République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Saad Dahleb de Blida 1

Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

Présenté en vue de l'obtention du
Diplôme de Master en Informatique
Spécialité: Traitement Automatique de la Langue (TAL)

Thème

Utilisation d'une approche non supervisée pour l'extraction automatique de mots-clés : Application aux textes Arabes

Réalisé par
BENBLIDIA Mehdi et
SERDOUNE Abderrahmane

Encadreur : MmeDROUA-Hamdani Ghania.
Promoteur: Mr CHERIF ZAHAR Amine.
Présidente du Jury : Mr FERFERA Sofiane.
Examinatrice : Mme YAKHLEF Hadjer

2018/2019

Résumé

A travers ce projet, nous présentons une approche hybride pour l'extraction automatique des mots clés dans les textes écrits en langue arabe. La méthode que nous proposons combine des critères statistiques tels que la fréquence avec d'autres linguistiques liés à la catégorie grammaticale.

L'expérience menée sur un corpus formé d'une collection de documents économiques, scientifiques et technologiques à montrer l'intérêt de combiner plusieurs critères pour améliorer le processus d'extraction des mots clés dans les textes arabes et en particulier l'intérêt de la catégorie grammaticale dans tel processus.

Mots clés

Traitement automatique de la langue arabe, Analyse morphologique, Mots pertinents, Pondération de termes, Catégorie grammaticale.

Abstract

Through this project, we present a hybrid approach for the automatic extraction of keywords in texts written in Arabic. The proposed method combines statistical criteria such as frequency of words with linguistic criteria related to the grammatical category.

The experiment conducted on a corpus formed of a collection of economic , technologic and scientific articles showed interest to combine several criteria to improve the process of extracting key words in Arabic texts and in particular the interest of the grammatical category in such process.

Keywords

Arabic language processing, Morpho-syntactic analysis, Relevant words, terms weighting, Grammatical category.

Remerciements

Au terme de notre cursus à l'université, nous invoquons Dieu et le remercions pour sa gratitude.

Nous tenons à transmettre notre reconnaissance à notre encadreur Mme DROUA Ghania et notre promoteur Mr CHERIF ZAHAR Amine qui nous ont guidé au cours de ce travail nous les remercions pour leurs confiance et leurs temps consacré pour mener à bien ce travail.

Nos remerciements s'adressent également aux membres du jury, qui ont bien voulu accepter d'évaluer notre projet.

Nous saisissons cette occasions pour complimenter nos professeurs pour les connaissances qu'ils nous ont transmises et l'éducation vertueuse qu'ils nous ont inculquée.

A tous ceux et celles qui ont contribué de près ou de loin à l'élaboration de ce mémoire qu'ils puissent trouver dans ce travail le témoignage de nos sincères gratitudes et profonds respects.

Mehdi et Abderrahmane

Dédicaces

Je dédie ce mémoire à :

A mes très chers parents,

Aucun mot ne pourrait exprimer la gratitude et l'amour que je vous porte.

A mes chers frères et à tous les membres de ma famille pour leur encouragement permanent, et leur soutien moral.

A mon binôme SERDOUNE Abderrahmane.

A tous mes amis et collègues du département math et informatique.

A tous ceux qui m'ont soutenu et qui me soutiennent encore.

BENBLIDIA Mehdi.

Dédicaces

Je dédie ce travail qui n'aura jamais pu voir le jour sans le soutien indéfectible et sans limites de mes chers parents qui ne cessent de me donner avec amour le nécessaire pour que je puisse arriver à ce que je suis aujourd'hui. Que dieu vous protèges et que la réussite soit toujours à ma portée pour que je puisse vous combler de bonheur .

Je dédie également ce travail à :

Mes chers frères et sœurs.

Mon binôme BENBLIDIA Mehdi.

Mes amis et collègues.

A toute la famille SERDOUNE.

SERDOUNE Abderrahmane.

Tables des matières

Liste des tableaux

Liste des figures

Introduction générale

Notre projet s'inscrit dans le cadre des travaux relatifs au traitement automatique de la langue Arabe. L'objet est de réaliser un système d'extraction automatique de mots clés dans les textes arabes en se basant sur la mesure d'importance des mots dans les textes. L'importance des mots sera déterminée en appliquant une combinaison de critère statistique distributionnel (fréquence) et d'autres linguistiques grammaticaux (types des termes).

En effet, l'extraction de mots clés est une tâche très importante pour les systèmes d'aide à la lecture, le résumé automatique et la traduction automatique…etc.

La mise en place d'un tel système nécessite une série de traitements automatiques comme la segmentation, la normalisation, la lemmatisation et la détection des catégories grammaticales et l'utilisation d'un analyseur morphologique et d'un étiqueteur grammatical

Problématique

Ces dernières années sont marquées par une augmentation énorme de la quantité d'information électronique arabe et dont l'accès à des informations pertinentes est devenu de plus en plus complexe et le besoin de développer des applications d'aide à la lecture est devenu incontournable.

La réalisation d'un système d'extraction de mots clés constitue un domaine à part entière se trouvant à la croisée du Traitement Automatique de la Langue (T.A.L) et de la Recherche d'Information (R.I).

Les travaux dans ce domaine existent déjà pour d'autres langues comme l'anglais ou le français, malheureusement pour l'arabe les travaux sont rares et les choses ne font que commencer. Nous essayons donc de contribuer dans ce sens en proposant une méthode de détection de mots clés pour les textes arabes.

Chapitre 01

Généralité sur la langue Arabe

1 Chapitre 01 : Généralité sur la langue Arabe

1.1 Introduction

La langue arabe est une langue dérivationnelle et flexionnelle. A l'origine, la langue arabe est la langue parlée par les Arabes. En plus, elle est la langue sacrée du Coran et de l'Islam. Du fait de la propagation de l'Islam et la diffusion du Coran, cette langue est devenue une langue liturgique. Elle est parlée dans 22 pays alors que le nombre de ses locuteurs est plus de 200 millions.

Dans ce chapitre, nous commencerons par présenter la langue arabe, son système d'écriture et sa morphologie. Ensuite, nous présenterons ces différentes catégories grammaticales. Enfin, nous donnerons un aperçu sur les particularités de la langue arabe.

1.2 Langue arabe

L'arabe est la langue officielle d'au moins 22 pays, parlée par plus de 200 millions de personnes à l'origine par les Arabes. C'est une langue sémitique (comme l'hébreu, l'araméen et le syriaque). Au sein de cet ensemble, elle appartient au sous-groupe du sémitique méridional. Du fait de l'expansion territoriale au Moyen Âge et par la diffusion du Coran, cette langue s'est répandue dans toute l'Afrique du nord et en Asie mineure. C'est aussi la langue de référence pour plus d'un milliard de musulmans.

Le développement de la langue arabe a été associé à la naissance et la diffusion de l'islam. L'arabe s'est imposée, depuis l'époque arabo-musulmane, comme langue religieuse mais plus encore comme langue de l'administration, de la culture et de la pensée, des dictionnaires, des traités des sciences et des techniques. Ce développement s'est accompagné d'une rapide et profonde évolution (en particulier dans la syntaxe et l'enrichissement lexical).

L'arabe peut être considérée comme un terme générique rassemblant plusieurs variétés :

- l'arabe classique : la langue du Coran, parlée au VIIe siècle
- l'arabe standard moderne (l'ASM) : une forme un peu différenciée de l'arabe classique, et qui constitue la langue écrite de tous les pays arabophones. L'ASM reste le langage de la presse, de la littérature et de la correspondance formelle, alors que l'arabe classique appartient au domaine religieux et est pratiqué par les membres du clergé

• les dialectes arabes : malgré l'existence d'une langue commune, chaque pays a développé son propre dialecte. Issus de l'arabe classique, leurs systèmes grammaticaux respectifs affichent de nettes divergences avec celui de l'ASM. On peut regrouper ces dialectes en quatre grands groupes:

1. les dialectes arabes, parlés dans la Péninsule Arabique : dialectes du Golfe, dialecte du najd, yéménite.

2. les dialectes maghrébins : algérien, marocain, tunisien, hassaniya de Mauritanie.

3. les dialectes proche-orientaux : égyptien, soudanais, syro-libano-palestinien, irakien (nord et sud)

4. la langue maltaise est également considérée comme un dialecte arabe.

L'arabe est un ensemble complexe dans lequel s'étendent des variétés écrites et orales répondant à un spectre très varié d'usages sociaux. L'ASM est la langue des médias officiels, de la communication écrite et de tout type de communication non spontanée. Elle se distingue des dialectes arabes par son système grammatical partagé avec l'arabe classique. L'ASM, quoique qu'elle soit considérée comme le symbole le plus puissant de l'unité arabe, possède des variations régionales. Nous reconnaissons un texte algérien vis-à-vis d'un texte égyptien ou d'un texte provenant des pays du Golfe. Cette variation est due aux différences qui ont lieu dans la formation de nouveaux vocabulaires. Mais elle est aussi la conséquence de l'histoire coloniale différente des régions impliquées. Les pays du Maghreb, par exemple, ont une tendance naturelle à regarder des exemples français, et le texte est largement influencé par la langue française même au niveau de la syntaxe et de la stylistique. Nous trouvons, par exemple **« الوزير الأول »** (le premier ministre français) au lieu du terme fréquent « **رئيس الوزراء** » (le président des ministres). Dans les pays arabes sans un passé colonial français, l'anglais remplace le français en tant que langue fournissant les modèles syntaxiques et stylistiques [1].

1.2.1 Système d'écriture

L'arabe s'écrit de droite vers la gauche en utilisant deux types de caractères : les lettres et les diacritiques. Les lettres arabes renferment 28 consonnes et des voyelles longues alors que les diacritiques servent à vocaliser le texte i.e. déterminer la phonétique de ses

mots pour une meilleure précision de la prononciation, Ces derniers se décompose en trois voyelles brèves et sept signes orthographiques qui s'ajoutent aux consonnes.

La représentation morphologique de l'arabe est complexe en raison de la variation morphologique et du phénomène d'agglutinement, les lettres changent de formes selon leur position dans le mot (isolée, initiale, médiane et finale). (**Tableau 1.1**) montre les 28 consonnes de la langue arabe et leurs différentes formes et positions.

Caractère	Initiale	Médiane	Finale	Isolé
Alif			ـا	ا
Ba	بـ	ـبـ	ـب	ب
Ta	تـ	ـتـ	ـت	ت
Tha	ثـ	ـثـ	ـث	ث
Jim	جـ	ـجـ	ـج	ج
Ha	حـ	ـحـ	ـح	ح
Kha	خـ	ـخـ	ـخ	خ
Dal			ـد	د
Thal			ـذ	ذ
Ra			ـر	ر
Zay			ـز	ز
Sin	سـ	ـسـ	ـس	س
Chin	شـ	ـشـ	ـش	ش
Sad	صـ	ـصـ	ـص	ص
Dhad	ضـ	ـضـ	ـض	ض
Tad	طـ	ـطـ	ـط	ط
Dha	ظـ	ـظـ	ـظ	ظ
Ayn	عـ	ـعـ	ـع	ع
Ghayn	غـ	ـغـ	ـغ	غ
Fa	فـ	ـفـ	ـف	ف
Qaf	قـ	ـقـ	ـق	ق
Kaf	كـ	ـكـ	ـك	ك
Lam	لـ	ـلـ	ـل	ل
Mim	مـ	ـمـ	ـم	م
Noun	نـ	ـنـ	ـن	ن
He	هـ	ـهـ	ـه	ه
Waw			ـو	و
Ya	يـ	ـيـ	ـي	ي

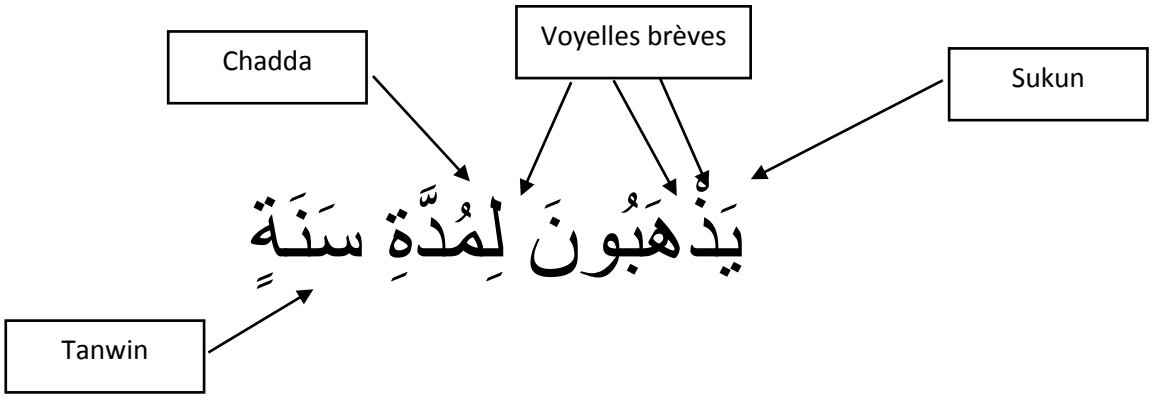
Tableau 1.1 : Les 28 consonnes de la langue arabe et leurs différentes formes et positions.

Les trois voyelles brèves sont :

- ✓ Fatha « ◌َ », elle surmonte la consonne et se prononce comme un « a » français.
- ✓ Damma « ◌ُ », elle surmonte la consonne et se prononce comme un « ou » français.
- ✓ Kasra « ◌ِ », elle se note au-dessous de la consonne et se prononce comme un « i » français.

Ces sept signes orthographiques sont :

- ✓ Sukun « ◌ْ » : ce signe indique qu'une consonne n'est pas suivie (ou muet) par une voyelle. Il est noté toujours au-dessus de la consonne.
- ✓ Les trois signes de tanwin : lorsque (la Fatha, la Kasra et la Damma) sont doublées, elles prennent un son nasal, comme si elles étaient suivies de « n » et on les prononce respectivement :
  - ➢ an « ◌ً » pour les Fathatan .
  - ➢ in « ◌ٍ » pour les Kasratan .
  - ➢ un « ◌ٌ » pour les Dammatan.
- ✓ Chadda « ◌ّ » : comme dans le français, l'arabe peut renforcer une consonne quelconque.
- ✓ Wasla « ◌ٱ » : quand la voyelle d'un Alif au commencement d'un mot doit être absorbée par la dernière voyelle du mot qui précède.
- ✓ Madda « ◌آ » : la madda (prolongation) se place sur l'Alif pour indiquer que cette lettre tient lieu de deux alifs consécutifs ou qu'elle ne doit pas porter le Hamza [2].



1.3 Morphologie arabe

Le lexique arabe comprend trois catégories de mots : verbes, noms et particules. Les verbes et noms sont le plus souvent dérivés d'une racine à trois consonnes radicales [3]. Une famille de mots peut ainsi être générée d'un même concept sémantique à partir d'une seule racine à l'aide de différents schèmes. Ce phénomène est caractéristique à la morphologie arabe. On dit donc que l'arabe est une langue à racines réelles à partir desquelles on déduit le lexique arabe selon des schèmes qui sont des adjonctions et des manipulations de la racine. Le Tableau 1.2 montre quelques exemples de schèmes appliqués aux mots **كتب** écrire et **حمل** porter. On peut ainsi dériver un grand nombre de noms, de formes et de temps verbaux [4].

Schèmes	**KTB**	**كَتَب**	**Notion d'écrire**
R1â-R2i-R3	KâTiB	كَاتِب	écrivain
R1a-R2a-R3a	KaTaBa	كَتَبَ	A écrit
MaR1-R2a-R3	maKTaB	مَكتَب	bureau
R1u-R2i-R3a	KuTiBa	كُتِبَ	A été écrit

HML	**حمل**	**Notion de porter**
HâMiL	حَامِل	Porteur
HaMaLa	حَمَلَ	A porté
maHMaL	مَحمَل	Brancard
HuMiLa	حُمِلَ	A été porté

Tableau 1.2 : Exemple de schèmes pour les mots **كتب** écrire et **حمل** porter [3].

Les lettres en majuscule (Ri) désignent les consonnes de base qui composent la racine.
Les voyelles (a, i, â……) désignent les voyelles et les consonnes en minuscule (m, …) sont des consonnes de dérivation utilisées dans les schèmes. La majorité des verbes arabes ont une racine composée de 3 consonnes. L'arabe comprend environ 150 schèmes ou patrons dont certains plus complexes, tel le redoublement d'une consonne ou l'allongement d'une voyelle de la racine, l'adjonction d'un ou de plusieurs éléments ou la combinaison des deux. Une autre caractéristique est le caractère flexionnel des mots : les terminaisons permettent de distinguer le mode des verbes et la fonction des noms [3].

1.4 Structure d'un mot Arabe

En arabe un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination d'éléments de la grammaire, la représentation suivante schématise une structure possible d'un mot. Notons que la lecture et l'écriture d'un mot se font de droite vers la gauche. Le

Tableau 1.3 suivant montre la structure du mot arabe

Post fixe	Suffixe	Corps schématique	Préfixe	Antéfixe

Tableau 1.3 : structure du mot arabe

- Antéfixes sont des prépositions ou des conjonctions.
- Préfixes et suffixes expriment les traits grammaticaux et indiquent les fonctions : cas du nom, mode du verbe et les modalités (nombre, genre, personne, …)
- Post fixes sont des pronoms personnels. [4]

Exemple أَتَتَذَكَرُونَنَا

Ce mot exprime la phrase en français: « Est ce que vous vous souvenez de nous? »

La segmentation de ce mot donne les constituants suivants:

أ	تَ	تَذَكَرُ	ونَ	نَا

Antéfixe : أ conjonction d'interrogation.

Préfixe : préfixe verbal du temps de l'inaccompli.

Corps schématique :تَذَكَرَ dérivé de la racine : **ذكر** selon le schème taR1aR2aR3a

Suffixe : **ون** suffixe verbal exprimant le pluriel

Post fixe : **نا** pronom suffixe complément du nom [4].

1.5 Structure des phrases Arabes

Il existe deux types de phrase en arabe : la phrase verbale et la phrase nominative. L'ordre des mots dans une phrase arabe déterminent son type :

Si la phrase est débutée par un verbe alors on dit qu'il est verbal par exemple :

المدرسة	إلى	التلميذ	ذهب
Ecole	à	élève	aller
Nom génitif	Particule génitif	Sujet	Verbe

Si la phrase est débutée par un nom ou par une particule on dit qu'il est nominatif par exemple :

المدرسة	إلى	ذهب	التلميذ
Nom génitif	Particule génitif	Verbe	Sujet

التلميذ	ذهب	المدرسة	إلى
Sujet	Verbe	Nom génitif	Particule génitif

1.6 Catégorie des mots

L'arabe considère 3 catégories des mots :

- **Le verbe** : entité exprimant un sens dépendant du temps, c'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble.
- **Le nom** : l'élément désignant un être ou un objet qui exprime un sens indépendant du temps.
- **Les particules** : entités qui servent à situer les événements et les objets par rapport au temps et l'espace, et permettent un enchainement cohérent du texte [4].

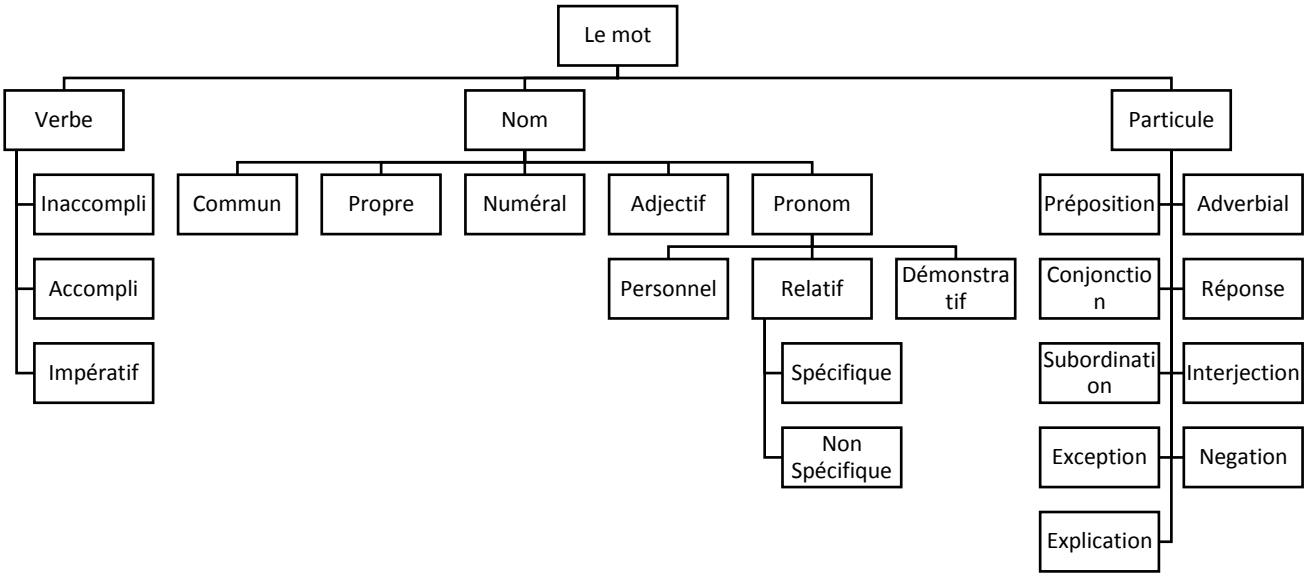


Figure 1.1 : Catégories des mots

1.6.1 Verbe

La plupart des mots en arabe, dérivent d'un verbe de trois lettres. Chaque verbe est donc la racine d'une famille de mots. Comme en français, le mot en arabe se déduit de la racine en rajoutant des suffixes ou des préfixes.

La conjugaison des verbes dépend de plusieurs facteurs :

- Le temps (accompli, inaccompli).
- Le nombre du sujet (singulier, duel, pluriel).
- Le genre du sujet (masculin, féminin).
- La personne (première, deuxième et troisième).
- Le mode (actif, passif).

Par exemple: **ك + ت + ب** K+T+B donne le verbe **كتب** KaTaBa. (Écrire).

Dans tous les mots qui dérivent de cette racine, on trouvera ces trois lettres K, T, B.

La conjugaison des verbes se fait en ajutant des préfixes et des suffixes, un peu comme en français.

La langue arabe dispose de trois temps :

- **L'accompli :** correspond au passé et se distingue par des suffixes (par exemple pour le pluriel féminin on a كَتَبنَ KaTaBna, (elles ont écrit) et pour le pluriel masculin on a كَتَبُوا KaTaBuu, (ils ont écrit).
- **L'inaccompli présent :** présente l'action en cours d'accomplissement, ses éléments sont préfixés **يكتب** yaKTuBu (il écrit), تَكتبُ taKTuBu , (elle écrit).
- **L'inaccompli futur :** correspond à une action qui se déroulera au future et est marqué par l'antéposition de **س** sa ou **سوف** sawfa au verbe (سَيَكتُبُ sayaKTuBu, il écrira, سوف يَكتُبُ sawfa yaKTuBu il va écrire) [4].

1.6.2 Nom

Les substantifs arabes sont de deux catégories, ceux qui sont dérivés de la racine verbale et ceux qui ne le sont pas comme les noms propres et les noms communs. Dans le premier cas, le fait que le nom soit dérivé d'un verbe, il exprime donc une certaine sémantique qui pourrait avoir une influence, comme exemple, dans la sélection des phrases saillantes d'un texte pour le résumé. [6]

La déclinaison des noms se fait selon les règles suivantes :

- ✓ **Le féminin singulier** : on ajoute le **ة**, exemple **صغير** petit devient **صغير ة** petite.
- ✓ **Le féminin pluriel** : de la même manière, on rajoute pour le pluriel les deux lettres **ا ت**, exemple **صغير** petit devient **صغيرات** petites.
- ✓ **Le masculin pluriel** : pour le pluriel masculin on rajoute les deux lettres **ين** ou**ون** dépendamment de la position du mot dans la phrase (sujet ou complément d'objet), exemple : **الراجع** revenant devient **الراجعين** ou **الراجعون** revenants.
- ✓ **Le pluriel irrégulier** : il suit une diversité de règles complexes et dépend du nom. Exemple : **طفل** un enfant devient **أطفال** des enfants. [4]

Le phénomène du pluriel irrégulier dans l'arabe pose un défi à la morphologie, non seulement à cause de sa nature non concaténa ive, mais aussi parce que son analyse dépend fortement de la structure [5] pour les verbes irréguliers.

Certains dérivés nominaux associent une fonction au nom :

- Agent (celui qui fait l'action),
- Objet (celui qui a subi l'action),
- Instrument (désignant l'instrument de l'action),
- Lieu.

Pour les pronoms personnels, le sujet est dans le verbe conjugué. Il n'est donc pas nécessaire (comme c'est le cas en français) de précéder le verbe conjugué par son pronom. On distinguera entre singulier, duel (deux) et pluriel (plus de deux) ainsi qu'entre le masculin et féminin. [4]

1.6.3 Particule

Ce sont principalement les mots outils comme les conjonctions de coordination et de subordination.

Les particules sont classées selon leur sémantique et leur fonction dans la phrase, on en distingue plusieurs types (introduction, explication, conséquence, …). Elles jouent un rôle important dans l'interprétation de la phrase [6].

. Elles servent à situer des faits ou des objets par rapport au temps ou au lieu, elles jouent également un rôle clé dans la cohérence et l'enchainement d'un texte.

Comme exemple de particules qui désignent un temps **منذ ,قبل ,بعد** pendant, avant, après, un lieu ou de référence **الذين** ceux, ….

Ces particules seront très utiles pour notre traitement à deux niveaux :

1-Elles font partie de l'anti dictionnaire qui regroupe les termes à ne pas prendre en considération lors de calcul de fréquence de distribution des mots.

2-Elles identifient des propositions composant une phrase.

Les particules peuvent avoir des préfixes et suffixes ce qui rajoute une complexité quant à leur identification. [4]

1.7 Particularités de la langue arabe

1.7.1 Voyelles

En Arabe écrit, les voyelles (signes diacritiques) sont omises et le résultat de cette omission est que les mots tendent à avoir un haut niveau d'ambiguïté qui risque de générer une certaine ambiguïté à deux niveaux :

- ✓ Sens du mot.
- ✓ Difficulté à identifier sa fonction dans la phrase, (différencier entre le sujet et le complément, ...).

Ceci peut influencer les fréquences des mots étant donné qu'elles sont calculées après la détection de la racine ou la lemmatisation des mots qui est basée sur la suppression de préfixes et suffixes. Lors du calcul des scores à partir des titres, il peut arriver que des mots soient considérés comme dérivants d'un même concept alors qu'ils ne le sont pas. Dans l'exemple, en utilisant la distribution des mots ou le titre avec ou sans lemmatisation, la phrase 3 aura un score le plus important alors que les phrases 1 et 2 semblent plus intéressantes, ce qui n'aurait pas été le cas avec un texte voyellé.

العنوان : اثر العلم	Titre : impact de la science
1- العلماء...	1- Les scientifiques
2- علميا...	2- Scientifiquement
3- بين العلم الوطني و العلم الأجنبي...	3- Entre le drapeau national et les drapeaux étrangers...

Tableau 1.4: Exemple sur l'effet du mot non voyelle « العلم » sur les extraits.

L'ambiguïté vient du mot **العلم** la science ou drapeau alors que voyellé on aura العِلْمُ la pour science et **العَلَمُ** pour le drapeau. Cette ambiguïté pourrait, dans certains cas, être levée soit par une analyse plus profonde de la phrase ou des statistiques (par exemple il est plus probable d'avoir « **العلم الوطني** » le drapeau national que la science nationale). De plus la capitalisation n'est pas employée dans l'arabe ce qui rend l'identification des noms propres, des acronymes, et des abréviations encore plus difficiles [7].

1.7.2 Agglutination

Contrairement aux langues latines, en arabe, les articles, les prépositions, les pronoms, etc. collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française.

Exemple : le mot arabe « **أتتذكروننا** » correspond en Français à la phrase "Est-ce que vous vous souvenez de nous ?". Cette caractéristique peut engendrer une ambiguïté au niveau morphologique. En effet, il est parfois difficile de distinguer entre une proclitique ou enclitique et un caractère original du mot. Par exemple, le caractère "و "dans le mot « **وصل** » (il est arrivé) est un caractère original alors que dans le mot « **وفتح** » (il a ouvert), il s'agit d'une proclitique. [7]

1.7.3 Irrégularité de l'ordre des mots dans la phrase

L'ordre des mots en arabe est relativement libre. D'une manière générale, on met au début de la phrase le mot sur lequel on veut attirer l'attention et l'on termine sur le terme le plus long ou le plus riche en sens ou en sonorité. 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. Ainsi par exemple, on peut changer l'ordre des mots dans la phrase (Tableau 1.5) pour obtenir deux phrases ayant le même sens. [7]

Verbe + sujet + complément	**فعل + فَاعل + متمم**	**Est allé le garçon a l'école**	**ذهب الولدُ إلَى المدرسة**
Sujet + verbe + complément	**فَاعل + فعل + متمم**	**Le garçon est allé à l'école**	**الولدُ ذهب إلَى المدرسة**
Complément + verbe + sujet	**متمم + فعل + فَاعل**	**A l'école est allé le garçon**	**إلَى المدرسة ذهب الولدُ**

Tableau 1.5: Exemple de combinaisons possibles d'inversion de l'ordre des mots dans la phrase.

La langue arabe n'est pas appuyée principalement sur les signes de ponctuations et les marqueurs typographiques, il est à noter que ces derniers ne sont pas utilisés de façon régulière dans les textes arabes actuels, et même dans le cas où ils y figurent, ils ne sont pas gérés par des règles précises d'utilisation.

Par ailleurs, nous pouvons trouver tout un paragraphe arabe ne contenant aucun signe de ponctuation à part un point à la fin de ce paragraphe. Ainsi, il convient de noter que la présence des signes de ponctuation ne peut pas guider la segmentation comme c'est le cas pour d'autres langues latines, telles que le français ou l'anglais. Ainsi, la segmentation de textes arabes doit être guidée non seulement par les signes de ponctuations et les marqueurs typographiques mais aussi par des particules et certains mots tels que les conjonctions de coordination, etc. [7]

1.7.4 Mots étrangers translittérés en arabe

Les translittérations en arabe de mots étrangers posent un problème, puisqu'ils n'ont pas de racine en arabe. Les mots translittérés sont considérés comme inconnus par l'analyseur. Quelques items étrangers méritent une attention particulière en raison de leurs fréquences élevées. Exemple : **أورو, دولار**... etc. [7]

1.7.5 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 [8]

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.

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

Tableau 1.6: Le système numérique arabe.

1.8 Conclusion

Dans ce chapitre, nous avons présenté les caractéristiques du texte arabe qui sont différentes par rapport à d'autres langues indo-européennes. L'Arabe se distingue par le lien étroit entre ses différents niveaux linguistiques : phonologique, morphologique, syntaxique et sémantique. Ces caractéristiques ont été traitées par différentes applications de traitement automatique de la langue arabe comme le résumé automatique et la traduction des textes arabes, …etc. De telles applications reposent sur des fonctions communes d'analyse syntaxique et morphologique que nous verrons dans le chapitre suivant.

Chapitre 02

Traitement Automatique de la Langue (TAL)

Et

Traitement Automatique de la Langue Arabe (TALA)

2 Chapitre 02 : TAL et TALA

2.1 Introduction

Le Traitements Automatique des Langues (TAL) est une discipline qui associe étroitement linguistes et informaticiens. Il repose sur la linguistique, les formalismes (représentation de l'information et des connaissances dans des formats interprétables par des machines) et l'informatique**.** Le but du TAL est réellement de "comprendre" le sens des phrases, les idées qui s'en dégagent et ce de manière à pouvoir "traiter" de la manière la plus optimale et la plus naturelle d'un point de vue humain ces phrases. [9].

Les outils du traitement automatique de la langue en général, et de la langue Arabe en particulier, se caractérisent par leur diversité au niveau des langages de développement utilisés, des entrées/sorties manipulées, des représentations internes et externes des résultats, etc. Cette diversité ne favorise ni l'interopérabilité entre ces différents outils, ni leur réutilisabilité dans de nouveaux contextes [10].

L'extraction automatique de mot-clé peut être officieusement décrite comme un processus par lequel une courte liste de mots-clés est extraite d'un texte beaucoup plus grand avec peu de perte d'informations. Il est utilisé pour fournir un procédé efficace pour les humains et les machines afin d'identifier rapidement le contenu et le type de textes et de documents.

Dans ce chapitre, nous présentons le traitement automatique des langues, leur niveau de traitement et les problèmes, le traitement automatique de la langue arabe et les différents problèmes de traitements. Ensuite, nous donnons un aperçu sur les méthodes d'extraction automatique des mots clés et en terminera avec une conclusion

2.2 Traitement Automatique de la Langue

Le traitement automatique des langues naturelle ou de la langue est un domaine de recherche qui se positionne à l'intersection de plusieurs disciplines : Informatique théorique, calcul statistique, linguistique, Intelligence artificielle …etc. Dont le principal objectif est la création des programmes informatiques capables de traiter automatiquement les données linguistiques, qui sont exprimées dans une langue dite naturelle, comme les textes écrits, ou bien les dialogues écrits ou oraux, ou encore tel que les unités linguistiques de taille inférieure à ce que l'on appelle habituellement des textes (par exemple : des phrases, des énoncés, des groupes de mots ou simplement des mots isolés). La langue naturelle désigne la langue parlée ou écrite par les êtres humains, par

opposition aux langages artificiels, informatiques, mathématiques ou logiques, par exemple.

Ces dernières décennies le TAL a connu une véritable ascension que ce soit sur le plan scientifique mais aussi socio-économique et cela par l'émergence de plusieurs firmes et de produits spécialisés, on parle aujourd'hui de Traduction automatique, correction automatique d'orthographe, résumé automatique et d'interrogation de base de données en langues naturelle, …etc [11].

N'importe quelles applications parmi celles citées précédemment, leurs réalisations passent principalement par différents niveaux (morphologique, syntaxique, sémantique et pragmatique) mais aussi par le développement de plusieurs modules important, où la réussite de l'application dépend pleinement de la performance de ces modules.

2.2.1 Bref historique du TAL

L'année	**Le traitement**
Années 50	Traduction automatique – débuts du TAL
1964	Rapport ALPAC
Années 60	Linguistique formelle (Chomsky, Montague) comme base pour le TAL. Applications basées sur des techniques linguistiques (Eliza, shrdlu) – Chomsky (grammaires formelles, analyseurs syntaxiques) ; sémantique procédural (Woods). Approches limitées à des domaines restreints. Non portables.
Années 70	Premières applications
Années 80	Approches symboliques. Applications utilisent des connaissances linguistiques et encyclopédiques extensives. Manquent de robustesse.
Années 90 et plus	Premiers corpus, approches statistiques, apprentissage automatique. Applications utilisent corpus de grande taille et méthodes statistiques.

Tableau 2.1: Bref historique du TAL [12].

2.3 Approches du traitement automatique de la langue naturelle

Les approches de traitement du langage naturel se divisent en trois catégories : symbolique, statistique et connexionniste. Dans cette section, nous examinons chacune

de ces approches en fonction de leurs fondements, leurs techniques typiques, des différences entre les aspects liés au traitement et au système, leur flexibilité et leur pertinence pour diverses tâches.

2.3.1 Approche statistique

Les approches statistiques utilisent diverses techniques mathématiques et utilisent souvent de grands corpus textuels pour développer des modèles approximatifs généralisés de phénomènes linguistiques basés sur des exemples concrets de ces phénomènes fournis par les corpus textuels sans ajouter de connaissances linguistiques ou mondiales significatives. Contrairement aux approches symboliques, les approches statistiques utilisent des données observables comme principale source de preuves.

Exemple : Le modèle statistique de Markov caché (HMM) est fréquemment utilisé hérité de la communauté de la parole. HMM est un automate à états finis qui possède un ensemble d'états avec des probabilités liées aux transitions entre états. Chaque état produit une des sorties observables avec une certaine probabilité.

Les approches statistiques ont généralement été utilisées dans des tâches telles que la reconnaissance de la parole, l'acquisition lexicale, l'analyse syntaxique, les collocations, la traduction automatique statistique, l'apprentissage statistique de la grammaire, … etc.

2.3.2 Approche symbolique

Les approches symboliques effectuent une analyse approfondie des phénomènes linguistiques et reposent sur une représentation explicite des faits sur le langage au moyen de schémas de représentation des connaissances bien compris et d'algorithmes associés. La principale source de preuves dans les systèmes symboliques provient de règles et de lexiques développés par l'homme.

Un bon exemple d’approches symboliques est celui de la logique ou des systèmes à base de règles. Dans les systèmes basés sur la logique, la structure symbolique se présente généralement sous la forme de propositions logiques. Les manipulations de telles structures sont définies par des procédures d'inférence qui préservent généralement la vérité. Les systèmes à base de règles consistent généralement en un ensemble de règles, un moteur d'inférence et un espace de travail ou une mémoire de travail. La connaissance est représentée sous forme de faits ou de règles dans la base de règles. Le moteur

d'inférence sélectionne à plusieurs reprises une règle dont la condition est remplie et l'exécute.

2.3.3 Approche connexionniste

Semblables aux approches statistiques, les approches connexionnistes développent également des modèles généralisés à partir d'exemples de phénomènes linguistiques. Ce qui sépare le connexionnisme des autres méthodes statistiques, c'est que les modèles connexionnistes associent l'apprentissage statistique à diverses théories de la représentation. Les représentations connexionnistes permettent donc la transformation, l'inférence et la manipulation de formules logiques. De plus, dans les systèmes connexionnistes, les modèles linguistiques sont plus difficiles à observer car les architectures connexionnistes sont moins contraintes que les architectures statistiques.

D'une manière générale, un modèle connexionniste est un réseau d'unités de traitement simples interconnectées dont les connaissances sont stockées dans les poids des connexions entre les unités. Les interactions locales entre les unités peuvent entraîner un comportement global dynamique, ce qui conduit au calcul.

2.4 Niveaux de traitement automatique de la langue :

Pour traiter le langage naturel on a besoin d'informations coordonnées et pertinentes sur la langue à des niveaux divers. Le plus souvent on a recours à cinq niveaux de connaissances sur une langue: phonologique, morphologique, syntaxique, sémantique et pragmatique. Ces niveaux se superposent, chacun apportant des problèmes spécifiques à résoudre relatif à un niveau donné.

La figure 2.1 montre ces différents niveaux de traitements :

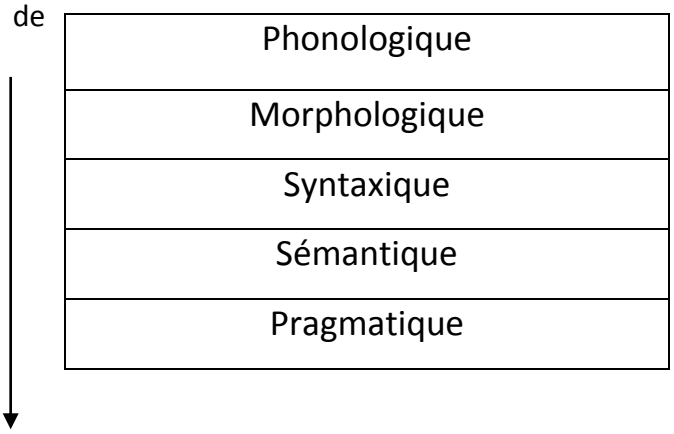


Figure 2.1 : Les Différents niveaux de traitement automatique de la langue.

L'entrée du système peut être soit vocale, soit écrite. Dans le premier cas, une étape de traitement phonétique est nécessaire. Les étapes qui suivent sont :

- ✓ **Le niveau phonologique** : La machine doit reconnaître les signaux acoustiques (domaine de la phonétique) et les identifier en tant que mots.

  Exemple : « Vincent mit l'âne dans un pré ou vingt cent mille ânes dans un pré ».

  Plus précisément, il s'agit de reconnaître dans le flot sonore les unités acoustiques élémentaires (phonèmes). La difficulté est que la forme acoustique d'un phonème varie selon plusieurs facteurs: le sexe, l'âge, la région, la fatigue, la peur, l'intensité de la parole, etc…
- ✓ **Le niveau morphologique :** il concerne l'étude de la formation des mots et de leur variation de formes.

  Exemple : « Aujourd'hui, parce qu'il a acheté un micro, Paul a payé la T.V.A »
- ✓ **Le niveau syntaxique :** il s'intéresse à l'agencement des mots et à leurs relations structurelles.

  Exemple : « Le boucher sale la tranche ».
- ✓ **Le niveau sémantique** : se consacre au sens des énoncés.

  Exemple : « Elle a mangé du poisson avec des amis ».
- ✓ **Le niveau pragmatique :** prend en compte le contexte d'énonciation.

  Exemple : « La mousse aux fraises est sur la table de l'avocat ». [13]

2.5 Domaines d'application du TALN

Parmi les diverses applications possibles, on peut citer principalement :

- Traducteur automatique.
- Résumé automatique de texte : résumer un texte signifie identifier le contexte et pondérer les parties significatives des autres.
- La réalisation d'interface en langue naturelle.
- L'indexation automatique de documents.
- L'enseignement assisté par ordinateur.
- Serveurs vocaux
- Correction lexicale et syntaxique
- Outils d'aide à la rédaction
- Dictée vocale
- … etc. [13]

2.6 Problèmes majeurs de TAL

2.6.1 Ambiguïté

Le langage naturel est ambigu, et ce à quelque niveau qu'on l'appréhende. Cette ambiguïté est un de ses traits caractéristiques. On peut d'ailleurs voir là le résultat d'un compromis inévitable entre d'un côté une capacité d'expression quasi illimitée, et de l'autre des contraintes liées à la limitation des ressources physiologiques mises en œuvre (taille de la mémoire à long et court-terme, densité de l'espace phonétique, contraintes articulatoires, etc.) [14]

Cette ambiguïté se manifeste par la multitude d'interprétations possibles pour chacune des entités linguistiques pertinentes pour un niveau de traitement, comme en témoignent les exemples suivants :

– ambiguïté des graphèmes (lettres) dans le processus d'encodage orthographique : comparez la prononciation du i dans lit, poire, maison.

– ambiguïté dans les propriétés grammaticales et sémantiques (i.e. associées à son sens) d'une forme graphique donnée : ainsi manges est ambigu à la fois morpho-syntaxiquement, puisqu'il correspond aux formes indicative et subjonctive du verbe manger, mais aussi sémantiquement.

– ambiguïté de la fonction grammaticale des groupes de mots.

– ambiguïté de la portée des quantificateurs, des conjonctions, des prépositions.

– ambiguïté sur l'interprétation à donner en contexte à un énoncé. [14]

2.6.2 Implicite

L'activité langagière s'inscrit toujours dans un contexte d'interaction entre deux humains, sensément dotés d'une connaissance du monde et de son fonctionnement telle que l'immense majorité des éléments de contexte nécessaires à la désambiguïsation mais aussi à la compréhension d'un énoncé naturel peuvent rester implicites. La situation change du tout au tout dès qu'une machine tente de s'insérer dans un processus de communication naturel avec un humain : la machine ne dispose pas de cette connaissance d'arrière-plan, ce qui rend la compréhension complète de la majorité des énoncés difficiles, voire impossible, si l'on ne dispose pas de bases de connaissances additionnelles, donnant accès à la fois à un savoir sur le monde (ou le domaine) en général (connaissance statique) et sur le contexte de l'énonciation (connaissance dynamique)[14].

2.7 Extraction automatique de mots clés

L'extraction de mots-clés est le processus de reconnaissance d'une courte liste de mots qui décrit et présente les idées ou les sujets les plus importants abordés dans un document textuel. C'est l'une des tâches du TAL étudiée par de nombreux chercheurs au cours des deux dernières décennies. Elle a été utilisée dans diverses applications de traitement du langage naturel, telles que les systèmes de récupération d'informations, la recherche dans une bibliothèque numérique, la gestion de contenu Web, la mise en cluster de documents et la synthèse de texte.

2.7.1 Mots clés

Les mots-clés sont une séquence d'un ou de plusieurs mots qui donnent une représentation des aspects principaux qui sont abordé dans d'un document. De ce fait, ils sont utilisés dans de nombreux domaines du Traitement Automatique des Langues (TAL). Ils peuvent faciliter la lecture d'un utilisateur en lui permettant d'aller d'un point clé à un autre lorsqu'ils sont mis en évidence dans un texte.

Les mots-clés sont aussi largement utilisés pour définir les requêtes au sein des systèmes de récupération d'informations (IR) car ils sont faciles à définir, réviser, mémoriser et partager. En comparaison avec les signatures mathématiques, les mots-clés sont indépendants de tout corpus et peuvent être appliqués à de multiples corpus et systèmes de RI.

Dans la plupart des articles, une liste de mots-clés est définie par l'auteur, cette liste représente les idées générales abordées dans le sujet de l'article. L'exemple suivant, montre une liste de mots-clés pour un article d'une revue économique :

الكلمات المفتاحية :مقياس التدريب، الموارد البشرية، التنوع، الشمولية، الاحتياجات التدريبية.

Mots clés: formation à l'échelle, les ressources humaines, la diversité, l'inclusion, les besoins de formation.

Key words: scale training, human resources, diversity, inclusiveness, training needs.

Figure 2.2: Un exemple des mots clés identifié par un auteur

2.7.2 Méthodes d'extraction automatique de mots-clés

L'extraction de mots-clés est une tâche qui consiste à analyser un document et à en extraire les aspects importants. L'extraction de termes-clés se focalise sur les unités textuelles qui composent ces phrases. Les unités textuelles sur lesquelles travaillent les systèmes d'extraction automatique de termes-clés sont appelées termes candidats. Ces derniers sont des mots ou des multi-mots (phrasèmes) pouvant être promus au statut de terme-clé.

L'extraction de mots candidats est une étape préliminaire de l'extraction de mots-clés, que ce soit pour les méthodes non-supervisées ou supervisées. Cette étape est importante, car si certains mots-clés du document analysé ne sont pas présents dans l'ensemble des termes candidats, alors ceux-ci ne pourront pas être extrait. Hulth [31].étudie des méthodes d'extraction des termes candidats. L'une de ces méthodes consiste à extraire tous les n-grammes, elle permet de retirer les termes contenant des mots outils. Dans ses expériences Hulth [31] montre que l'extraction de mots-clés à partir de n-grammes filtrés avec les mots outils donne les meilleurs résultats parmi les autres méthodes qu'elle propose.

Les travaux de Hulth [31] sont évalués avec un corpus dont les documents sont des résumés d'articles scientifiques. Cependant, dans d'autres domaines tels que la biomédecine, la nature des termes à extraire n'est pas la même. En effet, ce sont les acronymes et les entités nommées (noms de protéines par exemple) qu'il est nécessaire d'extraire en tant que termes-clés. Pour cela, l'extraction de termes candidats est spécifique au domaine d'application. Les méthodes d'extraction de termes-clés présentées dans ce chapitre traitent des documents supposés sans spécificités particulières, les méthodes d'extraction de termes candidats sont donc les mêmes que celles expérimentées par Hulth [31], mais il est envisageable de les adapter à des domaines présentant des spécificités particulières. Utilisés avec les méthodes non-supervisées, les termes candidats sont ordonnés selon un score d'importance obtenu soit à partir d'eux-mêmes, soit à partir de l'importance des mots qui les composent. Si une méthode s'appuie uniquement sur les mots, alors le score d'un terme candidat est généralement calculé en faisant la somme des mots qui le composent. Cependant, ceci n'est pas toujours juste, c'est donc un inconvénient important des méthodes travaillant sur les mots pour extraire les termes-clés. En effet, la sommation peut privilégier des termes qui contiennent beaucoup de mots non-importants vis-à-vis de termes contenant

très peu de mots, mais importants. Utilisés dans les méthodes supervisées, les termes candidats sont classés en tant que termes-clés ou non termes-clés grâce à des méthodes de classification.

Les méthodes existantes d'extraction automatique de mots clés peuvent être divisées en deux approches principales: Supervisée et non supervisée. La figure 2.3 montre les différents méthodes et approches d'extraction automatique de mot clés

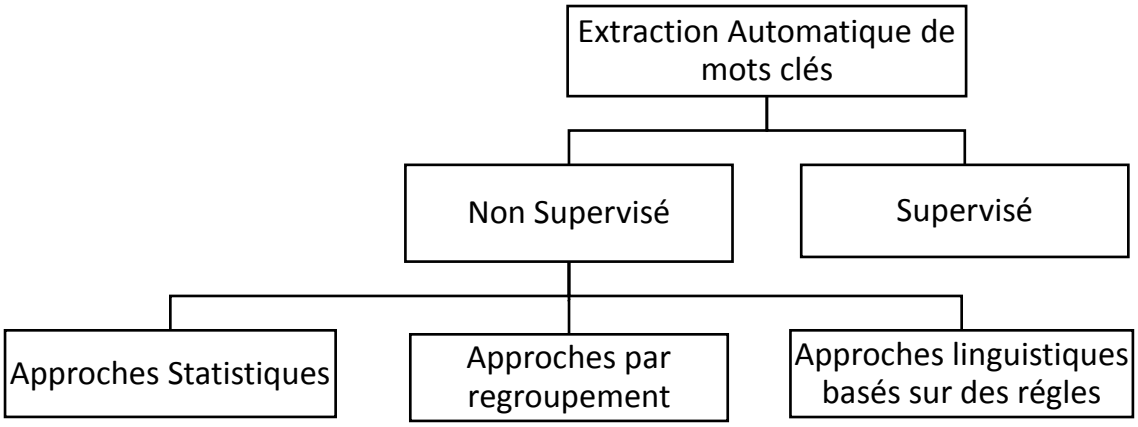


Figure 2.3 : Méthodes et approches d'extraction automatique de mots clés

2.7.2.1 Méthodes non supervisée

Les méthodes non-supervisées d'extraction de mot clés évite d'avoir à utiliser des documents annotés. Il utilise la modélisation du langage et l'analyse statistique pour sélectionner les mots-clés potentiels. Un mot clé candidat est souvent sélectionné en fonction de caractéristiques telles que sa fréquence dans le document, la position de sa première occurrence dans un document et ses attributs linguistiques, tels que sa tige et son tag de partie de la parole (POS) Les méthodes non supervisées sont en général indépendantes du domaine et moins chères, car elles ne nécessitent pas la construction d'un corpus annoté.

2.7.2.1.1 Approches statistiques

Cette approche comprend des méthodes simples qui ne nécessitent pas les données de formation. De plus, les méthodes sont indépendantes de la langue et du domaine. Les statistiques des mots du document peuvent être utilisées pour identifier des mots-clés: statistiques n-grammes, fréquence des mots, TF-IDF, cooccurrences de mots, arbre PAT (arbre Patricia; arbre de suffixe ou arbre de position), etc.

G. Salton et al, En 1975 [38], a proposé une méthode, l'analyse de la valeur de discrimination, qui classe les mots dans le texte en fonction de leur capacité à discriminer les documents d'une collection les uns des autres. La valeur d'un terme dans cette approche dépend de la variation de la séparation moyenne entre les documents individuels qui se produit lorsque le terme donné est affecté à l'identification du contenu. Les mots obtenant la plus grande séparation devraient être les meilleurs mots.

En 1995, J.D. Cohen [39], a proposé une approche permettant de tirer des termes d'index du texte. Il n'utilise aucune liste d'arrêt, aucun stemmer, ni aucun ne composant spécifique à une langue ou à un domaine, ce qui permet une application facile dans une langue ou un domaine avec une légère modification. La méthode utilise des comptes de n-grammes, ce qui donne une fonction similaire et plus générale qu'un stemmer.

En 2002, M. Ortuño et al [40]. Ont démontré que les mots importants d'un texte ont tendance à s'attirer et à former des groupes. Il fait valoir que l'écart type de la distance entre les occurrences successives d'un mot est un paramètre permettant de quantifier cette auto-attraction.

P. Carpena et al [41]. A proposé d'extraire automatiquement les mots-clés de textes littéraires en généralisant l'analyse statistique de niveau des systèmes quantiques désordonnés. Ils prennent en compte les fréquences des mots ainsi que leur distribution spatiale le long du texte. Ils se basent sur l'observation selon laquelle les mots importants sont significativement regroupés, alors que les mots non pertinents sont distribués de manière aléatoire dans le texte. Aucun corpus de référence n'est nécessaire dans cette approche et convient particulièrement aux documents uniques pour lesquels aucune information préalable n'est disponible [42].

L'inconvénient est que, dans certains textes professionnels, tels que santé et médecine, le mot clé le plus important puisse apparaître une seule fois dans l'article. L'utilisation de modèles habilités par les statistiques peut par inadvertance éliminer ces mots.

2.7.2.1.2 Approches linguistiques basées sur des règles

Ces approches sont généralement basées sur des règles et sont dérivées des connaissances / caractéristiques linguistiques. Ces approches peuvent être plus précises, mais nécessitent beaucoup de calcul et requièrent une connaissance du domaine en plus de l'expertise linguistique. Ces approches utilisent les caractéristiques linguistiques des

mots, principalement des phrases et des documents. L'approche linguistique comprend l'analyse lexicale, l'analyse syntaxique, l'analyse du discours, etc [42].

2.7.2.1.3 Approches par regroupement

L'objectif des approches par regroupement est de définir des groupes dont les unités textuelles partagent une ou plusieurs caractéristiques communes. Ainsi, lorsque des termes-clés sont extraits à partir de chaque groupe, cela permet de mieux couvrir le document analysé selon les caractéristiques utilisées.

Dans la méthode de Matsuo et Ishizuka [32], ce sont les termes qui sont regroupés. Parmi ceux-ci, seuls les plus fréquents sont concernés par le regroupement. Celui-ci s'effectue en fonction du lien sémantique entre les termes. Après le regroupement, la méthode consiste à comparer les termes candidats du document analysé avec les groupes de termes fréquents, en faisant l'hypothèse qu'un terme candidat qui co-occurre plus que selon toute probabilité avec les termes fréquents d'un ou plusieurs groupes est plus vraisemblablement un terme-clé.

Dans l'algorithme KeyCluster [33] Liu et al [35]. Utilisent aussi un regroupement sémantique, mais dans leur cas ils considèrent les mots du document analysé et ils excluent les mots outils. Dans chaque groupe sémantique, le mot qui est le plus proche du centroïde est sélectionné comme mot de référence. L'ensemble des mots de référence est ensuite utilisé pour filtrer les termes candidats en ne considérant comme termes-clés que ceux qui contiennent au moins un mot de référence [42].

2.7.2.2 Méthodes Supervisée

Le système d'extraction de mots-clés est formé pour déterminer si oui ou non un mot ou une phrase est un mot-clé. Un ensemble annoté de documents avec des mots-clés prédéfinis est toujours utilisé dans la phase d'apprentissage. Tous les termes et les expressions nominales du texte sont considérés comme des mots-clés potentiels, mais seuls ceux qui correspondent aux mots-clés attribués aux données annotées sont sélectionnés. Les principaux inconvénients de cette approche sont sa dépendance au modèle d'apprentissage, aux documents utilisés comme ensemble de formation et aux domaines des documents. En outre, les données de formation et les processus d'apprentissage prennent généralement beaucoup de temps [42].

KEA (algorithme d'extraction de phrases clés) a été développé par Frank et al [34]

. Dans ce système, un classificateur est construit sur la base du théorème de Bayes à partir de documents de formation, puis il est utilisé pour extraire des phrases clés à partir de nouveaux documents. KEA analyse le document d'entrée sur des limites orthographiques, par exemple : signes de ponctuation, nouvelles lignes, etc. pour trouver les expressions candidates. Deux caractéristiques sont utilisées: tf-idf et première occurrence du terme.

KPSpotter Song et al, combine le gain d'informations avec plusieurs techniques de traitement du langage naturel, telles que la première occurrence du terme et une partie de la parole. WordNet a été intégré à KPSpotter pour améliorer la précision de l'extraction.

La même année que les travaux de Hulth [31] sur le bien-fondé d'utiliser des traits linguistiques pour l'extraction automatique de mots-clés, Sujian et al.Proposent [36].une méthode utilisant un modèle d'entropie maximale dont l'un des traits repose sur les parties du discours des mots qui composent les termes candidats. Un modèle de maximum d'entropie consiste à trouver parmi plusieurs distributions, une pour chaque trait, laquelle a la plus forte entropie. La distribution ayant la plus forte entropie est par définition celle qui contient le moins d'informations, ce qui la rend de ce fait moins arbitraire pour l'extraction des mots-clés

(1)
$$\text{Score}(\text{terme}) = \frac{P(\text{oui}|\text{terme})}{P(\text{non}|\text{terme})}$$

(2)
$$P(\text{classe}|\text{terme}) = \frac{\exp\left(\sum_{\text{trait}} \alpha_{\text{trait}} \times \text{trait}(\text{terme}, \text{classe})\right)}{\sum_{c \in \{\text{oui},\text{non}\}} \exp\left(\sum_{\text{trait}} \alpha_{\text{trait}} \times \text{trait}(\text{terme}, c)\right)}$$

Le paramètre « α_{trait} » définit l'importance du trait auquel il est associé. Les Séparateurs à Large Marge sont aussi des classifieurs utilisés par les méthodes d'extraction automatique de mots-clés. Ils exploitent divers traits afin de projeter des exemples et des contre-exemples sur un plan, puis ils cherchent l'hyperplan qui les sépare. Cet hyperplan sert ensuite dans l'analyse de nouvelles données.

K. Zhang et al [37] A considéré l'extraction de mots-clés comme un problème de classification, dans lequel les mots / phrases d'un document devaient être classés en trois groupes: "bon mot-clé", "mot-clé différent" et "mot-clé mauvais". L'extraction de mots-clés a ensuite été réalisée à l'aide d'un modèle de classification SVM préalablement formé.

Bao Hong et al ont proposé une méthode améliorée d'extraction de mots clés (Extended TF). Ils ont utilisé les caractéristiques linguistiques de mots-clés tels que la fréquence des mots, une partie du discours, la fonction syntaxique des mots, la localisation apparue et la morphologie du mot. Sur la base des caractéristiques de chaque caractéristique, des pondérations ont été attribuées à différentes caractéristiques et le modèle SVM a été utilisé pour une optimisation ultérieure.

Des algorithmes d'extraction de mots-clés issus des deux approches ont été développés et mis en œuvre avec succès pour des documents dans les langues européennes (Rose et al. 2010; Liu et al [35] 2009; Matsuo et al. 2004 [32]). Cependant, bien que l'arabe soit l'une des principales langues internationales représentant environ 4% du contenu Internet, peu d'études sur l'extraction de mots-clés arabes ont été réalisées. El-Shishtawy et Al-Sammak (2009) ont présenté une méthode supervisée utilisant des connaissances linguistiques et des techniques d'apprentissage automatique pour extraire des mots-clés en arabe. Le système utilise un ensemble de données arabes annotées de 30 documents d'un domaine spécifique, compilées par les auteurs en tant que jeu de données d'apprentissage. Les mots-clés du jeu de données des documents utilisés pour évaluer leur système ont été attribués manuellement [42].

2.8 Catégorisation de textes

La catégorisation de texte, également appelée classification de texte, est le processus de catégorisation du texte en groupes organisés. En utilisant le traitement de langage naturel (NLP), les classificateurs de texte peuvent analyser automatiquement le texte, puis attribuer un ensemble de balises ou de catégories prédéfinies en fonction de son contenu.

La classification des textes devient une partie de plus en plus importante des entreprises, car elle permet d'obtenir facilement des informations à partir des données et

d’automatiser les processus métiers [30]. Certains des exemples les plus courants et des cas d'utilisation pour la classification automatique de texte sont les suivants:

Analyse des sentiments: processus permettant de comprendre si un texte donné parle positivement ou négativement d'un sujet donné (par exemple, à des fins de surveillance de la marque).

Détection de sujet: tâche qui consiste à identifier le thème ou le sujet d’un texte (par exemple, savoir si une critique de produit concerne la facilité d’utilisation, le support client ou la tarification lorsqu’on analyse les commentaires des clients).

Détection de la langue: procédure permettant de détecter la langue d’un texte donné (par exemple, savoir si un ticket de support entrant est écrit en anglais ou en espagnol pour acheminer automatiquement les tickets vers l’équipe appropriée) [30].

La figure 2.4 montre le processus qui résume en générale la classification

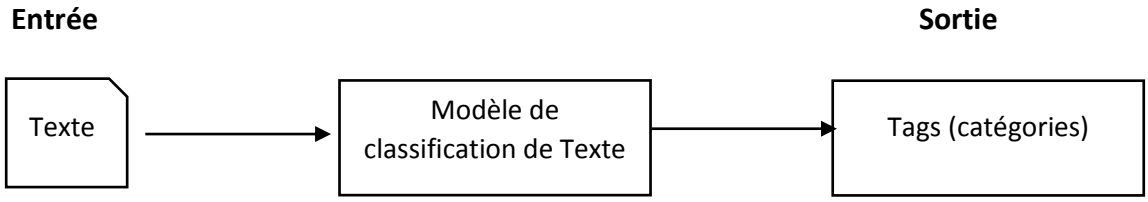


Figure 2.4 : Classification [30].

2.8.1 Fonctionnement de catégorisation du texte

La classification du texte peut être réalisée de deux manières différentes: la classification manuelle et la classification automatique. Dans le train, un annotateur humain interprète le contenu du texte et le catégorise en conséquence. Cette méthode peut fournir des résultats de qualité, mais elle prend du temps et coûte cher. Ce dernier s'applique à l’apprentissage automatique, au traitement du langage naturel et à d’autres techniques permettant de classer automatiquement le texte de manière plus rapide et plus rentable.

Il existe de nombreuses approches de la classification automatique du texte, qui peuvent être regroupées en trois types de systèmes différents:

- Systèmes à base de règles
- Systèmes basés sur l'apprentissage automatique [30].

2.8.1.1 Systèmes basés sur des règles

Les approches basées sur des règles classifient le texte en groupes organisés à l'aide d'un ensemble de règles linguistiques créées à la main. Ces règles indiquent au système d'utiliser des éléments sémantiquement pertinents d'un texte pour identifier les catégories pertinentes en fonction de son contenu. Chaque règle comprend un antécédent ou un motif et une catégorie prédite.

Comme par exemple, afin de classer les articles de presse en 2 groupes, à savoir Sports et Politique. Tout d'abord, il faut définir deux listes de mots caractérisant chaque groupe (par exemple, des mots liés à des sports tels que le football, le basketball, LeBron James, etc.) et des mots liés à la politique tels que Donald Trump, Hillary Clinton, Poutine, etc.…) Ensuite, pour classer un nouveau texte entrant, il faut compter le nombre de mots liés au sport qui apparaissent dans le texte et faire la même chose pour les mots liés à la politique. Si le nombre d'apparences de mots liées au sport est supérieur au nombre de mots liés à la politique, le texte est classé comme sport et inversement [30].

Les systèmes à base de règles sont compréhensibles par l'homme et peuvent être améliorés avec le temps. Mais cette approche présente des inconvénients. Pour commencer, ces systèmes nécessitent une connaissance approfondie du domaine. Elles prennent également beaucoup de temps, car la génération de règles pour un système complexe peut s'avérer très complexe et nécessite généralement beaucoup d'analyses et de tests. Les systèmes à base de règles sont également difficiles à maintenir et mal dimensionnés, car l'ajout de nouvelles règles peut affecter les résultats des règles préexistantes [30].

2.8.1.2 Systèmes basés sur l'apprentissage automatique

Au lieu de s'appuyer sur des règles élaborées manuellement, la classification de texte avec apprentissage automatique apprend à établir des classifications basées sur des observations passées. En utilisant des exemples pré-étiquetés en tant que données d'apprentissage, un algorithme d'apprentissage automatique peut apprendre les différentes associations entre des éléments de texte et qu'un résultat particulier (c'est-à-dire des balises) est attendu pour une entrée particulière (c'est-à-dire du texte).

La première étape vers la formation d'un classifieur avec l'apprentissage automatique est l'extraction de caractéristiques: une méthode est utilisée pour transformer chaque texte

en représentation numérique sous la forme d'un vecteur. L'une des approches les plus fréquemment utilisées est le sac de mots, où un vecteur représente la fréquence d'un mot dans un dictionnaire de mots prédéfini.

Ensuite, l'algorithme d'apprentissage automatique est alimenté avec des données d'apprentissage composées de paires d'entités (vecteurs pour chaque exemple de texte) et d'étiquettes (par exemple, sport, politique) afin de produire un modèle de classification: La figure 2.5 montre ce processus

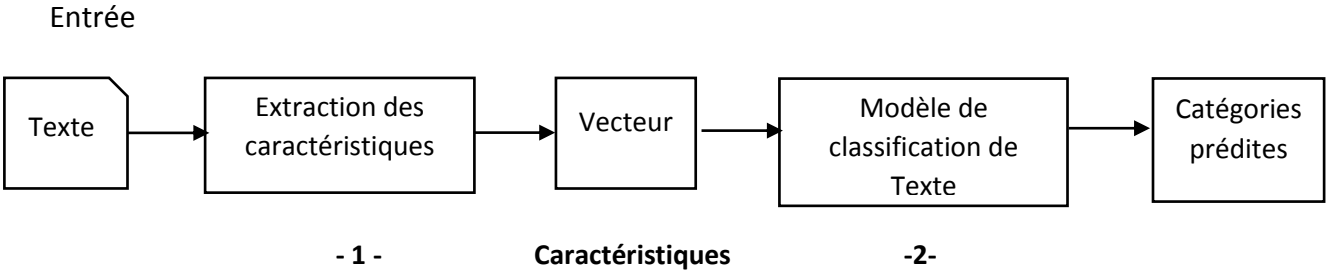


Figure 2.5 : Classification de texte avec apprentissage automatique [30].

1- Transforme chaque texte en un ensemble de caractéristiques sous la forme d'un vecteur

2- Les caractéristiques sont ensuite introduites dans le modèle de classification qui produit les catégories prédites

La classification de texte avec l'apprentissage automatique est généralement beaucoup plus précise que les systèmes de règles conçus par l'homme, en particulier pour les tâches de classification complexes. En outre, les classificateurs avec apprentissage automatique sont plus faciles à gérer et vous pouvez toujours baliser de nouveaux exemples pour apprendre de nouvelles tâches [30].

2.8.2 Algorithme de classification de texte

Certains des algorithmes d'apprentissage automatique les plus populaires pour la création de modèles de classification de texte incluent la famille des algorithmes naïfs bayes, les machines à vecteurs de support et l'apprentissage en profondeur.

- **Naive Bayes**

Naive Bayes est une famille d'algorithmes statistiques que nous pouvons utiliser lors de la classification de texte. Un des membres de cette famille est Multinomial Naive Bayes (MNB). Un de ses principaux avantages est que vous pouvez obtenir de très bons résultats

lorsque les données disponibles ne sont pas très importantes (environ deux mille échantillons étiquetés) et que les ressources informatiques sont rares.

En résumé, Naive Bayes est basé sur le théorème de Bayes, qui nous aide à calculer les probabilités conditionnelles d'occurrence de deux événements en fonction des probabilités d'occurrence de chaque événement. Cela signifie que tout vecteur représentant un texte devra contenir des informations sur les probabilités d'apparition des mots du texte dans les textes d'une catégorie donnée, afin que l'algorithme puisse calculer la probabilité que ce texte appartienne à la catégorie [30].

2.9 Evaluation sur la liste de mot clés

2.9.1 Evaluation manuelle

L'évaluation manuelle peut être faite en annotant manuellement, mais cela retourne généralement un état d'accords inter-annotateurs très faibles. D'autres prennent le parti d'accoler bout à bout des séquences appartenant à des textes différents ; les ruptures lexicales sont alors les ruptures entre textes. Dans la plupart du temps, l'évaluation manuelle est très couteuse donc les évaluations automatique ou semi-automatique sont considérées comme des bonnes alternatives.

2.9.2 Evaluation semi-automatique

L'évaluation semi-automatique peut être faite en comparant les résultats produits automatiquement (mot clés) par le système et d'autres produits manuellement par un expert humain, dans ce cas nous faisons appel par exemple à la métrique F-mesure qui calcule les scores de rappel et de précision.

2.9.3 Evaluation automatique

Afin d'évaluer automatiquement de tel système, d'autres scores ont été proposés, dont les plus usités sont les mesures Pk et WindowDiff. La mesure WindowDiff consiste à calculer la différence du nombre de ruptures dans une fenêtre glissante. Les mesures ROUGE, pour Recall-Oriented Understudy for Gisty Evaluation, ont été introduites par Lin. Ces mesures sont fondées sur la comparaison de n-grammes entre un ou plusieurs résumés de référence et un résumé à évaluer. Il n'existe pas un unique résumé de référence, et il est donc essentiel de comparer les résumés automatiques à plusieurs résumés de référence établis manuellement afin d'obtenir des mesures plus précises de

la qualité des résumés. Ces mesures nécessitent donc la rédaction de résumés de référence par un ou plusieurs experts au préalable de la mesure de qualité du résumé.

2.10 Traitement Automatique de la langue Arabe

Le Traitement Automatique de la Langue Arabe (TALA) est une discipline en pleine expansion, dans laquelle on voit de plus en plus de recherches et de technologies se soucier des spécificités de la langue Arabe [9] et proposer des outils nécessaires au développement de son traitement automatique. Par ses propriétés morphologiques et syntaxiques. La langue Arabe est considérée comme une langue difficile à maîtriser dans le domaine du TAL [15], [16]. Les recherches pour le TALA ont débuté vers les années 1970. Les premiers travaux concernaient notamment les lexiques et la morphologie.

Avec la diffusion de la langue Arabe sur le Web et la disponibilité des moyens de manipulation de textes Arabes, les travaux de recherche ont abordé des problématiques plus variées comme la syntaxe, la traduction automatique, l'indexation automatique des documents, la recherche d'information, etc.

2.11 Difficultés du Traitement Automatique de la Langue Arabe

En traitement automatique de la langue arabe rencontre quatre problèmes: la segmentation du texte, l'agglutination des mots et détection de racine, l'absence de voyelles à l'écrit et l'étiquetage grammatical. Pour chacun de ces problèmes, tout système de traitement automatique doit traiter et enlever certaine ambiguïté.

2.11.1 Segmentation

L'une des étapes principales pour le traitement automatique de la langue arabe est la segmentation des textes cette dernière consiste à découper le texte en unités qu'on aura défini et repéré au préalable. En effet, l'opération de cette étape consiste à déterminer les segments de ses éléments de base qui sont les caractères, en éléments formants différents niveaux structurels tels que : paragraphe, phrase, syntagme, mot graphique, mot-forme, morphème, etc.

Cependant, les particularités de la langue arabe, rendent la segmentation arabe toujours différente, il n'y a pas de majuscules qui marquent le début d'une nouvelle phrase et les signes de ponctuation, ne sont pas utilisés de façon systématique. [17]

2.11.2Agglutination des mots et Détection de la racine

Dans toute perspective de traitement automatique, le problème est de décomposer le mot en différentes parties et pour la langue arabe la plupart des mots sont composés par agglutination d'éléments lexicaux de base (proclitique + base + enclitique). Cette décomposition nécessite des connaissances de niveau supérieur en cas ou le mot accepte plusieurs segmentations.

Pour identifier la racine d'un mot, il va falloir connaître le schème par lequel il a été dérivé et supprimer les éléments flexionnels (antéfixes, préfixes, suffixes, post fixes) qui ont été ajoutés.

Le Tableau 2.2 montre la liste de préfixes et de suffixes que plusieurs d'entre eux ont été utilisés pour la lemmatisation de mots arabes

Préfixes							
لا	فبِ	للا	كمـ	بمـ	وتَ	بتَ	والـ
با	وا	لبِ	فمـ	لمـ	ستَ	يتَ	فالـ
	فا	وبِ	الـ	و	نتَ	متَ	بالـ
Suffixes							
ا	ةَ	ين	يةَ	هم	ته	وه	ات
	ه	يه	تك	هن	تم	ان	وا
	ي	يةَ	نا	ها	كم	تي	ون

Tableau 2.2: Liste des préfixes et suffixes les plus fréquents.

C'est important dans l'analyse morphologique de séparer et identifier des morphèmes semblables aux mots préfixés comme les conjonctions wa- و et fa- ف, des prépositions préfixées comme bi- ب et li- ل, l'article défini ال, des suffixes de pronom possessif.

L'étape d'analyse morphologique détermine un schème possible. Les préfixes et suffixes sont trouvés en enlevant progressivement des préfixes et des suffixes et en essayant de faire correspondre toutes les racines produites par un schème afin de retrouver la racine.

Lorsqu'un mot peut être dérivé de plusieurs racines différentes, la détection de la racine est encore plus difficile, en particulier en absence de voyelles.

Par exemple, pour le mot arabe ايمان AymAn les préfixes possibles sont : "Ø","A ا" et "Ay اي " et les suffixes possibles sont : "Ø" et "An ان ", sans compter que ce mot peut aussi représenter un nom propre إيمَان *Imène*.

En traitement automatique de la langue arabe on fait face à un problème qui est l'agglutination.

C’est quand les composantes du mot sont liées les unes aux autres et ce qui complique la tâche de l’analyse morphosyntaxique pour identifier les vrais composants du mot.

Par exemple, le mot أَلَمُهُمْ ALaMuhum (*leur douleur)* dans sa forme voyellée n’accepte qu’une seule segmentation : أَلَمُ + هُمْ (ALaMu+hum). Dans sa forme non voyellée المهم (ALMHM), le même mot accepte au moins les trois segmentations présentées dans le Tableau 2.3

Segmentation possible		**Traduction en français**
أ+ لم+ هم	A+LM+hm	*Les a-t-il ramassés*
ألم+ هم	ALM+hm ALM+hm	*Leur douleur* *Il les a fait souffrir*
أل+ مهم	AL+MHM	*L’important*

Tableau 2.3 : Exemple de segmentation du mot المهم

L'amplification de l’ambiguïté de segmentation s’opère selon deux façons :

D’abord, il y a plus d’unités ambiguës dans un texte non voyellé que dans son correspondant voyellé mais aussi, les unités ambiguës acceptent plus de segmentations dans le texte non voyellé.

De plus le fait de précéder la lemmatisation par la troncature des préfixes avant les suffixes (et réciproquement) peut influencer les résultats. En considérant l'exemple dans le Tableau 2.3**,** sur un texte où la notion de douleur est importante, le fait d'avancer la suppression des préfixes avant les suffixes les mots comme أَلَمُهُمْ*leur douleur (pour le pluriel)*, **ألمهما** *leur douleur* (pour le duel) exprimeront une toute autre notion. [17]

2.11.3Voyellation

L’absence quasi systématique de la voyellation dans les textes arabes est vraiment un problème. En effet, les signes de voyellation , apparaissent dans certains textes comme le Coran, et les hadiths ou bien dans les textes littéraires comme la poésie classique et ils sont notés sous la forme de signes diacritiques placés au-dessus ou au-dessous des lettre dans ce cas on dit qu’ils sont édités en graphie voyellée.

Lors de l’analyse automatique la non-voyellation dans les textes arabes engendre plusieurs cas d’ambiguïtés et des problèmes. En effet, l’ambiguïté grammaticale accroitre si le mot est non voyellé. Cela est dû au fait qu’un mot non voyellé possède plusieurs

voyellations possibles, et pour chaque voyellation est associée une liste différente de catégories grammaticales [17].

Le Tableau 2.4 montre un exemple du mot non-voyellé ktb | كتب possède 16 voyellations potentielles et qui représentent 8 catégories grammaticales différentes.

Mot Voyellé	Pré-notion	Notion d'écrire
كَتَبَ	Kataba	Il a écrit
كُتِبَ	Kutiba	Il a été écrit
كُتُب	Kutub	Des livres
كَتَب	Katob	Un écrit
كَتَّبَ	Kattaba	Il a fait écrire
كُتِّبَ	Kuttiba	Faire écrire-forme factitive
كَتِّبُ	Kkattibo	Fais écrire
كَتَبَّ	Katabba	Comme trancher

Tableau 2.4 : Exemple de voyellation de mot non-voyellé ktb | كتب

2.11.4 L'étiquetage grammatical

L'étiquetage grammatical est l'opération qui attribue à chacun des mots d'un texte la catégorie (nom, verbe, adjectif, article défini, etc.). La difficulté de l'étiquetage grammatical s'amplifie lorsque les textes visés se présentent sous leur forme non pas voyellée, mais partiellement voyellée ou encore totalement non voyellée, ce qui correspond au cas le plus courant.

2.12 Outils de traitement automatique de la langue arabe

Les outils de traitement automatique de la langue arabe sont l'ensemble des recherches et développements qui ont pour objectif à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication. Notre objectif dans cette section est de recenser les principaux outils de TAL en langue arabe.

2.12.1 Analyseurs morphologiques

- **DIINAR**

C'est une ressource linguistique très efficace de l'arabe, structurée comme une base de données, et opérant au niveau du mot graphique qu'on peut le dire le niveau morphologique. Leur nom complet est **DI**ctionnaire **IN**formatisé de l'**Ar**abe [18]

- **Buckwalter**

L'analyseur de buckwalter developpé par LDC (Linguistic Data Consortium) permet de segmenter chaque unité lexicale en une séquence du type préfixe-stem-suffixe. Il est constitué principalement de trois lexiques : préfixes (548 entrées), suffixes (906 entrées), et stem (78839 entrées). Ainsi, l'analyseur donne en sortie l'unité lexicale, sa catégorie morphosyntaxique et sa traduction anglaise [10].

- **Aramorph**

L'analyseur morphologique Aramorph segmente les unités lexicales, repère les différents composants et atteste son appartenance à la langue.

Exemple:" **وَالتلوث** " " et la pollution " l'analyseur découpera le proclitique و et dira que و est celui de la liaison. [10]

- **Snowball**

Snowball est un langage de traitement de petites chaînes conçu pour la création d'algorithmes de stemming à utiliser dans l'extraction d'informations. Arabic snowball stemmer a été conçu en 2016 et il utilise le modèle le plus récent par rapport à d'autres analyseurs morphologique gratuit [19]

2.12.2Analyseurs morphologiques à base de racine

- **L'analyseur de larkey**

L'approche est une analyse morphologique assouplie. Elle consiste à essayer de déceler les préfixes et les suffixes ajoutés à l'unité lexicale [20]

- **Shereen Khoja**

L'approche de Khoja consiste à détecter la racine d'une unité lexicale, d'une part, il faut connaître le schème par lequel elle a été dérivé et supprimer les éléments flexionnels (préfixes et suffixes) qui ont été ajoutés, d'autre part comparer la racine extraite avec une liste des racines préalablement conçue [21].

2.12.3 Concordanciers

Le concordancier a pour objectif de permettre l'exploration du corpus selon les traits proposés par l'analyse morphologique et selon les informations graphiques qui se trouvent dans le texte. Il prend en entrée un texte ou un ensemble de textes et il permet :

- La construction de listes de fréquences d'items, de racines ou tout autre trait de l'analyse morphosyntaxique, par ordre alphabétique ou par ordre fréquentiel.
- La construction d'une concordance.

2.12.4 Etiqueteur grammaticale (POS TAGGER)

- **Stanford- POS TAGGER**

Un analyseur en langage naturel est un programme qui calcule la structure grammaticale des phrases, par exemple, quels groupes de mots vont ensemble (en tant que "phrases") et quels mots sont le sujet ou l'objet d'un verbe. Les analyseurs syntaxiques probabilistes utilisent la connaissance de la langue tirée de phrases analysées à la main pour tenter de produire l'analyse la plus probable des nouvelles phrases. Ces analyseurs statistiques commettent encore des erreurs, mais fonctionnent généralement assez bien. Leur développement a été l'une des plus grandes avancées du traitement du langage naturel dans les années 90 [22] .

Tag	Meaning	Tag	Meaning
ADJ	Adjective	NNS	Noun, plural
CC	Coordinating conjunction	NOUN	Noun
CD	Cardinal number	PRP	Personal pronoun
DT	Determiner	PRP$	Possessive pronoun
DTJJ	Adjective with the determiner 'Al'	PUNC	Punctuation
DTJJR	Adjective, comparative with the determiner 'Al'	RB	Adverb
DTNN	Noun, singular or mass with the determiner 'Al'	RP	Particle
DTNNP	Proper noun, singular with the determiner 'Al'	UH	Interjection
DTNNPS	Proper noun, plural with the determiner 'Al'	VB	Verb, base form
DTNNS	Noun, plural with the determiner 'Al'	VBD	Verb, past tense
IN	Preposition or subordinating conjunction	VBG	Verb, gerund or present participle
JJ	Adjective	VBN	Verb, past participle
JJR	Adjective, comparative	VBP	Verb, non-3rd person singular present
NN	Noun, singular or mass	VN	Verb, past participle
NNP	Proper noun, singular	WP	Wh-pronoun
NNPS	Proper noun, plural	WRB	Wh-adverb

Figure 2.6: les différentes étiquettes d'un POS-Tagger

2.13 Prétraitements nécessaires pour le TALA

2.13.1 Encodage

La langue arabe est encodée suivant plusieurs formats d'encodage comme Unicode, ISO-8859-6, ou autres. Les textes recherchés et les requêtes peuvent être encodés différemment, afin de rendre ceux-ci incomparables. Par exemple, les documents sont représentés en Unicode (UTF-8) et les requêtes en ISO-8859-6 ou un autre encodage. Afin

d'apparier les documents avec les requêtes, nous devons réutiliser des outils de conversion entre différents encodages. Ainsi, tout a été transformé en format Unicode dans ce cas.

2.13.2 Unicode

Le standard Unicode est un mécanisme universel de codage de caractères. Il définit une manière cohérente de coder des textes multilingues et facilite l'échange de données textuelles. Il est obligatoire pour la plupart des protocoles de l'Internet, et mis en oeuvre dans tous les systèmes d'exploitation et langages informatiques modernes. Unicode est la base de tout logiciel voulant fonctionner aux quatre coins du monde.

À l'heure actuelle, les données Unicode peuvent être codées sous trois formes principales : une forme codée sur 32 bits (UTF-32), une forme sur 16 bits (UTF-16) et une forme de 8 bits (UTF-8) conçue pour faciliter son utilisation sur les systèmes ASCII préexistants [23].

2.13.3UTF-8

Afin de satisfaire les besoins des systèmes architecturés autour de l'ASCII ou d'autres jeux de caractères à un octet, le standard Unicode définit une forme en mémoire supplémentaire : l'UTF-8.

C'est une forme de mémorisation très fréquemment adoptée pour effectuer la transition des systèmes existants vers Unicode, et elle a notamment été choisie comme forme préférée pour l'internationalisation des protocoles d'Internet.

UTF-8 est un codage constitué de suites d'octets; les bits de poids le plus fort d'un octet indiquent la position de celui-ci dans la suite d'octets [23].

Autres caractéristiques importantes de l'UTF-8 :

- Conversion efficace à partir de ou vers un texte codé en UTF-16 ou en UTF-32.
- Le premier octet indique le nombre d'octets, ceci permet une analyse rapide du texte vers l'avant.
- Recherche rapide du début de tout caractère, quel que soit l'octet où l'on commence la recherche dans un flux de données, il suffit de consulter au plus quatre octets en amont pour reconnaître aisément le premier octet qui code le caractère.
- UTF-8 est un mécanisme de stockage relativement compact en termes d'octets.

2.14 Travaux relatifs

D'après les recherches que nous avons effectuées sur le web, nous n'avons pas trouvé des travaux indépendants qui traitent le problème d'extraction automatique de mots clés dans les textes arabes, par contre, il existe des travaux qui intègrent cette tâche dans leur processus de traitement. Les travaux recensés concernent particulièrement les sujets suivants :

- **Etude sémantique des mots-clés et des marqueurs lexicaux stables dans un corpus technique :** Ce travail présente les résultats d'une analyse sémantique quantitative des unités lexicales spécifiques dans un corpus technique, relevant du domaine des machines-outils pour l'usinage des métaux. L'étude vise à vérifier si et dans quelle mesure les mots-clés du corpus technique sont monosémiques.
- **Identification d'opinions dans les textes arabes en utilisant les ontologies :** Les opinions explicites peuvent être extraites par projection directe des concepts ontologiques sur le texte. Cependant, les opinions implicites ont besoin d'une exploration profonde de la couche sémantique de l'ontologie, en exploitant les relations entre les concepts, les individus et les attributs.
- **Résumé automatique :** Résumer un texte consiste à réduire ce texte en un nombre limité de mots. Le texte ainsi réduit doit rester fidèle aux informations et idées du texte original, et dans la mesure du possible rendre compte du style et de l'intention de l'auteur. Cette discipline, quoique très ancienne, est mal formalisée. Le processus de résumé est en effet dépendant à la fois du type de texte à résumer et de l'utilisation qui en sera faite. Ainsi, un résumé de type rapport d'activités sera dans la forme comme dans le fond radicalement différent d'un résumé d'une oeuvre littéraire, d'un résumé d'ouvrage scientifique, d'un résumé de dépêches ou d'une revue de presse.

2.15 Conclusion

Dans ce chapitre, nous avons explicité les différentes connaissances liées au TAL, TALA et les différentes méthodes et techniques utilisées dans l'extraction automatique de mots-clés dans les textes écrits. On a commencé par présenter le TAL, après on a défini les niveaux et approches de traitement de la langue ensuite nous avons mis le point sur les différentes méthodes et approches d'extraction automatique de mots clés, enfin nous

avons présenté le traitement automatique de la langue Arabe, les difficultés qui fait face à cette dernière et nous avons cité quelques outils et produits nécessaires de TALA

Dans ce qui suit, nous détaillerons notre approche pour l'extraction automatique de mots clés, nous travaillerons sur une collection de textes arabes non voyellé et écrits en arabe standard moderne.

Chapitre 03

Conception de Système

3 Chapitre 03 : Conception de Système

3.1 Introduction

L'extraction automatique de mots-clés est la tâche d'extraire un petit ensemble de mots, de phrases clés ou des segments clés d'un document qui permettent de décrire la signification du document. Un système d'extraction est efficace s'il permet d'extraire une liste d'expressions langagières qui représente globalement (assure une couverture thématique) le sens du document (texte ou article).

Notre système est basé sur une méthode d'extraction appliquée aux textes arabes, afin d'extraire automatiquement des mots clés à partir d'un document textuel, notre système suit tout un processus qui commence par une étape de segmentation et filtrage en passant par les traitements linguistiques (normalisation et lemmatisation) ensuite la pondération des termes et enfin la sélection des mots clés.

3.2 L'Approche proposée

Nous proposons une approche qui est une approche extractive, hybride et statistique pour l'extraction automatique des mots-clés dans des documents textuels.

Notre approche est qualifiée hybride par rapport à la combinaison du critère statistique et d'autre linguistique pour la détection des mots clés, le critère statistique est la distribution des occurrences des termes (fréquences). L'indice linguistique est la catégorie grammaticale (type) des termes.

L'importance des termes est quantifiée en se basant sur la fréquence d'occurrences qui se repose sur le calcul des fréquences d'apparition des termes dans chaque document. Un score final est attribué à chaque terme du texte en fonction de sa fréquence d'apparition dans le textes

Les termes sont classés en fonction de leurs scores finaux. Les scores les plus élevés déterminent les mots-clés candidats potentiels. Ensuite une étape de sélection est faite pour sélectionner les mots clés candidats selon des règles de sélections que nous avons suivies.

Enfin, le nombre final de mots clés est déterminé selon un taux de compression (nombre maximal de mots clés) qui peut être défini par l'utilisateur du système. Le nombre moyen

de mots clés dans les articles de notre corpus est 5, cette valeur est considérée comme valeur par défaut.

3.2.1 Caractéristiques du corpus

Notre corpus d'étude, de test est formé d'une collection documents qui sont collectés du site d'al Jazeera [24] et regroupé dans trois catégories : économie, technologie et scientifique. Les textes sont écrits en arabe standard moderne, un système d'écriture simple, et non voyellé. Le corpus est constitué d'environ 60 documents textuels et le nombre moyen de mots par document est environ 600 mots. Le tableau 3.1 qui suit montre les différentes caractéristiques de notre corpus

Catégories	Nombre de documents	Nombre moyens de mots dans un document	Exemples de Titres des documents
Technologie	20	[100-300] avec une moyenne 195 mots	1- مايكروسوفت أشعلت الإنترنت بإعلانها عن ويندوز 1.. الآن اتضح السبب 2- سكايب يصل إلى سماعات أمازون إيكو 3- هذا هو التصميم الجديد لهاتف آيفون 2019 4- أمازون تُقصي آلاف التجار من متجرها الإلكتروني بدون إشعار 5- فيسبوك تطور تقنية تبلغ تلقائيا بالمواد غير اللائقة
Economie	20	[220-420] avec une moyenne de 345 mots	1- كيف فشلت أميركا في ضبط صادرات النفط الإيرانية؟ 2- مجموعة العشرين تدعو لزيادة النفقات لإنعاش النمو 3- حيتان البتكوين.. من أكبر مُلاك هذه العملة الرقمية؟ 4- غلوبال فيلاج سبايس: استثمارات قطر النفطية بسواحل غويانا مجرد بداية 5- آي-دينار".. أول منصة إلكترونية إسلامية لتبادل العملة " الرقمية
Science	20	[250-550] avec une moyenne de 362 mots	1- اكتشاف السيانيد في النيازك يقدم أدلة على أصل الحياة 2- غدد جلدية تسمح لأنواع من أسماك القرش بالتوهج 3- خبراء البيئة يحذرون: مليون نوع بيولوجي مهددة بالانقراض 4- مضادات حيوية من جلد الضفدع البني

Tableau 3.1 : Caractéristique du corpus

En constituant cette collection de documents, nous avons pris en considération les traits linguistiques suivant :

Une liste de mots clés est associée à chaque texte, cette liste est défini par le ou les auteurs du texte : c'est la raison pour laquelle nous avons choisi ce type de texte avec les mots clés , en effet, la rareté ou même l'absence de corpus arabes annotés avec mots clés nous a poussé à choisir un type d'article annoté par l'auteur lui-même, cette annotation nous permettra non seulement de tirer quelques informations linguistiques sur les mots clés mais aussi d'évaluer notre système par rapport à la liste des mots clés référence défini par l'auteur.

La figure 3.1 montre un exemple de texte annoté avec des mots clés effectuer par auteur

قال الرئيس الصيني شي جين بينغ في افتتاح قمة مجموعة الدول العشرين صاحبة الاقتصادات الكبرى في العالم اليوم الأحد، إن المخاطر تتزايد في الاقتصاد العالمي بسبب ارتفاع مستويات الاستدانة، ودعا دول العالم إلى تجنب وضع الحواجز أمام التجارة.

وتبحث القمة في دورتها الـ 11 التي تنعقد في مدينة خانجو شرقي الصين سبل تحقيق نمو اقتصادي عالمي قوي ومستدام. ومن المنتظر أن يجدد قادة المجموعة تعهداتهم باستخدام السياسات الضريبية والإنفاق لتنشيط الاقتصاد العالمي المتباطئ، لكن ليس مرجحًا أن تتبنى قمة العشرين مبادرات جديدة لتعزيز النمو.

ويمثل السعودية في قمة العشرين الأمير محمد بن سلمان ولي ولي العهد السعودي الذي سيعرض رؤية المملكة لمستقبل اقتصادها المحلي ودورها في الاقتصاد العالمي، وفق "رؤية 2030" التي أعلنتها الرياض في أبريل/نيسان الماضي.

وتتضمن الخطط السعودية إنفاقا حكوميا بقيمة 270 مليار ريال (72 مليار دولار) خلال السنوات المقبلة على مشروعات لتنويع اقتصاد البلاد. ومن بين الموضوعات المطروحة للنقاش في القمة فائض الطاقة الإنتاجية في صناعة الصلب العالمية، وهي نقطة حساسة بالنسبة للصين أكبر منتج للصلب في العالم، بالإضافة إلى معوقات الاستثمار الأجنبي والتجارة العالمية، ومخاطر خفض قيمة العملات لحماية أسواق الصادرات.

من جانب آخر، قال رئيس المفوضية الأوروبية جان كلود يونكر اليوم إنه من الضروري أن تضع الصين آلية لمعالجة مشكلة فائض الطاقة الصناعية لديها، مضيفا أن "من غير المقبول" أن تفقد صناعة الصلب الأوروبية عدد الوظائف الذي فقدته في السنوات الأخيرة بسبب فائض الإنتاج الصين

الكلمات المفتاحية : الكجار، اقتصاد عالمي، قمة عشرين، خانجو، صين.

Figure 3.1 : Exemple sur un texte annoté avec des mots clés

La liste des mots clés définie par l'auteur est composée généralement de syntagmes nominaux (noms, adjectifs, adverbes), l'utilisation des verbes est très rare ou même inexistantes : La connaissance de la catégorie grammaticale des mots est indispensable est importante pour un système de détection automatique de mots clés, pour cette raison nous devions choisir un étiqueteur grammatical qui permet de faire l'étiquetage

des mots (POS Tagging). Nous avons choisi un analyseur morphologique « Snowball » qui utilise un modèle récent et puissant (sortie en 2016) et qui va nous permettre d'avoir un bon résultat lors de l'étape de lemmatisation et un étiqueteur grammatical « Stanford Pos Tagger » afin d'attribuer une catégorie grammaticale à chaque mot du texte.

Pour expliquer un thème, l'auteur débute ses paragraphes avec les mots les plus saillants, ces mots sont généralement ceux cités dans la liste des mots clés.

3.3 Architecture du système

Notre système d'extraction automatique de mot clés est fondé principalement sur des techniques d'extraction. La mise en œuvre fonctionnelle de notre système est représentée à la Figure 4.1 Pour extraire automatiquement une liste de mots clés d'un texte, une série de traitements est exécutée, en commençant par la segmentation, filtrage, normalisation, lemmatisation ensuite le calcul des fréquences et étiquetage grammatical et enfin la sélection des mots clés et la définition des catégories des textes à partir des mots-clés extraits.

La figure suivante (Figure 3.2) représente l'architecture globale de notre système :

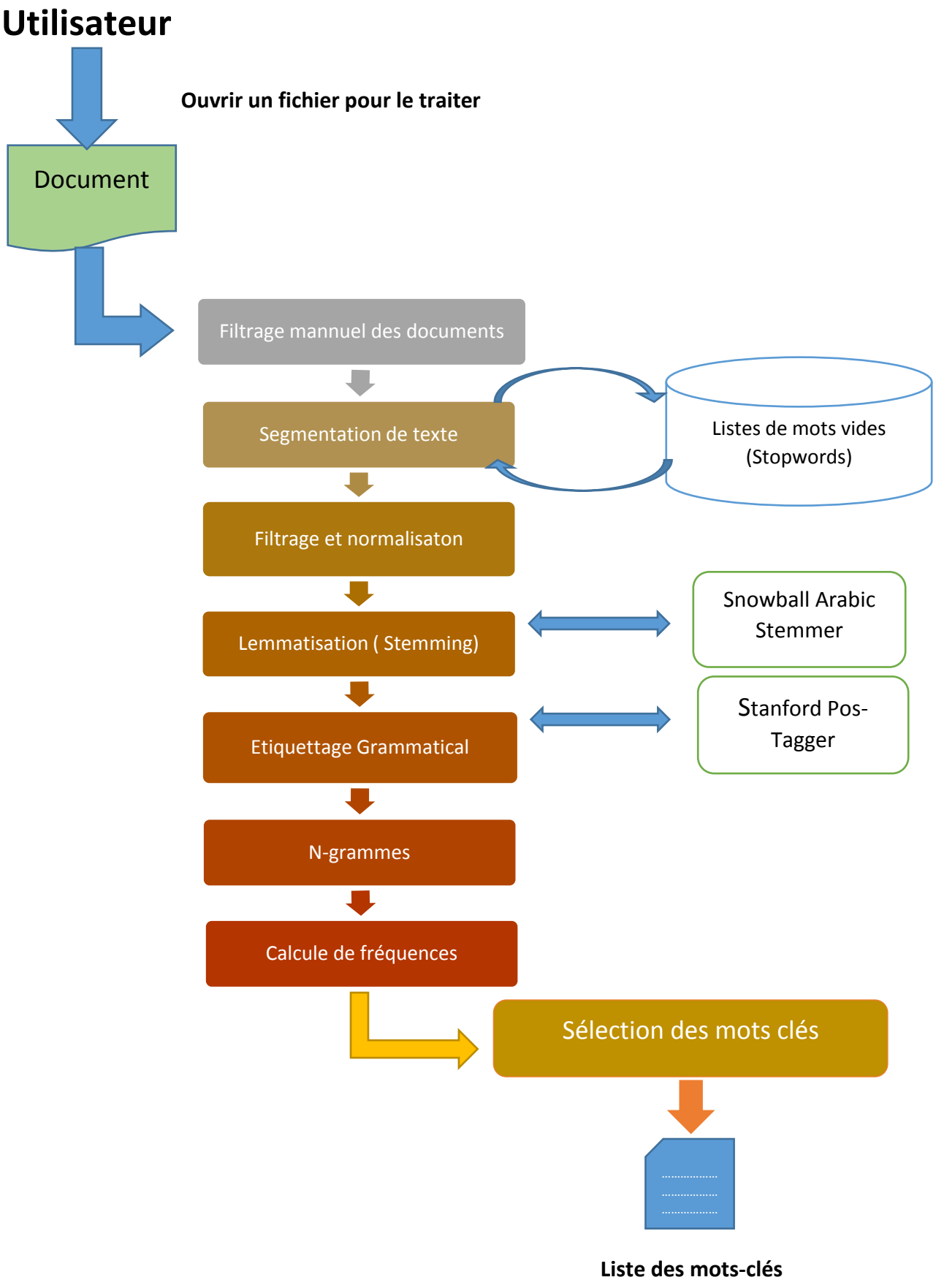


Figure 3.2: Architecture du système.

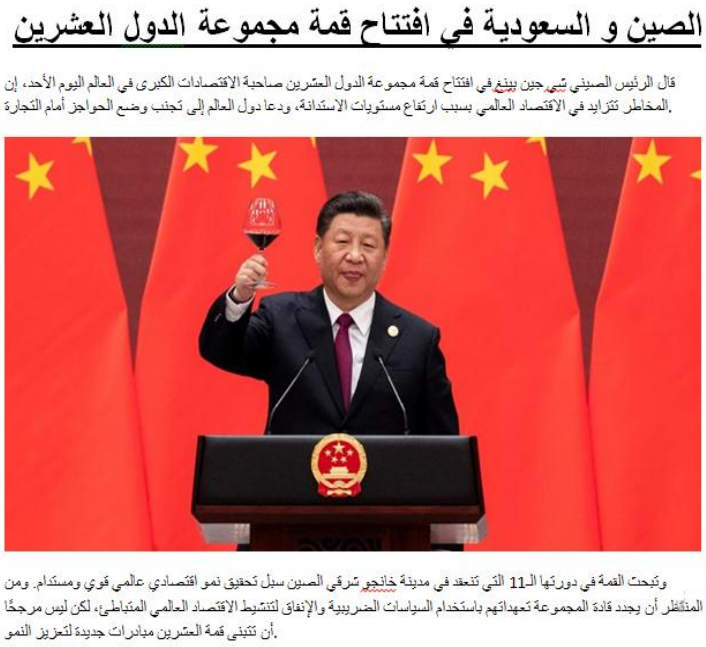
3.3.1 Prétraitements

3.3.1.1 Filtrage manuel des documents

Le filtrage manuel des documents est primordial car seulement le contenu textuel dans le document qui nous importe en phase de traitement. Le filtrage débute par la suppression des figures, images, les schémas inutiles, les adresses mails et les liens URL ensuit la conversion d'extension d'un article de type (.doc) fichier word à un fichier texte (.txt).

Exemple :

- La figure 3.3 montre le contenu du document avant le filtrage manuel


الصين و السعودية في افتتاح قمة مجموعة الدول العشرين

قال الرئيس الصيني تِبي جين بينغ في افتتاح قمة مجموعة الدول العشرين صاحبة الاقتصادات الكبرى في العالم اليوم الأحد، إن المخاطر تتزايد في الاقتصاد العالمي بسبب ارتفاع مستويات الاستدانة، ودعا دول العالم إلى تجنب وضع الحواجز أمام التجارة.

وتبحث القمة في دورتها الـ11 التي تنعقد في مدينة خانجو شرقي الصين سبل تحقيق نمو اقتصادي عالمي قوي ومستدام. ومن المنتظر أن يجدد قادة المجموعة تعهداتهم باستخدام السياسات الضريبية والإنفاق لتنشيط الاقتصاد العالمي المتباطئ، لكن ليس مرجحًا أن تتبنى قمة العشرين مبادرات جديدة لتعزيز النمو.

Figure 3.3 : Le contenu du document avant le filtrage manuel [24]

- Contenu du document après le filtrage manuel

"قال الرئيس الصيني شي جين بينغ في افتتاح قمة مجموعة الدول العشرين صاحبة الاقتصادات الكبرى في العالم اليوم الأحد، إن المخاطر تتزايد في الاقتصاد العالمي بسبب ارتفاع مستويات الاستدانة، ودعا دول العالم إلى تجنب وضع الحواجز أمام التجارة.

وتبحث القمة في دورتها الـ 11 التي تنعقد في مدينة خانجو شرقي الصين سبل تحقيق نمو اقتصادي عالمي قوي ومستدام. ومن المنتظر أن يجدد قادة المجموعة تعهداتهم باستخدام السياسات الضريبية والإنفاق لتنشيط الاقتصاد العالمي المتباطئ، لكن ليس مرجحًا أن تتبنى قمة العشرين مبادرات جديدة لتعزيز النمو"[24]

3.3.1.2 Encodage des textes

L'arabe est l'un des systèmes d'écriture le plus utilisé au monde, elle est écrite dans un style cursif qui est lu et écrit de droite à gauche. Chaque lettre arabe peut éventuellement adopter plusieurs formes.

Chaque caractère de l'écriture arabe a son propre ensemble de règles de jonction et peut, ou non, changer de forme d'aspect lorsqu'il a un autre caractère à sa gauche, à sa droite ou à sa gauche et à sa droite.

L'encodage unique des textes en format standard, permet de représenter les textes sans aucune déformation au niveau de caractère lors de la sauvegarde, de la lecture ou bien de l'écriture. Tous les textes de notre corpus sont représentés avec un encodage UTF-8.

3.3.2 Segmentation de texte

La segmentation que l'on peut l'appelée aussi tokénisation de texte ou analyse lexicale consiste à segmenter (diviser) le texte en segments. Cette fonction permet de couper les morceaux de texte comme les paragraphes en phrases, les phrases en mots, etc. Ces segments peuvent être introduits dans un analyseur morphologique pour un traitement ultérieur. Dans notre approche nous avons opté à une segmentation en token. Une frontière d'un token est facilement détectée par un blanc (espace) et/ou un saut de ligne.

Exemple :

- **Texte Arabe avant segmentation**

  (قال الرئيس الصيني شي جين بينغ في افتتاح قمة مجموعة الدول العشرين صاحبة)

- **Texte Arabe après segmentation**

  (قال ، الرئيس ، الصيني ، شي ، جين ، بينغ ، في ، افتتاح ، قمة ، مجموعة ، الدول ، العشرين ، صاحبة)

3.3.3 Filtrage

Cette étape consiste à éliminer tous les mots non significatifs. Pour chaque mot reconnu, on le compare avec un des éléments dans la base de données qui contient les mots vides ou « Stop words » (les mots non-significatifs) comme (كان, بعد, أن,…) si un mot en fait partie on le supprime. Les mots vides sont très courants dans un texte, cela n'aide pas lors de la sélection de mots clés car leurs fréquences d'apparition dans le texte vont être élevées.

Exemple :

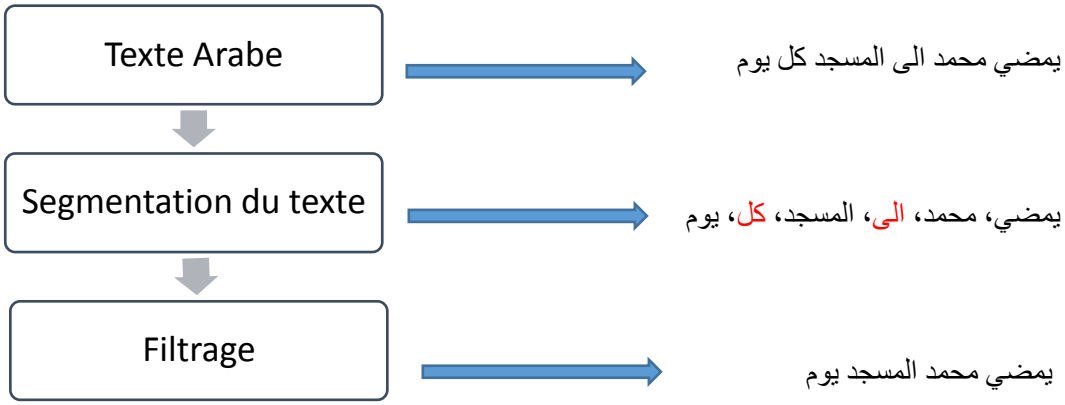


Figure 3.4 : Exemple représentatif du processus de filtrage

La base de données utilisée pour la comparaison afin de supprimer les stop words est incluse dans la bibliothèque de NLTK et nous avons rajouté un fichier qui contient au environ 200 autres stop words qui complètent la base de données de NLTK, car on a remarqué, en faisant les tests, que pas tous les mots vides sont éliminés. Après l'identification de ces derniers, nous les avons mis dans un fichier et rajouté avec le traitement afin d'avoir le résultat voulu. Le tableau 3.2 suivant montre les 200 stop words (mots vides) rajouter

Mots vides (Stop Words)							
بؤسا	وهذا	والذي	هو	نوفمبر	مساء	ليت	قال
بإن	وهو	وان	هي	نون	مع	مادام	قام
بئس	وهي	واهاً	هيا	نيسان	معاذ	ليس	قبل
باء	وَيْ	واو	هيهات	نيف	معه	ليسب	قد
بات	وُشْكَانَ	واوضح	هيّا	نَخْ	مقابل	م	قرش
باسم	ى	وبين	هَؤلاء	نَّ	مكانكم	مئة	قطّ
بان	ي	وثي	هَاتانِ	ه	مكانكما	مئتان	قلما
بخ	ياء	وجد	هَاتَيْنِ	هؤلاء	مكانكنّ	ما	قوة
بد	يفعلان	وراءَك	هَاتِه	ها	مكانَك	ما أفعله	ك
بدلا	يفعلون	ورد	هَاتِي	هاء	مليار	ما انفك	كأن
برس	يكون	وعلى	هَجْ	هاكَ	مليم	ما برح	كأنّ
بسبب	يلي	وفي	هَذا	هبّ	مليون	مائة	كأيّ
بسّ	يمكن	وقال	هَذانِ	هذا	مما	ماانفك	كأيّن
بشكل	يمين	وقالت	هَذَيْنِ	هذه	من	مابرح	كاد
بضع	ين	وقد	هَذِه	هل	منذ	ماذا	كاف
بطآن	يناير	وقف	هَذِي	هللة	منه	مارس	كان
بعد	يوان	وكان	هَيْهات	هلم	منها	مازال	كانت
بعدا	يوليو	وكانت	و	هلّا	مه	مافتئ	كانون

بعض	واكد	ولا	و6	هم	مهما	ماي	كثيرا
بغتة	والتي	ولايزال	وأبو	هما	نحن	مايزال	كذا
بل	ومن	ولكن	وأن	همزة	نحو	مايو	كذلك
بلى	وهب	ولم	وا	هن	نعم	متى	كرب
بن	بهذا	وله	واحد	هنا	نفس	مثل	كسا
به	بين	وليس	واضاف	هناك	نفسه	مذ	كل
بها	بَسْ	ومع	واضافت	هنالك	نهاية	مرّة	كلتا

Tableau 3.2 : Les 200 mots vides rajouté.

3.3.4 Traitement linguistique

3.3.4.1 Normalisation

Pour gérer les variations du texte qui peuvent être représentées en arabe, on applique différents genres de normalisation sur le texte. Par exemple, dans l'arabe écrit, on peut parfois trouver quelques voyelles présentes avec les mots. Donc, leur élimination est nécessaire. La normalisation met tous les mots sur un pied d'égalité et permet au traitement de se dérouler de manière uniforme. L'une des raisons pour ce prétraitement est que l'on a fréquemment tendance à mal écrire ces différentes formes de hamza. Ce genre d'erreurs est très répandu dans les textes arabes. Par exemple, le mot « أكل » est généralement écrit « اكل ». Aussi la lettre « ة » à la fin des mots qui peut être écrite de deux façons : « ة » ou « ه ». Les deux mots arabes « عادة » et « عاده » signifient le même mot (habitude) malgré que leur dernière lettre soit représentée différemment.

Afin de préparer les mots au prochain traitement (lemmatisation), ces derniers sont normalisés en effectuons les remplacements des lettres suivantes :

- Enlever la ponctuation et les chiffres.
- Retirer les Signes diacritiques (principalement voyelles faibles).
- Retirer les non-lettres arabes.
- Remplacer le إ ou le أ initial par l'alif nu ا.
- Remplacer le آ par le ا.
- Remplacer le ىء d'ordre par le ى
- Remplacer le ة final par le ه

La liste des signes de ponctuation, des signes diacritiques, et des non-lettres est celle incluse dans la bibliothèque NLTK.

3.3.4.2 Stemming (lemmatisation) :

Pour un mot significatif normalisé, on applique une lemmatisation en utilisant l'analyseur Snowball Stemmer qui consiste à détecter le stem (lemme) d'un mot et supprimer les éléments flexionnels (préfix et suffixes), ceci permet de retourner une liste d'items appelés *tokens.* Cette étape est en effet très importante pour le calcul de distribution des mêmes termes.

Exemple :

Mot	**Stem**
يسألون	سأل
Ils demandent	Il a demandé
المتاجر	متجر
Les magasins	Magasin

Tableau 3.3: Exemple de lemmatisation (Stemming)

3.3.4.3 Etiquetage grammatical (POS-TAG) :

Dans cette étape, l'étiquetage grammatical est très important et utile pour la suite des processus de traitements. L'étiquetage grammatical a pour objectif d'attribuer à chaque unité lexicale sa catégorie grammaticale (nom, verbe, …) à l'aide de l'analyseur Stanford-Postagger. Les étiquettes (catégories grammaticale) attribuées à chaque unité lexicale seront réutilisées pour sélectionner les N-grammes possibles. Le but est d'éliminer les entités qui ne vont pas être considérer comme des mot clés ou alors qui le sont rarement.

La figure 3.5 suivante montre une exemple d'étiquetage grammaticale

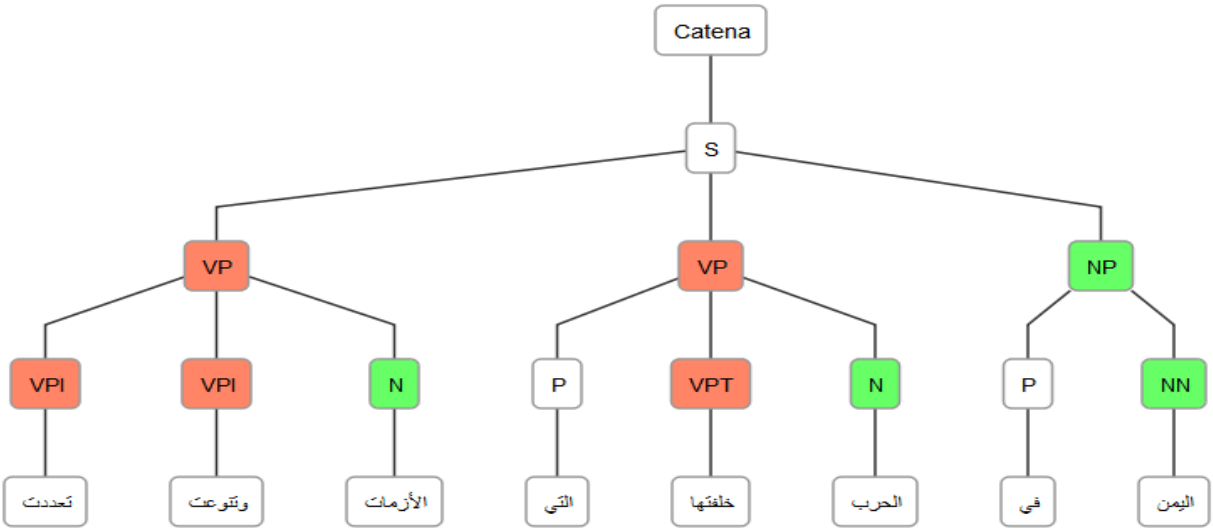


Figure 3.5 : Exemple d'étiquetage grammatical d'une phrase

A la fin de l'étiquetage, on a éliminé les verbes, les déterminants, les adjectifs numéraux, les adjectifs comparatifs et les pronoms personnelles.

3.3.5 N-grammes

Un mot clé est plus souvent trouvé comme une combinaison de noms et / ou d'adjectifs. En outre, le nombre de termes autorisés dans un mot clé est souvent limité à trois mots. De méme que, chaque phrase est traitée afin d'extraire tous les N-grammes possibles qui constituent une séquence de mots qui sont côte à côte avec une longueur maximale de trois mots. Seuls les N-grammes dont les membres sont étiquetés en noms ou des adjectifs sont conservés. La Figure 3.6 montre les unigrammes, les bi-grammes et les trigrammes d'une phrase.

Une phrase en Arabe (entrée) :	قام الرئيس الامريكي بزيارة الى المملكة الاردنية الهاشمية
Une phrase en Anglais (entrée): The American president visited the Hashemite Kingdom of Jordan	
Tokenization:	قام | الرئيس | الامريكي | ب | زيارة | الى | المملكة | الاردنية | الهاشمية
Unigrams:	الرئيس - الامريكي - زيارة - المملكة - الاردنية - الهاشمية
Bi-grams:	الاردنية الهاشمية - المملكة الأردنية - الرئيس الأمريكي
Tri-grams :	المملكة الاردنية الهاشمية

Figure 3.6: Exemple sur les N-grammes

3.3.6 Analyse statistique

3.3.6.1 Pondération des N-grammes

Cette phase consiste à associer à chaque terme dans la liste de N-grammes un score en fonction de la distribution de ses occurrences dans le texte. Le calcul du score attribué à chaque terme se repose sur sa fréquence d'apparition dans le texte. Afin de calculer l'importance de ce dernier dans un document, nous avons utilisés deux méthodes la première « Fréquence d'occurrences » et la deuxième « TF-IDF »

- **Fréquence d'occurrences**

L'attribution d'un score à chaque terme a été effectué en calculant la fréquence d'apparition de ce dernier dans la liste de N-grammes. Ensuite, Un classement a été réalisé selon l'ordre décroissant des scores des termes. A la fin, on a pris les 10 premiers termes qui ont les fréquences les plus élevé dans la liste des N-grammes

$$Poid(terme) = \sum terme \in D \qquad (1)$$

D : Le document analysé

- **TF-IDF (Term Frequency-Inverse Document Frequency)**

Notant que ***TF-IDF*** calcule le poids d'un terme dans un document par rapport à une collection de documents(Corpus).

$$TF - IDF(terme) = TF(terme) \times Log\left(\frac{N}{DF(terme)}\right) \qquad (2)$$

TF : représente le nombre d'occurrences d'un terme dans le document analysé.

DF : représente le nombre de documents dans lequel le terme est présent.

N : étant le nombre total de documents.

Pour cette méthode « TF-IDF » on a suivis les mêmes étapes de classement et le choix des 10 premiers termes que la première méthode « Fréquence d'occurrences **»**

Le tableau 3.4 suivant montre un échantillon des résultats obtenu

Mots	Poids (fréquences d'occurrences)	Mots	Poids (TF-IDF)
قمة	7	قمة	0.007593549440172498
اقتصاد	6	عالمي	0.006508756663004999
عالمي	6	اقتصاد	0.006508756663004999
عشرين	6	عشرين	0.005423963885837499
تجارة	5	تجارة	0.004339171108669999
اقتصاد, عالمي	4	اقتصاد, عالمي	0.004339171108669999
صين	3	قمة, عشرين	0.003254378331504995

Tableau 3.4: Exemple sur les résultats obtenu en utilisant les deux méthodes de pondération

3.3.6.1.1 Sélection des mots clés

La liste des n-gramme est réorganisée en fonction de leurs scores car les scores les plus élevés déterminent les mots-clés candidats potentiels. Le nombre de mots-clés extraits est défini par l'utilisateur. Nous initialisons ce taux à 5 (5 mots clés), ce nombre est généralement représentatif dans la collection de textes de notre corpus.

Notant qu'un mot clé n'est pas toujours un seul terme, il peut être composé de deux ou trois termes alors la sélection des mots-clés se fait selon les règles suivantes :

- Si deux N-grammes ont le même score, le plus long sera sélectionné.
- Si un N-gramme est sélectionné, toutes les combinaisons possibles de ses composants seront supprimées de la liste des N-grammes afin de garantir qu'un mot-clé extrait ne sera pas inclus dans un autre.

3.4 Classification de textes à partir des mots clés

La classification automatique de textes aussi appelée catégorisation automatique de textes consiste à apprendre à une machine de classer un texte dans la bonne catégorie en se basant sur son contenu. Dans notre projet on a choisi une méthode de classification

qui basé sur l'apprentissage automatique. Un algorithme d'apprentissage automatique peut apprendre les différentes associations entre des éléments de texte et qu'un résultat particulier (par exemple, des catégories) est attendu pour une entrée particulière (par exemple, du texte).

- Pour identifier la catégorie ou la classe à laquelle un texte est associé, on a suivi plusieurs étapes. Ces derniers concernent principalement de la classification avec l'apprentissage automatique et la prédiction des tags(catégories).

Dans la classification standard les données d'apprentissage se sont des textes catégorisés par contre dans notre projet nous avons utilisé des mots clés catégorisés (mots clés étiquetés) au lieu d'utilisé des textes catégorisés

Le processus comprend deux étapes :

- **Première étape : Classification avec l'apprentissage automatique**

Cette étape consiste à, premièrement donner une représentation numérique sous forme d'une matrice (ensemble de vecteurs) pour chaque mot clés étiqueté avec sa catégorie (par exemple, scientifique, technologique) et cette représentation va servir comme des données d'apprentissage (train data) a alimenté pour l'algorithme d'apprentissage automatique afin de produire un modèle de classification.

Une fois le model formé avec suffisamment d'échantillons d'entraînement, le modèle d'apprentissage automatique peut commencer à faire des prédictions précises sur les catégories.

- **Deuxième étape : Prédiction des catégories (Tags)**

La classification naïve bayésienne basée sur le théorème de Bayes est une classification probabiliste. Le classifier bayésien utilisé calcule la probabilité d'appartenance d'un texte à la catégorie x, il aura comme paramètres d'entrée deux évènements : En premier le modèle (de l'étape précédente) et deuxièmement le texte qu'elle devra donnée sa catégorie. Enfin pour faire la prédiction l'algorithme va calculer la probabilité d'appartenance du texte à la catégorie et identifiera la catégorie associée à ce dernier qui a eu le score le plus élevé

3.5 Conclusion

Dans ce chapitre, on a décrit notre système dont l'objectif était de concevoir un système capable d'extraire automatiquement des mots clés présents dans des textes arabes. La conception du système est divisée en 5 étapes applicables sur une collection de textes de teste que nous avons élaborer. Cette conception sera mise en fonction dans le chapitre qui suit.

Chapitre 04

Implémentation et Test

4 Chapitre 04 : Implémentation et test

4.1 Introduction

Dans ce chapitre, nous allons présenter l'implémentation de notre système d'extraction automatique de mots clés. 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.

4.2 Environnement de développement

On présente dans cette section, le langage de programmation PYTHON utilisé, et les environnements de développement.

4.2.1 Python

Python est un langage de script de haut niveau, structuré et open source. Ila été 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 fois plus court qu'un programme *C* ou *C++* (ou même *Java*), ce qui représente en général un temps de développement de 5 à 10 fois plus court et une facilité de maintenance largement accrue [29].

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

Python est un langage qui continue à évoluer, soutenu par une communauté d'utilisateurs enthousiastes et responsables, dont la plupart sont des supporters du logiciel libre. Parallèlement

à l'interpréteur principal, écrit en C et maintenu par le créateur du langage, un deuxième interpréteur, Jython, écrit en Java, est en cours de développement [25] .

4.2.2 Spyder

Spyder est un environnement scientifique puissant écrit en Python, pour Python, et 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 l'exploration de données, l'exécution interactive, l'inspection approfondie et les superbes capacités de visualisation d'un progiciel scientifique. En outre, Spyder offre une intégration intégrée à de nombreux logiciels scientifiques populaires, notamment NumPy, SciPy, Pandas, IPython, QtConsole, 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 [26].

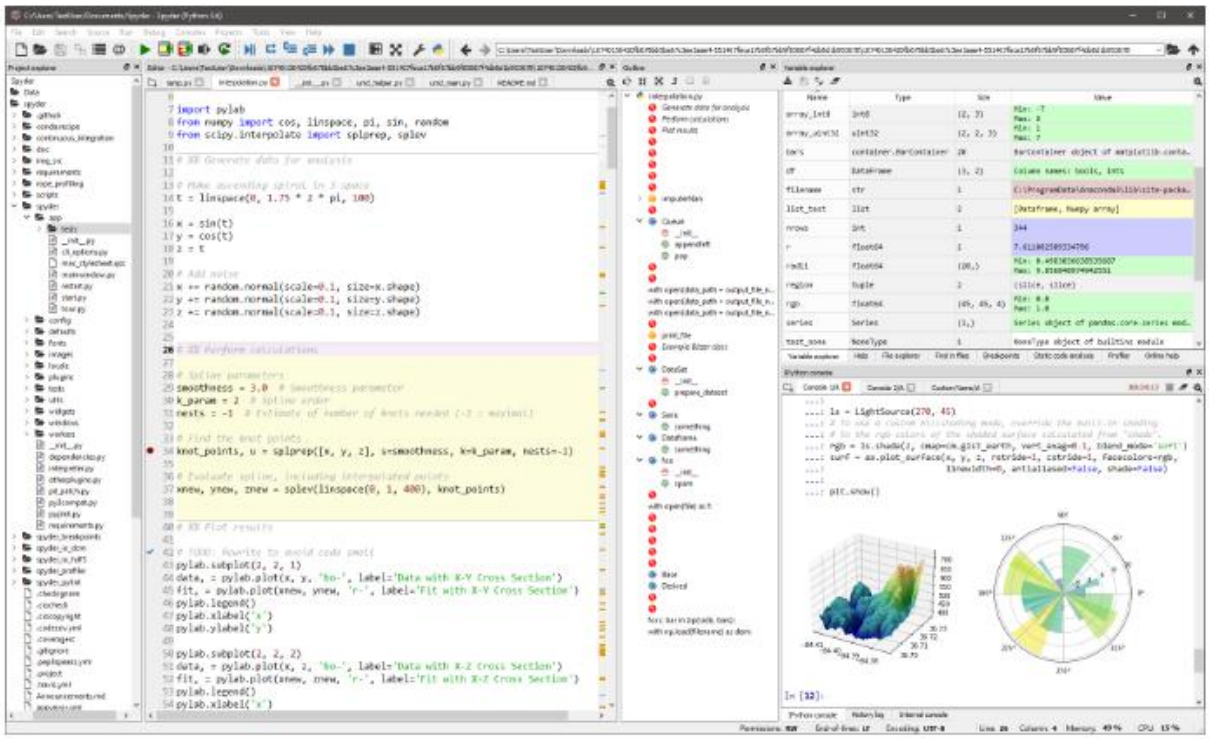


Figure 4.1 : Environnement Spyder.

4.2.3 PyQt5 interfaces graphiques

PyQt est l'une des liaisons Python les plus populaires pour le framework C ++ multi-plateformes Qt. PyQt développé par Riverbank Computing Limited. Qt est développé dans le cadre du projet Qt. PyQt fournit des liaisons pour Qt 4 et Qt 5.

PyQt est disponible en deux éditions: PyQt4, qui compilera contre Qt 4.x et 5.x et PyQt5, qui ne compilera que contre 5.x. Les deux éditions peuvent être construites pour Python 2 et 3. PyQt contient plus de 620 classes couvrant les interfaces utilisateur graphiques, la manipulation XML, la communication réseau, Animation 3D, graphiques, visualisation de données 3D et interface avec les magasins d'applications, les bases de données SQL, la navigation Web et d'autres technologies disponibles dans Qt [27].

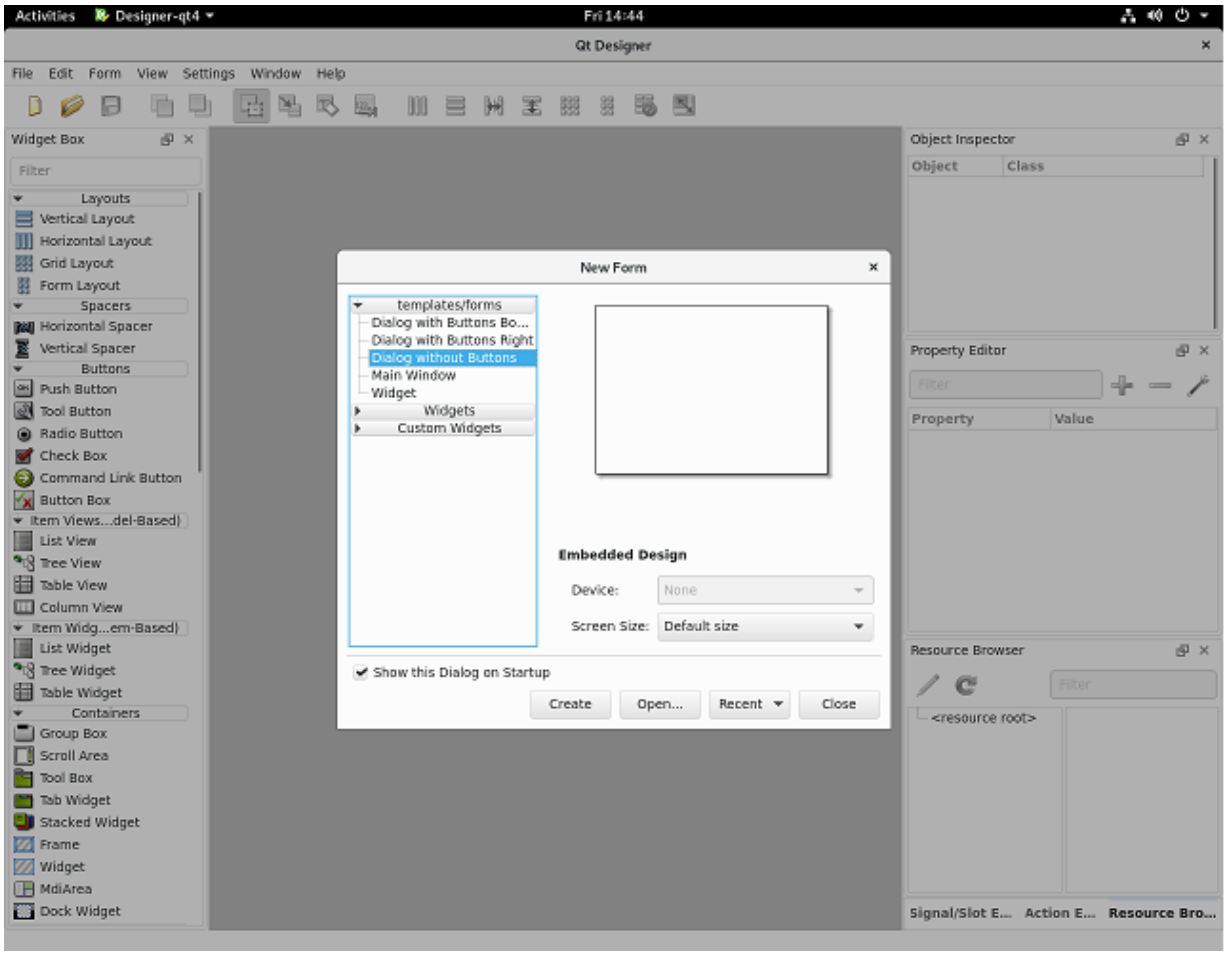


Figure 4.2 : PyQt5 designer

4.2.4 Natural Language ToolKit

NLTK est une plate-forme de premier plan pour la création de programmes Python utilisant des données en langage humain. Il fournit des interfaces faciles à utiliser avec

plus de 50 corpus et ressources lexicales tels que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la création de jetons, le suivi, le balisage, l'analyse et le raisonnement sémantique, et un forum de discussion actif.

Grâce à un guide pratique présentant les bases de la programmation, des sujets en linguistique informatique et une documentation complète sur les API, NLTK convient aux linguistes, ingénieurs, étudiants, enseignants, chercheurs et utilisateurs du secteur. NLTK est disponible pour Windows, Mac OS X et Linux. Mieux encore, NLTK est un projet gratuit, à code source ouvert et piloté par la communauté.

NLTK a été qualifié de « formidable outil d'enseignement et de travail en linguistique informatique utilisant Python » et de « formidable bibliothèque pour jouer avec le langage naturel ».

Le traitement automatique du langage avec Python constitue une introduction pratique à la programmation pour le traitement du langage. Écrit par les créateurs de NLTK, il guide le lecteur à travers les bases de l'écriture de programmes Python, du travail avec les corpus, de la catégorisation de texte, de l'analyse de la structure linguistique, etc. La version en ligne de ce livre a été mise à jour pour Python 3 et NLTK 3[28].

4.3 Description de l'application

Notre application développée en python à l'aide de l'environnement Spyder, il est muni d'une interface graphique à travers le PyQt5 Designer, l'utilisateur peut entrer des textes pour traitement et spécifié le nombre de mot clés. Pour réaliser ce projet nous avons utilisé les bibliothèques suivantes :

- **NTLK :** pour l'étape de segmentation, la normalisation, et les n-grammes.
- **Snowball stemmer:** pour l'étape de stemming.
- **Operator :** pour calcule de fréquence**.**
- **Stanford POS-Tagger** : pour l'étiquetage grammatical.
- **PyQt5 :** pour la création de l'interface graphique

4.4 Déroulement

Dans cette partie on présente les différentes étapes du processus d'extraction de mots clés par notre système, dès la sélection de textes jusqu'à l'extraction des mots clés. La figure suivante montre une la fenêtre principale de notre système.

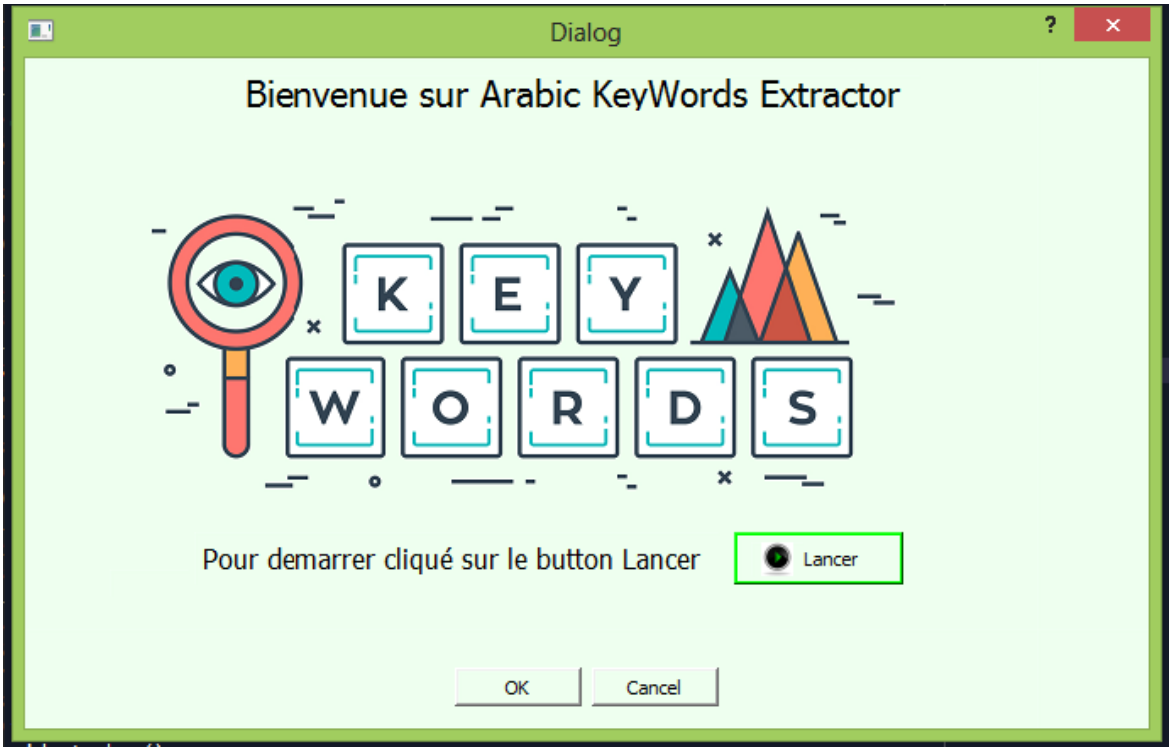


Figure 4.3 : Fenêtre principale de l'application.

4.4.1 Sélection des textes

Afin de sélectionner et utiliser les textes comme des données d'entrée pour notre système, nous avons établi un prétraitement sur les document en appliquant le processus de filtrage manuel et l'encodage en UTF-8 qui est très important. Les documents textuels sont enregistrés sous format « .txt ». Le traitement commence par la sélection d'un document pour le traiter afin d'en extraire les mots clés. Les figures suivantes montrent les étapes à suivre pour sélectionner les textes En appuy allant sur le bouton « Ouvir » cela va permettre de choisir l'un des documents pour le traiter

Etape 01 : En appuyant sur le bouton « Ouvrir » cela va permettre d'afficher la liste des documents

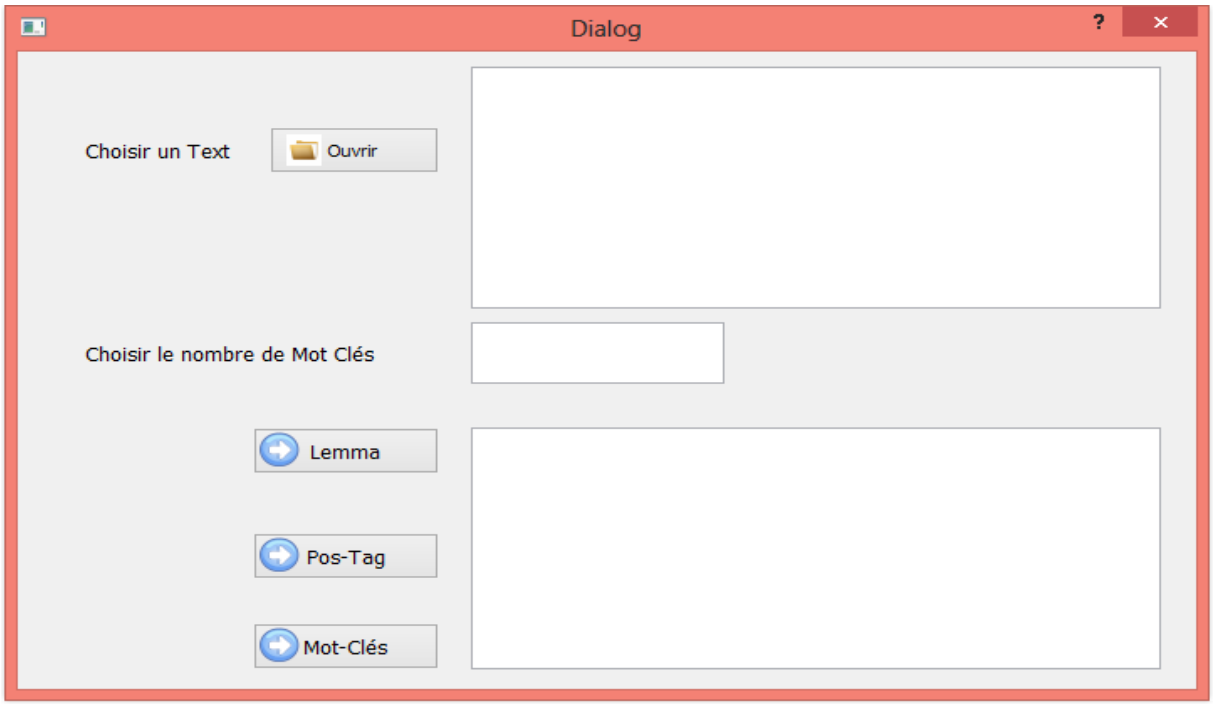
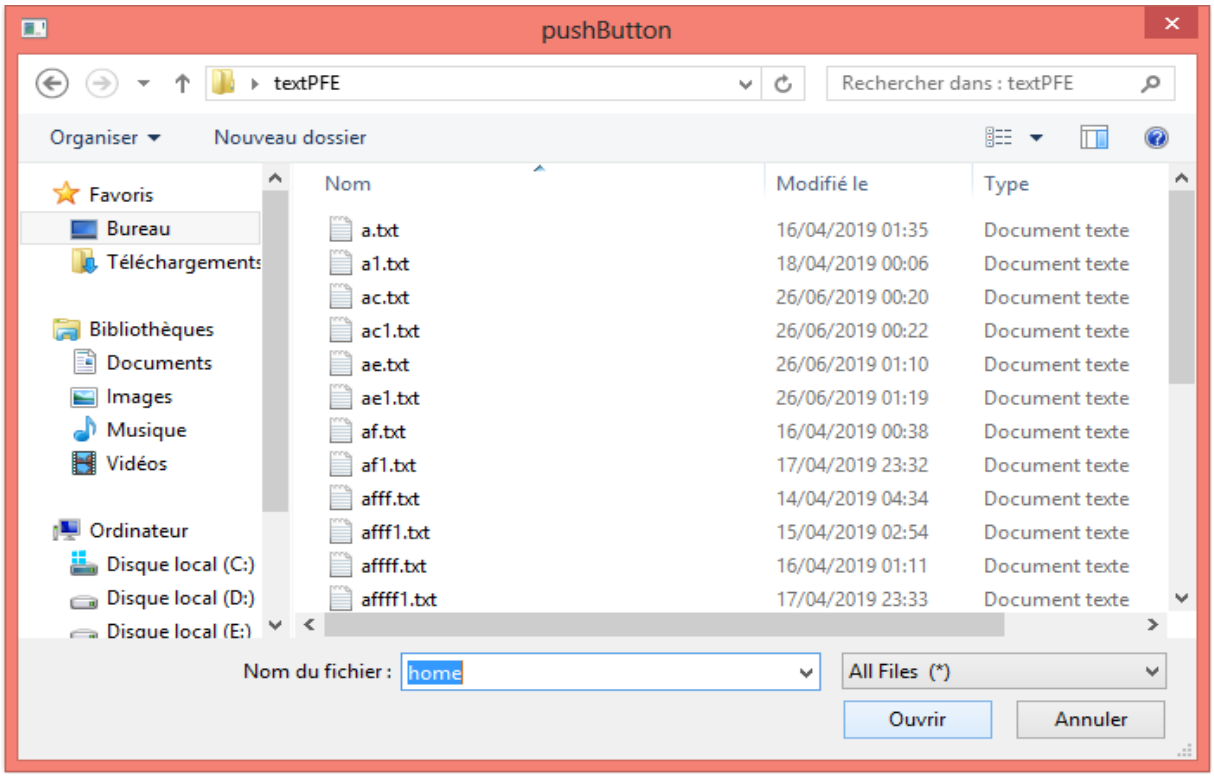


Figure 4.4: Etape 01 de la selection de textes

Etape 02 : Une liste de documents s'affiche et on valide le document choisit en cliquant sur « Ouvrir »



Figures 4.5 : Etape 02 de la sélection de texte

4.4.2 Traitement de texte

Notre système offre la possibilité d'effectuer des traitements sur le texte comme la segmentation, filtrage, lemmatisation, le pos-tag et l'affichage des textes prétraités. Les figures suivantes, indiquent les différentes étapes :

1- En premier, notre système commence par deux processus de traitements et qui s'exécute automatiquement quand on fait entré un document textuel dans ce dernier. Le premier traitement est la segmentation du texte et le deuxième est le filtrage et normalisation de texte.

2- Le deuxième traitement de texte est « La lemmatisation » : On clique sur le bouton « Lemma »

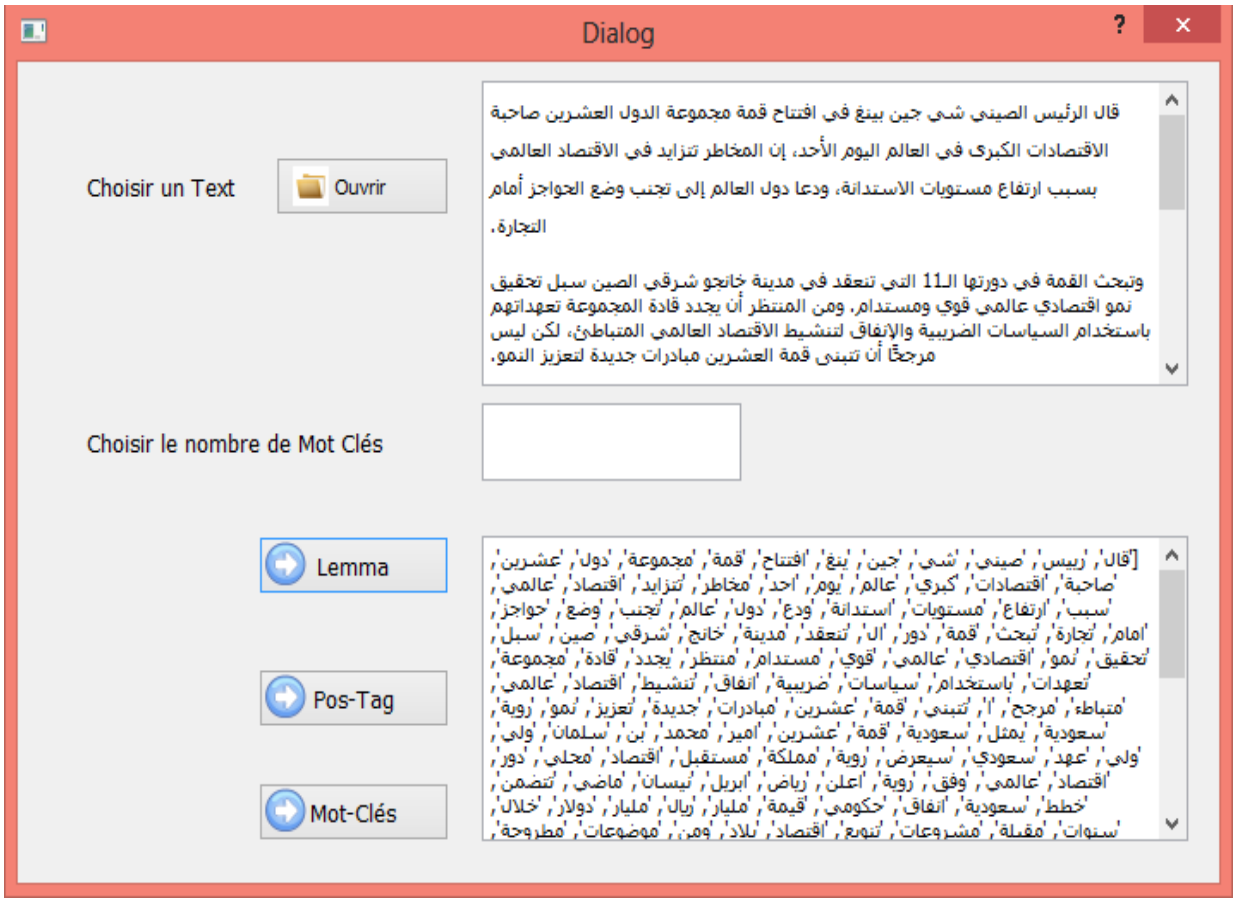


Figure 4.6 : Traitement de texte Lemmatisation.

3- Le troisième traitement de texte « L'étiquetage grammaticale Le pos-tag » : On clique sur bouton « Pos-Tag »

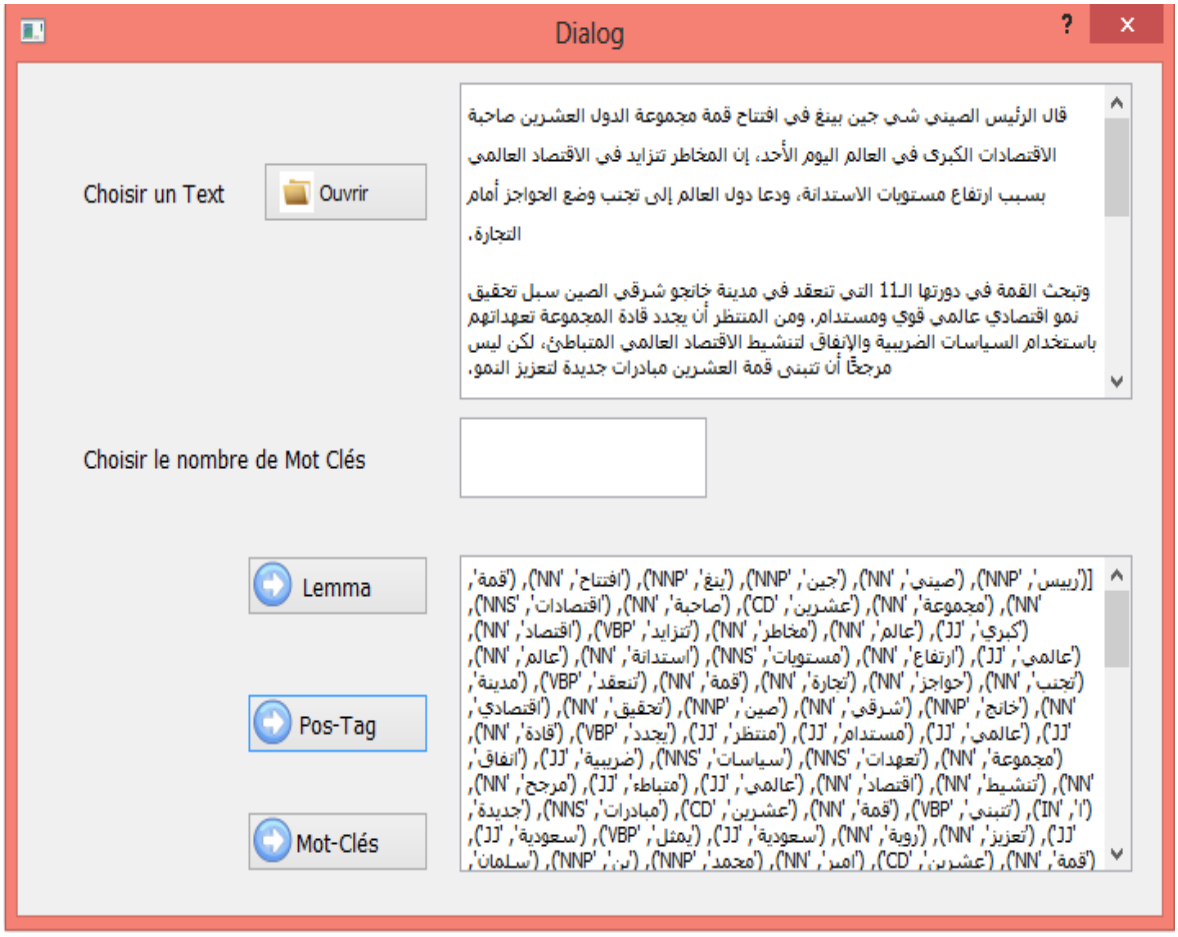


Figure 4.7 : Etiquetage grammaticale POS-Tagging

4.4.3 Extraction automatique de mot clés

Dans cette étape, on arrive à l'extraction automatique des mots clés (des expressions, mots) présentes dans les textes. Après avoir entré un document et après avoir passé par différentes étapes de traitement de textes, une simple clique sur le bouton ***Mot -Clés*** affichera la liste de mots clés.

Le nombre de mots clés favorables est initialisé à 5, cette valeur peut être aussi personnalisable.

Figure 4.8 : Extraction automatique de mots clés

4.5 Catégorisation de textes

Dans cette étape le système identifie automatiquement la catégorie du texte d'entré et cela en passant par une approche d'apprentissage automatique qui s'effectue en appliquant deux étapes principaux qui sont, la classification avec l'apprentissage automatique et la prédiction des catégories (Tags)

Figure 4.9 : Catégorisation du texte d'entrée

4.6 Evaluation du système

4.6.1 Evaluation globale du système

Nous avons fait une évaluation semi-automatique de notre système sur l'ensemble des documents de notre corpus

L'évaluation sert à comparer la liste des mots sélectionnés par le système et ceux définit préalablement par l'auteur. Nous avons fait une comparaison par mot en utilisant les mesures Rappel et Précision pour les deux méthodes de pondération, Fréquence d'occurrences et TF-IDF.

Les mesures Rappel, Précision sont calculées par les équations suivantes :

$$- Rappel = \frac{Corrects}{Corrects + Oubliés} \qquad (1)$$

$$- Précision = \frac{Corrects}{Corrects + Incorrects} \qquad (2)$$

Tel que :

*Correct*s : Nombre de mots sélectionnés par le système et par l'auteur.

Incorrects : Nombre de mots sélectionnés par le système et non pas par l'auteur.

Oubliés : Nombre de mots sélectionnés par l'auteur et non pas par le système

Nombre de documents	Rappel (frequence d'occurences)	Précision (frequences d'occurences)
1	1	0.83
2	0.8	0.8
3	0.66	0.8
5	0.66	0.8
6	0.66	0.8
7	0.71	0.83
8	0.71	0.83
9	0.71	0.83
10	0.6	0.5
11	0.66	0.66
12	0.66	0.8
13	0.66	0.57
14	0.42	0.6
15	0.42	0.6
16	0.55	0.83
17	0.6	0.6
18	0.83	1
19	0.66	0.66
20	0.57	0.66
21	0.57	0.8
22	1	1
23	0.875	1
24	0.6	0.85
25	0.66	0.8
26	0.6	0.6
27	0.85	1
28	0.83	0.83
29	0.83	0.83
30	0.71	0.83
31	0.88	1
32	0.83	1
33	0.5	0.6
34	0.75	0.85
35	0.42	0.5
36	0.57	0.57
37	0.57	0.66
38	0.42	0.6
39	0.71	0.71
40	0.83	0.83
41	0.42	0.5
42	0.83	1
43	0.66	0.8

44	0.71	0.83
45	0.87	0.87
46	0.85	0.85
47	0.75	0.85
48	0.8	0.8
49	0.85	0.85
50	0.83	0.83
51	0.8	0.8
52	0.57	0.8
53	0.83	0.62
54	0.66	0.8
55	0.6	0.6
56	0.57	0.8
57	0.85	0.85
58	0.85	1
59	0.66	0.75
60	0.6	0.8

Tableau 4.1 : Evaluation du système avec la méthode « Fréquence d'occurrences »

Nbr docs	Rappel (tf-idf)	Précision (tf-idf)
1	1	0.83
2	0.8	0.8
3	0.66	0.8
5	0.5	0.5
6	0.5	0.42
7	0.71	0.83
8	0.71	0.83
9	0.71	0.83
10	0.6	0.5
11	0.66	0.66
12	0.66	0.8
13	0.66	0.57
14	0.42	0.6
15	0.42	0.6
16	0.55	0.83
17	0.6	0.6
18	0.5	0.6
19	0.66	0.66
20	0.57	0.66
21	0.57	0.8
22	0.8	0.8
23	0.875	1
24	0.6	0.85
25	0.5	0.625
26	0.5	0.5
27	0.71	0.83
28	0.5	0.5
29	0.83	0.83
30	0.71	0.83
31	0.88	1

32	0.83	1
33	0.5	0.6
34	0.75	0.85
35	0.42	0.5
36	0.57	0.57
37	0.5	0.5
38	0.42	0.6
39	0.71	0.71
40	0.83	0.83
41	0.42	0.5
42	0.83	0.83
43	0.66	0.8
44	0.81	0.83
45	0.87	0.87
46	0.85	0.85
47	0.7	0.7
48	0.8	0.8
49	0.85	0.85
50	0.71	0.83
51	0.8	0.8
52	0.57	0.8
53	0.62	0.62
54	0.66	0.8
55	0.6	0.6
56	0.57	0.8
57	0.85	0.85
58	0.85	0.85
59	0.66	0.75
60	0.6	0.8

Tableau 4.2 : Evaluation du système avec la méthode « TF-IDF »

Afin de voir qui entre les deux méthodes « Fréquence d'occurrences » ou bien « TF-IDF » proposent meilleur résultats, on a fait une comparaison entre ces derniers. Nous avons constaté, en premier, que notre système d'extraction produit des listes de mots clés acceptables. La méthode « fréquence d'occurrence » propose meilleur résultats que la deuxième méthode « TF-IDF » , . La méthode TF-IDF est performante quand il y a un grand corpus de test à utiliser.

Par ailleurs, l'évaluation de tel système est généralement subjective car le style d'organisation des textes diffère d'un auteur à un autre et la liste des mots clés sélectionnés par l'auteur n'est pas toujours liée à l'importance des termes.

4.6.2 Evaluation système après catégorisation

Pour l'évaluation on a utilisé une seule mesure qui est la précision qui égale le pourcentage de textes prédits avec la catégorie correcte.

$$Précision = \frac{nombre\ de\ textes\ prédits\ avec\ la\ catégorie\ correcte}{nombre\ de\ textes\ total}.$$

Catégories	Précision de classification
Technologie	0.85
Science	0.91
Economie	0.797

Précision de classification Total	0.857

Tableau 4.3 : Evaluation du système après catégorisation

4.7 Conclusion

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

Notre système a pour rôle d'extraire les mots clés à partir des articles écrit en langue arabe.

Conclusion Générale

L'Extraction automatique de mot clés dans les textes arabe a pour but de détecter et d'extraire les unités les plus saillantes d'un texte. Ces unités ou termes sont généralement cachés derrière des mots, des phrases et des paragraphes. Une expression de mot clés est l'unité la plus petite à partir de laquelle les termes sont identifiés. Les mots exprimant des idées générales sur le texte (comme un résumé), la démarche d'extraction est basée, en premier lieu, sur l'extraction des segments (tokens), et ensuite, l'identification des termes candidats.

A travers notre projet, nous avons pu réaliser un système capable d'extraire automatiquement les mots clés à partir des textes écrit en langue arabe. Nous avons étudié cette langue d'un point de vue informatique en faisant ressortir les traits linguistiques et statistiques à partir du texte.

Nous avons utilisé une méthode non supervisé hybride basée sur la combinaison des critères : fréquences pondérées et catégories grammaticales des termes.
La méthode que nous avons adoptée s'est avérée adaptable et nous avons pu faire nos expérimentations et évaluations sur un corpus qui regroupe un ensemble de documents économiques, scientifiques et technologiques écrit en Arabe

Références

[1] Mohamed Hedi Maaloul, Approche hybride pour le résumé automatique de textes.Application à la langue arabe, Oct 2017

[2] Aïda KHEMAKHEM,"Arabic LDB : une base lexicale normalisée pour la langue arabe" mémoire présenté en vue de l'obtention du diplôme de MASTER en Systèmesd'Information et Nouvelles Technologies en 2 Novembre 2006, Université de Sfax, Faculté desSciences Economique et de Gestion, Tunisie.

[3]. K. Darwish: Building a shallow Arabic Morphological analyzer in one day. Proceedings of the workshop on computational approaches to semitic languages in the 40th annual meeting of the association for computational linguistics (ACL-02), Philadelphia, PA, USA.

[4] Larkey L. S., Ballesteeros L. and Connell M., improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis, in proceeding of the 25th annual international conference on research and development in information retrieval (SIGIR 2002), tampere, finland, august 2002

[5] Y.Kadri, A.Benyamina, systéme d'analyse syntaxico semantique du langage arabe,mémoire d'ingénieur, université d'Oran Es-séria, 1992

[6] Y. Kadri, Recherche d'Information Translinguistique sur les Documents en Arabe, Thèse Ph.D, Université de Montréal 2008.

[7] Mohamed Hédi Maâloul, Approche hybride pour le résumé automatique de textes. Application à la langue arabe. 18 décembre 2012.

[8] Wajdi Zaghouani, " 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 en Mars 2008, Université de Montréal.

[9]. Benoît TROUVILLIEZ, Traitement Automatique des Langues (TAL), Intelligence Artificielle (IA), Analyse sémantique et Clusterings , 31 mars 2010.

[10]. Siham Boulaknadel, 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, soutenue le 18 octobre 2008,Université de Nantes, France,.

[11]. M. A. Chéragui, Y. Hoceini ET M. Abbas, "A Morphological Analysis of Arabic Language based on Multicriteria Decision Making: TAGHIT System", IEEE, International Conference On Machine and Web Intelligence, (2010). [2] M. A. Chéragui, Y. Hoceini et M. Abbas, "Une Approche Multicritères pour lever l'ambiguïté Morphologique dans le Texte Arabe » COSI : Colloque d'optiomisation des Systèmes d'Informations, (2010).

[12] Claire Gardent, Traitement des Langues Naturelles (TAL) , Septembre 2011, ENS Cachan.

[13] Traitement automatique des langues pour l'accès au contenu des documents?Christian Jacquemin?, Pierre Zweigenbaumy? LIMSI-CNRS BP 133 91403 ORSAY Cedex FRANCE jacquemin@limsi.fr et http://www.limsi.fr/Individu/jacquemi/ y DIAM : Service d'informatique médicale, DSI/AP-HP et Département de Biomathématiques, Université Paris 6 91, bd de l'Hôpital, 75634 Paris Cedex 13 pz@biomath.jussieu.fr et http://www.biomath.jussieu.fr/~pz/

[14] François Yvon, Une petite introduction au Traitement Automatique des Langues Naturelles, Support du cour.

[15]Larkey L. S., Ballesteros L. and Connell M., Improving Stemming for Arabic Information Retrieval: Light Stemming and Cooccurrence Analysis, *In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, August 2002.*

[16]M. Aljlayl and O. Frieder, On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach, *In 11th International Conference on Information and Knowledge Management (CIKM), November 2002, Virginia (USA).*

[17] Mohamed Hedi Maaloul. Approche hybride pour le résumé automatique de textes.Application à lalangue arabe.. Traitement du texte et du document. Université de Provence - Aix-Marseille I, 2012.

[18] Joseph Dichy, Ramzi Abbès, « Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1 », Université Lumière Lyon 2, ICAR-CNRS JADT 2008 : 9emeJournées internationales d'Analyse statistique des Données Textuelles.

[19] https://snowballstem.org/

[20] Larkey Leah S, Margaret E. Connell, «Arabic Information Retrieval atUMass in TREC-10 Centre de recherché d'information Département de l'informatique, Université de Massachusetts

[21] Shreen Khoja, Porger Garside, and Gerry Knowles « A tagset for the morphosynactic tagging of Arabic ». Article présenté en corpus linguistique 2001, Université de Lancaster, UK, Mars 2001.

[22] https://nlp.stanford.edu/software/tagger.shtml

[23] http://www.tuteurs.ens.fr/faq/utf8.html

[24] https://www.aljazeera.net/

[25] https://docs.python.org/3/license.html

[26]https://docs.spyder-ide.org/

[27].Mark Summerfield's book, Rapid GUI Programming with Python and Qt,

[28] https://www.nltk.org/

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

[30] https://monkeylearn.com/text-classification/

[31] HULTH, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge.In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing

[32] MATSUO, Y. et ISHIZUKA, M. (2004). Keyword Extraction from a Single Document Using Word Co-occurrence Statistical Information.

[33] LIU, Z., LI, P., ZHENG, Y. et SUN, M. (2009). Clustering to Find Exemplar Terms for Keyphrase Extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1,

[34]FRANK, E., PAYNTER, G., WITTEN, I., GUTWIN, C. ET NEVILL-MANNING, C. (1999). Domain-Specific Keyphrase Extraction

[35] LIU, Z., CHEN, X., ZHENG, Y. et SUN, M. (2011). Automatic Keyphrase Extraction by Bridging Vocabulary Gap. In Proceedings of the 15th Conference on Computational Natural Language Learning.

[36] SUJIAN, L., HOUFENG, W., SHIWEN, Y. ET CHENGSHENG, X. (2003). News-Oriented Keyword Indexing with Maximum Entropy Principle.

[37] ZHANG, K., XU, H., TANG, J. et LI, J. (2006). Keyword Extraction Using Support Vector Machine.

[38] G. Salton, C. S. Yang, C. T. Yu, —A Theory of Term Importance in Automatic Text Analysis , Journal of the American society for Information Science, 1975.

[39] J. D. Cohen, —Highlights: Language and Domain-independent Automatic Indexing Terms for Abstracting Journal of the American Society for Information Science, 46(3): 1995

[40] M. Ortuño et al., —Keyword detection in natural languages and DNA , Europhys. 2002

[41] J.P. Herrera, P.A. Pury, —Statistical keyword detection in literary corpora , The European physical journal, 2008

[42] P. Carpena et al., —Level statistics of words-Finding keywords in literary texts and symbolic sequences , Physical Review E, 79, 03512(R), 2009

[43] Adrien Bougouin. État de l'art des méthodes d'extraction automatique de termes-clés. Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), Jun 2013, Sables d'Olonne, France.