الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
جامعة عمار ثليجي بالاغواط
Université Ammar Telidji Laghouat
كلية العلوم
Faculté des Sciences
قسم الإعلام الآلي
Département d'Informatique

Mémoire de fin d'étude
pour l'obtention de diplôme de Master en informatique.

Domaine : Mathématiques et Informatique.
Filière : Informatiques.
Option : Système d'Information et Décision.

Réalisé par:

Chennouf Mohammed

Thème

Réalisation d'une boite à outils dédiée aux systèmes de détection de plagiat pour la langue Arabe

---

Soutenu le 29/09/2020, devant le jury composé de :

Mme. Hadda Cherroun	Phd.	Université Laghouat	`Rapporteur`
Mme. Badra Kerrouche	MAA	Université Laghouat	`Présidente`
M. Younes Guellouma	MCA	Université Laghouat	`Examinateur`

Nº d'ordre ..... Année universitaire 2019/2020

Remerciements

Tout d'abord, nous remercions notre Allah de nous avoir donné la force et le courage pour acheves se travail.

Le travail réalisé tout au long de cette mémoire n 'a été possible qu'avec l'aide et soutien de nombreuses persones. Je profite de l'occasion qui m'est donnée pour leur exprimer toute ma gratitude.

Nous remercions le corps administratif de notre département et tous nos professeurs qui ont été la raison de notre succès.

Nous remercions tout particulièrement notre enseignante Mme Cherroun Hadda pour ses conseils et sa patience avec nous pour que ce travail ait lieu.

Toute la famille et nos amis Hamza, Amine, Tarzi, Mourad, Mostafa, Soufiane, Ahmed pour être une source d'inspiration. tous les gens qui nous guident pour réaliser ce travail avec une façon ou d'une autre.

Chennouf Mohammed

Dédicaces

A ma trés cher mère, et trés cher Père..

Avec tout mon amour et tout mon effection. je vous remercie de m'avoir guidée. conseillée et cru en moi. et surtout d'avoir fait de moi ce que je suis devenue ajourd'hui. Merci Papa, et merci maman pour la pression que tu as su exercer sur moi pour obtenir mon master. j'espére que ce travail vous rendra encore plus fiers.

A la mémoire de mes grands parents..

Je n'ai pas eu la chance de profiter pleinement de vous, mais on se retrouvera un jour, je vous dédie ce travail ..

Chennouf Mohammed

Abstract

Scientific plagiarism is considered to be one of the most widely spread practices in literary and scientific communities, and the most infringement to the scientific integrity that is supposed to be provided by the scientific researcher.
In this is work, we have participited in the conception of toolbox for researchers and academic users . It helps to detect plagiarism.
The researchers in the field of scientific detection of plagiarism can experiment with known and additive algorithms and methods of their own creation and comparison with other algorithms.
Our toolbox conception and realisation reply on modularity , oriented object paradigm while using open source languages and platforms.

Keywords: Scientific plagiarism, The researchers, detection of plagiarism, toolbox.

ملخص

يعتبر الانتحال العلمي (السرقة الأدبية) من أكثر الظواهر انتشارا في الأوساط الأدبية و العلمية، و أكثرها إساءة إلى الأمانة العلمية التي من المفترض توفرها في الباحث العلمي.
في هذه المذكرة قمنا بإنشاء أداة للباحثين والأساتذة تساعد على كشف الإنتحال العلمي ويستطيع الباحثون في مجال كشف الانتحال العلمي تجربة خوارزميات معروفة وإضافة خوارزميات وطرق من إنشائهم وخاصة بهم وكذا مقارنتها بخوارزميات أخرى

الكلمات المفتاحية : الإنتحال العلمي، السرقة الأدبية، أداة كشف الإنتحال، خوارزميات .

Résumé

Le plagiat scientifique est considéré comme l'un des phénomènes les plus fréquents dans les milieux littéraires et scientifiques, et le plus grave pour l'intégrité scientifique supposée fournir par le chercheur scientifique.
Dans ce Mèmoire, nous avons créé une boîte à outils pour les chercheurs et les professeurs qui permet de découvrir le plagiat.
Les chercheurs dans le domaine de la détection scientifique du plagiat peuvent expérimenter des algorithmes connus et additionnels Algorithmes et méthodes de leur propre création et comparaison avec d'autres algorithmes

Mot clés: plagiat scientifique, détection du plagiat, une boîte à outils, méthodes.

Table des matières

Table des figures

Liste des abréviations

NLTK	Natural Language Toolkit
TALN	Traitement automatique du langage naturel
UML	Unified Modeling Language

Introduction générale

Contexte de recherche et problématique

La disponibilité massive de documents en ligne et l'arrivée de sociétés de rédaction académique sur Internet augmentent l'incidence du plagiat, en particulier dans les universités. En outre, ces dernières années, l'expansion d'Internet facilite également l'accès aux documents du monde entier (rédigés en langues étrangères) et à des outils de traduction automatique de plus en plus puissants, ce qui accentue la progression d'un nouveau type de plagiat : le plagiat translingue. Cela a incité de nombreux chercheurs à tenter de trouver une solution à ce problème.

Dérivé du latin "plagiarius" qui signifie "kidnappeur, séducteur, voleur littéraire".[10] Du mot anglais "plagiary" est "celui qui prend injustement les mots ou les idées de quelqu'un".

La détection automatique du plagiat a fait l'objet d'une attention particulière dans le cadre de la mise au point de systèmes de détection du plagiat à petite et à grande échelle comme contre-mesure possible. Dans le cas d'un document textuel, la tâche d'un système de détection de plagiat est de découvrir si le document est copié en tout ou en partie à partir d'autres documents sur le Web ou de tout autre dépôt[11].

Finalement, de nombreux outils ont été créés dans ce domaine, cependant, il y a un manque d'outils qui supportent la langue arabe.

Objectifs de la recherche

Notre travail de recherche rentre dans le cadre d'un projet plus grand qui est le developement d'une plateforme dedié la detection du plagiat .
les utilisateurs de cette plateforme peuvent ètre essentiellement des chercheurs, des étudiants et des professionnelles du domain acadèmique .
Les buts principaux de ce projet golbale(voir figure) sont nombreux :

- Introduire et Présenter des informations générales sur le plagiat et ses formes.
- Mettre en évidence le domaine de la détection automatique du plagiat et des méthodes utilisées, et le manque d'applications supportant l'arabe.
- Essentielement concevoir et puis implémenter l'essentiel d'une boite à outils dédiée aux chercheurs s'intéressant à la détection de plagiat. Les fonctionalités de cette boite à outils doivent ètre modulaires pour permettre aux chercheurs d'introduire de nouvelles méthodes.
- Offrir aux chercheurs une collection de ressources textuelles pour pouvoir valider et tester leur méthodes.

— Implémenter et intégrer certaines méthodes de détection de plagiat de la littérature.

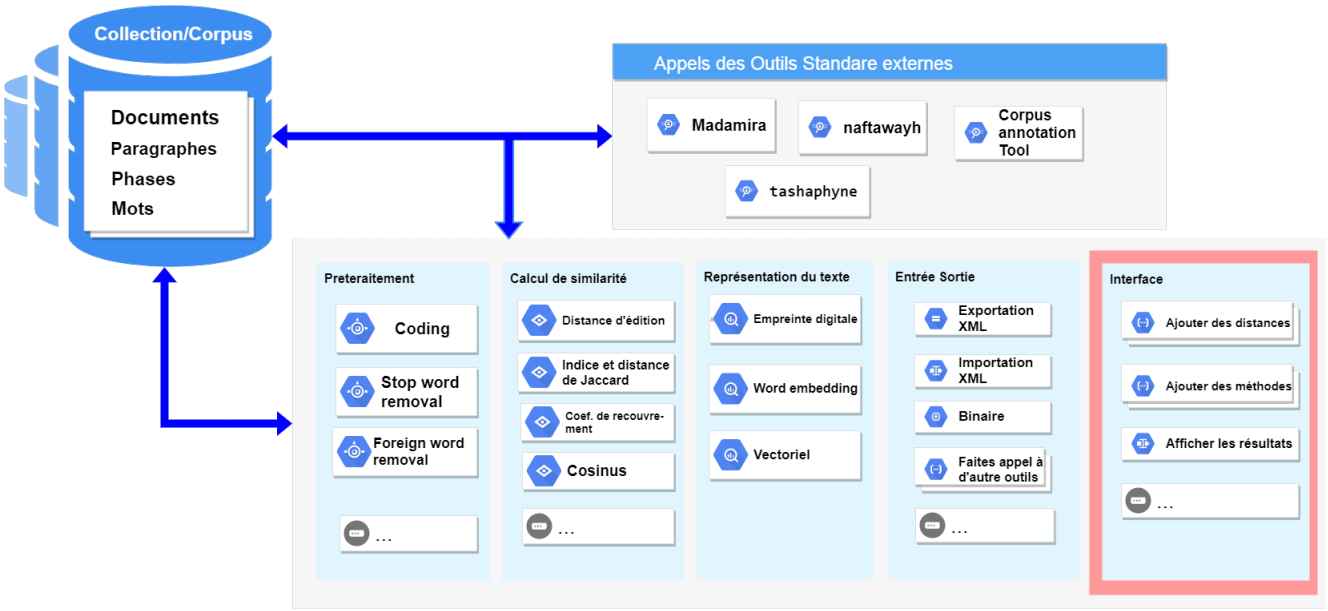

FIGURE 1 – Aperçu de projet globale

Notre contribution dans ce grand projet est la conception de la boite à outils ainsi que l'implémentation de certaines fonctionnalités telque le pretraitement des textes et calcul de similarité.

Organisation du mémoire

Nous avons divisé ce maniscript en trois chapitres selon le plan méthodologique suivant :

Le premier chapitre donne un aperçu général du plagiat, quelques définitions du plagiat, les types de plagiat et une description de certains termes associés.

Le deuxième chapitre ce chapitre n est compose a deux partie :

— la première partie l'étude et l'analyse de besoins de notre boite à outils, la deuxième partie pour la conception de cette boite à outils .

Le troisième chapitre Nous décrivons et présenter les environnements et logiçiels utilisés ainsi que nos propres codes.

Chapitre 1

Généralités sur le plagiat et les outils existants

1.1 Introduction

Le plagiat, et la facilités qu'il induise, tente une grande majoritéd'étudiants , chercheurs et même professeurs, et avec la croissance de ce phénomène dans la communauté scientifique, de nombreuses questions se posent. Dans ce chapitre, nous répondrons aux deux questions les plus importantes :

Q'est-ce que le plagiat ?
Quels sont ses types ?
Quels sont les outils existant ?

1.2 Définitions et terminologies

A fin de bien illuster le plagiat et ses variates , nous présenterons dans cette section ses définitions avec une description de certains des termes associés à ce domaine.

1.2.1 Plagiat

Le dictionnaire en ligne d'Oxford [1] définit le plagiat comme : "la pratique de prendre les idées ou le travail d'un autre et de les faire passer comme étant les siennes".

Le dictionnaire en ligne Larousse [2], définit également le plagiat comme :"Acte de quelqu'un qui, dans le domaine artistique ou littéraire, donne pour sien ce qu'il a pris à l'œuvre d'un autre."

de plus , selon Fishman[1], le plagiat est défini comme suit : " le plagiat est l'utilisation des idées, des concepts, des mots ou des structures et les intégrer à son propre travail sans en mentionner la provenance ".

À partir des définitions ci-dessus, nous concluons que l'utilisation de tout travail, qu'il s'agisse de notre travail précédent ou des travail d'autres personnes sans mentionner la

1. `https://www.lexico.com/definition/plagiarism`.(consulté le 16/04/2020)
2. `https://www.larousse.fr/dictionnaires/francais/plagiat/61301`.(consulté le 16/04/2020)

source, est considérée comme du plagiat. Il est donc important de définir certains termes liés à ce domaine.

1.2.2 Citation

Millet [12], définit une citation comme étant : " un fait de parole (ou d'écriture), par définition individuelle et unique, qui est repris comme tel -cité- par un autre locuteur ou une infinité de locuteurs ".

Par ailleurs, l'action cite est définie dans le dictionnaire Larousse comme : " Action de citer, de rapporter les mots ou les phrases de quelqu'un ; paroles, passage empruntés à un auteur ou à quelqu'un qui fait autorité. " [3]. Donc, une citation est la reproduction d'un extrait d'un texte ou un écrit antérieur dans la rédaction d'un nouveau texte.

1.2.3 Référence

Selon Baudry[2], une référence est une " action de se référer ou de renvoyer à un article, à un passage, à une chose ayant quelques rapports ". Référencer une source dans un document représente un indicateur d'existence de source d'une information, texte, parole ou autre, ayant un rapport direct ou indirect avec ce document. La référence doit contenir suffisamment d'information pour identifier de façon unique les sources telles que : un article, un livre, ou tout autre document. Ainsi, une référence doit contenir : un numéro (comme [12] ou bien [Chaouch, 2005]), un nom du/des auteurs, un titre, le nom de l'éditeur, et la date de parution , et le numéro international normalisé du livre (ISBN) (dans le cas d'un livre). Dans le cas d'un article : le nom du/des auteurs, le titre de l'article, le titre de la revue, le volume, le numéro de l'édition, la date et la page[3].

1.2.4 Bibliographie

La bibliographie est une liste structurée des sources citées (références), notamment des livres, des ouvrages, des articles de journaux, ou autres documents utilisés pour la préparation d'un document scientifique. La bibliographie se trouve généralement à la fin d'un article de journal, d'un livre ou d'un article d'encyclopédie[4].

1.3 Plagiat textuel

Le phénomène du plagiat s'est accru dans la communauté universitaire et la recherche dans ce domaine a considérablement progressé, ce qui a donné lieu à de nombreux articles et recherches sur les définitions et les types de plagiat et leur prévention. Ce qui est important ici, c'est le plagiat du texte, et c'est le plagiat impliquant la réutilisation d'un ouvrage écrit (travail écrit) sans mention de la source.

Sur la base de ces recherches, le plagiat se présente sous de nombreuses formes, telles que [3] :

— Présenter comme sien un travail original de quelqu'un d'autre.

3. `https://www.larousse.fr/dictionnaires/francais/citation/16228`. (consulté le 16/04/2020)

— Intégrer dans un travail des passages de textes, des données, des résultats expérimentaux ou même des images provenant de sources externes sans en mentionner la source.

— Résumer une idée créative de quelqu'un d'autre en l'exprimant avec d'autres mots tout en omettant d'en mentionner la provenance.

1.4 Formes de plagiat textuel

Dans cette section, nous présentons différentes formes de plagiat textuel et quelques exemples pour chacune d'entre elles.

1.4.1 Duplication

La duplication est la copie directe de phrases ou de passages d'un texte publié sans citation, également appelée copier/coller[6].

1.4.2 Substitution par des synonymes

Remplacer des mots par des synonymes est une forme de plagiat, Et de son nom, elle a utilisé des synonymes pour paraphraser des expressions.La figure 1.1 montre la différence après avoir reformulé une phrase en arabe en utilisant des synonymes.

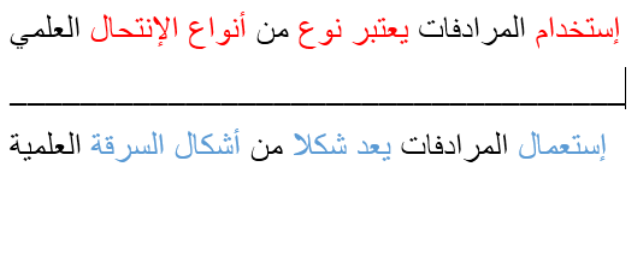

FIGURE 1.1 – Exemple de plagiat en remplaçant des mots par des synonymes

1.4.3 Reformulation paraphrastique

La paraphrase ou la reformulation paraphrastique, consiste à reprendre un texte original sans altération de son contenu en utilisant le changement de son vocabulaire par l'ajout, la suppression ou la substitution de mots par des synonymes[?]. La figure 1.2 illustrie un example de reformulation paraphrastique

1.4.4 Traduction

Le plagiat par traduction aussi appelé plagiat translingue consiste à faire une transformation manuelle ou automatisée d'un texte original d'une langue à une autre sans mentionner la source[8]. Par exemple traduire la définition de Traitement Automatique du Langage Naturel (TALN) [4] du français vers l'anglais comme l'illustre la figure 1.3.

4. `https://fr.wikipedia.org/wiki/Traitement_automatique_du_langage_naturel` (consulté le 27/04/2020)

تُعتبر طريقة إعادة ترتيب شكل الجُملة من أبسط الطرق في إعادة الصياغة
فهي لا تحتاج سوى التقديم والتأخير

في إعادة صياغة الجملة نحتاج فقط لطريقة بسيطة و ذلك بإعادة
ترتيبها أي أننا نقوم بالتقديم أو التأخير.

FIGURE 1.2 – Exemple de plagiat par reformulation paraphrastique

Définition originale en français

Le traitement automatique du langage naturel, abrégé en TALN, est une discipline s'appliquant au domaine de l'informatique et du langage. Il est utilisé par exemple pour les traductions, la reconnaissance vocale ou encore les réponses automatiques aux questions[a].

Définition traduite en anglais

The automatic processing of natural language, abbreviated to NLP, is a discipline that applies to the field of computer science and language. It is used, for example, for translations, voice recognition or automatic answers to questions.

a. fr.wikipedia.org. (consulté le 27/04/2020)

FIGURE 1.3 – Exemple de cas de plagiat translingue

1.4.5 Fantôme littéraire

Fantôme littéraire appelé aussi *ghostwriting* est une autre forme de fraude académique, souvent utilisée par les étudiants dans le domaine universitaire. L'étudiant paie une personne tierce ou une entreprise sur le net ou cyber café pour lui faire ses devoirs écrits, réaliser son mémoire de master ou même sa thèse de doctorat [3, 9].

1.4.6 Assemblage

L'assemblage est une copie de passages provenant de sources multiples et leur mélange dans la nouvelle œuvre sans citation. Également connu sous le nom de "patchwork" .[3].

1.4.7 Auto-plagiat

C'est le cas lorsque quelqu'un réutilise de grandes parties de son propre travail, sans mentionner le texte original.. La figure 1.4 illuster un autp-plagiat effectué par le docteur physicien Étienne Klein. [5].

1.5 Logiciel de détection de plagiat

Dans cette section, nous listons quelques outils de détection automatique du plagiat, en nous concentrant sur ceux qui supportent la langue arabe,

5. https://blogs.mediapart.fr (consulté le 27/04/2020)

L'origine, par Etienne Klein

Dans nos contrée, la métaphore du fleuve a accompagné presque toute l'histoire de la pensée du temps et continue d'irriguer notre façon de l'évoquer et de le représenter : ... au temps même les propriétés de la ligne par laquelle on le représente. Kant l'avait déjà vu :"Nous représentons la suite de temps par une ligne qui se ... second sont toujours successives.".[a]

a. "Le temps est-il affaire de conscience ?", Etienne klein, European Psychiatry, Novembre

L'emprunt, par Etienne Klein

la métaphore du fleuve a accompagné presque toute l'histoire de la pensée du temps - du moins en Occident - et continue d'irriguer notre façon de l'évoquer et de le représenter : ... au temps les propriétés de la ligne par laquelle on le représente. En d'autre termes, le fait de d'écrire le temps par une ligne lui confére i|sp facto des "probléme de ligne" Kant l'avait reléve :"Nous représentons la suite de temps par une ligne qui se ... second sont toujours successives.".[a]

a. "L'instant présent, unique mais banal",Etienne Klein,Pour La Science,octobre 2010.

FIGURE 1.4 – Un exemple d'un auto plagiat effectué par le docteur physicien Étienne Klein.

l'étude de ces outils repose sur leur fonctionement et leur popularité.

1.5.1 L'outil Urkund

Selon le site web d'Urkund : "Urkund est un système entièrement automatique d'apprentissage automatique de reconnaissance de texte conçu pour détecter, prévenir et gérer le plagiat, quelle que soit la langue dans laquelle vous écrivez".

Urkund analyse les documents envoyés à l'aide de l'apprentissage automatique, et détecte non seulement les similitudes avec d'autres sources, mais aussi l'utilisation de paraphrases et de substitutions par des synonymes [6].

1.5.2 L'outil Copyleaks

Copyleaks Est un service de détection de plagiat en ligne. vérifier le texte et le contenu de sites depuis une URL. il fournit un ensemble d'algorithmes de détection et utilise l'API Web de Google pour enrichir ses recherches[14].

— Ne Possède pas sa propre base de données.

— C'est un outil payent pour la vérification textuelle et URL mais qu'il dispose d'une version limitée gratuite pour la vérification de 9 essais.

1.5.3 L'outil Aplag

Il s'agit d'une abréviation pour *Arabic Plagiarism* et est considéré comme un logiciel de détection du plagiat pour les textes en arabes, il est publié par le département d'informatique de l'université du Roi-Saoud, année 2011. Il s'appuie sur la représentation logique des textes sous forme de paragraphes, de phrases et de mots afin que chaque phrase et chaque mot prenne des nombres entiers qui l'expriment dans l'ordre où ils apparaissent dans le texte [13].

6. www.urkund.com(consulté le 06/05/2020)

1.5.4 L'outil Turnitin

Turnitin, l'un des plus anciens logiciels de détection de plagiat publié en 1998, par un groupe de chercheurs de l'université de Californie "UC Berkeley". Et il est devenu un programme affilié d'Iparadigms LLC(Limited Liability Company). [7].

Turnitin maintient une base de données de comparaison qui comprend :

— Le contenu web actuel et archivé qui est accessible au public.

— Livres, journaux et revues (grâce à ses partenariats avec des éditeurs, des bases de données des bibliothèques, des collections de référence numériques et des publications par abonnement).

— Documents d'étudiants envoyés à Turnitin.

Turnitin utilise un algorithme de comparaison pour trouver des chaînes de mots identiques à celles de son entrepôt de données. Cela signifie que le jugement humain, basé sur divers facteurs et considérations, est nécessaire pour déterminer si un cas de plagiat s'est produit. Turnitin ne peut donc créer que des rapports d'originalité qui montrent le degré de similarité entre un travail soumis et les sources de contenu de la base de données [8].

1.5.5 L'outil PlagScan

PlagScan est un programme commercial (2000 premiers mots gratuits) de détection du plagiat de texte traite toute langue qui utilise le système de codage scientifique UTF8, donc il traite des textes arabes. Il est principalement basé sur la recherche sur Internet (Microsoft Bing) et dans les bases de données des éditeurs (articles scientifiques, revues académiques), et en option il peut rechercher dans n'importe quelle base de données définie par l'utilisateur [9].

1.5.6 L'outil Plagiarisma

Plagiarisma est la plus simple application web de détection de plagiat. La version gratuite comporte un nombre limité de vérifications de plagiat, et plus de 190 langues sont supportées entre elles l'arabe. Il vous permet de le faire : Copier/Coller ou type texte dans le champ approprié afin qu'il soit vérifié, vérifier l'URL fournie, ou le fichier téléchargé de votre ordinateur [10].

Plagiarisma recherche les doublons de votre texte dans Google, Yahoo, Bing, Scholar et Books. Si vous n'utilisez aucun de ces moyens, il est impossible pour Plagiarisma de savoir si vous êtes en train de plagier.

1.5.7 L'outil PlagAware

PlagAware est un moteur de recherche sur le plagiat. Il utilise le moteur de recherche classique (Google) pour détecter et scanner le plagiat, et fournit différents types de rapports qui aident l'utilisateur ou le propriétaire du document à décider si son document a été plagié ou non.

7. https://www.turnitin.com/fr (consulté le 15/07/2020)
8. www.algonquincollege.com (consulté le 06/05/2020)
9. www.plagscan.com(consulté le 06/05/2020)
10. Site de Plagiarisma(consulté le 11/03/2020)

1.6 Conclusion

Dans ce chapitre, nous avons présenter un ensemble de définitions liées aux plagiats textuel , ainsi que certaines termes et concepts assoçiés. Enfin, de plus nous avons rapports certains de ses formes , le définissant et le classant chacun à sa manière, mais avec la large diffusion de ce phénomène dans le domaine scientifique, il est nécessaire de trouver des moyens de le prévenir et de le contrôler et surtout de le détecter.

Chapitre 2

Analyse de besoins et la conception

Dans ce chapitre nous avons fait plusieurs approches pour collecter les besoins et nous avons mené un questionnaire au prés des professeurs et chercheurs a qui est déstiné l'outil de détection plagiat pour la langue arabe, Par la suite nous décrivons notre conception en utilisant les diagrammes d'UML.

2.1 Analyse de besoins

L'analyse de besoins c'est une étape trés importante avant la conception et le développement nous avons utilisé deux approches pour faire une bonne étude, En effet nous avons utilisé un questionnaire traitant des questions a propos l'outil de détection plagiat avec les chercheurs et les professeurs pour collectons les besoins fonctionnels .
Conçernant les besoins liés a l'élaboration l'interface de nous avons tavaillé avec la méthode de prototypage.

2.1.1 Processus d'analyse des besoins

Le questionnaire est un outil méthodologique comportant un ensemble de questions qui s'enchaînent d'une manière structurée.
Notre questionnaire comporte a 10 questions pour les Chercheurs et les Professeurs pour collecter les informations et les besoins (voir l'annexe).
Parmi les questions plus importantes nous mentionnons :
Quel type d 'interface vous avez souhaité ?

- Desktop
- Web
- Hybride
- autre ..

Quels sont les logiciels vous utilisez pour détecter le plagiat ?

- Aplag
- Turnitin/ ethenticate
- Copyleaks
- PlagAware
- Urkund
- PlagScan
- Plagiarisma

— autre ..

À quel niveau votre méthode préférable traite-t-elle les textes saisies ?

— niveau des caracetèrs

— niveau des mots

— niveau des phrases

— autre ..

Quels types de pretraitements vous voulez avoir dans l'application ?

— Parts-of-Speech

— Tokenized Forms

— Lemmas

— Base Phrases

— Named Entities

— tout

Synthèse des résultats de l'enquète

Suite à un ensemble de dix réponses , nous avons dégagé un ensemble de besoins fonctions dont l'essentiel est :

Besoin 1 : Les utilisateurs ont besoin d'un outil de détection du plagiat, et ils souhaitent que support principalement la langue arabe.

Besoin 2 : La possibilité d'ajouter de nouvelles méthodes pour détecter le plagiat est sollicitée.

Besoin 3 : La possibilité de comparaison deux méthodes est une exigence.

Besoin 4 : L'interface souhaite est de type application web.

Besoin 5 : La possibilité d'appel à un ensemble d'outil externes est exprimé.

Besoin 6 : Les utilisateurs expriment une particulière attention ce que les entrées sorties soit en plusieurs formats spécialement via xml

2.2 Prototypage

Le prototypage est la démarche qui consiste à réaliser un prototype, Le prototype est un exemplaire incomplet et non définitif de ce que pourra être le produit ou l'objet final. Il existe deux types de prototypes : le prototype vertical met en œuvre certaines fonctionnalités de l'application afin que l'utilisateur puisse réaliser complètement un scénario typique d'utilisation du logiciel,le prototype horizental correspond au développement de la partie graphique de l'interface homme-machine, c'est parfois une simple maquette sur papier.
(Pour plus de détails sur le prototypage voir l'annexe 3.5)

Nous avons utilisé le prototype pour définir les modules requis par l'utilisateur et les fonctions dont il a besoin dans la boîte à outils et pour faciliter l'utilisation de la boîte à outils.

2.2.1 Conception prototypes

Nous avons choisi le prototype horizontal pour décrire à l'utilisateurs l'interfaçcage. Dans cette section, on se limite juste décrire les interfaçes initiales et finales : (Le lecteur pourra voir plus de détails sur les protoypes des interfaces intermediaires dans l'annexe 3.7,3.6...etc)

— Les interfaces initiales(version initiale) :

Cette interface a été la première que nous avons réalisée, mais il nous est devenu évident qu'elle devait être modifiée et que des nouvelles interfaces devaient être mises en place, faciles à utiliser et plus détaillées.

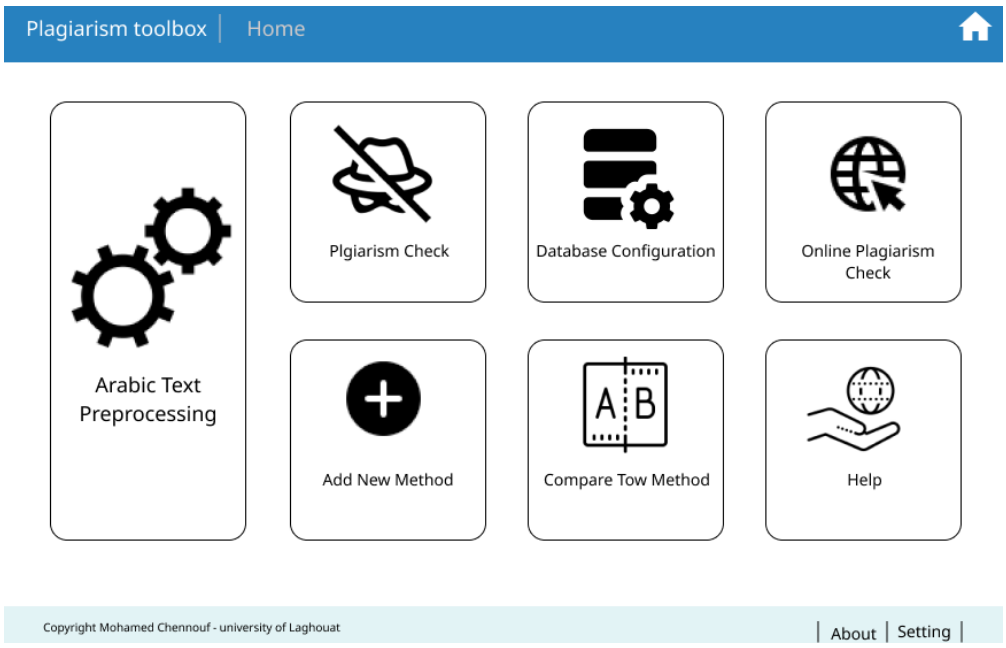

FIGURE 2.1 – la page d'Acceuil de la boite à outils : Version initial

— Les interfaces finales (version finale) :

Après avoir passé en revue plusieurs interfaces et modifications, nous sommes arrivés à une interface finale qui répondait à tous les besoins d'ergonomie, d'enchainement et de fonctionnalités des utilisateurs.

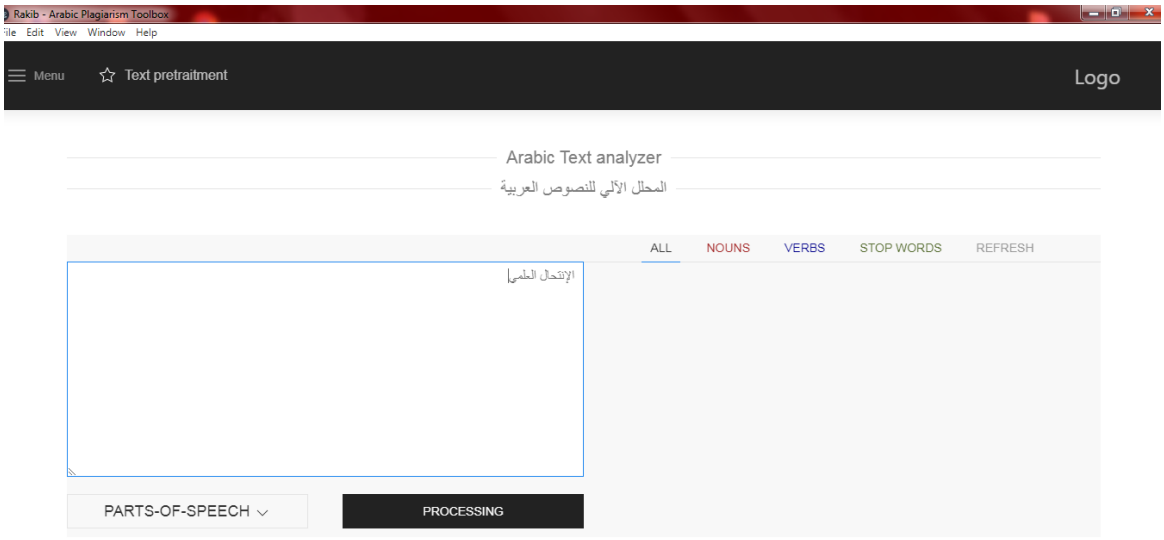

FIGURE 2.2 – Outil détection plagiat (RAKIB) : Version finale

2.3 La conception

Nous avons choisi de faire la modélisation de notre boîte à outils avec le langage UML (Unified Moddeling Language), l'UML est un langage de modélisation unifié est un langage de modélisation graphique et textuel permettant de comprendre et de décrire les exigences, de spécifier et de documenter les systèmes, d'esquisser des architectures logicielles, de concevoir des solutions et de communiquer des points de vues [1],[5].
UML (Unified Langage Modelling) est un langage formel et normalisé, Il permet une plus grande précision et garantit la stabilité, l'UML est un puissant support de communication.
A fin de décrire notre conception on vas utiliser principalement trois types de diagrammes : diagramme de cas d'utilisation, diagrammes des classes et les diagrammes de séquences.

1. http://www.uml.org

2.4 Présentation diagramme de cas d'utilisation

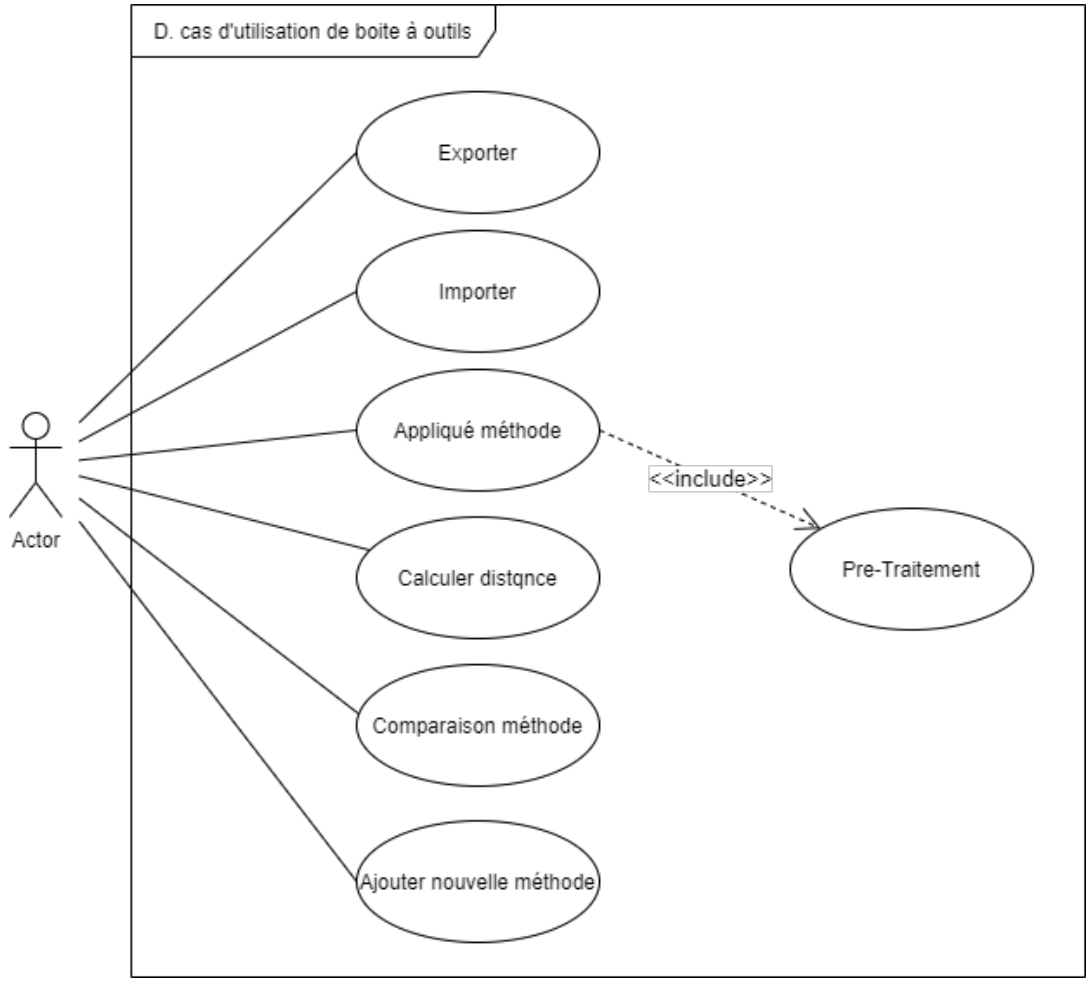

FIGURE 2.3 – Diagramme de cas d'utilisation de boite à outils

2.5 Présentation diagrammes de séquences

Il permet de décrire les scénarios de chaque cas d'utilisation en mettant l'accent sur la chronologie des opérations en interaction avec les objets.
Concernant ce type de diagrammes, nous avons choisi de décrire 3 instances de diagrammes de séquences :

— Appliquer une méthode XX.
— Ajouter une méthode XX.
— Saisie d'un texte.
— Ajouter d'un outil externe.

2.5.1 Appliquer méthode(x)

Scénario : Appliquer méthode (x).

— Après l'accès à l'interface, l'utilisateur peut appliquer une méthode.
— Il fait l'opération de segmentation ,pre-traitement et aplique méthode x.

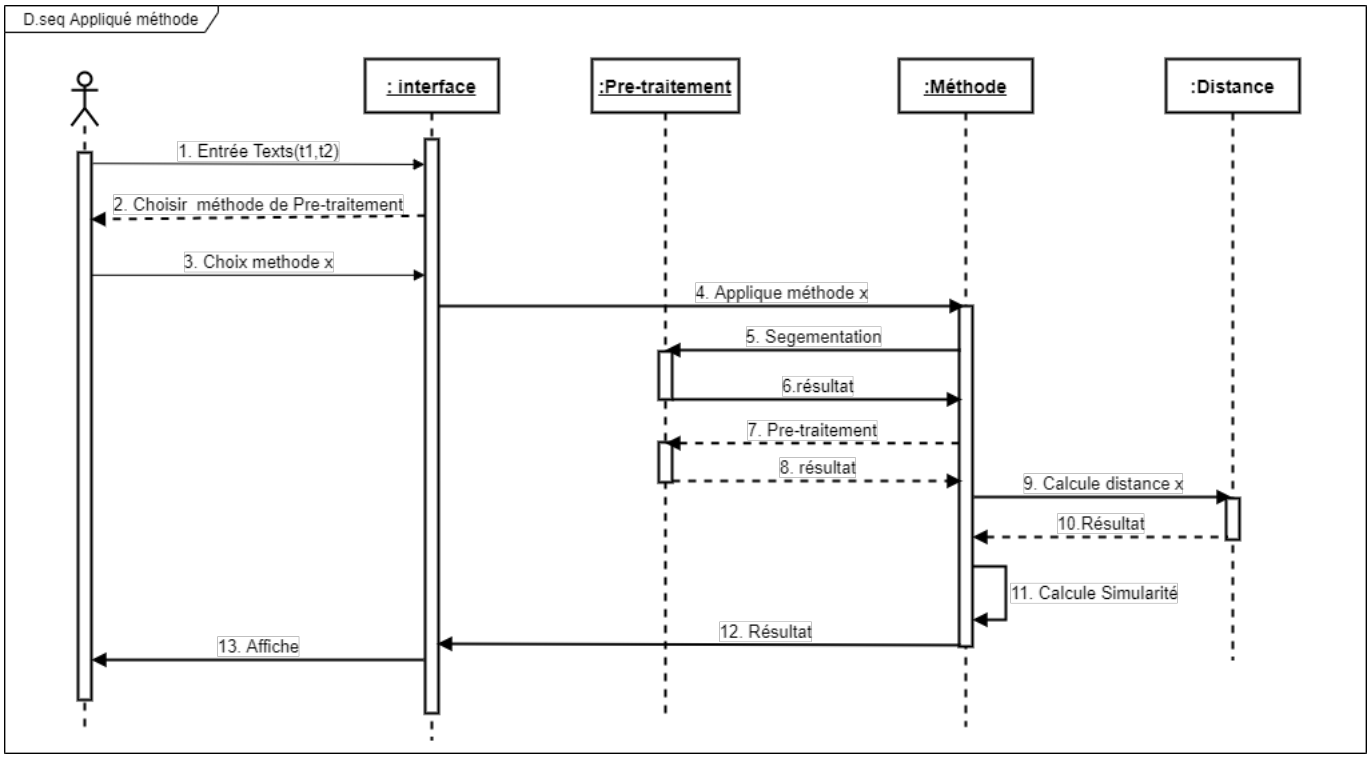

FIGURE 2.4 – Diagramme de séquence "Appliquer méthode XX"

2.5.2 Ajouter nouvelle méthode

Scénario : Ajouter nouvelle méthode.

— Après l'accès à l'interface, l'utilisateur peut demande ajouter une nouvelle méthode.

— Si n'existe pas un erreur le méthode aas ajouté avec succées.

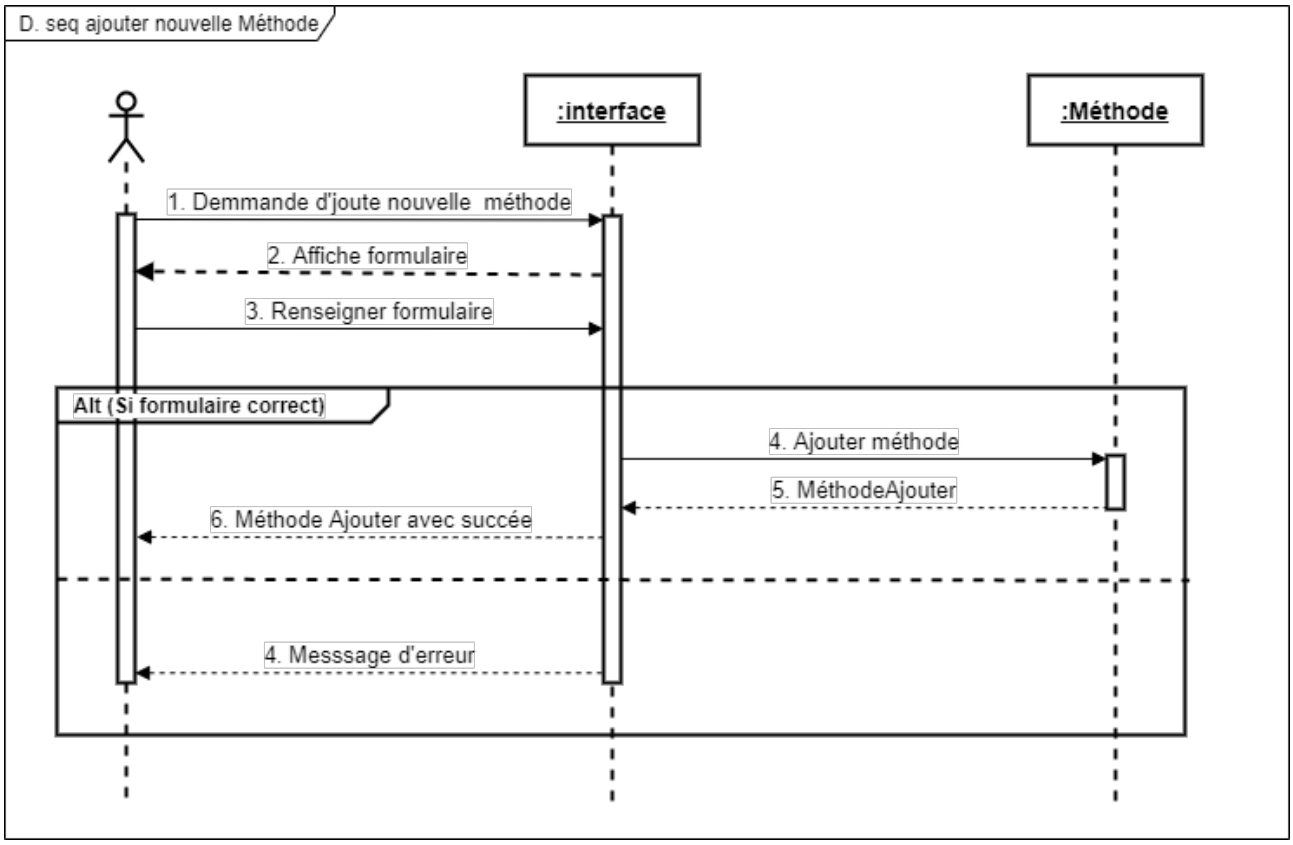

FIGURE 2.5 – Diagramme de séquence "ajouter méthode"

2.5.3 Ajouter un outil externe

Scénario : Ajouter un outil externe .

— Après l'accès à l'interface, l'utilisateur peut demande ajouter une nouveau outil externe.

— Si n'existe pas un erreur l'outil as ajouté avec succées.

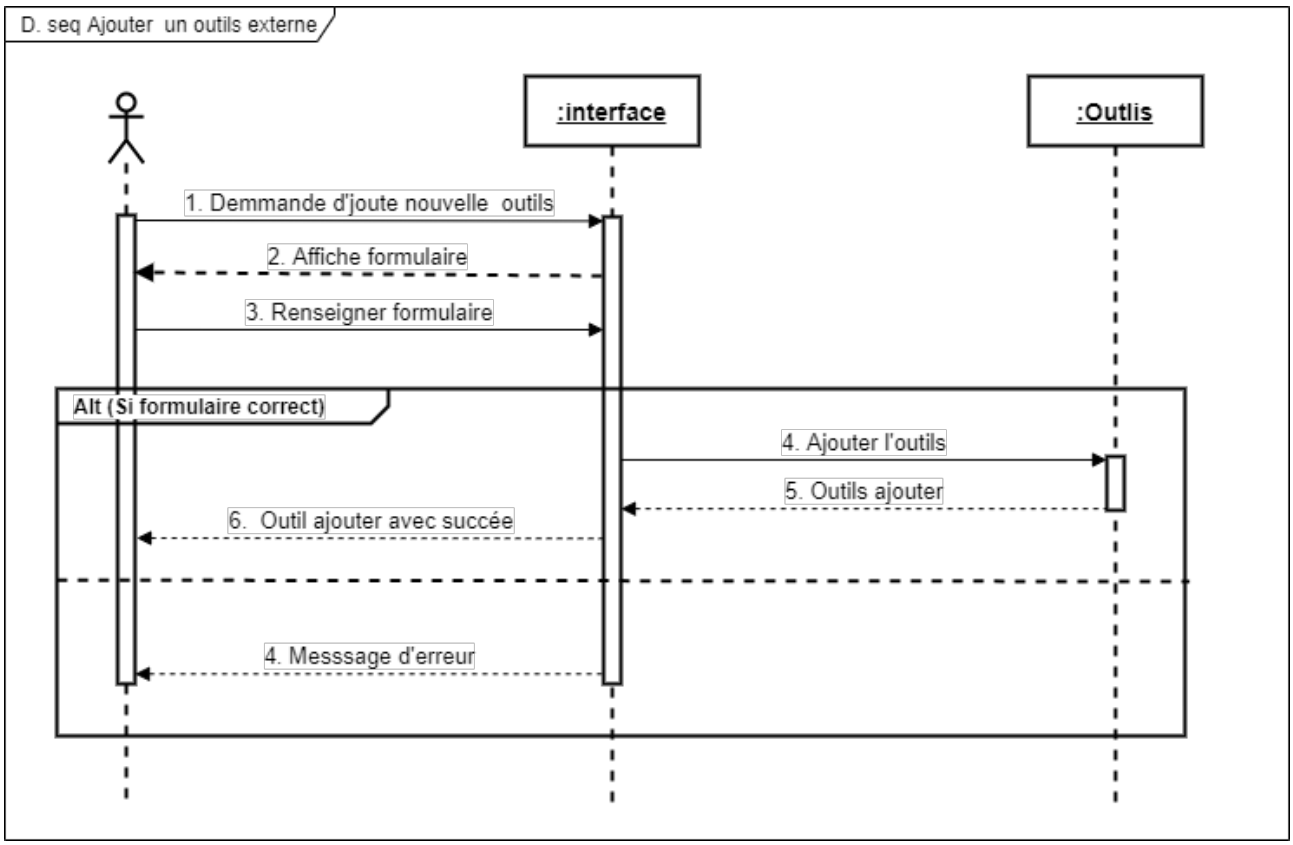

FIGURE 2.6 – Diagramme de séquence "Ajouter un outil externe"

2.6 Présentation diagramme de classes

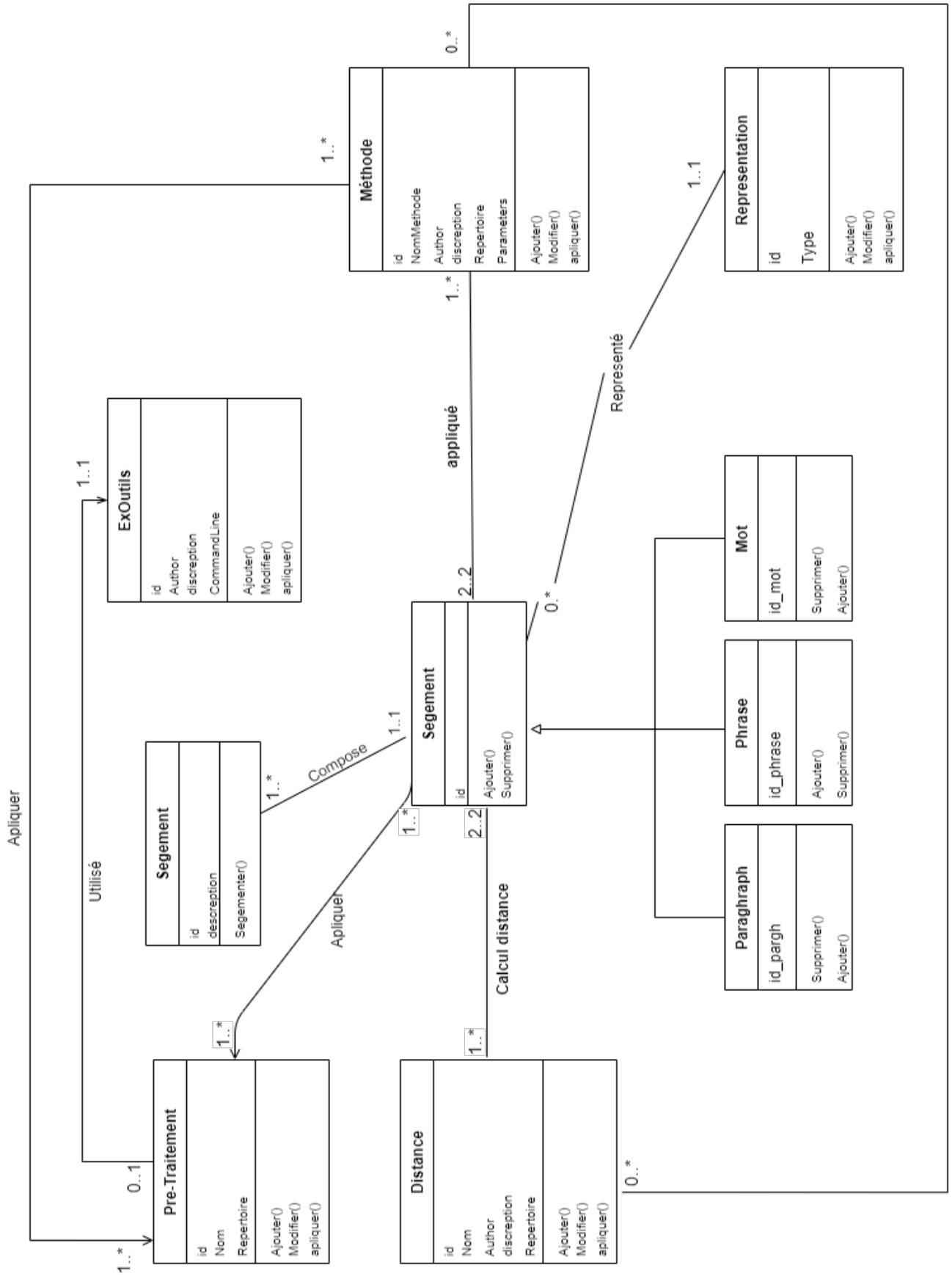

FIGURE 2.7 – Diagramme de classes de la boite à outils

2.7 Conclusion

Ce chapitre a été consacré au design de notre boite à outils. On a décrit l'étude de besoins en addition à la présentation de l'aspect conceptuel d'un boite à outils travers les différents diagrammes décrits en UML, des diagrammes de séquence qui illustre le processus de visualisation des cas d'utilisation et diagramme qui représente les différentes classes dans notre boite à outils.

Chapitre 3

L'implementation

Dans ce chapitre, nous présenterons l'environment de développement et les différents programmes qui sont utilisés pour développer notre application et ses fonctionnalités. Nous commençons par présenter l'environnemment de développement puis on présente le produit développé a travers ses interfaces.

3.1 Langage et environment de développement

Nous avons développé notre application essentiellement avec language python :

Python

Python est un langage de programmation trés riche par rapport aux autres langages de programmation Il est facile a apprendre , et il possède des structures de données haut niveau, il est efficace de la programmation orientée objet.
Parmi les caractéristiques on trouve :

- L'utilisation de python gratuit.
- Syntaxe de python et simple. il Combine des types de données évoluées.
- Python est multi-platforme (utilisée dans plusieurs systém d'exploiatation).

Electron

Electron permet de créer des applications bureautique est un framework open source qui utilise HTML, CSS et Javascript , Il est géré par Chromium et Node js.

L'électron a plusieurs propriétés, dont les plus importantes sont :
C'est une mult-plateforme qui permet de réaliser des applications pour Linux, MacOS et Windows, Il comporte également des éléments natifs tels que des menus et des notifications, ainsi que des outils de développement utiles pour le débogage et les rapports de panne.

Nodes js

NodeJS est une plateforme basée sur le moteur JavaScript de Chrome qui permet de développer des applications en utilisant JavaScript.

Java Script

JavaScript est un langage de programmation de scripts principalement employé dans les pages web interactives mais aussi pour les serveurs avec l'utilisation (par exemple) de Node.js . Il s'exécute à l'aide d'un programme spécial appelé "Moteur Javascript". C'est un langage orienté objet à prototype.

HTML/CSS

HTML signifie « HyperText Markup Language » qu'on peut traduire par « langage de balises pour l'hypertexte ». Il est utilisé afin de créer et de représenter le contenu d'une page web et sa structure. CSS est l'un des langages principaux du Web ouvert et a été standardisé par le W3C. Ce standard évolue sous forme de niveaux (levels), Nous avons utilisé CSS3, qui est découpé en modules plus petits.

3.2 Les outils et les bibliothèques dédiés la détection plagiat dans la langue Arabe

Parmi les outils et les bibliothèques de la détection de plagiat on a déployé :
Madamira qui est un outil pour traitement et segmentation dans la langue arabe et les bibliothèques de Python NLTK(Natural language toolkit),naftawayh,pyarabic .

Naftawayh

Naftawayh[1] est une bibliothèque Python pour le marquage des mots arabes (classification des mots) en types (noms, verbes, mots vides) développé par *Taha Zerrouki*[15], ce qui est utile dans le traitement du langage, notamment pour l'exploration de textes. Naftawayh fonctionne selon la structure des mots arabes, et la capacité de deviner la classe de mots, à travers certains signes.

Pour installer Neftawayh on utilise la commande : `pip install neftawayh`

pour utilise la bibliothéque :

```
import naftawayh.wordtag as wordtag
```

la figure aux dessous illustre une sortie de cette commande : v= verbe , n= nom, t= stop mots

1. Naftawayh: Arabic Word Tagger Site web.

```
>>> import naftawayh.wordtag
>>> word_list=(u'بالبلاد', u'بينما', u'أو', u'انسحاب', u'انعدام',
u'انفجار', u'البرنامج', u'بانفعالاتها', u'العربي', u'الصرفي',
u'التطرف', u'اقتصادي', )
>>> tagger = naftawayh.wordtag.WordTagger();
>>> # test all words
>>> list_tags = tagger.word_tagging(word_list)
>>> for word, tag in zip(word_list, list_tags):
>>>     print word, tag
بالبلاد n
بينما vn3
أو t
انسحاب n
انعدام n
انفجار n
البرنامج n
بانفعالاتها n
العربي n
الصرفي n
التطرف n
اقتصادي n
```

FIGURE 3.1 – Exemple d'utilisation bibliothéque Neftawayh

Natural Language Toolkit (Nltk)

Nltk, est une boîte-à-outils permettant la création de programmes pour l'analyse de texte. Nltk a été créé par Steven Bird et Edward Loper

pour installer nltk on utilise la commande :`pip install nltk`

pour l'utilisation de ntlk : `import ntlk`

Pyarabic

Pyarabic[2], une bibliothèque spécifique la langue arabe developpé en Python par *Taha Zerrouki*[16], fournit des fonctions de base pour manipuler les lettres et le texte arabes, comme la détection des lettres arabes, des groupes de lettres arabes et de leurs caractéristiques, la suppression des diacritiques, etc.

pour installer Pyarabic on utilise la commande :`pip install Pyarabic`

pour l'utilisation de Pyarabic : `import Pyarabic`

3.3 Présentation du produit (l'outil développé)

L'outil a été baplisé RAKIB provenant du mot arabe **رقيب** pour illustrer sa fonction principale de détection du plagiat.
Dans ce qui suit on présente un ensemble de ses fonctionnalités.
L'interface d'acceuil de l'outil Rakib permet de saisie d'un texte manuellement pour lui associe des pretraitement tels que : post-tag , lemmatisation , tokinesation et base phrase (voir la figure suivante).

2. PyArabic: Python Library for Arabic's documentation.

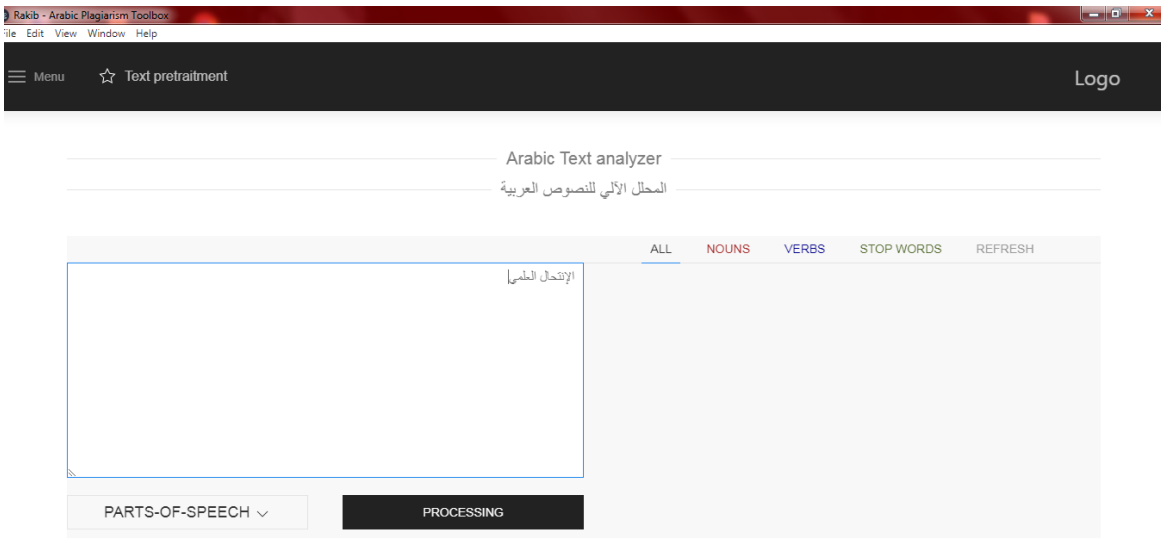

FIGURE 3.2 – Outil détection plagiat (RAKIB) : Interface d'acceuil

Dans ce menu, l'utilisateur peut accéder à certains fonctions de la boite à outils : La première est le "text prétraitement" et cette interface demarre comme page d'acceuil .
Dans l'interface "plagiarism détection" , l'utilisateur peut vérifier le texte arabe et le fichier arabe pour le plagiat en choisissant.
Dans l'interface "méthode configuration" ,l'utilisateur peut ajouter une nouvelle méthode ou comparer deux méthodes.

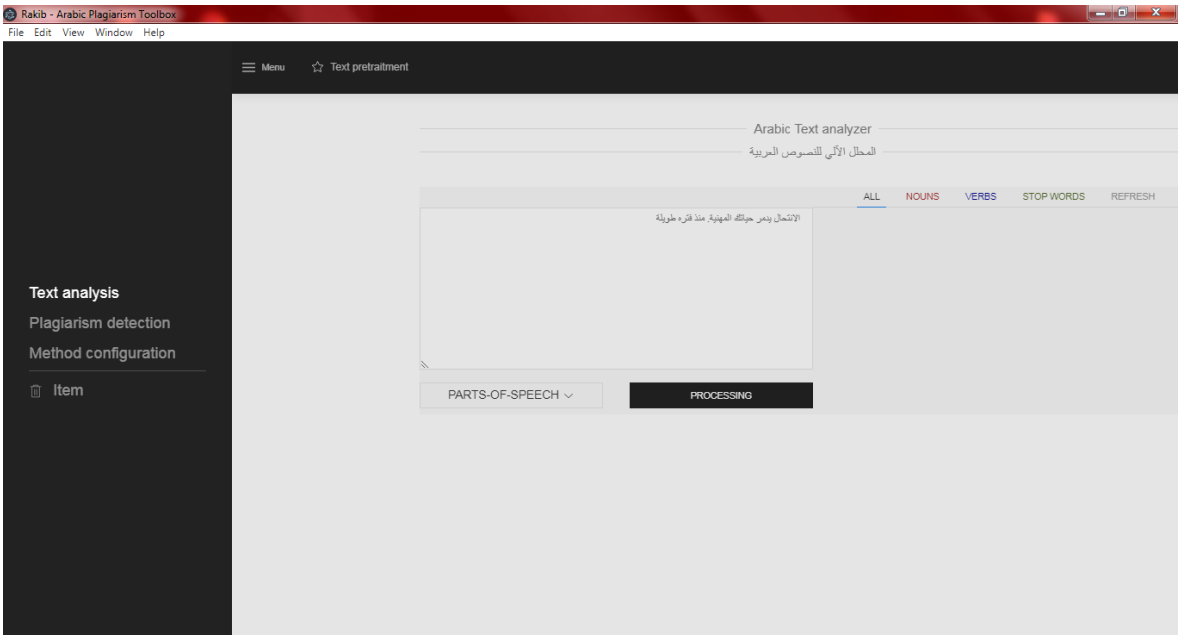

FIGURE 3.3 – Outil détection plagiat (RAKIB) :interface prétraitement

L'interface suivant permet aux chercheurs saisie manuallement ou importer fichier et afficher les types de plagiat a détecté

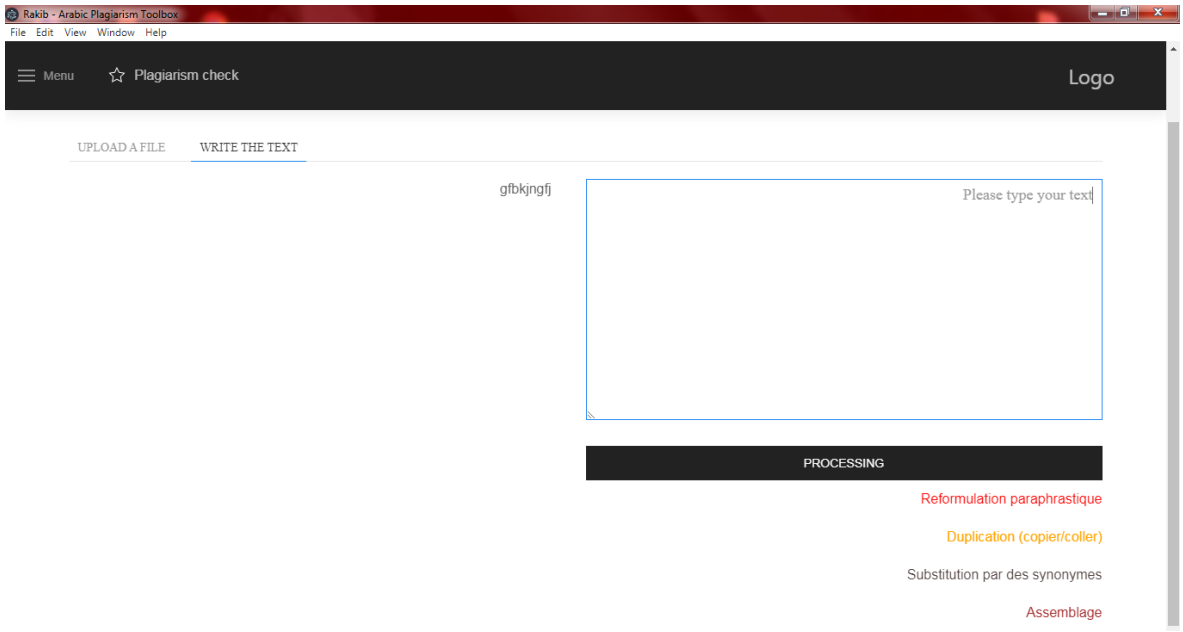

FIGURE 3.4 – Outil détection plagiat (RAKIB) :Interface détecte types de plagiat (RAKIB)

Conclusion et perceptives

Dans ce mémoire nous avons développe une boite a outil pour les chercheurs et professeurs dédié a détection le plagiat pour la langue arabe, nommée **راقب**(en anglais "Rakib"). L'interface est développé en utilisant *Electron* et l'infrastructure en *python/node.js*.
Elle est conçue pour aider les chercheurs à tester et à comparer leur méthode de détection du plagiat avec d'autres méthodes. La principale fonction de cette boîte à outils est d'analyser les textes arabes (Tokenization, stop-words removal, Lemmas ..), l'interface de détection du plagiat donne à l'utilisateur la possibilité de copier/coller ou de charger des documents et de choisir la méthode de détection pour le traitement, l'interface de configuration des méthodes de détection du plagiat permettant d'ajouter de nouvelles méthodes de détection en remplissant un formulaire, l'interface de comparaison des méthodes il suffit de choisir deux méthodes déjà ajoutées dans la boîte à outils et d'effectuer la comparaison.

Perspective : Suite aux travaux effectués, plusieurs points restent à développer et à améliorer, parmi lesquels citons :

— Alimenter la base de données par les textes (corpus, collection) pour faciliter la comparaison des méthode une échelle plus importante.

— Évaluer les fonctionnalité de la boite outils par experts de détection de plagiat pour perfectionner son fonctionnement.

Bibliographie

[1] Teddi Fishman, "We know it when we see it" is not good enough : toward a standard definition of plagiarism that transcends theft, fraud, and copyright, 4th Asia Pacific Conference on Educational Integrity , September 2009.

[2] Frédéric Baudry, Grammaire comparée des langues classiques : contenant la théorie élémentaire de la formation des mots en sanscrit, en grec et en latin avec références aux langues germaniques, volume 1, 1868.

[3] El Moatez Billah NAGOUDI, Détection automatique de plagiat, Thése de doctorat Université Telidji Laghouat, 2017.

[4] Anna Ritchie, Simone Teufel, and Stephen Robertson, How to find better index terms through citations. In Proceedings of the workshop on how can computational linguistics improve information retrieval ?, Association for Computational Linguistics, p. 25–32, 2006.

[5] Joseph Gabay, David Gabay " UML 2 analyse et conception : Mise en oeuvre guidée avec études de cas" , Dunod 2008.

[6] Paul Clough et al, Old and new challenges in automatic plagiarism detection, Plagiarism Advisory Service, University of Sheffield 2003.

[7] Jérémy Ferrero, Similarités Textuelles Sémantiques Translingues : vers la détection automatique du plagiat par traduction, La communauté université grenoble alpes, 2017.

[8] Debora Weber-Wulff, Test cases for plagiarism detection software. In Proceedings of the 4th International Plagiarism Conference, 2010.

[9] Lawrence R Ness et al, The ethical implications of plagiarism and ghostwriting in an open society, Journal of Social Change Volume 9, 2017.

[10] Borisov A et al, Research into plagiarism cases and plagiarism detection methods, Scientific journal of Riga Technical University Vol.44, 2010.

[11] Imtiaz Hussain Khan et al, "A framework for plagiarism detection in arabic document", Computer Science & Information Technology, 2015.

[12] Olivier Millet, Dictionnaire des citations, Librairie générale française, 1997.

[13] Mohamed El Bachir Menai, Manar Bagais APlag : A Plagiarism Checker for Arabic Texts, Université King Saud, The 6th International Conference on Computer Science & Education, 2011.

[14] Hussain A Chowdhury, Dhruba K Bhattacharyya Plagiarism : Taxonomy, Tools and Detection Techniques, Dept. of CSE, Tezpur University,

[15] Zerrouki TahaArabic Word Tagger,Article, 2010.

[16] Zerrouki Taha An Arabic language library for Python, Article, 2010.

Annexe

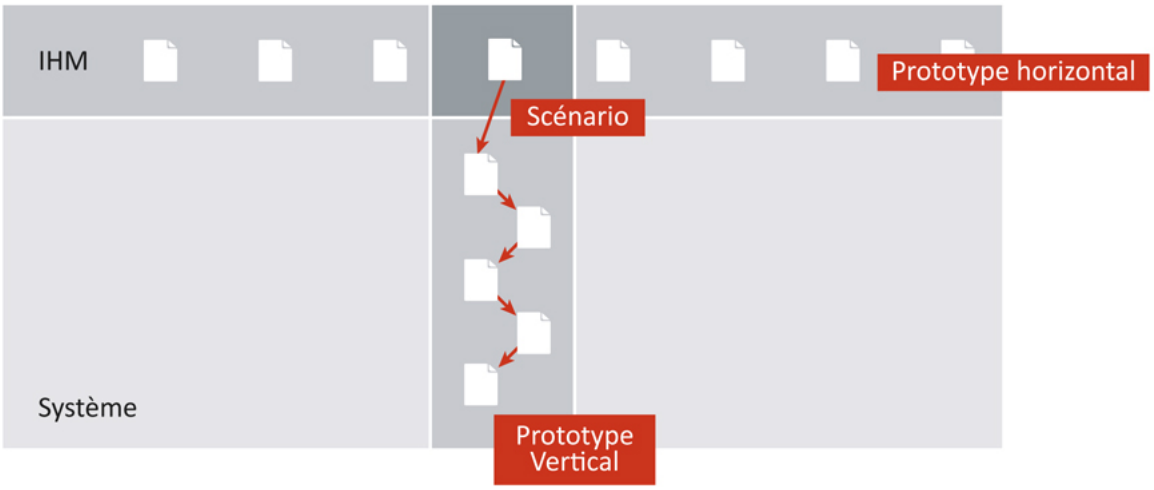

FIGURE 3.5 – Types de prototypes

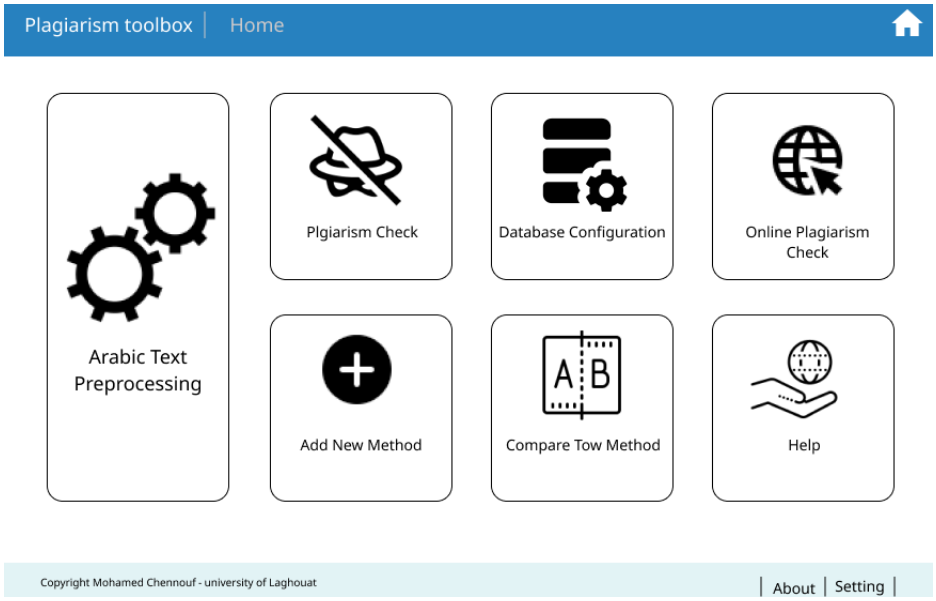

FIGURE 3.6 – L'interface d'acceuil

Questionnaire pour logiciel anti_plagiat

Bonjour,

Le questionnaire auquel vous participerez a été créé dans le cadre d'un projet de création d'une boîte à outils dédiée aux systèmes de détection du plagiat. Nous vous invitons à répondre le plus sérieusement possible aux questions qui vous seront posées.

Nous vous remercions par avance de votre aide.

*Obligatoire

1. Adresse e-mail *

_____

2. 1/ quel le type d 'interface vous avez souhaité? *

Plusieurs réponses possibles.

- [ ] Desktop
- [ ] Web
- [ ] Hybride
- [ ] Autre : _____

3. 2/ quels sont les outils auquel vous voulez ?

Plusieurs réponses possibles.

- [ ] Madamira
- [ ] NLTK
- [ ] NEFTAWAYH
- [ ] Autre : _____

4. 3/ Quel sont les logiçiels vous utiliez pour détecter le plagiat ? *

Plusieurs réponses possibles.

- [ ] Aplag
- [ ] Turnitin/ ethenticate
- [ ] Copyleaks
- [ ] PlagAware
- [ ] Urkund
- [ ] PlagScan
- [ ] Plagiarisma
- [ ] Autre : _____

5. 4/ Prend-il en charge la langue arabe? *

Une seule réponse possible.

- ◯ 0%
- ◯ 25%
- ◯ 50%
- ◯ 75%
- ◯ 100%
- ◯ Autre : _____

6. 5/ Quelles sont les méthodes qui vous préferez utiliser pour detecter plagiat ? *

Plusieurs réponses possibles.

- [ ] Empreinte Digitale
- [ ] analyse des occurrencesdes mots
- [ ] Moteur de Recherche
- [ ] word Embedding
- [ ] Autre : _____

7. 6/ À quel niveau votre méthode préferable traite-t-elle les textes saisies ? *

Plusieurs réponses possibles.

- [ ] niveau des caracheters
- [ ] niveau des mots
- [ ] niveau des phrases
- [ ] Autre : _____

8. 7/ Comment voulez-vous prétraiter le texte(la granularité de votre analyse)? *

Plusieurs réponses possibles.

- [ ] aux niveau paragraphes
- [ ] aux niveau phrases
- [ ] aux niveau aux mots
- [ ] Autre : _____

9. 8/ Comment vous souhaitez affichent les resultats de votre méthode pretraitement? *

Plusieurs réponses possibles.

- [ ] Afficher dans une interface séparée
- [ ] Enregistrer dans un fichier .txt
- [ ] en format binaire

10. 9/Comment voulez vous faire rentrer vos données textuelles ? *

Plusieurs réponses possibles.

- [ ] via la saisie à travers un éditeur
- [ ] via un fichier txt (.txt)
- [ ] via un fichier XML (.xml)
- [ ] un fichier rtf (.rtf)
- [ ] Importer à partir d'une collection
- [ ] Autre : _____

11. 10/ Quels types de pretraitements vous voulez avoir dans l'application ? *

Plusieurs réponses possibles.

- [ ] Parts-of-Speech
- [ ] Tokenized Forms
- [ ] Diacritized Forms
- [ ] Lemmas
- [ ] Base Phrases
- [ ] Named Entities
- [ ] tout

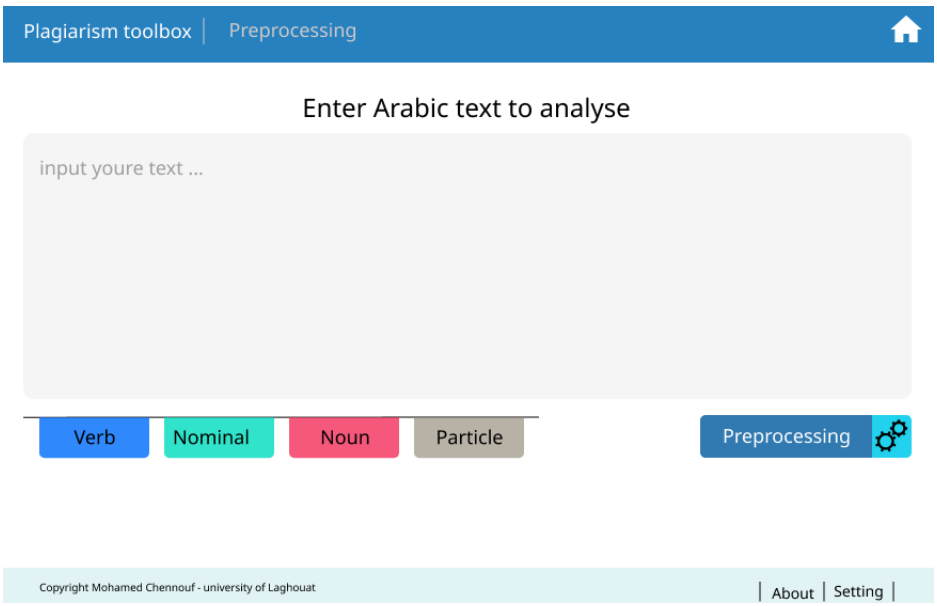

FIGURE 3.7 – L'interface de pretraitement

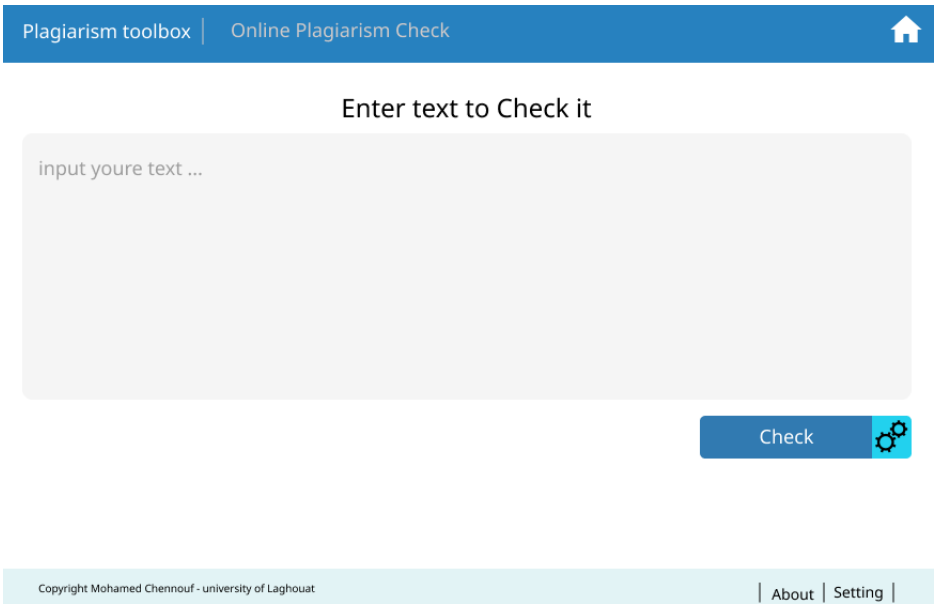

FIGURE 3.8 – L'interface de Détection plagiat

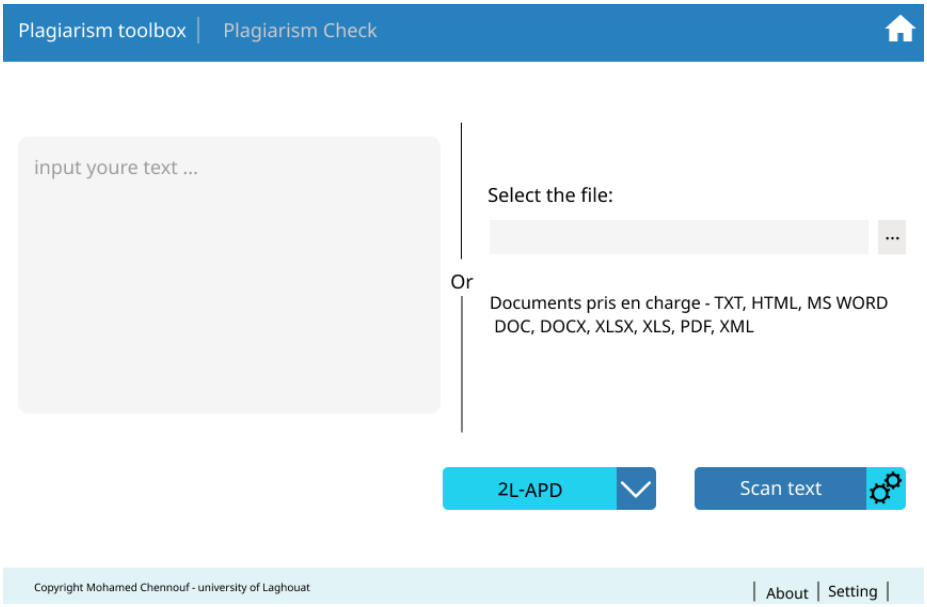

FIGURE 3.9 – L'interface de méthodes

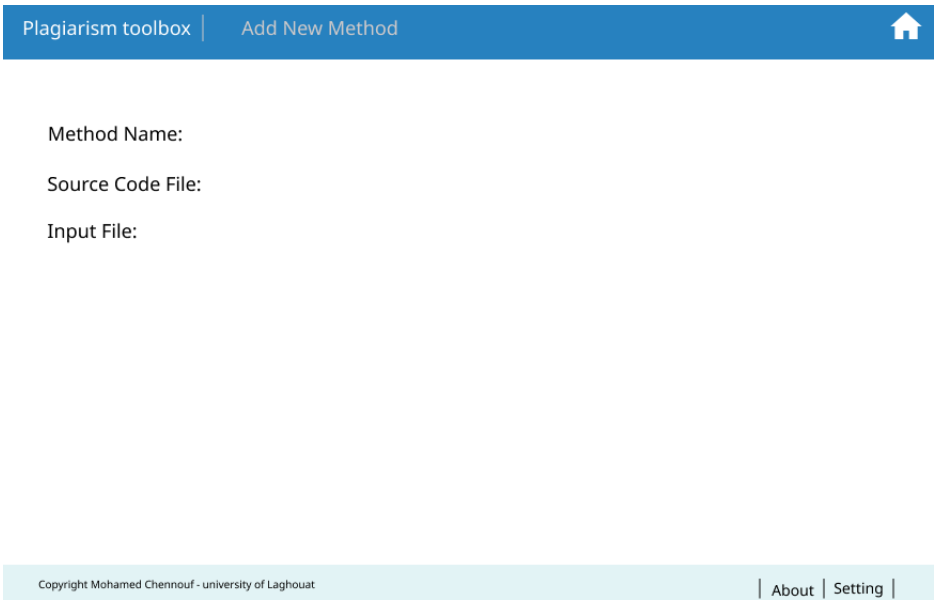

FIGURE 3.10 – L'interface de insertion méthode

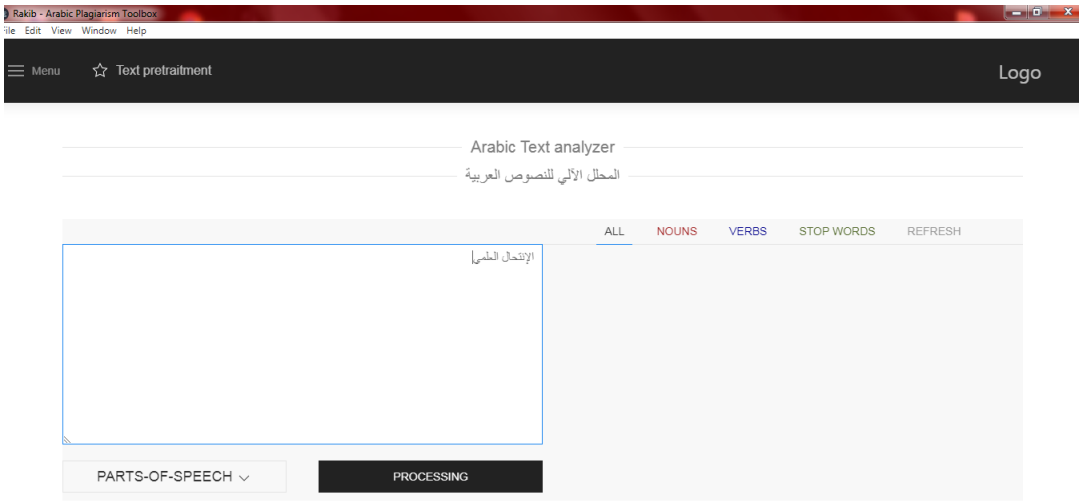

FIGURE 3.11 – L'interface home de RAKIB

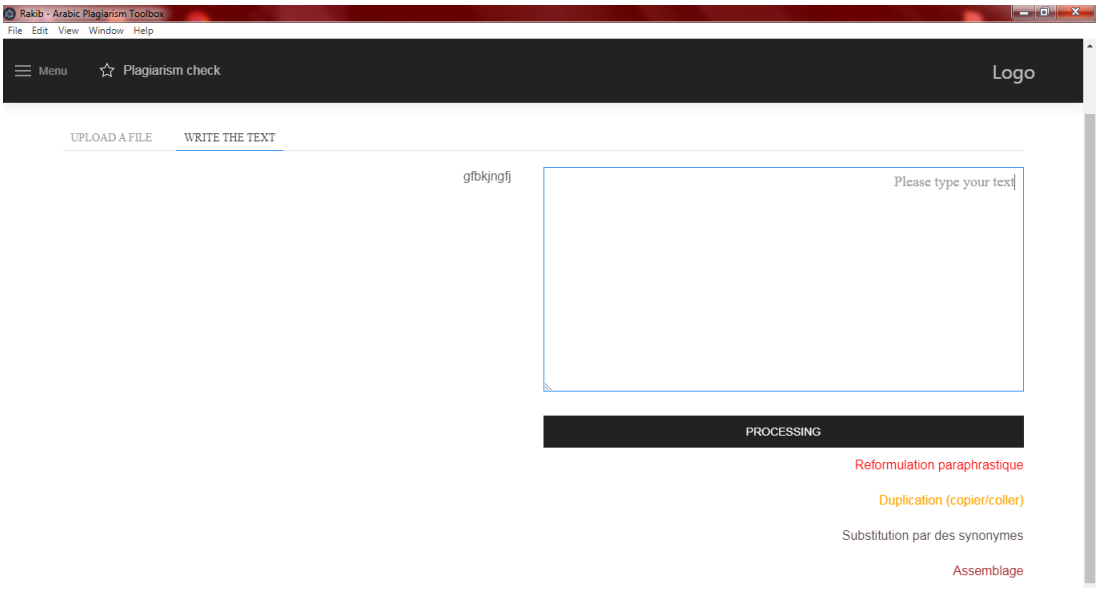

FIGURE 3.12 – L'interface pretraitement de RAKIB

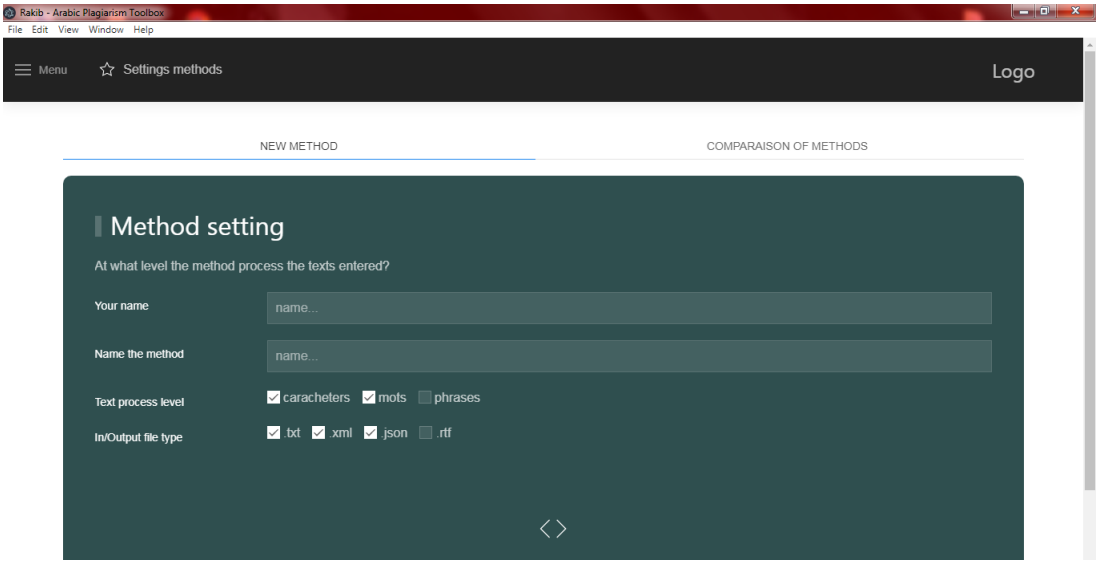

FIGURE 3.13 – L'interface methodes de RAKIB

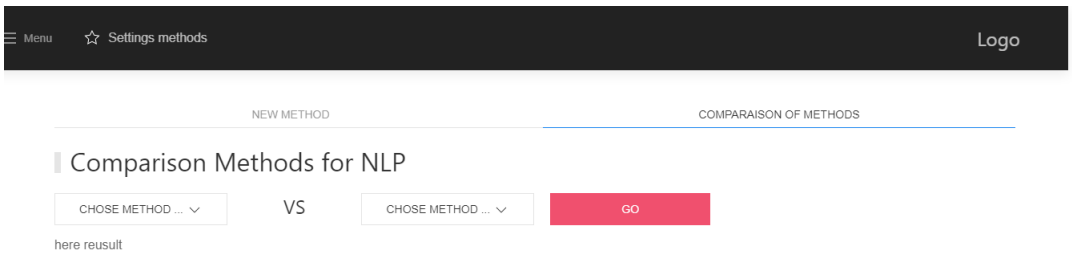

FIGURE 3.14 – L'interface comparaison methodes de RAKIB