République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieure et de la Recherche Scientifique

Université Kasdi Merbah Ouargla

Faculté des nouvelles technologies de l'information et de la communication
Département d'informatique et technologie de l'information

Mémoire de Master en Informatique

Spécialité : Informatique Industrielle

Thème

Prétraitement automatique des textes écrits dialecte Algérien

Présenté par
KHAZENE Mabrouka, SMAHI Hadjer

Devant le jury :

Président	Rouagat Wahab	Professeur	Université de Ouargla
Examinateur	Bouhyaoui Nasria	M.C.B	Université de Ouargla
Encadreur	Benkhelifa randa	M.A.A	Université de Ouargla

Année universitaire : 2021/2022

Remerciements

Tout d'abord, nous remercions Dieu tout-puissant de nous avoir donné la force et le courage pour réaliser ce modeste travail.

Nous adressons nos sincères remerciements à Benkhalifa Randa pour ses conseils, notamment sur L'aide précieuse que vous nous avez apportée tout au long du travail.

Nous remercions également chaque membre du jury pour cet honneur, qu'ils nous font en acceptant de juger nos travaux.

Avant de fermer notre page de remerciements, nous n'oublions pas de remercier nos parents pour leur patience et leur soutien inconditionnel sans lesquels nous n'aurions pas pu terminer notre travail.

Nous terminons en remerciant nos familles, nos collègues et toutes les personnes qui ont contribué de près ou de loin à leurs encouragements et leur soutien moral à l'issue de ce travail.

Résume

La langue arabe moderne standard (AMS) est la langue officielle des pays arabes dont elle est utilisée dans les officielles tels que l'éducation, et les administrations. La langue mère de ces nations est le l'arabe dialectale (AD). Cette dernière est largement utilisé dans la vie quotidienne des citoyens arabe (interactions quotidiennes, discours politiques, publicités, émissions, films, et communication à travers les médias sociaux, etc.). Depuis l'avènement des réseaux sociaux, la plupart des utilisateurs préfèrent publier, commenter, et envoyer leurs messages dans leurs langage familier local au lieu d'utiliser AMS ou une autre langue comme le français ou l'anglais. L'AD diffère totalement de l'AMS, ce qui traduit la nécessité de réaliser de nouveaux outils adéquats pour les taches de Traitement automatique du langage naturel (TALN) comme le prétraitement automatique du texte, ou encore l'analyse de sentiment.

L'objectif de ce travail est la réalisation d'un nouvel outil de prétraitement automatique des textes adapté au dialecte arabe algérien. Cet outil permet de résoudre quelques problèmes liés au TALN.

Mots clés

Le dialecte arabe Algérien, Traitement automatique du langage naturel, Traitement de texte.

Abstract

The modern standard Arabic language (MSA) is the official language of the Arab countries of which it is used in the official ones such as education, and the administrations. The mother tongue of these nations is Dialectal Arabic (AD). The latter is widely used in the daily life of Arab citizens (daily interactions, political speeches, advertisements, shows, films, and communication through social media, etc.). Since the advent of social networks, most users prefer to post, comment, and send their messages in their local colloquial language instead of using AMS or another language such as French or English. AD is totally different from AMS, which reflects the need to create new tools suitable for Automatic Natural Language Processing (NLP) tasks such as automatic text pre-processing, or even sentiment analysis.

The objective of this work is the realization of a new automatic text pre-processing tool adapted to the Algerian Arabic dialect. This tool solves some problems related to NLP.

Keywords

The Algerian Arabic dialect, Natural language processing, Text Processing.

Table des matières

Liste des tableaux

Liste des figures

Introduction générale

Introduction générale

Aujourd'hui, le monde a été témoin d'une énorme explosion de données et d'une révolution scientifique et technologique résultant de la diffusion rapide de l'information. Les réseaux sociaux tels que Facebook, Twitter et YouTube sont devenus un élément essentiel de notre vie quotidienne. En effet, ils sont de plus en plus utilisés pour véhiculer des messages et des idées en générant des tonnes de données sur les utilisateurs et leurs interactions. L'importance de ces données est qu'elles contiennent une bonne fraction de messages contenant des informations y compris les avis, les intérêts et les préférences des utilisateurs. Ces informations jouent un rôle important dans tous les domaines de la vie, économique, scientifique et social.

La langue est un moyen d'interaction, un élément essentiel pour la communication entre les humains. Parmi les langues, on trouve la langue arabe, qui est la langue officielle des pays arabes. Cependant, les habitants du monde arabe ne communiquent pas dans la langue formelle sauf dans les situations officielles, mais dans la communication quotidienne, ou en utilisant les médias sociaux, chaque pays utilise son propre dialecte. Parmi les dialectes arabes, nous trouvons le dialecte algérien, qui a rencontré à son tour plusieurs problèmes dans le traitement automatique du langage naturel (TALN), vue sa mixture (arabe, berbère, français, etc.). Le prétraitement des textes est une étape indispensable du processus d'analyse des textes en prenant les textes bruts et les transformant en un format pouvant être compris et analysé par les ordinateurs et l'apprentissage automatique. C'est pour cela qu'il est indispensable de construire un nouvel outil de prétraitement des textes écrites en dialecte algérien en permettant la résolution de quelques problèmes du TALN liés à ce langage.

C'est ce qui nous a conduit à établir ce travail, qui à son tour est divisé en trois chapitres. Où dans le premier chapitre nous présentons l'état de l'art, en abordant les domaines de l'intelligence artificielle, et du traitement automatique du langage naturel, le concept de prétraitement automatique du texte, la langue arabe et ses types, les problèmes et défis rencontrés par les chercheurs dans le traitement automatique des dialectes arabes. Quant au deuxième chapitre, nous expliquons les étapes de traitement d'un texte en dialecte algérien pour la conception du système. Dans le troisième et le dernier chapitre, nous présentons les résultats obtenus grâce à l'application issue de ce travail.

Chapitre 1
Etat de l’art

Chapitre 1 : Etat de l'art

1. Intelligence artificielle

Le terme d'intelligence artificielle est apparu pour la première fois en 1956 lors d'une conférence à l'université américaine Darmath, par le chercheur américain (Mccarthy)[1]. pour donner un signal de départ à la recherche artificielle, qu'il considère comme un domaine à part entière dans le domaine des sciences informatiques, où cette Un chercheur a présenté un concept d'intelligence artificielle À l'époque, il la définissait comme la science et l'ingénierie de la fabrication de machines intelligentes. Ce chercheur a modifié sa définition en 2007 en rapprochant le domaine de l'IA aux programmes informatiques intelligents, en déclarant : "L'intelligence artificielle est la science et l'ingénierie de fabrication de machines intelligentes, en particulier des programmes informatiques." **[1]**

- **Définition :** De nombreux chercheurs ont contribué à présenter différents concepts du terme intelligence artificielle, et cette différence est due à l'affiliation scientifique différente et à l'environnement technologique dans lequel il est né,

- **Quelques définitions**

Bodun a introduit en 1978 que l'intelligence artificielle est "la science d'obtenir des machines et des systèmes informatiques, pour effectuer des tâches qui nécessitent de l'intelligence, si elles sont effectuées par des humains".

Quant à **Charniak et McDermott**, ils ont proposé en 1985 une brève définition de « l'étude des collèges réels par l'utilisation de modèles informatiques »

Dans sa définition de l'intelligence artificielle, **Mughali** a déclaré : "L'intelligence artificielle est basée sur l'analyse de tâches intelligentes, telles que la réflexion, et enseigne de nouvelles compétences".

En général, malgré les différentes définitions Pour les raisons susmentionnés, l'intelligence artificielle peut être considérée comme une machine intelligente si elle est caractérisée par : la capacité d'apprendre, la compréhension d'ambiguïté, le traitement des données complexes, la rapidité de réagir, le réfléchissement, et la capacité d'analyser et de conclure.

2. Traitement du langage naturel

- **Définition :** Certains chercheurs ont présenté une explication du traitement du langage naturel, en tant que domaine de recherche et d'application, car il explore la manière dont les ordinateurs peuvent être utilisés pour comprendre et traiter des textes naturels. **[2] ,[3],[4],[5].**

Tel que défini **[1],** la TALN est un groupe de techniques informatiques, pour représenter et analyser des textes naturels, dont le but est de traiter un langage de type humain pour une gamme d'applications. Expliquez **[6]** que le terme TALN est généralement utilisé pour décrire la fonctionnalité des composants logiciels et du matériel dans un système informatique qui analyse le langage écrit ou parlé.

2.1. L'objectif du traitement du langage naturel

Le langage naturel est défini comme très ambigu, et la machine ne peut pas facilement le comprendre, et le but du traitement et de la programmation du langage naturel est la possibilité de dialoguer avec la machine, et sa facilité, en levant l'ambiguïté qu'est la capacité à comprendre en plus d'une manière, car le langage naturel contient plusieurs types d'ambiguïté **[7] :**

- ✓ **Ambiguïté lexicale :** C'est une ambiguïté d'un mot, un mot peut être à la fois un verbe, un nom et un adjectif, comme (sylver).
- ✓ **Ambiguïté grammaticale :** Cette ambiguïté se situe au niveau de la phrase dans son ensemble, lorsqu'elle est analysée de différentes manières. Par exemple, "L'homme a vu la fille avec l'endoscope" semble ambiguë si l'homme a vu la fille à travers l'endoscope ou l'a vue tenant l'endoscope. Ambiguïté sémantique : ce type d'ambiguïté se produit lorsque le sens du mot est mal interprété.

2.2. Etapes du traitement du langage naturel

1) Traitement morphologique

C'est la première étape de la TALN. Leur but est de diviser les parties de l'entrée linguistique en groupes de jetons pour les paragraphes, les phrases et les mots.

2) Analyse syntaxique

C'est la deuxième étape de la TALN. Le but de cette étape est double : vérifier si la phrase est bien formée ou non, et la décomposer en une structure qui montre les relations grammaticales entre les différents mots.

3) Analyse sémantique

C'est la troisième étape de la TALN. Le but de cette étape est de tirer le sens exact, ou vous pouvez dire le sens du dictionnaire à partir du texte. La faisabilité du texte est vérifiée.

4) Analyse pragmatique

C'est la quatrième étape de la TALN. L'analyse pragmatique adapte des objets/événements réels, existant dans un contexte donné avec des références à des objets obtenus lors de la troisième étape (analyse sémantique).

Le schéma suivant illustre les étapes ou étapes logiques du traitement du langage naturel :

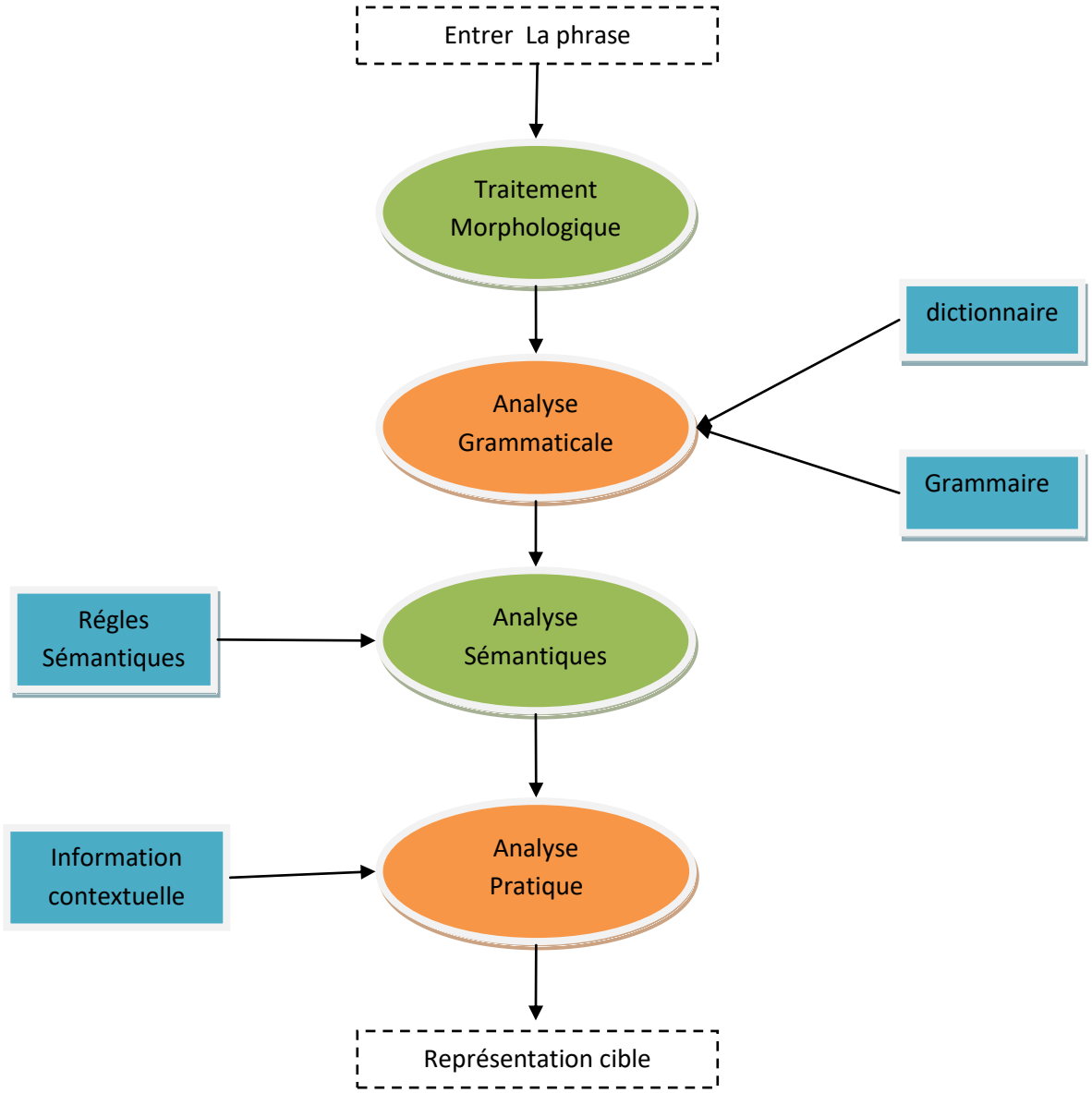

Figure 1.1 : Etapes du traitement du langage naturel

2.3. Tâches de traitement du langage naturel

Le traitement du langage naturel a de nombreuses tâches, notamment : Le traitement primaire, la traduction et la classification des voleurs, où le traitement primaire est plus important encore, à travers lequel nous pouvons effectuer d'autres tâches.

2.4. Le prétraitement de texte

Définition : le prétraitement est une technologie de transformation de fichiers qui garantit que les données brutes sont converties dans un format facile à comprendre. Ceci est assuré en passant par une série d'instructions pendant le prétraitement **[8]**

C'est l'une des branches du TALN, son objectif principal est de prendre du texte brut et de le convertir en une forme simplifié, afin qu'un ordinateur puisse le comprendre et d'extraire des informations pertinentes de ce texte. Ce prétraitement consiste à nettoyer le texte et à supprimer toutes les informations non textuelles, telles que les chiffres, les symboles, etc., et ensuite à convertir le texte en un codage compréhensible pour l'ordinateur sans problème.

3. La Langue arabe

3.1. Aperçu les dialectes arabes

Est utilisé comme un outil de transmission qui transmet la religion islamique à La genèse de la langue arabe dans la péninsule arabique classée par langues sémitiques Association, composée de: phénicien, syriaque, amharique, hébreu, etc. Le vaste patrimoine littéraire arabe, qui remonte à avant l'Islam (Ve et VIe siècles), est né dans la vaste aire géographique s'étendant du Moyen-Orient (Asie) à l'Afrique du Nord (Afrique), dominée par la langue arabe. expansion géographique et régionale et les conquêtes islamiques ont conduit à la propagation rapide et frappante de la langue arabe géographiquement dans plusieurs continents, y compris l'Afrique (la région du Maghreb et certains pays africains comme la Somalie et Djibouti), l'Asie (le Moyen-Orient et les États du Golfe), **[9]**

La langue arabe est considérée comme une langue riche et complexe à la fois, en raison du nombre de ses utilisateurs, qui dépasse les 400 millions de locuteurs répartis dans plus de 22 pays, et c'est ce qui en fait l'une des langues les plus toutes les parties du monde. La culture et la pensée du patrimoine du monde arabe, Avec l'avènement de la technologie et d'Internet, c'est la quatrième langue la plus utilisée sur Internet. La langue arabe se caractérise par la présence d'un certain nombre de types principaux, représentés en

1/ Arabe Standard **:** qui est la langue du Saint Coran (le Livre Saint de l'Islam et des Musulmans),

2/ Arabe Standard Moderne **:** qui est souvent utilisé dans les discours, les réunions officielles, l'éducation et les administrations en général .

3/ Dialecte arabe **:** Il est utilisé dans la communication quotidienne, les réunions informelles et les réseaux sociaux.

- **Arabe classique**

L'arabe classique (AC) est connu pour être fondamentalement la langue du Saint Coran, et ici sa grandeur est évidente. Le (AC) est le premier et le plus ancien des trois types de la langue arabe, qui est une approche linguistique dont les règles ont été définies au début du VIIIe siècle.), la langue du Coran (l'étape des conquêtes islamiques , et la phase post-coranique (de l'avènement de l'Islam au début du XVIIIe siècle).**[9]**

La langue Arabe classique se distingue du reste des langues, sa grande capacité à utiliser lettres d'orthographe, ce qui manque à d'autres langues étrangères, comme la lettre Dhad (la langue arabe classique est appelée la langue de dhad), **[10]** Il est également utilisé uniquement dans le patrimoine culturel passé et apparaît également dans les formalités, telles que le Saint Coran, les textes classiques, l'histoire, la poésie, les preuves anciens, etc. **[9]**

L'arabe classique n'est pas une langue commune de dialogue, mais elle est enseigné dans les établissements d'enseignement, et des cours de formation sont organisés pour la littérature arabe, et il existe encore à ce jour, et est utilisé dans les races religieuses et officielles **[9].**

- **Arabe standard moderne (ASM)**

C'est une forme moderne de la langue arabe classique. C'est la langue officielle commune du monde arabe **[11].** Elle est comprise par la majorité des ses utilisateurs et locuteurs, mais ne s'acquiert pas comme la langue maternelle, elle est considérée comme langue officielle. La multiplicité des littératures est à l'origine de la prédominance de ce modèle dans les médias, notamment les magazines et rapports, et il est également utilisé dans les médias audiovisuels tels que la radio et la télévision.

Il n'a pas de pratique directe et spontanée et n'est utilisé qu'aux situations officielles, et comme c'est la langue commune des pays arabes, c'est la seule langue de communication arabe environnementale. L'arabe standard moderne est dérivé directement de la langue.

- **Arabe dialectal**

la langue berbère et la langue française suite au colonialisme. Malgré l'importance du dialecte et ses utilisations fréquentes pour raconter des histoires, des drames, des poèmes, des chansons et divers programmes télévisés, il n'est pas enseigné dans les écoles, contrairement aux médias informels qui l'utilisent dans les feuilletons télévisés.

Le dialecte a C'est le terme dialecte arabe qui fait référence à la langue arabe régionale et est utilisé dans la communication informelle quotidienne, y compris les dialectes résultant de la fusion de l'arabe et des langues locales.

L'influence culturelle ou le processus d'arabisation qui se produit sur la langue est dû au colonialisme, aux mouvements migratoires et commerciaux, et les médias ont un rôle, surtout à notre époque. Parmi les exemples, nous constatons que le dialecte algérien est influencé par rabe est considéré comme la langue maternelle des Arabes, et ses formes et règles diffèrent d'un pays à l'autre selon la région, et il est également en retard sur les dialectes d'un pays.

4. Dialecte arabe algérien

La population de l'Algérie a atteint plus de 45 millions de personnes en 2022 [12]. Qui est situé en Afrique du Nord(pays du Maghreb), car sa langue principale est le dialecte algérien, qui est l'un des dialectes les plus difficiles et le plus proche de l'arabe classique par rapport aux autres dialectes arabes, car c'est le seul dialecte qui prononce toutes les lettres de l'alphabet de A à Z. Ils communiquent également avec elle dans leur vie quotidienne et via les sites de réseaux sociaux et Échangent des SMS entre eux. Et utilisé dans les chansons, les publicités, les films et les séries.

4.1. Présence sur les réseaux sociaux

Le nombre d'utilisateurs des sites de réseaux sociaux en Algérie est estimé à 25 millions de personnes, soit un taux de 56,5%, soit plus de la moitié de sa population, puisqu'il a connu une augmentation de 13,6% de 2020 à 2021. Face book est le les plus utilisées par 97,9% En ce qui concerne les langues utilisées ou les plus fréquemment utilisées sont (arabe, français)**[12].**

4.2. Caractéristiques

Le dialecte algérien est le dialecte le plus proche de la langue arabe car il existe des mots inspirés de la langue arabe, en plus d'être un mélange linguistique car il contient de nombreuses langues comme l'italien, le turc et le français. **[14].**

Les Algériens communiquent avec eux dans leur vie quotidienne et sur les réseaux sociaux, et cela ne se fait pas dans l'enseignement et les communications officielles **[15].** Dans cette partie, nous allons montrer les caractéristiques lexicales et syntaxique.

- ✓ **Caractéristiques lexicales**

Dans le dictionnaire du DAA on trouve de nombreux mots empruntés à plusieurs langues (arabe, berbère, français, turc, espagnol) en raison des événements historiques qui ont traversé le pays et y ont laissé leur empreinte, c'est ce qui le distingue de la langue arabe. Le tableau suivant montre ce mélange, en présentant quelques mots utilisés de différentes origines.

Tableau 1.1 : Quelques mots du dialecte algérien

Dialecte Algérien	Prononciation	Français	Origine
طرشي	Turshi	Douces villas	Türk
الزوالي	Zawali	Pauvre	
بكوش	Bakosh	Muet	
بالطو	Balto	Manteau	
سمانا	Sman	Semaine	Espagnol
صباط	Sabat	Chaussure	
قيرة	Gira	Problème	
صافا	Cava	Çava	Français
نورمال	Normal	Normal	
بيا	Bya	Bien	
طابلة	Tabla	Table	
كوزينة	Kozina	Cuisine	
فريجيدار	Frejidare	Réfrigérateur	
هيدورة	Hayduora	Peau sacrificielle	Berbère
شوشة	Shosha	Cheveux frontaux	
مكشرد	Mkashrid	Cheveux peignés	
شلاغم	Shlaghim	Moustaches	

- ✓ **Caractéristiques syntaxique**

1) Pronoms

- **Pronoms personnels**

Le dialecte algérien a des pronoms personnels spéciaux où il y a des cas dans lesquels les pronoms masculins et féminins ont le même pronom comme indiqué dans le tableau 2 avec une explication des différences entre eux et les pronoms personnels de la langue arabe.

Tableau 1.2 : Pronoms personnels en dialecte algérien

		Arabe classique	dialecte algérien	Français
-Conférencier	**-Seul**	**- أنَا**	**انا**	**-Je**
	-Pluriel	**- نَحْنُ**	**حْنا**	**-Nous**
-Destinataire	**-Féminin singulier**	**- أنت**	**أنت**	**-Tu**
	-Masculin singulier	**- أنتَ**	**أنتَ**	
	-Féminin, Masculin	**- أنْتُما**	**انْتوما**	**-Vous**
	-Féminin Pluriel	**- أنْتُنَّ**		
	-Masculin Pluriel	**أنْتُم**		
-Absent	**-Féminin singulier**	**-هيَ**	**هيَ**	**-Elle**
	-Masculin singulier	**-هوَ**	**هوَ**	**-Il**
	-Féminin, Masculin	**-هُما**	**هوما**	**-Ils**
	-Féminin Pluriel	**-هنّ**		**-Elles**
	-Masculin Pluriel	**-همْ**		**-Ils**

- **Pronoms démonstratifs**

Tableau **3** nous comparaison des pronoms démonstratifs entre le dialecte algérien et la langue arabe.

Tableau 1.3 : Pronoms démonstratifs en dialecte algérien

	Arabe classique	dialecte algérien	Français
-Féminin	**هذه**	**هادي / هاذي**	**Celle-ci**
	تلك	**هاذيك هاديك**	**Celle-là**
-Masculin	**هذا**	**هادا**	**Celui-ci**
	ذلك	**هاداك**	**Celui-là**
-Masculin singulier	**أولائك**	**هادوك**	**Ceux-là**
	هؤلاء	**هادو**	**Ceux-ci**

2) Forme définitive de la phrase

L'ordre des mots dans la phrase commence soit par le verbe, soit par le sujet, soit par l'objet Ce qui indique la flexibilité du dialecte algérien dans le tableau **4 [16]** , nous expliquons avec un exemple des types d'arrangement.

Tableau 1.4 : Disposition des mots dans la phrase

Arrangement	La phrase est en dialecte algérien	Français
SVO	الولد راح للمسيد	Le garçon est allé à l'école
VSO	راح الولد للمسيد	
OVS	للمسيد الولد راح	
OSV	الولد للمسيد راح	

3) Négation

Dans le dialecte algérien, « ماشي , mashi» et « ما , ma» sont utilisés pour la négation, puisque « ماشي » correspond à « ليس » en arabe classique et « ما» correspond à « ما » ou « لم ».

- **Négation en utilisant (ما)**

- Nous ajoutons "ش" ou "يش" à la fin du mot destiné à la négation.

- **Négation en utilisant(ماشي)**

-(ماشي) est utilisée avant le mot ou la phrase destinée à la négation.

Le tableau **5** montre quelques exemples de phrases déclaratives **[16].**

Tableau 1.5 : Phrases déclaration avec leur Négation

Dialect Algerien	Arabe classique	Français
- لعبت -ما لعبت	- لعبت - ما لعبتش / لم تلعب	-Elle a joué -Elle n'a pas joué
- راهي مريضة - ما راهيش مريضة	- انها مريضة - ليست مريضة	-Elle malade -Elle n'est pas malade
- هوما كتبو - ماشي هوما كتبو	- هم كتبوا - ليسوا هوم من كتبوا	-Lls ont écrit -Ce ne sont pas ceux qui ont écrit
- هوما كتبو - ماشي هوما الي كتبو	- هم كتبوا - ليسوا هوم اللذين كتبوا	-Lls ont écrit -Ce ne sont pas ceux qui ont écrit
-الولد مريض - ماشي مريض الولد	- الولد مريض -ليس الولد مريض	-Le garçon est malade -Le garçon n'est pas malade
- الولد مريض - الولد ماشي مريض	- الولد مريض -ليس الولد مريض	-Le garçon est malade -Le garçon n'est pas malade

4) Interrogation

En Algérie, les phrases interrogatives sont exprimées de deux manières

1-Toute phrase peut être une phrase interrogative, et cela se comprend par la prononciation de la phrase.

Exemple :

؟ رايح تخدم , vas tu au travail ?

2-En utilisant des pronoms interrogatifs

وين رايح تخدم ؟ , Où vas-tu travailler ?

Dans le tableau **6**, nous mentionnons les mots utilisés à interrogation et leurs équivalents dans arabe classique.

Tableau 1.6 : Pronoms interrogation

dialecte algérien	Arabe classique	Français
كفاش	كيف	Comment
وين	اين	Ou
علاه / وعلاش	لماذا	Pourquoi
شكون	من	Qui
شحال	كم	Combien
فواش / فاش	في ماذا	En quoi
منين	من اين	D'où
وقتاش	متى	Lorsque
باش	بماذا	Avec quoi

4.3. Prétraitement des textes

4.3.1. Les travaux connexes

Tableau 1.7 : Prétraitement

Les travaux	**Caractéristiques**						**Dialecte**
	Prétraitement						
	Suppression URL	**Suppression Sympol**	**Suppression lettres longues**	**Suppression lettres répets**	**Suppression les mots vides**	**Stemming**	

[17]	✓	✓		✓		✓	**Marco**
[18]	✓	✓			✓	✓	**Tunisie**
[19]	✓	✓		✓	✓	✓	
[14]	✓	✓		✓	✓	✓	**Algérien**
[15]		✓	✓	✓		✓	**Algérien**
[20]		✓		✓			

Comme nous le constatons dans le tableau **7**, il n'y a pas de travail qui traite de toutes les étapes de prétraitement du texte, puisque dans notre travail nous aborderons toutes ces étapes avec la modification de l'étape suppression les mots vides et stemming**.**

Dans ce chapitre, nous avons présenté les types de la langue arabe, où nous avons abordé le dialecte et défini le dialecte algérien, nous avons donc étudié ses caractéristiques et comment il est traité.

Chapitre 2

Méthodologie

Chapitre 2 : Méthodologie

Dans ce chapitre, nous aborderons une description générale d'outil proposés en connaissant les entrées, comment elles sont traitées et les sorties que nous obtenons, puis nous expliquerons l'environnement et les outils utilisés.

1. L'architecture d'application

La structure de l'application se compose de trois parties : entrée, prétraitement et sortie.

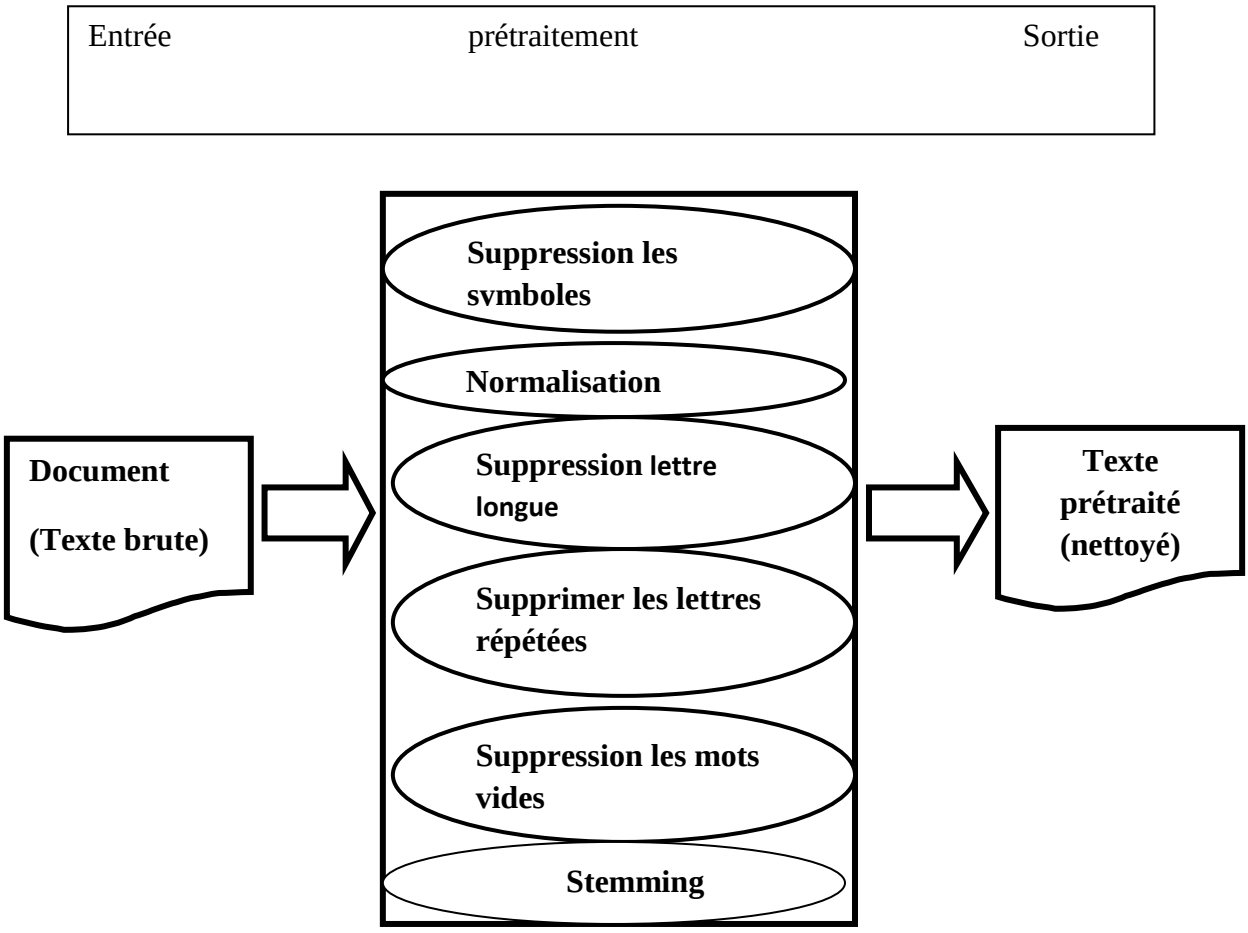

Figure 2.2 : Prétraitement

1.1. Les entrés

❖ **Les fiches de l'extension**

Il s'agit d'un document de traitement de texte créé avec Microsoft Word ou d'autres programmes de traitement de texte tels que Open office writer ou Apple Page.

1.2. Prétraitement

- **Phase initiale :** Nous supprimons le contenu des URL, hashtags , espaces blancs et Les signes de ponctuation .
  - **Exemple d'URL :** https ://www.google.fr/
  - **Les signes de punctuation : ! , ; . ? ***

- **Normalization**

Normalization est suppression des accents de lettre en convertissant tous les accents en une seule lettre comme mentionné dans l'exemple

Exemple : إ, أ, آ, ا est remplacée par ا

- **Suppression des lettres longues**
  Exemple : باهيـــــة est remplacée par **باهية** (bahya)

- **Suppression des lettres répétées**
  Exemple : مليييييحة (mliiiiiha) est remplacée par **مليحة** (mliha)

Avant ce processus, les machines vont considérer que les deux mots sont différents, donc ils sont représentés par des vecteurs différents au lieu de les représenté avec le même vecteur, et que ce sont à l'origine le même mot.

- **Suppression des mots vides**

Les mots vides sont des mots couramment utilisés dans le dialecte algérien, car leur suppression n'affecte pas le texte ou le sens du texte, comme la suppression de prépositions, afin de réduire considérablement le temps de traitement et d'apprentissage **[21].**

Nous avons utilisé une liste de mots vides proposée par Damazouz **[22]**

"... أكيما, شغل , هيهات, ضركا, كاش,تسما".

Nous l'avons modifié en supprimant ces mots **"ماكانش , ماكاش ماكان, مكاينش , مش ,مكانش مكاين"** car ils reflètent le sens du texte "Négation", donc les supprimer change le sens du texte.

Exemple :

avant la suppression : **"ليوم مكانش قراية"**

après : **"ليوم قراية"**

- Algorithme liste mots vides

  Nous appliquons l'étape de suppression de certains mots de la liste de mots vides proposée par Damazouz pour obtenir une nouvelle liste plus précise.

```
Fonction      Liste_mot_vides() :String
Var     liste_mot_vides :String
Début
    Motvid ← list mot vide
    Nomotvid← "مكانش ماكاش مكاينش مكاش"
    NomotvidD← Diviser_en_mots(Nomotvid)
    motvidN← normalization(Motvid)
    motvidD← Diviser_en_mots(motvidN)
    For i ← i dans  motvidD
       If ( i != NomotvidD) alors
          liste_mot_vides ← liste_mot_vides +  "  " + i
          liste_mot_vidD ← Diviser_en_mots( liste_mot_vides )
return  liste_mot_vidD
```

Figure 2.3 : Algorithme List mots vides

- Algorithme mots vides

  On supprime une liste de mots vides précédemment obtenus d'un texte donné

```
Fonction      Suppression_mot_vides(text) :String
Var     new :String
Début
    textN← normalization(Motvid)
    textD← Diviser_en_mots(motvidN)
    For i ← i dans  textD
       If ( i != Liste_mot_vid() ) alors
          new ← new +  "  " + i
return  new
```

Figure 2.4 : Algorithme mots vides

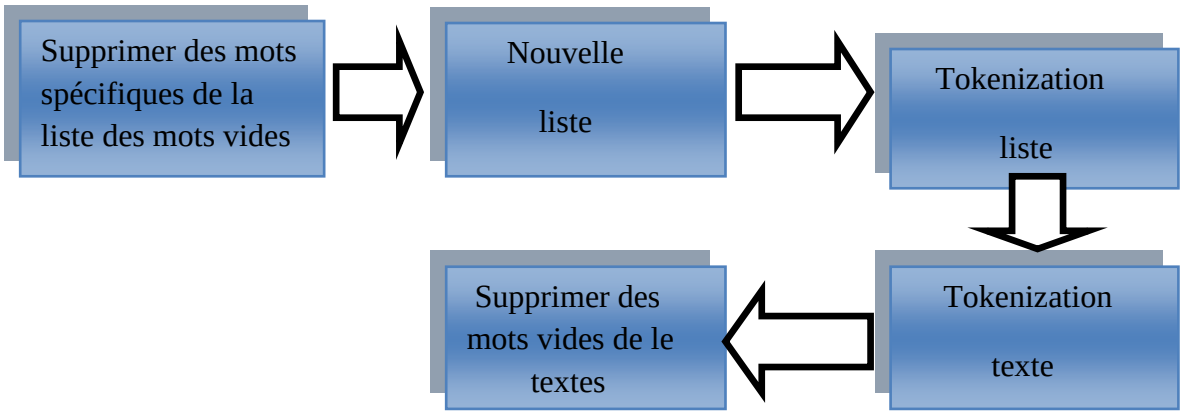

Figure 2.5 : Schéma d'algorithme de mots vides

- **Stemming**

La dérivation est le processus de réduction d'un mot à sa racine en supprimant ou en coupant les préfixes et suffixes présents dans le mot (début, fin) **[23].**

❖ **Préfixes**

au début du mot liste **" ال يـ فل بل"**

Entré : يخدم, فللفرجدار, المعلم

Sorté : خدم, لفرجدار, معلم

❖ **Suffixes**

à la fin du mot liste**" و وا ات ين"**

Entré : رايحين , نلعبو

Sorté : رايح , نلعب

❖ **Passer du pluriel au singulier :**

Un mot commençant par "نـ" ou "تـ" ou "ا" et se terminant par "و" ou "وا"

Entré : احفضو , تقروا

Sorté : حفض , قر

❖ **Supprimer "ة"**

Supprimer "ة" s'il y a un "يـ" ou une lettre avant, et avant la lettre "يـ"

Entré : مليحة , باهية

Sorté : مليح , باهي

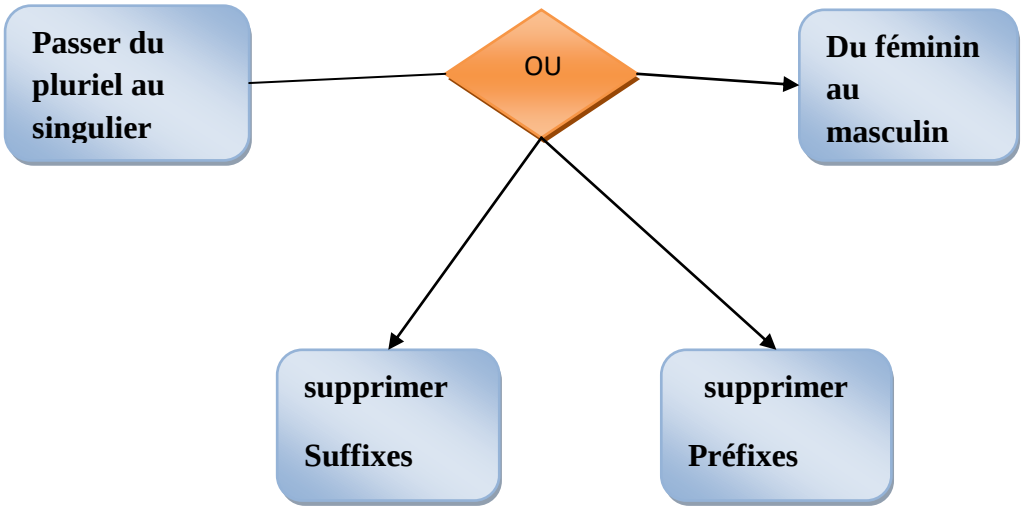

Figure 2.6 : Schéma d'algorithme stemming

1.3. Sortie

Les résultats sont un texte simple et propre qui peut être converti après avoir appliqué les étapes de prétraitement du texte mentionnées précédemment.

2. Architecture matérielle

Nous avons développé notre application sur un PC ayant les caractéristiques suivantes **:**

1) Processeur : Intel core i3-6006U 2. 0 GHZ.

2) RAM : 4 GO .

3) Système d'exploitation : Windows 10.

3. Outils utilisés

3.1. Langage utilisé

Python

Python est un langage de programmation facile à apprendre qui est utilisé dans plusieurs domaines, tels que la gestion de données volumineuses, des calculs complexes et des programmes indépendants à l'aide d'interfaces graphiques. Il possède également des bibliothèques qui aident l'utilisateur à travailler sur des projets.

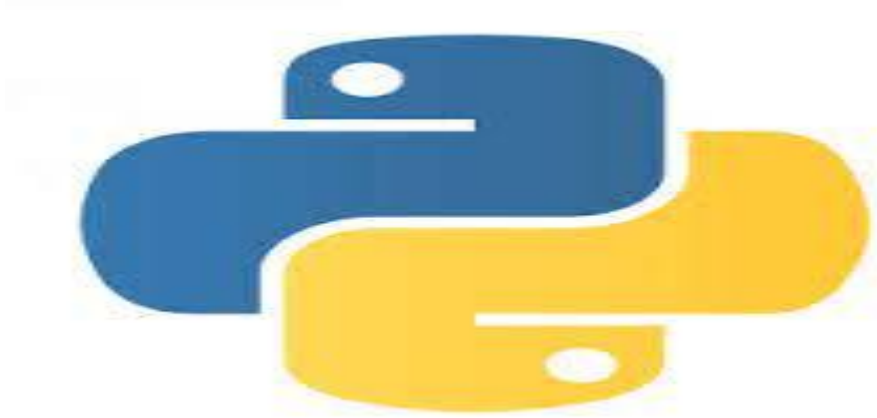

Figure 2.7 : PYTHON

3.2. Bibliothèques utilisées

NLTK

Nltk (Naturl langage Toolkit) est une bibliothèque de logiciels pour le traitement du langage naturel.

Figure 2.8 : NLTK

Pandas

Pandas est une bibliothèque open source sous licence BSD fournissant des structures de données et des outils d'analyse de données hautes performances et faciles à utiliser pour le langage de programmation Python.

Figure 2.9 : Pandas

PyQt6

Pyqt6 est une bibliothèque Python pour créer une interface graphique à l'aide des outils Qt.

Figure 2.10 : PyQt6

- **Gensim**

Gnesim est une bibliothèque de modélisation de données et de recherche de similarités utilisée dans le traitement du langage naturel (TAL) et la recherche d'informations.

Figure 2.11 : Gensim

4. L'environnement de développement

PyCharm

Est un environnement de développement intégré pour le langage Python utilisé dans la programmation informatique, produit par la société tchèque JetBrains. Il fournit l'analyse de code, un débogueur graphique, un testeur d'unité intégré, l'intégration avec des systèmes de contrôle de version (VCS) et prend en charge le développement Web avec Django ainsi que la science des données avec Anaconda**[10].**

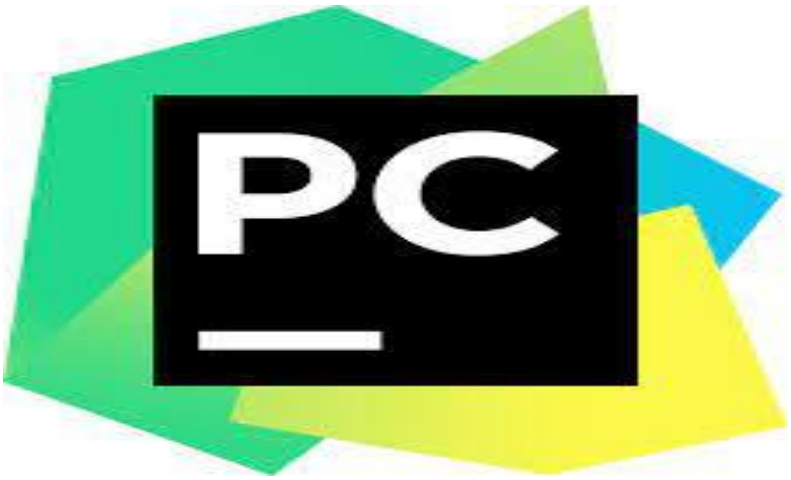

Figure 2.12 : PyCharm

Dans ce chapitre, nous avons expliqué les étapes de conception du système pour exécuter l'application avec l'environnement et les outils, et dans le chapitre suivant, nous parlerons de l'étape de mise en œuvre et des résultats obtenus.

Chapitre 3 :

Implémentation et Résultats

Chapitre 3 : Implémentation et Résultats

1. Application

1.1. Étapes pour exécuter l'application

1. L'utilisateur choisit les données sur lesquelles le traitement sera appliqué depuis l'ordinateur ou les écrit directement dans l'interface.

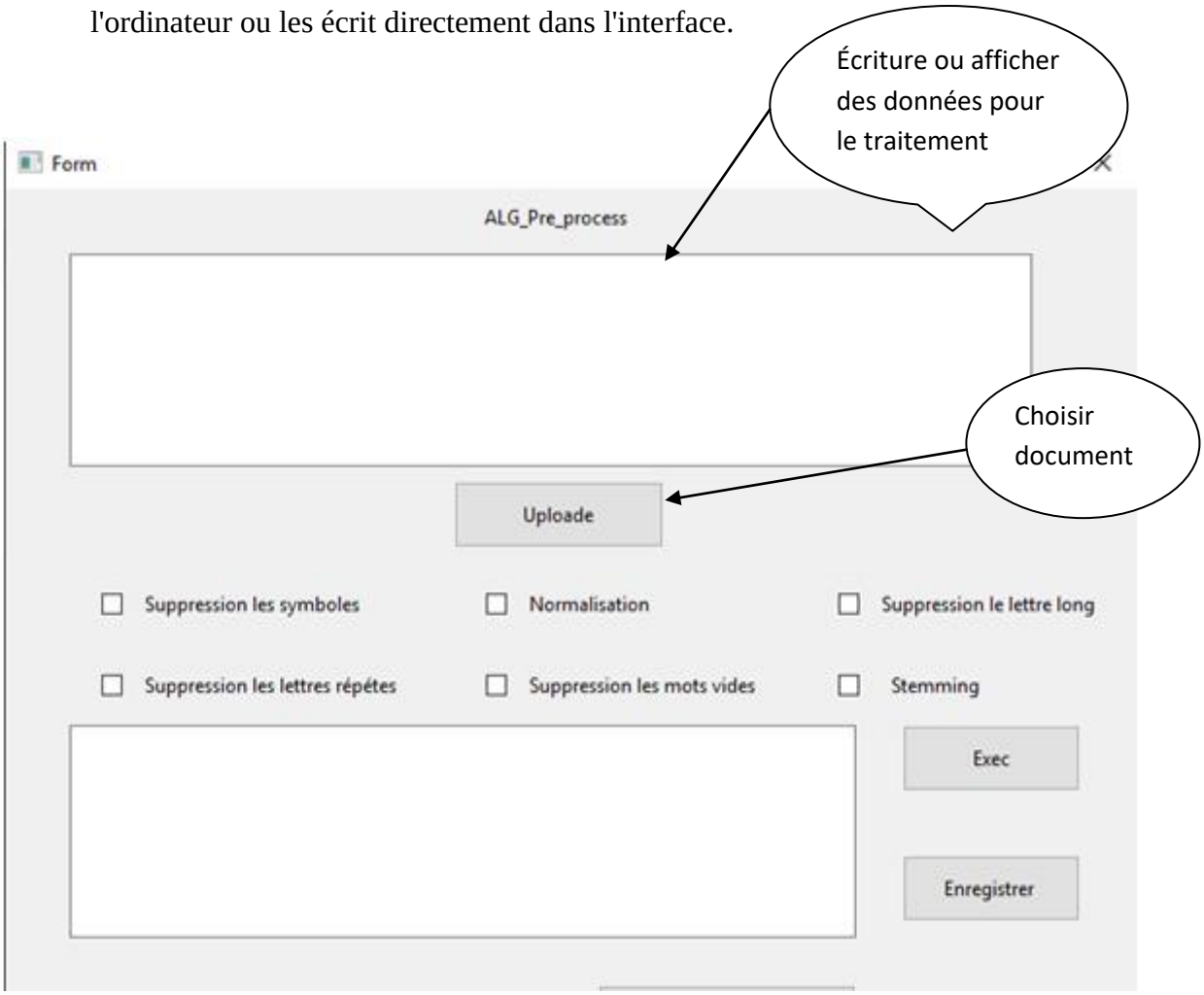

Figure 3.13 : Identification des données

2 Les fonctions choisies

- L'utilisateur choisit les fonctions pour traiter ses données afin que nous affichions les résultats de chaque fonction séparément.

  ✓ Suppression les symboles

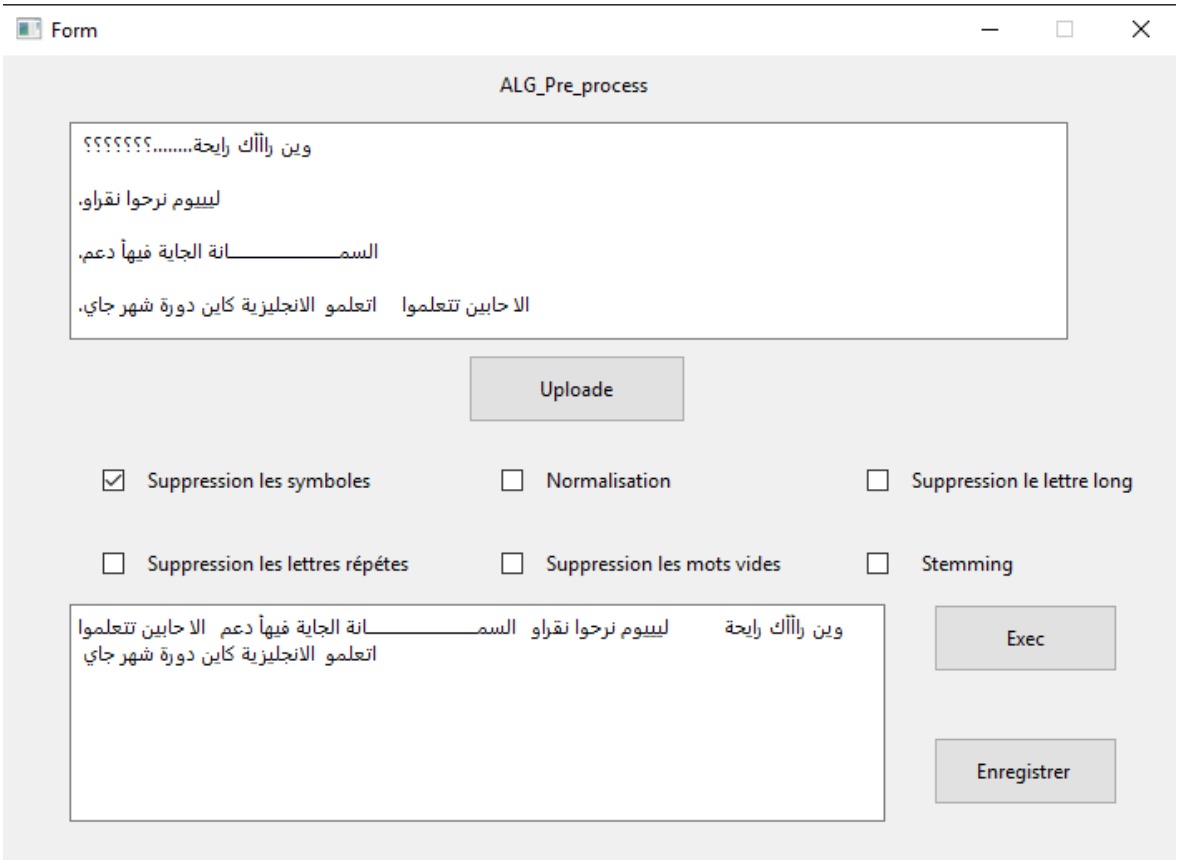

Figure 3.14 : Suppression les symboles

✓ Normalisation

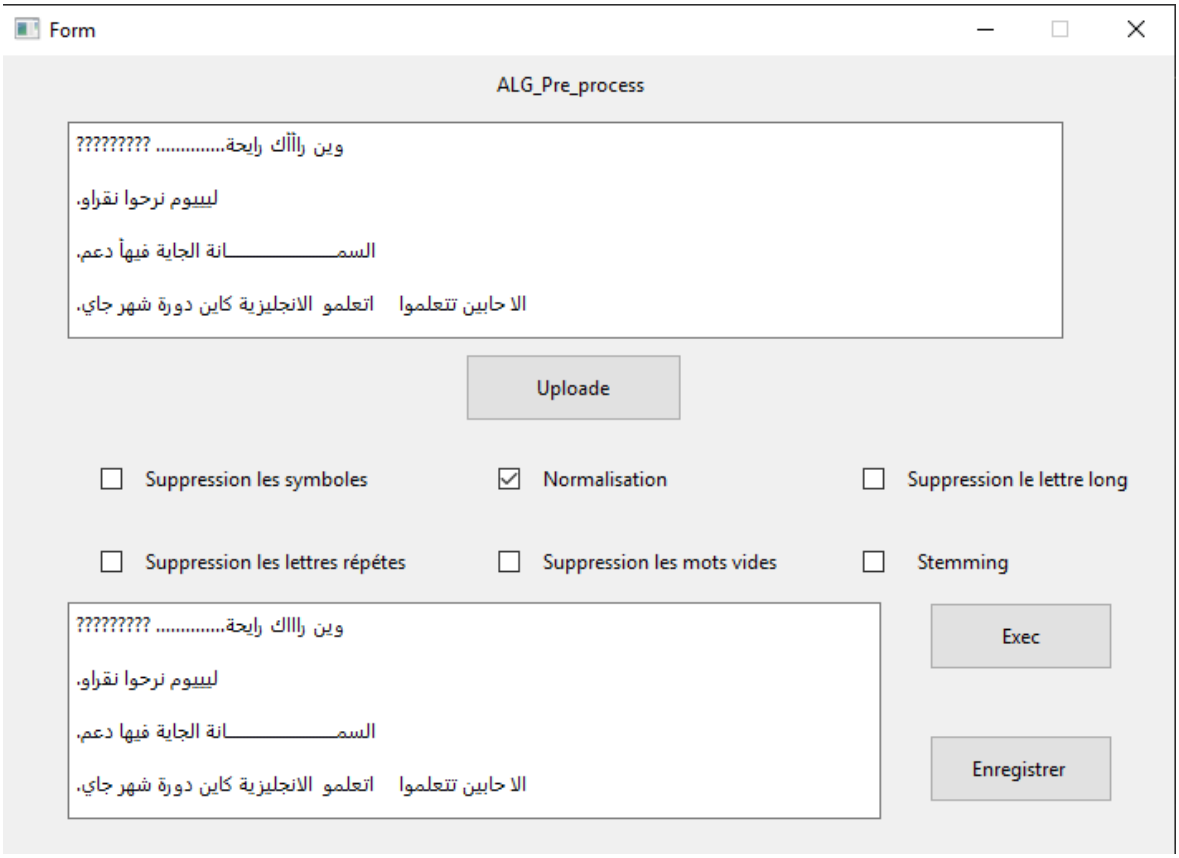

Figure 3.15 : Normalisation

✓ Suppression le lettre long

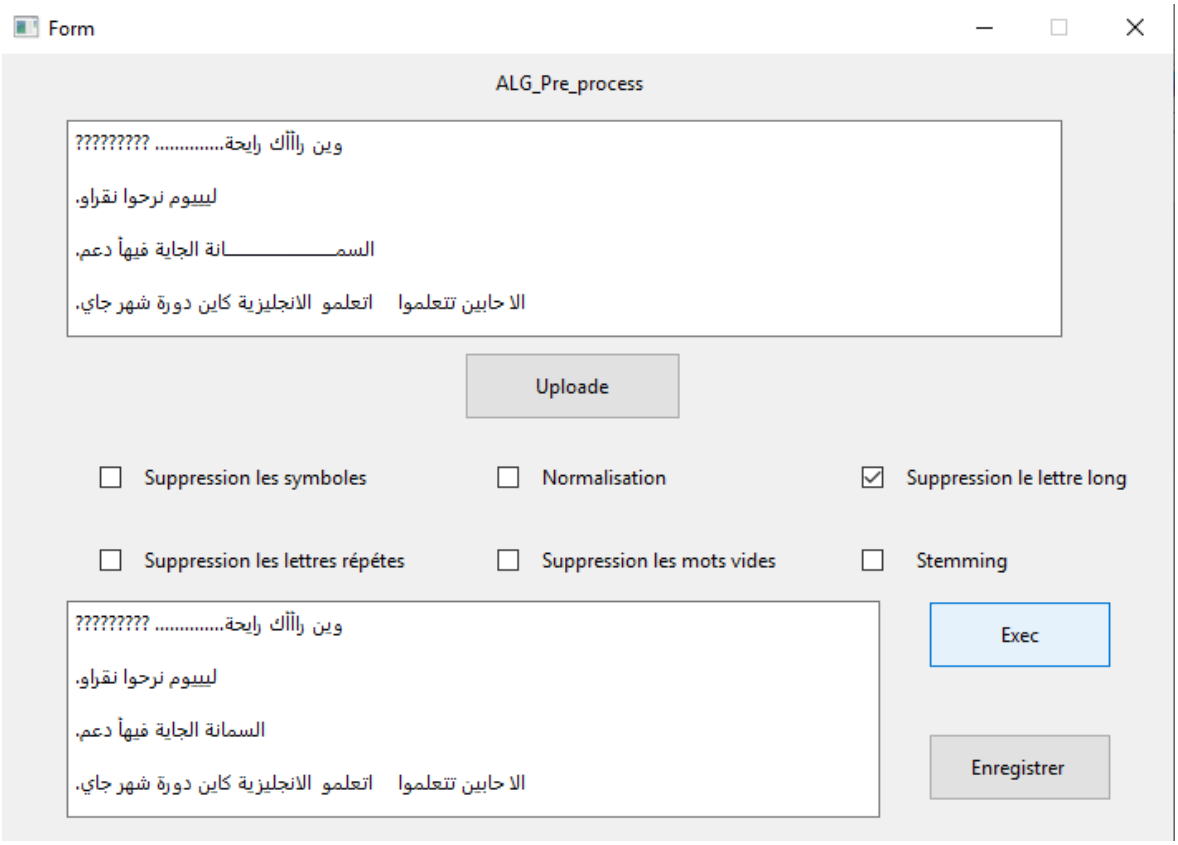

Figure 3.16 : Suppression le lettre long

- ✓ Suppression les lettres répètes

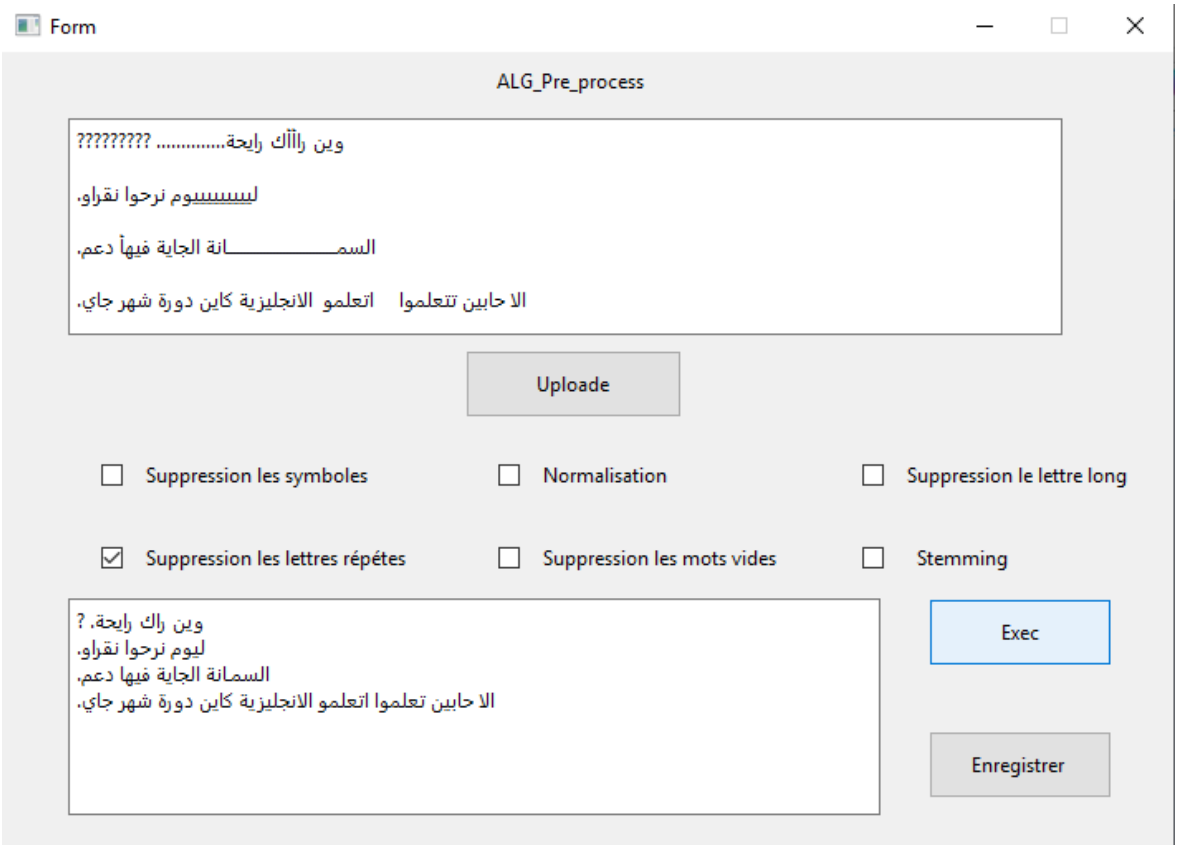

Figure 3.17 : Suppression les lettres répétes

- ✓ Suppression les mots vide

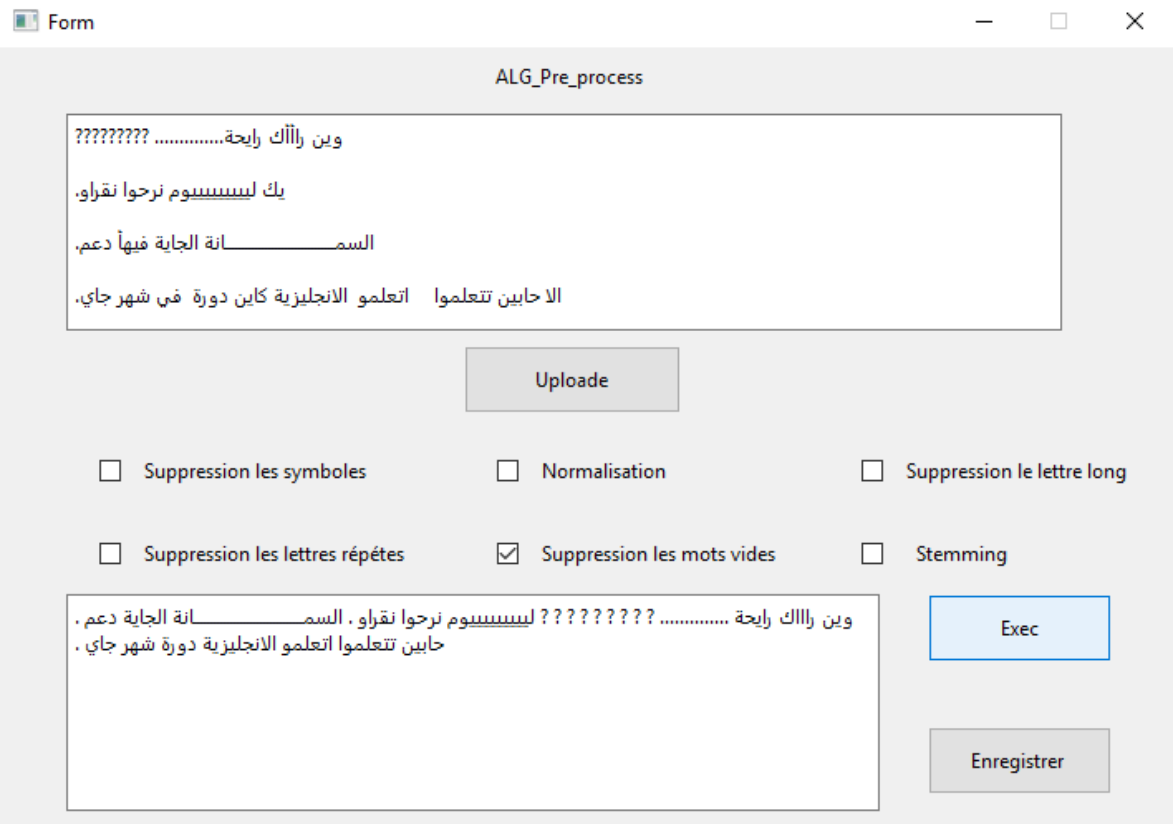

Figure 3.18 : Suppression les mots vides

- ✓ Stemming

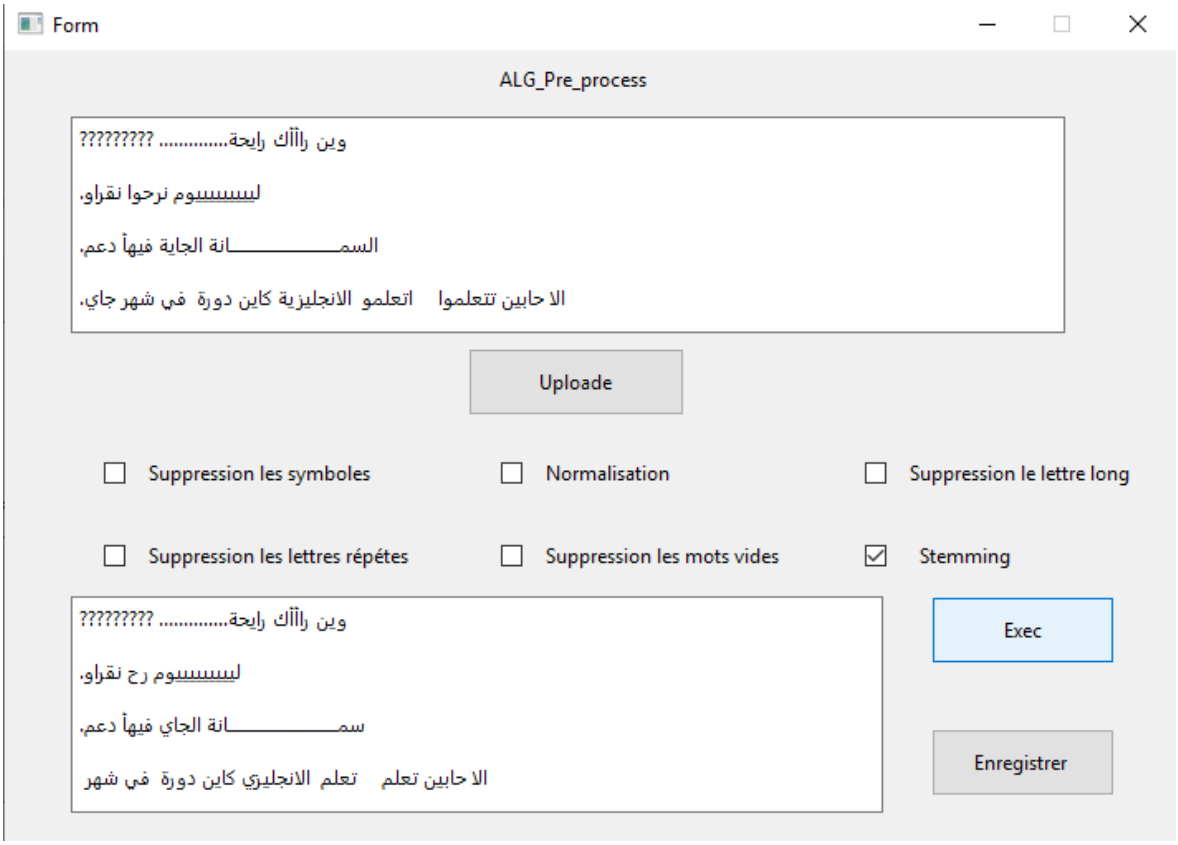

Figure 3.19 : Stemming

- Les résultats de données sont simples et propres lorsque toutes les étapes de traitement mentionnées ci-dessus sont appliquées.

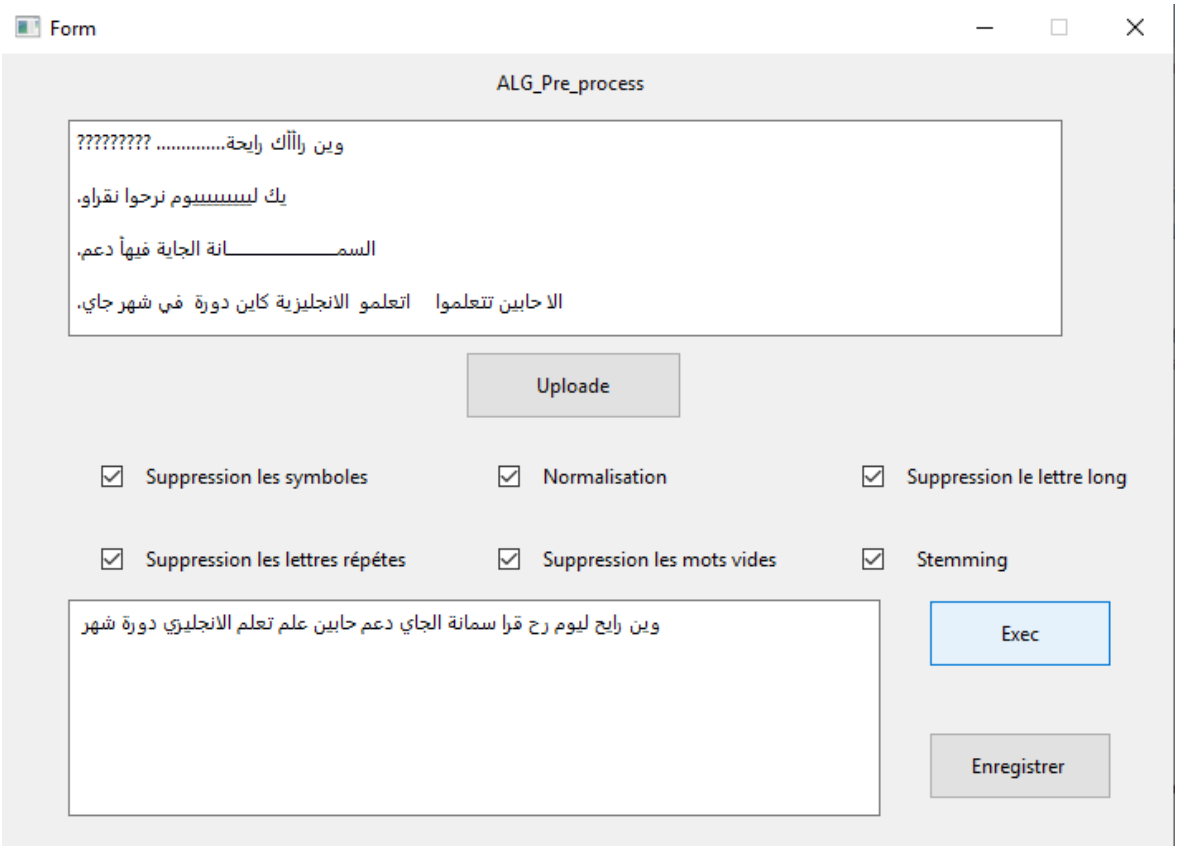

Figure 3.20 : Résulta

3. Enregistrer les données de traitement

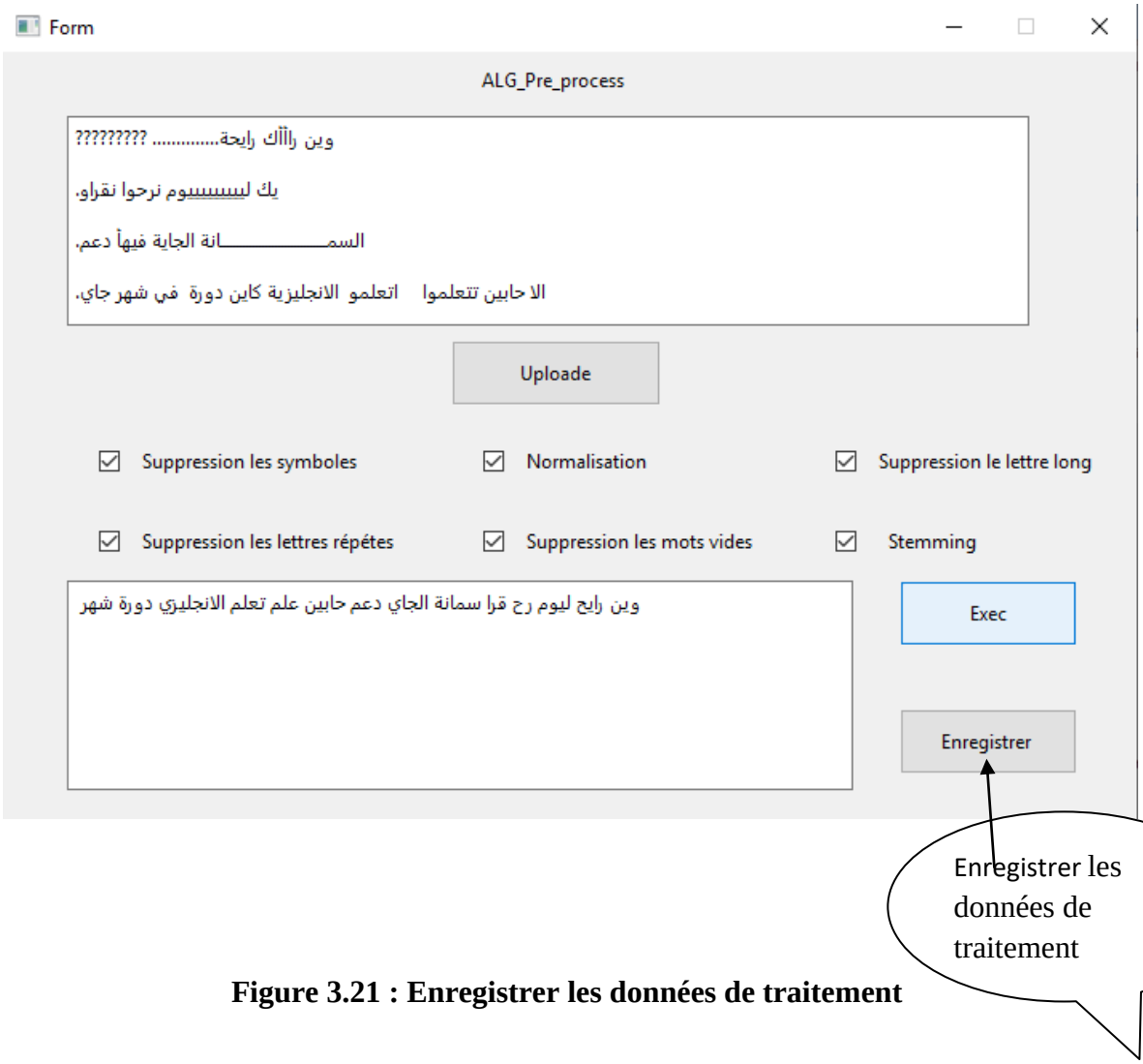

Figure 3.21 : Enregistrer les données de traitement

Conclusion générale

Conclusion générale

En conclusion, le prétraitement automatique du texte est l'une des branches du TAL dont le but est de représenter un texte d'une façon simple et significatif pour l'apprentissage automatique et l'extraction d'informations.

Malgré l'existence de quelques travaux dans ce domaine, ils comportent tous des limites que notre application a pu lever **:**

- J'ai supprimé les mots de la liste mots vides proposée par damazouz car ils ne sont pas mots vide afin d'augmenter la précision et l'exactitude du traitement.

- Appliquer des règles pour restaurer les mots à leurs origines d'origine en supprimant les préfixes et les suffixes.

Nous avons atteint le but de ce travail en fonction des résultats obtenus, et il peut être amélioré dans d'autres travaux.

Références

[1] E.D. Liddy, Natural Language Processing, 2001

[2] N. Kaur1, V. Pushe and R Kaur,"Natural Language Processing Interface for Synonym", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.7, July- 2014, pp. 638-642 ,ISSN 2320–088X

[3] S. Vijayarani1, J. Ilamathi and Nithya, "Preprocessing Techniques for Text Mining - An Overview", International Journal of Computer Science & Communication Networks, Vol.5, issue.1, pp. 7-16 7 ISSN: 2249-5789

[4] S. Jusoh and H.M. Alfawareh, "Natural language interface for online sales", in Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007),Malaysia: IEEE, November 2007, pp. 224-228

[5] E.K. Ringger, R.C. Moore, E. Charniak, L. Vanderwende, and H Suzuki, "Using the Penn Treebank to Evaluate Non-Treebank Parsers", In Proceedings of the 2004 Language Resources and Evaluation Conference (LREC), 2004, Lisbon, Portugal.

[6] **P. Jackson and I. Moulinier**,"Natural Language Processing for Online Applications": Cambridge University press, New York.2012, page 7-9.

[7] www.tutorialspiont.com mai 2022

[8] https://aitech.om/text-analysis/?lang=ar avril 2022

[9] Zribi, I. (2016). *Traitement automatique du dialecte tunisien: construction de ressources linguistiques* (Doctoral dissertation, Université de Sfax (Tunisie)).

[10] https://mawdoo3.com/مفهوم-_اللغة_العربية_الفصحى avril 2022

[11] Abidi, K., & Smaïli, K. (2017, December). An empirical study of the Algerian dialect of Social network. In *ICNLSSP 2017-International Conference on Natural Language, Signal and Speech Processing*.

[12] https://hijra.news/عدد-سكان-الجزائر/ **mai 2022**

[13] https://www.echoroukonline.com/احصائيات-مستخدمي-الانترنت-وش-اخر avril 2022

[14] Moudjari, L., Akli-Astouati, K., & Benamara, F. (2020, May). An Algerian corpus and an annotation platform for opinion and emotion analysis. In 12th Language Resources

and Evaluation Conference, LREC 2020 (pp. 1202-1210). European Language Resources Association.

[15] Guellil, I., Azouaou, F., Saâdane, H., & Semmar, N. (2017). Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien. Revue TAL.

[16] Harrat, S., Meftouh, K., Abbas, M., Hidouci, W. K., & Smaili, K. (2016). An algerian dialect: Study and resources. International journal of advanced computer science and applications (IJACSA), 7(3), 384-396.

[17] Abidi, K. (2019). La construction automatique de ressources multilingues à partir des réseaux sociaux: application aux données dialectales du Maghreb (Doctoral dissertation, Université de Lorraine).

[18] Ali, C. B., Mulki, H., & Haddad, H. (2018). Impact du Prétraitement Linguistique sur l'Analyse de Sentiment du Dialecte Tunisien (). In Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN (pp. 383-392).]

[19] MOSTEFA-CHEBRA, M. B., ESI, M. S. B., DEROUAZ, M. M. L., & BELLIK, M. S. Mémoire de projet de fin d'étude.

[20] Guellil, I., Adeel, A., Azouaou, F., & Hussain, A. (2018, July). Sentialg: Automated corpus annotation for algerian sentiment analysis. In International conference on brain inspired cognitive systems (pp. 557-567). Springer, Cham.

[21] Tamene, A. Classification Automatique des documents textuels.

[22] https://github.com/Damazouz/Algerian-Arabic-stop-words avril 2022

[23] Soukaina, M. I. H. I., Ismail, E. L., AREZKI, S., & LAACHFOUBI, N. (2020). MSTD: Moroccan Sentiment Twitter Dataset. International Journal of Advanced Computer Science and Applications, 11(10).