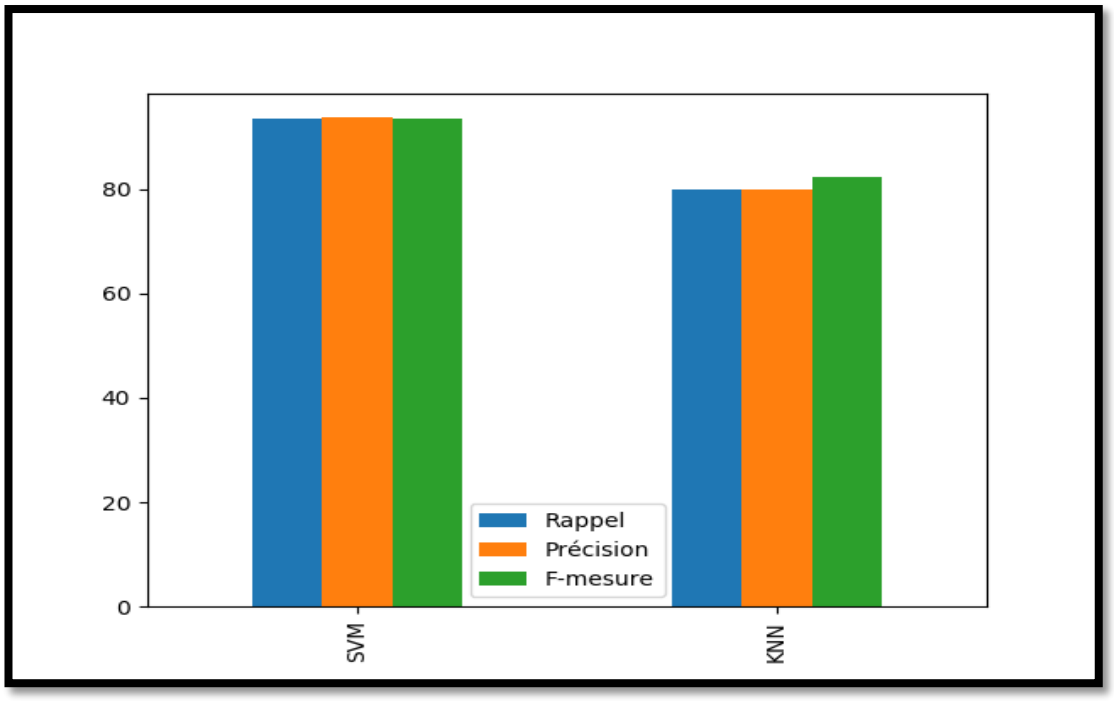


Figure 18:Rappel, Précision et F-mesure du KNN et SVM

5.4.3 Discussion des résultats :

Les expérimentations montrent qu'une bonne indexation de textes donne des bons résultats de la performance des méthodes d'apprentissage automatique, c'est-à-dire le succès de ces méthodes est justifié par un bon processus qui consiste à définir une représentation du corpus par des termes (sac à mots) après un prétraitement linguistique (segmentation, élimination des mots inutiles, extraction des radicaux…) du texte et un prétraitement statistique (calculer les attributs des vecteurs, on utilise la méthode TFIDF) .

Pour l'étape d'apprentissage utilisant la méthode SVM "Support Vector Machine" et la méthode KNN (k Nearest Neighbors) montre le succès de ces méthodes par le calcul d'évaluation (Rappel, précision et F-mesure).

Les résultats obtenus montrent que l'utilisation des SVM donne des bons Résultats par rapport au KNN.

Conclusion

Au cours de ce chapitre, nous avons présenté l'environnement de développement de notre système ainsi que les différentes interfaces graphiques qui a travers lesquelles nous pouvons superviser les différents traitements du système.

Notre système a pour rôle d'implémenter les deux méthodes d'apprentissage automatique SVM et KNN a des textes arabes non structurés après une indexation.

On conclura d'après les résultats obtenus, que la méthode SVM donne des bons résultats.

Conclusion Générale

Trouver ce que vous avez besoin n'a jamais été une tâche facile, surtout si vous recherchez un contenu dans l'un des mille documents que vous avez sur votre ordinateur, et pour compliquer la tâche pour les programmeurs et les chercheurs , le contenu est en arabe et nous savons tous à quel point il est difficile de traiter ça.

Dans cette étude nous avons présente le rôle d'implémenter des deux méthodes d'apprentissage automatique SVM et KNN a des textes arabes non structurés après une indexation. Nous avons utilisé notre propre corpus ; le corpus est composé de 120 documents appartenant à 3 catégories et de 60 requêtes. Les documents et les requêtes de l'ensemble de données ont été prétraités en supprimant les mots vides et les normalisés, puis les documents ont été représentés à l'aide du modèle d'espace vectoriel basé sur l'ensemble des mots-clés extraits de chaque document et le schéma de pondération normalisé tf × idf. Pour mesurer la similitude entre deux documents, la mesure de produit scalaire a été utilisée. Enfin, pour évaluer et mesurer l'efficacité de l'indexation, nous avons utilisé des deux méthodes d'apprentissage automatique SVM et KNN. Les résultats montrent que SVM et kNN est applicable au texte arabe ; en utilisant kNN et SVM sur notre ensemble de données, nous avons obtenu des résultats très satisfaisants.

Enfin, nous proposons comme perspectives :

- Ajouter d'autres langues pour rendre le système Multilingue afin que le programme puisse comprendre les requêtes d'autres langues plutôt que l'arabe.
- Utiliser les ontologies (wordnet) ou les dictionnaires pour enrichir la représentation des textes et les requêtes.
- Prendre en considération la complexité pour un accès aux données d'une façon précise et rapide.

References

[1]Mr Abderrezak BRAHMI (2013) [”Contribution à la Recherche Intelligente sur le Web : Indexation Sémantique des Textes Non-Structurés”, Thèse de Doctorat en Sciences en Informatique, Université des Sciences et de la Technologie , Oran]

[2] Lancaster F.-W (1998) [Indexing and abstracting in theory and practice. Library Association Publishing. London]

[3] Michèle Hudon (1998) [Indexation et langages documentaires dans les milieux archivistiques à l’ère des nouvelles technologies de l’information]

[4] Bruno Bachimont (2007) [Ingénierie des connaissances et des contenus : le numérique entre ontologies etdocuments, Lavoisier, Hermès]

[5] GASMI Mounira (2009) ["Utilisation des ontologies pour l’indexation automatique des sites Web en Arabe ", Mémoire Présenté pour l’obtention du diplôme de MAGISTER Spécialité : Informatique, UNIVERSITE KASDI MERBAH OUARGLA.]

[6]Catherine Roussey (2010) ["Une méthode d'indexation sémantique adaptée aux corpus multilingues ", thèse de Doctorat Laboratoire d'Ingénierie des Systèmes d'Information (LISI) de l'INSA de Lyon]

[7]Haïfa Zargayouna (2005) ["Indexation sémantique de documents XML", thèse de Doctorat, Université Paris XI]

[8]Nathalie Hernandez et al. (2008) [RI et Ontologies – Etat de l’art 2008, CNRS, INP Toulouse, Université Paul Sabatier, N° IRIT/RR—2008-14—FR]

[9]Domingos Ruiz Lepores (2011) ["Des grandes classifications au Web de données et l’émergence de l’indexation sémantique: le cas du tagging sémantique dans le portail», mémoire présenté pour obtenir le Titre professionnel "Chef de projet en ingénierie documentaire" INTD niveau I en 8 décembre 2011, École Management et Société- Département CITS, INTD.]

[11]Isra HAMADA (2013) [”Utilisation de "WordNet" pour indexation sémantique & recherche d’information”,mémoire présenté par FACULTE DES SCIENCES & TECHNOLOGIE,DÉPARTEMENT DE MATHÉMATIQUES & INFORMATIQUE ,UNIVERSITÉ Dr. MOULAY TAHAR , SAIDA]

[12] Siham Boulaknadel (2008) [” Traitement Automatique des Langues et Recherche d’Information en langue arabe dans un domaine de spécialité :Apport des connaissances morphologiques et syntaxiques pour l'indexation”,THÈSE DE DOCTORAT en Informatique, Université de Nantes, France]

[13] A. W. Saad M., (2010) ["Arabic Text Classification Using Decision Trees," presented at the Workshop on computer science and information technologies CSIT‟2010, Moscow - Saint-Petersburg, Russia]

[14] C. D. Manning and P. Raghavan, (2009) ["An Introduction to Information Retrieval Draft," Online edition. Cambridge University Press]

[15] M. Lan, C. Tan, H. Low, and S. Sung, (2005)["A comprehensive comparative study on term weighting schemes for text categorization with support vector machines," presented at the Special interest tracks and posters of the 14thinternational conference on World Wide Web, Chiba, Japan]

[16]Hachemi Hadjira,Rimouche Nour El Houda (2013) [”Moteur de recherche Sémantique”,mémoire présenté par Faculté des Sciences,Département d’Informatique, Université Abou Bakr Belkaid, Tlemcen]

[17] LAURE SOULIER (2014) [“Définition et évaluation de modèles de recherche d’information collaborative basés sur les compétences de domaine et les rôles des utilisateurs”, Thèse de DOCTORAT, L’UNIVERSITÉ DE TOULOUSE, France]

[18] ABDERRAHIM Mohammed Alaeddine (2016) [” Exploitation des Ontologies dans les Systèmes de recherche d’informations Arabes”, Thèse de DOCTORAT, Université Aboubakr Belkaïd, Tlemcen]

[19]BENZATER Nebia (2014) [”Anallyse Morphollogiique du Textte Arabe pour Son Indexattiion Sémanttiique”,mémoire pour obtention du diplôme de Magistère en

Informatique, Université des Sciences et de la Technologie - Mohamed Boudiaf – Oran]

[20] Fouad Soufiane Douzidia (2004) [" Résumé automatique de texte arabe" ,Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Scen informatique ,Université de Montréal]

[21]Aïda KHEMAKHEM (2006) ["ArabicLDB : une base lexicale normalisée pour la languearabe" mémoire présenté en vue de l'obtention du diplôme de MASTER en Systèmesd'Information et Nouvelles Technologies, Université de Sfax, Faculté desSciences Economique et de Gestion, Tunisie]

[22]Ahmed Al-hamlawi, (2005) [" شذا انعزف في فنّ انصّزف ", livre publié en 17 décembre 2005]

[23]Mostapha Al-Glayini, (2007) [" جامع انذروس انعزبية " , livre édité en Bierut, Lebanon]

[24]LABIADALI (2017) [''SÉLECTION DES MOTS CLÉS BASÉE SUR LA CLASSIFICATION ET L'EXTRACTION DES RÈGLES D'ASSOCIATION",L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES]

[25]S. Khoja, R. Garside, and G. Knowles. (2001) [A Tagset for the Morphosyntactic Tagging of Arabic". Proceedings of the Corpus Linguistics. Lancaster University (UK)]

[26] Saïd OUATIK EL ALAOUI(2015) [Approches statistique et sémantique pour l'accès à l'information dans les collections textuelles en langue arabe [thèse de doctorat,UNIVERSITÉ MOHAMMED V – AGDAL, Rabat, maroc]

[27]Gherabi Sara (2014) ["CLASSIFICATION AUTOMATIQUE DES TEXTES ARABE (ARABIC OPINION POLARITY)",mémoire présenté par FACULTÉ DES MATHÉMATIQUES ET DE L'INFORMATIQUE,Département d'informatique, UNIVERSITÉ DE M'SILA]

[28]Slim MESFAR, (24 Novembre 2008) [" Analyse Morpho-Syntaxique Automatique et Reconnaissance des entités nommées En Arabe Standard", thèse de Doctorat, Université De Franche-Comté.]

[29]F. Debili, H. Achour et E. Souissi. (2002) [De l'étiquetage grammatical à la voyellation automatique de l'arabe. In Correspondances IRMC : Institut de Recherche sur le Maghreb Contemporain, volume vol 71, pages 10–26, Tunis]

[30]L. Hadrich Belguith, L. Baccour et M. Ghassan. (2005) [Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2005 - Dourdan France, vol. Vol. 1, pages 451–456]

[31]Atef Ben Youssef (01 juillet 2008) [" Méthodes Mixtes pour la Traduction Automatique Statistique " mémoire présenté en vue de l'obtention du diplôme de MASTER 2 en Modélisation et traitements automatique en Industries De la Langue : parole, écrit, apprentissage Orientation Recherche, Université STENDHALGrenoble3, Laboratoire d'informatique de Grenoble Équipe GETALP]

[32]Wajdi Zaghouani,(2008) [" Le repérage automatique des entités nommées dans la langue arabe : vers la création d'un système à base de règles", Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.A. en linguistique, Université de Montréal]

[33]P. Schauble. (1997) [Multimedia Information Retrieval : content-based Information Retrieval from Large Text and Audio Databases. Kluwer Academic Publishers]

[34]S. Khoja and R. Garside (1999) [Stemming Arabic text. Computing Department, Lancaster University, Lancaster]

[35] Liu, Z., Peng, L., Yabin, Z., & Maosong, S. (2009) [Clustering to find exemplar terms for keyphrase extraction ,Actes de la conférence 2009 sur les méthodes empiriques dans le traitement du langage naturel (pp. 257-266).]

[36] Saeed Raheel and Joseph Dichy, (2010) [An empirical study on the feature qs type effect on the automatic classification of Arabic documents, présenté de la 11eme Conférence internationale sur la linguistique informatique et le traitement intelligent de texte, CICLing'80 (pp. 673–686).]

[37] Thabtah F., Eljinini, M., Zamzeer, M., & Hadi, W. (2009) [Naïve bayesian based on Chi square to categorize Arabic data. Présenté à la 11eme conférence de l'International Business Information Management Association (IBIMA) sur l'innovation et la gestion des connaissances dans les économies à deux voies, IBIMA'2009 (pp. 930–935)]

[38]Bawaneh, M. J., Alkoffash, M. S., & Al Rabea, A. I. (2008) [Arabic text classification using K-NN and Naive bayes. Journal of Computer Science, 4, 600-605]

[39] Kanaan G., Yaseen, M., Al-Shalabi, R., Al-Sarayreh, B., & Mustafa, A. (2009) [Using EM for text classification on Arabic. « Actes de la deuxième Conférence internationale sur les ressources et outils de la langue arabe »]

[40]Jamoussi, S. (2009) [Une nouvelle représentation vectorielle pour la classification sémantique.p50]

[41] Jean-Louis Monino et Soraya Sedkaoui (1er février 2016) [*Big Data, Open Data et valorisation des données*, Londres, ISTE Éditions p158]

[42]LABIADALI, (2017) [''SÉLECTION DES MOTS CLÉS BASÉE SUR LA CLASSIFICATION ET L'EXTRACTION DES RÈGLES D'ASSOCIATION", L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES]

[43]Berrani, S.-A., Amsaleg, L., & Gros, P. (2002) [Recherche par similarité dans les bases de données multidimensionnelles : panorama des techniques d'indexation. Ingénierie des systèmes d'information (RSTI série ISI-NIS), 7(5-6), pp 65-90]

[44]Dhilip Subramanian (2019,8 juin), [« A Simple Introduction to K-Nearest Neighbors Algorithm"] https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e.

[45] Salouan, S. Safi and B. Bouikhalene (en 2015) ["Printed Noisy Greek Characters Recognition Using Hidden Markov Model, Kohonen Network, K Nearest Neighbours and Fuzzy Logic ",mémoire présenté par Departement of mathématique et d'informatique, Sultan Moulay Slimane University, Beni Mellal, MOROCCO.]

[46]Jaillet S, Teisseire M, Dray G (2003) [Adéquation des modèles de représentation aux méthodes de catégorisation. LIRMM-CNRS -ISIM - Université Montpellier]

[47] Aouine Mohammed (2009) [”CATEGORISATION AUTOMATIQUE DE TEXTE ARABE”, Mémoire Présenté par département d’informatique, UNIVERSITE DE GUELMA]

[48]Laroum S., Béchet N., Hamza H., Roche M.(2009) [Classification automatique de documents bruités à faible contenu textuel. Revue des Nouvelles Technologies de l'Information]

[49] N. Bechet, I. Bayoudh (2008) [Quelles connaissances linguistiques permettent d’améliorer la classification de blogs avec les k-ppv ?, Quatrième Atelier : Qualité des Données et des Connaissances]

[50] G. Dziczkowski (2008) [Analyse des sentiments : système autonome d'exploration des opinions exprimées dans les critiques cinématographiques. Thèse de doctorat, Écoles Supérieures des Mines de Paris]

[51] Bassam Al-Salemi, Masri Ayob, Graham Kendall ,Shahrul Azman ,Mohd Noah (19,07,2019) [« RTAnews: A Benchmark for Multi-label Arabic Text Categorization »]
https://mendeley.figshare.com/articles/dataset/RTAnews_A_Benchmark_for_Multi-label_Arabic_Text_Categorization/8964563/1

[52]SOUIDI Abdelhakim (le 26 mai 2016) [”Indexation contrôlée des textes biomédicaux orientée par l’extraction de connaissances”,mémoire présenté par Faculté de Technologie,Département de Génie Biomédical ,Université Abou Bakr Belkaïd,Tlemcen]

[53]A. K. Farahat and M. S. Kamel, (2010) ["Enhancing document clustering using hybrid models for semantic similarity," in Proceedings of the eighth workshop on text mining at the tenth SIAM international conference on data mining. SIAM, Philadelphia, pp. 83-92]

[54]J. Sanger, (2007) [The Text Mining handbook: advanced approaches in analyzing unstructured data, Cambridge University Press]

[55]X.-B. Xue and Z.-H. Zhou, (2009) ["Distributional features for text categorization," Knowledge and Data Engineering, IEEE Transactions on, vol. 21, pp. 428-442]

[56]V. Amala Bai and D. Manimegalai, (2010) ["An analysis of document clustering algorithms," in Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on, pp. 402-406, Ramanathapuram]

[57] Bird, Steven, Edward Loper and Ewan Klein (2009) [Natural Language Processing with Python. O'Reilly Media Inc]

[58] https://docs.python.org/3/license.html , « depuis le site officiel de python. »

[59] https://www.anaconda.com/what-is-anaconda, « depuis le site officiel anaconda »

[60] https://docs.spyder-ide.org , « depuis me site officiel spyder »