People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Mohamed Khider University Biskra
Faculty of Exact Sciences and Sciences of Nature and Life
Computer Science department

Thesis

Presented for

Master's degree graduation in Computer Science

Option : **Artificial Intelligence (AI)**

Title

Ontological approach to detect plagiarism between Arabic documents based on semantic similarity

Presented by :

Khelili Mohamed Akram

Presented on 20/06/2019, in front of the jury composed of :

Jury's President	**Dr. ....**	**University of Biskra**
Supervisor	**Dr. ....**	**University of Biskra**
Assistant Supervisor	**Dr. ....**	**University of Biskra**
Examiner	**Dr. ....**	**University of Biskra**

Academic year : **2018 − 2019**

Acknowledgements

Before all, I thank **ALLAH** for giving me the strength, courage and hope to do this work.

I would like to express my gratitude to my supervisors **Dr.Rezeg Khaled** and **Dr.Zouaoui Samia** that I wish to express to them my sincere thanks for the quality of their supervision, their availability and for their advice and remarks. I really appreciate their intellectual and human qualities.

My sincere thanks also go to the members of the jury who honored me with their presence and for agreeing to review and improve the quality of this modest work and to enrich it with their proposals.

I thank My teacher **Dr.Aichi Asma** for the help that she gave me and her support. I thank all my teachers who taught me all of these years of study. I would like to express my gratitude to them.

And of course i will not forget the dearest people to me : my family and my friends, thank you for being always there for me and for supporting and encouraging me.

Thank you for everything. Thank you from the heart.

Finally, a big thankfulness to all those who has helped me, and supported me from near and far, with one way or another during all these years.

Thanks everyone.

Dedication

I dedicate this work to my dear parents for their love, their support and sacrifice, their patience, and their encouragement. May God protect them.
To my brother Samer and my sisters.
To my grandfather and my grandmother.
To my aunts and my uncles.
To all my family.
To my friends : Ali, Oussama, Amir, Othmane, Ihab, Mondher, ... who are always present by my side, I want to thank you for all these moments together. I am very happy to know you.
To all my loved ones and everyone who loves me.
To all my colleagues.
To all my teachers.

Khelili Mohamed Akram

Abstract

In the context of Arabic plagiarism detection systems (APDS) using an Arabic ontology, and permitting these systems to support semantic representation, to better verify the originality of the research and meet the needs of researchers, this thesis aims to semantically index documents by selecting the best elements from Arabic WordNet. We have analyzed each sentence in the document, have extracted Part-of-Speech (PoS), have extracted Arabic WordNet synonyms for each word and then we have created the first index. The latter is then used by the Lucene program to create the second index, which is used to detect plagiarism in Arabic documents. Our experience is based on a corpus of Arabic text, which we have manually created with the interaction of an expert. The results obtained showed us that the proposed system has proved its performance and its efficacy in detecting the plagiarism in Arabic documents.

key words : Arabic Documents, Arabic Ontology, Semantic Indexing, Semantic Similarity, Plagiarism Detection.

Contents

Bibliography **78**

List of Figures

List of Tables

Chapter 1

General Introduction

With the development and the growth of information and communication technologies, which have been a direct cause of the development of modern published media and in the light of the explosion of knowledge and information that we are experiencing today, the user can easily find the information and use it without gives attention to its copyrights, this is what create new phenomenon which called plagiarism.

Currently, Scientific plagiarism is one of the most widely found phenomena in literary and scientific communities, which should not be used by the scientific researcher.The phenomenon of plagiarism and the diffusion of ideas, words or researches of an another author has became widely disseminated and unacceptable . This phenomenon does not only prevail within the academic community (teachers and students), but also outside such as the press, literary works and movies.

To protect the intellectual rights, many researchers give a great attention to develop an accurate tools or applications of plagiarism detection to verify the originality of research. In the literature, we found that many tools have proven very successful in Latin languages, especially with English, but with Arabic there is a lack of applications that support it, especially for the plagiarism detection with semantic representation.

For this reason, it is important to develop programs, tools and algorithms to detect scientific plagiarism that can protect and preserve the moral and material rights of all parties, especially if it is not intentional.

Work context :

Our work is interested to develop an ontological tool to detect plagiarism between Arabic documents. We focus on the semantic aspect, which is a branch of linguistics.

We study two aspects: the meaning of words and the relationships between words, and at the same time, we want to represent semantically documents by using Arabic WordNet (AWN). The AWN consists in categorizing and helping us to explain the relations between the different concepts in the same field.

The main objective is to calculate the semantic similarity between two Arabic documents in order to detect if there was plagiarism, that is to say, a flight of ideas.

Chapter 2

Ontology

2.1 Introduction

Nowadays, the use of ontologies in information systems has become more and more popular in various fields, such as web technologies, multi agent systems, natural language processing, . . . etc. Artificial intelligent researchers have initially borrowed the word " ontology " from Philosophy, then the word spread in many scientific domain. In this chapter we will present general idea about the ontology (definition , creation , motivation,. . .).

2.2 Definition

The word "Ontology" has different meanings in different communities :

- **The first meaning**, refers to a philosophical discipline, which treats the structure and the nature of reality [7]. Aristotle defined Ontology as the science of " being qua being "[7]. Which aim at discovering and modeling reality under a certain perspective.
- **The second meaning**, refers to the most widespread use in computer science, the meaning of ontology here is like a special kind of information object or computational artifact.

Gruber defined ontology as " an explicit specification of a conceptualization " [8]. It is a model for describing the concepts and relationships between them in a hierarchical way.

2.3 The components of an ontology

An ontology generally consists of the following elements [9][10]:

- **The concepts:** these are the words used in a field of study and whose meanings must be determined. These concepts are defined in the description of the ontology through the classes.
- **The properties of a concept:** these are the characteristics of a concept. These properties can be invisible in an ontological schema (diagram that represents an ontology), they can be also described through actions that complete the ontology.
- **Relationships between concepts:** they can be in different types (inheritance for example). These relationships are usually represented in an ontological schema by arrows between the different concepts. In the description of the ontology, they are described by the object properties of the classes defining the concepts concerned.
- **Instances of concepts:** Instances of concepts are the individuals of the ontology.

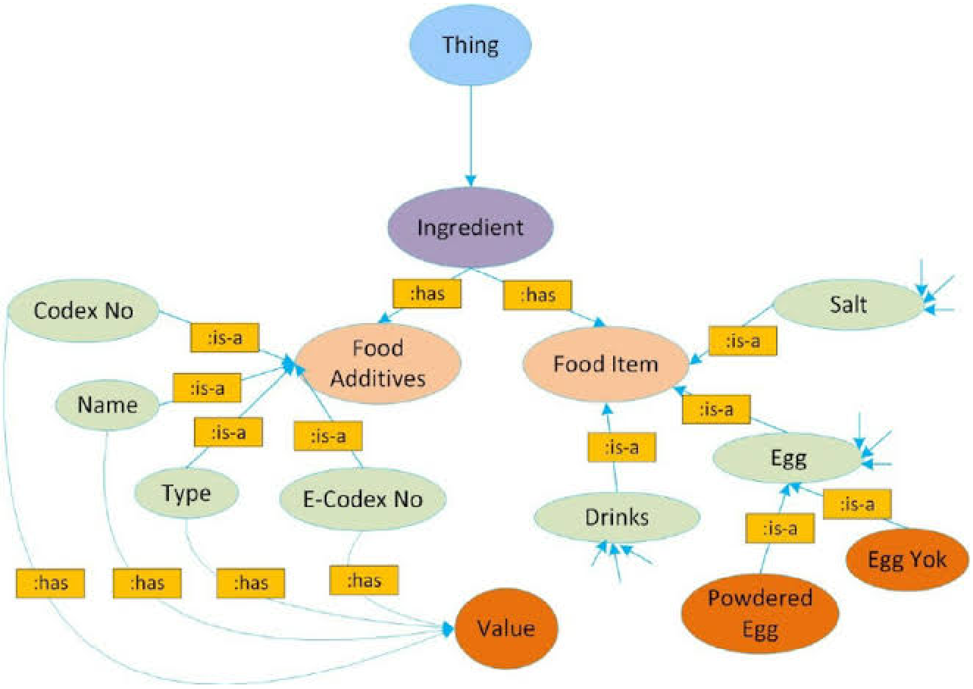


FIGURE 2.1: Example of ontology [1].

2.4 Ontology description languages

According to [10] the ODL contains :

2.4.1 The RDF language

Resource Description Framework (RDF) is the semantic web core language, published by the W3C consortium on 10/02/2004. It is a graphical model for formally describing Web resources and their metadata, so that allow the automatic processing of such descriptions. RDF's role is to link each basic concept to other definitions in order to make sense of it. RDF documents can be written in different syntaxes, including XML. It is possible to use other syntaxes to express the triplets. RDF is simply a data structure consisting of nodes and organized into graphs. Although RDF/XML (the XML version proposed by W3C) is only a serialization of the model, it is often called RDF [10].

2.4.1.1 RDF structure

An RDF structured document is a set of triplets.The components of an RDF triplet are[10]:

1. **The subject :** represents the resource to describe.

2. **The object :** represents a data or other resource.

3. **The predicate :** represents a property type applicable to this resource.

2.4.1.2 RDF syntax

There are 3 different RDF syntaxes [10]:

1. **RDF:** description of web resources (metadata).

2. **RDF Schema (RDFS):** vocabularies for describing ontologies.

3. **XML syntax:** exchange of metadata and schemas.

2.4.2 The OWL language

Ontology Web Language (OWL) is, like RDF, a language that takes advantage of the syntactic universality of XML. Based on the syntax of RDF/XML, OWL offers a way to describe web ontologies. It is precisely an ontology language. RDF and RDFS provide to the user the ability to describe classes and properties[10].

OWL integrates tools for comparing properties and classes, identity, equivalence, cardinality, symmetry, transitivity, disjunction,...etc. Thus, OWL offers machines a greater ability to interpret web content due to a broader vocabulary and a real formal semantics[10].

2.4.2.1 Syntax of OWL

OWL syntax based on RDF/XML, which explains the presence of all elements of the RDF Schema language. In general, the improvements made by OWL are mainly elements that make it possible to better express the semantic value of ontology, elements that express constraints and restrictions[10]. Among the improvements of OWL we find[10]:

- The notion of restriction (OWL Restriction) which makes it possible to define a constraint.
- The notion of cardinalities (OWL MinCardinality and OWL MaxCardinality) which makes it possible to define the minimum and maximum cardinalities.
- The concept of object properties (OWL ObjectProperty) which defines the relationships between concepts.

2.5 Protégé-2000

Protégé-2000 is an open-source, based frame ontology editor developed at Stanford University's medical informatics department. It is written in Java and supports plug-ins. Protégé can be used to create a hierarchy of classes and instances of those classes [11].

The user interface divided into different tabs which offer different views on the current model: the browser class tab to create and view properties of classes, the instance tab to create and view instances,...etc. New tabs can be added via the plug-in mechanism[11].

Protégé also offers a Java API that can be used from any Java program to access to Protégé model without the Protégé user interface[11].

2.6 Ontology vs Database

Recently, computers dominate the world, and massive quantity of data appeared and stored in computer warehouses. For this reasons, databases and ontologies have been widely applied.

While databases are well known and used as a part of the everyday working routine, ontologies are gaining its popularity gradually. Both of them become integral elements of our live.

2.6.1 Database

Definition : Database is a collection of data organized in such a way as to be easily accessible, administered and updated[12].

2.6.2 Unique Name Assumption (UNA)

The definition of this term asserts that there is only one word available for one entity from the real world[13].

2.6.3 Close World Assumption (CWA)

The CWA is utilized by systems that include complete information, these are mostly database applications [14] [15]. For example, companies of hiring cars database, which enable us to find the availability of specific car(Golf 7). If the database does not include the car name data, a clear result will be returned (0 or NULL), and the interpretation is that no car which its name Golf 7 is available .

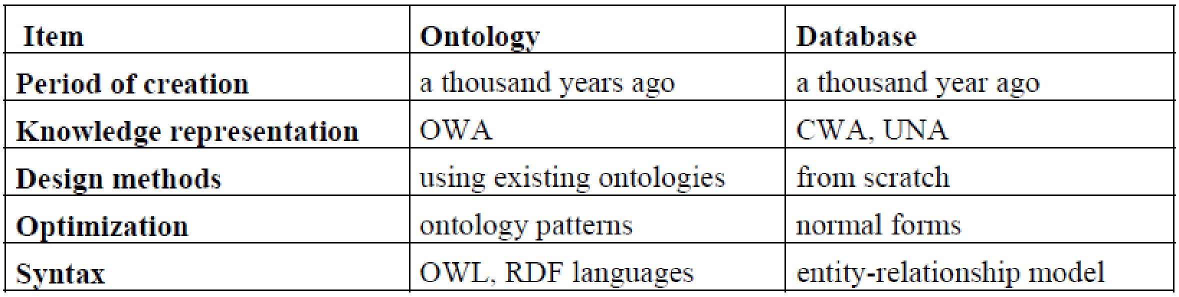
2.6.4 Open World Assumption (OWA)

The OWA is used when the system contains incomplete information. This concept represents concrete knowledge and indicates how new information can be found[14][15]. For example, if a given driver's license does not include information about the talent of the driver, we cannot be sure whether the driver has experience in driving until additional information to confirm or refute the hypothesis found .

Generally, we can say that the CWA returns "0" which the information is missing and the OWA returns "I do not know"[14][15].

2.6.5 Difference between ontology and database

1. The first difference between ontologies and databases concerns the OWA and the CWA. While The ontologies use the OWA system of knowledge representation, the databases used CWA [2]. A database exploits the UNA for naming entities. Any information missing in a database system has the value of "0". Any item of information missing in an ontology system is considered unknown [2].



Item	Ontology	Database
Period of creation	a thousand years ago	a thousand year ago
Knowledge representation	OWA	CWA, UNA
Design methods	using existing ontologies	from scratch
Optimization	ontology patterns	normal forms
Syntax	OWL, RDF languages	entity-relationship model

FIGURE 2.2: Comparison between database and ontology [2].

2. The second difference is the aim for which they are created. While ontologies are focused on adding meaning and comprehension to facilitate the communication between human and machine, databases concentrate on data storage. The databases used for data storage [2].

3. The third difference is the creation methods. A database system is created from zero .To design an ontology system, we try to benefit of existing ontologies or system structuring upon an existing ontology [2]. To create database system, it should delete redundant data from the tables and reduce the complexity by applying normalization of tables. Database system uses set of rules, for transformation of entities and relationships between tables as normal forms [16]. The creation of ontology does not use normal forms. The ontology creation method based on design patterns. These patterns are different than normal forms [16]. These patterns are :

    - Structural pattern
    - Syntactic pattern
    - Content pattern
    - Presentation pattern
    - Consideration pattern
    - Corresponding pattern

2.7 Ontology creation

According to[2] the main phases of the ontology creation are : specification, conceptualization, implementation, and maintenance.

2.7.1 Specification

A specification of an ontology provides a characterisation that is independent of how the ontology is implemented. It makes us know about what the ontology is designed for, rather than how the ontology supports this reasoning [17].

2.7.2 Conceptualization

A conceptualization is an abstract simplified view of some selected part of the world, containing the objects, concepts, and other entities that are presumed of interest for some particular purpose and the relationships between them [18].

2.7.3 Implementation

Implementation is a realization of a technical specification of the ontology through computer programming and deployment[19].

2.7.4 Maintenance

A maintenance is the process of checking, servicing, repairing or replacing of necessary equipment in order to update the ontology .For example : add new terms or new relations . . . [20].

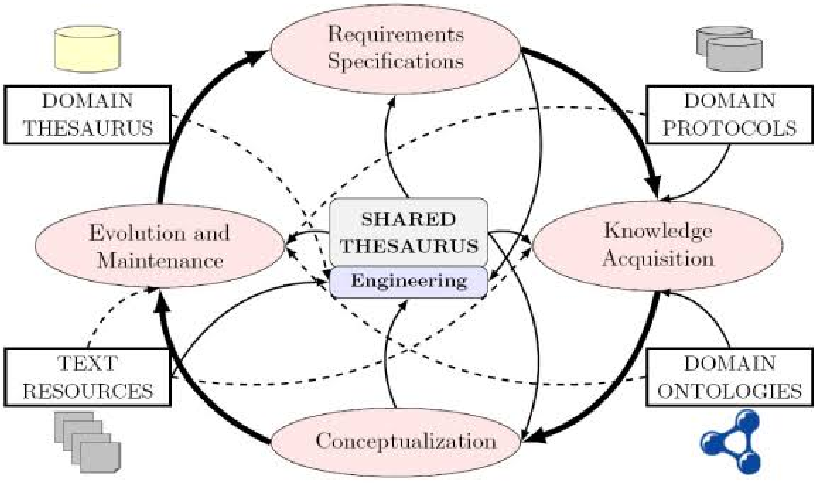


FIGURE 2.3: Life cycle of the ontology [3].

In general, to integrate an ontology you should follow this steps [2] :

1. **Identify the integration possibility :**

   The framework is being applied to build the ontology should allow for some kind of knowledge reuse. In certain cases, integration may involves rebuilding the ontology in different framework where the ontology is available.

   In some situations, this may be cost-effective, but in others it could be more profitable to build from scratch a new ontology that perfectly matches the present needs and purposes than to pursue the rebuilding and adaptation of a pre-existent one.

2. **Identify the modules :**

   The modules (building blocks) needed to build a future ontology are identified, that is, the sub-ontologies in which the future ontology should be divided are defined (in integration, the modules are obviously related to ontologies).

3. **Identify the assumptions and ontological commitments:**

   The presented aspects are described in the conceptual model and in the specification requirements document of the future ontology. This is one of the activities where the actual documentation of the ontology can be crucial to facilitate better, faster, and easier reuse. The assumptions and ontological commitments of the building blocks should be compatible with those found for the resulting ontology.

4. **Identify the knowledge to be represented in each module :**

It is necessary to determine what knowledge should be represented in each building block. At this stage, only an idea is provided of what the modules that will compose the future ontology should be like in order to recognize whether available ontologies are adequate to be reused. A list of essential concepts is identified.

5. **Identify the candidate ontologies :**

   To choose the candidates, all available ontologies are analyzed according to a series of features. At this stage, only a very general analysis is performed.

   Some of the features are: strict requirements (the domain, availability, formalism paradigms in which the ontology is available), the main assumptions and ontological commitments, and the main concepts represented. If an ontology does not have adequate values for these properties, it cannot be considered for integration. The properties are used to eliminate ontologies.

   Other features include desirable requirements or information : where is the ontology available, at what level is the ontology available, what kind of documentation is obtainable (such as technical reports or articles), where is the documentation accessible. If some of the properties and desirable requirements exhibit appropriate values, then the ontology is a better candidate.

6. **Obtain the candidate ontologies :**

   Getting the desired candidate ontologies includes the processes of their representation and also the acquirement of all available documentation. It is preferable to work with the representation knowledge level of the ontology.

   However, in most cases, only the implementation level representation is available. We can then use the reengineering procedure or pursue reconstruction via the available documentation.

7. **Study and analyze the candidate ontologies :**

   To analyze the candidate ontologies, we need to perform two activities:

   - **Technical evaluation of the candidate ontologies:**

     It is important to consider certain features, such as: what knowledge is missing, what knowledge should be removed, what knowledge should be reallocated, what knowledge source changes should be performed, what documentation changes should be performed, what definition changes should be made, and what practice changes should be carried out.

- **User assessment of the candidate ontologies :**

  The overall structure of the ontology and the distinctions upon which the ontology is built to assess whether they are relevant and required, the relation used to structure the knowledge in the ontology to assess whether it is the desired one, the naming convention rules used to assess whether the reuse simplified and promoted, the quality of the definitions, the quality of the documentation of the ontology, and the knowledge pieces represented. All of this points must be checked by the user .

8. **Choosing the source ontologies :**

   From among the candidate ontologies that passed the strict requirements and those which scored best in the oriented integration technical evaluation and user assessment, we have to choose the source ontology that best suits our needs and purpose. The best candidate is the one that can be better or more easily adapted to become the baseline ontology.

   Sometimes more ontologies can be chosen if each one focuses on different elements of the given domain. The choice of source ontologies should be divided into two stages:

   The first phase, in which one tries to find the candidate ontologies best suited for integration (considering general features, development features, and content features).

   The second phase, where it is necessary to tackle the compatibility and completeness of the preliminarily chosen ontologies in relation to the desired resulting ontology.

9. **Apply the integration operations :**

   When the appropriate ontologies reused within one particular integration process are found, the knowledge of these ontologies should be integrated. The related integration operations specify how the knowledge from an integrated ontology will be included and combined with the knowledge in the resulting ontology or modified before its inclusion.

10. **Analyze the resulting ontology :**

    After the knowledge integration, one should evaluate and analyze the resulting ontology. Besides, exhibiting an adequate design and compliance with the evaluation criteria, the ontology should have an overall uniform level. The resulting ontology should be consistent and coherent all over.

2.8 Motivation for the use of the ontology

Ontologies are needed for the prevention and resolution of communication issues between heterogeneous systems, knowledge sharing, and information fusion. They facilitate the information's integration and interoperability between heterogeneous knowledge and information sources while maintaining a high level of abstraction.

Typical reasons for the development and use of ontologies are listed in the following summary[21]:

- To share common understanding of the information structure between people or software.
- To enable reuse of the domain knowledge.
- To make the domain assumptions.
- To separate the domain knowledge from the operational knowledge.
- To analyze the domain knowledge.

2.9 Conclusion

After what we have seen about the ontology in this chapter, we conclude that the ontology is a model for describing the concepts and relationships between them in a hierarchical way, and there is possibility to reuse pre-existing one. For this reasons we are interested to use the ontology in our system as a knowledge source in semantic part of our system.

Chapter 3

Plagiarism

3.1 Introduction

Due to the great extent of development in the technology's world and communication, plagiarism has become a significant challenge. Plagiarism has been found everywhere : on different levels of academic writing (school, institute, university,... etc.), engineering, medicine, music, painting, literature,. . . etc.

In this chapter we will try to give general idea about plagiarism .

3.2 Definition

1. **'Plagiarism'** derives from the Latin word **'plagiarius'**, which mean **'kidnapper'** or **'abductor'**. plagiarism is an act of stealing someone else's work and lying about it afterward[22].

2. According to[23] plagiarize is :

   - "To steal and pass off the ideas or words of another one as own ideas".
   - "To use another's production without citation the source".
   - "To commit literary theft".
   - "To present as new and original an idea or product derived from an existing source".

3.3 Citation

A “citation” is to reference all external ideas and material that you put it in your work . The citation contains[22]:

- Information about the author.
- The title of the work.
- The name and location of the company that published your source copy.
- The date your copy was published.
- The page numbers of the material you are borrowing.

3.3.1 The importance of citing sources

Citation is very important because[22] :

- It is the only way to use other people’s work without plagiarizing.
- It helps everyone wants to know more about the ideas mentioned and their originality.
- It distinguishes between the personal ideas and the ideas of external sources.
- It supports your work and your ideas .
- It shows the research amount you have done.

3.4 Intellectual property

According to [24] the term " Intellectual Property " refers to the mind’s work : inventions, literary and artistic works, designs and models, and logos, names and images used in the trade. It has two branches :

- literary and artistic property.
- industrial property.

3.4.1 Literary and artistic property

Literary and artistic property applies the mind's work, is composed of copyright and neighboring rights [24].

3.4.2 Industrial property

Industrial property includes, on the one hand, the utilitarian creations, such as the patent of invention. On the other hand, the distinctive signs in particular trademark, domain name and the designation of origin [24].

3.5 Copyright laws

Copyright laws exist to protect our intellectual property. They make it illegal to reproduce someone else's expression or ideas or information without permission. This can include music, images, written words, videos, and a variety of other media.

At one time, a work was only protected by copyright if it included a copyright trademark (the © symbol). Anyone who reproduces copyrighted material improperly can be prosecuted in a court of law[25].

3.6 The different forms of plagiarism

There is two main forms of plagiarism [22] :

1. Sources not cited

   - In case when the writer replaces words of another's work with other words have the same meaning like synonyms and declares that this work come back to him.
   - In case when the writer copies part or full text, without changes, from another's writer work.
   - In case when the writer copies part of text from several different sources and makes it homogeneous.
   - In case when the writer reuses his or her previous work.

2. Sources cited
    - In case when the writer mentions an author's name for a source, but did not put one of the specific information concern the source, this is what makes difficulties to find the source.
    - In case when the writer did not use the quotation marks in parts when he or she copies word-for-word.
    - In case when their is no original work, which means that this work just like documentation.
    - In case when the writer uses quotation marks and cites sources in some places, but paraphrases other parts from those sources without citation.

3.7 What make the student plagiarize

According to [26] a lot of influences and pressures lead the student to plagiarise such :

- Poor time management skills due to the increasing competition for student's time arising from work and study and care for his children or his little brothers .
- The perception that the academic responsible has little enthusiasm for this subject this is what makes the student less motivated for the work.
- The external pressure to succeed from parents or peers, or for financial reasons.
- The curiosity of testing the system and taking it on.
- Cultural distinctions in educational and display skills, because in some countries the use of another's work without citation considered as legal practice.

According to [27] plagiarism "saves time and effort, improves results, and shows considerable initiative on the part of the plagiarist" . The student think that the plagiarism does not aim to steal another's property, it just about the spread of information and knowledge.

3.7.1 Conclusion

In this chapter we introduce the plagiarism which is one of the points that we will deal with in our thesis especially in Arabic language. At the end, we have to say that the plagiarism is to steal someone else's work and say that this work refer to you. It has multiple kinds what make difficulties to detect it and it makes a real challenge for the researchers.

In the next chapter we will present the plagiarism in the Arabic language.

Chapter 4

Plagiarism and Arabic Language

4.1 Introduction

The great revolution of data streams and the numerization facilitated for people the search for information in different fields of knowledge. As a consequence, there are various types of plagiarism issues as we mentioned in the previous chapter.

Arabic language also suffer from the phenomenon of plagiarism, this is what create a new challenge for researchers to detect the plagiarism in Arabic document .

In this chapter, we have focused on the Arabic language and the plagiarism detection techniques in Arabic documents.

4.2 Arabic language

The word " **Plagiarism** " in Arabic language means "إنتحَال", it is also considered as unethical practice because it includes arrogation of authorship [28] .

4.2.1 Characteristics of Arabic language

The Arabic language is the fourth most widely spoken language in the world. It has lot of specificities that make it so different compared to other languages [29], some of this characteristics are listed below [29] :

- Arabic language has 28 characters [... , ث , ت , ب , أ], Three of them are vowels, [ى , و , ا] and others are consonants.
- The position of the character in the word might changes its form. For example : [ي , Y] in [Write , يكتب , yaktob].
- The most different charactiristics of Arabic language is : written from right to the left and it does not have capitalization. There are two types of writing : (رُقْعَة Rogaa and نَسْخ Naskh) .
- Arabic documents are read and understood clearly by adding some diacritics above or under each character in word, [دّ , دْ , دِ , دُ , دَ], for example [قَرَعَ Karaa , knock], while [Kara قَرَعْ , Zucchini].
- The root of every word in Arabic has just three characters, and new words is formed by adding some suffixes [name, verb, number,. . . etc]. For example: [Wrote, Kataba كَتَبَ], [مَكْتَبْ maktb, office]. Person and verb have three forms (singular, dual, and plural).

4.3 Plagiarism techniques and related works

The field of plagiarism detection (PD) contains many works that have been proposed to help searchers to avoid the use of works or ideas without permission. The highest percentage of these works describes the techniques and tools of the English language.

However, there is a limited researches done to address documents written in Arabic and especially with the exploitation of ontology as semantic resource to detect the plagiarism .
According to [30] the most techniques and works used in PD are :

- Latent Semantic Analysis (LSA) : is a technique used to describe relationships between a set of documents and terms they contain. In this technique, words that are close in meaning are assumed to occur close together [31].
  A matrix is constructed in which rows represent words, and columns represent documents. Every document contains only a subset of all words [31].
- Singular Value Decomposition (SVD) : is " a factorization method of real or complex matrix " [32], is used to reduce the number of columns while preserving the similarity structure

among rows.

This decomposition is time consuming because of the sparseness of the matrix. Words are compared by taking the cosine of the angle between the two vectors formed by any two rows. Values close to 1 represent very similar words, while values close to 0 represent very dissimilar words[32] .

- Stanford Copy Analysis Mechanism (SCAM) : is based on a registration copy detection scheme. Documents are registered in a repository and then compared with the pre-registered documents[33].

  The architecture of the copy detection server consists of a repository and a chunker. The chunking of a document breaks up a document into sentences, words or overlapping sentences [33].

- String tiling, finding the joint coverage for a pair of files and parse tree comparison [34][35]. Usually these techniques work in pairs of files, so the comparison routine should be called for each possible file pair found in the input collection [36][37] .

- Fuzzy-set IR model used by Salha Mohammed Alzahrani and Naomie Salim to calculate the degree of similarity and compared it with a threshold value to judge whether two statements are the same or different [38][39].

- A fingerprint is a set of integers created by hashing subsets of a document to represent its key content. Techniques to generate fingerprints are mainly based on k-grams (a k-gram is a contiguous substring of length k) which serve as a basis for most fingerprint methods [29].

- The winnowing algorithm is an algorithm to select document fingerprints from hashes of k-grams. To obtain the fingerprint of a document, the text is divided into k-grams, the hash value of each k-gram is calculated, and a subset of these values is selected to be the fingerprint of the document [30].

- Randa K. [34] has developed APD Tool stand-alone desktop tool base on Winnowing local document Fingerprinting Algorithm. it has been adaptive for Arabic and tested using three essays written by a class of student. She has concluded that ADP is an efficient solution to minimize student coping.

- Mohamed .El Bachir in 2012 implemented a prototype of APlag in Java it is based on a new comparison algorithm that uses heuristics to compare suspect documents at different

hierarchical levels to avoid unnecessary comparisons [38].
He evaluated its performance in terms of precision and recall on a large data set of Arabic documents, and showed its capability in identifying direct and sophisticated copying, such as sentence reordering and synonym substitution [36].
He presented and discussed a series of experiments to demonstrate its effectiveness on a large set of Arabic documents. The results indicate that APlag has the capability to detect precisely exact copy, change in sentence structure, and synonym replacement [30].

- " Bing " a search engine, they developed a system to detect plagiarism in both Arabic and English languages using "Bing" search engine. The system which relies on plagiarism detection algorithm is effective and can support both Arabic and English languages [40].
  This algorithm reduces the unuseful comparison between texts, since it compares only between cue-phrases surrounding words which forms the logical and natural boundaries of text sentences [40].

- Alzahrani And AL have produced an Arabic plagiarized detection (APD) tool especially for working with Arabic language [41].
  APD (Arabic Plagiarism Detection) tool uses the Internet to help professors and teachers in e-learning systems identify stolen intellectual property by utilizing Google API to find similar documents on the web [42].
  The typical workflow in APD paradigm has two major steps:

  - The first step, students submit their assignments in Arabic to the system, which in turn will be stored into reports database [42].

  - The second step, the teacher triggers APD tool via a user interface to check the assignments for plagiarism [42].

  Then, the tool will compare the documents against the intra corpus collection, which probably contains the previous assignments [42].
  Moreover, APD tool searches the web to give similar resources as well. An automatic report will be generated that contains highlighted plagiarized parts and a list of similar resources ranked from highest to lowest [39].

- The work presented in [43], aims to identify the plagiarism in Arabic based text documents using Boolean ranked queries.

The proposed solution is based on a search engine structure in order to reduce the cost of pairwise similarity, which is called Iqtebas 1.0. It is a primary solid and complete piece of work for plagiarism detection in Arabic documents [43].
For the indexing process, they use the winnowing n-gram fingerprinting algorithm to compute fingerprints for each sentence and to reduce the index size. Then, the search engine is supposed to return the n most similar sentences to the query sentence, which works on the fingerprints level [43].

4.4 Conclusion

In this chapter we introduce some related works concerning the detection of plagiarism in Arabic document. We notice that the Arabic language is very difficult what makes its treatment very hard and needs a lot of experience.

Based on what we have studied on the related works of plagiarism detection in Arabic documents, we have found that these studies are influenced by various factors, for example, the quality of the obtained results depends on the quality of the corpus used and the treatment techniques deployed.

Moreover, we found that the most works are not compared with each other using the same data set, which raises the question of how to determine the degree of good results obtained in relation to other works.

In addition, the experiments established were carried out using a small corpus, it therefore necessary to construct large corpus with normalized statistics in relation to queries that can be used to confirm the utility of the techniques or methods developed to improve the obtained results compared to other studies.

Chapter 5

Conception of El Momayaz System

5.1 Introduction

The field of plagiarism detection contains many works that have been proposed to help searchers to avoid the use of works or ideas without permission. The highest percentage of these works describes the techniques and tools of the English language.

However, there is a limited researches done to address documents written in Arabic and especially with the exploitation of ontology as semantic source to detect the plagiarism .For this reason, we are interested to develop system of detecting plagiarism in Arabic documents based on ontology .

In this chapter we will describe our system, ranging from a description that explains in a general way the operation and the principle of the system (general design) to a description that explains each part of the system in detail (detailed design).

5.2 Purpose of the system

The objective of our work is to calculate the degree of plagiarism in Arabic documents using Arabic WordNet as semantic source. The work consists of cleaning the document from unuseful words, then indexing the document twice in order to facilitate the counting of similarity between the documents and reduce the size of the document.

5.3 Overall design

In this section, we will introduce the principle and general architecture of our system.

5.3.1 General principle of the system

The figure 5.1 presents our principle of the system .

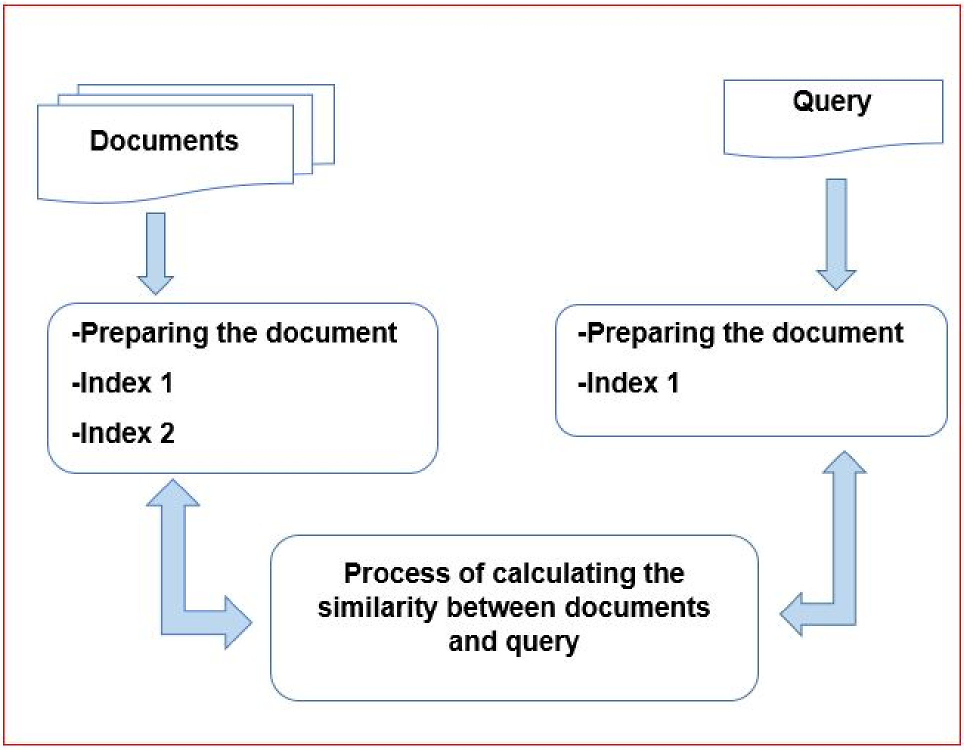


FIGURE 5.1: General principle of the system.

5.3.2 General architecture of the system

In our system, we aim to detect the plagiarism in Arabic documents, we have used three layers (see figure 5.2) :

- The first layer (NLP Layer), it prepares the query and the documents that was extracted from STS by applying some process that simplify the query and data set to make it ready for the next layer .

- The second layer (Indexing Layer), in this layer we index our data (documents and the query)that has been treated in the NLP layer, and store it in the index table (Index Table Of Documents, Index Table Of Query)
- The third layer (Semantic Similarity Layer), in this layer we obtain the similar document and calculate the degree of similarity between this document and the query, then send this result to the user .

In the next section we will discuss the details of each layer .

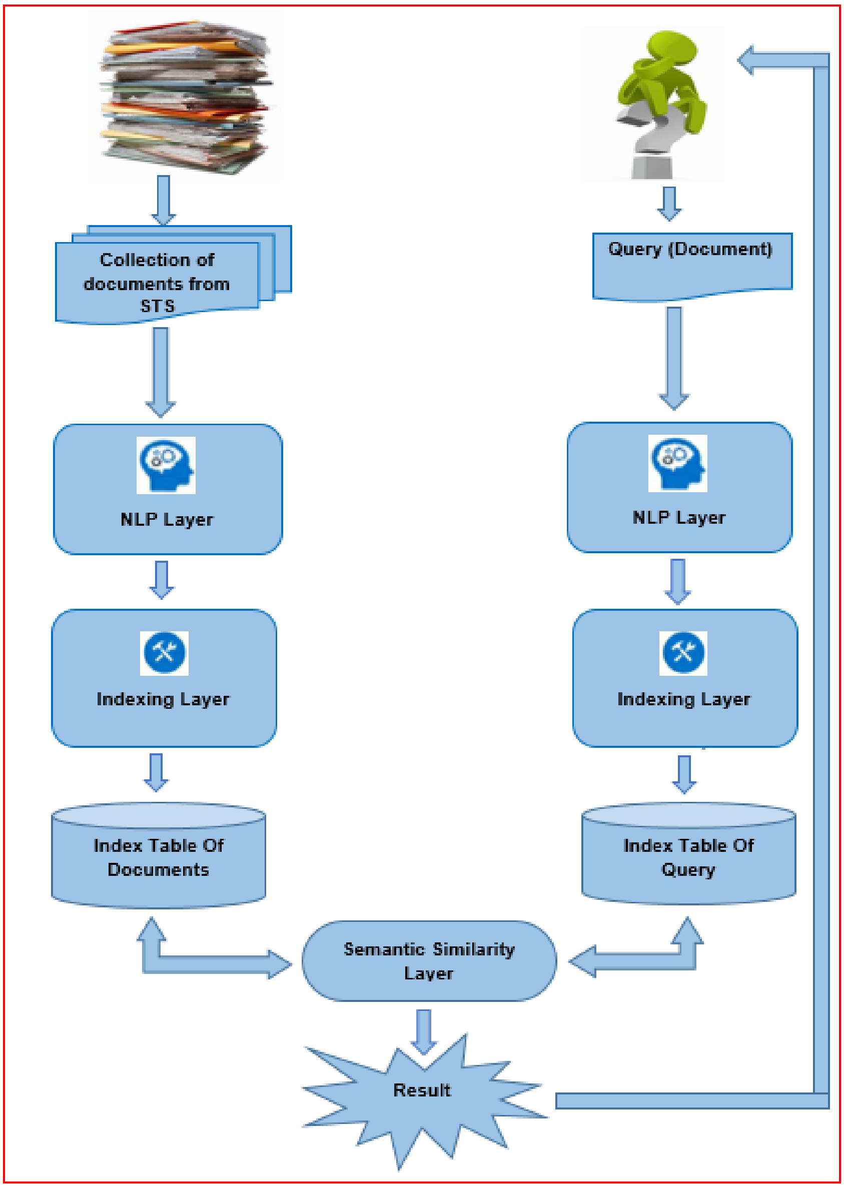


FIGURE 5.2: General architecture of the system.

5.4 Detailed design

As it is shown in the title, this part will be devoted to detail the different layer and the different processes of our system.

Figure 5.3 shows the detailed architecture of our system.

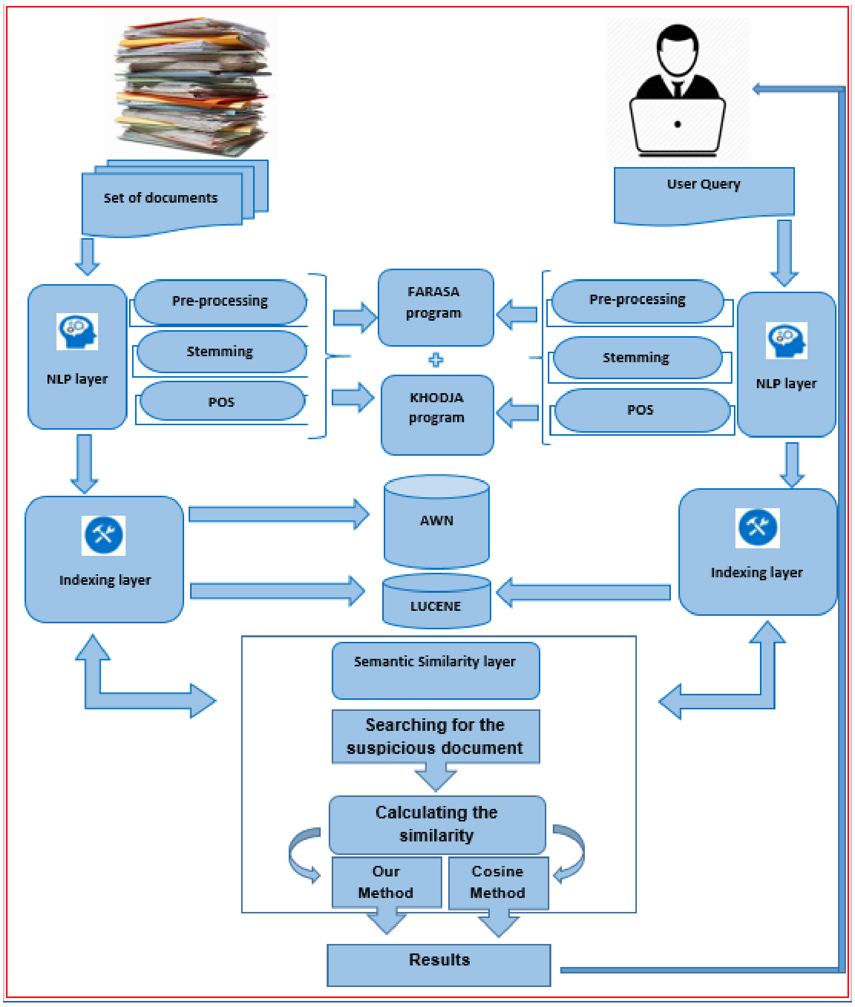


FIGURE 5.3: Detailed design.

5.4.1 NLP layer

Natural language processing (NLP) : it is a theoretically motivated range of computational techniques for analyzing and representing natural texts at one or more levels of linguistic analysis [44].

We have used three processes in this layer :

1. Pre-processing : this module based on morphological analysis which divides the document into (words, entities, or attributes) by :

   - The lexical analysis (Tokenization): it is the process of converting the text of a document into a set of terms. A term is a group of characters that constituts a significant word. The lexical analysis makes it possible to recognize the spaces of words' separation, numbers, punctuations,. . . etc [45].
   - Stop-Word Removal : Stop words are too frequent words that are not very significant, which makes the search slower and increases the size of the document. These words do not make sense and they are not useful, so they must be eliminated.

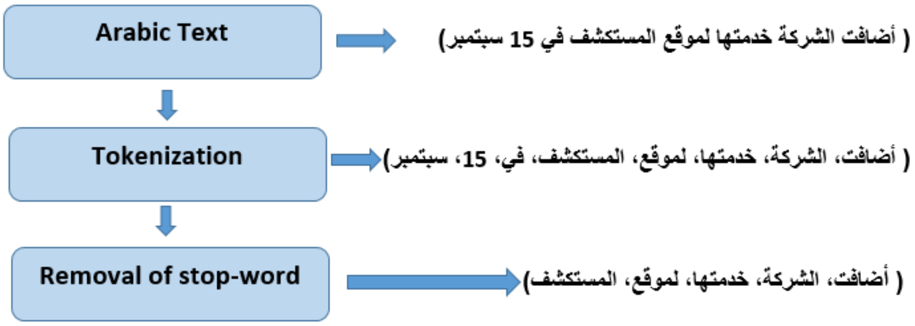

Arabic Text
(أضافت الشركة خدمتها لموقع المستكشف في 15 سبتمبر)
Tokenization
(أضافت، الشركة، خدمتها، لموقع، المستكشف، في، 15، سبتمبر)
Removal of stop-word
(أضافت، الشركة، خدمتها، لموقع، المستكشف)

FIGURE 5.4: Example of NLP.

2. Stemming : Stemming is the process of removing any affixes from words, and reducing these words to their roots [46].

   The stemming algorithms are used in many NLP applications .The purpose of Arabic stemming is to extract the stems or roots of the different Arabic words [47].

   As many stemming algorithms have been built for the Arabic language, we used Khoja stemmer [47] because it is one of the best stemmer tools and the most used in Arabic language. The Figure 5.5 shows us an example of stemming .

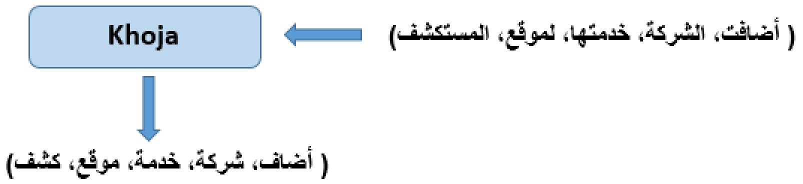

Khoja
(أضافت، الشركة، خدمتها، لموقع، المستكشف)
(أضاف، شركة، خدمة، موقع، كشف)

Figure 5.5: Example of stemming.

3. Part-Of-Speech (POS) : according to [48] POS is a category of words (noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection,. . .) .
   Farasa is an Arabic word processing toolbox available online on the internet. It consists of the segmentation module (tokenization), POS tagger, Arabic text Diacritizer, Dependency Parser,. . . etc. In our work, it is useful to use Farasa to extract the POS of each word in order to use it in the synonyms search with AWN.

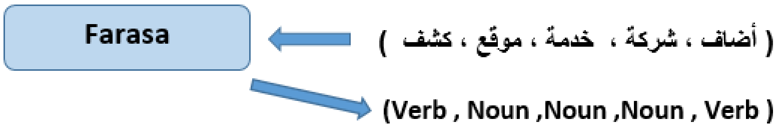

Farasa
(أضاف ، شركة ، خدمة ، موقع ، كشف)
(Verb , Noun ,Noun ,Noun , Verb)

Figure 5.6: Example of POS.

5.4.2 Indexing layer

The indexation of the text is an important step for each NLP task in order to obtain the best possible results [49]. For each document, each sentence is prepared by applying a set of NLP operations. In our system we have used the tow type of index for data set and we have just applied the second type for the query.

1. The first index(Using Arabic WordNet): In our system we have used AWN to extract the synonyms of each word using the associated POS tag, from the previous step.
   Arabic WordNet is an Arabic semantic knowledge source based on the structure and the content of the Princeton WordNet (PWN). It is a lexical resource for Modern Standard

Arabic (MSA), which is widely used by internet users in the Arabic world.
It was created in 2006, and extended in 2015 to use wordnet in multiple languages. The current version contains 9,916 Synsets, 17,785 words and 37,335 Senses. The Synsets are a set of Arabic words with their synonyms and semantic relations [50].

2. The second index (Using Lucene) : The next step is to index the documents for the second time. Recent experience have shown that the most effective method for indexing is the selection of the document frequency (DF)[51]. For this, we have used a characteristic selection technique by using Lucene program, is an open source library written in Java based on TF/IDF[51]. It is a project of the Apache Foundation made available under Apache license[51].

   The main use of Lucene is for indexing and searching for text [51]. Our objective is to analyze and reduce the time of the indexing sentence, which is an essential step that must be done quickly for data extraction. This proposal can be performed by using Lucene.
   It is a technology suitable for the development of our application that requires full text search, especially for the plagiarism detection in Arabic documents. It offers incremental indexing as fast as batch indexing and reduces the size of the indexed text [51].

   - TF (Term Frequency): This measure calculates the local importance of the term in a document (calculate the number of term occurrences in the document) [52].
   - IDF (Inverse of Document Frequency): This measure calculates the importance of a term throughout the collection. The overall weighting of a term is expressed in terms of the total number of documents in the corpus and the number of documents containing that term [52].
     This method is usually expressed as follows: log (N/DF), where DF is the number of documents containing the term and N is the total number of documents in the collection [52].

   The TF*IDF measure gives a good approximation of the importance of the term in the document, particularly in documents of uniform size [52] .

We represent the document in our system with the stem of each word plus its synonyms in order to enrich our document and get the different meaning of the word .

5.4.3 Semantic similarity layer

In this layer we have tow processes :

1. Searching for the suspicious documents : in this process we have used Lucene Search option that allows us to extract the similar documents.

2. Calculating the similarity : The similarity calculation process consists of comparing the representation of the query (document) with the representations of the documents.
   It calculates for each query pair documents, a measure called resemblance score that reflects the degree of similarity between the query and the document considered. This score is calculated from a similarity function.
   After getting the similar document we calculate the degree of similarity between this document and the query using tow methods:

   (a) Cosine method : The objective of this method is to represent the documents as vectors, then calculate similarity between two documents by computing the Cosine between their vectors as follows:
   $Sim = \cos(\theta) = \frac{A.B}{\|A\|\|B\|}$

   which mean : $$Sim = \cos(\theta) = \frac{\sum_{i=1}^{k} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{k} B_i^2}}$$
   where A and B are vectors presented as follow :
   $A = A_1, \ldots, A_i, \ldots, A_n$
   $B = B_1, \ldots, B_i, \ldots, B_k$

   (b) Our method is based on representing the documents and the query in index-table, then get the percentage of query's words represented in the index-table found in each document's index .
   The formula of our method is :
   $Sim = \frac{\sum_{j=1}^{k} V_j}{K}$
   where A and B are vectors represented as follow :
   $A = A_1, \ldots, A_i, \ldots, A_n$
   $B = B_1, \ldots, B_j, \ldots, B_k$

$$V_j = 1 \ if \ B_j \in A$$

$$V_j = 0 \ if \ B_j \notin A$$

5.5 Conclusion

In this chapter we have focused on the architecture of our system and the different layers of our system and we have also presented our method and cosine method. In the next chapter we will evaluate our method and we will show the results .

Chapter 6

Implementation

6.1 Introduction

In order to realise our project we have to use some tools and environments that can help us. So in this chapter we will introduce some definitions of this tools and environments, then we will show the interfaces of our project and the obtained results.

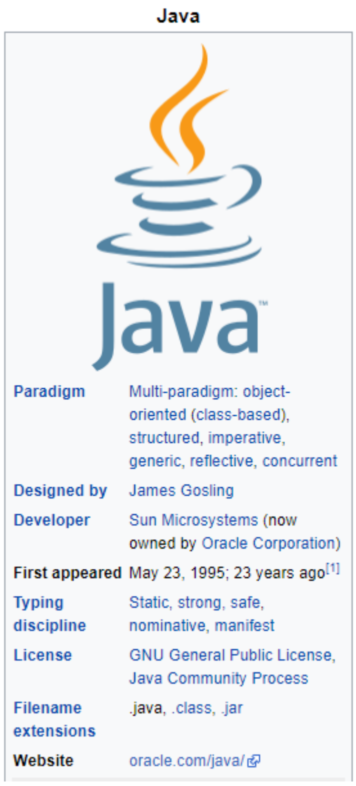
6.2 Environments, Tools, and Apis used

6.2.1 JAVA

Java is a widely used programming language expressly designed for use in the distributed environment of the internet. It is the most popular programming language for Android smartphone applications and is among the most favored for edge device and internet of things development [4].

Java was designed to have the look and feel of the C++ language, but it is simpler to use than C++ and enforces an oriented object programming model. Java can be used to create complete applications that may run on a single computer or be distributed among servers and clients in a network [4].

It can also be used to build a small application module or applet for use it as part of a webpage. The aim goal of java is to let application developers "write once, run anywhere" [4].



Java

Paradigm	Multi-paradigm: object-oriented (class-based), structured, imperative, generic, reflective, concurrent
Designed by	James Gosling
Developer	Sun Microsystems (now owned by Oracle Corporation)
First appeared	May 23, 1995; 23 years ago[1]
Typing discipline	Static, strong, safe, nominative, manifest
License	GNU General Public License, Java Community Process
Filename extensions	.java, .class, .jar
Website	oracle.com/java/

Figure 6.1: General information about Java [4].

6.2.1.1 Java platforms

There are three key platforms upon which programmers develop Java applications [53]:

1. Java SE Simple, stand alone applications are developed using Java Standard Edition. Formerly known as J2SE, Java SE provides all of the APIs needed to develop traditional desktop applications [53] .

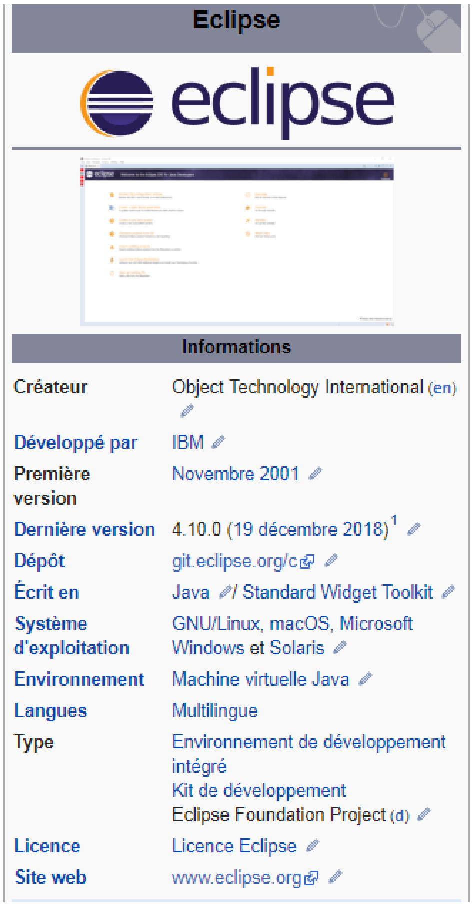
2. Java EE. The Java Enterprise Edition, formerly known as J2EE, provides the ability to create server side components that can respond to a web based request response cycle. This arrangement allows the creation of Java programs that can interact with based internet clients, including web browsers, based CORBA clients and even REST and SOAP based web services [53] .

3. Java ME. Java also provides a lightweight platform for mobile development known as Java Micro Edition, formerly known as J2ME. Java ME has proved a very popular platform for embedded device development, but it struggled to gain traction in the smartphone development arena. In terms of smartphone development, Android has become the mobile development platform of choice [53] .

6.2.2 Eclipse

In our project we use Eclipse Luna which is version of eclipse and it is the latest Interred Development Environment(IDE). Also, Eclipse support many languages such as : XML , HTML, C, C++, JavaScript, Ruby and PHP [5].

Eclipse has lot of characteristics like : (color editor, multi-language projects ,refactoring ,interface graphical editor and web page). Eclipse is available for Windows, Linux, Solaris (in x86 and SPARC), Mac OS X or under an independent operating system version (requesting a Java virtual machine). An environment Java Development Kit (JDK) is required for Java developments [5].



Eclipse

Informations

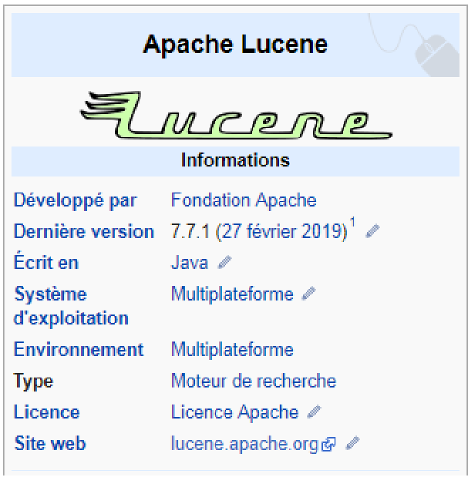
Créateur	Object Technology International (en)
Développé par	IBM
Première version	Novembre 2001
Dernière version	4.10.0 (19 décembre 2018)[1]
Dépôt	git.eclipse.org/c
Écrit en	Java / Standard Widget Toolkit
Système d'exploitation	GNU/Linux, macOS, Microsoft Windows et Solaris
Environnement	Machine virtuelle Java
Langues	Multilingue
Type	Environnement de développement intégré Kit de développement Eclipse Foundation Project (d)
Licence	Licence Eclipse
Site web	www.eclipse.org

FIGURE 6.2: General information about Eclipse [5].

6.2.3 Lucene

Lucene is an open source library written in Java. It is a project of the Apache Foundation made available under Apache license. It supports Ruby, Perl, C ++, PHP, C # and Python languages [6].

It is used in some search engines. The main use of lucene is for indexing and searching for text [6].



Apache Lucene

Lucene

Informations	
Développé par	Fondation Apache
Dernière version	7.7.1 (27 février 2019)[1]
Écrit en	Java
Système d'exploitation	Multiplateforme
Environnement	Multiplateforme
Type	Moteur de recherche
Licence	Licence Apache
Site web	lucene.apache.org

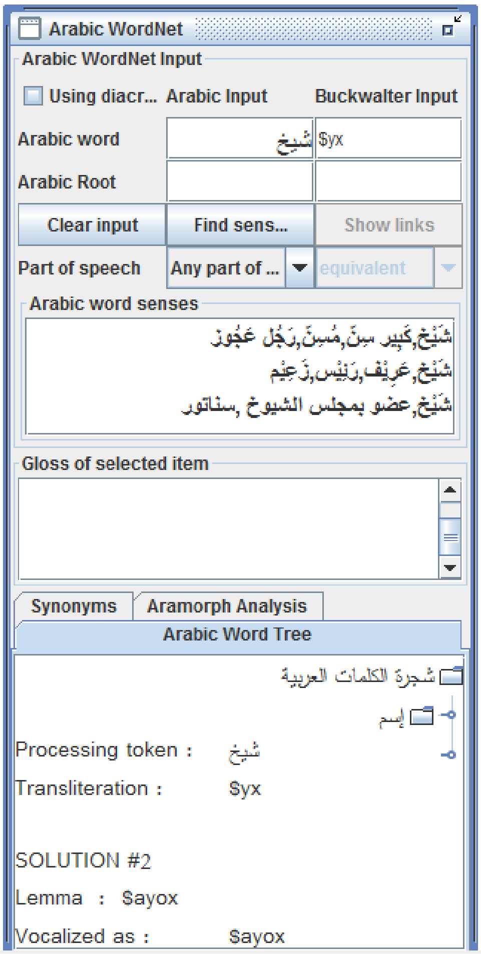
FIGURE 6.3: General information about Lucene [6].

6.2.4 ArabicWordNet

Arabic WordNet (AWN) is Arabic semantic knowledge source based on the structure and content of the Princeton WordNet (PWN) and mapped directly onto PWN 2.0 and EuroWordNet (EWN). Most of the synsets of AWN should be linked to English WN, and the structure of AWN hierarchy followed the same WN topology [54].

AWN used in many Arabic Natural Language Processing (ANLP) and Arabic Information Retrieval applications to find common characteristics between concepts. AWN 2.0 was released in January of 2008 [54].

The database structure of AWN contains four entity types: item, word, form and link. Items are the synsets, each item has unique identifier and brief description called gloss. A word entity is a word sense. A form entity contains lexical information. A link represents the relation between synsets, examples of relation type are : related_to, has_hyponym, verb_group, has_holo_member and has_derived[54].



Arabic word: شيخ $yx

شَيْخ,كَبِير سِنّ,مُسِنّ,رَجُل عَجُوز
شَيْخ,عَرِيف,رَئِيس,زَعِيم
شَيْخ,عضو بمجلس الشيوخ ,سناتور

شجرة الكلمات العربية
إسم

Processing token : شيخ
Transliteration : $yx

SOLUTION #2
Lemma : $ayox
Vocalized as : $ayox

FIGURE 6.4: The AWN interface.

6.3 The interfaces of our system

In this section, we present some windows of our application to illustrate how our system works and describe each part of these interfaces.

6.3.1 The first interface

The first interface in our system is presented below :

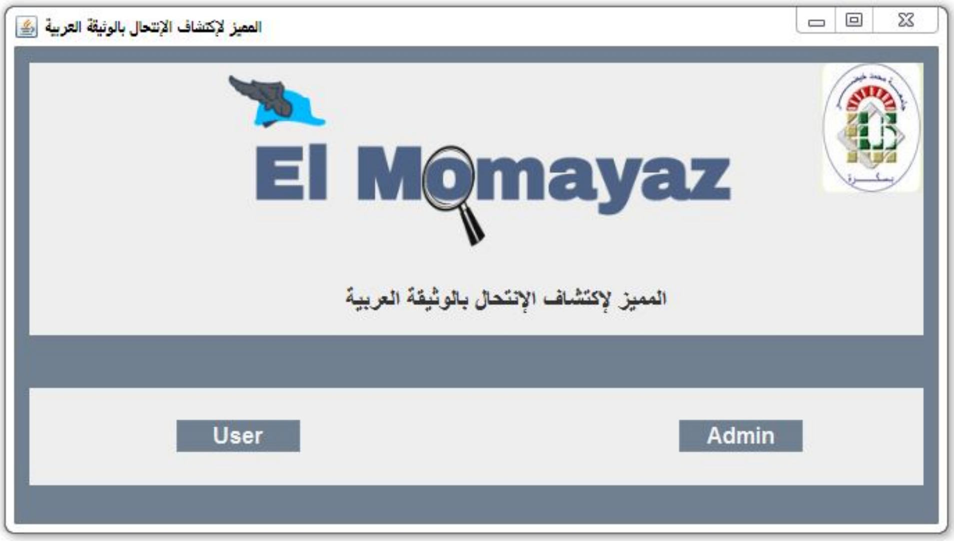

المميز لإكتشاف الإنتحال بالوثيقة العربية
El Momayaz
المميز لإكتشاف الإنتحال بالوثيقة العربية
User
Admin

FIGURE 6.5: The first interface.

1. The first thing that attract us in this interface is the word "El Momayaz" as it is shown below :

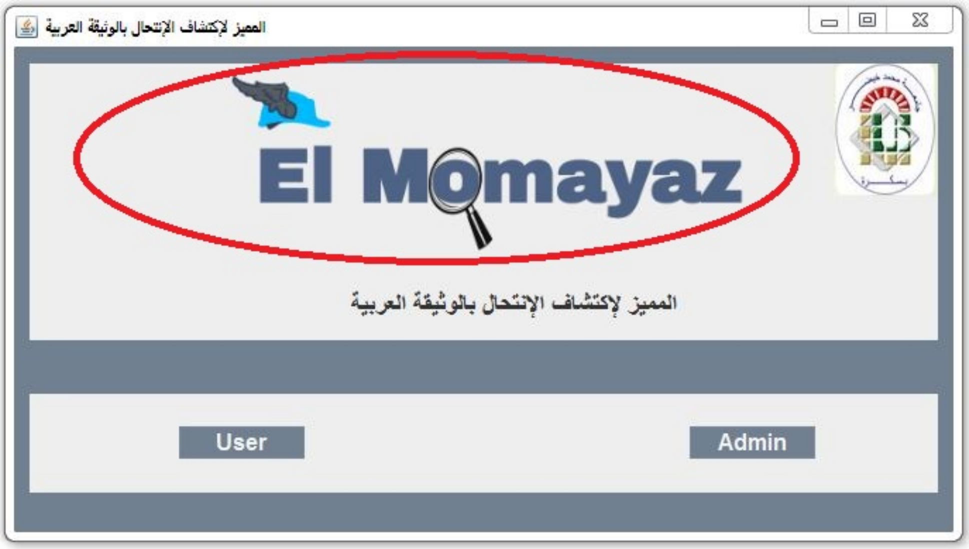

المميز لإكتشاف الإنتحال بالوثيقة العربية
El Momayaz
المميز لإكتشاف الإنتحال بالوثيقة العربية
User
Admin

FIGURE 6.6: The first interface (1).

We choose this word as a logo for our system to specify our system and give it a value.

2. The second thing that we notice in this interface is the title " المميز لإكتشاف الإنتحال بالوثيقة العربية ". We consider " المميز لإكتشاف الإنتحال بالوثيقة العربية" as title of our system and we choose the Arabic to write this title to make the viewer to our system understand that this

system is related to the Arabic language.

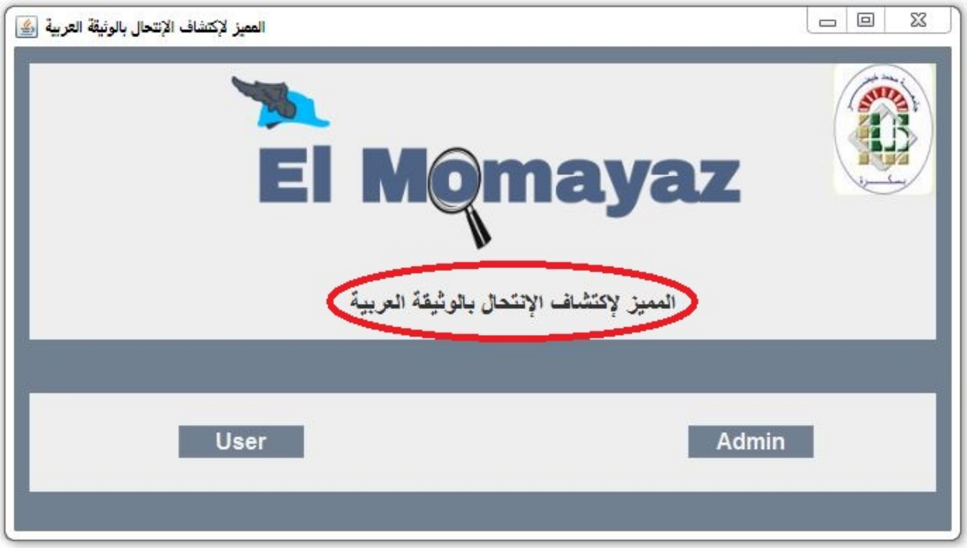

المميز لإكتشاف الإنتحال بالوثيقة العربية
El Momayaz
المميز لإكتشاف الإنتحال بالوثيقة العربية
User
Admin

FIGURE 6.7: The first interface(2).

3. We also mentioned the symbol of our university " جامعة محمد خيضر " to refer that this system are realised in this university.

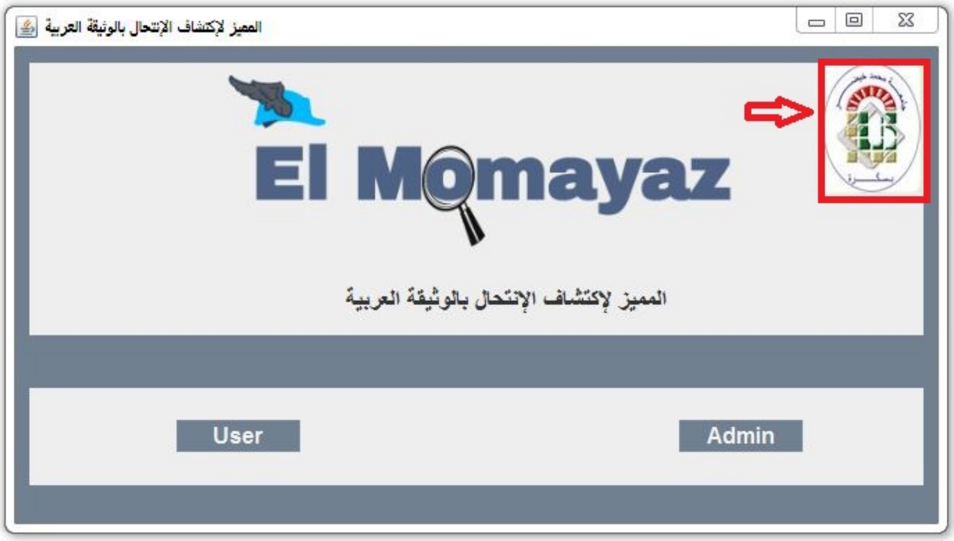

المميز لإكتشاف الإنتحال بالوثيقة العربية
El Momayaz
المميز لإكتشاف الإنتحال بالوثيقة العربية
User
Admin

FIGURE 6.8: The first interface(3).

4. The first button in this first interface is "User", by simple click on this button we will go to the "User area" (Figure 6.50) that we will talk about it later.

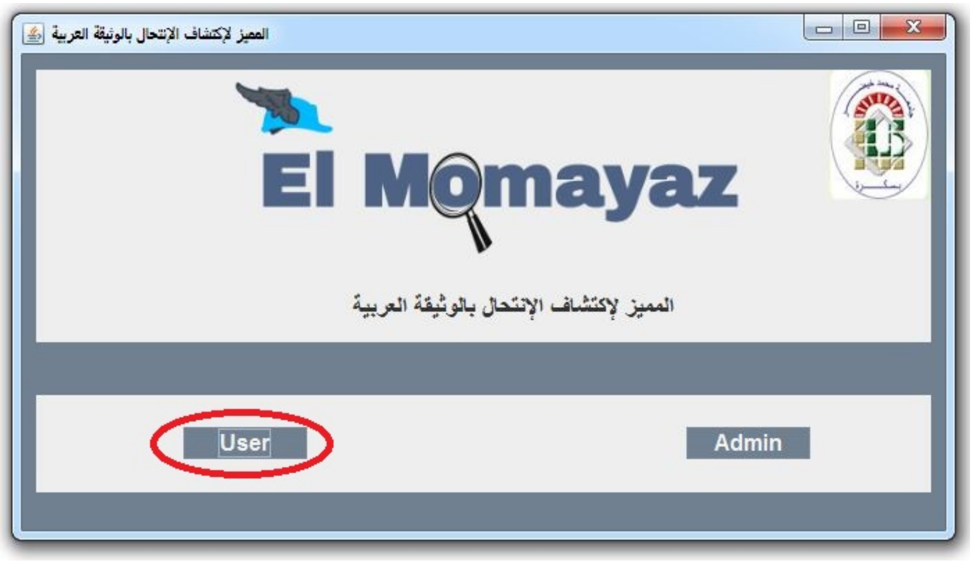

المميز لإكتشاف الإنتحال بالوثيقة العربية
El Momayaz
المميز لإكتشاف الإنتحال بالوثيقة العربية
User
Admin

FIGURE 6.9: The first interface(4).

5. The other button in this interface is "Admin", when we click on this button new interface will appear which is the "Login interface " (Figure 6.11).

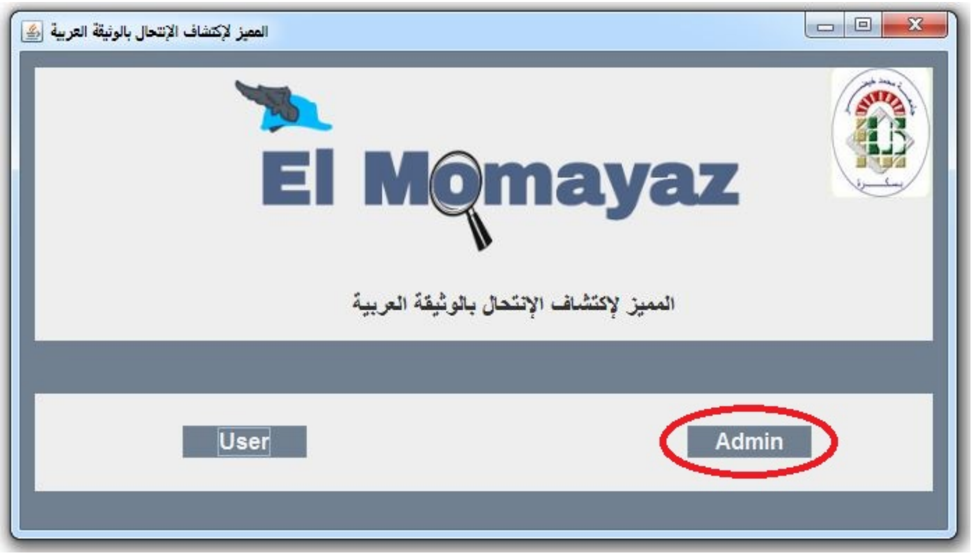
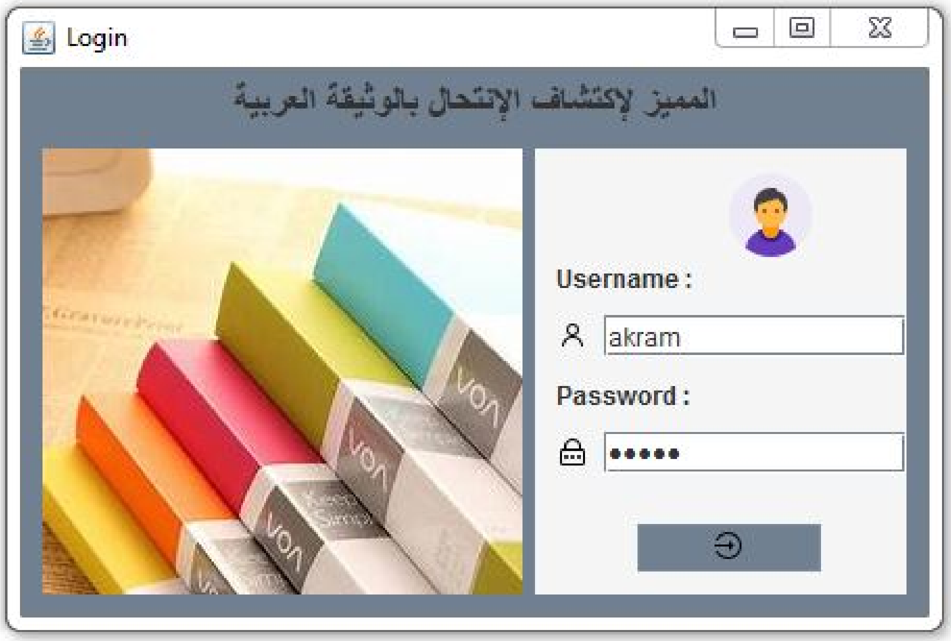

المميز لإكتشاف الإنتحال بالوثيقة العربية
El Momayaz
المميز لإكتشاف الإنتحال بالوثيقة العربية
User
Admin

FIGURE 6.10: The first interface(5).

6.3.2 The Login interface

After clicking on "Admin" button in the first interface, the interface (Figure 6.11) will appear.



المميز لإكتشاف الإنتحال بالوثيقة العربية

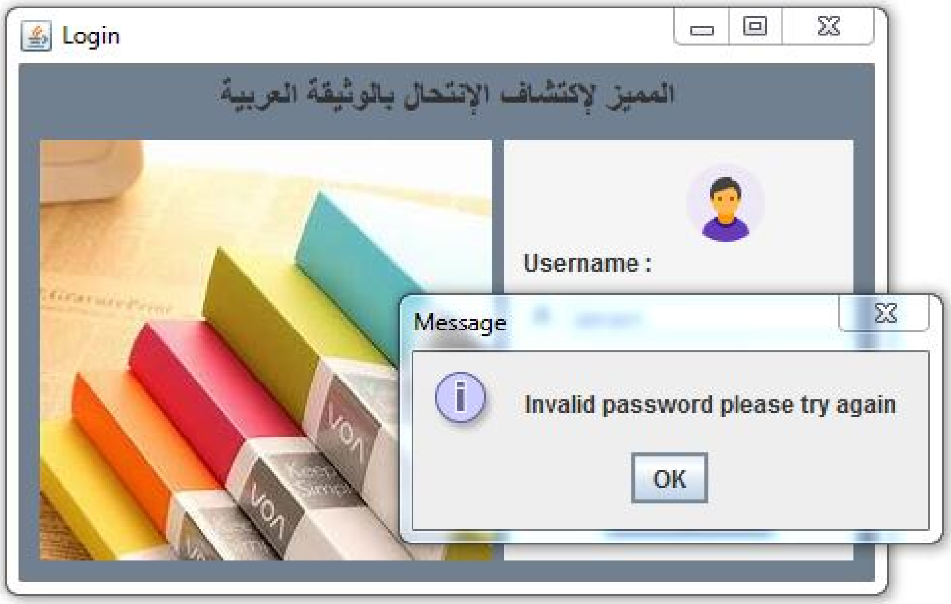
Username : akram

Password : •••••

FIGURE 6.11: The Login interface.

– When we enter the correct username and the password the "Admin interface" will display (Figure 6.15).

– If we enter an incorrect username or incorrect password an error message will display (Figure 6.12, Figure 6.13).

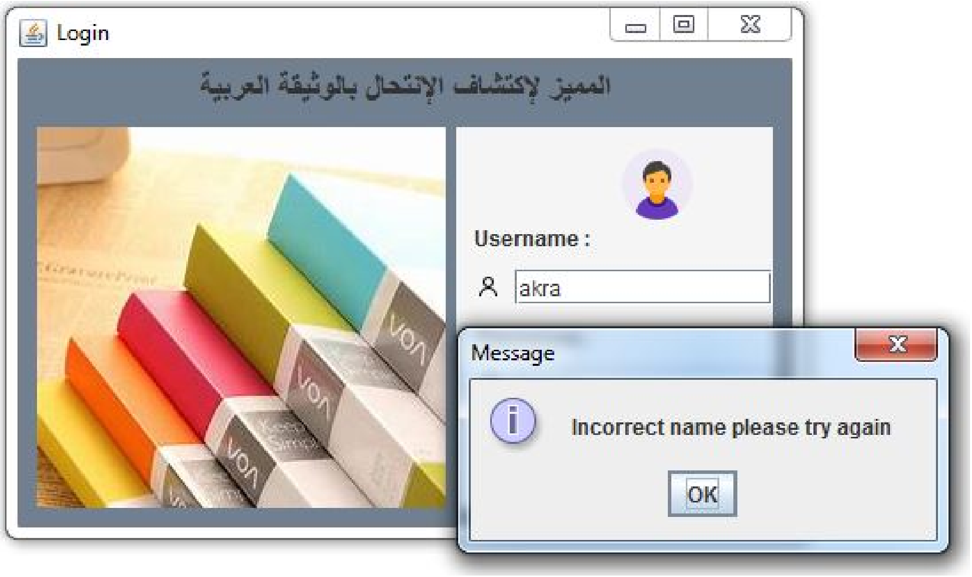


المميز لإكتشاف الإنتحال بالوثيقة العربية

Username :

Invalid password please try again

FIGURE 6.12: The Login interface(1).



Login

المميز لإكتشاف الإنتحال بالوثيقة العربية

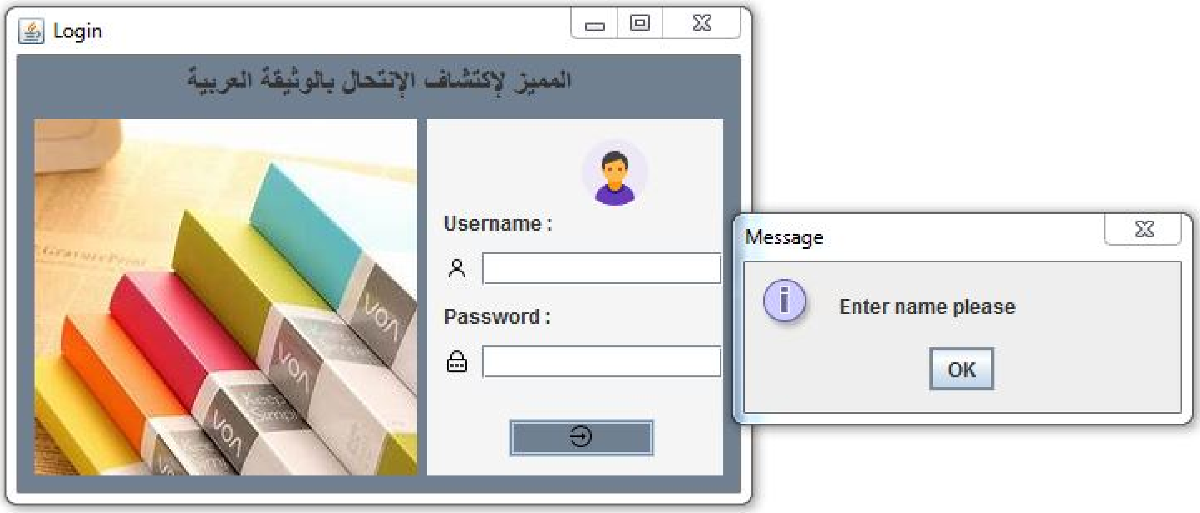
Username : akra

Message

Incorrect name please try again

Figure 6.13: The Login interface(2).

– In case when we do not enter username and the password, another error message will appear (Figure 6.14).



Login

المميز لإكتشاف الإنتحال بالوثيقة العربية

Username :

Password :

Message

Enter name please

Figure 6.14: The Login interface(3).

6.3.3 The Admin interface

The Admin interface presented in the figure bellow :

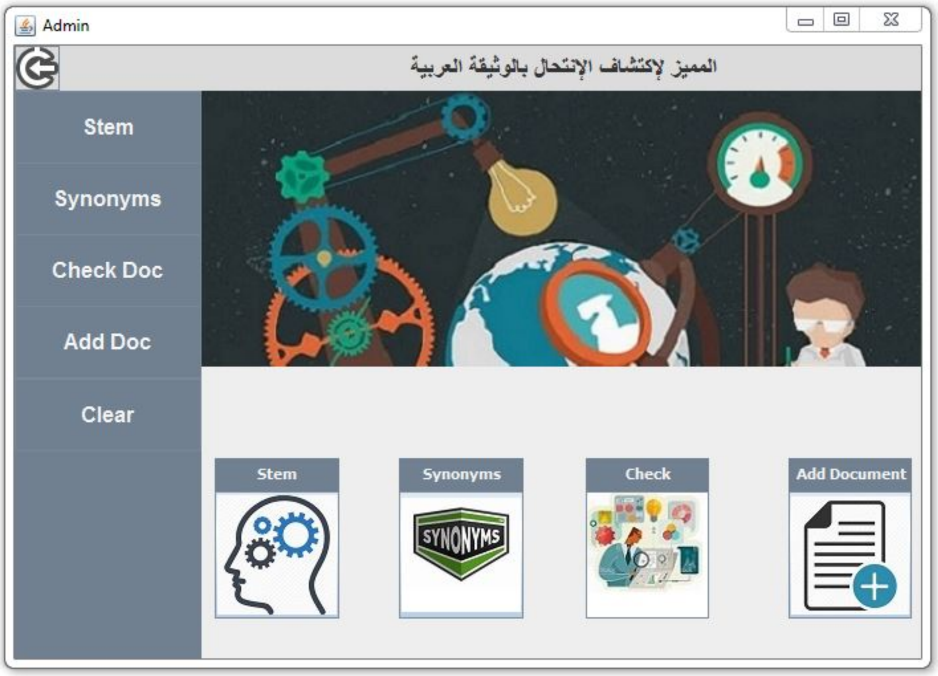
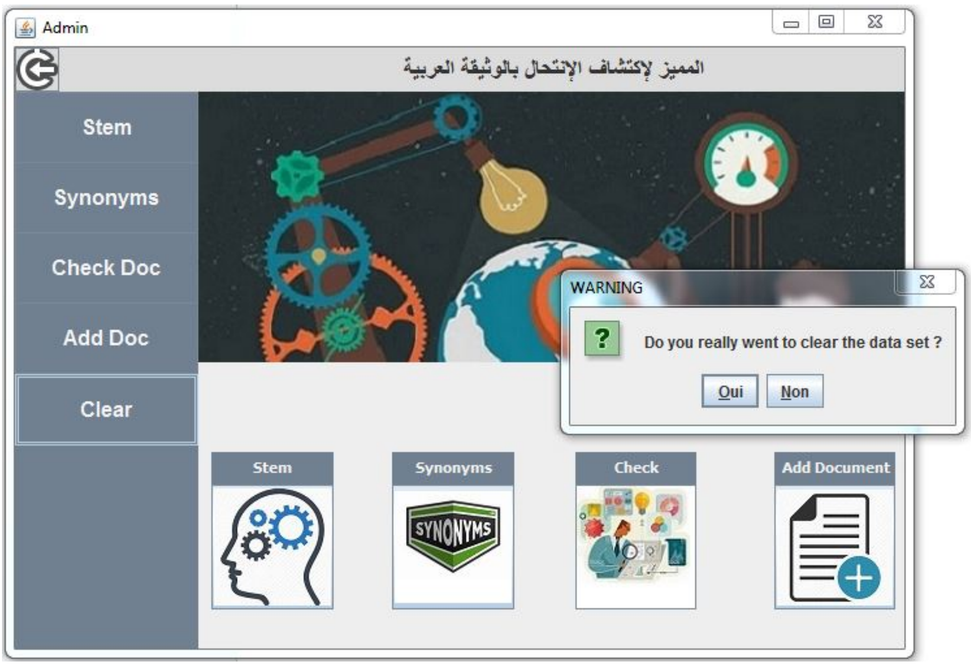


FIGURE 6.15: The Admin interface.

– As we see there is some buttons in this interface :

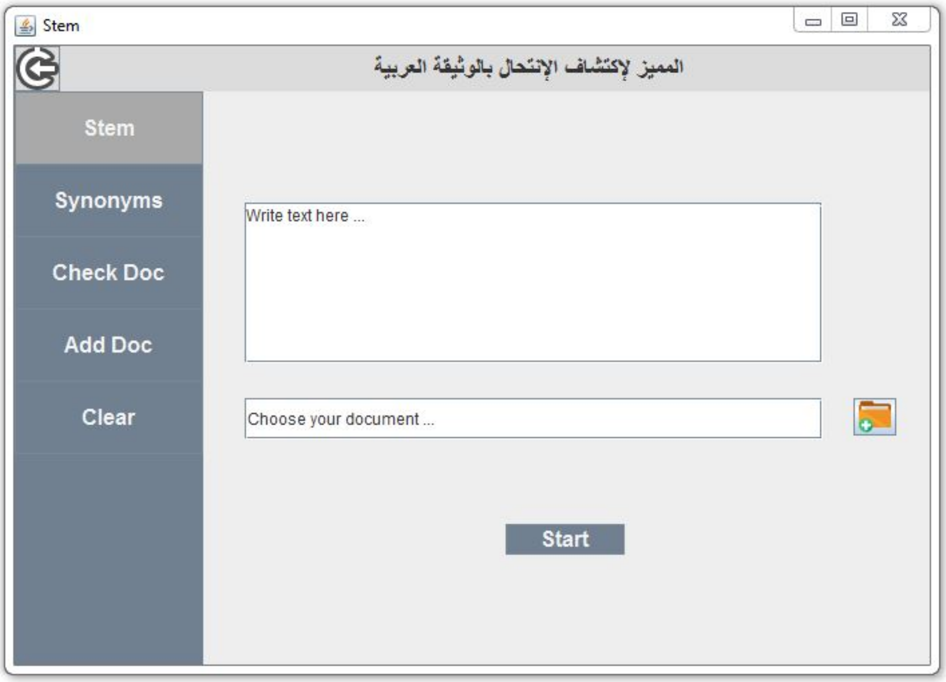
- Stem button : this button displays the "Stem" interface (Figure 6.17).
- Synonyms button : this button displays the "Synonyms" interface (Figure 6.22).
- Check Doc button : this button displays the "Check Document" interface (Figure 6.24).
- Add Doc button : this button displays the "Add Document" interface (Figure 6.43).
- Clear button : this button will clear all documents indexed in our directory (Figure 6.16).



Do you really went to clear the data set ?

FIGURE 6.16: Clear data set.

6.3.4 The Stem interface

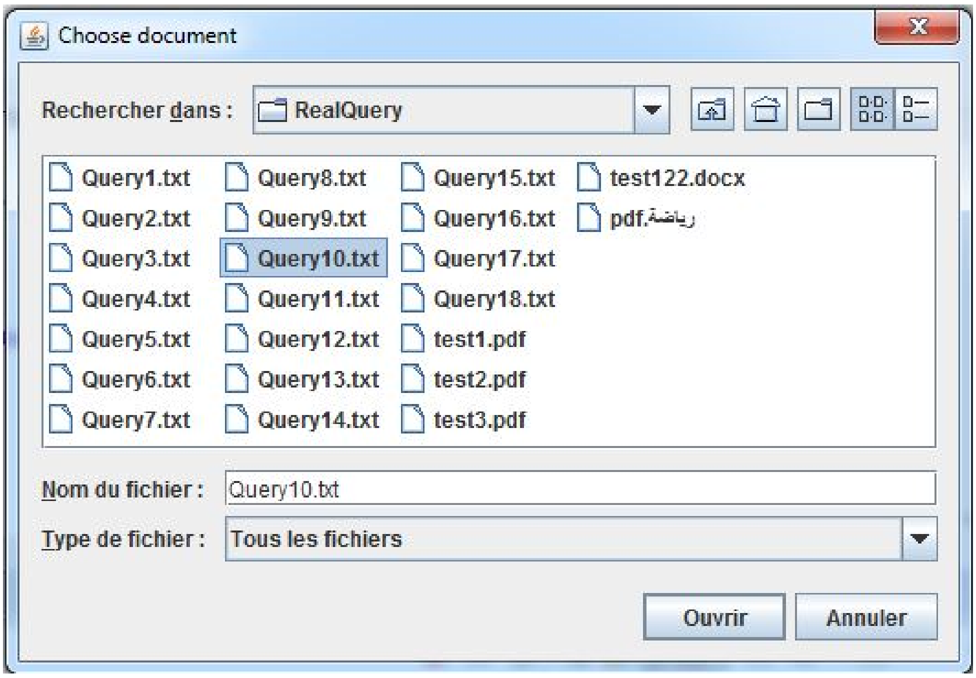


Write text here ...

Choose your document ...

FIGURE 6.17: The Stem interface.

– When we click the choose document button, "Choose document" interface will display (Figure 6.18).



Rechercher dans : RealQuery

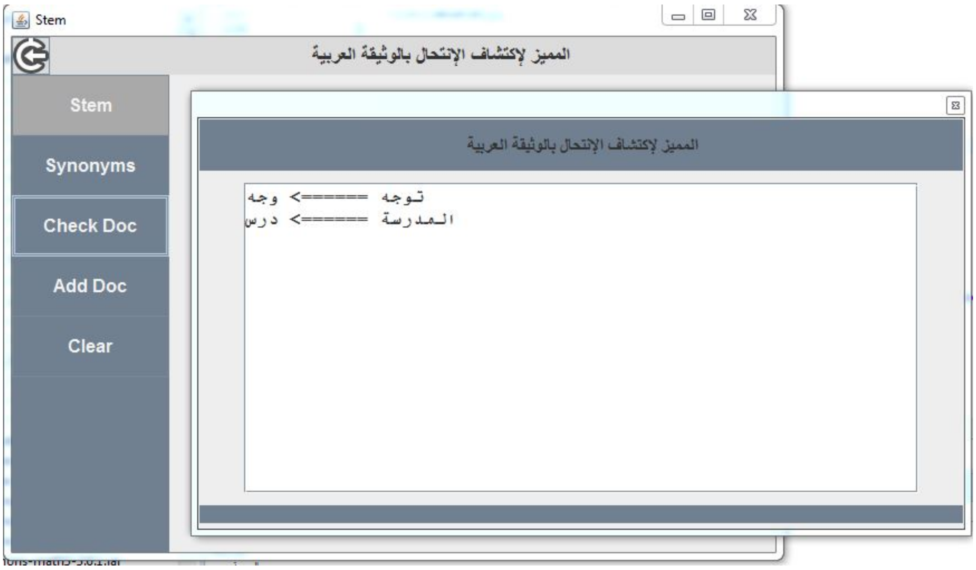
Query1.txt
Query2.txt
Query3.txt
Query4.txt
Query5.txt
Query6.txt
Query7.txt
Query8.txt
Query9.txt
Query10.txt
Query11.txt
Query12.txt
Query13.txt
Query14.txt
Query15.txt
Query16.txt
Query17.txt
Query18.txt
test1.pdf
test2.pdf
test3.pdf
test122.docx
رياضة.pdf

Nom du fichier : Query10.txt

Type de fichier : Tous les fichiers

FIGURE 6.18: The Choose Document interface.

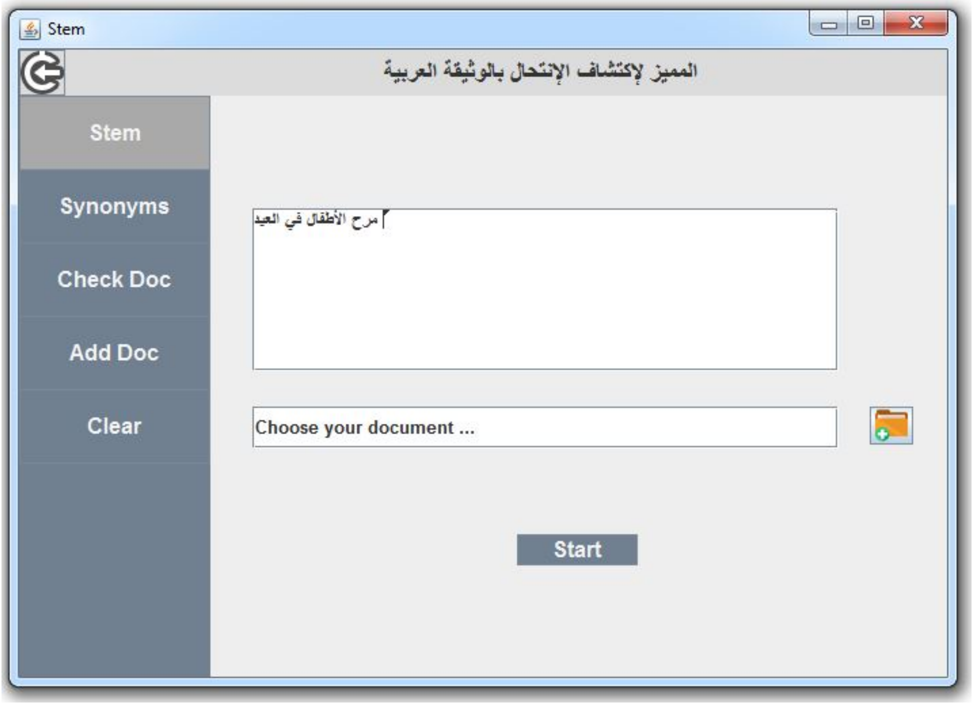
– After choosing document and click "Start" button new interface will display (Figure 6.19).


Stem
المميز لإكتشاف الإنتحال بالوثيقة العربية
Stem
Synonyms
Check Doc
Add Doc
Clear
المميز لإكتشاف الإنتحال بالوثيقة العربية

```
تـوجه ======> وجه
الـمدرسة ======> درس
```

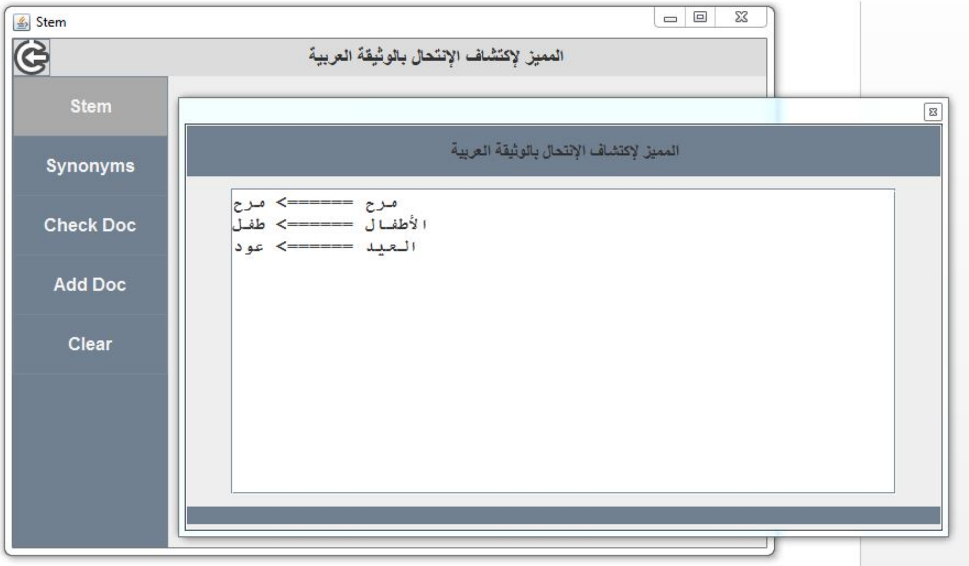
FIGURE 6.19: The Stem of document.

– Also we have the possibility of writing text in the text area (Figure 6.20) and apply the Natural Language Processing (NLP) by clicking on the "Start" button. The result of this operation will be shown in Figure 6.21.



مرح الأطفال في العيد

FIGURE 6.20: The Stem interface (text example).


Stem
المميز لإكتشاف الإنتحال بالوثيقة العربية
Stem
Synonyms
Check Doc
Add Doc
Clear
المميز لإكتشاف الإنتحال بالوثيقة العربية

```
مرح ======> مرح
الأطفال ======> طفل
العيد ======> عود
```

FIGURE 6.21: The Stem of text.

6.3.5 The Synonyms interface

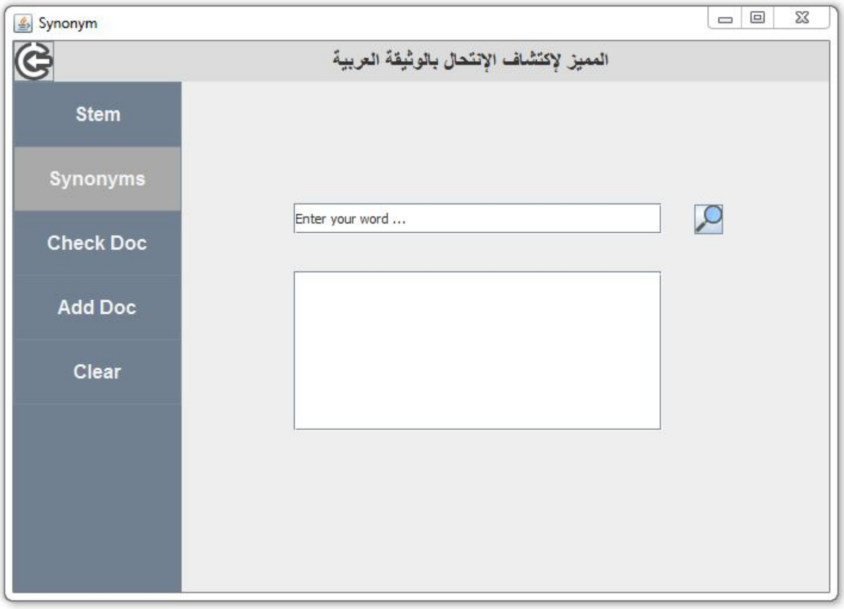
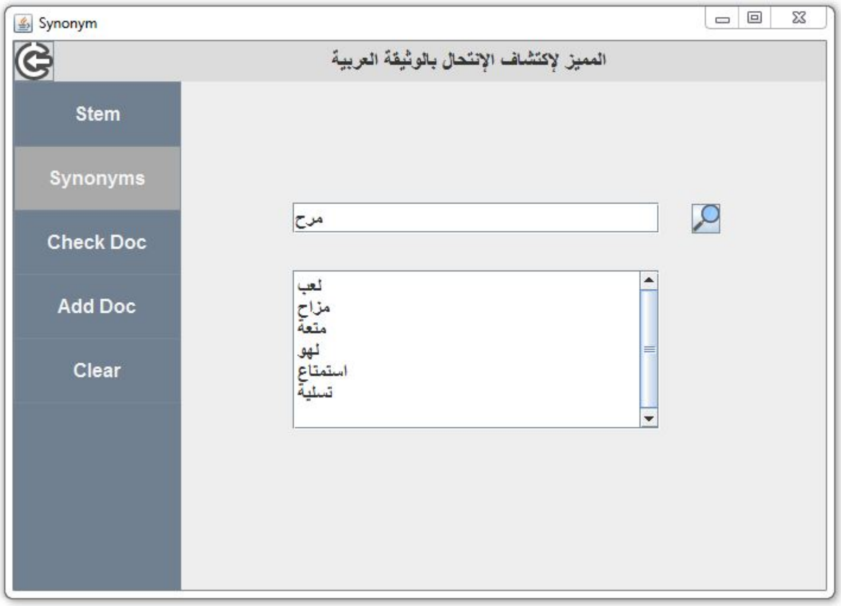


FIGURE 6.22: The Synonyms interface.

The Figure 6.22 aims to exploit the Arabic WordNet (AWN) as dictionary by writing word in the text area and click the search button, the result will be shown in the second text area like the example bellow :



مرح

لعب
مزاح
متعة
لهو
استمتاع
تسلية

FIGURE 6.23: The Synonyms interface (example).

6.3.6 The Check Document interface

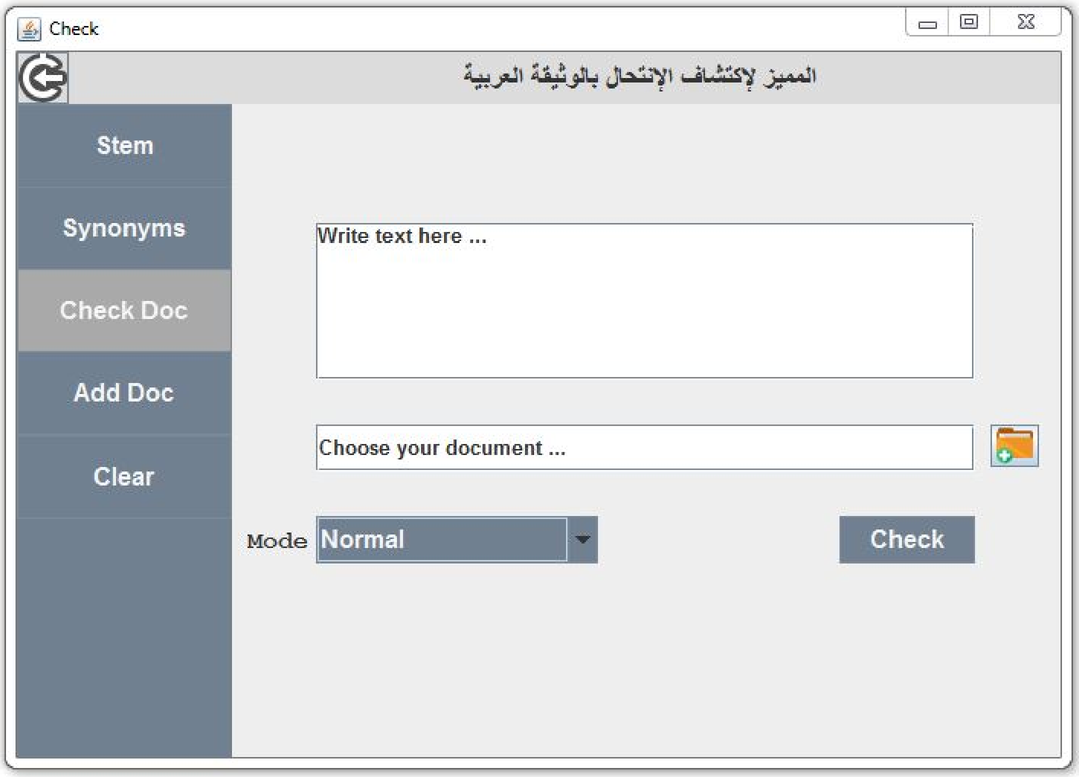
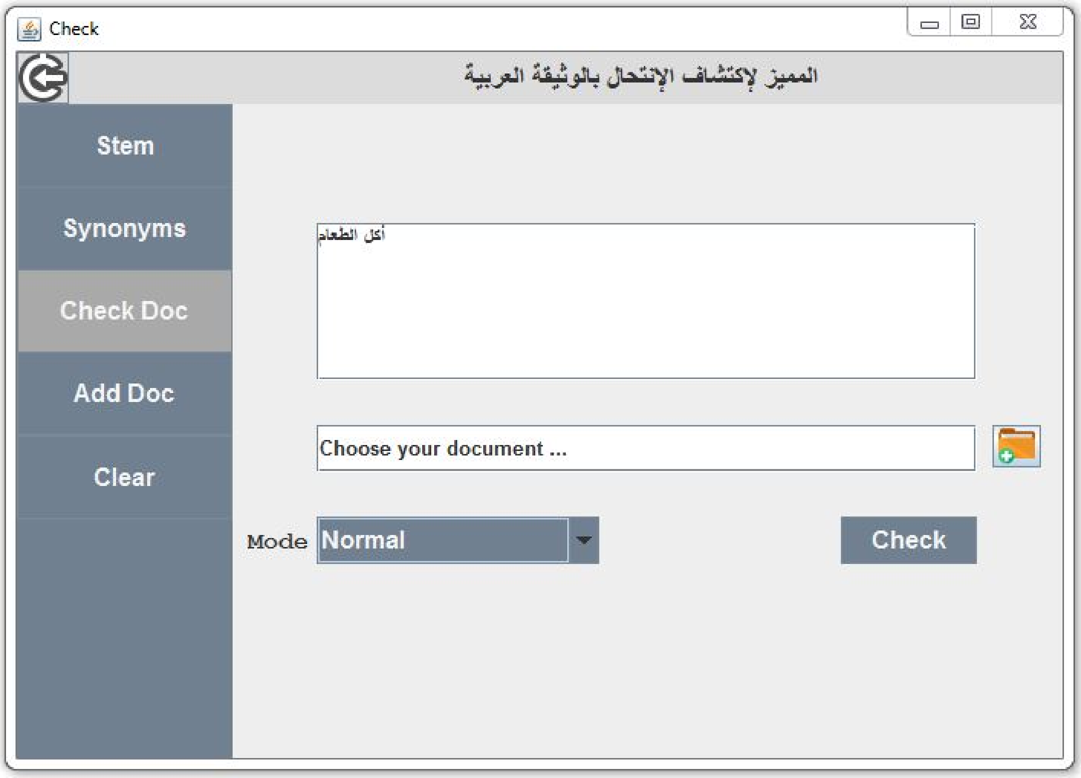


FIGURE 6.24: The Check Document interface.

- In this interface the admin has the possibility to write text in the text area and choose the check mode :

  1. Normal mode : in this mode, the system compares syntactically the text entered and similar document in the data set. Then, highlights the words repeated in the text and the similar document like the example bellow :

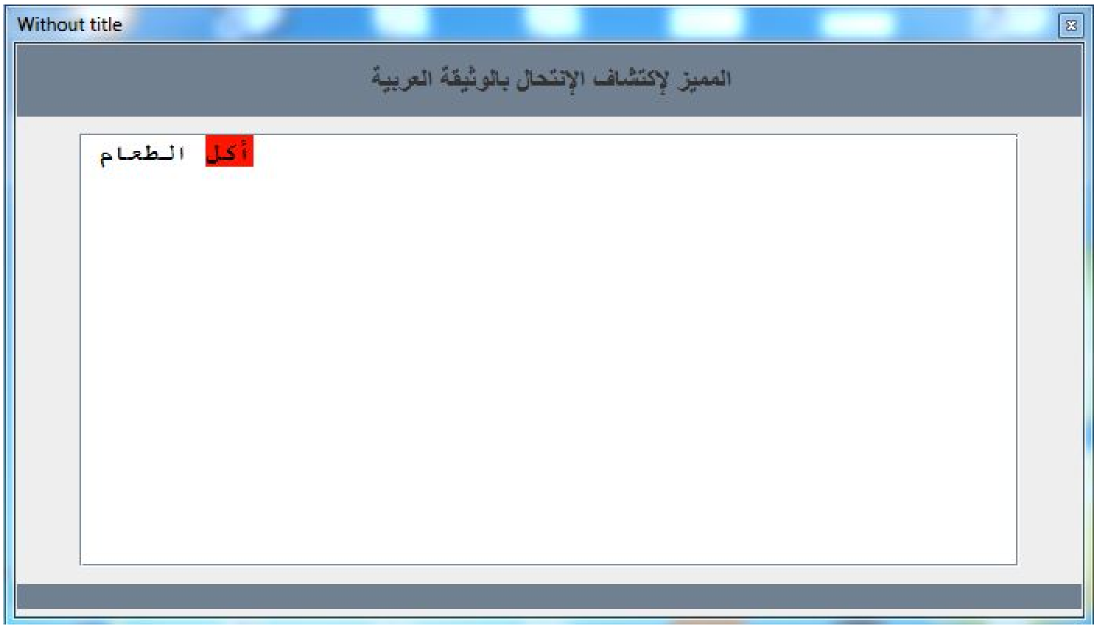


أكل الطعام

Choose your document ...

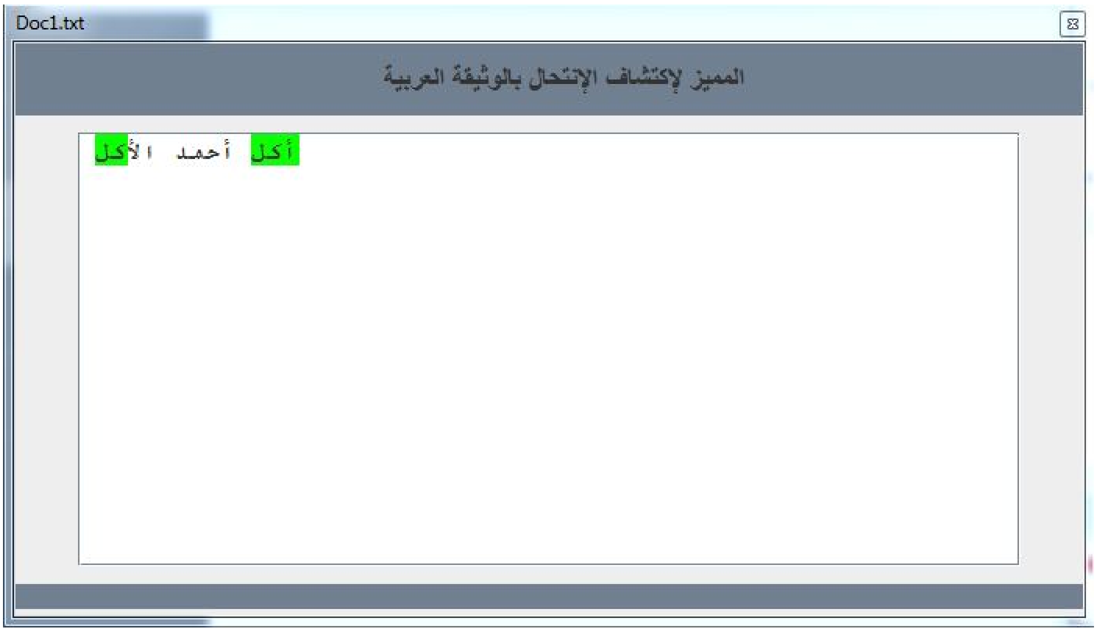
Mode Normal

FIGURE 6.25: The Check Document interface (text example).



أكل الطعام

FIGURE 6.26: The repeated words in the entered text.



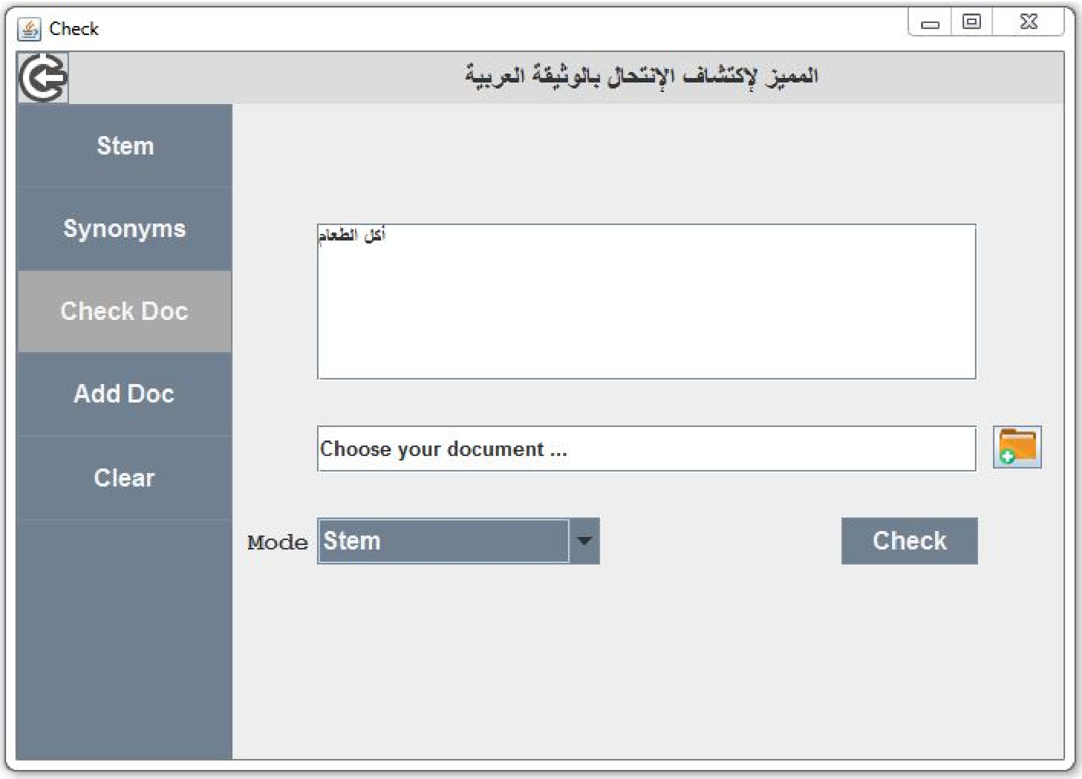
Doc1.txt

المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل أحمد الأكل

FIGURE 6.27: The repeated words in the similar document.

2. Stem mode : in this mode, the system applies the NLP process for the text in order to get the stem of each word in the text, then compares it with similar document in the data set and highlights the words that have the same root like the example bellow:



Check

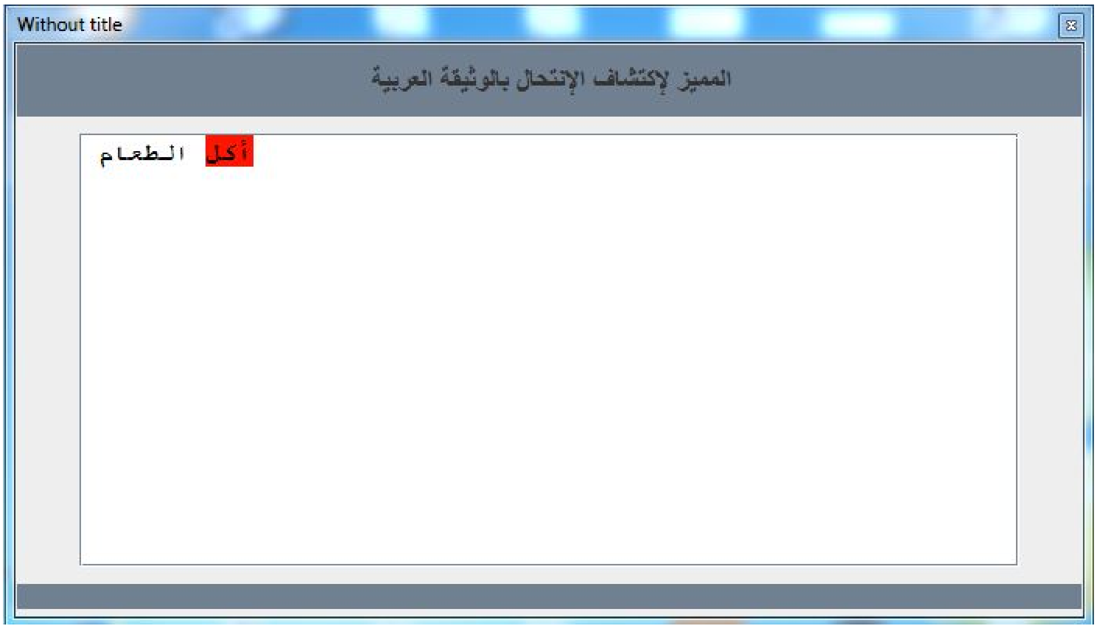
المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل الطعام

Choose your document ...

Mode Stem

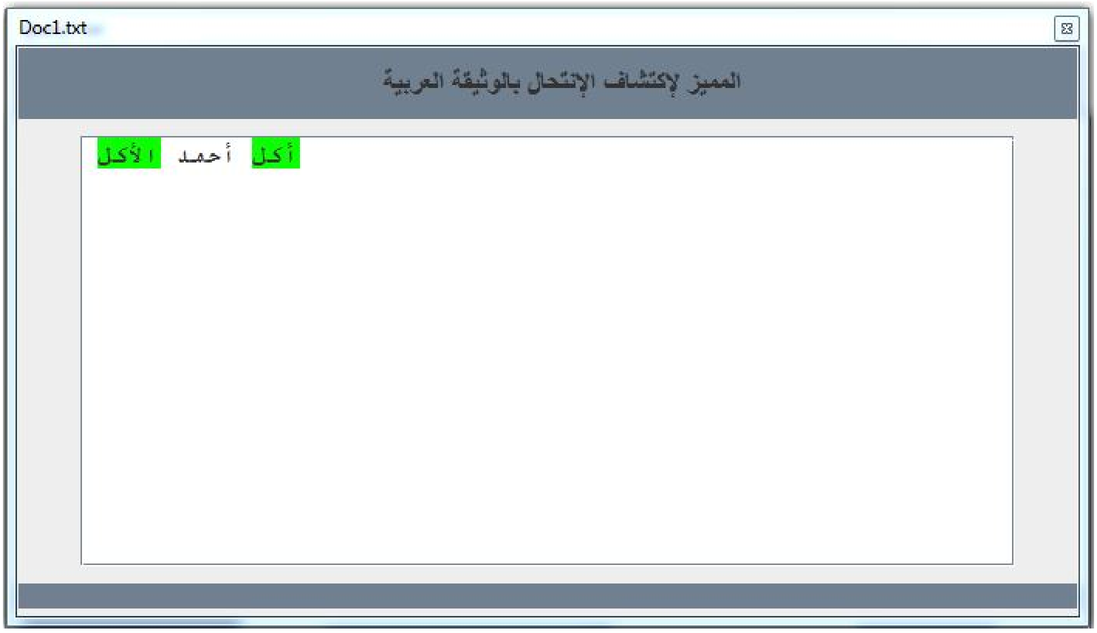
FIGURE 6.28: The Check Document interface (example) .



المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل الطعام

Figure 6.29: The repeated words in the entered text.

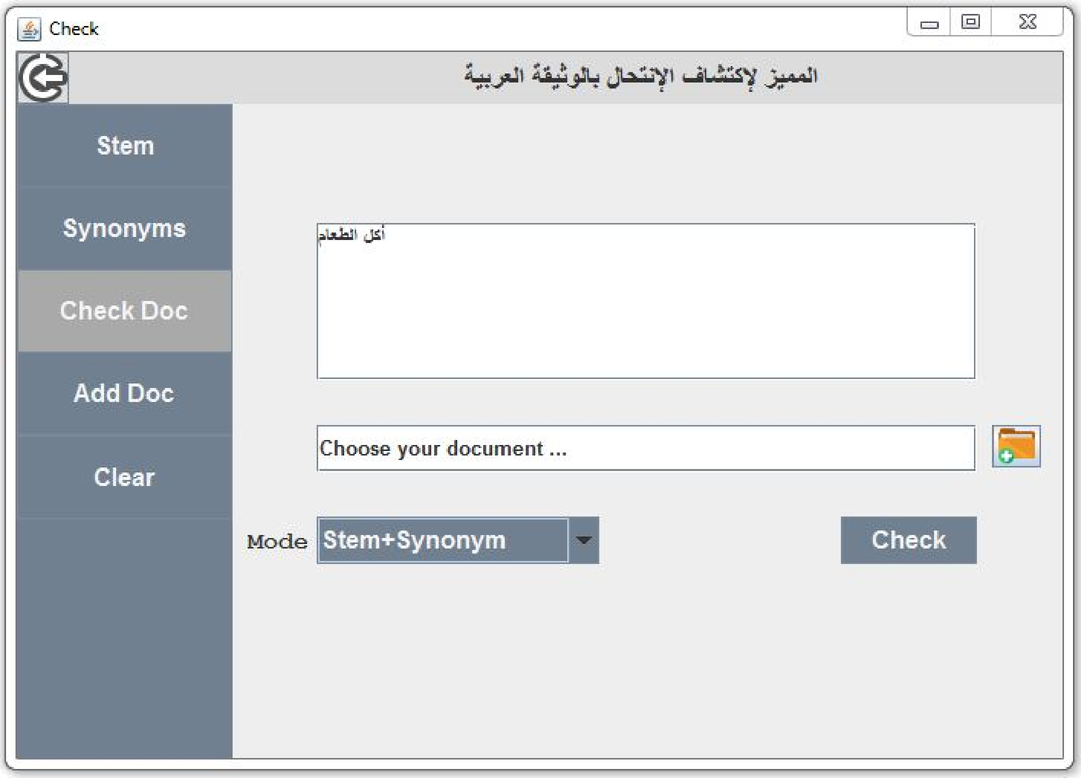


المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل أحمد الأكل

Figure 6.30: The repeated words in the similar document .

3. Stem+Synonym mode : in this mode, the system gets the stem of each word in the text, then compares it with similar document in the data set and highlights the words that have the same root or in case of synonym like the example bellow :

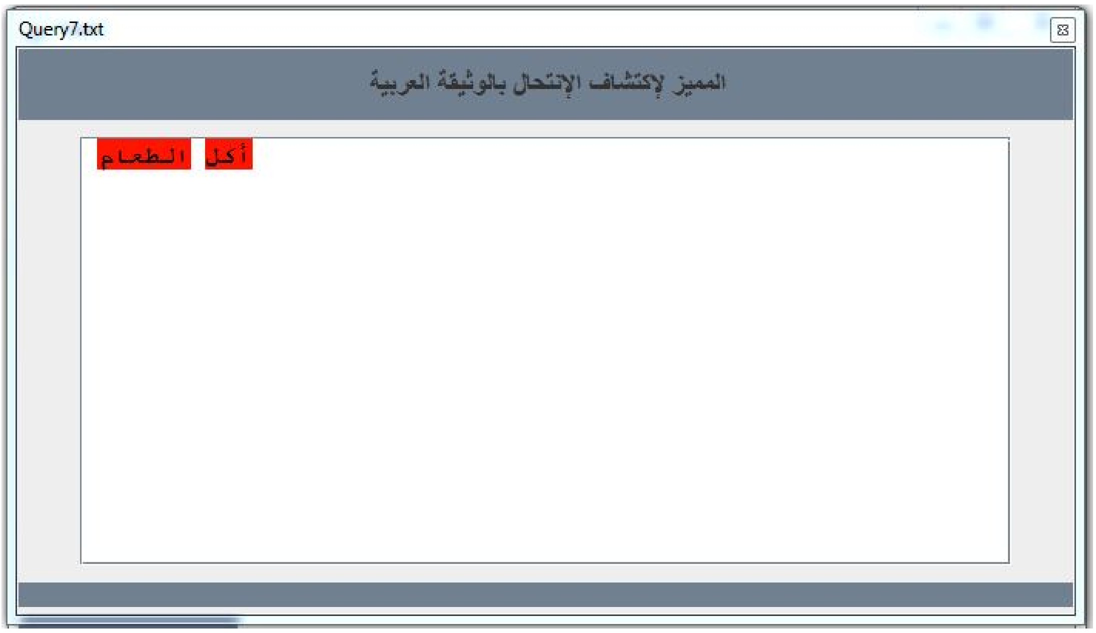


أكل الطعام

Choose your document ...

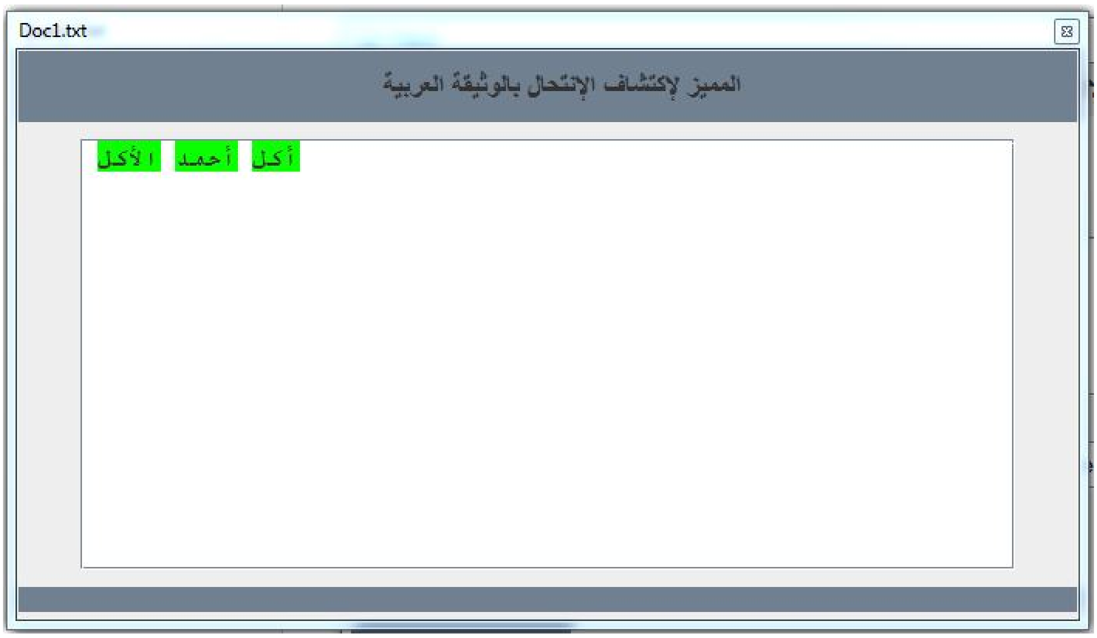
Mode Stem+Synonym

FIGURE 6.31: The Check Document interface (example) .



أكل الطعام

FIGURE 6.32: The repeated words in the entered text .



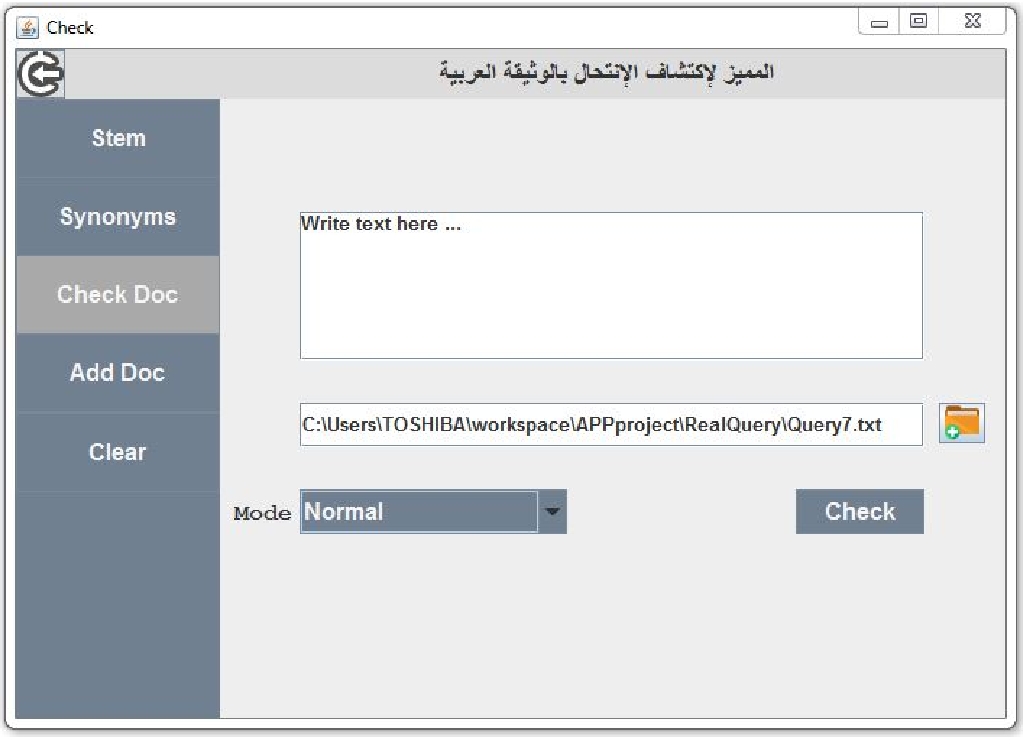
المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل أحمد ا لأكل

FIGURE 6.33: The repeated words in the similar document .

- Also we have the possibility of choosing the document that we aim to check it with three modes :

  1. Normal mode : in this mode, the system compares syntactically the document choosed and the similar document in the data set and highlights the words repeated in the selected document and the similar document like the example bellow:



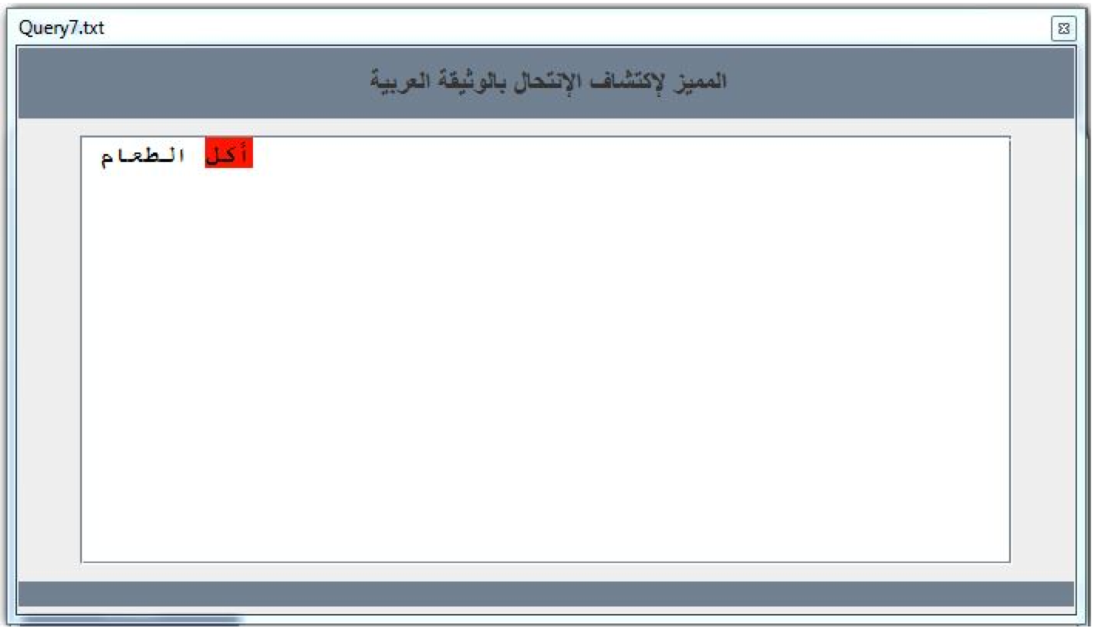
المميز لإكتشاف الإنتحال بالوثيقة العربية

Write text here ...

C:\Users\TOSHIBA\workspace\APPproject\RealQuery\Query7.txt

Mode Normal

FIGURE 6.34: The Check Document interface (document example).

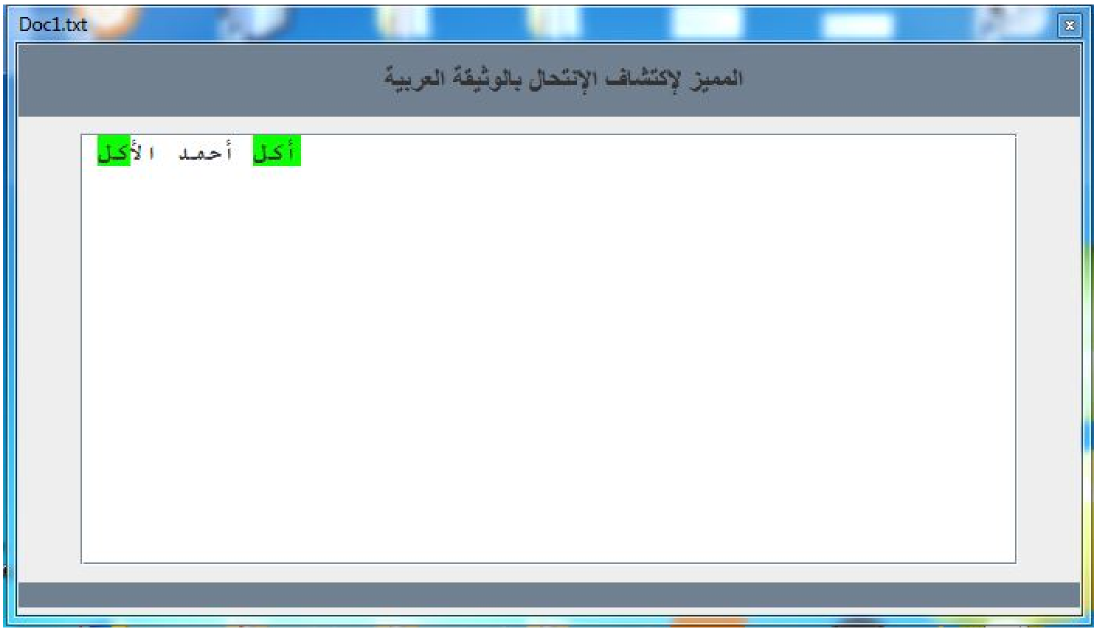


Query7.txt

المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل الطعام

FIGURE 6.35: The repeated words in the selected document.



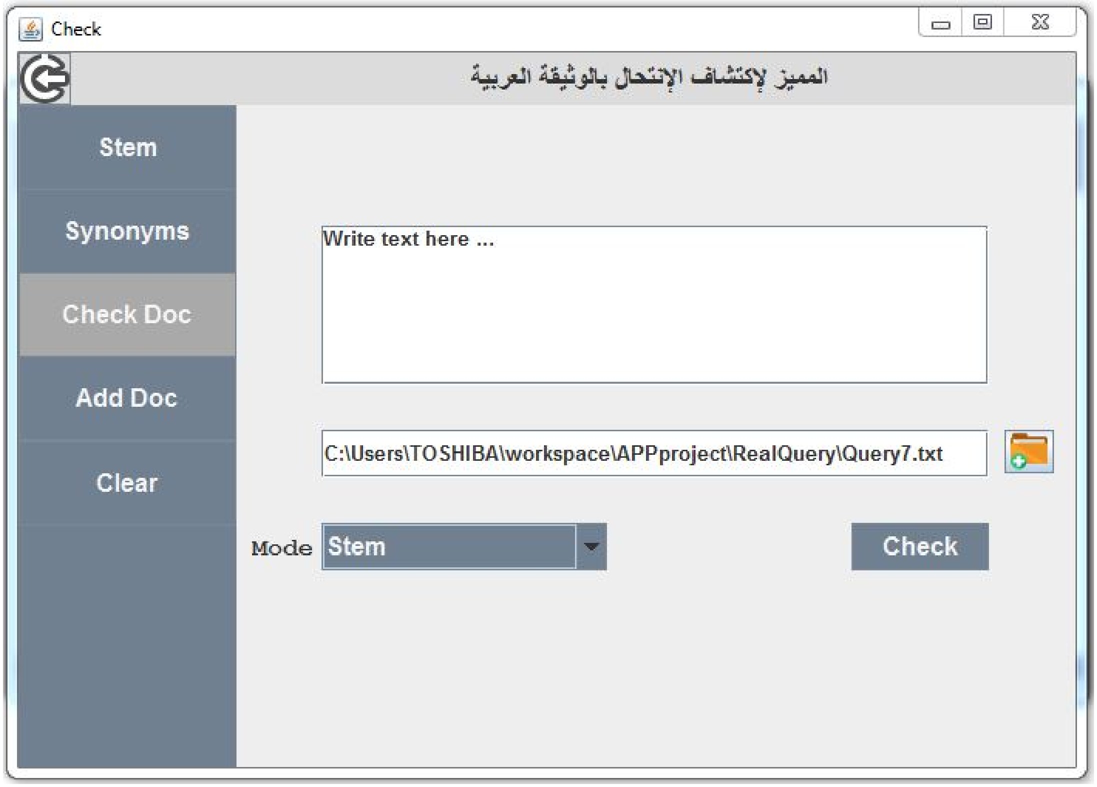
Doc1.txt

المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل أحمد الأكل

FIGURE 6.36: The repeated words in the similar document.

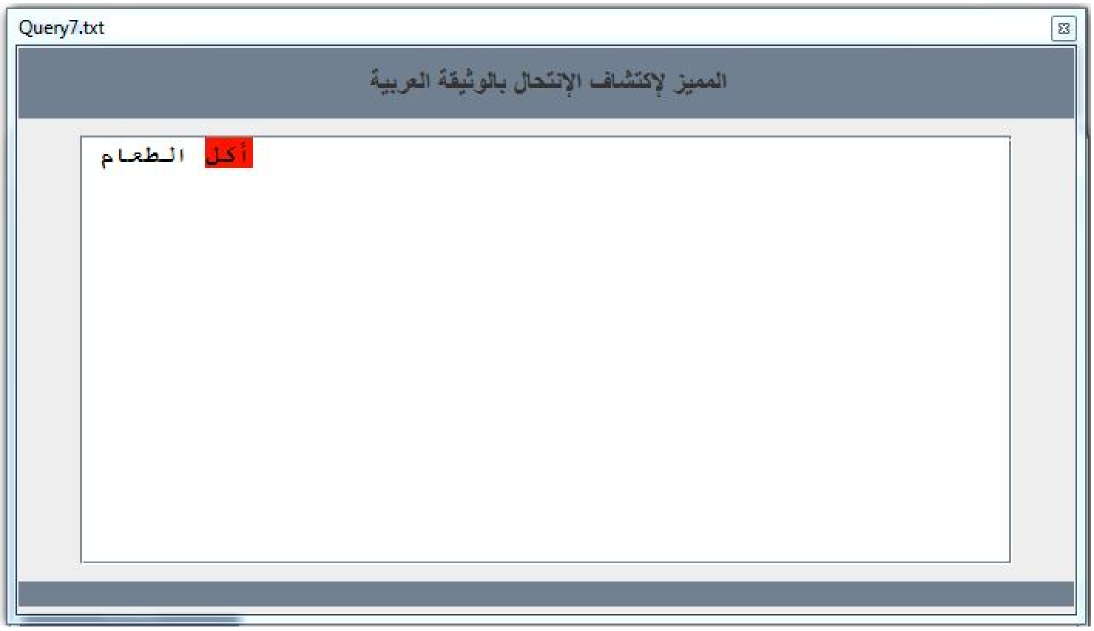
2. Stem mode : in this mode, the system applies the NLP process for the selected document in order to get the stem of each word in this document, then compares it with the similar document in the data set and highlights the words that have the same root like the example bellow:



C:\Users\TOSHIBA\workspace\APPproject\RealQuery\Query7.txt

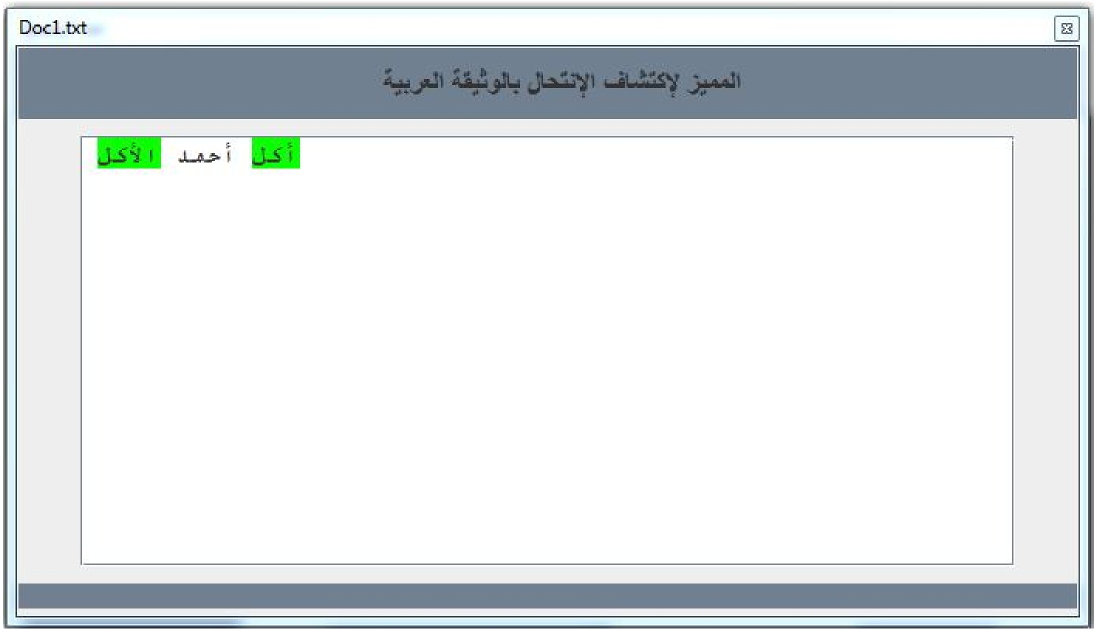
Mode Stem

FIGURE 6.37: The Check Document interface (document example).



أكل الطعام

FIGURE 6.38: The repeated words in the selected document.



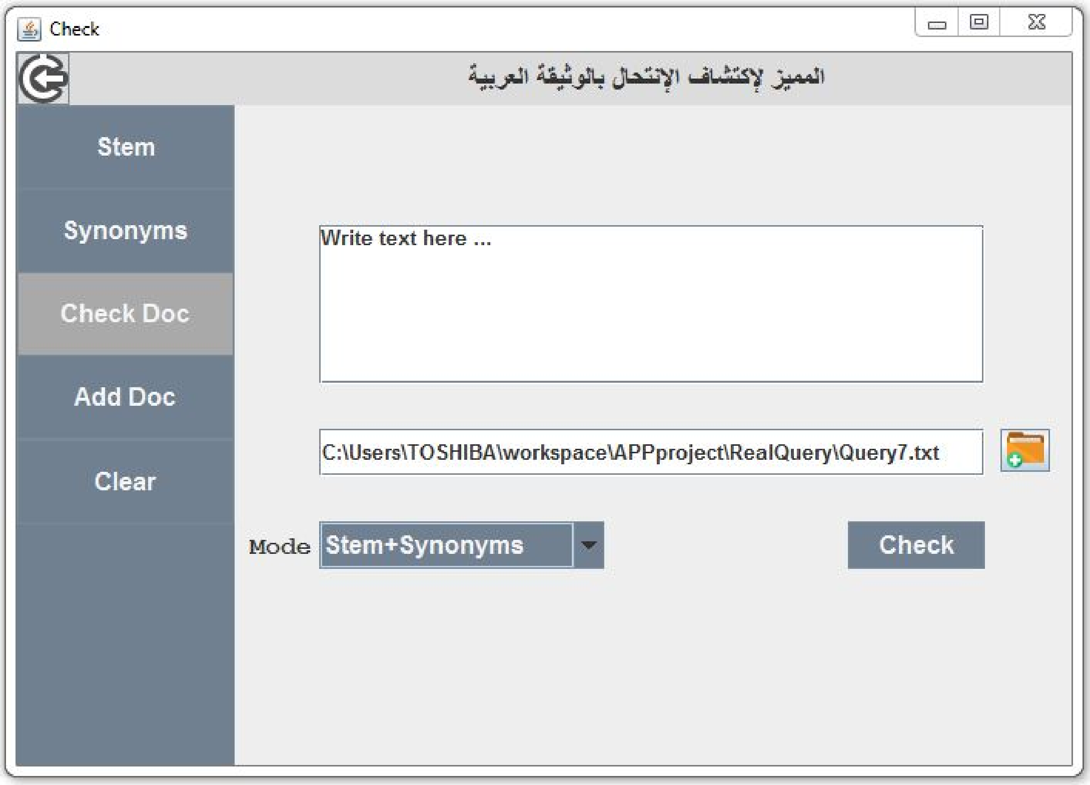
Doc1.txt

المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل أحمد الأكل

FIGURE 6.39: The repeated words in the similar document.

3. Stem+Synonym mode : in this mode, the system gets the stem of each word in the selected document, then compares it with the similar document in the data set and highlights the words that have the same root or in case of synonym like the example bellow:



Check

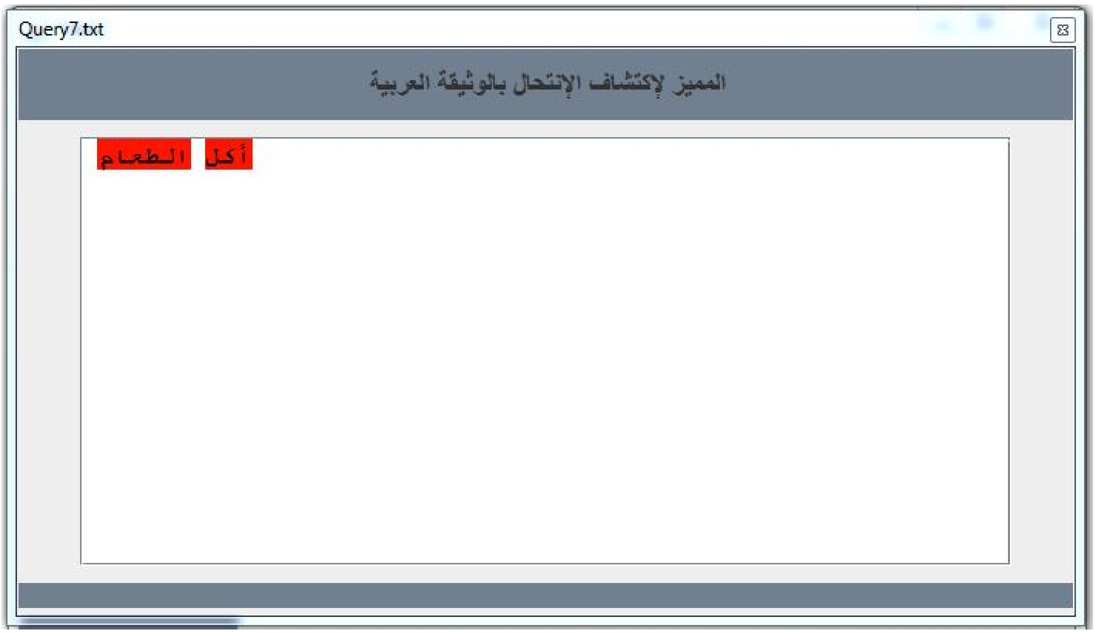
المميز لإكتشاف الإنتحال بالوثيقة العربية

Write text here ...

C:\Users\TOSHIBA\workspace\APPproject\RealQuery\Query7.txt

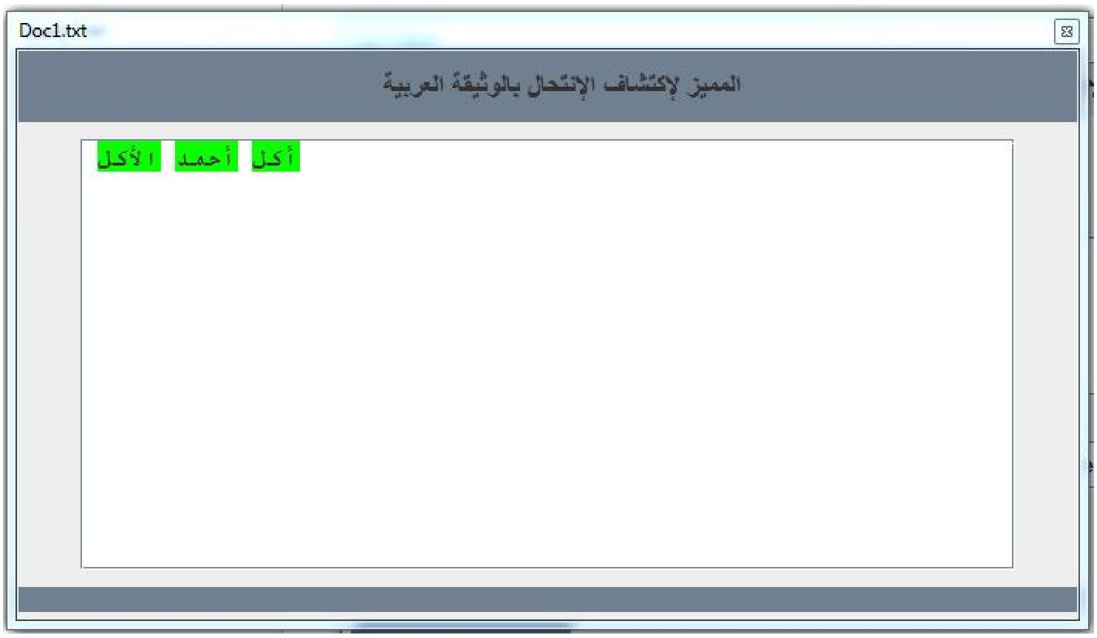
Mode Stem+Synonyms

FIGURE 6.40: The Check Document interface (document example).



أكل الطعام

Figure 6.41: The repeated words in the selected document.



أكل أحمد الأكل

Figure 6.42: The repeated words in the similar document.

6.3.7 The Add Document interface

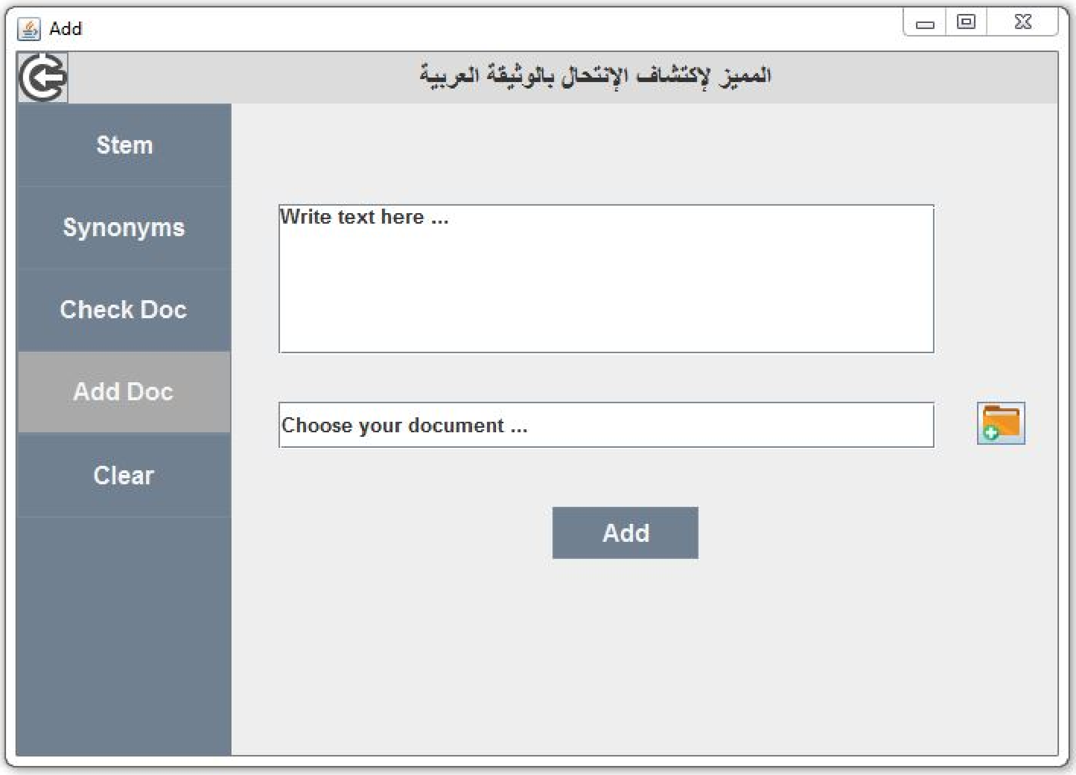
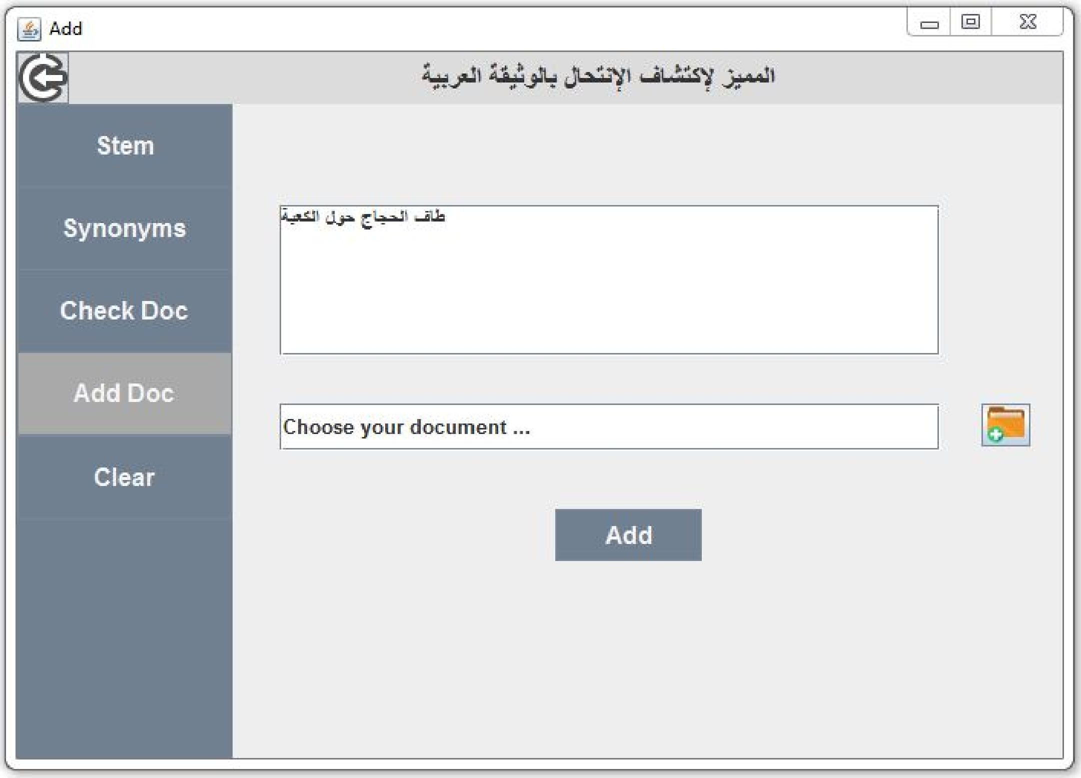


Figure 6.43: The Add Document interface.

In this interface the admin can add new document in his data set by writing text in the text area and saves it as new document or by selecting document from directory .

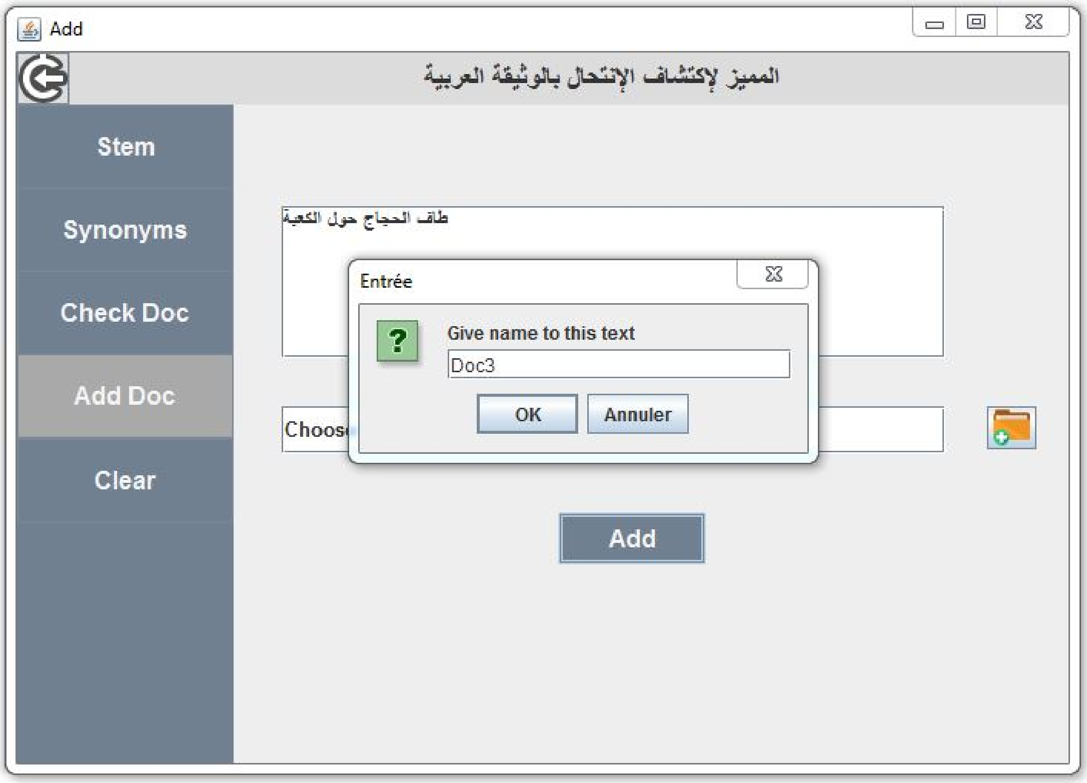
- The case of writing in text area :



طاف الحجاج حول الكعبة

Choose your document ...

Figure 6.44: The Add Document interface(text example).



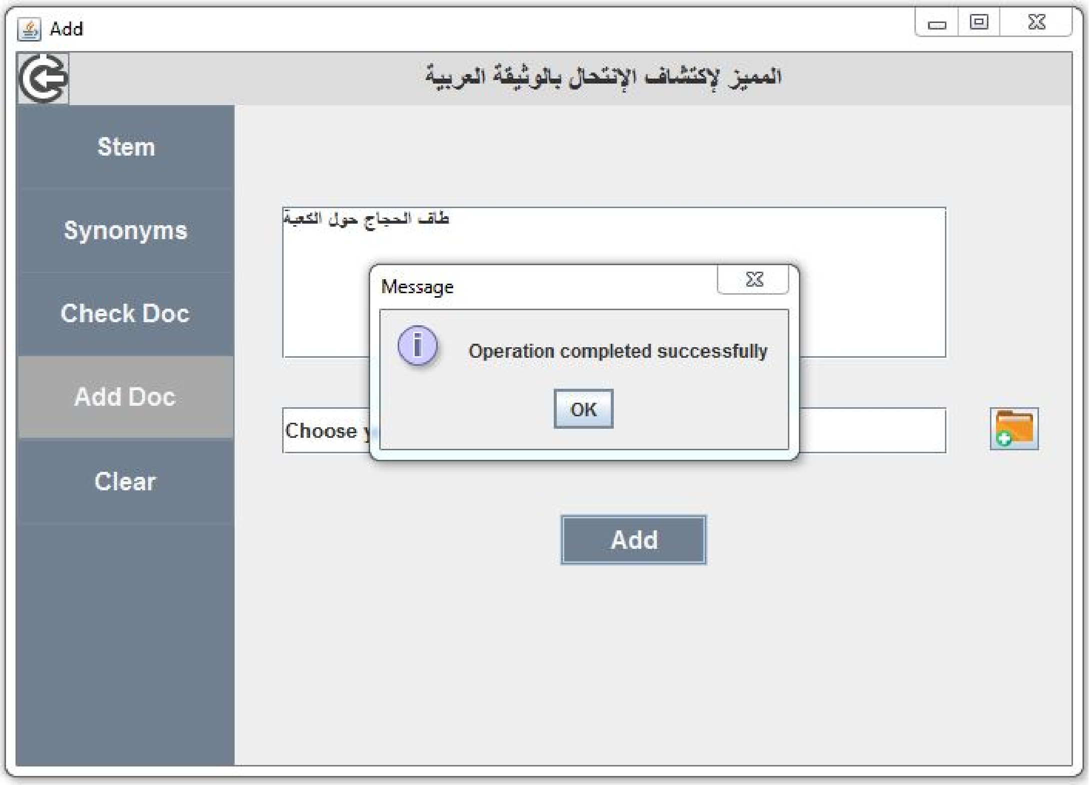
طاف الحجاج حول الكعبة

Give name to this text

Doc3

Choos

Figure 6.45: The Add Document interface(choosing name for the document).



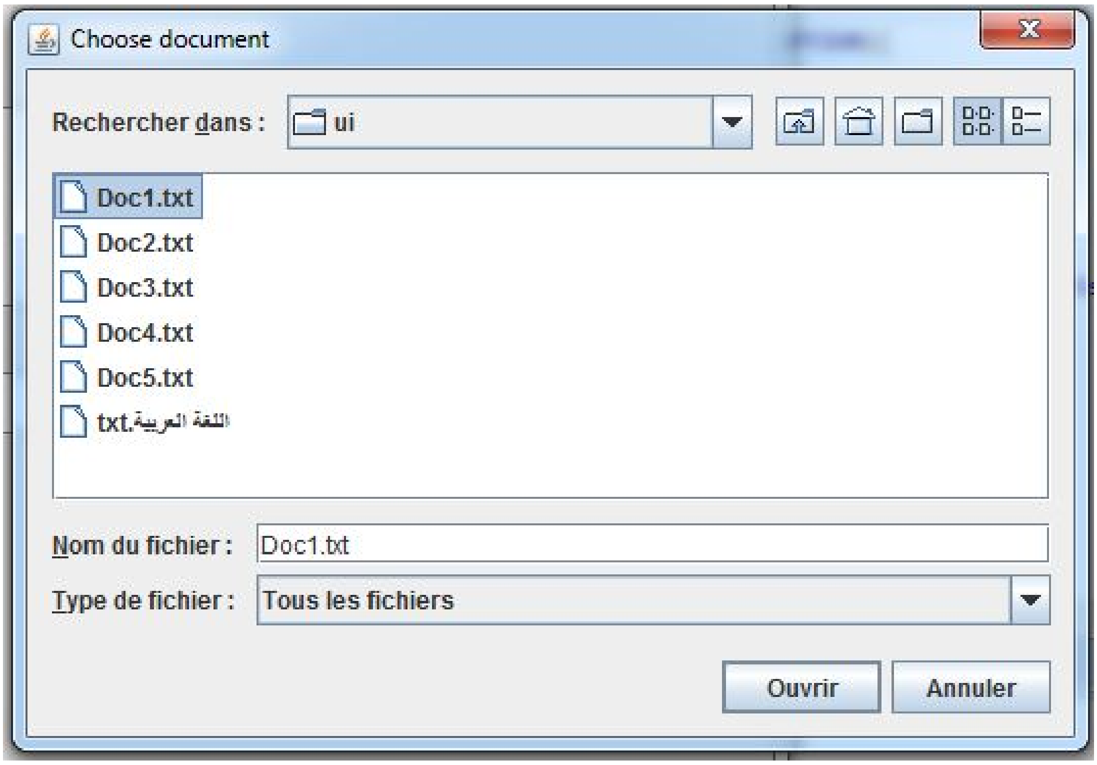
طاف الحجاج حول الكعبة

Choose y

Operation completed successfully

FIGURE 6.46: The Add Document interface(finishing the add).

- The case of selecting document :



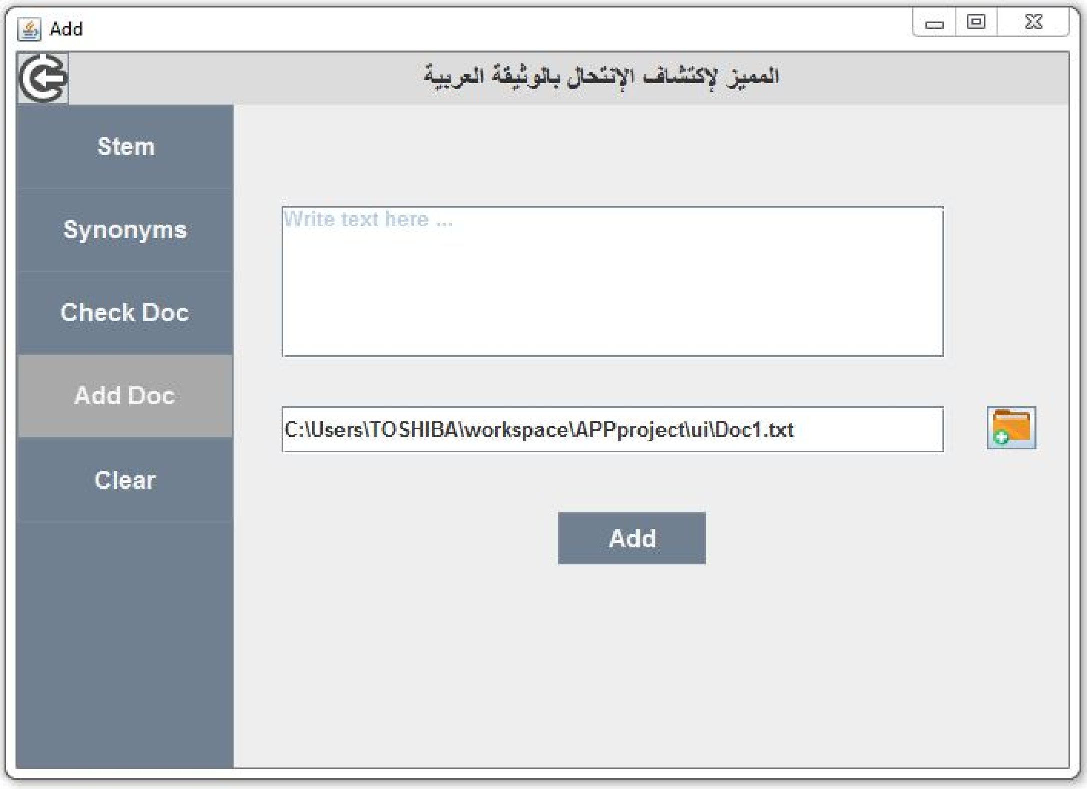
Rechercher dans : ui

Doc1.txt
Doc2.txt
Doc3.txt
Doc4.txt
Doc5.txt
اللغة العربية.txt

Nom du fichier : Doc1.txt

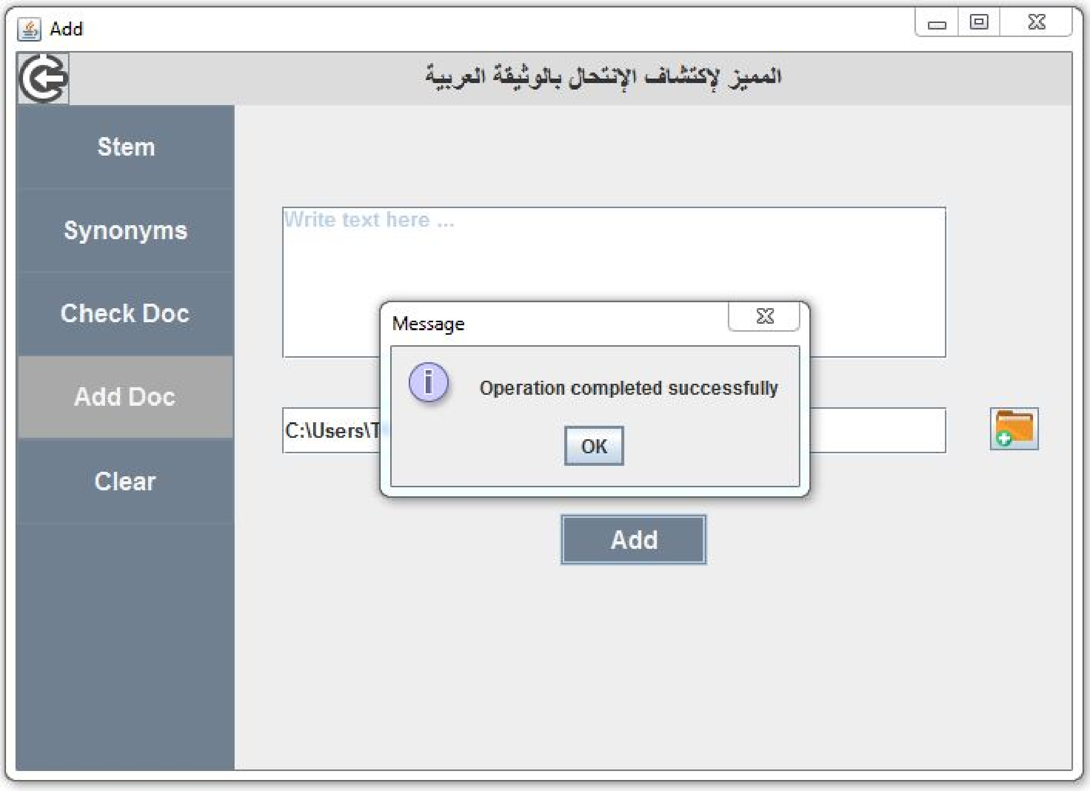
Type de fichier : Tous les fichiers

FIGURE 6.47: The Add Document interface(selecting document).



C:\Users\TOSHIBA\workspace\APPproject\ui\Doc1.txt

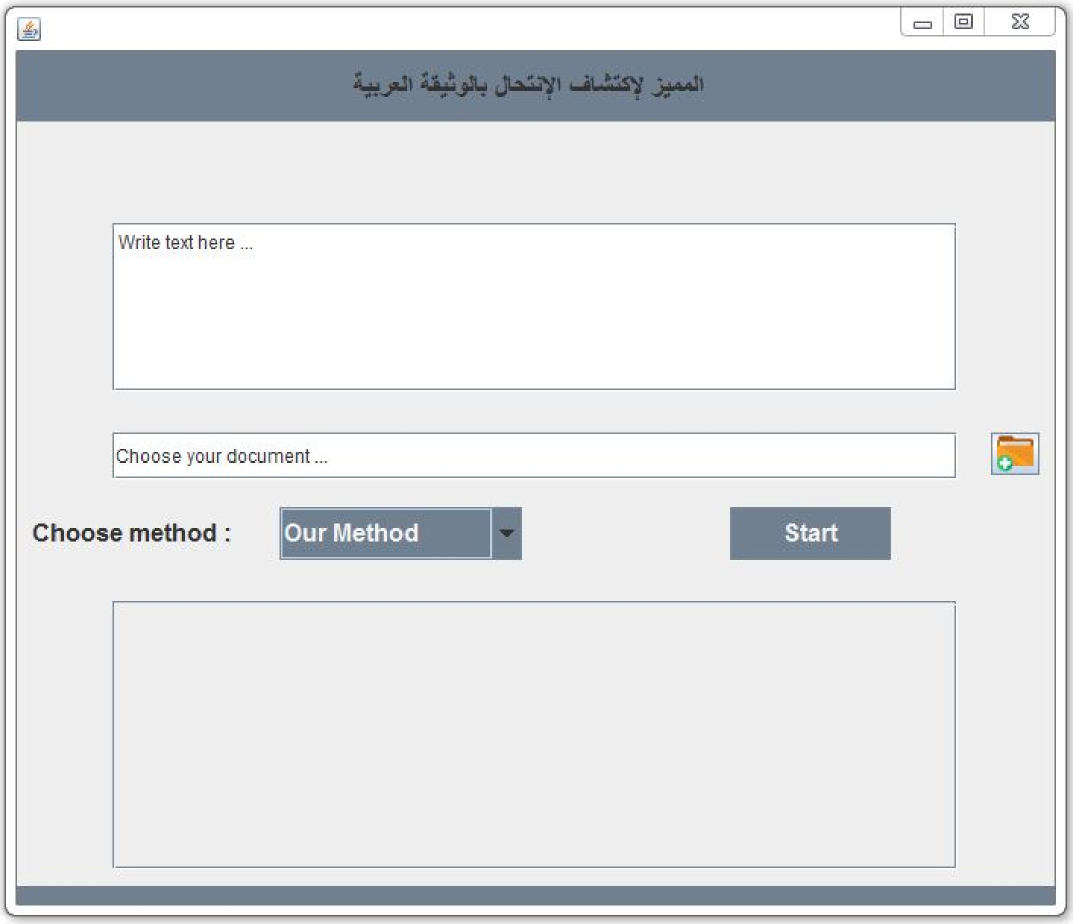
FIGURE 6.48: The Add Document interface(document selected).



Operation completed successfully

FIGURE 6.49: The Add Document interface(finishing the add).

6.3.8 The User interface



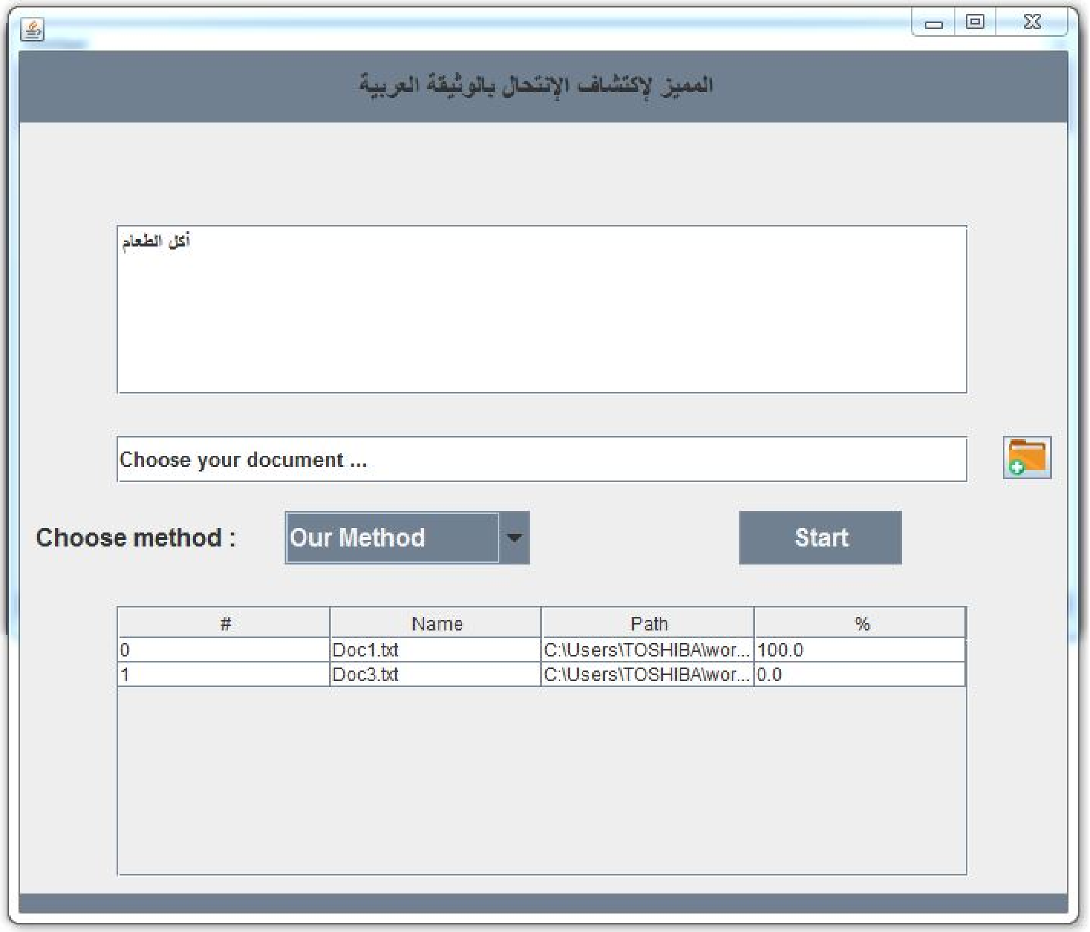
Choose method : Our Method

FIGURE 6.50: The User interface.

– The user interface is presented in figure 6.50. The user in this interface has the possibility to :

- Write his text in the text area and choose the method that he wants as it shown in the example bellow :

  1. In case when the user chooses "Our Method" and clicks the "Start" button, the system starts calculating the similarity between this text and all the documents in the data set using our method and shows the result in the table(see Figure 6.51).

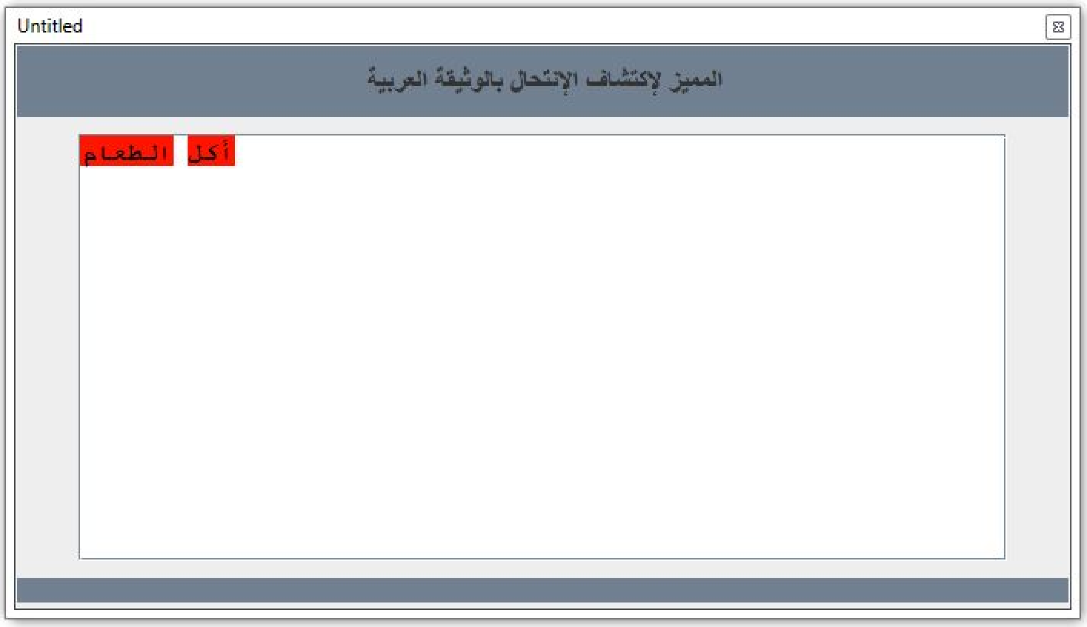


أكل الطعام

#	Name	Path	%
0	Doc1.txt	C:\Users\TOSHIBA\wor...	100.0
1	Doc3.txt	C:\Users\TOSHIBA\wor...	0.0

FIGURE 6.51: The example of choosing "Our Method".

Also the system displays the entered text with highlighted words that are plagiarised (Figure 6.55), and displays the most similar document with highlighted words that are mentioned in the entered text (Figure 6.56).

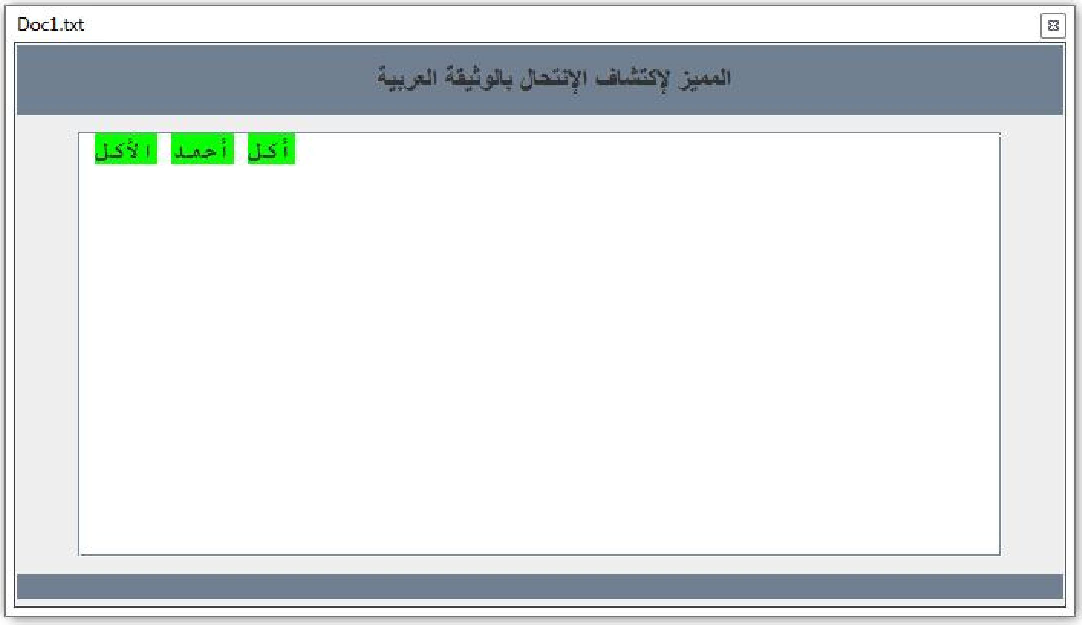


Untitled

المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل الطعام

FIGURE 6.52: The highlighted words in the entered text.



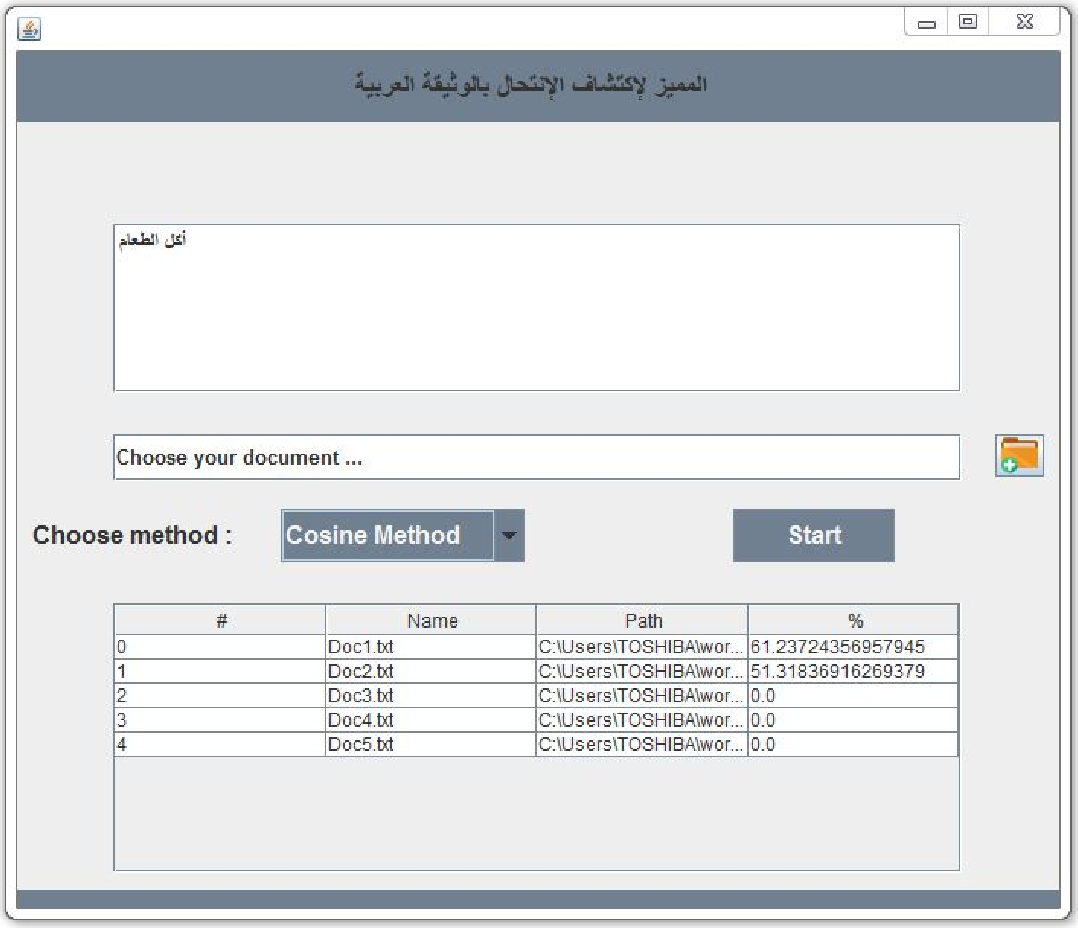
Doc1.txt

المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل أحمد الأكل

FIGURE 6.53: The highlighted words in the most similar document.

2. In case when the user chooses "Cosine Method" and clicks the "Start" button, the system starts calculating the similarity between this text and all the documents in the data set using the cosine method and shows the result in the table(see Figure 6.54).



المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل الطعام

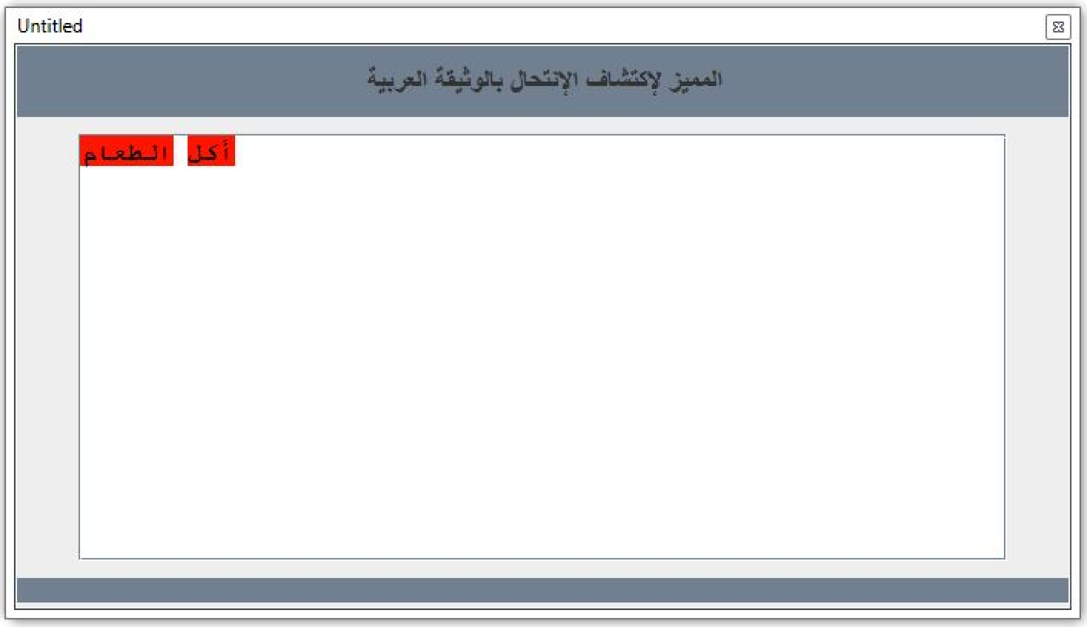
Choose your document ...

Choose method : Cosine Method

#	Name	Path	%
0	Doc1.txt	C:\Users\TOSHIBA\wor...	61.23724356957945
1	Doc2.txt	C:\Users\TOSHIBA\wor...	51.31836916269379
2	Doc3.txt	C:\Users\TOSHIBA\wor...	0.0
3	Doc4.txt	C:\Users\TOSHIBA\wor...	0.0
4	Doc5.txt	C:\Users\TOSHIBA\wor...	0.0

FIGURE 6.54: The example of choosing "Cosine Method".

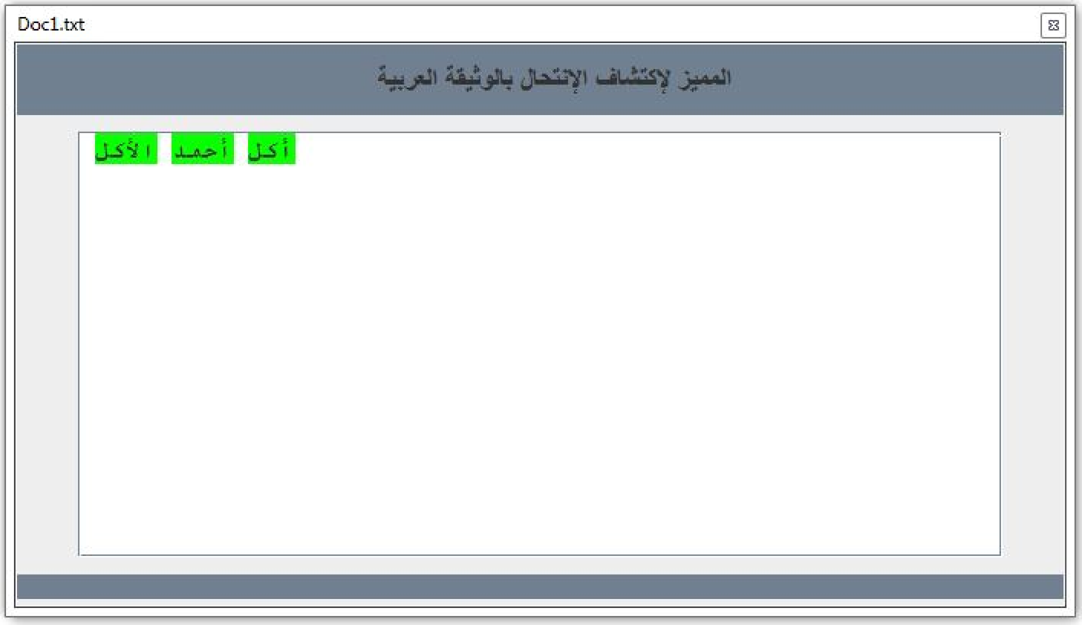
Also the system displays the entered text with highlighted words that are plagiarised (Figure 6.55), and displays the most similar document with highlighted words that are mentioned in the entered text (Figure 6.56).



المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل الطعام

FIGURE 6.55: The highlighted words in the entered text.

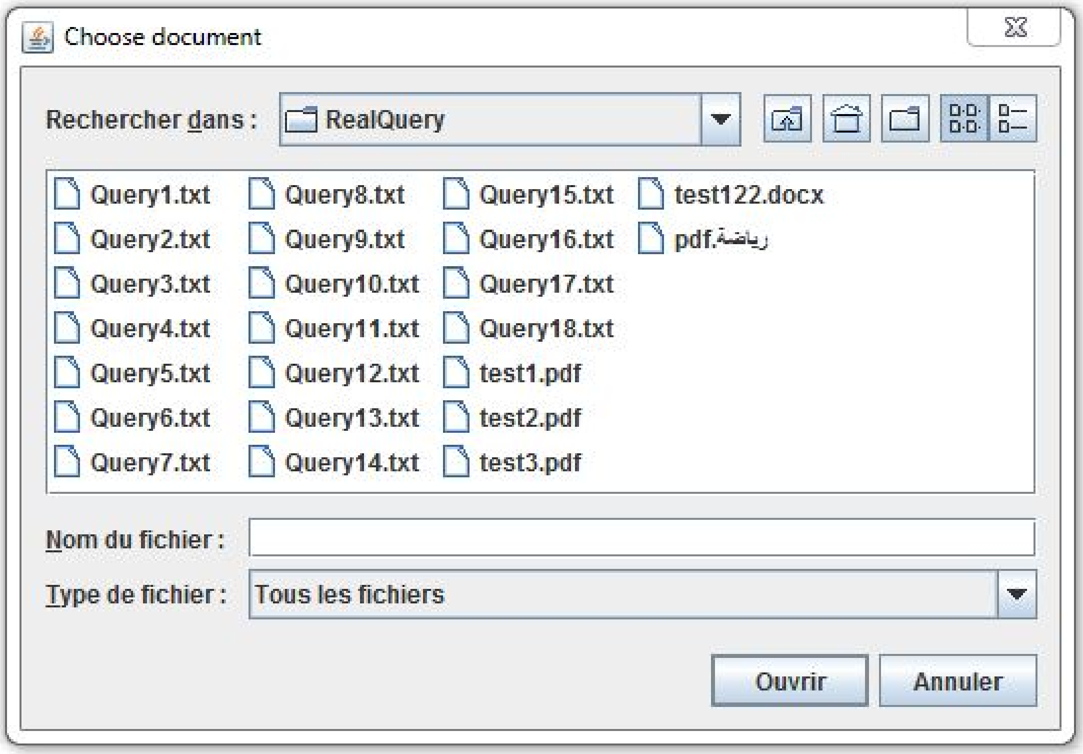


المميز لإكتشاف الإنتحال بالوثيقة العربية

أكل أحمد الأكل

FIGURE 6.56: The highlighted words in the most similar document.

- Select document from directory (Figure 6.57), and apply on it one of the available methods.



Choose document

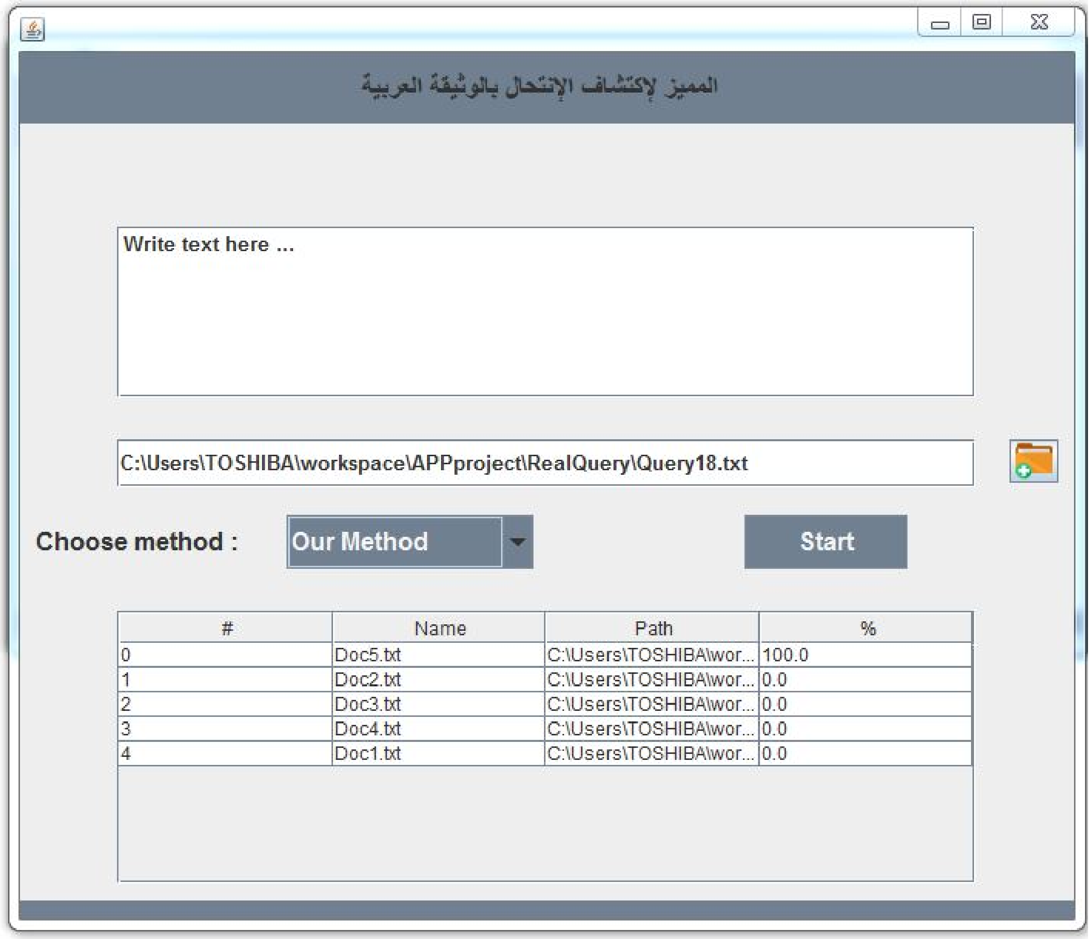
Rechercher dans : RealQuery

Nom du fichier :

Type de fichier : Tous les fichiers

FIGURE 6.57: Choosing document.

1. In case when the user chooses "Our Method" and clicks the "Start" button, the system starts calculating the similarity between the selected document and all the documents in the data set using our method and shows the result in the table(see Figure 6.58).



المميز لإكتشاف الإنتحال بالوثيقة العربية

Write text here ...

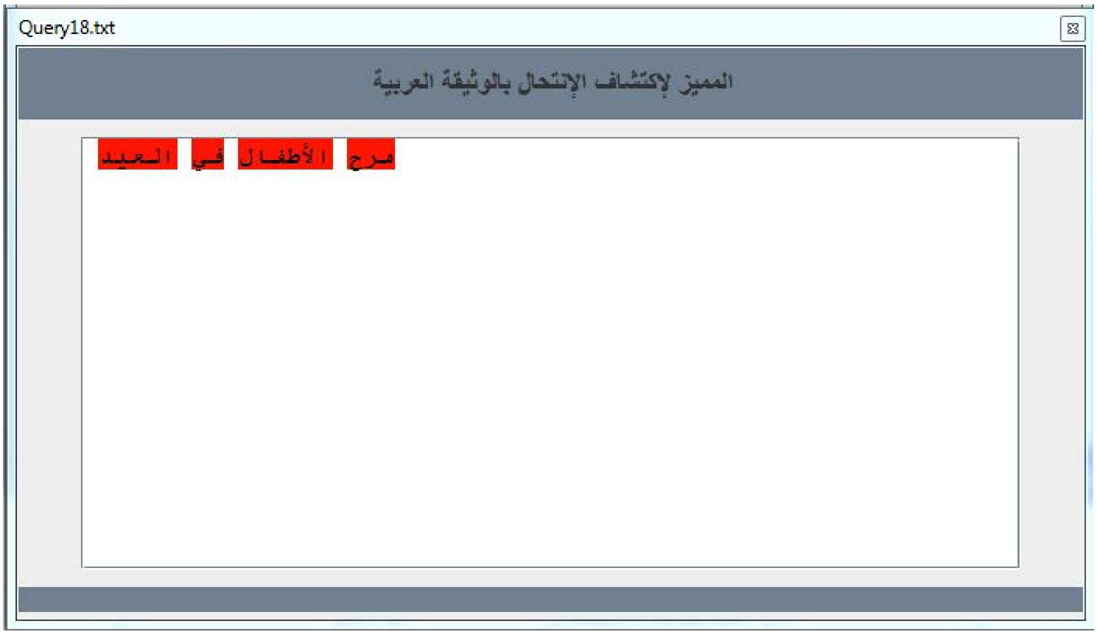
C:\Users\TOSHIBA\workspace\APPproject\RealQuery\Query18.txt

Choose method : Our Method

#	Name	Path	%
0	Doc5.txt	C:\Users\TOSHIBA\wor...	100.0
1	Doc2.txt	C:\Users\TOSHIBA\wor...	0.0
2	Doc3.txt	C:\Users\TOSHIBA\wor...	0.0
3	Doc4.txt	C:\Users\TOSHIBA\wor...	0.0
4	Doc1.txt	C:\Users\TOSHIBA\wor...	0.0

FIGURE 6.58: The example of choosing "Our Method".

Also the system displays the selected document with highlighted words that are plagiarised (Figure 6.62), and displays the most similar document with highlighted words that are mentioned in the selected document (Figure 6.63).

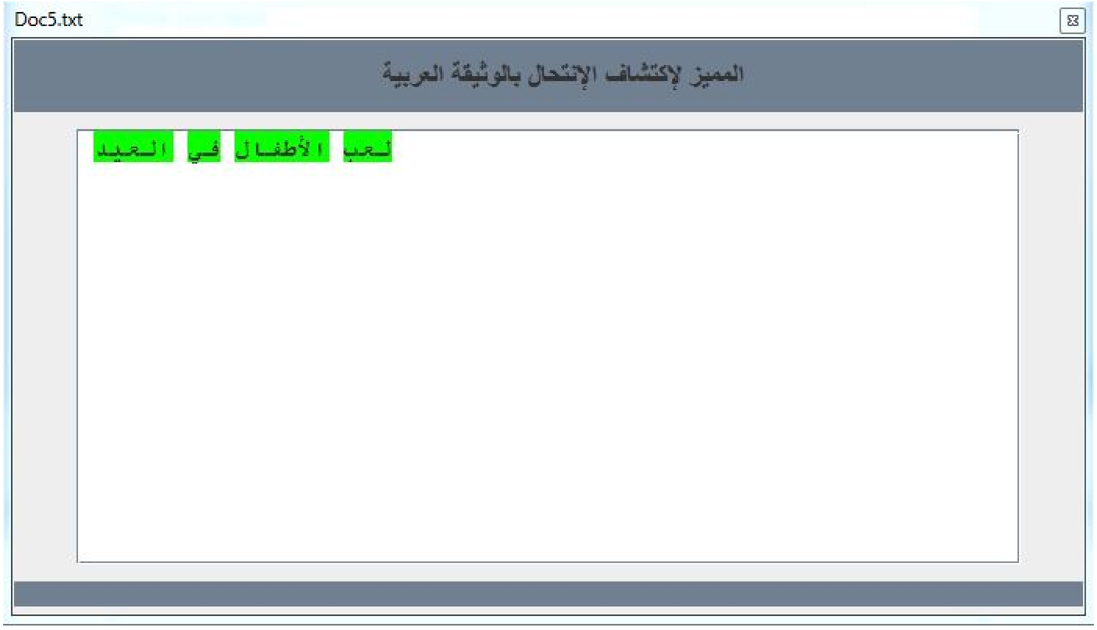


Query18.txt

المميز لإكتشاف الإنتحال بالوثيقة العربية

مرح الأطفال في العيد

FIGURE 6.59: The highlighted words in the selected document.



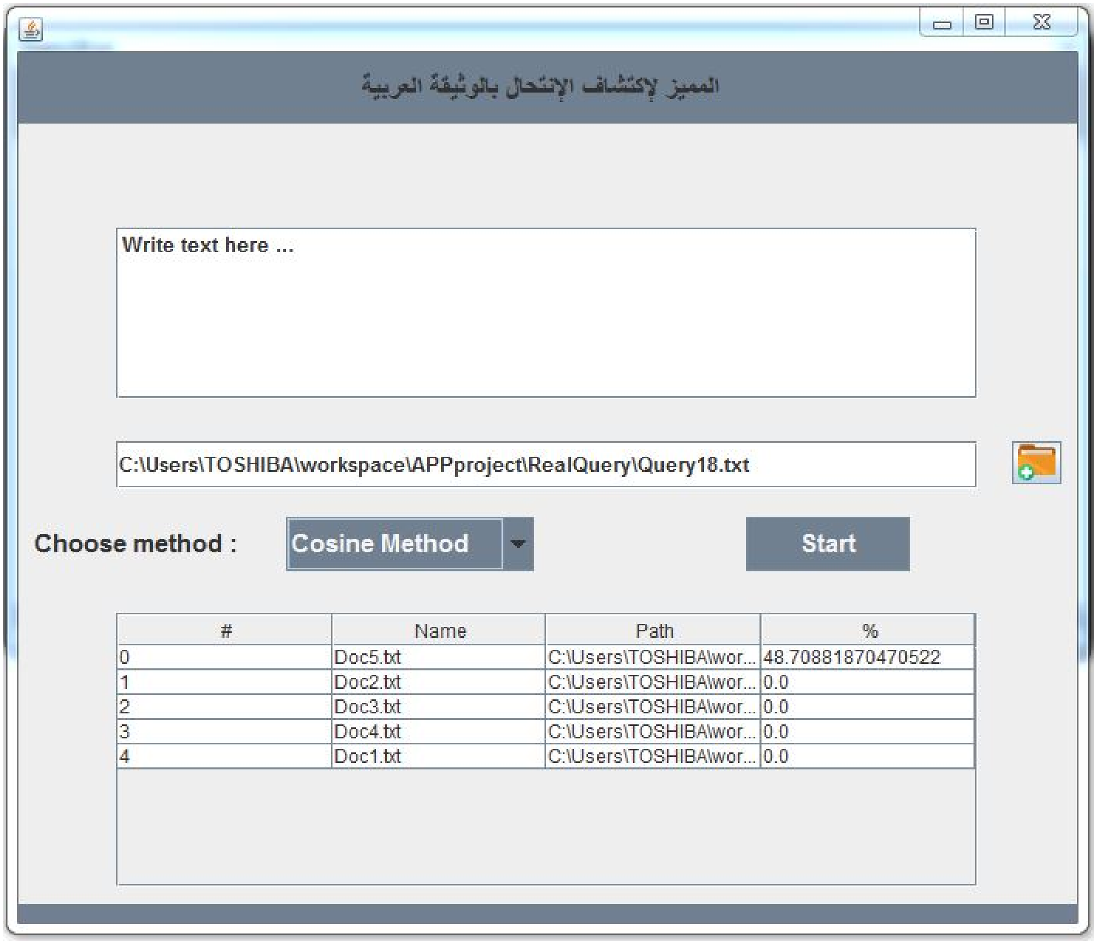
Doc5.txt

المميز لإكتشاف الإنتحال بالوثيقة العربية

لعب الأطفال في العيد

FIGURE 6.60: The highlighted words in the most similar document.

2. In case when the user chooses "Cosine Method" and clicks the "Start" button, the system starts calculating the similarity between the selected document and all the documents in the data set using the cosine method and shows the result in the table(see Figure 6.61).



المميز لإكتشاف الإنتحال بالوثيقة العربية

Write text here ...

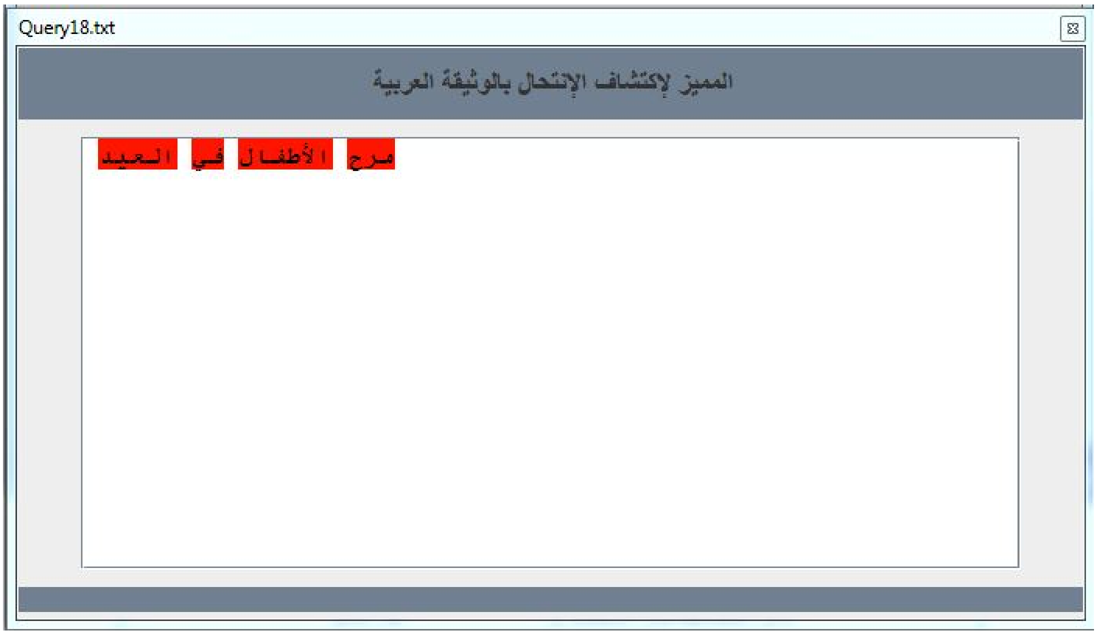
C:\Users\TOSHIBA\workspace\APPproject\RealQuery\Query18.txt

Choose method : Cosine Method

#	Name	Path	%
0	Doc5.txt	C:\Users\TOSHIBA\wor...	48.70881870470522
1	Doc2.txt	C:\Users\TOSHIBA\wor...	0.0
2	Doc3.txt	C:\Users\TOSHIBA\wor...	0.0
3	Doc4.txt	C:\Users\TOSHIBA\wor...	0.0
4	Doc1.txt	C:\Users\TOSHIBA\wor...	0.0

FIGURE 6.61: The example of choosing "Cosine Method".

Also the system displays the text of the selected document with highlighted words that are plagiarised (Figure 6.62), and displays the most similar document with highlighted words that are mentioned in the selected document (Figure 6.63).

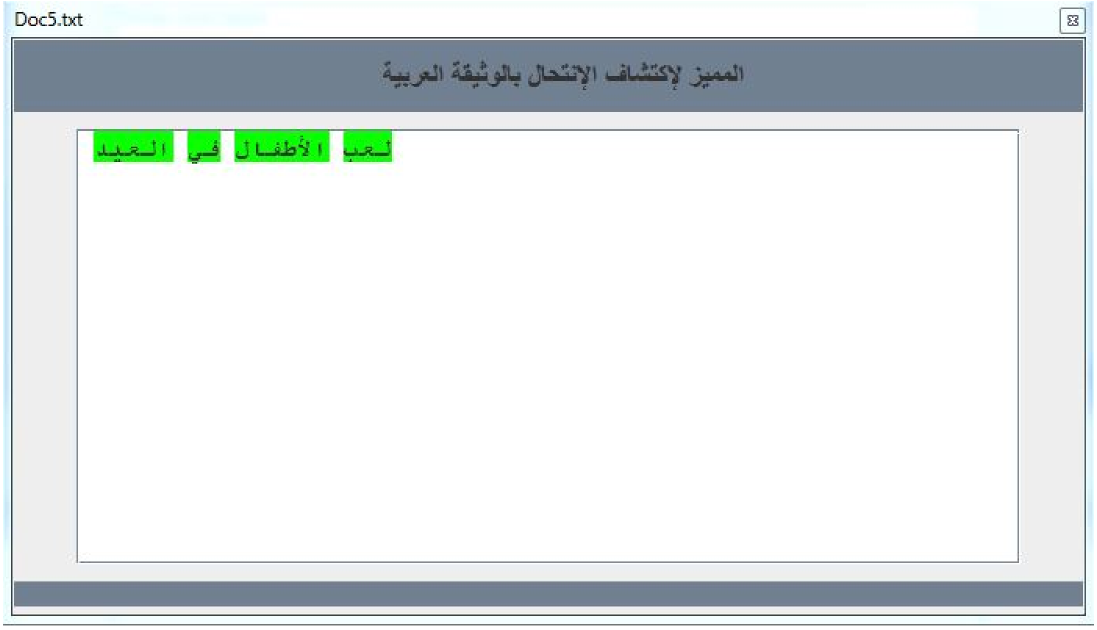


Query18.txt

المميز لإكتشاف الإنتحال بالوثيقة العربية

مرح الأطفال في العيد

FIGURE 6.62: The highlighted words in the selected document.



Doc5.txt

المميز لإكتشاف الإنتحال بالوثيقة العربية

لعب الأطفال في العيد

FIGURE 6.63: The highlighted words in the most similar document.

6.4 Evaluation and Results

In order to measure effectively the performances of our system, we have used a real data set provided by [55]. In fact, we have used a data set of 510 pairs of sentences drawn from STS 2017, which will assess the ability of our system to determine the degree of semantic similarity between monolingual sentences in Arabic language to detect the plagiarism.

The pair sentences have been manually tagged by four annotators on a scale from 0 to 5, with 0 wich indicates that the semantics of the sentences are completely independent and 5 wich signifies semantic equivalence.

The proposed system was tested for detecting plagiarism in Arabic text files. These files or documents are created from STS 2017 data set (see Table 6.1).

The number of documents	510
The total number of words	8971
The number of indexed words	5464
The number of tokens	3507

Table 6.1: Statistics of the corpus

In the following, we are about to explain our experiment for indexing Arabic documents and calculating the semantic similarity between them and the research text. In order to evaluate the proposed approach, two types of different indexation methods have been done. We will examine them separately to measure the contribution of each type to prove the performance of the PD. These indexation methods are:

- Indexing the original text (Index0): An indexing via Apache Lucene without any change of documents or queries.
- Indexing the text with synonyms (Index1): An indexing via Apache Lucene after the extraction of synonyms by using AWN ontology and the stem of each word.

In our expirementation, we have used a list of 141 queries, each one represents one of pair sentences with a score value greater than or equal to 4, which signifies the equivalent meaning to use it for detection the plagiarism in semantic similarity measurement.

The Lucene search API takes a search text (query) and returns a set of documents ranked by relevancy with documents most similar to the query having the highest score. The Table 6.2 presents the number of documents retrieved for each query .

ID Of The Query	Total Number of Documents	Number of Relevant Documents	Number of Selected Documents	Number of Relevant Selected Documents
1	510	397	40	35
2	510	304	07	07
3	510	293	53	49
4	510	335	19	17
...	...	...	...	...
141	510	203	35	34

TABLE 6.2: The number of documents returned

Then, we calculate the similarity between the search text (query) and the five most retrieved documents using our method and the cosine similarity method in order to make comparison between them. The results are presented in Table 6.4.

6.4.1 Our method algorithm

This algorithm represents our method to detect the plagiarism in Arabic dosuments.

Algorithm 1: Our Method

```
while $i \neq Nbr$ do
    $T1 \leftarrow Tokenization(Doc[i])$
    $S1 \leftarrow Stem(T1)$
    $S2 \leftarrow Synonym(S1)$
    $IndDoc[i] \leftarrow Indexer(S2)$
end while
$Q \leftarrow Tokenization(Query)$
$Sq \leftarrow Stem(Q)$
$IndQuery \leftarrow Indexer(Sq)$
while $j \neq Nbr$ do
    $Sim[j] \leftarrow OurMethod(IndQuery, IndDoc[j])$
end while
$SimilarDoc \leftarrow Max(Sim)$
```

6.4.2 Cosine method algorithm

Algorithm 2: Cosine Method

```
while $i \neq Nbr$ do
  $T1 \leftarrow Tokenization(Doc[i])$
  $S1 \leftarrow Stem(T1)$
  $S2 \leftarrow Synonym(S1)$
  $VecDoc[i] \leftarrow Vectorization(S2)$
end while
$Q \leftarrow Tokenization(Query)$
$Sq \leftarrow Stem(Q)$
$VecQuery \leftarrow Vectorization(Sq)$
while $j \neq Nbr$ do
  $Sim[j] \leftarrow Cosine(VecQuery, VecDoc[j])$
end while
$SimilarDoc \leftarrow Max(Sim)$
```

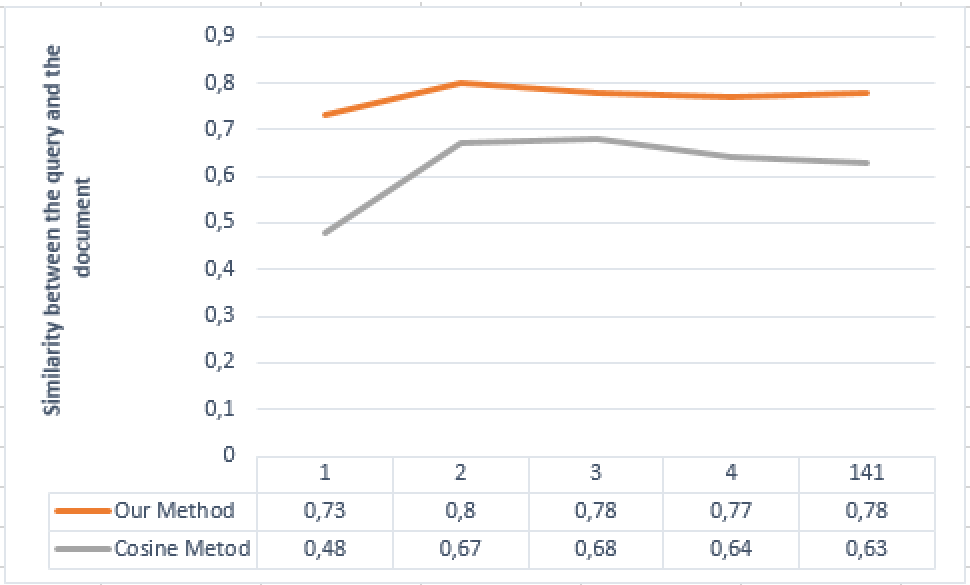
ID Of Query	Recall	Precision
1	0.088	0.87
2	0.023	1.00
3	0.167	0.92
4	0.050	0.89
...	...	...
141	0.167	0.97

Table 6.3: The precision and recall of each query

ID Of The Query	Document	Human judgement / 5	Cosine Method	Our Method
1	doc1	4/5	0.48	0.73
2	doc402	4.75/5	0.67	0.8
3	doc387	4.25/5	0.68	0.78
4	doc87	4.25/5	0.64	0.77
...	...	...	...	...
141	doc85	4.25/5	0.63	0.78

Table 6.4: Comparison between similarity methods

The results described in the Table 6.3 prove the Precision and Recall, because in all cases the most relevant document will be mentioned in the first five obtained documents in the result. Also , by simple comparison of the obtained results in Table 6.4 and the Figure 6.64 we can prove the performance of our measure in comparison with the cosine method and human judgment. We can therefore say that our measurement is better than the cosine method. This measure can be used to detect plagiarism between Arabic documents.



	1	2	3	4	141
Our Method	0,73	0,8	0,78	0,77	0,78
Cosine Metod	0,48	0,67	0,68	0,64	0,63

FIGURE 6.64: Comparison between cosine method and our method.

6.5 Conclusion

In this chapter we have presented how we have implemented our system by discribing the environments, tools and apis used to realise our system (Java, Eclipse ,AWN,Lucene,. . .). Also we have shown the interfaces of our system and we have explained how our system work.

Also, the obtained results show us that Apache Lucene with semantic representation is the best method of indexing to detect plagiarism in Arabic texts. Therefore, we have applied our method of calculating the similarity between documents using human judgement and the results proved the performance of our method compared to the cosine method.

Chapter 7

General Conclusion

In recent years, the supply of Arabic language material has grown considerably on the Internet (websites, theses, scientific articles, databases, digital documents,. . . etc.) due to the massive use of technologies of information and communications. This facilitates the search for information or ideas and helps the spread of plagiarism phenomenon in Arabic documents.

In this thesis, we have developed a new ontological approach of detecting the plagiarism in Arabic documents, this approach has been proved its force for detecting plagiarism documents. The idea of this thesis is to exploit a lexical resource (Arabic WordNet) to index the documents in order to prove the retrieval results. We have extended Lucene to prove the quality of plagiarism detection in Arabic documents using AWN ontology. This approach can reduce the time required to analyze and index documents and provides a powerful based semantic search capability for plagiarism detection.

In addition, our approach has used Apache Lucene to create the index on a dynamic collection of documents and reduces the size of the index to increase the efficiency of the search process. The performance of the proposed system is confirmed by the values of the similarity measurement scores calculated between the text and the documents to detect the plagiarism and it gives better results compared to Cosine approach.

As future work, we will focus on constructing a new ontology with a large number of classes, properties and individuals for supporting a large number of Arabic terms and concepts. Despite the growing interest in the field of semantic similarity measurement, we are also interested to build an adequate data set with human judgements of experts for the Arabic language, especially when changing the words' sense in the sentence level by using synonyms or homonyms to use it in the evaluation phase.

Also, we are interested on constructing new root extracting tools of the Arabic word .

Bibliography

[1] Example of ontology. URL https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcSRzxV93VPvjxnD1AFu0rUymQEf2Tyvjr3pd6DOetso3Q_MiAVoBQ.[Visitedat29/01/2019].

[2] Michal Sir & Zdenek Bradac & Petr Fiedler. Ontology versus database.

[3] Life cycle of ontology. URL https://www.researchgate.net/figure/The-Thesaurus-within-the-Ontology-Life-Cycle-Solid-.[Visitedat29/01/2019].

[4] Java. URL https://en.wikipedia.org/wiki/Java_%28programming_language%29.[Visitedat08/04/2019].

[5] Eclipse. URL https://fr.wikipedia.org/wiki/Eclipse_(projet).[Visitedat09/04/2019].

[6] Lucene. URL https://lucene.apache.org.[Visitedat09/04/2019].

[7] N. Guarino and P. Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. in n. mars, editor, towards very large knowledge bases: Knowledge building and knowledge sharing, pages 25–32. ios press, amsterdam. Dummy Publisher, 1995.

[8] T. R. Gruber. A translation approach to portable ontologies.knowledge acquisition, 5(2):199–220. 1993.

[9] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. international journal of human computer studies ,43(5–6): 907–928. 1995.

[10] Les languages de description des ontologies. URL https://www.researchgate.net/publication/324209660_Les_langages_de_description_des_ontologies_RDF_OWL.[Visitedat16/02/2019].

[11] Prof. Dr. Christiane Floyd Dr. Carola Eschenbach. Design and implementation of an ontology for knowledge assessment. 04/04/2005.

[12] Base de données. URL https://whatis.techtarget.com/fr/definition/Base-de-donnees/[Visited:07/01/2019].

[13] Mackworth A Poote, D. &amp. Unique names assumption.artificial inteligence: Foundations of computal agents [online]. URL http://artint.info/html/ArtInt_302.html[Visited:09/01/2019].

[14] L. 2009 Dutra. Closed-world assumption.wikipedia: the free encyclopedia [online]. URL http://en.wikipedia.org/wiki/Closed-world_assumption.[Visited:11/01/2019].

[15] C. 2006 Matthews. Open-world assumption. wikipedia: the free encyclopedia [online]. URL http://en.wikipedia.org/wiki/Open-world_assumption.[Visited:12/04/2019].

[16] Ontology Design Patterns .30 May 2014. URL http://ontologydesignpatterns.org/wiki/.[Visited:15/01/2019].

[17] Specification. URL https://en.wikipedia.org/wiki/Specification_(technical_standard).[Visitedat05/02/2019].

[18] Conceptualization. URL https://educalingo.com/en/dic-en/conceptualization.[Visitedat14/02/2019].

[19] Implementation. URL https://en.wikipedia.org/wiki/Implementation.[Visitedat14/02/2019].

[20] Maintenance. URL https://en.wikipedia.org/wiki/Maintenance_(technical).[Visitedat05/02/2019].

[21] D. 2000 Noy, N. amp; McGuinness. Ontology development 101: A guide to creating your first ontology. knowledge systems laboratory stanford university [online]. URL http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html.[Visited:17/01/2019].

[22] What is plagiarism. URL https://www.Turnitin.comandResearchResources.Whatisplagiarism.pdf.[Visited:20/01/2019].

[23] Plagiarize in Merriam dictionary. URL https://www.merriam-webster.com/dictionary/plagiarize.[Visited:20/01/2019].

[24] Propriété intellectuelle. URL https://fr.wikipedia.org/wiki/Propri%C3%A9t%C3%A9_intellectuele.[Visitedat30/02/2019].

[25] Singapore Polytechnic (Copyright laws). URL http://www.sp.edu.sg/departments/asd/hk_1261.htm.[Visited:20/01/2019].

[26] Joint Information Systems Committee (JISC) (2003). Why do students plagiarise?[online]. URL http://www.jisc.ac.uk/index.cfm?name=plagiarism_why.[Visited:12/06/2004].

[27] B. (1996) Duguid. The unacceptable face of plagiarism [online]. URL http://media.hyperreal.org/zines/est/articles/plagiari.html.[Visited:12/06/2004].

[28] chichi.S Bensalem. i=I, Rosso.P. A new courpus of the evaluation of arabic intrinsic plagiarism detection. CLEF2013, Valencia- Spain. Clef2013. clef-initiative.eu/ diapositive1-clef2013.pdf, p10.

[29] M. El Bachir Menai. Detection of plagiarism in arabic documents. I.J. Information Technology and Computer Science, 2012, 10, 80-89 .Published Online September 2012 in MECS. URL http://www.mecs-press.org/.

[30] A Survey Of Plagiarism Detection Methods Information Technology Essay 11 2013. URL https://www.researchgate.net/publication/311114973_A_SURVEY_OF_PLAGIARISM_DETECTION_FOR_ARABIC_DOCUMENTS.[Visitedat08/01/2019].

[31] Dumais S.T. Latent semantic analysis [j]. annual review of information science and technology, 2005: 38-188, doi:10.1002/aris. 1440380105.

[32] Singular value decomposition. URL https://en.wikipedia.org/wiki/Singular_value_decomposition.[Visitedat20/03/2019].

[33] M. Montes-y L. Villase P. Rosso F. Sanchez-Vega, E. Villatoro-Tello. Determining and characterizing the reused text for plagiarism detection. Contents lists available at SciVerse ScienceDirect , Expert Systems with Applications 40 (2013) 1804–1813.

[34] Randa. K. A plagiarism detection tool for arabic text document. Thesis of Master Degree ,Sudan University of Science and Technology , 2010, Sudan.

[35] Garcia-Molina H. SCAM Shivakumar N. A copy detection mechanism for digital documents [c]. in: Proceedings of the 2nd international conference on theory and practice of digital libraries. Austin,Texas, USA, June 1995.

[36] M. Bagais M. Menai. Aplag: A plagiarism checker for arabic texts. The 6th International Conference on Computer Science & Education (ICCSE 2011) , IEEE 2011.

[37] Plagiarism. What is plagiarism ? URL http://www.plagiarism.org/plagiarism-101/what-is-plagiarism.[Visitedat14/03/2019].

[38] Salha Mohammed Alzahrani and Naomie Salim. Plagiarism detection in arabic scripts using fuzzy information retrieval. Proceedings of 2008 Student Conference on Research and Development (SCOReD 2008), 26-27 Nov. 2008, Johor, Malaysia.

[39] N. Salim S. M. Alzahrani. Statement-based fuzzy-set ir versus fingerprints matching for plagiarism detection in arabic documents. In Proc. of the 5th Postgraduate Annual Research Seminar (PARS09), Johor Bahru, Malaysia, 2009.

[40] M. Dashash K. Omar, B. Alkhatib. The implementation of plagiarism detection system in health sciences publications in arabic and english languages. International Review on Computers and Software (I.RE.CO.S.), Vol. 8, N. 4 ISSN 1828-6003 April 2013.

[41] Ikdam AlHami Izzat Alsmadi and Saif Kazakzeh. Issues related to the detection of source code plagiarism in students assignments. International Journal of Software Engineering and Its Applications Vol.8, No.4 (2014), pp.23-34.

[42] Kamal Mansoor Jambi Imtiaz Hussain Khan, Muazzam Ahmed Siddiqui and Abobakr Ahmed Bagais. A framework for plagiarism detection in arabic documents. Dhinaharan Nagamalai et al. (Eds) : CSEA, DKMP, AIFU, SEA – 2015 pp. 01–09, 2015. CS & IT-CSCP 2015.

[43] A.Jadalla and A.Elnagar. A plagiarism detection system for arabic text-based documents. 2012,springer-Verlag Berlin Heidelberg, PAISI 2012,LNCS 7299,pp. 145153, 2012.

[44] Elizabeth D. Liddy. Natural language processing. In Encyclopedia of Library and Information Science, 2nd edition. NEW YORK. Marcel Decker, Inc. Syracuse University,2001.

[45] LAMRAOUI. Y. Recherche intelligente des informations dans le coran. Graduation thesis. AbouBakr Belkaid University – Tlemcen 2010/2011.

[46] Shereen Khoja. URL http://zeus.cs.pacificu.edu/shereen/research.htm.[Visitedat24/03/2019].

[47] S G.R. Khoja. Stemming arabic text. 1999, computing Department,Lancaster University, Lancaster. URL https://www.sciencedirect.com/science/article/pii/S1319157815000166.

[48] Part of speech. URL https://en.wikipedia.org/wiki/Part_of_speech.[Visitedat26/03/2019].

[49] K. M. Jambi I. H. Khan, M. A. Siddiqui and A. A. Bagais. A framework for plagiarism detection in arabic documents. 2015, dhinaharan Nagamalai et al. (Eds) : CCSEA, DKMP, AIFU, SEA – 2015 pp. 0109.

[50] H. Rodriguez M. Alkhalifa P. Vossen A. Pease S. Elkateb, W. Black and C. Fellbaum. Building a wordnet for arabic. 2006, in Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

[51] S. Addagada. Indexing and searching document collections using lucene. 2007, university of New Orleans Theses and Dissertations.1070. URL https://scholarworks.uno.edu/td/1070.

[52] HLAOUA. L. Reformulation de requêtes par réinjection de pertinence dans les documents semi-structurés. PhD thesis in computer science carried out at Paul Sabatier University, 2007.

[53] Java definition. URL https://www.theserverside.com/definition/Java.[Visitedat08/04/2019].

[54] Mohammad Ghandi Aldiery. The semantic similarity measures using arabic ontology. January, 2017.

[55] Semeval-2017 task 1. Semantic textual similarity. URL http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools.[Visitedat27/02/2019].