



A text uniqueness checking system for Armenian Language

Gohar Tomeyan

Final project report submitted to the School of Technology and Management of Bragança
to obtain the degree of Master in Information Systems

Relatório Final do Trabalho de Projecto de Engenharia Informática apresentado à
Escola Superior de Tecnologia e de Gestão do Instituto Politécnico de Bragança.

Supervised by/Trabalho orientado por:

Prof. Maria João Varanda Pereira

This dissertation does not include the criticisms and suggestions made by the Jury.

Esta dissertação não inclui as críticas e sugestões feitas pelo Júri.

Bragança

2017



A text uniqueness checking system for Armenian Language

Gohar Tomeyan

E-mail: goharikyan93@gmail.com

a35480@alunos.ipb.pt

Final project report submitted to the School of Technology and Management of Bragança
to obtain the degree of Master in Information Systems

Relatório Final do Trabalho de Projecto de Engenharia Informática apresentado à
Escola Superior de Tecnologia e de Gestão do Instituto Politécnico de Bragança.

Supervised by/Trabalho orientado por:

Prof. Maria João Varanda Pereira

Polytechnic Institute of Bragança

Camus de Santa Apolonia

5301-854 Bragança, Portugal

June

2017

Acknowledgements

I would like to start by thanking my supervisor Professor Maria João Varanda Pereira for providing me with the opportunity to work in this research and for her encouragement, support, and supervision at all levels.

I am grateful to Professor Gevorg Margarov for all his help and guidance from Armenia.

I want to express my appreciation to my mother and sister for the Armenian language assistance provided.

I would like to thank my friends from Armenia for tested system.

Finally, a special thanks to Erasmus+ ICM project for supporting the research collaboration between IPB and NPUA.

THANK YOU ALL!

Abstract

The goal of this dissertation is to develop a tool to analyze the similarity of Armenian texts. The idea is to compare two texts or to compare a text with a set of texts and detect the possibility of plagiarism. This system will be used in academic contexts but can also be useful in other situations. In the academic context it is very important to evaluate the uniqueness of reports, scientific papers and other documents that are everyday disseminated on the web. There are already several tools with this purpose but not for Armenian texts.

Resumo

O objetivo desta dissertação é desenvolver uma ferramenta para analisar a semelhança de textos em arménio. A ideia é comparar dois textos ou comparar um texto com um conjunto de textos e detectar a possibilidade de plágio. Este sistema poderá ser usado em contextos académicos, mas, também pode ser útil em outras situações. No contexto académico, é muito importante avaliar a singularidade de relatórios, artigos científicos e outros documentos que são todos os dias divulgados na web. Já existem várias ferramentas com este propósito mas não para a linguagem arménia.

Նամառոտագիր

Տեղեկարկական տեխնոլոգիաների զարգացմանը զուգընթաց ավելացել են նաև գրագողության դեպքերը: Նաշվի առնելով այն հանգամանքը, որ կան գրագողությունը ստուգող մի շարք համակարգեր, բայց ոչ մի համակարգ նախատեսված չէ հայերեն տեքստերի ունիկալության վերլուծություն համար, խնդիր դրվեց մշակել այնպիսի համակարգ, որը կապահովի տեղեկարկական համակարգերում տեքստերի ունիկալության վերլուծությունը, ինչպես նաև թույլ կտա համեմատել և հայտնաբերել գրագողության առկայությունը: Աշխատանքի նպատակն է ուսումնական գործընթացում ունիկալությունը ստուգող համակարգերի կիրառումը, քանի որ շատ կարևոր է գնահատել ավտոմատությունների, ռեֆերատների, կուրսային աշխատանքների և այլ տեքստերի ունիկալության աստիճանը: Այս նախագիծը հնարավորություն կտա մշակել և հիմնավորել հայերեն տեքստերի ունիկալության համակարգչային վերլուծությունը և կանխել գրագողությունը հայերենում:

Contents

Acknowledgements	v
Abstract	vii
Resumo	ix
Նամանորագիր	xi
Acronyms	xvii
1 Introduction	1
1.1 Motivation	4
1.2 Work purpose	4
2 Related work	7
2.1 Antiplagiat.ru	7
2.2 ETXT –Antiplagiat	8
2.3 PlagScan	9
2.4 Turnitin.com	10
2.5 The created system	11
3 The main solutions	13
3.1 Natural Language Processing	14
3.2 Levels of plagiarism detection system	15
3.3 Algorithms and methods of plagiarism detection	17

3.4	System Implementation	18
4	Created System	21
4.1	Normalization alphabet	21
4.2	Finding the same text	22
4.3	Choosing keywords domain	25
4.4	Stop words removal	27
4.5	Stemming	28
4.6	Synonymizer	30
4.7	Finding plagiarism with translation	33
5	Web Application	37
5.1	Web Application development	37
6	Tests	43
7	Conclusion and Future work	47
7.1	Conclusion	47
7.2	Future work	48
	Bibliography	49
A	User Guide	i
A.1	Set keyword domain and comparing	i
A.2	Comparing two File	iii
A.3	Set Database	iv
A.4	Translator	v

List of Figures

2.1	Comparative analysis	11
3.1	Steps of plagiarism detection	16
3.2	Structure of database	20
3.3	Structure of database	20
4.1	Normalization alphabet	22
4.2	Comparing two or more documents	23
4.3	Comparing two documents	24
4.4	Keyword generation	26
4.5	Most common stop words	27
4.6	Stop words removal process	28
4.7	Stemming example	29
4.8	Stemming	30
4.9	Before synonymize	31
4.10	Replacement with synonyms	32
4.11	Adding synonyms and explanation	32
4.12	Google Translate	34
4.13	Translation	35
5.1	Main page	38
5.2	Registration page for the teacher	39
5.3	Log In Page	40
5.4	After Log In	40

5.5	Download	41
5.6	Contact page	41
6.1	Plagiarism detection by percentage	44
A.1	Comparison of keywords	i
A.2	Comparison of many documents	ii
A.3	Edit and see the keywords	ii
A.4	Comparison of two documents	iii
A.5	Using stemming and stop word removal	iv
A.6	Add synonyms	iv
A.7	Translator	v

Acronyms

NLP Natural Language Processing

ASCII American Standard Code for Information Interchange

TF Term Frequency

TF-IDF Term Frequency-Inverse Document Frequency

MVC Model View Controller

MSSQL Microsoft Structured Query Language Server

SQL Structured Query Language

HTML Hypertext Markup Language

RESTful Representational state transfer

SEO Search Engine Optimization

Chapter 1

Introduction

Internet is getting more widespread in our life and in our activities. However, visiting different Web sites, we see that all the found articles or other materials are very similar. Besides, there are many thesis, term papers, research and other scientific works on the Internet. If formerly it was necessary for the students to take advantage of the published books and literature, now it is enough to write the name of the subject in the search engines and we can find thousands of items.

The most common objects of plagiarism are texts, separate expressions, thoughts, inventions, facts described in novels. Scientific spheres include a large amount of ready works, course works and articles, in which we can make several changes and achieve results. That kind of change is called plagiarism. In order to avoid similar situations there is a use of plagiarism detection system [15].

At first, what does plagiarism mean? There are many definitions of plagiarism. The scientific and educational sphere plagiarism is the form of deception, which means to appropriate other ideas, passages from another work or author. This is a forgery that implies the violation of copyright laws. Plagiarism is a steal and pass off the ideas of another as one's own, using another's manufacture without lending the source, present as a new and original an idea taking from an existing source [16, 13]. At the legal point, plagiarism is a direct privatization of the text. Legally the plagiarism is text digestion, while the digestion of subjects and ideas can't be considered as plagiarism. The only

thing, which is not allowed, is the whole copy of the text. But often the whole text is translated and presented as an original. Thus, the plagiarism which is done by translation is widespread.

Usually in order to conceal the plagiarism people carry out several steps, for example text morphological change, lexical change, reduction of the text up to some words, sentences, pictures or formulas, text syntactical changes, movement of the sentences, punctuation marks change, spaces are replaced with transparent letters, and also create and use synonyms.

There are two more important types of plagiarism:

- Textual plagiarisms are passing off another's texts as own, and usually are done by students and researchers in academic research works, where documents are similar to the original documents, reports, scientific papers and art design.
- A source code plagiarism also, is done by computer science students. The students try to copy all or parts of source code written by someone else as one's own. These types of plagiarism is difficult to detect because the Internet has stored many source codes and students can copied the source code, and execute several steps, for example just changing the name of the methods, variables, values and achieve results.

Plagiarism detection and prevention became one of the educational problems in some universities, schools and institutions, because most of the students or researchers use another's ideas without pointing the source. That is why lots of resources can be found on the Internet. It is so easy for them to use one of the search engines to search for any topic and to copy from it without mention the document owner. So all academic fields must use plagiarism detection softwares to stop or to eliminate students cheating, copying and modifying documents. If they know that they will be found, they will stop copying.

Some types of plagiarism can be detected easily by using some of the recent plagiarism detection tools, which is available on the Internet. However for some of the expert plagiarists who use some the anti-plagiarism softwares, there is necessity of more efforts to detect the plagiarism. Plagiarism is practiced not only by student but also there are

some staff members who like to publish papers in which some parts are directly copied or partially modified in order to increase the number of publications.[14]

There is a big number of checking softwares for plagiarism detection but still they have some limitations as they cannot prove plagiarism. They show evidences that the documents has been plagiarized from another document or sources, it only shows the similarity and give hints to some other documents. That is why if the paper is published globally in some international journal, but the universities research centers still do not take any action against plagiarism detection that help people to cheat more and more.

Copyrights and legal aspects for use of published documents also can be covered by using plagiarism software, so it can show whether this person has legally or illegally copied the documents or not and it also show the whether this person has permission from the owner to use this document or not. Plagiarism detection is also one of the most important issues to journals, research center and conferences; they are using advanced plagiarism detection tools to ensure that all the documents have not been plagiarized, and to save the copyrights from violation for the publishers.[14]

The classification of the text plagiarism detection methods for natural language depends on the: complexity and number of documents processed by the used methods. The complexity of the used method can divided in a two types:

- Superficial (word by word): realized without any knowledge of the linguistic rules and a document structure or language.
- Structural: when words can be replaced by a synonym or the verbs are used in another tense or the text is in another language.

Next type is the number of documents processed by the used methods. This type can divided in a four types:

- Singular: this type usually is used to calculate the similarity of the documents.
- Paired: Two documents are processed together to compute the frequency.
- Multidimensional: N documents from a corpus are processed together to compute.

- Corpal: All documents contained in a corpus are processed together to compute the frequency.

Therefore, the research questions of this dissertation are: It is possible to construct a plagiarism detection system for Armenian Language? Is it possible to detect more than one plagiarism levels?

1.1 Motivation

The cases of plagiarism have raised along with the development of information technologies. Students often present ready works as their own. There are many plagiarism detecting systems. Existing plagiarism detecting multilingual systems are not intended for Armenian language.

The main goal of this work is to construct a system to detect plagiarism in Armenian texts. With the ever-increasing availability and accessibility of the Internet, students are able to access a multitude of resources in support of their studies. However, this has also led to an increase in their ability to cheat through plagiarizing text and claiming it as their own. To avoid such situations, it was decided to develop a system that will automate the uniqueness analysis of the work done by students in the learning process and will allow teachers to quickly detect the presence of plagiarism. And because there is not this kind of systems for Armenian language, this one will be used in lot of universities and will be useful for lecturers.

1.2 Work purpose

In order to analyze the similarity between two documents it is necessary to use some techniques of natural language processing [31, 24]. The first step will be to compare the texts word by word but this work must go further. Everyone knows that the people that use the texts of other people change it a little bit to dissimulate the plagiarism. So, one of the most important part of this work is to define plagiarism levels and what must

be checked in each level. Then, construct a tool implementing those plagiarism levels detection for document written in Armenian language.

Chapter 2

Related work

There are many automatic systems to detect plagiarism, such systems are Antiplagiat.ru, eTXT, PlagScan, CheckforPlagiarism.net [26, 14], Turnitin, etc. Here we describe as comparison analysis of some textual softwares and discuss about each program separately.

Analyzing the main characteristics of these tools we decided to use six parameters to be evaluated in each tool: if the tool checks in database or in the web or both, if the tool checks the possibility of use synonyms or sentence structure modification, if the tool allows multiple document comparison, the supported languages and the possibility to detect plagiarism through translation.

2.1 Antiplagiat.ru

One of the famous online service is Antiplagiat, [2, 1] used for textual plagiarism detection. The system used in the universities of Russia. The main features of Antiplagiat are:

- Checking in the database: the system searches from its own database for plagiarisms of students academic documents and analyze them. There is a limit up to 3000-5000 words for free version.

- Checking on the Internet: it isn't able to search on the Internet.
- Checking of the synonyms and sentence structure: the system Antiplagiat is able to detect, reduced, and replaced words, sentences and paragraphs. But system doesn't able to recognize the synonyms replacement and system doesn't detect text morphological changes. If spaces are replaced by transparent letters, they will be visible to computer. The replacement of English letters with Russian is also detected. The change of punctuation marks has no influence on the work of the system.
- Multiple Document Comparison: Antiplagiat offers can compare of multiple documents.
- Supported Languages: supports Russian as primary language.
- Plagiarism with translation: it isn't supported.

2.2 ETXT –Antiplagiat

ETXT–Antiplagiat [7, 3] has a system and online server. The main features of ETXT–Antiplagiat are:

- Checking in the database: the system searches from its own database for plagiarisms of students academic documents and analyze them.
- Checking on the Internet: the system gives the opportunity to search on the Internet.
- Checking of the synonyms and sentence structure: the system ETXT is able to detect, reduce and replace words, sentences and paragraphs. ETXT-Antiplagiat does not support synonym and sentence structure checking. Matching parts of the text will be indicated with the respective colors by system. The program can easily detect non-unique texts. To avoid the system students need to make changes in the text and use synonyms, etc.
- Multiple document comparison: ETXT can compare multiple documents.

- Supported languages: many languages.
- Plagiarism with translation: it isn't supported.

2.3 PlagScan

PlagScan [19] is a (available online and on-premises) plagiarism detection software, used by academic institutions and businesses. PlagScan servers teachers and professors to identify plagiarism and educate students on the appropriate usage of sources in academic works as well as protecting copyrights of texts [14]. The main features of PlagScan are:

- Checking in the database: PlagScan compares submitted texts with billions of documents on the internal archives. The accessible search index expands on a daily basis. Each year, PlagScan checks several million documents for plagiarism.
- Checking on the Internet: the system gives an opportunity to search on the web sources. Users can either register as single user, or as an organization, which enables further setting options. After scanning a submitted text for plagiarism, the software provides the user with a detailed report, indicating potential plagiarism and listing the sources of similarities PlagScan compares your document with billions of others and highlights relevant correlations between them. It enables our users to identify fraudsters by checking documents with own online plagiarism detection platform in real time.
- Checking of the synonyms and sentence structure: PlagScan doesn't support synonyms recognition, and checking structure of sentences.
- Multiple Document Comparison: PlagScan can compare multiple documents.
- Supported Languages: PlagScan supports all the language that use Latin or Arabic characters.
- Plagiarism with translation: it isn't supported.

2.4 Turnitin.com

Turnitin [20] is the most famous online service for plagiarism prevention on over the world. The system is used in ten thousand institutions in 126 countries, and many teachers and students are actively using this program for checking originality of texts [18]. The main features of Turnitin are:

- Checking in the database: Turnitin has an ever-growing database for checking documents stored in own database, where students and teachers submit documents for checking. That excludes the possibility to copy texts from other students of the last years.
- Checking on the Internet: the system can find the materials on the web sources.
- Multiple document comparison: Turnitin.com can compare multiple documents.
- Supported Languages: the papers can be submitted to Turnitin in the 30 languages: for example Chinese, Japanese, Czech, Danish, Finnish, French, German, Hungarian, Italian, Polish, Portuguese, Romanian, Serbian, Slovak, Slovenian, Spanish, Swedish, Arabic, Greek, Hebrew, Farsi, Russian, and Turkish.
- Plagiarism with translation: if the papers submitted in Portuguese language the program first will check the text only in with Portuguese database, which only support Portuguese language research works, term papers, etc., and it will check just Portuguese language web pages. If a paper is in English translated from Portuguese, then it will be checked again in the English database.

In addition to, critics have claim that use of Turnitin plagiarism detection software violates educational privacy and intellectual property copyright laws.

2.5 The created system

In this Figure 2.1 we are describing the disadvantages and advantages of the analyzed tools. For example, all tools did not use a synonymizer, and only one tool can found the plagiarism with translation (only from English texts), that is the Turnitin. Therefore, to avoid such situations it was decided to develop a system that will automate the uniqueness analysis of the work done by students in the learning process and will allow teachers to detect quickly the presence of plagiarism.

Name	Compare in database	Compare on the Internet	Languages	Translation	Synonymize
Antiplagiat.ru	+	-	Russian	-	-
ETXT-Antiplagiat	+	+	Many	-	-
PlagScan	+	+	Germany French English Spain	-	-
Turnitin.com	+	+	Many	+	-

Figure 2.1: Comparative analysis

And because there is not this kind of systems for Armenian language, this one will be available for lot of universities and will be useful for lecturers.

The main features of the proposed tool are:

- Checking in the database: the program will check the students papers in the own database, where each year will be uploaded the research works done by student and all works will be stored in database.
- Checking on the Internet: the program doesn't give the opportunity to search on

the web sources, but teachers can upload documents based on the web sources, and prevent the plagiarism based on the Internet.

- Checking of the synonyms and sentence structure: Program will allow these steps: normalization alphabet, choosing keyword domain, stop word removal, stemming, which will be use to search the correct forms of words and replace with the synonyms. Our system will also support synonym recognition.
- Multiple Document Comparison: Our system will compare one document with more documents and will show the percentage of plagiarism possibility.
- Supported Languages: Armenian.
- Plagiarism with translation: The program will detect Russian and English text translations, and will compare with Armenian sources. The translation is based on the Google translator.

Chapter 3

The main solutions

To achieve the assigned goal it is necessary to solve the following tasks:

1. review the existing algorithms for detecting plagiarism in the texts,
2. review existing methods to conceal the fact of plagiarism, as well as methods of dealing with them,
3. develop a method of searching plagiarism in Armenian texts that is resistant to modifications when detecting,
4. define different plagiarism levels,
5. create a software tool based on the developed methods, which provides plagiarism detection with the possibility of visualizing the borrowed pieces of text in the scanned document and in the source document and a percentage calculation.

In this chapter, we will describe the main steps of our tools. We've already researched some tools for plagiarism detection, but not for Armenian texts.

The main aim of this dissertation is to use of Natural Language Processing (NLP) (Natural Language Processing) techniques [25] for detection the possibility of plagiarism in Armenian texts.

Armenian language belongs to Indo-European languages. Armenian alphabet was created by Mesrop Mashtoc in 405 ADs, the main purpose was to translate the Bible in

Armenian, because the Bible was not yet accessible in the native language. The Armenian language has 39 letters and 36 phonemes. The state language of the Republic of Armenia is Armenian. The Armenian language consists of two dialects Western Armenian and Eastern Armenian. Armenian diaspora around the world speak in Western Armenian. As a diasporic language, Western Armenian is not an official language of the country unlike Eastern Armenian.

Since this dissertation deals with a plagiarism detection system for Armenian texts, the program involves the main features of the Armenian language. The Armenian language has difficult structure. For example Armenian has 7 cases while English has only 5. Unlike English in Armenian with every case the noun ending is going to change, and etc.

Armenian has a unique writing system.

3.1 Natural Language Processing

First the program must involves the use of Natural Language Processing techniques[31], and not only for usually comparing texts word by word but also to detected rewritten texts. Natural Language Processing includes semantic and syntactic changes, stop word removal, stemming, lemmatization, punctuation removal and etc., as part of the pre-processing stage [34].

If the text has semantic and syntactic changes, the plagiarism detection systems do not work well. In order to detect such changes, linguistic techniques must be considered. The already existing systems for plagiarism detection show if the text is unique or not, but programs couldn't detect intelligent plagiarism, when ideas are presented in different words, replacement with synonyms, translation, etc.

Translation plagiarism is very extended, because students can also translate the text from one language to another without pointing the original source. For example we haven't many materials about Information Systems in Armenian language and students carry out translation from English or Russian texts including automatic translation (for example Google and another translators) and manual translation (which can be done by

students who knows some languages).

3.2 Levels of plagiarism detection system

Possible modifications of the text plagiarism depends on the language used, and during the analysis of the text, we should take into account the specifics of the given language. Dependencies are manifested in the difference between the rules of sentence structure and language opportunities to translate to other words of the same meaning. Detecting plagiarism should be made by possible modifications when detecting, and the system must be able to allocate specific pieces of borrowed text, as well as the corresponding fragments of the source text.

In order to process an algorithm, it is important to determine two aspects:

- standards of determining the similarity of texts (form and content),
- determining the level of similarity and its threshold value (when the text isn't a copy)

Technical uniqueness of text is a threshold value, which is usually measuring by percentage and show if the text have duplications or not. The text that has an 100% technical uniqueness, is not unique yet (de facto it can be unique also from about 0). For example, write off the thought of another person, and that is not unique, measuring by other words without pointing original source. However, there are some exceptions too, factual unique texts can be technical unique for 50%. For example, the author's work is unique, when includes exceptional materials that are written from 0. A work is not unique when it includes citations, expressions, technical terminus and etc.

The main steps of this work the find sentences exactly the same, normalization alphabet, choosing keywords domain, stop-word removal, stemming, synonym recognition and find plagiarism with translation. The steps of the plagiarism detection tools you can see in Figure 3.1. Stop-word removal it is like remove common words. Stop-words is a very frequent words that we are using. The usual way of determining what counts as

a stop-word is just to use a dictionary that lists them [34]. The motivation for using synonymy recognition comes from considering human behavior, whereby people may seek to hide plagiarism by replacing words with appropriate synonyms. The system must be able to detect similar changes, and also contain synonyms and stemmers for the Armenian language.

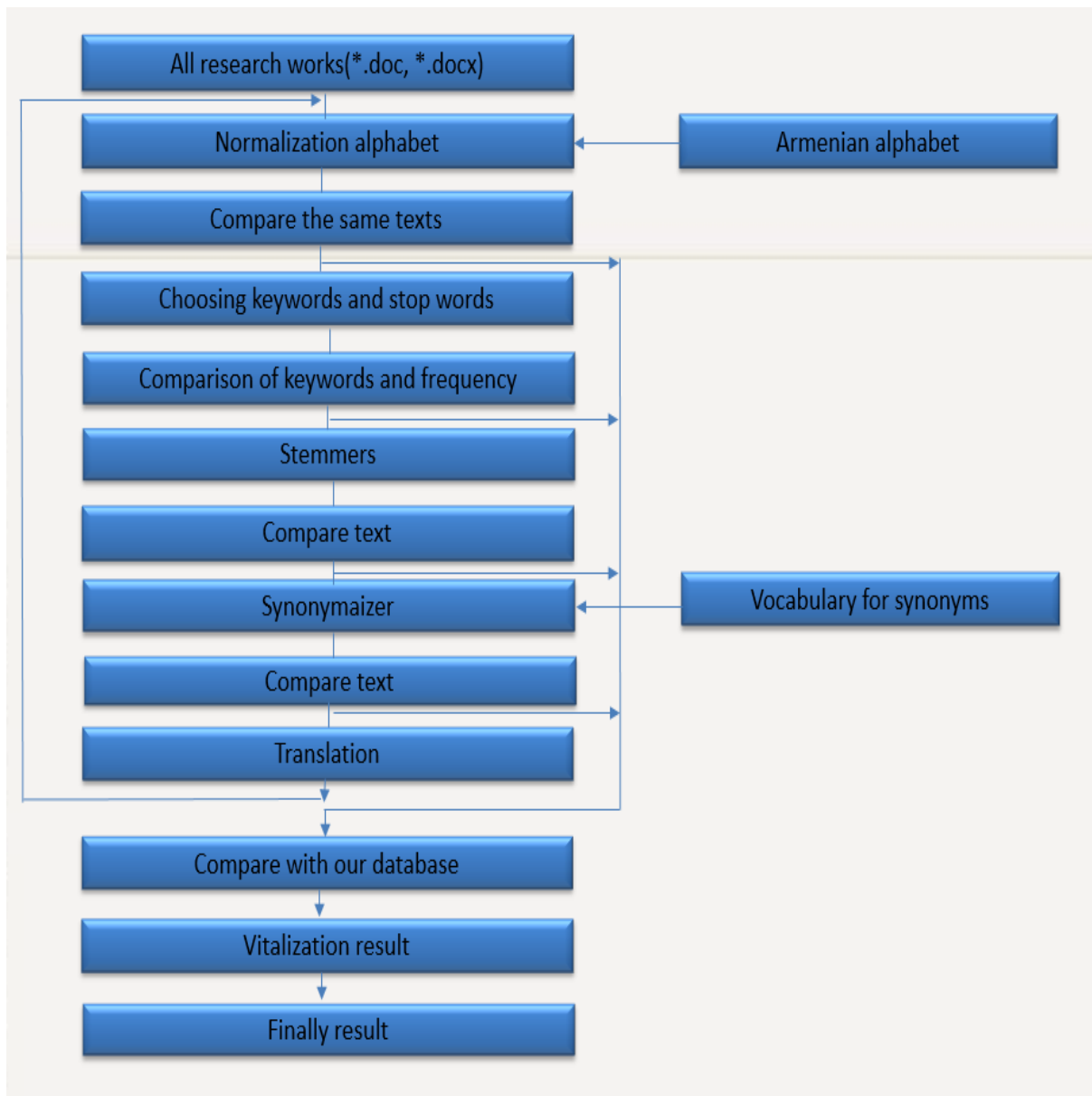


Figure 3.1: Steps of plagiarism detection

In every step the user can compare text.

3.3 Algorithms and methods of plagiarism detection

Search engines, special programs and online services have their own plagiarism detecting systems. However, their ideas are similar and are based on data searching and comparing available texts. In many algorithms [4] are included options, which are being changed. For example, in the option length of shingles [23] used for detecting text uniqueness, the text is divided into separate words and expressions. This technique consists on amounts to reducing each document to a series of numeric codes, such as hash codes, based on sequences of words. The point is that the text is divided into separate fixed length parts (from 3 to 8 words), and the plagiarism detecting system is checking the existence of the shingle on the Internet. So, the uniqueness of papers depends on shingle length.

One of the approaches is based on lexical principles. IMatch [11, 10] signature is calculated for those parts of the text, which has inversion sentences. Two documents are considered to be similar, if their IMatch signatures coincidence.

Another approach is a linguistic method, which is called “keyword” method [30]. In this case to create new documents suitable keyword collections are designed. If the signatures match to each other that means the texts are similar. Although it is hard to process the algorithm, but it detect similarities easily. Sometimes are used uniqueness checking systems that are based on classical methods of information searching, such as Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), etc. TF, TF*IDF [5] is often used as a weighting factor in information retrieval and text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document, but it is offset by the frequency of the word in the corpus, which helps to adjust the fact that some words appear more frequently in general. The Jaccard index [6], also known as the Jaccard similarity coefficient (originally coined coefficient de communalité by Paul Jaccard), is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets.

At first, our program carries out the comparison word by word. The most significant

principle is lexical analysis, as well as linguistic methods. To detect lexical changes we have used steamers, which are based on Porters algorithm [32] taking into account the features of Armenian language. The algorithm gives an opportunity to delete verb endings, noun ending and other types of endings. In our program is used the idea keyword, which gives an opportunity to organize searching in our database very quickly. Keywords have special meaning and they are chosen and formed according to each subject. To replace synonyms we use Armenian dictionary.

3.4 System Implementation

The search for detecting plagiarism should be carried out in the local directory of documents. The implementation is done using the language C sharp, to develop the project we used the Windows Form Application for creating Desktop Application and Asp.net MVC [9] to develop the Web Application, we used MSSQL [8] and Google translate for detecting plagiarism on the Internet with translation. Teachers can upload other files in their own choice to the same directory.

Now we will explain why we have chosen .Net. .Net gives many opportunities for the development of applications for the Windows platform, as well as for the other operating systems, too. .Net supports many programming languages, which have many common functionality across different type of applications. This framework is a software layer which sits between visual studio and operating system. Therefore Microsoft.Net is a common layer for all the Microsoft programming languages like C#, Visual Basic and even F# and etc. We decide to use Windows Form and its functionalities for making Desktop application.

In order to make the Web application ASP.NET Model View Controller (MVC) is used. MVC(Model-View-Controller) is a framework used by developers for developing that has simple and effective design. The controller controls the logical part of every page, making MVC a lightweight framework. The main features of ASP.net MVC are: provides full control over the HTML, code are clearly defined in MVC, simple integration

with JavaScript frameworks, use Bootstrap following the design which is necessary for the web design, Representational state transfer (RESTful) URLs that enables Search Engine Optimization (SEO).

We use Microsoft Structured Query Language Server (MSSQL) database for keep information. MSSQL it is supported by .Net. With SQL Server, we can have better tools integration between Entity Framework and our database schema. Entity Framework is an Object Relational Mapping framework that is open source for ADO.NET. Object Relational Mapping framework automatically creates classes based on database tables, and the contrariwise is also true, that is, it can also automatically generate necessary Structured Query Language (SQL) to create database tables based on classes. We use Bootstrap to design web application, which gives an opportunity to make web development faster and easier, as well as to create extensive and beautiful documentation for Hypertext Markup Language (HTML) elements and to use jQuery plugins.

Now we will describe database structure. The database has several tables, which contain the information about lecturer's, university, and for synonyms. In below on the Figure 3.2 can be seen the tables that are used to create the Web application. Here is showed all fields of two tables: teachers and universities. The teacher table contains personal data of teachers and of the university where they work.

The main idea to kept synonyms is to extend our database for synonyms. The tables structure for synonymizer can be seen in the Figure 3.3. We have two tables to stored data. In first table are stored words, and in second table meaning of the words.

Plagiarism detection systems never publish the information about strategy and methods of system, because that is a private information. That is the reason why in our dissertation we don't show the use case diagrams to keep security of our program.

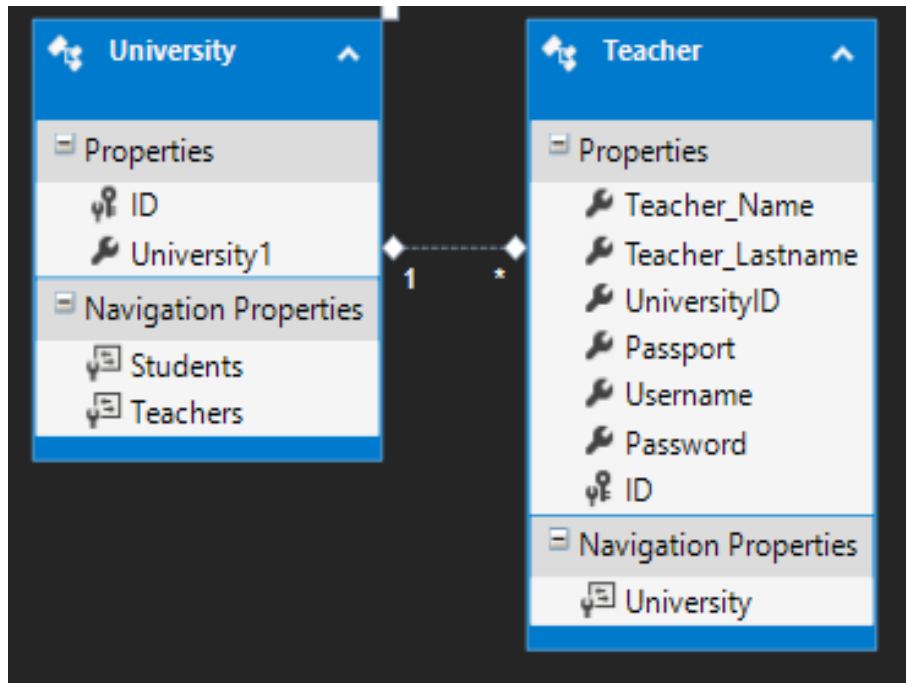


Figure 3.2: Structure of database

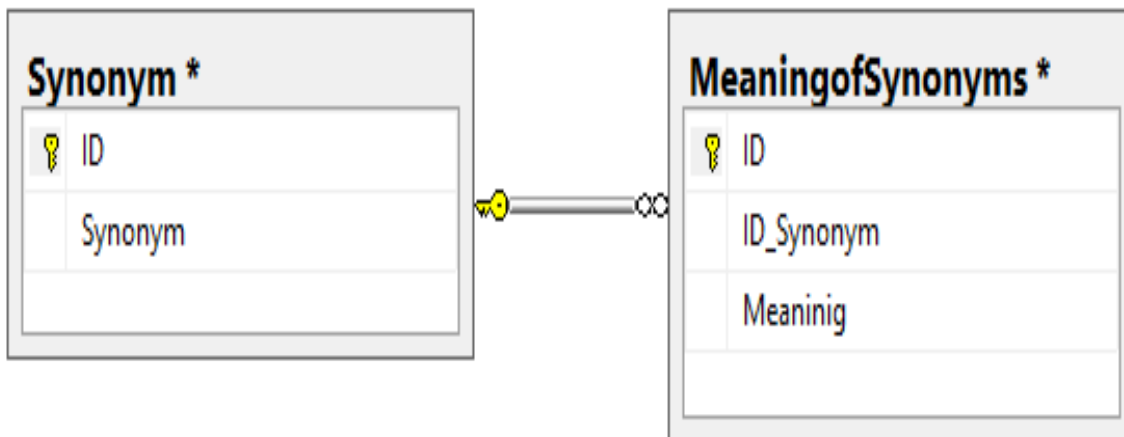


Figure 3.3: Structure of database

Chapter 4

Created System

In this chapter we will represent the main steps of this work to find exactly the same sentences, choosing keywords domain, stop word removal, synonyms recognition and translation for finding the possibility of plagiarism. The program can compare only *.doc, *.docx in our database. Database will be expanded by teachers when they upload and check the students documents(including translated documents). In that sense, the system allows to carry out searching based on previous years works.

The first step of the plagiarism detection tool are shown in 4.1 and the following sections will describe each other step in detail.

4.1 Normalization alphabet

Often students can replace the letters with another letters, for example, some systems are not able to detect if there is Russian “а” letter or English “a” letter. We have some letters which are similar to another letters, for example Armenian “հ” it seems like English “h”.

This program is able to find other letters and point out in another color. An example can be see in the Figure 4.1.

Usually for the normalization of the alphabet, appropriate dictionaries are used. Each fragment (word or symbol, depending on the stage), is compared and when they match the

համարվում → համարվում

Figure 4.1: Normalization alphabet

character word is replaced or deleted. The search must be carried out in the local database to avoid oversights. In our program the first alphabet that is checked is the Armenian one. Unless it is in Armenian the system detects it and underlines with corresponding color. Checks are carried out through American Standard Code for Information Interchange (ASCII) codes. If it is no Armenian, letters it will be pointed out in red color. The program includes the Armenian letters, and letters are comparing through the ASCII code. When the program point out letters, which are written in another language by red color, the teacher will see the result, is able to replace manually the letters into Armenian. After this taken steps, teacher can compare them. If teacher does not replace them, sometimes the system will not be able to recognize and will consider as another word. The main objective of alphabet normalization is to avoid that.

4.2 Finding the same text

First important part of plagiarism detection is to find exactly the same text [22]. This type of plagiarism involves those cases when students can copy word by word and use one or more original sources without pointing the source. Besides on the Internet there are many available information about everything and users have an easy and comfortable search engine for accessing texts in different domains. For students it is achievable by copy and paste [29] operation .

There are many reasons why students lie and do plagiarism. First reason is that many students don't know what is the plagiarism and what will should do when coping. Second type of students know what is the plagiarism, but don't thought that is wrong. And third type of students are interested in the shortest and possible way for doing the work.

For copy-detection on the Internet it's a good way to use the method of shingles[23]

and improvement of shingles.

The program can find exactly the same text and show the percentage, whether there are matching parts. The comparison is realized word by word. At first for comparing, we need to delete all characters except the “:”, which shows that the sentences are completed in Armenian language. This algorithm is used to split the source text into sentences. Separation is carried out by punctuation marks such as a point, exclamation mark, question mark then the text is compared sentence by sentence and if there is a match it will indicate plagiarism existence otherwise continues to perform the next action.

The program can compare two or more files. If we want to compare many documents, we will need only to choose the subject, after that keywords are extracted and we can evaluate the possibility of plagiarism. If we are comparing more files, we can see the result, which is presented on the Figure 4.2.

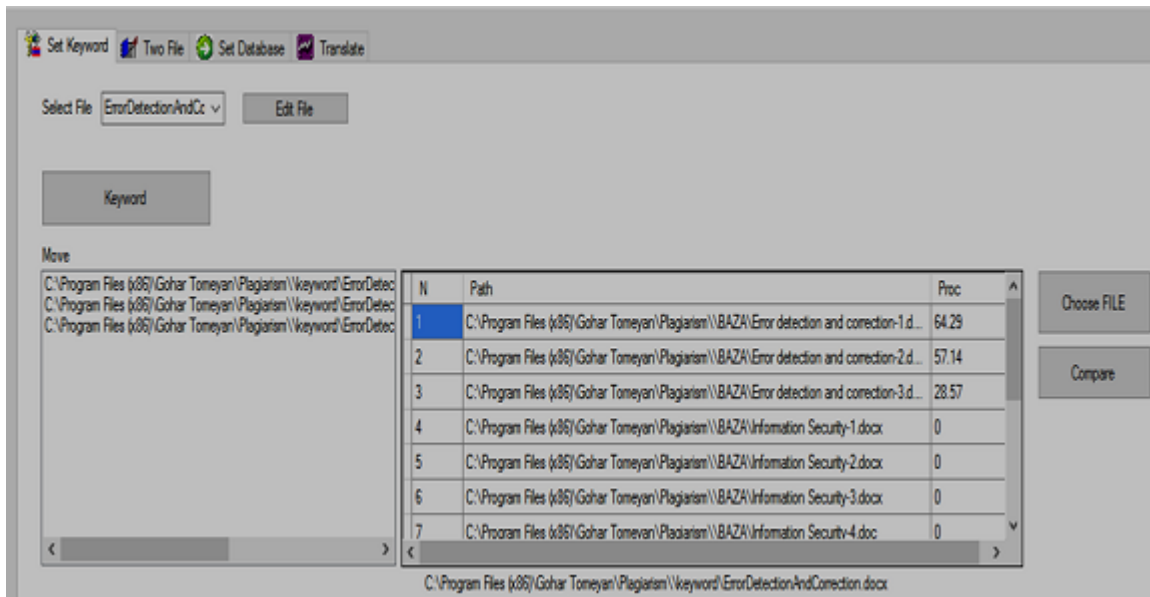


Figure 4.2: Comparing two or more documents

If we are comparing two files, we can see the result, which is presented on the Figure 4.3.

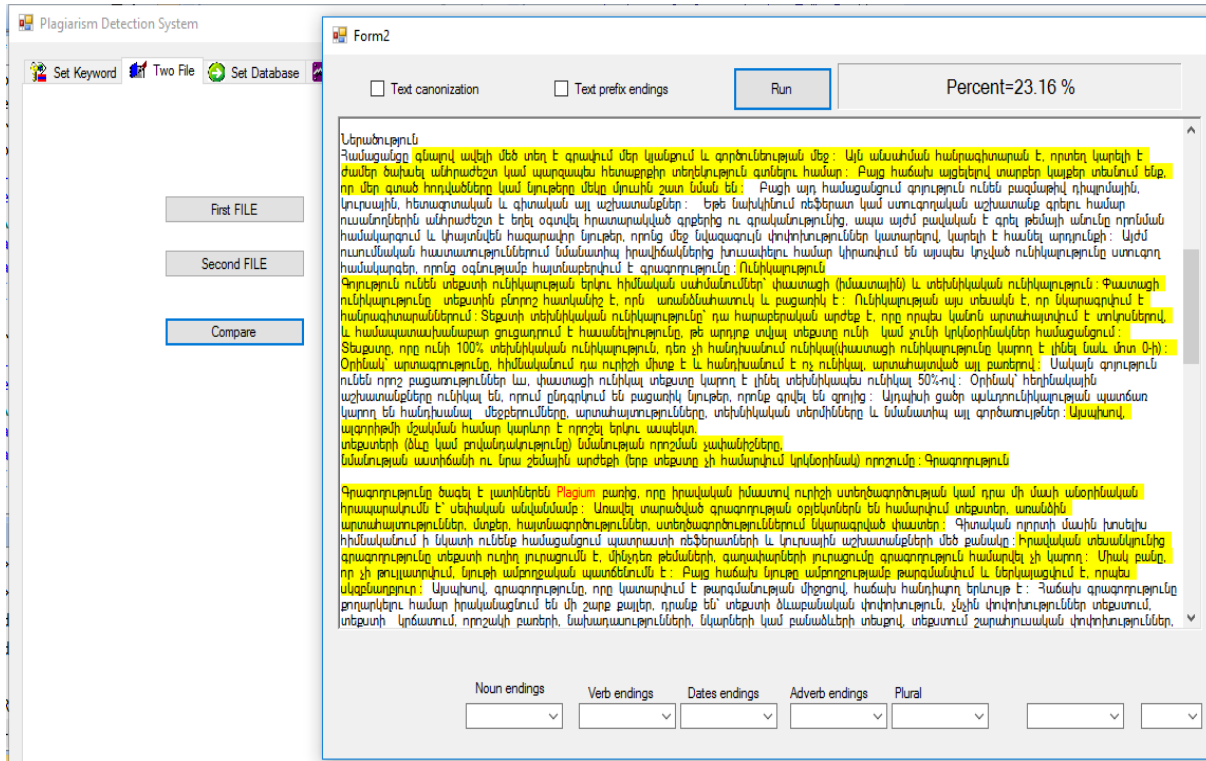


Figure 4.3: Comparing two documents

After carrying out the comparison the final result is presented which is expressed either as a percentage or corresponding references. If the percentage is more than 70% it will be considered as plagiarism. This value was observed when comparing a set of plagiarized documents.

4.3 Choosing keywords domain

Keywords [17] are important things for searching in our database as well as for searching on the Internet. Keywords acts secretly for the successful library searching. We use keyword all the time with Google or with other search engine of our choice. For the effective search in our database we also decided to use a keyword system, which gives an opportunity to compare the texts, which have the same keywords.

If two texts have lots of keywords in compare, these means that they belong to the same knowledge domain. Only comparisons between documents of the same domain make sense. For this reason, we need to work on this topic in order to look for the most corresponding keywords for the search. This is one of the important things to remember about database research. Taking as example the sentence: “Search engines and online services have their own plagiarism detecting systems”, the key elements are “search engines”, “online services”, “plagiarism detecting systems”. Therefore, in this case “have their own” are not keywords.

In our system we have already keywords for some subjects, which are kept in Microsoft Word and saved by special name, for example name of the subject. The program also give the opportunity to upload a new file, which contains the keywords and synonyms of a domain. Teachers can edit already existing keyword files. The program works like this: if we want to add keyword for any subject, first we need to put password, and after choosing the name of the subject, upload file. Depend on the fact who will enter the password, the profile of the user will be different. If is an administrator he can add keywords to existing files with the help of corresponding window, but teacher must have other interface to do it.

When the user choose a file to compare, the system will generate a new folder, and put there only that files that we have in our database and which have the same keywords (belongs to the same knowledge domain). The comparison result is presented by percentage. We do not need to compare all files, we must only compare texts, which have the same keywords. Each subject has separately keywords that are kept in separate documents. In Figure 4.4 we can see the page of the choosing, generating and comparing

the papers. Administrator of the page can add the keywords through a specific form.

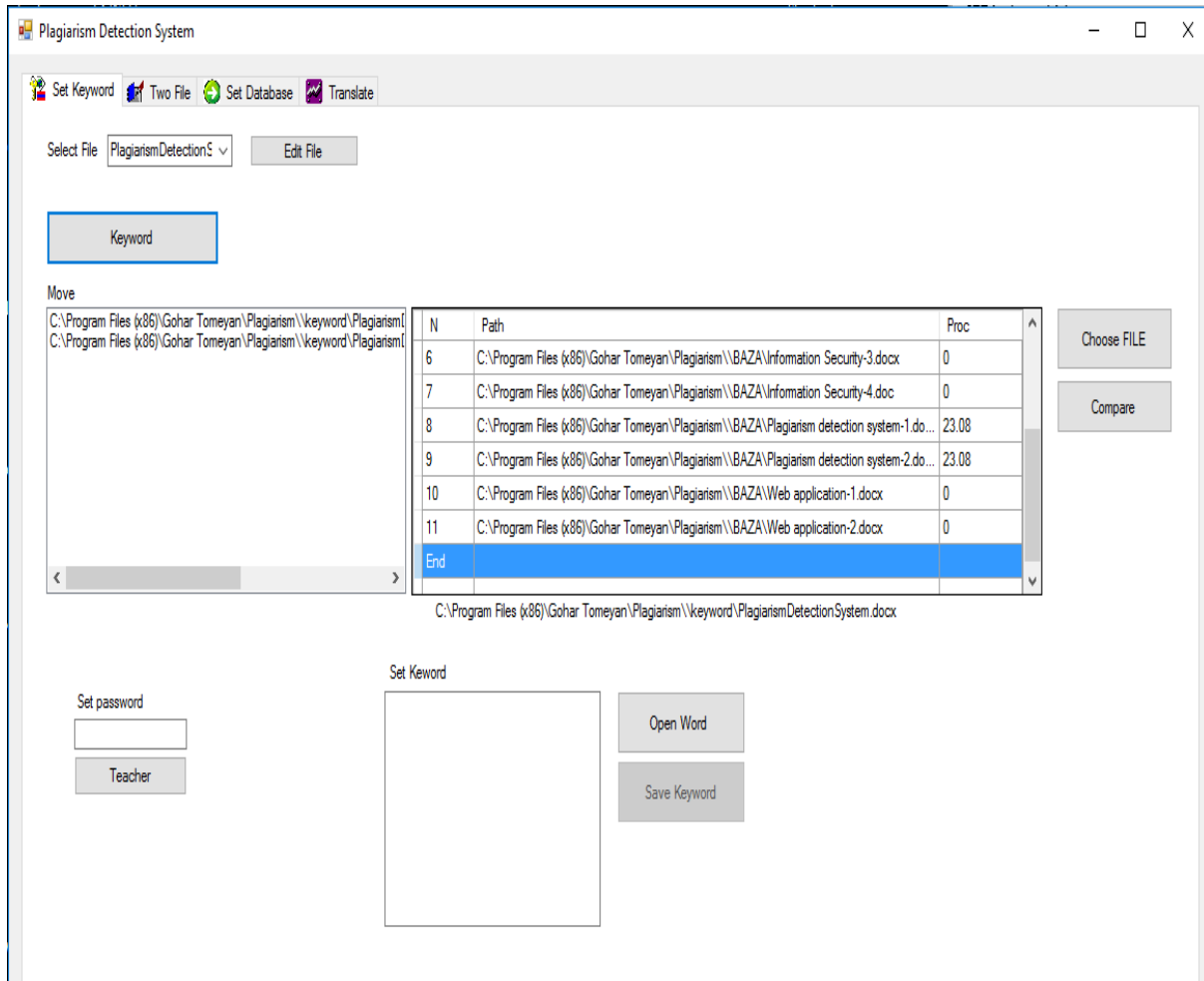


Figure 4.4: Keyword generation

For example the plagiarism detection system subject has only 2 documents which contain same keywords, and after when we will choose document for comparing it will compare only first document with two others and will show the result by percent. That is showed on the Figure 4.4. Keywords are kept in the Folder keyword, which is located in the system directory.

4.4 Stop words removal

Stop word [12] consists in removing common words. Stop words are the most frequent words in texts. The usual way of determining what counts as a stop word is just to use a dictionary that lists them. However, for Armenian language we have not a dictionary. Stop word do not indicate the topic of the document in any way and they carry almost no meaning.

այդ	էիր	միջև	հետ	իրենից	կար
այլ	էիք	նա	հետո	մասին	միայն
այն	եր	նաև	վրա	բայց	որոնց
այս	ըստ	նրա	մեջ	բացի	այնպիսի
այլևս	թե	նրանք	կարևոր	հաճախ	ինչպիսի
դու	թև	նրանց	կա	երբևիցե	ինչպիսիք
դուք	թերևս	որ	որի	հանձինս	ի
դրանք	իսկ	որք	հետևյալ	այսպիսով	իրարից
դրանց	իր	որոնք	այստեղ	երթ	ինչևիցե
եմ	կամ	որպես	այնտեղ	ու	մինչ
են	համար	մոտ	ինչպես	ևս	որտեղ
ենք	երբևէ	որևէ	նրան	երբեք	երբեք
ես	մի	որպեսզի	որոշ	երբ	երբեմն
եք	մինչև	և	իրեն	եք	դրա
այնի	չի	դա	սա	ինչ-որ	ինչևէ
այնպիսի	չէ	որոշակի	քանի որ	քանզի	ոչ
այնքան	բոլոր	բոլորովին	որից		

Figure 4.5: Most common stop words

In this table are given the most frequently 100 stop words for Armenian language.

The program includes these words and we can delete stop words, see frequency and compare. In the Figure 4.6 is presented the stop words removal process. If we want to delete the stop words we need only to choose the second "text canonization" check box and run, and the program will show the result and percent of the text removed.

Stop words are saved in our database, and in the future it can be extended.

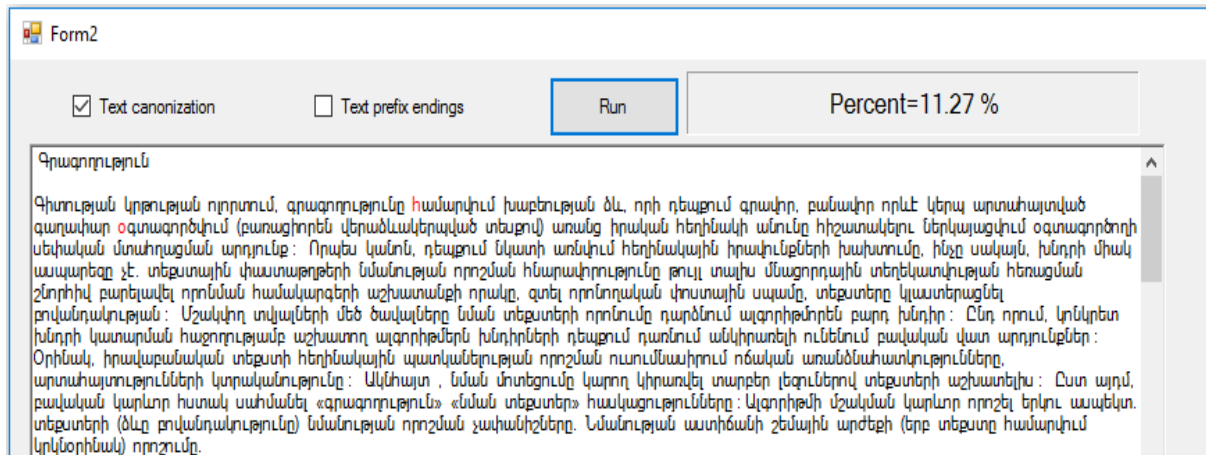


Figure 4.6: Stop words removal process

4.5 Stemming

One of such important and necessary things on computer linguistics is the operation of the using steamers [12]. Stemming are usually used in Information Retrieval systems. There are many algorithms for stemming and this process is for removing the suffix and prefix for the words. The main algorithm of the stemmer, is the Porter algorithm [33] written by Martin Porter for English language and other languages. For English language Porter2 algorithm is used now, which is an improvement of Porter.

The most famous project for stemming is Snowball [28], it is a small string processing language designed for creating stemming algorithms, which are now supported by C# ISO C, Java and Python. C# console application uses three steamers for Russian, English, German. But snowball has totally steamers for 14 languages: English, French, German, Hungarian, Italian, Portuguese, Russian and others.

In our program, stemming will be used for searching the basic forms of words and replacing with the synonyms.

The best way for determining steamer it is just using the dictionary. Armenian language has very hard structure. Now we will compare English and Armenian language and difference between them. The project Snowball contains the old version Armenian suffix and prefix, but Armenian language has endings too, when in Armenian language

we delete suffix or prefix, the words will change its meaning. But for English language endings and suffix have same meanings. For example, if words finished in “-ed” , just in English we can delete “-ed” suffix and words will not change the meaning. English has quite strict and rather inflexible word order, if we compare languages Armenian grammar word order will be quite flexible. For example: "I want to learn languages.", here there is only one grammatically correct way to express this sentence in English. Yet, if we translate this sentence with accurate Armenian grammar rules, we will have 4 different choices. For this sentence can be Armenian 4 choices. Another difference we have 7 cases of nouns in Armenian, but English has only 5, and each case has many and different endings. Prepositions are short words (on, in, to) that usually stand in front of nouns and sometimes also in front of gerund verbs in English. For example if we have a “in the program” this sentence, in Armenian language it will be only one word “cragrum”, where “-um” is ending. While in English, the plural is formed by adding (-s, -es.) to the singular, in Armenian, to form the plural of nouns and adjectives we add (-er, ner): The suffix “-er” is added to the end of monosyllabic nouns. The suffix “-er” is added to the end of polysyllabic nouns.

Համարվում → Համարվ

Figure 4.7: Stemming example

In our program, teachers can see all endings. If teacher wants to compare two files to know the possibility of plagiarism, he can delete all endings and for that he needs only to choose the second "text prefix endings" check box and run. All endings will be deleted, therefore the system will give the percentage without endings. The Figure 4.8 presents all the endings of nouns, numerals, plurals, verbs and pronouns, which the user can see.

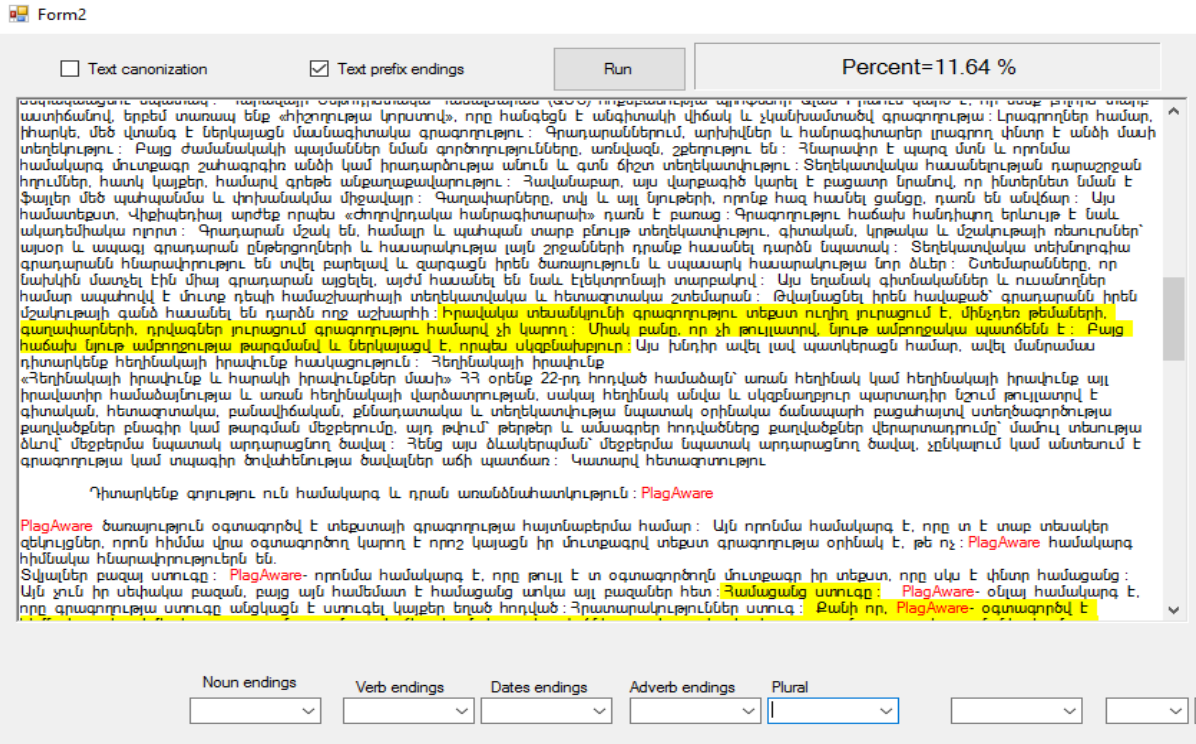


Figure 4.8: Stemming

4.6 Synonymizer

People use different words to write the same idea. The most important thing for NLP is the identification of synonyms. Students can replace the words by synonyms to conceal plagiarism, because people carry out everything to hide plagiarism by replacing words by synonyms, and plagiarism detection systems can't easily find similarity. Then, most of the common plagiarism detection methods couldn't achieve good result, because the text will be different from the original source.

There are several existing synonymizers [25]. On the Internet we can find several web services, which contains the synonyms of Armenian language, but source code is private, and all existing words has lot of explanation about each word, which is not necessary for our program.

In plagiarism detection program we need only synonyms and the explanation of the words are not necessary. And several synonymizer can replace by synonyms automatically,

but the meaning of the text will be changed. It is not a good idea to use automatic system, which complicates the work of the teacher, since every word may have many synonyms and different meanings for different sphere.

We decide to create synonymizer to automate and facilitate the process, and add in our database synonyms and meanings of synonyms. The final decision of replacement each word without changing the meaning of sentences is up to the teacher. The main concept is to use synonyms but to keep the meaning of the text.

When we will chose the file and compare two documents,after using stemming we can replace with synonyms. The program has an option which points out words in red color and replace with synonyms. Teacher has the opportunity to point out words in red color, choose the meaning which corresponds to the context and save changes. As presented in the Figure 4.9, after choosing the word in the right side panel(where the user can see the meaning of synonyms) and the user can choose the corresponding word and save it, then the user can “Run” the program and see result by percentage. We will present in Figure 4.9 a short example to explain the functionality of the system. This simple sentence, have two possible candidates for replacement with synonyms. This sentence is taken from the second file, which contains the same sentences from an original source.

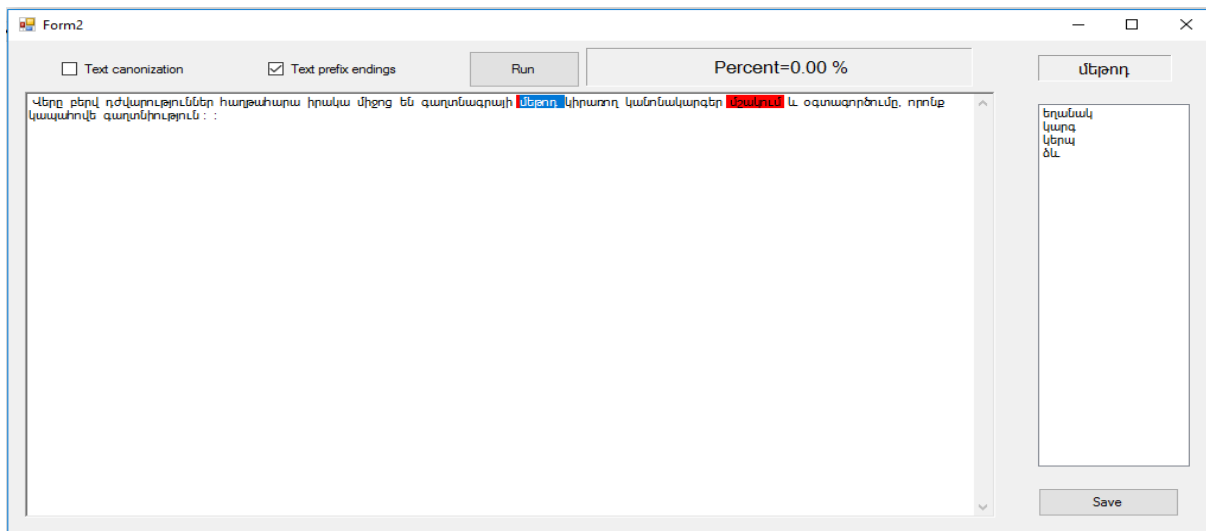


Figure 4.9: Before synonymize

After replacement the words, as we see in the Figure 4.10, the result is as expected text pointed out in yellow color, which means that is not an original text.

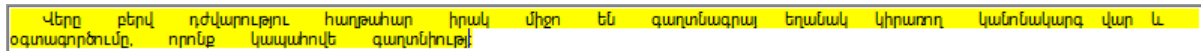


Figure 4.10: Replacement with synonyms

In the program we have solution for teachers adding synonyms and explanation of the synonyms. In the Figure 4.11 we can see the interface for that.

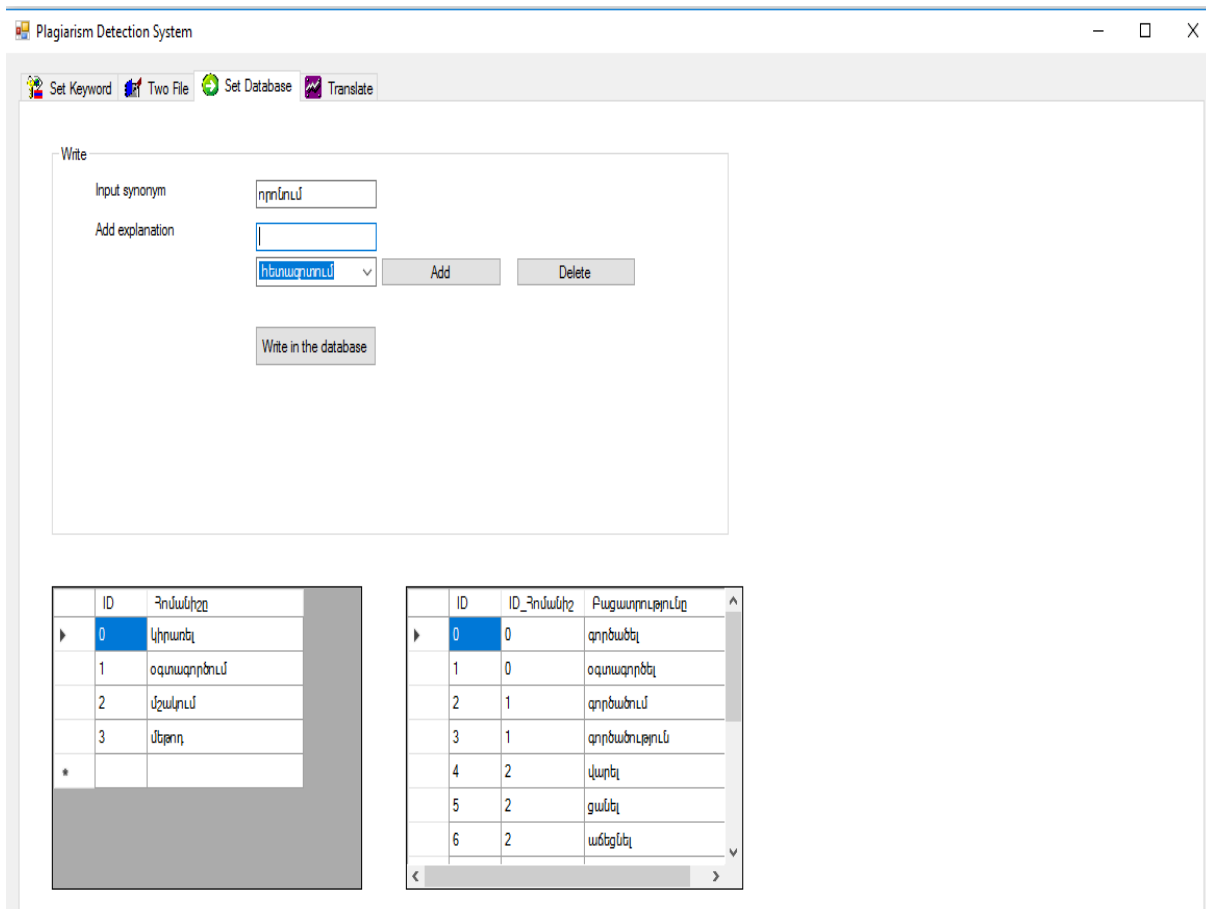


Figure 4.11: Adding synonyms and explanation

Teachers can add and see the synonyms, which are existing in our database. The teacher can only delete his own synonyms. At first, teacher need to write synonym, explanation and choose the add button, after in combo box we can see all the explanation. If its correct, the teacher can write in database.

4.7 Finding plagiarism with translation

Plagiarism can be done by people, who can translate the text from one language to other without referencing to the original source. Translated plagiarism [21, 29] can include two types of translation: automatic and manual translation. Google Translator is a free multilingual translation service, which is developed by Google. Google translator gives huge opportunity to translate text, images, sites, or real-time video from one language into another. It has a web interface, it is supported for Android and iOS apps and another operation systems. It supports 100 languages for translation at different levels. It does not translate from one language to another. It often translates at first the text into English and then to the target language. Google translator looks for various documents to make the best translation when generating.

By detecting samples in documents, which Google has in its database it chooses the most intelligent version to make appropriate translation. For improvement translation there is Translate Community platform. That allows to select up to five languages to help improve translation. Users can verify translated phrases and translate phrases in their languages to and from English. Sometimes for natural languages like Armenian Google translator doesn't work very well. It can give different translations for the same text. The reason is, when a person translates the text, he can make change in the Google translator result. Google translator can take profit of the changing and then in the next translation the Google will give already a right result.

Google gives a huge opportunity to make changes and optimize the natural language processing for the Google translate. Plagiarism with translation is the most serious problem in Armenian, because is difficult to find sources in mother language. That is a huge academic problem. The best way to hide plagiarism is using manual translation, but it is very hard work.

Plagiarism with translation is very difficult to detect. There are many kinds of problems: the translated words, can have many meanings, and the translator translates all words automatically, and system has to find which one is the correct word. If people

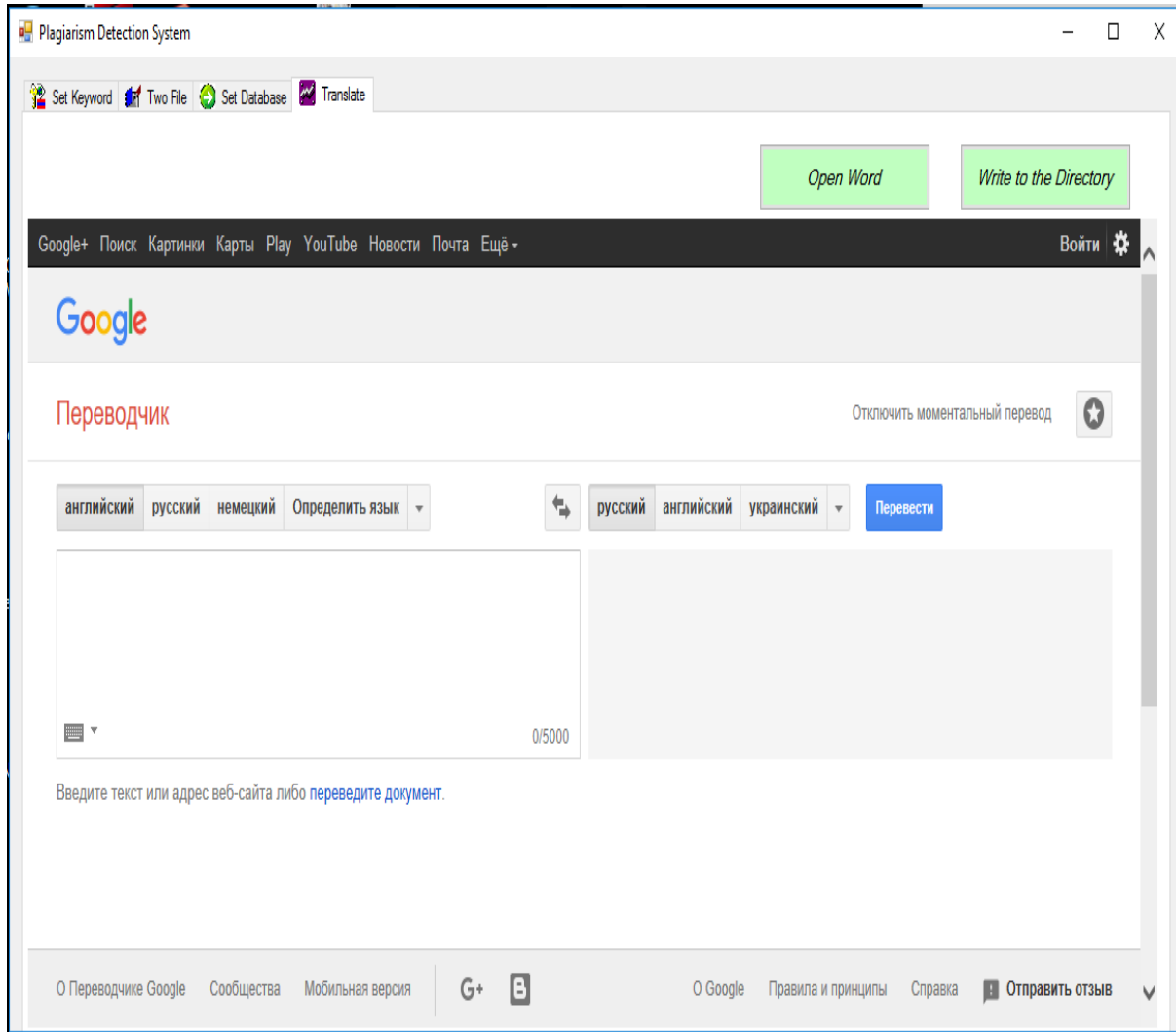


Figure 4.12: Google Translate

knows many words, they easily can hide plagiarism with the help of synonyms.

Each Natural language have different grammar structure and rules, so a word by word translation is not possible and completely. Translation for Armenian language is not working effectively, it's enough only for understanding but not for plagiarism detection.

We don't have much information in Armenian language on the Internet, and students often translate the documents from Russian and English texts, and present as own idea. Usually students use Google translate, for that reason, we include Google translate in our program. Translation will work if the user has connection to the Internet to translate the

document. Teachers must copy and paste the translated text on the Microsoft Word, and after which upload that file to our database. And teacher can take the same steps for this document: choose keywords domain and compare with many documents or compare only two documents, stop word removal, stemming and synonym recognition.

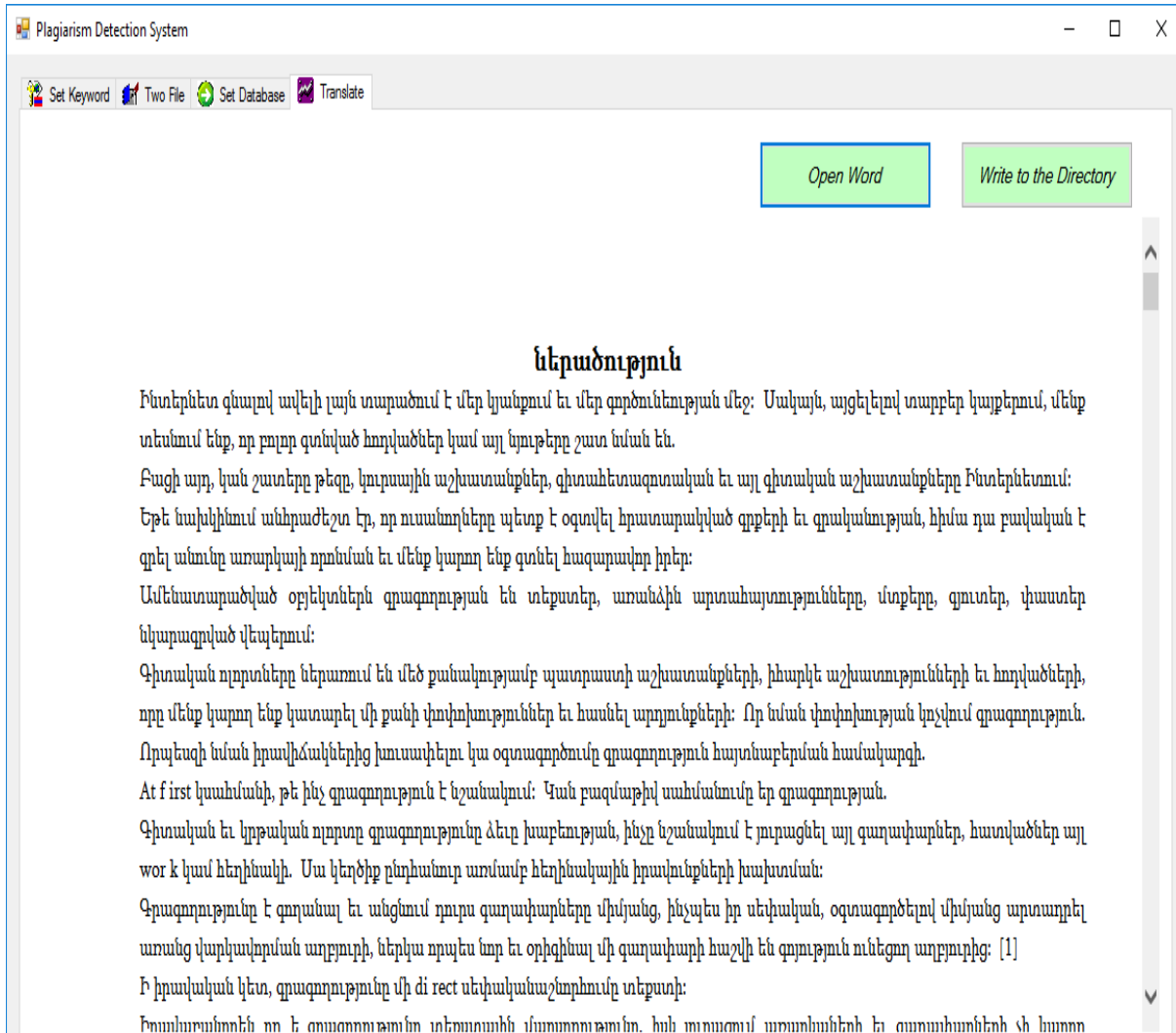


Figure 4.13: Translation

The User Guide of the system is in appendix.

Chapter 5

Web Application

The Internet and the World Wide Web gives many approaches for sharing and searching information. People can't imagine how can they live without Internet and information. But it's clearly a big problem when people passing off another person's idea as their own, and not only educational, but about other spheres too. Internet everyday growth brings to the fact that new web sites and web applications are created. Consequently, we decided to create web application to extend our program on the Internet. The program will be useful for teachers who work in the universities. The web application gives an opportunity of registration, download and install system with database, and also download the user guide get acquainted how system is work, as well as contact page, for sending question about doubts or suggestions. Now we have two kind of users: administrator and teachers.

5.1 Web Application development

To develop Web application we decide to create web application using ASP.NET MVC. The web application will be useful for lecturers. The program now has two type of users: first is the administrator of web application and second are the lecturers.

In the main page people can get acquainted with steps of our program, see the page about us and about strategies that are used in this program, send message administrator and know about interesting and troubling questions. The main page can be seen in Figure

5.1 .

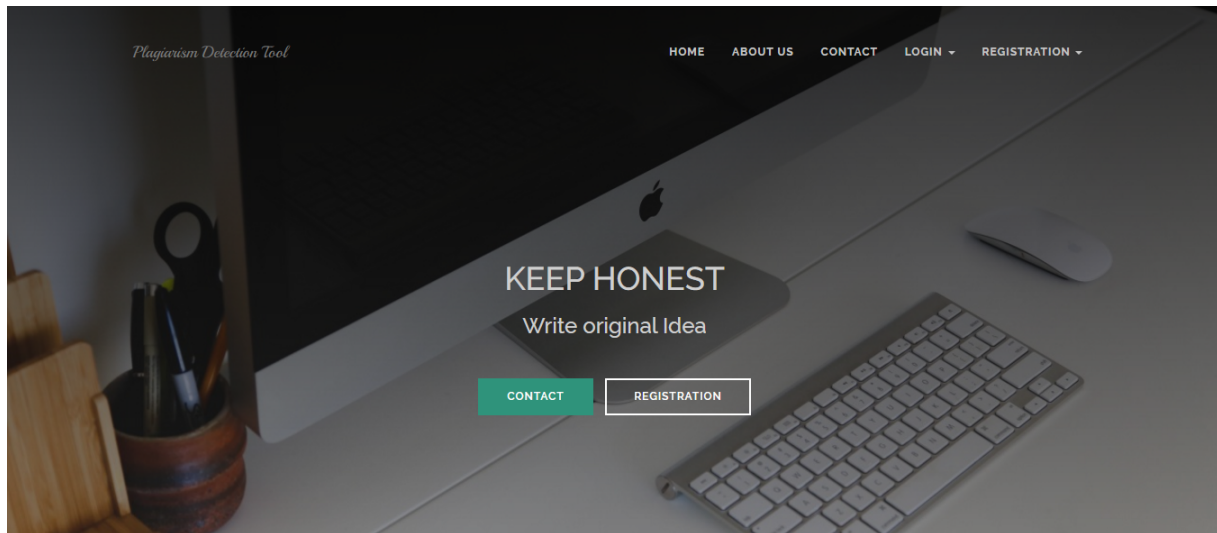


Figure 5.1: Main page

To use many websites, users need to register on the system and fill personal data. In our website it's important, because if people doesn't do registration on the system, they can't use our program. For the basic registration important data is name and email address, but its not enough for registration in our program. The program need to recognize the user profile. To distinguish administrator and lecturers we have a special key for registration. That key is given only to those lecturers who works at the university. One of the most important part in web application it is to implement security. Role management it's one of the important thing to manage authorization process, which includes that users have access to specify the resources. Roles give the flexibility to change permissions for example to add and remove users, or change permission, without making changes throughout the site.

Our program gives solution for the registration lecturers. Lecturers need to fill all fields to secure registration process. If lecturer don't fill some fields, the application will show several errors. One of the important thing for registration is Email confirmation, because if lecturer forget the password, he/she can only reset password with the help of

email. Next important thing is to store passwords. The best method of storing passwords is the use of hash function. Hashing is a one-way function, which means when we hash a password it is very difficult to get the original password back from the hash, exception several standard hash functions. If we are hashing with standard function we can easily use a reversible attack, which can be carried out with the help of dictionary, where it will look up hashed passwords. Now there are many hash algorithms [27] reversible , for example MD5 or other algorithms. Best way to store passwords is using the standard hash function, salt and then encrypt the hash to prevent dictionary attacks.

The program must be able to check all input fields. For example in the “name” field you can’t put numbers. The password must contain at least contain 6 characters letters, numbers, uppercase and lowercase. The registration page we can see in the Figure 5.2.

Registration Page

[Back To Index](#)

Figure 5.2: Registration page for the teacher

After Registration, lecturer can Log In and see other functionalities. The Log In page of our application can be seen in the Figure 5.3

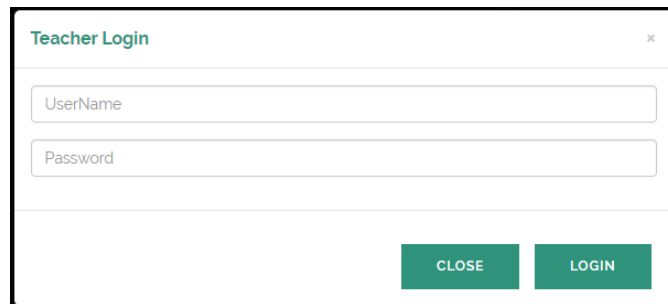
A screenshot of a 'Teacher Login' dialog box. The dialog has a title bar with 'Teacher Login' and a close button. It contains two input fields: 'UserName' and 'Password'. Below the fields are two buttons: 'CLOSE' and 'LOGIN'.

Figure 5.3: Log In Page

When teachers are registered on the system they can see the new page, which is showed in the Figure 5.4.

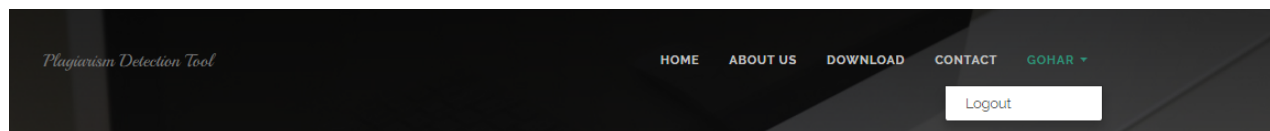


Figure 5.4: After Log In

Where he can download, see the explanation about system and User Guide for using the system for the first time (Figure 5.5).

Certainly user has the opportunity to send a message with the help of contact page, and find the answers for questions, as can be seen in Figure 5.6.

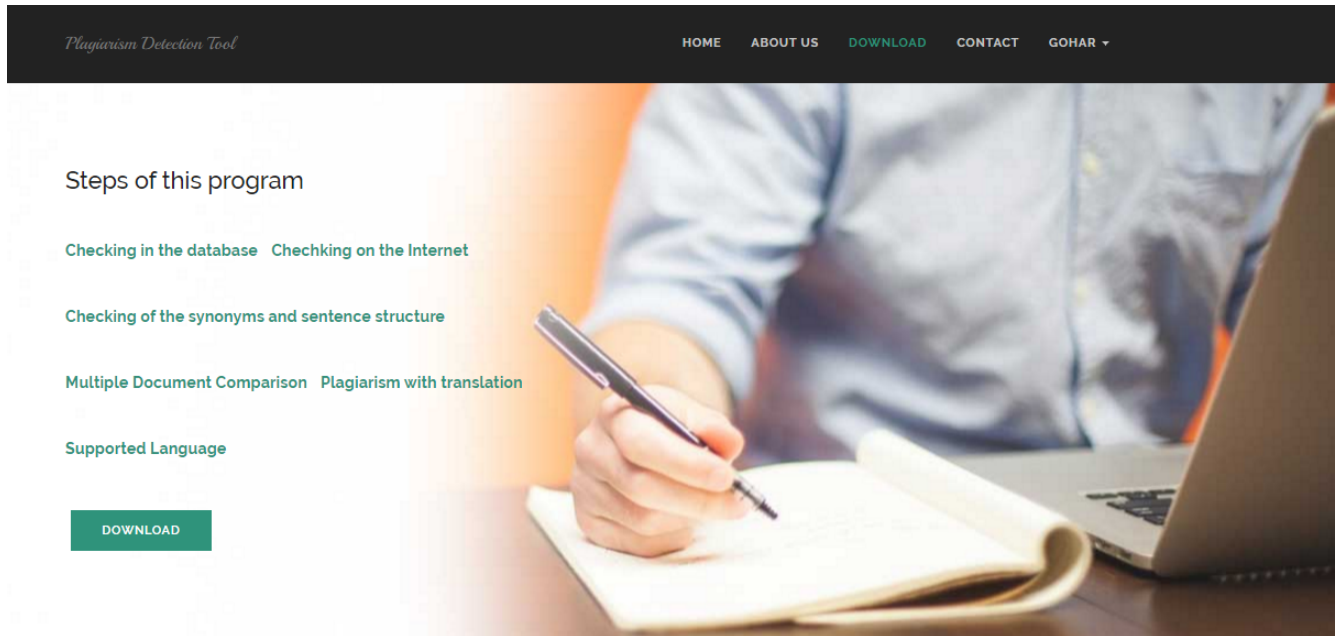


Figure 5.5: Download

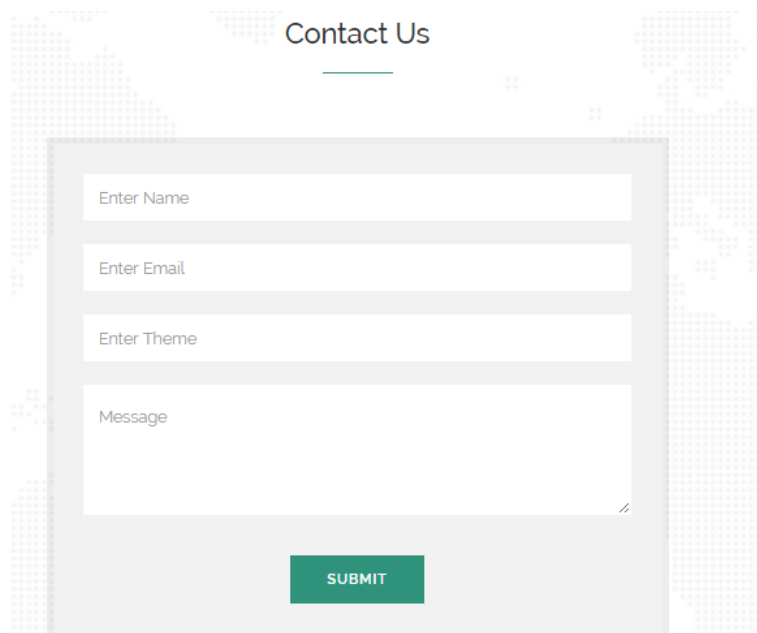


Figure 5.6: Contact page

Chapter 6

Tests

It is intended to perform program testing on the basis of existing texts using Beta testing. While testing the existing disadvantages, we also test linguistic failures. As can be seen, the program includes many levels of plagiarism detection. And each step is represented separately in chapter 4 and user guide. Each window has different steps and different procedures. In the first page, we can compare more documents, but we need to choose the subject, if the subject is upsilons, teacher should add the keyword for subject, and after generate the folder for that subject. In the second window teacher needs to compare two documents, use stemming, stop word removal, and synonym recognition. In the third window teacher can add new synonyms, or delete synonyms, which he/she add. And the last procedure is to find the possibility of plagiarism with translation, upload documents, and after repeat the whole steps.

The tests are done by students from Armenia who study now in IPB and specialist for testing linguistic approach. The students have added their own keywords and documents. Each student has added 5 documents, for example: thesis, term papers, research works, which was done in Armenian, as well as keywords and synonyms of keywords, which are chosen by students, depend on the specialty. The count of common documents now is 60, which include the following knowledge domains: Information System, Error Detection and Correction, Information Security, Chemical Engineering, Cloud Storage, Plagiarism Detection, Business Management.

Each student is testing the system separately. After testing students gave the opinions and advances about system. And first opinion is the system is useful for lecturers. The main disadvantage is retained to hard interface of desktop application, which is very difficult to use without user guide. System gives opportunity of installation package to download the user guide. Another disadvantage is the system has few synonyms, which will be added in near future or we will use already existing synonymizer for Armenian texts. Each student compared own documents and saw the result. In the Figure 6.1 you can see the comparison result for this subject: Plagiarism Detection.

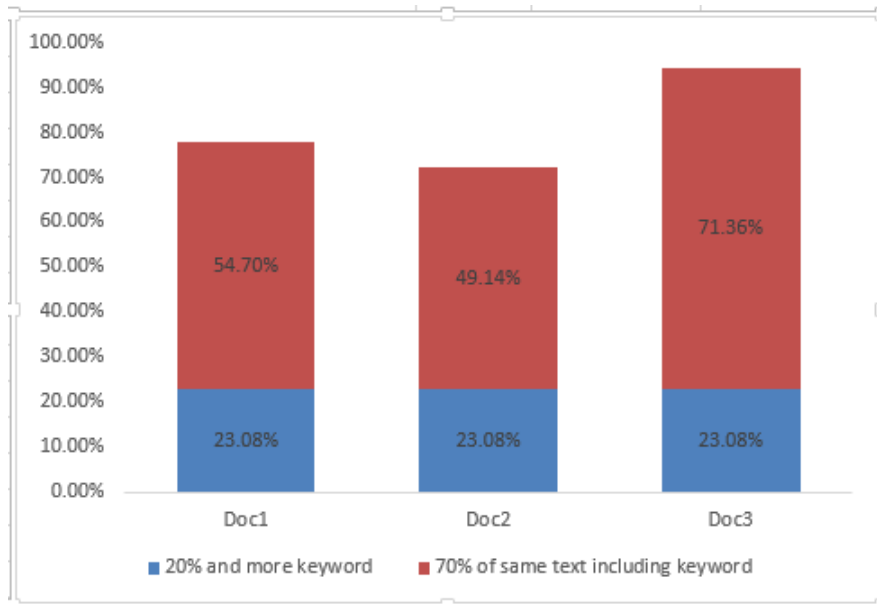


Figure 6.1: Plagiarism detection by percentage

Doc1, Doc2, Doc3 are the documents that are used to compare with the original document.

There is percentage for several steps. For example, choosing keywords, if the original document contain 20% keywords, new folder will be created , with documents which have the same keywords. If the keywords are only 10% in original document we decide that documents don't belong to that knowledge domain. If we are comparing many documents the percentage will show the percentage of keywords + possibility of plagiarism together(in this study the possibility of plagiarism is only considering the comparison

about same text). For many documents if the percentage is 70% and more it will have considered plagiarism. How we can see in the Figure 6.1 the percentage of keywords for each documents is 23.08%, and after compare one with many documents the percentage will be for keywords + possibility of plagiarism. The result will be: first document 77.78%, second 72.22%, third 94.44% possibility of plagiarism. It was considered as plagiarism, because the result higher than 70%.

When comparing two documents that contain 60% of the same text we consider that it is candidate to plagiarism. After comparison more documents we decided to compare the main document with the document which has the highest possibility of plagiarism and see the details about possibility of plagiarism in text. The possibility was shown without keyword, the percentage is 74.45%. It was consider as plagiarism, because the result higher than 60%. This is an limit that we consider when comparing two document. The system does not contain information about references, and teachers have to second manually and check references, to make a decision, because that documents which are candidate to be plagiarism, can contain many references.

The main difference of different subjects is to add the keywords and compare. If in our database we don't have that kind of documents, teacher needs to add new documents and after compare again. At the end, the system was successful and it was able to detect the similarities between documents in a very effective way.

Chapter 7

Conclusion and Future work

This dissertation described the proposed plagiarism detection system for Armenian documents. The program compares two and more documents. Program now allows these steps: normalization alphabet, choosing keyword domain, stop word removal, stemming, and program is able to detect the replacement by synonyms and find plagiarism with translation.

7.1 Conclusion

Our plagiarism detection system compares the texts in directory, which is extended owing to teacher's uploaded files. The local component carries out detailed similarity computations to detect if the given document was plagiarized from the documents retrieved from the Web or in a local directory. To prove the thesis, we already construct a system implementing different plagiarism levels. That levels are to normalize alphabet, find the exactly same text, compare to or more documents, choosing keywords(knowledge domain), stemming, stop-words removal, recognition synonyms and plagiarism with translation. The user can follow all of these steps but he can also stop when he want and just do some of them, change the keywords domain, introduce synonyms and so on. It was also created a Web application, which will be extended and available not only for teachers but for all the users in the future.

7.2 Future work

In the future work, the system must be optimized in order to copy with:

- syntactical changes detection,
- comparison with Web sources, without having to copy to the directory, using Shingle method.
- make the web page also available to students(with important restrictions),
- finding references,
- online use.

And also, in future work, more similarity measures can be included in the system comparison model. More tests to the system will be performed.

Bibliography

- [1] ЗАО Садист. Антиплагиат. www.antiplagiat.ru.
- [2] О.М.Замятина П.И. Мозгалева, К.В. Гуляева. Информационные технологии для оценки компетенций и организации проектной деятельности при подготовке технических специалистов. *Информатизация образования и науки*, (4):30–46, 2013.
- [3] ЗАО Садист. etxt Антиплагиат, программа проверки текста на уникальность. www.etxt.ru/antiplagiat/.
- [4] И.В.Сегалович И.В. Зеленков. Сравнительный анализ методов определения нечетких дубликатов для web-документов. In *Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»*, pages 166–174, 2007.
- [5] Е.В.Шарапова Р.В. Шарапов. Система проверки текстов на заимствования из других источников. In *Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL '2011): Труды XIII Всеросс. научн. конф. Воронеж: ВГУ*, pages 233–238, 2011.
- [6] Н.В. Неелова. Функция удаления нейтральных слов при вычислении нечетких дублей лексическим методом Джаккарда. In *Оргкомитет конференции*, volume 18, page 149, 2009.

- [7] Е.С. Чиркин. Системы автоматизированной проверки на неправомерные заимствования. *Вестник Тамбовского университета. Серия: Гуманитарные науки*, 2013.
- [8] М.А. Хорошилова. Выбор системы управления базами данных с позиции обеспечения информационной безопасности. *им. НЭ Баумана Издательство МГТУ им. НЭ Баумана*, 2017.
- [9] Э.Троелсен. *Язык программирования C# 2010 и платформа .NET 4*. Вильямс edition, 2010.
- [10] O. FRIEDER et al. A. CHOWDHURY. Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems (TOIS)*, 20(2):171–191, 2002.
- [11] J. Alspector A. Kolcz, A. Chowdhury. Improved robustness of signature-based near-replica detection via lexicon randomization. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 605–610. ACM, 2004.
- [12] M.S. Binwahlan et al. A.H. Osman, N. Salim. An improved plagiarism detection scheme based on semantic role labeling. *Applied Soft Computing*, 2012.
- [13] Amalia. Performance evaluation of free anti-plagiarism software. *Proceedings of The 3rd Annual International Conference Syiah Kuala University (AIC Unsyiah) In conjunction with The 2nd International Conference on Multidisciplinary Research (ICMR)*, 2013.
- [14] V. Snasel A.M.E.T. Ali, H.M.D. Abdulla. *Overview and Comparison of Plagiarism Detection Tools*. CEUR Workshop Proceedings, Ostrava - Poruba, Czech Republic, department of computer science, vsb-technical university of ostrava, 17 edition, 2011.
- [15] C.N. Bhusari A.N. Pai. Plagiarism detection system. *International Journal of Innovations in Engineering and Technology (IJJET)*, 1 February 2013.

- [16] M.Y.M. Chong. A study on plagiarism detection and plagiarism direction identification using natural language processing techniques. 2013.
- [17] P. Clough. Plagiarism in natural and programming languages: an overview of current tools and technologies. 2000.
- [18] J. Touras D.W. Wulff, C. Moller. Plagiarism detection software test 2013. *Abgerufen am*, 2013.
- [19] PlagScan GmbH. Plagscan - plagiarism checker. www.plagscan.com.
- [20] iParadigms;LLC. Turnitin - technology to improve student writing. turnitin.com.
- [21] J.Kasprzak and M. Brandejs. Improving the reliability of the plagiarism detection system. *Lab Report for PAN at CLEF*, pages 359–366, 2010.
- [22] E.W. Brown J.W. Cooper, A.R. Coden. Detecting similar documents using salient terms. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 245–251. ACM, 2002.
- [23] C.L. Giles K.Williams. Near duplicate detection in an academic digital library. In *Proceedings of the 2013 ACM symposium on Document engineering*, page 91–94, 2013.
- [24] E.Sutinen M. Mozgovoy, T. Kakkonen. Using natural language parsers in plagiarism detection. *SLaTE*, October 1-3 2007.
- [25] M.E.B. Menai. Detection of plagiarism in arabic documents. *International journal of information technology and computer science (IJITCS)*, 2012.
- [26] S.A.Hiremath M.S.Otari. Plagiarism detection-different methods and their analysis: Review. *International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN*, August 2014.

- [27] D. Mazières N.Provos. A future-adaptable password scheme. In *USENIX Annual Technical Conference, FREENIX Track*, pages 81–91, 1999.
- [28] M.F. Porter. Snowball: A language for stemming algorithms. `snowball.tartarus.org`.
- [29] C.N. Mahender R.R.Naik, M. B.Landge. A review on plagiarism detection tools. *International Journal of Computer Applications*, 125(11), 2015.
- [30] A. Melkov et al. S. Ilyinsky, M. Kuzmin. An efficient method to detect duplicates of web documents with the use of inverted index. In *Proc. 11th Int. World Wide Web Conference (WWW'2002)*, 2002.
- [31] A.Abraham S.M. Alzahrani, N. Salim. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2012.
- [32] R. Pohlmann W. Kraaij. Porter's stemming algorithm for dutch. *Informatiewetenschap*, pages 167–180, 1994.
- [33] P. Willett. The porter stemming algorithm: then and now. *Program*, 40(3):219–223, 2006.
- [34] C. Fox Z. Ceska. The influence of text pre-processing on plagiarism detection. *Association for Computational Linguistics*, 2011.

Appendix A

User Guide

A.1 Set keyword domain and comparing

Each subject has separate keywords that are kept in separate documents. In Figure A.1 we can see the page of the choosing, generating and comparing the papers. If we want to extract keyword for each subject, we need to choose subject, and after chose "Keyword" button. In the Move part you can see generation folder and that automatically put there keywords for that subject. The result can be seen in Figure A.1 .

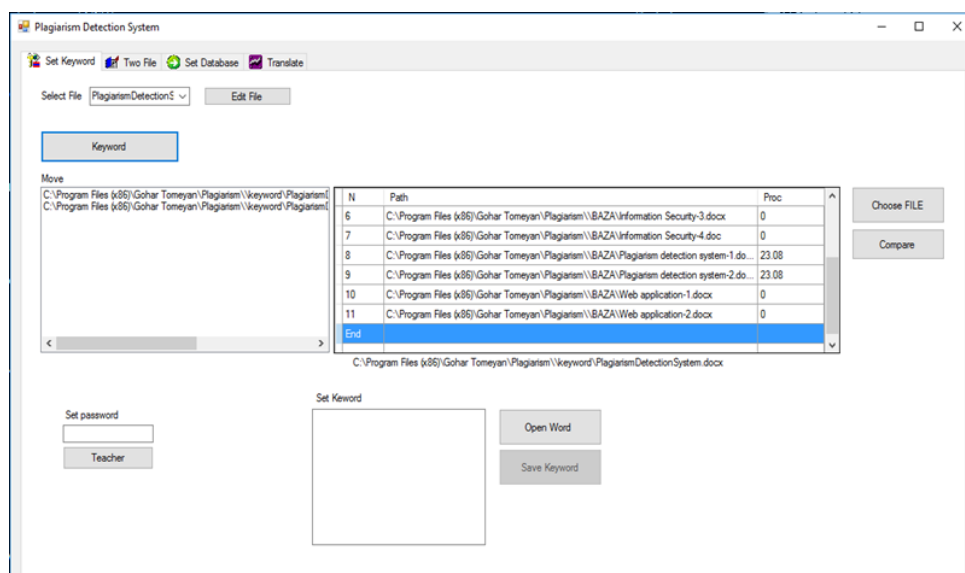


Figure A.1: Comparison of keywords

In the right part, you can see the result about how many percent are there in each document with same keywords that will appear after after choosing document and compare, the result is presented in Figure A.2.

N	Path	Proc
1	C:\Program Files (x86)\Gohar Tomeyan\Plagiarism\Keyword\PlagiarismDetectionSystem\P...	76.47
2	C:\Program Files (x86)\Gohar Tomeyan\Plagiarism\Keyword\PlagiarismDetectionSystem\P...	100
End		

D:\2015-2016\Magistr. tezi fayler\հոդված\Հոդված.docx

Figure A.2: Comparison of many documents

To see and edit file we need to put password and after that teacher can upload the file with the help of “Teacher” button. Then, the teacher can see the uploaded file name in the list (using password of teacher), he can create his own keyword folder and the system will be prepared to compare documents in this new knowledge domain. Administrator can add the keyword directly from our program The result can be seen in Figure A.3.

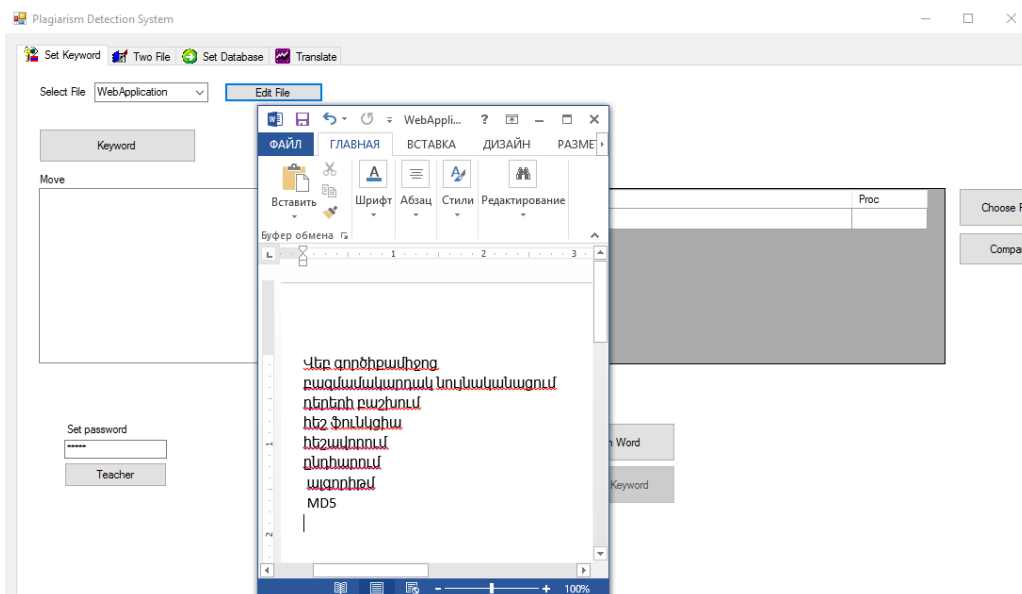


Figure A.3: Edit and see the keywords

After choosing the keyword domain, we need chose the file, which we want to compare and program will compare with files, which have the same keywords in the system database.

A.2 Comparing two File

Here teacher can compare only two files. First she/he need to choose the original document, and other from own database. Now the systems compares the texts that are in a directory, which is located in the directory of the Program that is the installation file. For example C:\ProgramFiles\ (x86)\GoharTomeyan\Plagiarism\BAZA.

After choosing, you can see the result by click on the button "Run", and we can see percentage of the same text, and exactly the same text in yellow color, which is presented on the Figure A.4.

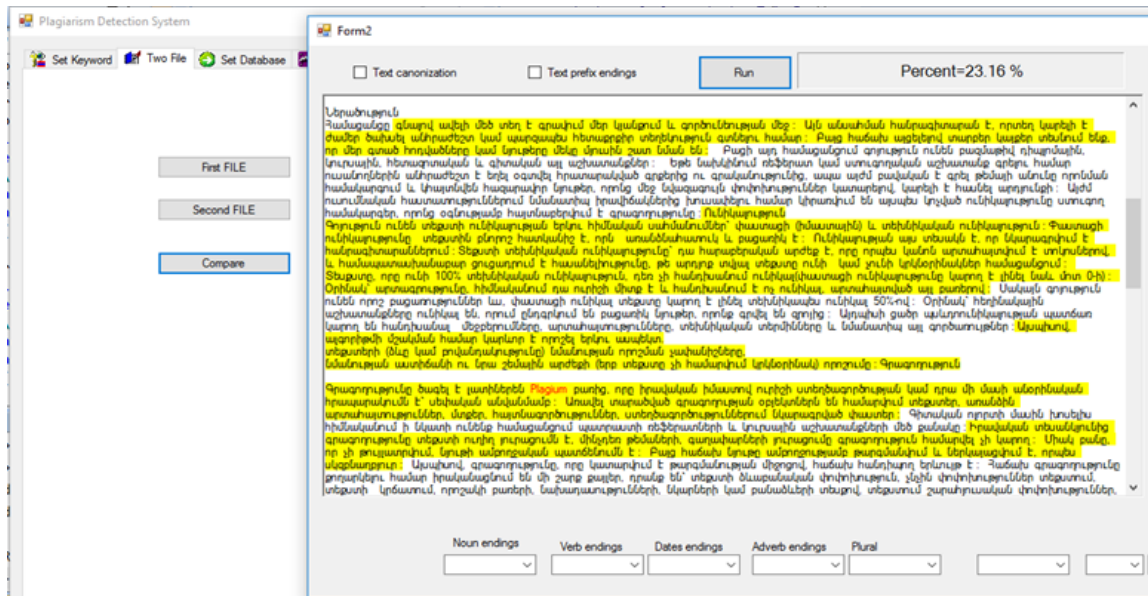


Figure A.4: Comparison of two documents

If we want to delete stop-words, you need to choose “Text canonization” check box and run. If we want to delete endings we need to choose second check box. In above in list box we can see all endings for Armenian language. If we want to replace words

with synonyms, the program has an option, which points out words in red color. And replace with synonyms. You has the opportunity to point out words in red color, choose the meaning, which corresponds to the context and save changing. After we can run and see result on the Figure A.5 by percentage.

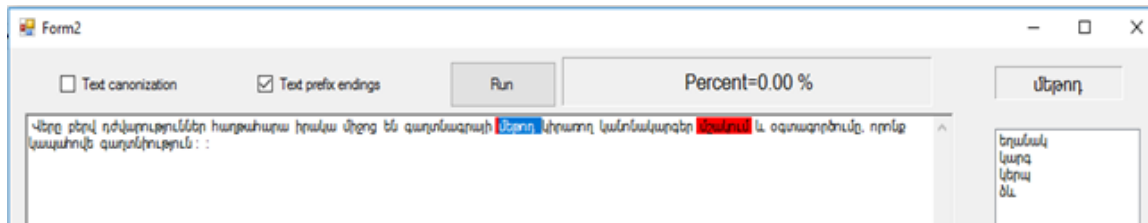


Figure A.5: Using stemming and stop word removal

A.3 Set Database

Teachers can add and see the synonyms, which are existing in our database. The teacher can delete synonyms, which he/she wrote, only he/she need to click on the “Delete Button”. At first teacher need to write synonym, explanation and choose the add button, after in combo box we can see all the explanation, if its correct you can write in database. The result can be seen in Figure A.6

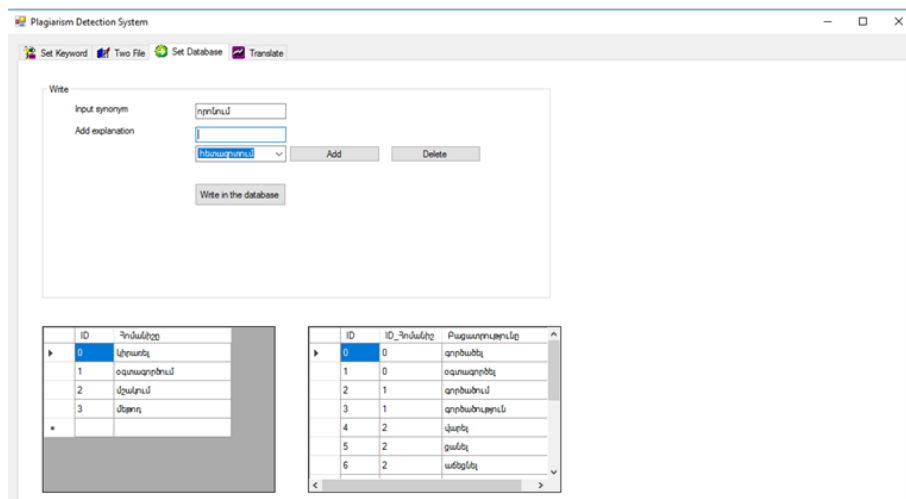


Figure A.6: Add synonyms

A.4 Translator

If we want to translate the document we need to choose the document, when program finished translation, the text must be copied to Microsoft Word(opened from our system), and after that the file can be uploaded to the database. Then, the document can proceed to the other steps of the system:: choose keywords and compare with many documents or compare only two documents, stop word removal and synonym recognition (Figure A.7).

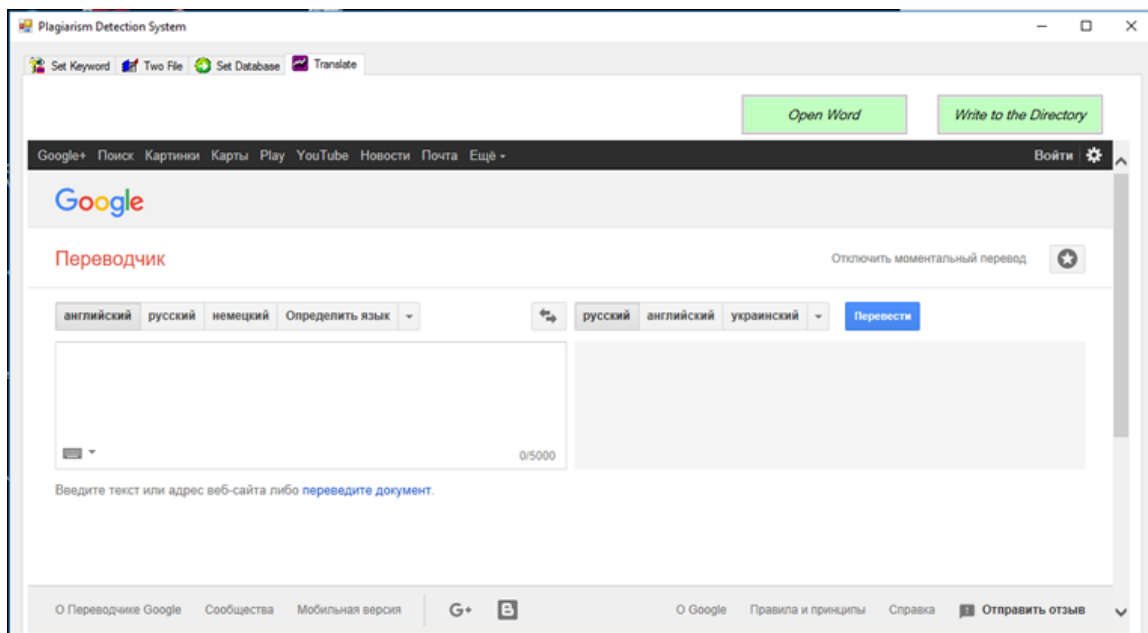


Figure A.7: Translator

