SUDAN UNIVERSITY
FOR SCIENCE AND TECHNOLOGY

College of Computer Science and Information Technology

Developing an Interpretation Corpus for Scientific Terms (Medical and Computer Science Terms)

تطوير معجم مفسر للمصطلحات العلمية (المصطلحات التقنية والطبية)

A Thesis submitted in partial fulfillment of the requirements of Master of Science degree in computer sciences

By: **Azhar Raheem Mohammed**

Supervision of

Dr. Wafaa Faisal Mukhtar

March 2019

DECLARATION

I hereby declare that the work reported in this M.Sc. thesis titled as “Designing translations corpus for scientific terms” submitted College of Computer Science and Information Technology / Sudan University of Science and Technology, is an authentic record of my work carried out under the supervision of Dr. Wafaa Muktar. I have not submitted this work elsewhere for any other degree.

Azhar Raheem Mohammed	Dr. Wafaa Mukhtar
______________________	______________________
Student	Supervisor

DEDICATION

To my family specially to whom I carry his name with pride my father Raheem Mohamed, and to the greatest woman I ever met that always support me, my mother, who supported me to reach advanced educational levels, and to my brother and sisters as they are my strength and thanks for everything.

I also would like to extend my sincere thanks to my colleagues, lecturers and the others for contributing and supporting me directly and indirectly. Thanks for your support, comments and advice.

Finally, thanks to every person's commitment in making this project successful.

ACKNOWLEDGEMENT

I wish to express my deepest appreciation to all those who helped me, in one way or another, to complete this project. First and foremost I thank God almighty who provided me with strength, direction and purpose throughout the project.

Special thanks to my project supervisor Associate Professor Dr. Wafaa Mukhtar for all her patience, guidance and support during the execution of this project. Through her expert guidance, I was able to overcome all the obstacles that I encountered in these enduring ten months of my project.

In fact, she always gave me immense hope every time I consulted with her over problems relating to my project.

Abstract

The Corpus is one of the most important tools in learning and understanding new terms. There are many corpuses used for the English language which can help interpreting general terms. The scientific terms may have different definitions according to their uses.

This research highlighted this problem and proposing a specialized corpus for technical and medical terms. A database was collected using crawling methods from dedicated dictionaries to provide scientific definition and uses for these terms. These definitions were filtered manually and a scientific corpus was implemented. A search engine was designed to evaluate the usage of the developed corpus.

Medical and technical terms were easily interpreted and the result of the search engine was displayed in both Arabic and English languages. The search engine web page was evaluated and proved to satisfy the users. The databases can be extended and more fields of science can also be added.

المستخلص

المعجم هو احد اغلب الادوات المهمة في التعليم وفهم المصطلحات الجديدة. هناك عدة معاجم مستخدمة للغة الانكليزية حيث يمكن ان تساعد في تفسير مصطلحات بشكل عام . المصطلحات العلمية من الممكن ان تختلف في التعريف طبقا لاستخدامها, ولا يكون هناك معجم خاص لحقل معين لكل العلوم. هذا البحث يسلط الضوء على هذه المشكلة ويقترح معجم خاص للمصطلحات التقنية والطبية. قاعدة البيانات قد جمعت باستخدام طريقة الزحف من قواميس محددة لكي يتم تجهيز تعريفات علمية واستخدامات لهذه المصطلحات. هذه التعاريف قد تم فلترتها يدويا وبعد ذألك تم تنفيذ المعجم العلمي. ماكنة البحث قد صممت لتقييم استخدام المعجم المطور. المصطلحات التقنية والطبية قد فسرت بشكل سهل والنتيجة ان ماكنة البحث عرضت النتائج بلغتين (عربي وانكليزي) ماكنة البحث في صفحة الويب قد تم تقيمها وتجهيزها لإرضاء المستخدمين. قاعدة البيانات يمكن ان توسع ويتم اضافة اكثر من حقل لمختلف العلوم في المعجم نفسه.

Table of Contents

List of Figures

Chapter One

Introduction

INTRODUCTION

1.1 Preface

It is well known that new life is highly affected by the factor of information technology, and technology plays an important role in today's human society development.

Based on this fact, it is indispensable to take advantage of the modern technological facilities in aiding, and in terminology learning method.

A corpus is an arbitrary sample of language term, aims to be a systematic account of the lexicon of a language [kilgarriff ,2005]. Terminology corpus has become is one of the most important tools, method in learning. The terminologies, and the traditional method which is papers are replace by advanced method which is electronic one [nomass,2013], electronic corpus is more intelligence and sufficient rather than paper dictionary this electronic corpus can be used and accessed by mobile, tablet and computer. E-Corpus has the potential to be a useful instrument which helps in many fields in education level.

The process of translation is increasing every day with appearing of new terms continuously, the issue is arising with the need of translate scientific terms of different science, which considered one of the aspects in the translation section.

1.1.1 Corpus purpose

A corpus is made for studying languages and terms in many scientific field, It as with for the study in many scientific fields which can be provide the meaning of many terms, Other collections of corpora are made for other purposes will be discuss later.

Thus a well-designed corpus will reflect these purposes that made for, and the contents of the corpus should be chosen to support it is purpose.

1.2 Problem statement

One of the most significant challenges that learners face during the process of science learning is terms learning. Term has been known as a central point in any field. Inadequate learning knowledge of the terms led to problems related to the source of knowledge .thus some issues led more investigation such as:

1. The need of translating scientific terms in a correct manner.

2. Most of this term is consist of group of letter that most dictionaries do not support the translation of this kind of written abbreviation.

1.3 Research questions

The main research questions to improve the design corpus and implementation it, in order to answer the main research question, the following will need to be answered:

1- It can work on interpreting words and terms together?

2- Does the web application design have been used more than science?

1.4 Objectives

This project aims to:

1- To design corpus that can interpreting words and terms together.

2- Develop web application for corpus that contains scientific terms in many sciences.

1.4 Scope of research

1- This research is restricted with that to create web application that with many design and coding tools.

2- With queries search engine the application contain as sample two science fields which is computer science and medicine.

3- Use software and programming languages like PHP and HTML, Java script, MYSQL, Ajax.

1.5 Methodology of research

The figure (1.1) is drawn to indicate that the system is built on the web application in order to search about the terms, figure out meaning, and give a description about each of these terms.

The application is built with the instruments each one have specific function and process, for the design and the functionality of the application the HTML and PHP programming language is used to create the page and design, second the corpus need to create database that can be used to store the terms and the meaning of this terms.

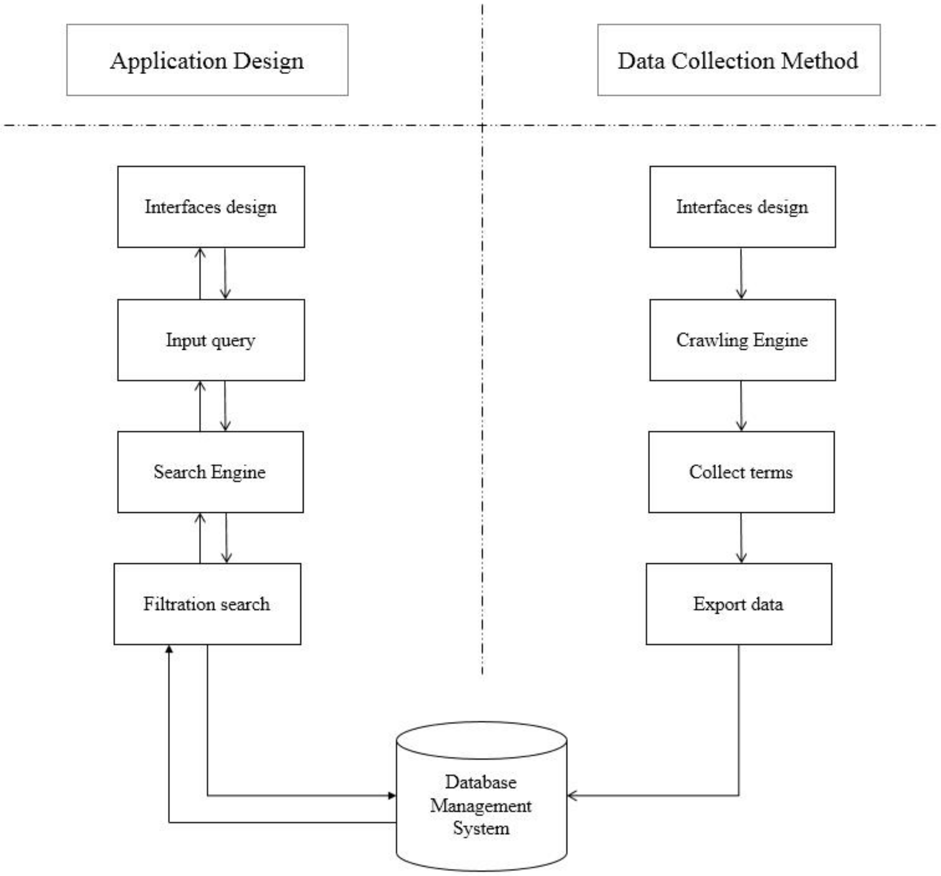


Figure 1.1: Research Methodology

1.6 Organization of thesis

This thesis is organized as follow. Chapter two presents the main concepts, theoretical focus that in the area of project and general information about system architecture related work. Chapter three presents research methodology, research design and procedures were presented as well as equipment. Chapter four which contain result and discussion about the implementation of work and finally the conclusion and future work are Chapter five.

CHAPTER TWO

LITERATURE REVIEW

Chapter two

Literature Review

2.1 Introduction

This chapter focuses on providing historical information, theoretical base information for the study such as operation and function of corpus, classification and tools and then related work in the field is also covered.

2.2 Brief history of Corpus

The earliest attempts to create 'corpora', was in beginning of the 17th century when the very first attempts were made by Samuel Johnson and his companions. They created a set of examples of the using real language in order to create what would be today called a corpus [sulc,1999] similar to Oxford English Dictionary.

After that many project was established in area of non-electronic corpora such as Survey of English Usage (SEU) Corpus, the aim which was to collect samples of texts and voice records (about 5.000 word entries), in late of 1960^{s} when computers were introduced into many scientific fields. The potential of electronic corpora was appeared at that time in period between the 60^{s} and 1980^{s}; it was called 'The First Generation Corpora'.

The Second Generation Corpora is began at late of 1980s and the early 1990s by establishing the first project of Birmingham University and led by J. Sinclair. Collins Birmingham University International Language Database (COBUILD) was primarily meant to serve as a text bank for a new dictionary of English. After that many corpus project are established the most famous one was known as British National Corpus (BNC).

2.3 E-Corpus Overview

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research [wynne,2005].

The basic resource for corpus linguistics is a collection of texts, called a corpus. Corpora can be of varying sizes, are compiled for different purposes, and are composed of texts of different types. All corpora are homogeneous to a certain extent; they are composed of texts from one language or one variety of a language or one register, etc. They also are all heterogeneous to a certain extent, in that at the very least they are composed of a number of different texts. Most corpora contain information in addition to the texts that make them up, such as information about the texts themselves, part-of-speech tags for each word, and parsing information [macdonald,2014].

2.3.1 Types of corpora

This section attempt to introduce some of the main types of corpus, so due to the fact there is mainly three types of the corpus are discussed this article [bowker,2002] and they are as follow:

2.3.1.1 General [reference] corpora

This kind of corpus is in defining what a general corpus is, one may encounter more descriptions since "within the hierarchy of corpus types, a general corpus appears to be the superordinate in the hierarchy" [pearson,1998], The text in the general corpus are homogeneous, and designed to be represent the characteristic vocabulary of a language.

Reference corpora are at the heart of the future development of corpus-based work in Europe and elsewhere. Reference corpora in several languages, constructed on similar principles, form a group of comparable corpora, for example (**The Bank of English**).

2.3.1.2 Monitor corpora

This type of corpus is corpus where text or terms are filtered "scanned on a continuing basis" to extract data from database [atkins,1992], Monitor corpora were originally meant to remain the same size with the new material relegating the old one. This model gave rise to the idea of rate of flow as the best way of managing the corpus. Instead of setting, say, 10 million words as the proper proportion of that genre, the setting could just as easily be 10 million words a year.

The language would flow through the machine, so that at any one time there would be a good sample available, comparable to its previous and future states, for example **(Corpus of Contemporary American English**).

2.3.1.3 Special purpose corpora

A special purpose corpus is one that focuses on a particular aspect of language. It could be restricted to the following state for:

1- LSP of a particular subject field, to a specific text type.

2- To a particular language variety.

3- To the language used by members of a certain demographic group.

2.3.1.4 Written corpora

Written corpora contain written texts, i.e. texts that were written for the purpose of being read. When designing a written corpus, it is also necessary for a linguist to decide what types of texts the corpus should contain. The analysis and processing of various types of corpora are also the subject of much work in computational linguistics, speech recognition and machine translation, where they are often used to create hidden Markov models for part of speech tagging and other purposes. Also there is many other type of corpus but the above is the main type while other types is (**Monolingual corpora, Bi- and multilingual corpora, Comparable corpora, Parallel corpora, Diachronic corpora, Open corpora, closed corpora**).

2.3.2 Corpus processing tools

In this subsection, main corpus processing software is listed as well and their basic features are described too. Today, corpora are now processed exclusively by means of specialized software tools, which is great progress terms researchers who once had to search and process terms and text manually.

Computer aided corpus-processing tools not only save money; time and human work but also provide accuracy of meaning. There are a variety of software tools for processing corpora such as Wordfisher, Corpus Builder, BNC, Gestrolex or Shoebox. Some of them may contain. A one specialized tool, whilst others may be fairly capacious and very complex. All most of these software tools are quite user friendly, while others, particularly at the beginning of the computer revolution, were not easy to use and it require some computer uses knowledge [quirion,2003].

2.4 Related Work

Numerous research works are going on these days in this field for better improvement in the performance and functionality of Corpus systems.

The development of electronic corpus have been covered by [meitie,2017] which present deals with the development of English to Manipuri electronic dictionary which is based on a database model.

The method that applied was dictionary developed in the search which is one of the most popular searching techniques. The software implemented is PHP as the front end for programming and MYSQL database as the Back-end, the Electronic-Dictionary.

Those developed will gain positive remark from computer experts and Manipuri language and terms learners. Also show that the E-dictionary can be used to help in teaching process smooth and it is a time-saving application with friendly user interface.

In the other perspective [macdonld, 2014] discusses the resources and methodologies used by corpus and then moves on to some key observations relating to comparative frequency and to patterning. Also they mentioned the importance of corpus linguistics for linguistic theory and present some of the applications that related to the corpus research. They give reviews on the corpus software development and methodologies and the corpus translator. As computer resources, particularly web-based ones come within the reach of the ordinary translator, language learner.

Terms search or linguist, the shoeing of many detailed information about Electronic dictionaries viewed in [zheng,2016], he state that the E-dictionary have been become more and more attractive, accepted and popular to learners at different levels.

Using electronic dictionaries in scientific field for search about terms has gradually become an alternative to many. As for teachers, helping students tap into electronic dictionaries effectively is one of the best ways to help them become independent, lifelong language learners.

The functionality of electronic dictionaries and why they are popular classes will be introduced. Later some of the current issues related to the integration of electronic dictionaries into instruction and learning will be identified and discussed, although recent dictionaries for the market have been developed for their function, design features, the main concern of all users, lexicographers and meta-lexicographers of these corpus or dictionaries is the functional quality of the dictionary product.

The functional quality of dictionaries and the scientific assessment this topic of this paper [metruk,2016]. The functional quality of corpus is defined in section number two. the next section contain current methodological approach to assessing the functional quality of texts in the fields of web design, the design and document design is also have been covered and discussed and its relevance for dictionary design is indicated. Finally is showed that with depth details how this methodology can be used to design dictionaries and to assess the functional quality of the design features of existing dictionaries, another idea for provide the terms meaning and words electronic pocket dictionaries as a language learning tool is discussed in this thesis [jian, 2009].

On study have been implement among university students in Hong Kong and Taiwan. The targets of the studies are group which included students. Speed of reference was found to be the main motivator for using these types of dictionary. Finally, multimedia content was ranked least important.

> The results of this study have implications for the design of electronic dictionaries and for the teaching of second languages with electronic dictionaries. For device the developer can getting focus on enhance the accessing speed against multimedia. Educators should ensure that the device functionality matches the language proficiency level of the students.

The Tilde Dictionary Browser (TDB) [deksne, 2013], an innovative dictionary browsing environment for a wide range of users such as: language learners, language teachers, translators, and casual users, TDB also provides information from different online resources, such as terminology dictionaries, as well as integrates the machine translation facility, also this work represent simple search and browsing, which can supports different language and terms technology driven services that facilitate better retrieval method of requested entries.

TDB can be used on different platforms, including mobile devices and the Web, for future work they show suggested closer integration with machine translation that allowing users to translate a full document instead of a phrase, sentence, or small fragment of text.

Here the article [al-sulaiti ,2005] displays the important tips that about the development of the Corpus of Contemporary Arabic (CCA), including design, collation and deployment of the initial version. Also give a brief report on the initial stage of the development of the Corpus of Contemporary Arabic (CCA). also it other feature for their corpus which is that spoken texts and parallel texts and this method important to be included and can be achieved within the second stage of development of the corpus, this in addition illustrated the different aspects of the procedure that would follow to create any corpus. In same hand [atkins ,1992] discussed important things about corpus design criteria which is start by creating object of a corpus, and the constituents of it, texts, and then constraints on the sort of documents, after that browse the practical stages in the process of establishing a corpus. Also pointed out the major difficulties in defining the population of texts of corpus, at final they focused on the context and the functions of the corpus and corpus when it’s implemented.

2.5 Related work summary

Paper	objectives	Methodology	Tools
Corpus Linguistics [atkins,1992]	To display methodologies used by corpus linguists and then moves on to some key observations relating to comparative frequency and to patterning.	Corpora are interrogated through the use of dedicated software, the nature of which inevitably reflects assumptions about methodology in corpus investigation	Web-based App Html PHP
The use of electronic Dictionaries in EFL Classroom [zheng ,2016].	To provide detailed information about Electronic dictionaries, To using electronic dictionaries in scientific field for search about terms has gradually become an alternative too many things.	Use of electronic dictionaries EFL (English as a Foreign Language) in classrooms among Chinese University students.	Survey
Corpus-Based Terminology Extraction [alexan dre ,2005]	To presented a way to generate a term extractor taking as prior knowledge a training corpus where the terms are identified.	The extraction is usually done using a hand-crafted automaton. In this work, we want to generate automatically the automaton from the training corpus.	hand-crafted automaton method
The role of electronic pocket dictionaries As an English learning tool among Chinese students. Journal of Computer Assisted Learning	This study have been implement among university students in Hong Kong and Taiwan. The target of the studies are group which included students. To provide new way for	Survey based on two section A questionnaire was designed comprising two parts. The first part of the questionnaire was completed by all the respondents and addressed	-------

[jian,2009]	using these types of dictionary with multimedia content was ranked least important. To design of electronic dictionaries and for the teaching of second languages with electronic dictionaries.	general issues related To dictionary use. The second part of the questionnaire was only completed by respondents who reported using an electronic dictionary regularly.	
The modern electronic dictionary that always provides an answer[deksne, 2013]	To innovative dictionary browsing environment for a wide range of users: language learners, language teachers, translators, and casual users TDB also provides information from different online resources, such as terminology dictionaries, as well as integrates the machine translation facility	To Develop dictionary software that is able to provide useful information for all types of search queries and information needs, including many problematic cases when searched items do not have any direct matches in the dictionary data.	XML Web Application tool JavaScript HTML
Extending the corpus of contemporary Arabic [al-sulaiti ,2005]	To discuss important tips that about the development of the Corpus of Contemporary Arabic (CCA) , including design, collation and deployment of the initial version. Also give a brief report on the initial stage of the development of the Corpus of Contemporary Arabic (CCA.	Set up infrastructure and prototype sampler corpus for the International Corpus of Arabic, an international collaborative research program to parallel the International Corpus of English.	Web pages PHP HTML JavaScript
	To discuss important things	Specifications and design.	- - -

Corpus design criteria[atkins,1992]	about corpus design criteria which is start by creating object of a corpus, and the constituents of it, texts, and then constraints on the sort of documents, after that browse the practical stages in the process of establishing a corpus.	- Hardware and software. - Data capture and mark-up. - Corpus processing. - Corpus growth and feedback
	To points out the major difficulties in defining the population of texts that the corpus will sample, at final is focused on the context and the functions of the corpus and corpus when it’s implemented.	

2.6 Summary

This chapter have been discussed the basic information about corpus such as operation and function of corpus, classification, tools and the historical information, also the published articles have been mentioned finally the related work that have been done in the area of the project.

The difference between my work and the work of [deksne ,2013] is that they worked a large area of users and took into account many of the information and experiences of teachers and translators to develop their own corpus, in the work [al-sulaiti ,2005] they developed infrastructure is a basis for the establishment of a universal Corpus for the Arabic and English language, My work is characterized by simulating a range of specific purpose that may not exist in previous corpuses.

Chapter Three

METHODOLOGY

Chapter three

METHODOLOGY

3.1 Introduction

This chapter demonstrates the research design and procedures were presented as well as instrumentation, and the explanation of the methodology to be followed to achieve the research objectives.

To implement the desired system architecture, many tools and techniques will be needed and used to perform many tasks; the figure gives summary of the tools and method that will be used to achieve these tasks.

3.2 Methodology and tools

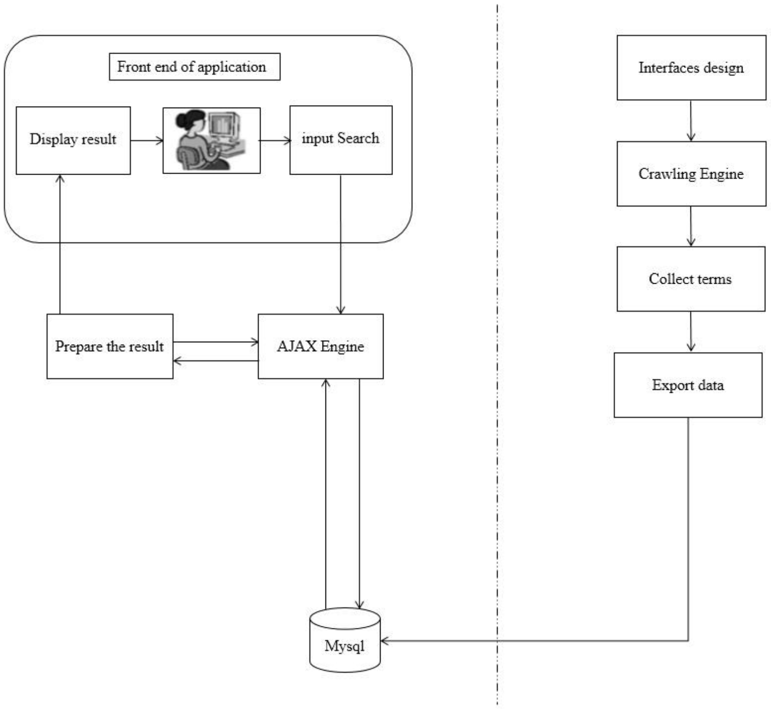


Figure 3.1: System Methodology view

The figure 3.1 shows the method the followed to create the system , the system contain two part first is concern about collect the data for the application and Demonstrate as clearly the crawling method in the internet by using the site that needs to get the terms form it. The step is started with user when open the browser which contain input field on for insert the website URL(It should be noted that the best five websites containing a set of terms and definitions have been used, these sites have been used to complete the data collection stage) and the with filter and crawl function the content the site will passed to the search engine process which collect the terms tags and then this term will be filtered to remove the tags and store this data and terms in the excel file, this file contain all data wanted data and not wanted so is need to be next steps which is refinement. The refine step is will be to remove any undesirable data and content, after this step as the result it get the data which is terms and the meaning of terms in excel sheet than can be easily imported to database, that can be used later in the query of search for the terms.

For the second part of the application provide web page application this page used to get quires about terms and search and the meaning, this page contain one input filed that used to insert the desired term, this input field is directly connect with AJAX application for search engine this method allow to application to provide prediction about the term that the user looking for according to the letter that the user going to insert them specifically is AJAX look according to the first letters matched that user insert them and then provide all probability for this match, the method can help the user to recognize the terms.

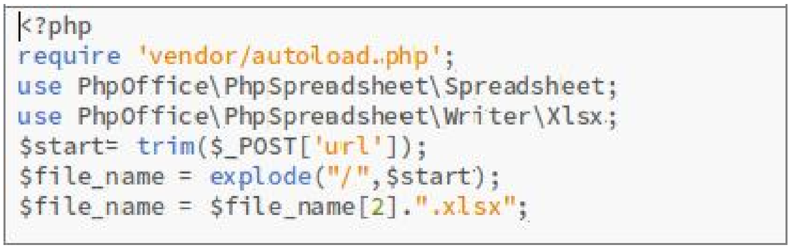
The application is built with the instruments each one have specific function and process, for the design and the functionality of the application the HTML and PHP programming language is used to create the page and design, second the corpus need to create database that can be used to store the terms and the meaning of this terms in this case we use PhpMyAdmin database to store this data, so until here almost done for the search engine that used to search on the application about terms the Ajax and JavaScript search engine which can be provide flexible search engine with the feature of prediction about the terms while it is writing.

Finally after all this component is collect together and the system have been build the data of the corpus is will be collect with the intelligent method which known as CRAWLING, this method used to filter the websites page and get the content of the pages and get the term tag and then the filter the tag and get the data that contained inside the tags, this data in next stage will be export as excel file and then import imported to be stored in database.

This application have been divide into two part the first is concern about how to crawl in the internet by using the site that needs to get the terms form it, all this procedure is illustrate in figure shows.

3.2.1 Crawling main engine

Import necessary file for the function



```
<?php
require 'vendor/autoload.php';
use PhpOffice\PhpSpreadsheet\Spreadsheet;
use PhpOffice\PhpSpreadsheet\Writer\Xlsx;
$start= trim($_POST['url']);
$file_name = explode("/",$start);
$file_name = $file_name[2].".xlsx";
```

In these first line of code the calling and include of the functions, the auto-loader function required to create and load the Excel file also two function which is need with which is:

use PhpOffice\PhpSpreadsheet\Spreadsheet;

use PhpOffice\PhpSpreadsheet\Writer\Xlsx;

The first sub-function is spread sheet which is excel file which can be created when you trigger the crawling button. Then the calling of the extension of spreadsheet method which is (.xsl), second one, as seen there is variable called start which have value the URL for the website the want is to crawling in it.

Well; after create the excel sheet file the next needing is to create an array used for crawled data which this array used to store data after crawled and then put these data in excel sheet which by:

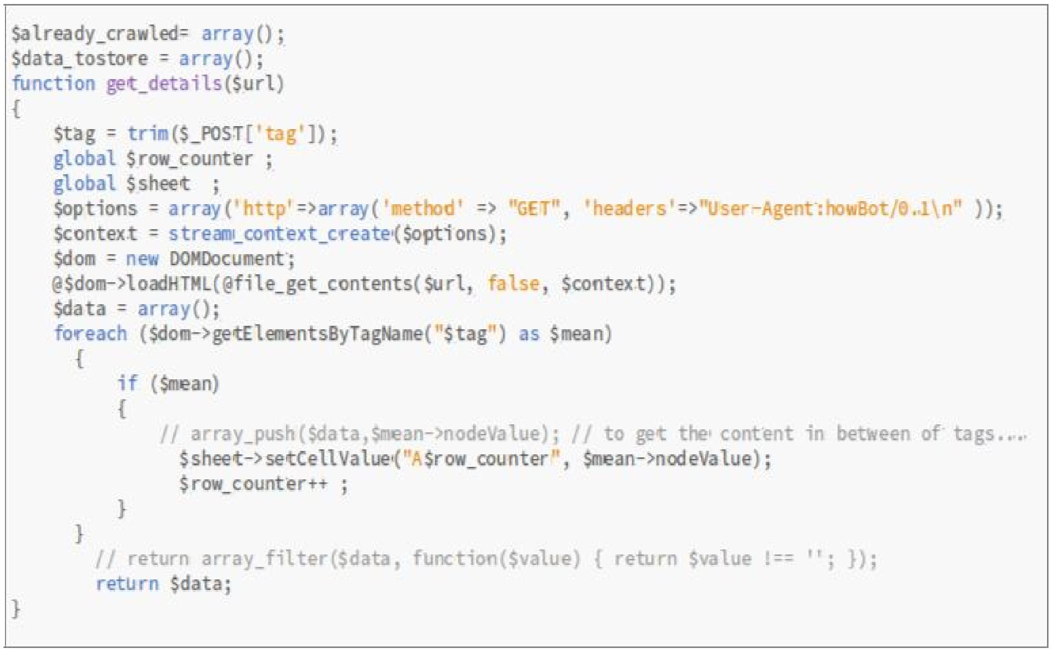
$already_crawled= array();

$data_tostore = array();

The figure 3.2 below explain the step about how to get the URL and the tag from user and passing this URL to the Crawl function to mining on it.

3.2.2 Extract and store the data

After import the necessary files and functions the next step will be is to create the function, main function the responsible form the crawling method, where is called to get URL and then return the data (terms) from the website.



```
$already_crawled= array();
$data_tostore = array();
function get_details($url)
{
    $tag = trim($_POST['tag']);
    global $row_counter ;
    global $sheet  ;
    $options = array('http'=>array('method' => "GET", 'headers'=>"User-Agent:howBot/0.1\n" ));
    $context = stream_context_create($options);
    $dom = new DOMDocument;
    @$dom->loadHTML(@file_get_contents($url, false, $context));
    $data = array();
    foreach ($dom->getElementsByTagName("$tag") as $mean)
      {
          if ($mean)
          {
              // array_push($data,$mean->nodeValue); // to get the content in between of tags....
                $sheet->setCellValue("A$row_counter", $mean->nodeValue);
                $row_counter++ ;
          }
      }
        // return array_filter($data, function($value) { return $value !== ''; });
        return $data;
}
```

Figure 3.2: Extract and store the data

In this main function have two global variable called

```
global $row_counter ;

global $sheet   ;
```

After that the following line will take place

```
$options = array('http'=>array('method' => "GET",

'headers'=> "User-Agent:howBot/ 0.1\n" ));
```

This line contain the array variable called option which get the data as method Get form this method get with user-agent value 0.1. these value or version of user method used by the browser for get and retrieve the data.

```
$context = stream_context_create($options);
$dom = new DOMDocument;
@$dom->loadHTML(@file_get_contents($url, false, $context));
$data = array();
```

The above line is used to create context for the option array and to get inside this context or data have to create new JQuery function called DOM Document, this function allowing to load the HTML code and getting the content of HTML tag as show in third line the create new data array , $data = array().

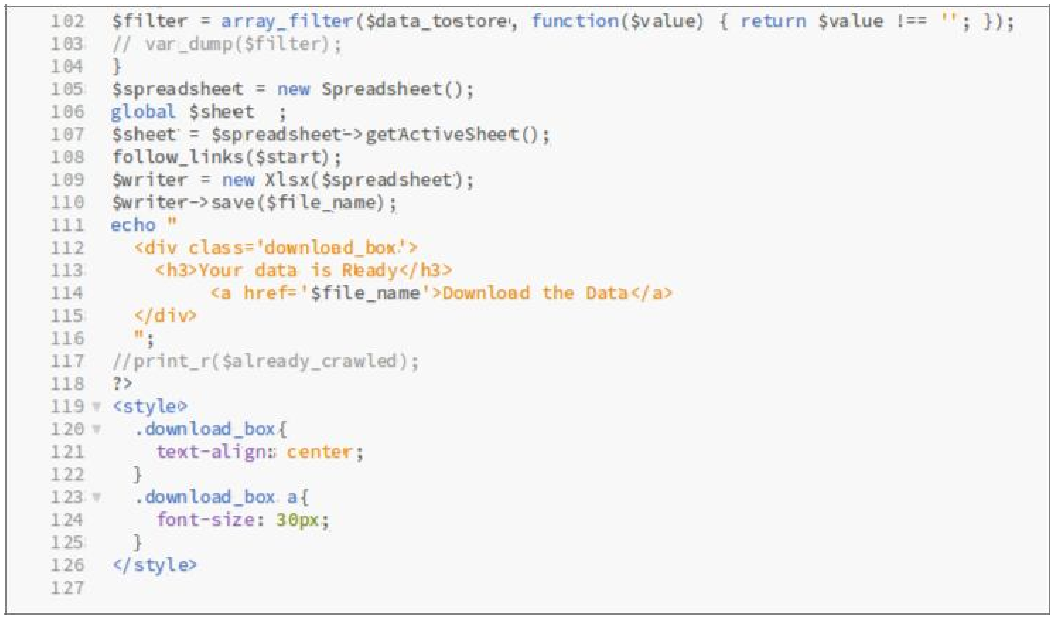
To filter the data inside the tag the DOM function have been used, for example the term are locate in the Paragraph HTML tag (p), can get this data by using predefined function which is getElementsByTagName('p') which only return the data the inside the html tag (P).

```
for each ($dom->getElementsByTagName('p') as $mean)
{
if ($mean)
{
//  array push($data, mean->node Value); // to get the content in between of
tags...
$sheet->setCellValue("A$row_counter", $mean-
>nodeValue); $row_counter++ ;
   }
  }
// return array_filter($data, function($value) { return $value !== '';
}); return $data;
}
```

For the if statement if in case if get value reassign the value of sheet variable to be converted into cell and the cell will be filled by the value of the $mean that returned by the for each loop

```
for each ($dom->getElementsByTagName('p') as $mean)
```

Then its need to increase the row-counter by one to be inserts the values until the loop finish. Then called the function or the variable to be executed, return $data; after that this long code which used to prepare the URL to crawling on it which by clear the error and unwanted data in these URL, the code of this step in Appendix.



```
$filter = array_filter($data_tostore, function($value) { return $value !== ''; });
// var_dump($filter);
}
$spreadsheet = new Spreadsheet();
global $sheet  ;
$sheet = $spreadsheet->getActiveSheet();
follow_links($start);
$writer = new Xlsx($spreadsheet);
$writer->save($file_name);
echo "
  <div class='download_box'>
    <h3>Your data is Ready</h3>
         <a href='$file_name'>Download the Data</a>
  </div>
  ";
//print_r($already_crawled);
?>
<style>
  .download_box{
    text-align: center;
  }
  .download_box a{
    font-size: 30px;
  }
</style>

```

Figure 3.3: Function of extract and download excel sheet

Here the variable is created for filter they are and store the data, and then define the following function to the clear any null value in the array

function($value) { return $value !== ''; }

After it should create the spread sheet variables, this step done by which can be activate by the *(getActiveSheet)* predefined function .

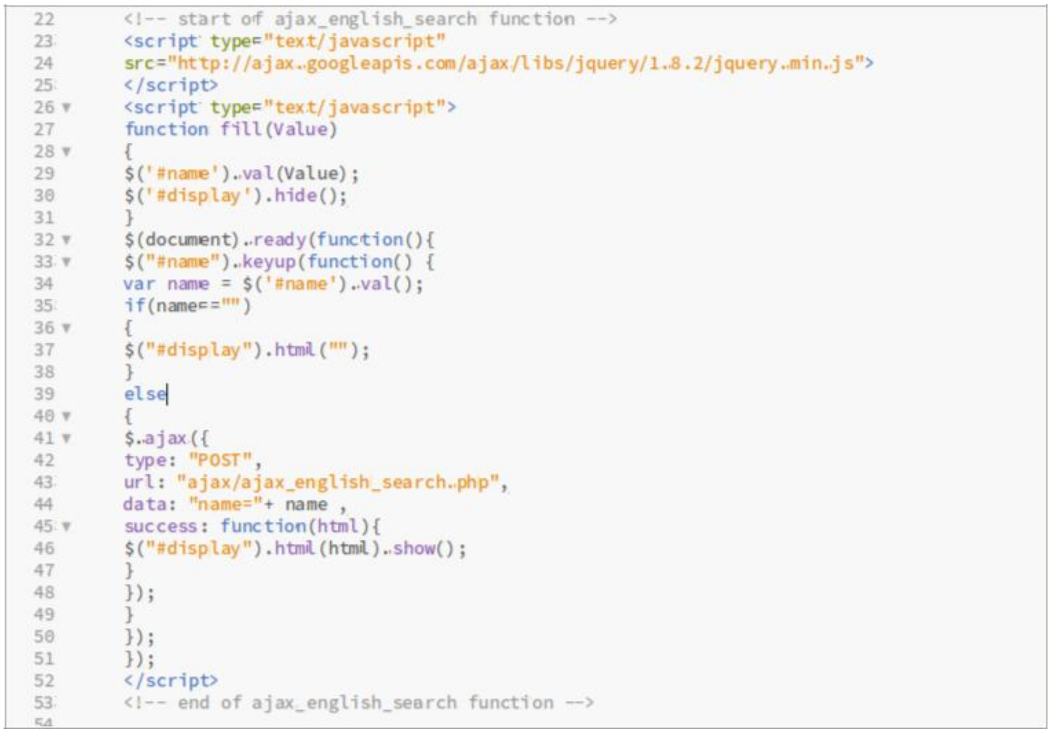
Then the creation of variable call writer to write the data on the Xlsx file which is excel sheet.

$writer = new Xlsx($spreadsheet);

$writer->save($file_name);

Finally for the final page, download link is created to download the excel sheet which is contain the crawled data only, that have been mining about it by using the element Tag (P).

After this discuss the important notice about the how to crawling data for the corpus here the final result for this section for application.



```
<!-- start of ajax_english_search function -->
<script type="text/javascript"
src="http://ajax.googleapis.com/ajax/libs/jquery/1.8.2/jquery.min.js">
</script>
<script type="text/javascript">
function fill(Value)
{
$('#name').val(Value);
$('#display').hide();
}
$(document).ready(function(){
$("#name").keyup(function() {
var name = $('#name').val();
if(name=="")
{
$("#display").html("");
}
else
{
$.ajax({
type: "POST",
url: "ajax/ajax_english_search.php",
data: "name="+ name ,
success: function(html){
$("#display").html(html).show();
}
});
}
});
});
</script>
<!-- end of ajax_english_search function -->
```

Figure 3.4: Main AJAX key pass function

To create the above function the below steps are followed:

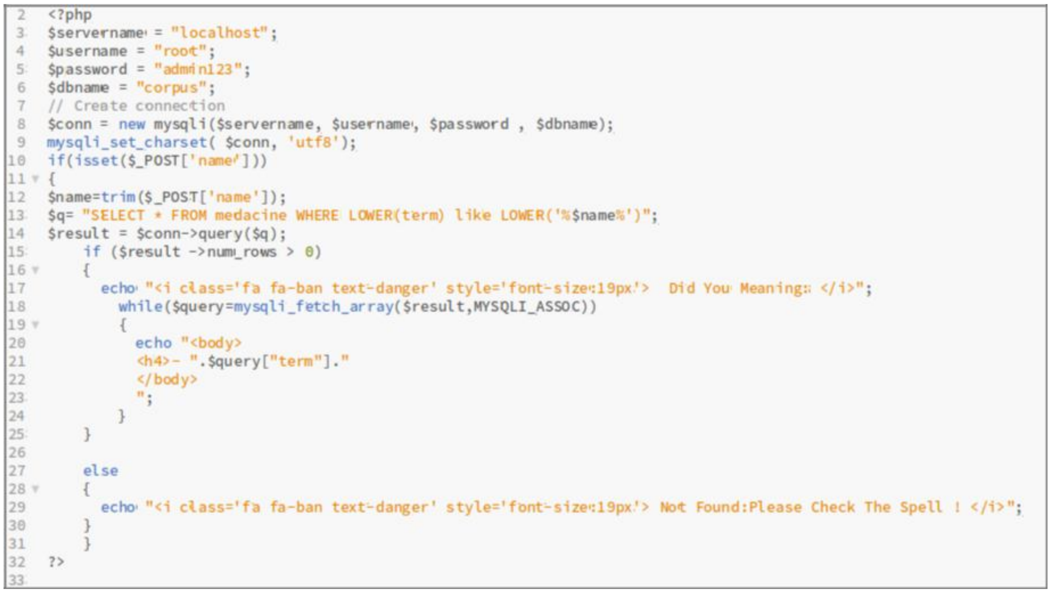
1. Call Google API for AJAX which can run AJAX which is on the above script which contains URL.

2. Create of the function named fill which take one value this value is the name of the input field on the form.

3. If the entered data is null as shown if (name ==””) then echo nothing in the Div that have Id = display.

4. Else if the form name filed not null then if the type of get data = POST take the parameter to AJAX folder which name.
5. Finally if function is success print these result on the Div which have Id = display, these print value by use show function.

After set the connection to database which contain the username and password and the name of the database the will search inside it, the figure 3.6 show these steps as follow.



```php
<?php
$servername = "localhost";
$username = "root";
$password = "admin123";
$dbname = "corpus";
// Create connection
$conn = new mysqli($servername, $username, $password , $dbname);
mysqli_set_charset( $conn, 'utf8');
if(isset($_POST['name']))
{
$name=trim($_POST['name']);
$q= "SELECT * FROM medacine WHERE LOWER(term) like LOWER('%$name%')";
$result = $conn->query($q);
    if ($result ->num_rows > 0)
    {
      echo "<i class='fa fa-ban text-danger' style='font-size:19px'>  Did You Meaning: </i>";
        while($query=mysqli_fetch_array($result,MYSQLI_ASSOC))
        {
          echo "<body>
          <h4>- ".$query["term"]."
          </body>
          ";
        }
    }

    else
    {
      echo "<i class='fa fa-ban text-danger' style='font-size:19px'> Not Found:Please Check The Spell ! </i>";
    }
    }
?>
```

Figure 3.5: Ajax and java script search engine

After if the user set the term or post the term as is define the above condition (if statement), if posting of the term for search, the term will be passed through the function is called trim () this function will change the case of the character in two wise (Capital or Small).

Then the query will be execute always will be select where equal to the variable which is name which is = the term that user insert it.

3.3 Searches, Software, and Methodologies

The proposed Corpus items are interrogated through the use of dedicated software, the nature of which inevitably reflects assumptions about methodology in corpus investigation. At the most basic level, corpus software:

Searches the corpus for a given target item, counts the number of instances of the target item in the corpus and calculates relative frequencies.

It is clearly that corpus methodologies are considered of quantitative. Indeed, corpus has been criticized for allowing only the observation of relative quantity and for failing to expand the explanatory power of linguistic theory.

For search bar any corpus, including a ‘raw’ corpus the raw use to looking about the word or the terms most search a Corpus software offer single search word or term to find sets of words and meaning [macdonald,2014].

3.4 Validation steps

After coding and design the desired system with two part, testing will be performed to this components to make sure the system application is running probably with fully functionality, first the testing for crawling engine by using many website URL, this method proof the crawling engine is work probably, for the corpus dictionary application the database will be fill with the terms, for each field 1000 term will be export to database with the meaning for each term.

3.5 Summary

This chapter has illustrations up to the research methodology which includes a detailed description instrumentation (tools and the system environment) that used to implement the proposed system, Also provided the system process overflow and how to validate and result presentation system, at the end of the above system processes, it should to presenting the results in order to compare them against objectives of research that must be achieved.

Chapter Four

RESULTS AND DISCUSSIONS

Chapter Four

Results and Discussions

4.1 Introduction

This chapter display discussion and result design of corpus and build the corpus system after install and design needed package for carrying out this work, it has been parted in section first implement crawl search engine to collect the data for corpus, second display that the final design of corpus.

4.2 Interaction with Corpus

The designed corpus operate in two mode , First the data mining or data collection method .In the first step create of web application page that is used to collect the terms form the internet gathered from websites page through multiple filter functions and scripts. In the second step is concern with the user of corpus, contain user simply friendly user interface that allow to user to search about the terms with assistance of many prediction method that can help to get desired term with effective and efficient manner.

4.2.1 Crawling (data collection) Mode

Simply in this mode the uses of crawling data mining method, it can help to gather the data (terms) for the corpus, as seen in the following figure 4.1 this is the main page for this mode.

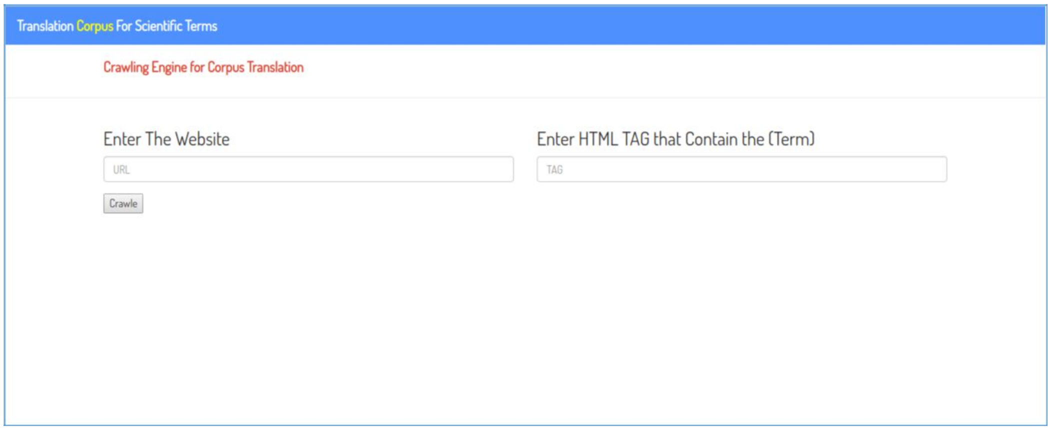

Translation Corpus For Scientific Terms
Crawling Engine for Corpus Translation
Enter The Website
URL
Enter HTML TAG that Contain the (Term)
TAG
Crawle

Figure 4.1: Crawling data mining engine

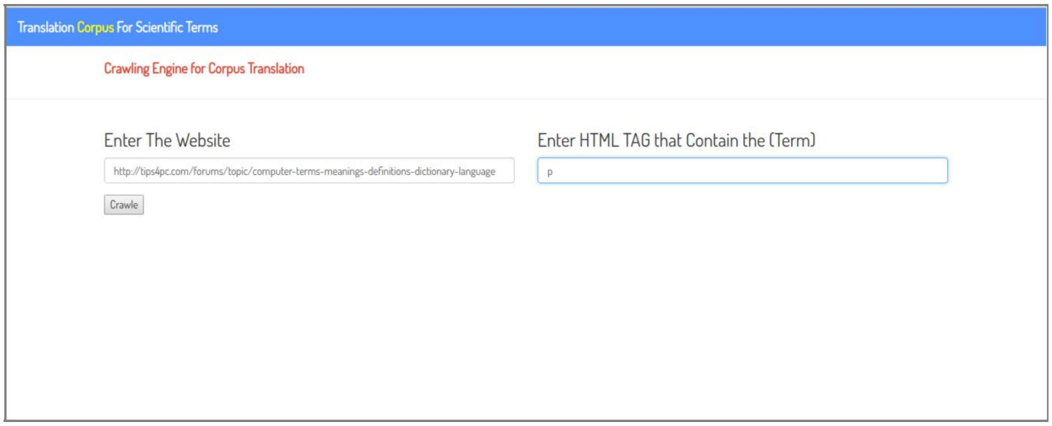
4.2.1.1 Crawling data mining engine

The above figure is show the main page for data collection for corpus, To operate with this data collector needed to specify two things; The website URL that needs of get the terms from it and the HTML tag that contain this terms, this method allow to mining function to retrieve the content of the tags quickly with clear data that are inside.

After the specify the needed item then, press the crawl button which is trigger many function in background .This function in order to mining in the given URL Based on the inserted tag, After all this procedure is done the out of this application is an Excel sheet file that contain the terms and the meaning of this terms which is will passed to the next step that demonstrated in chapter 3, This step is the refinement step, allowing to check and remove any undesired data, The main function the perform the mining step is illustrated below in next step.

4.2.1.2 Crawling mode

The figure 4.2 is display the main page for gather the data for application, as seen it contain the URL, here put the URL and the tag for search about terms.

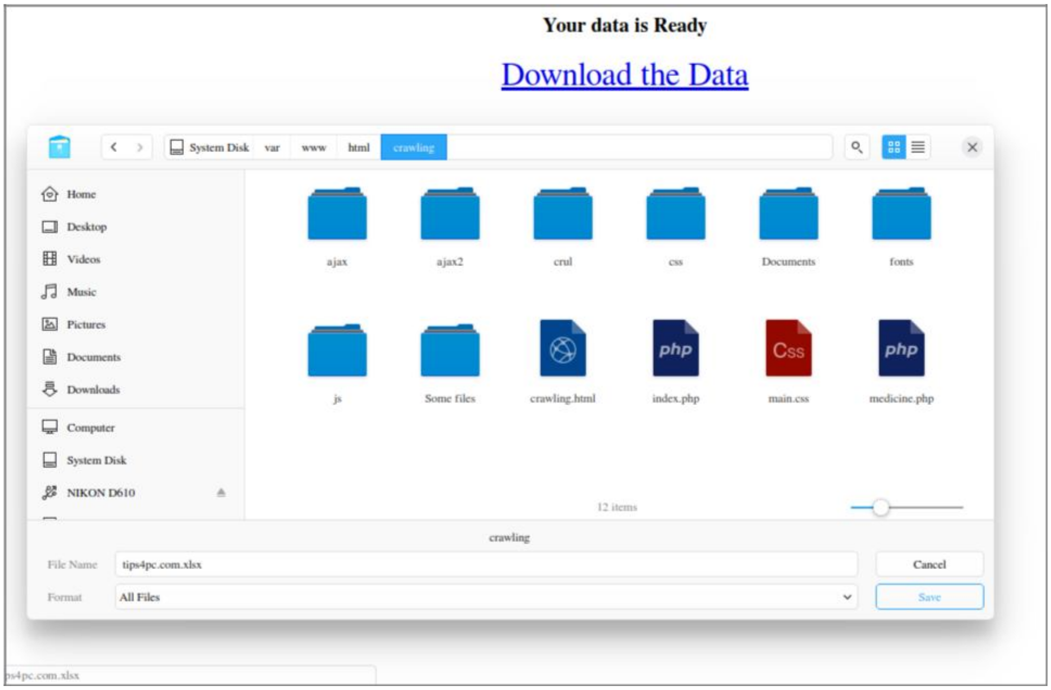

Translation Corpus For Scientific Terms
Crawling Engine for Corpus Translation
Crawle

Enter The Website: http://tips4pc.com/forums/topic/computer-terms-meanings-definitions-dictionary-language

Enter HTML TAG that Contain the (Term): p

Figure 4.2: Crawling method

After press the crawling button, waiting for seconds until the filter function done and the show the below page as showing in below figure 4.3.



Your data is Ready

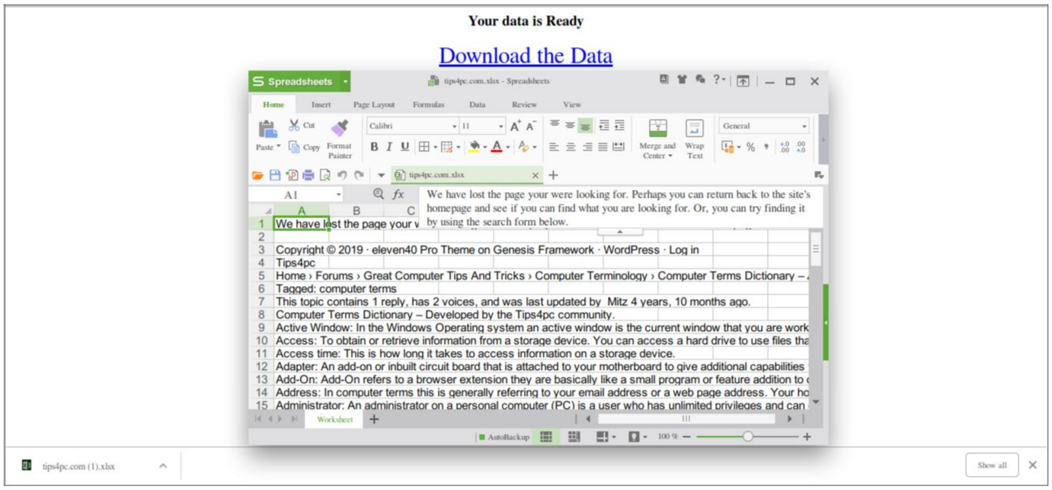
Download the Data

File Name: tips4pc.com.xlsx

Format: All Files

Figure 4.3: download the excel sheet data

The above figure the page after the crawling engine is done and the excel sheet is ready to be download, after press Download the Data link the file will be download to device as shown in figure 4.4 below.



Your data is Ready

Download the Data

We have lost the page your were looking for. Perhaps you can return back to the site's homepage and see if you can find what you are looking for. Or, you can try finding it by using the search form below.

1 We have lost the page your v
2
3 Copyright © 2019 · eleven40 Pro Theme on Genesis Framework · WordPress · Log in
4 Tips4pc
5 Home › Forums › Great Computer Tips And Tricks › Computer Terminology › Computer Terms Dictionary – A
6 Tagged: computer terms
7 This topic contains 1 reply, has 2 voices, and was last updated by Mitz 4 years, 10 months ago.
8 Computer Terms Dictionary – Developed by the Tips4pc community.
9 Active Window: In the Windows Operating system an active window is the current window that you are work
10 Access: To obtain or retrieve information from a storage device. You can access a hard drive to use files tha
11 Access time: This is how long it takes to access information on a storage device.
12 Adapter: An add-on or inbuilt circuit board that is attached to your motherboard to give additional capabilities
13 Add-On: Add-On refers to a browser extension they are basically like a small program or feature addition to
14 Address: In computer terms this is generally referring to your email address or a web page address. Your ho
15 Administrator: An administrator on a personal computer (PC) is a user who has unlimited privileges and can

Figure 4.4: download and view the data excel sheet

4.2.2 Corpus application

This part of the system contain the main page for the corpus that allow to user to query and search about the terms and their meaning, Figure 4.5 below is show the main page for this application.

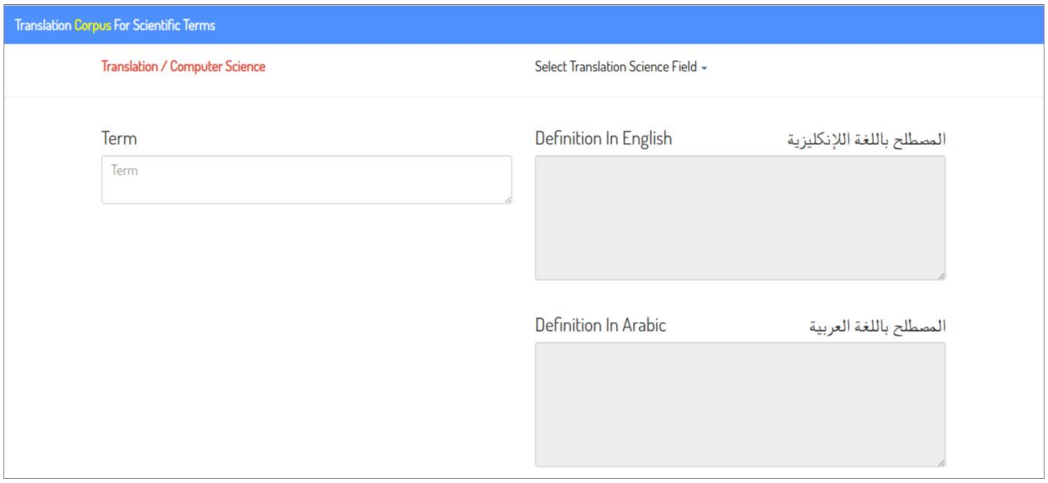
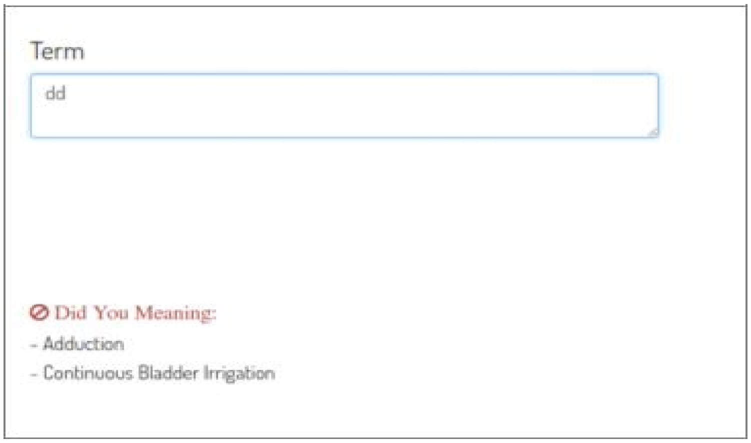

Translation Corpus For Scientific Terms
Translation / Computer Science
Select Translation Science Field
Term
Term
Definition In English
المصطلح باللغة اللإنكليزية
Definition In Arabic
المصطلح باللغة العربية

Figure 4.5: Corpus Application main page

The figure shows is show that the main page for the user of the corpus application, This page is used user to return the meaning of any terms, As show this page contain drop down list this list allow to user to select the scientific field that related to terms, when this field is selected next step for user is to enter the term, when this term have been insert while the user is writing the term letters this trigger the Ajax files this file contain the AJAX detection and search engine where is responsible for take the term that inserted by the user and then by the java script engine and AJAX search engine it looking for the term inside the database, if the term not found it show the prediction for this term like did mean.

For if there is more than one row of match for the user input with while loop will print this result, else if the result equal zero that mean will print the else condition to user.



Term

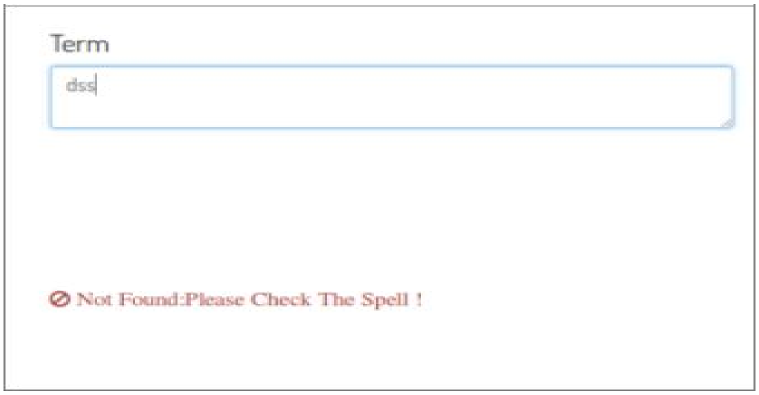
dd

⊘ Did You Meaning:

- Adduction
- Continuous Bladder Irrigation

Figure 4.6: Ajax predication engine

The figure 4.6 show that the result of search about term as shown that the prediction method in did mean bar, this method provide to the user list of closest term for the user insert.



Term

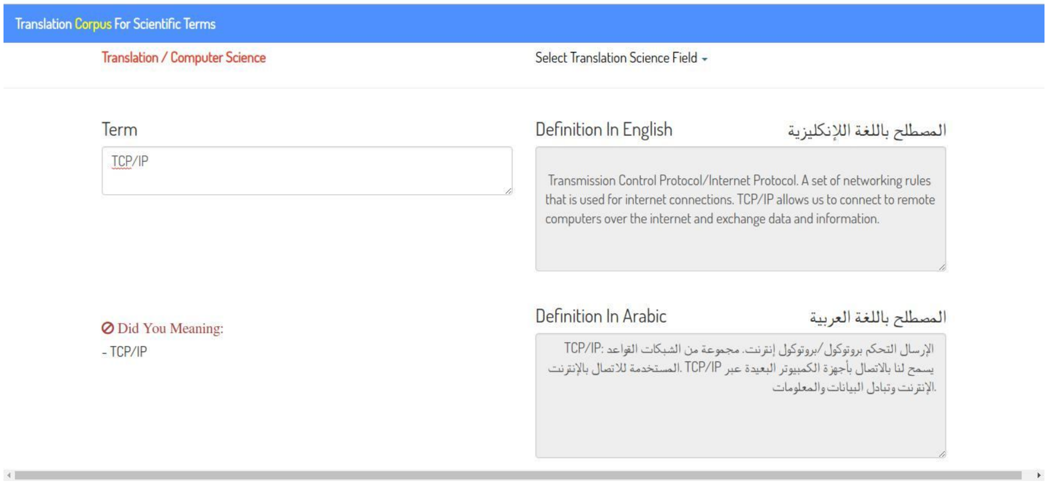
dss

⊘ Not Found:Please Check The Spell !

Figure 4.7: Unspecific term search result

In the case of there is no match for the entering of the user term the above 4.6 figure will be appear, and notified the user to check out the spell of the inserted term.

After above explanation of the main function that made up the application, the user will be ready to use the application to search about terms in specific scientific terms, the figure 4.7 is show the main page for this application with the sample of search about term.



Translation Corpus For Scientific Terms

Translation / Computer Science

Select Translation Science Field ▾

Term

TCP/IP

Definition In English — المصطلح باللغة اللإنكليزية

Transmission Control Protocol/Internet Protocol. A set of networking rules that is used for internet connections. TCP/IP allows us to connect to remote computers over the internet and exchange data and information.

⊘ Did You Meaning:
- TCP/IP

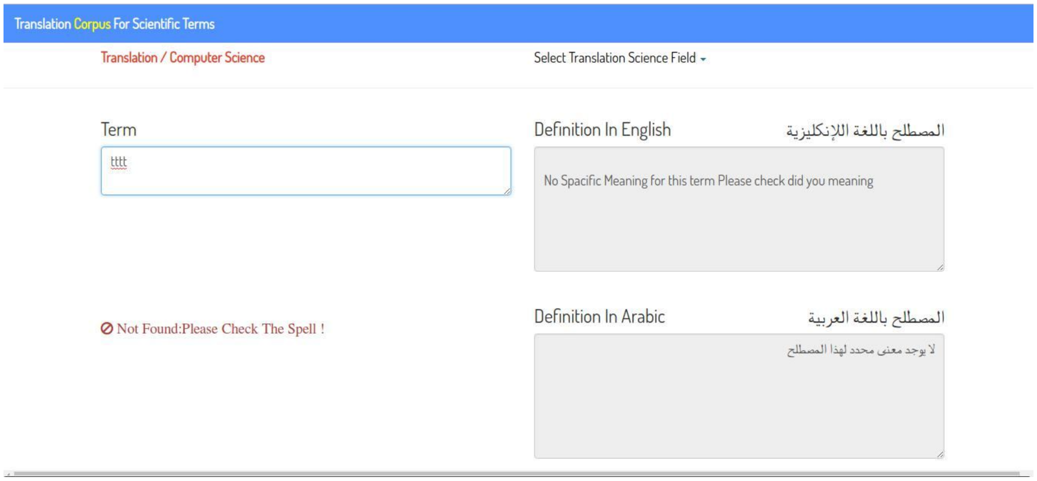
Definition In Arabic — المصطلح باللغة العربية

الإرسال التحكم بروتوكول /بروتوكول إنترنت. مجموعة من الشبكات القواعد :TCP/IP
يسمح لنا بالاتصال بأجهزة الكمبيوتر البعيدة عبر TCP/IP .المستخدمة للاتصال بالإنترنت
.الإنترنت وتبادل البيانات والمعلومات

Figure 4.8: Corpus Application main page

Is clearly show the above figure is illustrate the main page for the corpus application, in this sample the user here check about computer term is TCP/IP term the result of this scenario is all the function of application is working probability, as seen the prediction about what the user want to look for is working is shown is the left bottom of page, and the definition of the term is retrieved in both language (Arabic, English).

In the case if the user enter term that not recognize or there is syntax error in spell of this term, the figure 4.8 will appear to the user with the following description and instructions.



Translation Corpus For Scientific Terms
Translation / Computer Science
Select Translation Science Field

Term

tttt

Definition In English المصطلح باللغة اللإنكليزية

No Spacific Meaning for this term Please check did you meaning

⊘ Not Found:Please Check The Spell !

Definition In Arabic المصطلح باللغة العربية

لا يوجد معنى محدد لهذا المصطلح

Figure 4.9: Term not found syntax error

4.3 Chart (*Findlay user interface*)

This chart presents a group of users who have been presented with the application. It has been used for a period of not less than an hour and has been tested in a large number of terms. This work was done for five users of the preliminary studies students with four questions as follows:

1 - What is the portion of satisfaction on system?

2 - Do you increase your knowledge when using system?

3 - Your proposal to improve system?

question	User1	User2	User3	User4	User5
1	65%	80%	50%	90%	40%
2	Somewhat	yes	Little	yes	no
3	Extended the database	Edit by Experts in both fields	Use methods to improve resource retrieval	Only use more information in database	Use a good way to retrieve and accurately identify sources

4.4 Result and Discussion

There are a lot of electronic dictionaries for terms, putting in the mind that, this work will be helpful to the person who wants to learn and search about the scientific terms, this method in this thesis is achieved by using web application tools and some data mining based of web tool such as Ajax, crawling and JavaScript method, the electronic dictionary database.

The interaction steps is started with user when open the browser which contain input field on for insert the website URL and the with filter and crawl function the content the site will passed to the search engine process which collect the terms tags and then this term will be filtered to remove the tags and store this data and terms in the excel file, this file contains all data wanted data and not wanted so is need to be next steps which is refinement. The refine step is will be to remove any undesirable data and content, after this step as the result it get the data which is terms and the meaning of terms in excel sheet than can be easily imported to database, that can be used later in the query of search for the terms.

For the application provide web page application provided to get quires about terms and search and the meaning, this page contain one input filed that used to insert the desired term, this input field is directly connect with AJAX application for search engine this method allow to application to provide prediction about the term that the user looking for according to the letter that the user going to insert them specifically is AJAX look according to the first letters matched that user insert them and then provide all probability for this match, the method can help the user to recognize the terms.

The application Composed of the following items.

Crawling	Crawling data mine method to collect data for corpus
Given term	The given terms means the inputting English word.
Meaning in Arabic	Transliteration of the English terms in Arabic.

Prediction method	Provide prediction about the term while writing.
Export excel sheet	Extract and download excel sheet for variety website.

The Electronic corpus also supported auto search facility which means if type “add” the dictionary will search for words starting with add.....such as address, add-on ...so on.

4.4.1 Technical considerations

The application was designed considering the computational efficiency of the system execution and to be scalable, i.e., enabling the possibility of increasing the number of texts that conforms the corpus without producing a degradation of the performance, the web application has designed due to the needs of the target users and to allow the easy tool packing and redistribution. For its development used JavaScript, PHP and AJAX as a programming language since, this way, the obtained a platform independent software

4.4.2 System features

As a result of the corpus requirements, the initial version of the system allows:

Basic term search: i.e., looking for a word across the collection. This can be applied to the whole set of indexed database tables. As the result of a query all the occurrences of a word are showed.

Advanced search: this method was achieved with many queries and mining tool such as AJAX and JavaScript, whereby provide prediction about the term while writing.

Terms list generation: By aiding of JQuery coding that allow to crawling for terms form different website and generate list of terms that can be exported to database.

Report generation: The system allows to export the search results (crawling) to a excel sheet that editable by users.

The tool is being developed according to the corpus aims. Simultaneous development ensures text changes may be added to improve both. The fact that it is scalable provides the opportunity to enlarge the corpus.

4.5 Summary

This chapter presented the result of the implementation for carry out this work, also covered and explains the how to deal with application, at final present the complete shape and design for corpus and how to use this application.

Chapter Five

CONCLUSION & RECOMMENDATION

Chapter Five

Conclusion & Recommendation

5.1 Conclusion

Using technology in learning has become a real necessity nowadays. This paper has reviewed briefly how technology can be utilized in developing corpus the terms for the science learner, different methods for using technology in improving were discussed thoroughly.

The aim of this thesis was to create a glossary of terms used in many scientific fields, the first section of implementation is discuss about how to collect the data (terms) for the application this method achieved by create a web application page that use crawling engine, the engine that was design is use the URL web site to collect the terms form specific web site.

The second part of this study is the corpus application this application was design by using PHP programming language, PhpMyAdmin database, AJAX and other programming language mentioned in chapter three.

In conclude the work present electronic corpus in addition to simple search and browsing, this application can be used on different platforms, including mobile devices and the Web.

Currently, the application contain two scientific field with all terms dictionary, However, more fields can be easily incorporated and added, also with respect to functionality, many major extensions are added such as AJAX search engine and prediction method, finally with user-friendly interface as is consider in this application this feature help the user to cooperate and use the application very easily method.

5.2 Recommendation and tasks for the future

Discovering teaching and learning tools that save time and can contribute to learner achievement can help motivate teachers to learn more about effective uses of technology.

As a result, the following concluding remarks and recommendations can be recorded:

For functionality, two major extensions are planned. Firstly, we plan to support specialists and terms learners with extended content corpora, which contain many other scientific fields. Secondly, closer integration with machine translation is planned, thus allowing users to translate a full document instead of a phrase, sentence, or small fragment of text.

Reference

[1] Nomass, B. B. (2013). The impact of using technology in teaching English as a second language. English language and literature studies, 3(1), 111.

[2] Kilgarriff, A. (2005). Putting the corpus into the dictionary. In Proceedings MEANING Workshop..

[3] Wynne, M. (Ed.). (2005). Developing linguistic corpora: a guide to good practice (Vol. 92). Oxford: Oxbow Books.

[4] MacDonald, L. (2014). Phraseological Units in a Comparable Corpus of English and Machine Translations of Academic Writing.

[5] Šulc, Michal. Korpusová lingvistika. Praha: Karolinum, 1999. (all citations from this book are this thesis author's translations from Czech.)

[6] Bowker, L., & Pearson, J. (2002). Working with specialized language: a practical guide to using corpora. Routledge.

[7] Pearson, J. (1998). Terms in context (Vol. 1). John Benjamins Publishing.

[8] Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. Literary and linguistic computing, 7(1), 1-16.

[9] Quirion, J. (2003). Bowker, Lynne and Jennifer Pearson. Working with Specialized Language: A Practical Guide to Using Corpora. TERMINOLOGY-AMSTERDAM-, 9(2), 299-302.

[10] Meitei, S. P., & Devi, H. M. (2017). Development Of English To Manipuri Electronic Dictionary: A database approach. International Journal of Innovations & Advancement in Computer Science (IJIACS), 6(3).

[11] MacDonald, L. (2014). Phraseological Units in a Comparable Corpus of English and Machine Translations of Academic Writing.

[12] Zheng, H., & Wang, X. (2016). The use of electronic dictionaries in EFL classroom. Studies in English language teaching, 4(1), 144-456.

[13] Metruk, R. (2016). Determining the priority in vocabulary when learning English through electronic dictionaries. Xlinguae–European Scientific Language Journal, 9(4), 2-8.

[14] Jian, H. L., Sandnes, F. E., Law, K. M., Huang, Y. P., & Huang, Y. M. (2009). The role of electronic pocket dictionaries as an English learning tool among Chinese students. Journal of Computer Assisted Learning, 25(6), 503-514.

[15] Deksne, D., Skadina, I., & Vasiljevs, A. (2013). The modern electronic dictionary that always provides an answer. In Electronic lexicography in the 21st century:

thinking outside the paper: proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia (pp. 421-434).

[16] Al-Sulaiti, L., & Atwell, E. S. (2005). Extending the corpus of contemporary Arabic. In Proceedings of the CL'2005 Corpus Linguistics Conference. UCREL, Lancaster University.

[17] Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. Literary and linguistic computing, 7(1), 1-16.

[18] Alexandre Patry and Philippe Langlais(2005). Corpus_based Terminology Extraction .

APPENDIX A

Ajax Folder:

The file contain the AJAX detection and search engine where responsible for take the term that inserted by the user and then by the java script engine and AJAX search engine will looking for the term inside the database, if the term not found it show you the prediction for this term like did you mean that.

There are two files which are

1- ajax_arabic_search.php

2- ajax_arabic_search.php

They are almost the same on search in Arabic column in database and another one is search in English columns.

The code:

```
<?php

$servername = "localhost";
$username = "root";
$password = "admin123";
$dbname = "corpus";
// Create connection
$conn = new mysqli($servername, $username, $password , $dbname);
mysqli_set_charset( $conn, 'utf8');
if(isset($_POST['name']))
```

```
{
$name=trim($_POST['name']);
$q= "SELECT * FROM computer WHERE LOWER (term) = LOWER
('$name')"; $result = $conn->query($q);

        if ($result ->num_rows > 0)
        {

while($query=mysqli_fetch_array($result,MYSQLI_ASSOC))
{
    echo $query["in_english"];
}

}

        else
        {
            echo "No Spacific Meaning for this term Please check did you meaning";
        }
        }

?>
```

For the did mean file:

As always These first line show that the connection off database which contain the username and password and the name of the database the will search inside it.

After the it create the variable the will define the connection, we set the character for database to the utf8 to be able to search on the Arabic term.

The code :

```
<?php

$servername = "localhost";
$username = "root";
$password = "admin123";
$dbname = "corpus";
// Create connection
$conn = new mysqli($servername, $username, $password , $dbname);
mysqli_set_charset( $conn, 'utf8');
if(isset($_POST['name']))
{
$name=trim($_POST['name']);
$q= "SELECT * FROM computer WHERE LOWER(term) like
LOWER ('%$name%')";
$result = $conn->query($q);

    if ($result ->num_rows > 0)
    {
      echo "<i class='fa fa-ban text-danger' style='font-size:19px'> Did You
Meaning: </i>";
```

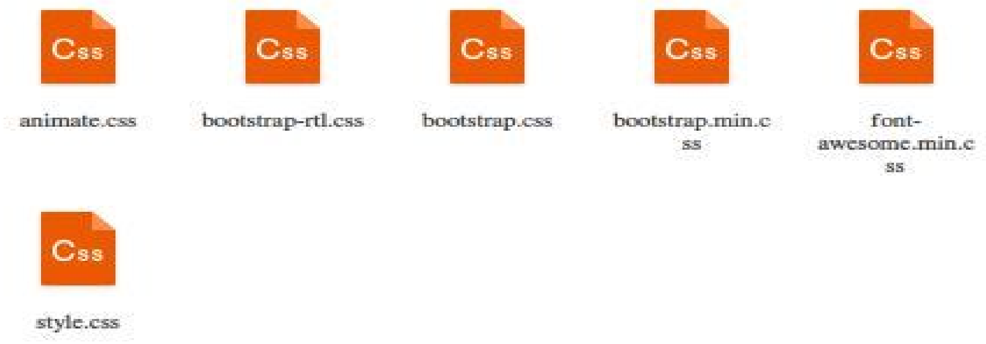
```
            while($query=mysqli_fetch_array($result,MYSQLI_ASSOC))
            {

               echo "<h4>- ".$query["term"];
            }

      }

      else
      {
         echo "<i class='fa fa-ban text-danger' style='font-size:19px'> Not Found: Please Check The Spell ! </i>";
      }
      }

?>
```

APPENDIX B

CSS code

CSS Folder

For the CSS Folder is contain many file as shown



All this file belong to the bootstrap design CSS and animation which can be need to design the application, something may need to consider the file of **style.css** this file code on it the customize style sheet the use to create the frontend of application.

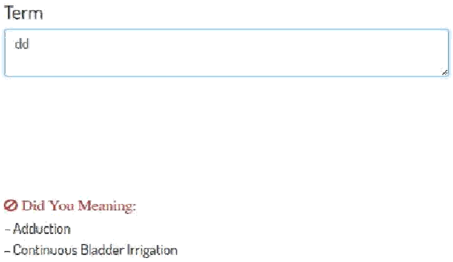
From the code above of **style.css**

- For the body style set the body font type as shown with name to snas-serif, which is located in the font Folder.

- Then for the heading type also set the same fonts

- After that set the navigation bar for any link inside it the color of text will be white
- Also have form in this form there is text area which set the text inside this area will be at all 50px for the size.

Finally for the did you mean field as shown in bottom figure set the margin for the top of the div will be 50px;



Term

dd

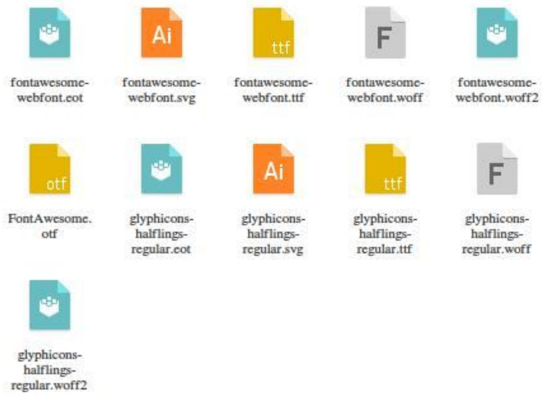
⊘ Did You Meaning:

- Adduction
- Continuous Bladder Irrigation

It is easy to understand the CSS file, can check the other files in Folder.

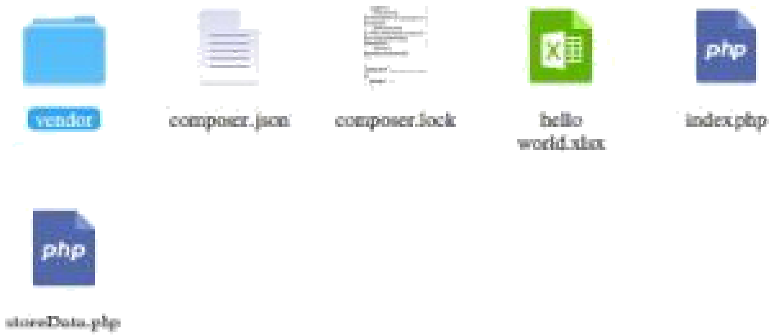
Fonts Folder

This folder contain all the font file that needed in the main page

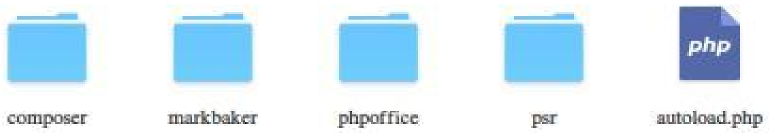

Ai
ttf
F
otf
fontawesome-webfont.eot
fontawesome-webfont.svg
fontawesome-webfont.ttf
fontawesome-webfont.woff
fontawesome-webfont.woff2
FontAwesome.otf
glyphicons-halflings-regular.eot
glyphicons-halflings-regular.svg
glyphicons-halflings-regular.ttf
glyphicons-halflings-regular.woff
glyphicons-halflings-regular.woff2

Data Mining Crawling Engine

The folder of the Crawling Engine called cruel and it contain many folders and file as shown below



The folder vendor is contain the main classes and the necessary needed files for the crawling engine, and is contain the following.



The folder composer is contain the composer engine for data which is user to create the excel file after the crawling engine.

And the folder php office the folder is more important file for create the excel file for Microsoft office, and they have auto load file which could be used to load the data on the file.

Code of crawling main page

```
<!DOCTYPE HTML>

<html>

<head>

    <meta http-equiv="content-type" content="text/html" />

    <meta name="author" content="GallerySoft.info" />

     <meta http-equiv="X-UA-Compatible" content="IE=edge">

     <meta name="viewport" content="width=device-width, initial-scale=1">
```

```
      <link href="css/bootstrap.css" rel="stylesheet"/>
      <!--<link href="css/bootstrap-rtl.css" rel="stylesheet"/>-->
      <link href="css/font-awesome.min.css" rel="stylesheet"    />

      <link href="css/style.css" rel="stylesheet"/>
      <link href="css/media.css" rel="stylesheet"/>
      <link href="css/hover.css" rel="stylesheet"/>
      <link href="css/animate.css" rel="stylesheet"/>
      <link href="https://fonts.googleapis.com/css?family=Dosis"
       rel="stylesheet"> <!--[if lt IE 9]-->
       <script src="js/html5shiv.min.js"></script>
       <script src="js/respond.min.js"></script>
     <!--[endif]-->
    <title>Translation Corpus</title>

    </head>

<body >

<!--- start of nava bar -->
     <nav class="navbar navbar-inverse navbar-fixed-top
         "> <div class="container-fluid">
         <!-- Brand and toggle get grouped for better mobile display -->


         <div class="navbar-header">
            <button type="button" class="navbar-toggle collapsed" data-
toggle="collapse" data-target="#ournavbar" aria-expanded="false">
              <span class="sr-only">Toggle
              navigation</span> <span class="icon-
              bar"></span> <span class="icon-bar"></span>
              <span class="icon-bar"></span>
            </button>
```

```
                <a class="navbar-brand" href="#" style="color:white"><b>Translation
                    <span    style="color:yellow">Corpus</span>    For    Scientific
Terms</b></a>
            </div>
        </div>
    </nav>
    <!--- end of nava bar -->

    <!-- Start of main page form -->

    <br>
    <br>
    <br>

    <!-- start of header -->

        <div class="row">
            <div class="col-sm-12">

                <div class="col-sm-5 col-sm-push-1">
                    <h4><span style="color:#dd4b39"> <b>Crawling Engine for
Corpus Translation</b> </span><h4>
                </div>

            </div>
        </div>
        <hr>
    <!-- end of header -->
```

```
  <!-- English term --->
    <div class="row">
                            <div class="col-sm-12">
                                <form role="form" action="crul/index.php"
method="POST">
                                    <div class="form-group col-sm-5
col-sm-push-1">
                                        <h3>Enter The Website</h3>
                                        <input type="text" class="form-control"
rows="1" placeholder="URL" name="url" required>
                                    </div>
                                    <div class="form-group col-sm-5
col-sm-push-1">
                                        <h3>Enter HTML TAG that Contain the
(Term) </h3>
                                        <input type="text" class="form-control"
rows="1" placeholder="TAG" name="tag" required>
                                    </div>
                    <div class="form-group col-sm-5 col-sm-push-
                      1"> <input type="submit" value="Crawle">
                    </div>
                                </form>
                            </div>
    </div>

  <!-- End of English term --->

<script src="js/jquery-3.2.1.min.js"></script>
<script src="js/bootstrap.min.js"></script>
<script src="js/plugins.js"></script>
<script src="js/wow.min.js"></script>
```

```
<script src="js/jquery.nicescroll.min.js"></script>

<script src="js/test.js"></script>

<script>new WOW().init();</script>

</body>

</html>
```

Index code of crawling engine:

```
<?php

require 'vendor/autoload.php';

use PhpOffice\PhpSpreadsheet\Spreadsheet;

use PhpOffice\PhpSpreadsheet\Writer\Xlsx;

$start= trim($_POST['url']);

$file_name = explode ("/",$start);

$file_name = $file_name [2].".xlsx";


$already_crawled= array();

$data_tostore = array();

function get_details($url)

{

        $tag = trim($_POST['tag']);

        global $row_counter ;

        global $sheet    ;

        $options = array('http'=>array('method' => "GET",
'headers'=>"User-Agent:howBot/0.1\n" ));

        $context = stream_context_create($options);

        $dom = new DOMDocument;

        @$dom->loadHTML(@file_get_contents($url, false,

        $context)); $data = array();

        foreach ($dom->getElementsByTagName("$tag") as $mean)

            {
                if ($mean)
                {
                        // array_push($data,$mean->nodeValue); // to get the content in
between of tags...
                        $sheet->setCellValue("A$row_counter", $mean-
                        >nodeValue); $row_counter++ ;
                }
        }
              // return array_filter($data, function($value) { return $value !== ''; });
              return $data;
}


function follow_links($url)
{
        global $already_crawled;
        global $data_tostore ;
        global $row_counter ;
        $row_counter = 0 ;

        $options = array('http'=>array('method' => "GET",
'headers'=>"User-Agent:howBot/0.1\n" ));
        $context = stream_context_create($options);
        $doc = new DOMDocument();
        @$doc->loadHTML(@file_get_contents($url, false,
        $context)); $linklist = $doc->getElementsByTagName("a");

        foreach ($linklist as $link)
        {
                $l = $link->getAttribute("href")."\n";
```

```
if (substr($l, 0 , 1) == "/" && substr($l, 0, 2) !== ("//"))
{

    $l = parse_url($url)["scheme"]."://".parse_url($url)["host"].$l;

}

else if (substr($l, 0, 2) == ("//"))
{

    $l = parse_url($url)['scheme'].$l;

}

else if (substr($l, 0, 2) == ("./"))
{

    $l = parse_url($url)['scheme']."://".parse_url($url)["host"].dirname(parse_url($url)["path"]).substr($l, 1);

}

else if (substr($l, 0, 1) == ("#"))
{

  $l = parse_url($url)['scheme']."://".parse_url($url)["host"].parse_url($url)["path"].$l;
```

```
        }

        else if (substr($l, 0, 3) == ("../"))
        {

            $l = parse_url($url)['scheme']."://".parse_url($url)["host"].$l;

        }

        else if (substr($l, 0, 5) != "https" && substr($l, 0, 4) != "http")
        {

            $l = parse_url($url)["scheme"]."://".parse_url($url)["host"]."/".$l;

        }

        if (!in_array($l, $already_crawled))
        {
            $already_crawled[] = $l;
            #echo $l;
            array_push($data_tostore,get_details($l));
        }

    }
$filter = array_filter($data_tostore, function($value) { return $value !== ''; });
// var_dump($filter);
}
$spreadsheet = new
Spreadsheet(); global $sheet ;
```

```
$sheet = $spreadsheet->getActiveSheet();
follow_links($start);
$writer = new Xlsx($spreadsheet);
$writer->save($file_name);
echo "
    <div class='download_box'>
       <h3>Your data is Ready</h3>
                  <a href='$file_name'>Download the Data</a>
    </div>
    ";
//print_r($already_crawled);
?>
<style>
    .download_box{
        text-align: center;
    }
    .download_box a{
        font-size: 30px;
    }
</style>
```

s