



Sudan University of Science and Technology
College of Computer Sciences and Information Technology

**Building the Multilingual Hadith Corpus to Enhance
Performance of Information Retrieval System for
Hadith**

**بناء مجاميع متعددة اللغات للحديث بغرض تحسين كفاءة نظام
استرجاع الاحاديث النبوية**

BY

Samah Mohamed Osman Hassan

A dissertation submitted in Partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

(Computer Science)

Supervisor

Dr.Eric Atwell

July 2017

Declaration

I hereby declare that: (1) The above thesis is my own unaided work, both in conception and execution, and that apart from the normal guidance of my supervisor, I have received no assistance apart from that stated below; (2) Except as stated below, neither the substance or any part of the thesis, has been submitted in the past, or is being, or is to be submitted for a degree in the University, or any other University.

I am now presenting the thesis for examination for the Degree of Ph.D. in Computer Science. I also grant the University free license to reproduce the above thesis in whole or in part, for the purpose of research.

Samah Mohamed Osman Hassan

Name

1-July-2017

Date

حديث شريف

قال رسول الله صلى الله عليه وسلم "إنما الأعمال بالنيات وإنما لكل امرئ ما نوى
فمن كانت هجرته إلى دنيا يصيبها أو إلى امرأة ينكحها فهجرته إلى ما هاجر إليه"

DEDICATION

To my dears

(Parents, brothers, sisters and teachers)

To those who help me in this research. With my appreciation and love.

Acknowledgment

I wish to express my deep gratitude and sincere thanks to my supervisor Dr. Eric Atwell for his kindness, valuable guidance, and constructive criticism .Also; I would like to express gratitude to all my teachers, colleague and everyone who contributed to this work.

I would like to acknowledge the hard work of the numerous volunteers (language learners, data collectors, evaluators, Islamic scholar) who contributed their time and effort to this project. I also owe special thanks to the colleagues who worked very hard on Hadith matching with the translation of the Multilingual Hadith Corpus.

Lastly, special Acknowledgments and sincere are extending to my husband and my families for their patience and support.

ABSTRACT

Information retrieval (IR) systems retrieve relevant information relating to a specific query by the user, and this requires the extraction of related unstructured information from data which may be texts, sound, images. In this context, an important problem facing information retrieval, in particular from text files, is reliance on exact matching of the word or words in the query and the same words in a specific text file. This leads in many cases to the loss of results where files contain synonyms with words in the query which may be useful to the user. This dilemma appears in most information retrieval systems for unstructured text data, and with most languages, especially with regard to the Arabic language. This research will deal with the problem of information retrieval from the Hadith across many languages, by building a parallel corpus with multiple languages containing the Hadith in Arabic as well as translated texts in English, French and Russian. We have built a parallel corpus containing the text of 2030 Arabic Hadith along with the translation in English, French, and Russian languages. Thus the parallel corpus contains 8120 Hadith consisting of 2,470,913 words. Our matching algorithm is applied into the data for the retrieval process, calculating the weight of the words in the query based on their importance and then comparing this with the existing documents, which have been processed to calculate the importance of words in each document. Then a similarity coefficient is calculated from the particular query and existing documents. To improve performance, the system has a dictionary of words with identifying all files that contain those words as an inverted index. We built a web portal to allow user search via World Wide Web.

We designed and evaluated the proposed solution by using a selection of important concepts, for which we have pre-determined the results manually without referring to the system. The evaluation work calculates both the average precision and average recall for each language.

The results showed that the proposed method has good results for retrieval in all four languages: the average precision and average recall of the Arabic language were 96.5% and 82%, consequently for the English language they were 98.4% and 90%, the French language were 97.5% and 91.7% and the Russian language they were 98% and 91%.

المستخلص

يعرف استرجاع المعلومات علي انة عملية استرجاع المعلومات ذات العلاقة بناء علي استعمال معين من قبل المستخدم المعين,وان ذلك يتطلب استخلاص هذه المعلومات ذات الصلة من بيانات غير مهيكلة قد تكون (نصوص , صوت , صور).وفي هذا السياق فان من أهم المشاكل التي تواجه استرجاع المعلومات وبصورة خاصة النصية منها ان الملفات ال مسترجعة تعتمد صيغة التطابق ال فعلي للكلمة او الكلمات الموجودة في الاستعلام والبحث عن نفس الكلمات في الملفات المعينة وذلك يؤدي في اغلب الح الات الي فقدان تلك الملفات التي تحتوي علي مصطلحات مرادفة لتلك الموجودة في الاستعلام المعين والتي قد تكون أكثر إفادة للمستخدم,ومن الملاحظ ان هذه المعضلة تظهر في اغلب انظمة استرجاع البيانات وباغلب اللغات وخاصة فيما يخص اللغة العربية .وفي هذا البحث سوف يتم التعامل مع هذه المشكلة فيما يخص نصوص الحديث النبوي الشريف عبر بناء مدونة لغوية متوازية متعددة اللغات تحوي الاحاديث النبوية باللغة العربية بالاضافة الي تراجم النصوص باللغات الانجليزية ,الفرنسية والروسي

لقد قمنا ببناء مدونة متوازية تحتوي علي 2030 نص من الاحاديث العربية متبوعه بتراجمها من اللغات الانجليزية والفرنسية والروسية .ونتيجة لذلك تكون مدونة الحديث المتوازية تحتوي علي مايقارب 8,120 حديث وعلي ما يقارب 2,470,913 كلمة.

قمنا باقتراح خوارزمية تماثل تطبق علي البيانات أثناء عملية الاسترجاع عبارة عن حساب وزن الكلمات علي حسب اهميتها بالنسبة للاستعلام المعين ومن ثم مقارنتها مع المستندات الموجودة والتي تم تخزينها ايضا علي حساب اهمية وزن الكلمات في كل مستند ومن ثم حساب معامل التشابة بين الاستعلام المعين والمستندات الموجودة واسترجاع الملفات ذات النسبة الأعلى وعرضها علي المستخدم وتكون مرتبة تنازليا من الأعلى نسبة الي الاقل ,ولتسريع زمن معالجة البيانات تم بناء قاموس يحتوي علي الكلمات مع تحديد كل الملفات التي تحتوي علي تلك الكلمات.بالاضافة إلي بناء موقع علي الانترنت يمكن المستخدمين من البحث .

تم تصميم وتقييم الحل المقترح باستخدام مجموعة مختارة من المفاهيم المهمة والتي تم تحديد نتائجها مسبقا بصورة يدوية من غير الرجوع للنظام ومن ثم تم عمل التقييم بحساب كلا من متوسط الدقة والاسترجاع .

النتائج أوضحت ان الطريقة المقترحة ادت الي تحسين في كفاءة اداء نظام الاسترجاع وذلك بنسبة كبيرة بالنسبة الي الاربع لغات المقترحة فان متوسط الدقة والاسترجاع للغة العربية كان علي التوالي 82 % and 96.5 (% وللغة الانجليزية كان (90% and 98.4 %) وللغة الفرنسية كان (91.7 % and 97.5 %) وللغة الروسية كان (91% and 98 %)

TABLE OF CONTENTS	PAGE
	No.
DECLARATION	I
DEDICATION	III
ACKNOWLEDGMENT	IV
ABSTRACT (ENGLISH)	V
المستخلص	VI
TABLE OF CONTENTS	VII
LIST OF TABLES	XI
LIST OF FIGURES	XIII
LIST OF SYMBOLS/ABBREVIATIONS	XV
LIST OF APPENDICES	XVI
CHAPTER ONE INTRODUCTION	1
1.1 Preface	1
1.2 Why Collect Multilingual Hadith	2
1.3 Information Retrieval	2
1.4 An Overview of the Hadith	3
1.5 Definition of Concept	3
1.6 Understanding the Concepts of the Hadith	5
1.7 Language Selection	5
1.8 Corpus Linguistics	5
1.9 Problem Statement	6
1.10 Research Significance	6
1.11 Research Objectives	7
1.12 Research Contributions	7
1.13 Research Question	8
1.14 Research Hypothesis	8
1.15 Research Methodology	8
1.15.1 Phase One	8
1.15.2 Phase two	9
1.15.3 Phase Three	9

1.15.4 Phase four	9
1.15.5 Phase five	9
1.16 Research Scope	9
1.17 Thesis Organization	10
CHAPTER TWO LITTERATEUR REVIEW	11
2.1 Introduction	11
2.2 Information Retrieval Strategies	11
2.2.1 Vector Space Model	11
2.2.2 TF-IDF measure	11
2.2.3 Cosine Similarity	12
2.2.4 Inverted Index	12
2.3 Determining the vocabulary of terms for text processing	13
2.3.1 Definition of Tokenization	13
2.3.2 Definition of Stemming	13
2.3.3 Remove stop words	13
2.3.4 Definition of Normalization	14
2.4 Parallel Corpora	14
2.5 Related Work	15
2.5.1 Parallel Corpora for Medium Density Languages	15
2.5.2 Assessment of a Significant Arabic Corpus	15
2.5.3 The arTenTen project	16
2.5.4 Open Source Arabic Corpora	16
2.5.5 Arabic Learner Corpus	16
2.5.6 Quranic Arabic Corpus	17
2.5.7 The KSUCCA Corpus	17
2.5.8 The ICA Corpus	18
2.5.9 Quran 'Search for a Concept' tool and website	18
2.5.10 The KACST Corpus	18
2.6 Text Analysis Tools	19
2.6.1 WordSmith Tool	19
2.6.2 aConCorde Tool	19
2.6.3 Sketch Engine	20
2.6.7 Xaira Tool	20

2.6.8 The XML technology	20
2.6.9 The Sklearn	21
2.7 Search Tools for Hadith	21
2.7.1 Search Truth Tool	21
2.7.2 AL Muhaddith Search Engine	22
2.7.3 Dourous search tool	22
2.7.4 Hadith Encyclopedia	25
2.8 Key ideas from the Hadith search tools	27
2.9 Summary	28
CHAPTER THREE METHODOLOGY	29
3.1 Introduction	29
3.2 The Methodology	30
3.2.1 Phase one: Design Requirements for MHC	30
3.2.1.1 Important of Hadith	30
3.2.1.2 Survey result	31
3.2.2 Phase two: Collect the Data	43
3.2.2 Data Collection	43
3.2.3 Data Cleaning	44
3.2.4 Data Preprocessing	45
3.2.5 File Generation	45
3.2.6 Data Annotation	47
3.2.3 <i>Phase Three :Vector space model</i>	49
3.2.3.1 English Corpus	54
3.2.3.2 Arabic Corpus	56
3.2.3.3 French Corpus	57
3.2.3.4 Russian Corpus	58
3.2.4 Phase Four: MHC in Sketch Engine	59
3.2.4.1 Steps for building parallel corpora	61
3.2.5 Phase Five: Web application Design	64
3.2.5.1 Software Requirements	64
3.2.5.2 Search Algorithm	65
3.2.5.3 Search by Arabic Concept	66
3.2.5.4 Search by English Concept	67

3.2.5.5 Search by Russian Concept	68
3.2.5.6 Search by French keyword	69
3.2.5.7 Download	70
3.2.5.8 Contact Us	71
3.3 Summary	71
CHAPTER FOUR RESULTS AND DISCUSSION	72
4.1 Introduction	72
4.2 Arabic Stemming	72
4.3 English Stemming	74
4.4 French Stemming	74
4.5 Russian Stemming	74
4.6 Information retrieval system evaluation	77
4.7 Quantitative Evaluation	78
4.8 Qualitative Evaluation	91
4.9 Issue to discuss	93
4.10 Summary	93
CHAPTER FIVE CONCLUSION AND FUTURE WORKS	94
5.1 Introduction	94
5.2 Conclusion	95
5.3 Future Works	95
5.4 Main Contributions	95
Bibliography	96
Reference	97
Appendices	103
List of Publications	122

LIST OF TABLES

Table No.	TITLE	Page No.
Table 2.1:	The advantages and disadvantages for the tools	22
Table 2.2:	The finding and limitation for the related works	23
Table 2.3:	The features in Hadith search tools	27
Table 3.1:	The XML tags for Multilingual Hadith Corpus	48
Table 3.2:	Column name in the MS-Excel file for the MHC	48
Table 3.3:	The number of Ahadith by words in the Corpus	49
Table 3.4:	The Inverted Indexes for the English corpus	53
Table 3.5:	The Inverted Indexes for the Arabic corpus	53
Table 3.6:	The Inverted Indexes for the French corpus	54
Table 3.7:	The Inverted Indexes for the Russian corpus	54
Table 4.1:	Number of the terms for Arabic text before and after stemming process	73
Table 4.2:	Number of the terms for English text before and after stemming process	74
Table 4.3:	Number of the terms for French text before and after stemming process	75
Table 4.4:	Number of the terms for Russian text before and after stemming process	76
Table 4.5:	The gold standard for our Search	78
Table 4.6:	The correct relevant documents for the selected English gold standard	80
Table 4.7:	The calculation of precision & recall for English gold standard	81
Table 4.8:	The precision & Recall and F-measure for English gold standard	81
Table 4.9:	The correct relevant documents for Arabic gold standard	82
Table 4.10:	The calculation of precision & recall for Arabic	83
Table 4.11:	Result of precision & Recall for Arabic gold standard	84
Table 4.12:	The correct relevant documents for the gold standard of French	85
Table 4.13:	The calculation of precision & recall for French	86
Table 4.14:	The precision & Recall & F-Measure for French gold standard	86
Table 4.15:	The correct relevant documents for the gold standard of Russian	87
Table 4.16:	The calculation of precision & recall for Russian	88
Table 4.17:	The precision & Recall for Russian gold standard	88

Table 4.18: The differences in Recall, Precision, and F-Measure between the languages	89
Table 4.19: The evaluation criteria for computational Hadith search tools	91
Table 4.20: The evaluated Hadith search tools with the HCSE	92

LIST OF FIGURES

Figure No.	Page No.
Figure 2.1: A snapshot of the Search Truth tool	24
Figure 2.2: A snapshot of AL Muhaddith Searches Engine	24
Figure 2.3: A snapshot of the Dourous Tool	25
Figure 2.4: A snapshot for the Error result in Dourous Tool	26
Figure 2.5: A snapshot of Hadith Encyclopedia in Russian	26
Figure 3.1: The MHC Muslims & non-Muslims Participants	32
Figure 3.2: The MHC participation of both gender	32
Figure 3.3: The MHC participants of different age	34
Figure 3.4: The MHC participants of different occupations	34
Figure 3.5: The MHC religious and educational motives	36
Figure 3.6: The MHC Hadith explanation in Arabic	36
Figure 3.7: The MHC respond to having the Meaning of words in Arabic	37
Figure 3.8: The MHC moral gained from Hadith	38
Figure 3.9: The MHC one website feature	39
Figure 3.10: The MHC Hadith in different languages	40
Figure 3.11: The MHC Hadith Sources	42
Figure 3.12: The MHC Hadith Classification	42
Figure 3.13: The MHC Searching Facilities	43
Figure 3.14: Snapshot of the XLTools	44
Figure 3.15: Plain text, for example, Arabic Hadith	45
Figure 3.16: Plain text, for example, English Hadith	46
Figure 3.17: Plain text, for example, French Hadith	46
Figure 3.18: Plain text, for example, Russian Hadith	47
Figure 3.19: The angle(θ) between two vector documents	50
Figure 3.20 : The similarity between Q and d2 in the vector space	51
Figure 3.21 : The English documents in the csv file	55
Figure 3.22 : A proposed feeding algorithm	56
Figure 3.23 : The Arabic documents in the csv file	57
Figure 3.24 : The French documents in the csv file	58

Figure 3.25: The Russian documents in the csv file	59
Figure 3.27: Snapshot of parallel Corpora in the Sketch Engine	60
Figure 3.28: Example ALIGNSTRUCT for the HadithArabic Corpus	62
Figure 3.29: Example ALIGNSTRUCT for the HadithEnglish Corpus	62
Figure 3.30: Example of ALIGNSTRUCT for the HadithFrench Corpus	63
Figure 3.31: Example of ALIGNSTRUCT for the HadithRussian Corpus	63
Figure 3.32: The search Tool interface	65
Figure 3.33 : A propsed search Algorithm	66
Figure 3.34:Snapshot for Arabic search concept “الايمان”	67
Figure 3.35: Snapshot for English search concept “knowledge”	68
Figure 3.36 : Snapshot for Russian search concept “омовения”	69
Figure 3.37: Snapshot for French search concept “foi”	70
Figure 3.38: The different options of download the MHC	70
Figure 3.39: Snapshot of the contact form	71
Figure 4.1: The change of number of terms for Arabic before and after stemming process	73
Figure 4.2: The change of number of terms for English before and after stemming process	75
Figure 4.3: The change of number of terms for French before and after stemming process	76
Figure 4.4:The change of number of terms for Russian before and after stemming process	77
Figure 4.5:The improvement of precision & recall and F-measure for English gold standard	82
Figure 4.6: The improvement of precision & recall & F-Measure for Arabic gold standard	84
Figure 4.7: The improvement of precision & recall and F-Measure for French gold standard	87
Figure 4.8: The improvement of precision & recall and F-Measure for Russian gold standard	89
Figure 4.9: The differences in Recall, Precision and F-Measure between the languages	90

LIST OF SYMBOLS/ABBREVIATIONS

CS: Cosine Similarity

HCSE: Hadith Corpus Search Engine

HTML : Hyper Text Markup Language

IDF :Inverse Document Frequency

ICA: International Corpus for Arabic

IDE : Integrated Development Environment

IR :Information Retrieval

KACST : King Abdulaziz City for Science and Technology

MHC: Multilingual Hadith Corpus

MVC : Model View Controller

NLTK: Natural Language Tool kit

PDF : Portable Document Format

POS :Part of Speech

TXT : Plain text format

TF: Term Frequency

WWW: World Wide Web

XML : eXtensible Markup Language

LIST OF APPENDICES

Page NO

Appendix A :MHC survey.....	109
Appendix B: Examples of MHC files format.....	112
Appendix C: The programming code	119
Appendix D: Hadith source books.....	127

CHAPTER I

INTRODUCTION

1.1 Preface

The aim of this research is to design Multilingual Hadith Corpus(MHC) to improve the retrieval accuracy of hadith text in different languages. Although the Hadith is one of the important religious books of the world, there is little computational analysis performed on it. This is down to many reasons; the absence of adequate morphological analyzers for Classical Arabic (the language of the Hadith), the nature of the text itself as it is not an ordinary text that we can put to standard machine processing but rather a text that is compiled in some special way in terms of linguistic structures that can reveal different meanings across the ages. On the other hand Corpus is an increasingly important area in applied linguistics . Hadith for Muslims comes in the second level after Quran. Hadith is considered the second source of religion practices and legislations for Muslims. Hadith explain the full life of Muslim mention in Prophet Mohammed (Peace be upon him) words which are called (Sunnah).So all Muslims irrespective of their age, occupation, nationality, their mother tongue language and place of residence must learn about Hadith, in addition knowing the meaning of each word as well as the moral of these Hadith.

1.2 Why collect multilingual Hadith

People have a drawer to deal with the Hadith directly from books and this applies to its own language Arabic or uses for translations in other languages such as English, French, and Russian. On the other hand, some work had been done by groups or by individuals in the work of the electronic scanning of many modern books have been incorporated computerized and has been available on the internet in the form of (PDF) and can be downloaded and viewed, but these books are found in several different locations, and to my knowledge until today no one worked to collect this AHadith in different languages in one place. Besides , the purpose of the collection of a large number of Hadith, including this word meanings for each word and a full explanation of the Hadith translation of talk in several languages that will allow Muslims easily accessible, used to identify the Hadith and the meanings of the words that come from the Hadith, as well as being able to find a translation in English, French and Russian. In addition, there are many words and meanings, interpretations and different associated semantics to each other making it easier for linguistics and researchers in the definition of vocabulary, the coherence of data, data analysis, concepts and phrases contained in the Hadith .The target users to use this source is the Muslims who wish to learn or studying Hadith as education. Also, the hadith corpus can be used by non-Muslims who would like to get to know about the great of Islam and its values, researchers and linguistics.

1.3 Information Retrieval

In recent years, there has been an increasing interest in Information Retrieval system and finding different ways to enhance the performance and accuracy of the

data which had been sending back to the users depends on their query (Shanahan et al, 2006). There are many different definitions for the IR, one defined by (Salton, 1968):

“Information retrieval is a field concerned with the structure, analysis, Organization, storage, searching, and retrieval of information”

Information Retrieval (IR) is devoted to finding "relevant" documents; not finding simple matches to patterns so to do that search query consisting of a keyword express user's information need the main interface of the IR system provides the user with an input field for the query. Then, all matching documents that have the query's term are found and displayed back to the user. In our approach, we focus on how rank the relevant document using will merge of two algorithms the first one is TF-IDF term frequency inverse document frequency and cosine similarity (Jbara, 2010). Therefore to achieve our goal some steps has been followed as we see below. Our method could be described as an IR system that manipulates the query in a manner that guarantees a better performance.

1.4 An Overview of the Hadith

There is persuasive evidence that the hadith plays a crucial role in regulating Muslims life (Al Imam, 2000). There has been a lot of work done in the creation of Arabic corpora, with many of them focusing on the QURAN. This is very beneficial, with many of the quranic corpora being excellent, but mentions within these corpora of the hadith, the words of prophet Mohammed (Peace be upon him), was very rare. The hadith for muslims is second in importance only to the Quran. In the Islamic Rules (Shreeaa Alislamia), the Hadith is considered the second source of religious knowledge for Muslims , as in the Hadith you will find teachings on all areas of the life of Muslims mentioned in the prophet Mohammed (Peace be upon him) words. The Hadith guides Muslims in how to be good Muslim the prophet Mohammed (Peace be upon him) explain everything necessary for Muslims to live their life: how to eat, How to drink ,How to sleep, how deal with other people ,how to pray, how to obey Allah and how to do everything else be it minor or major.

Therefore, a multilingual Hadith corpus would be useful for Muslims all around the world as it will allow them to know what each word is, what it means, and what it teach us about our religion.

Hadith is Arabic single word (plural is Ahadith) are collections of the reports claiming to word what the prophet Mohammed(peace be upon him)said (Hadith,2016), the original Hadith language is Arabic and we decided to select three different languages. As a result plus Arabic, we take Hadith in English, French, and Russian

1.5 Definition of Concept

Concept search is an automated search concepts information used for full-text search electronic information relevant provisions of concepts on how research organized and stored information. In other words, ideas retrieving information in response to a search query terms to relevant ideas in the query text (Giunchiglia,2009).

WordNet dictionary is derived concept as defined in the abstract or general idea or special occasions. Knowledge is not an easy task set for each area. On the other hand to clarify the ambiguity of taking (Bennett,2005) have two different definitions for ‘concept’:

“understanding of the word ‘concept’. Those who are doubtful about the idea of precision and universality tend to regard a ‘concept’ as something rather close to a natural language term, as something which may be vague or ambiguous and may be disputed over by people with different world views”

“ a concept is an abstract entity that is largely independent of the vagaries of natural language: only in special circumstances can a concept within a formal system be regarded as the referent of a natural term ”

For our work will consider the second definition of ‘concept’ as a term so regarding our work in Hadith consider each term as ‘concept’ notice that Hadith is very sensitive text and every word is important.

1.6 Understanding the Concepts of the Hadith

The importance of the Hadith for Muslims comes in the second level after Quran. Hadith is considered the second source of religion practices and legislations for Muslims. Hadith show Muslims ways of doing everything in life (AL Imam Majd, 2000). So Muslims must understand the concepts and meanings of the talk is a matter of great importance (AL Imam Majd, 2000) which require us to permanent work and continuing to seek to clarify and facilitate dealing with everything related to Hadith and make it available on the electronic network in the best pictures, easy to access and May God accept our work.

1.7 Language selection

We had decided to select the languages of our corpus as the Arabic is the original language of Hadith text. On the other hand the united nations official languages, with many Muslim language-speakers. the U.N. has only six official languages (Official Languages, 2016): Arabic, Chinese, English, French, Russian and Spanish; but there are relatively few Muslims speaking Chinese or Spanish. In addition, We needed volunteer informants in the languages chosen; researcher can cover Arabic and English, and we had volunteer assistants to advise on French and Russian.

1.8 Corpus Linguistics

Corpus linguistics can be defined as the study of language through the use of large collections of machine-readable texts, referred to by the term 'corpora'. Corpus linguistics is not a branch of linguistics, but rather a methodology that can be used to study all the different aspects of language, such as syntax, semantics, pragmatics, and speech. The basic corpus methodology was well known in linguistics for a long time, but what is different now is the increasingly large scale of the using of corpora in linguistics studies. This is due to the recent increasingly large advancements in technology, especially the massive production of computers and software that has occurred of late. The combination of corpora and computers as a means of studying

languages massively changed the way we analyze linguistics phenomena. Linguists are forever curious about different language structures and their functions. In the past, many theories and interpretations have been proposed to explain linguistics phenomena, but the scale of the data at hand was too small to prove or show much when considering the infinity of language. Thus, although results of such traditional studies were accurate, obtaining the results themselves was not very easy. In addition, past studies were more focused on investigating language structure rather than on language use itself (Al Sulaiti, Atwell, 2006).

1.9 Problem Statement

One of the informational retrieval system problem that in many cases users want the best answer to an information need among many documents that contain certain words. In our research we had been worked to had more improvement in the documents relevant retrieval for the specific users search for Hadith text in web search. The problem for this research are state as following:

[1] Lack of accurate information retrieval (IR) system in Hadith text for Arabic and other languages. Beside difficulties to finding "relevant" documents depend on the user query with more than one word in Arabic like “انما الاعمال بالنيات” and same problem with the other languages.

[2] Lack of Hadith resources on the world wide web in different file format like (.xml,html etc) that will allow different users can use Hadith in a different type of corpora tools analysis.

1.10 Research Significance

The purpose of this research is to enhance the performance of search for Ahadith (plural of “Hadith”) by creating Hadith Corpus Search Engine(HCSE) which will allow users to search and find Ahadith in the Arabic text plus their translation in languages. On the other hand to achieve our goal we decided to collect our dataset by building the multilingual Hadith Corpus(MHC) which will focus on

the Hadith text only .we want to investigate around 8,000 Hadith along with their translation in different languages, the weight for each and how important that word to specific text and how that can affect the search performance and how the result are achieved users satisfaction. Besides Ahadith included in our research are from ALBIKHARI book. Building this search engine will allow Muslims who access it to find Ahadith as well as being able to find translations for each text in English, French, and Russian from ALBIKHARI. The Hadith Corpus Search Engine (HCSE) could be used by Muslims who want to learn or teach the Hadith and other people who want to know about Islam.

1.11 Research Objectives

In order to achieve the our goals.The researcher defined a number of objectives as following

1. To develop the Multilingual Hadith Corpus based on the design criteria.
2. To developed Ahadith search engine in different languages.

1.12 Research Contributions

The following are the list of contributions of this research:

1. Developing a Multilingual Hadith Corpus based on the survey design to focuses on the nine aspects of user requirements.
- 2.To enhance the search for Hadith in the information retrieval system .
- 3.Develop website to hold the entire MHC as open source ,access file in three different formats (.txt , .html ,xml).

1.13 Research Question

In this research, we try to answer the following question

- 1- How the Hadith text must be represented to fit our model?
- 2- How the corpus compiling in linguistic Tools?
- 3-Can the proposed solution enhances the performance of Information retrieval system for Ahadith?

1.14 Research Hypothesis

For Muslims, the second most important source of knowledge for the Islamic Rules (Shreeaa Al-Islamia) is the Hadith. Consequently, finding Hadith explanations along with word meanings in Arabic for those who can read the Arabic language, and the translations in different languages for none-Arabic readers, will be an important contribution in the field of building Islam-related corpora.

1.15 Research Methodology

The system is implemented in five phases as follows:

1.15.1 Phase One

This phase was to include data collecting, data preprocessing and file generation for the entire corpus. On the other hand selecting and organizing the texts, which will include written texts in multilingual Arabic and for each text the translation in English, French, and Russian. We had been collected around 5,00,000 word. This process of selecting the texts had been extended over the three years of the project.

1.15.2 Phase two

Make the MHC working with the most known linguistics analysis tool called sketch engine creating the parallel corpora using the structure is defined as ALIGNSTRUCT in the sketch engine website it is free.

1.15.3 Phase Three

In this phase we plan to take the entire corpus feed into the TF-IDF algorithm and convert each document into vector that allows us to use the mathematical representation of the TF-IDF and after that generated the inverted index for all the document to improve the search scanning, Finally calculated the coefficient similarity between the user query and the documents find the best relevant documents for the user.

1.15.4 Phase four

In phase four the researcher had explained how we had applied vector space model into our corpus, calculating the TF-IDF weights and finally find the cosine similarity between the relevant documents

1.15.5 Phase five

Creating the website, which it had contained the Multilingual Hadith Corpus. Besides, the website will be held the search engine which allows users search in the entire corpus by the four different languages.

1.16 Research Scope

Provide search engine works with higher accuracy for the precision and recalls under specific consideration in the four languages Arabic, English, French, and Russian.

Ability to find Arabic Hadith along with the translation for each Hadith in three languages: English, French, and Russian.

1.17 Thesis Organization

This thesis is organized into five chapters as follows:

Chapter two: Provides some of the issues discussed and the importance of body design corpus. In addition, this chapter deals with some of the analysis tools used with the corpus. In addition, this chapter deals with the design and construction-related studies corpus. Finally we were discussed some of the challenges and outstanding issues.

Chapter three : Provides the methodology which has been following to accomplish our goal. Our method ordered in four phases which will be described in details in chapter three.

Chapter four: Provides the result and discussion of our algorithm implementation. By calculating the precision and recall for the entire corpus in the four different languages showing the improvement result.

Finally chapter five: Provide the close conclusion of our works and the recommendations of the future works and shortly described of our contributions.

CHAPTER II

LITTERATEUR REVIEW

2.1 Introduction

In this chapter, we had reviewed previous work of the designed corpora monolingual and multilingual corpora. Besides, the researcher was investigating if these corpora have any part of Hadith. On the other hand, we had discussed some analysis corpus tools and compare them. Finally, we had reviewed some search tools for Hadith had been found on the internet for the purpose of finding a way to access Ahadith, not for the research purpose.

2.2 Information Retrieval strategies

2.2.1 Vector Space Model

The vector space model computes a measure of similarity by defining a vector that represents each document, and a vector that represents the query. The model is based on the idea that, in some rough sense, the meaning of a document is delivered by the words used. If one can represent the words in the document by a vector, it is possible to compare documents with queries to determine how similar their content is. If a query is considered to be like a document, a cosine similarity (section 2.11.3) that measures the similarity between a document and a query can be computed. Documents whose content, as measured by the terms in the document, correspond most closely to the content of the query are judged to be the most relevant (Grossman et al,2012).

2.2.2 TF-IDF Measure

In the IR systems the TF-IDF stand for Term Frequency Inverse Document

Frequency used as statistical measure to identify the importance of specific word for document located within a group of texts or corpus, and the increase in the importance of the word commensurate with the number of times they appear in the designated document, but balanced with the number of times they arise in total documents or in the entire corpus (Ramos, 2003). The TF is the number of time the words appear in the document and the IDF indicate the number of time the word appears in the entire corpus (Ahmed et al, 2015). So that if the term appears rare in the document that means the term are important and give the term high score but if the term appears many times in the documents or in all the documents that mean the term not important to the document and give the term low score. Consequently, the common terms like “the”, “and”..etc they appear all the time in all the documents means not only they are not important but also do not identify the document.

2.2.3 Cosine Similarity

The most successful of these is the cosine correlation similarity measure. The cosine correlation measures the cosine of the angle between the query and the document vectors. When the vectors are normalized so that all documents and queries are represented by vectors of equal length, the cosine of the angle between two identical vectors will be 1 (the angle is zero), and for two vectors that do not share any non-zero terms, the cosine will be 0 (Croft et al, 2010). The cosine measure is defined as:

“There is no theoretical reason why the cosine correlation should be preferred to other similarity measures, but it does perform somewhat better in evaluations of search quality” (Croft et al, 2010).

2.2.4 Inverted Index

In Information Retrieval system an inverted index is a terminology used to index data structure storing a mapping from content, such as terms to its locations in a document or a set of documents and in our case Inverted index is a mapping process map each term into the number that number represents the document id that contains the specific term. On the other hand, we used inverted indexes are considered the most efficient and flexible index structure (Büttcher et al, 2016).

2.3 Determining the vocabulary of terms for text processing

There are different ways to process the text we mentioned some of them as follows:

2.3.1 Definition of Tokenization

This process is to convert the whole text or document into small sections or parts, each of which is called a token. This division depends on the white spaces and separations in a basic way and also throwing away certain characters, such as punctuation, and urges in the following languages: Arabic, English, French, and Russian. These tokens are often referred to as terms or words, and there is one definition for the token by (Manning et al, 2008):

“A token is an instance of a sequence of characters in some particular document that is grouped together as a useful semantic unit for processing”.

2.3.2 Definition of Stemming

Stemming is a process of eliminating affixes from a word, ending up with the stem. For example, the stem of cooking is a cook, and a good stemming algorithm knows that the ing suffix can be removed. Parts of the most commonly used word search engine for the indexing term. Instead of storing all forms of the word, search engines can store the stems, significantly reducing the size of the index, which improves the accuracy of the recovery (Perkins, 2014). The NLTK have different type of stemmer classes like *PorterStemmer* (Porter,1980), *LancasterStemmer* and *SnowballStemmer*.

2.3.3 Remove stop words

There are common words which whose effect is small value in the selection of the specific document or whether to decide if the document is relevant or not in the information retrieval system. These words are called stop words (Manning et al, 2008). Besides the strategy of eliminating these words, especially when we building of the search engine improve the speeds up the process was suggested by (Alsaleem,2011) & (Al-Shalabi et al,2006). So that for each language there is a list of

a word called stop words list and the NLTK comes with built in stop words list for each language in our case we interested in Arabic, English, French, and Russian.

2.3.4 Definition of Normalization

Text normalization is the process of converting text into a canonical form, he would not have had. Normalization of text to store gold processing allows separation of the problems. After our documents are divided into tokens, the simplest thing is also signs in the request match tokens in the token list document (Manning et al, 2008). However, there are many cases where two tokens are not quite the same, but you can get a match. For example, if you are looking for the KSA, hopefully, could link documents K.S.A. On the other hand normalization does not provide enough information for some languages like Arabic (Saad , Ashour,2010)

2.4 Parallel Corpora

"A parallel corpus can be defined as a corpus that contains source texts and their translations. Parallel corpora can be bilingual or multilingual" (McEnery, Xiao, 2007).

McEnery mentions that the parallel and comparable corpora 'offer specific uses and possibilities' for contrastive and translation studies: There is a new perspective on the comparative languages and other ideas that can be observed in unlikely in the monolingual corpora. Multilingual corpora can be used in the investigation of the differences between the languages and can also be used in dictionaries, learning the language, translation and linguistics research (McEnery, Xiao, 2007).

Parallel corpora are resources important for a wide range of applications in the field of corpus linguistics and natural language processing (Heja, 2010). The interest given in parallel corpora has increased considerably, notably due to the boom in the study of information retrieval systems in recent years. OPUS (Tiedemnn,Nygaard,2004) tries to imagine the scientific community a wide range of parallel corpora are available for free in many languages. The main objective is to collect parallel documents from several zones and pre-processing, which are directly useful for

applications such as statistical machine translation, multilingual terminology extraction and a multilingual search engine.

2.5 Related Work

There were a lot of work had been done in the corpora development area we mentioned some of them as follow:

2.5.1 Parallel corpora for medium density languages

In this project building, the bilingual corpus contains the 50M words Hungarian-English. Beside that building a methodology to work with bilingual alignment sentence using the Hungarian-English by doing the tokenization for each word then find the translation from the dictionary then did the alignment for each word in both the language. researcher collect their data from different sources like Literary texts, Religious texts, International Law, Movie captioning, Software internationalization, Bilingual magazines, Annual reports, corporate home pages(Varga et al,2007).

In this project, the researcher developed an algorithm for unsupervised extraction paraphrase from the multiple English translations. During the processing of appropriate penalty Aligned the paraphrase mining on the assumption that the proposals contained in the records that seem aligned rewritten in similar contexts. then automatically deducted it, relationships are ranked good indicators paraphrase sentences surrounding the identical contexts extracted and filtered based on their predictive power. Then these settings are used to identify new features. In addition, paraphrasing the vocabulary of the process of syntactic paraphrasing science, paraphrases syntactic model extracted taught together. Paraphrased extract, and then applied to the corpus and used to examine the scope of the new rules. This iterative algorithm continues to find new circumlocutions (Barzilay et al,2001)

2.5.2 Assessment of a Significant Arabic Corpus

In this project, they created an Arabic corpus by collecting the data from Al-Hayat Newspaper. The Arabic which they collected was Modern Arabic text, and data was gathered from different subjects from 7 different categories. All the text was electronic, and so they could and did investigate the common error spelling mistakes. The literature predicted this would happen, and it may be a problem for language engineering applications. The corpus is however still useful in giving a background for the developments of the techniques discussed. For this corpus, they repeated Yahya's (1989) experiment which was conducted on a small part of a text, and which denoted that Arabic datasets would be much more lax than comparable English ones. This approach may affect the success of standard techniques on Arabic data. However the dataset was not included in the other newspaper (Goweder, De Roeck, 2001).

2.5.3 The arTenTen project

The project's name is "arTenTen Arabic Corpus and word Sketches", and through they developed a large website working with a sketch engine using the MADA tools and tagged with POS. They included billions of Arabic words from different texts, data which had been gathered in 2012, and which was a difficult and time-consuming task (Arts et al, 2014). However, there were still billions of Arabic words not included and it is difficult to update the program as it worked only with sketch Engine.

2.5.4 Open Source Arabic Corpora

In this project, they created an Open Source Arabic Corpora by gathering data from different websites by offline explorer, HTTrack. The method used aimed to modify corpus HTML/XML files into UTF-8 encoding using "Text encoding converter" by WebKeySoft. They considered 10 categories for their work: economics, history, education and family, sport, health, astronomy, law, stories and cooking recipes, with a total 22,429 documents being used (Motaz, 2010). However, some categories were not including, for example, self-development or geography.

2.5.5 Arabic Learner Corpus

In this project, a website was built for the corpus. The data was collected from

learners of Arabic in the KSA native tongue and none native speakers of Arabic were males aged between 16 and 28 years old. All the data was written based on information received from 92 students of 24 nationalities, and it contained 31272 words and 215 written essays. They collected the data and analyzed their mistakes, and they compared the mistakes between the native and none native speakers. The results attained will help teachers and researchers involved in the learning of Arabic. For the annotation of the text they used Lemma, SALMA, and POS, and they developed an error Type tool called ALC. However this corpus does have some negatives – it covers only males, ignoring females, and it focuses on people within a very small range of age (Alfaifi, Atwell,2013).

2.5.6 Quranic Arabic Corpus

This project annotated all of the Quran, showing the morphology, syntax and grammar for any word in the Quran. Any words from any chapters could be selected and the grammar, syntax, and morphology plus the translation for each verse would be shown. To do this the researchers used POS tagging, a form of Natural Language Computing Technology, to find the dependence of the semantic of each word by using math theory. They created a website which had all the verses of the Quran on it (Dukes et al,2010). Whilst this project was one of the best available for looking at the Quran none of the Hadith was included, and it was only useful for Arabic speaking Muslims.

2.5.7 The KSUCCA Corpus

There had been a lot of work done designing of corpora, But there is only one corpus mention clearly that contain Hadith and that is the project is done by Maha Alrabiah for her Ph.D. project in king Saud University design of a Classical Arabic corpus (KSUCCA). She had collected around 50 Million words in her corpus for classical Arabic only. The data is generally classified into six categories, the Religion, Literature, Linguistics, Science, Biography, and Sociology. The corpus file is using UTF-8 for character encoding (Alrabiah et al,2013).she said her corpus contains 44 document of Hadith.

2.5.8 The ICA Corpus

This project was built with the aim to be software for an International Corpus for Arabic which planned to have 100 million words in it. The data was collected from different Arabic countries for different categories, and the researchers in doing so explained why we need a corpus for the vocabulary, regulation, semantics, Natural Language Processing and other language studies. All the files in the corpus were put in the document format(.doc). HTML was used to clean the text and the structural type of the text files was marked. The annotated structural type used were not standard used in all the other corpora it was only described for this tool. The ICA software allowed users to insert a new document, and the searching process and specify were available to each user. However, this software is not available for everyone as it is private software used only by specifically allowed people (Alrabiah et al,2013).

2.5.9 Quran 'Search for a concept' tool and website

Concept tool for Quran by Noura Abbass had allowed the user to search in holy Quran by two-way first one search by keyword and second by the concept, she generated a tree-view concept for Quran from 'Mushaf Al Tajweed' ontology of the Quran. She claims that her tool is more accuracy than another search tool for the Quran (Abbas,2009). Quranic concept search tool is a search for a semantic and syntactic function to recover syntactic information extends with semantics, thus controlling the benefits of retrieving information both syntactic and semantic is used. They show that the combined approach yields better results than the search for syntactic information (Giunchiglia,2008).

2.5.10 The KACST Corpus

This corpus was built to be a great source of Arabic language and form for the lack of Arab corpora. It contains over 700M words, covering a long period, from pre-Islamic until the day of the created the corpus. The data collection was covering time,

region, medium and domain as Al Thubaity mention. Beside the King Abdulaziz City for Science and Technology (KACST) was an open source and available on the World Wide Web with the ability to search the entire corpus depends on the classification, plus some tools had been created specifically to the corpus (Al Thubaity,2015).

2.6 Text Analysis Tools

There has been a lot of analysis and tagged tools build to analysis the corpus. We have described some of them and for each one there is advantages and disadvantages

2.6.1 WordSmith Tool

Wordsmith is text analysis tool that was started last century but that is still improving; at the time of writing its developers had released Wordsmith 6.0, by Mike Scott. It is not free software, but it allows linguists to investigate their corpus via catching the duplication of words, finding dependency words, and searching through the corpus. The tool had three functions; the first one was built by concordance to be the concord function, the second function was built to the register of the repetition of each word by looking at the keyword, and the third one was built to register words using the Wordlist. WordSmith did not work with Arabic text until the 5.0 and even then the Arabic came out in the wrong order (Roberts et al, 2005). Finally, WordSmith 6.0 functioned correctly with Arabic, viewing the text in the right way. I loaded some of the Arabic Hadith onto Wordsmith 6.0.

2.6.2 aConCorde Tool

The aConCorde software was designed by Andrew Roberts in 2013 and the software is free to download and was programmed by java. aConCord works with the text file then produces an analysis of how many words there are in the document and the repetition of each word within the document aConCorde can use different kinds of file types, for example, XML, HTML, plain text and Word files (Roberts et al, 2005) (

Roberts,2009).

2.6.3 Sketch Engine

Sketch Engine is a web-based corpus query system which allows the user to view many corpora in the website or the user can upload their own corpora. Sketch Engine has several functions used for analysis of the words, such as wordlist to see the frequency of each word in the entire corpus, a concordance to create queries about the contexts of words, and many other functions linguistic researchers can use (Kim et al, 2015). The Sketch Engine provided a tool which worked with any corpus marked by any tagged technique, such as Lemma or POS. It had various functions to analyze the corpus, but the most important two were:

1. Concordances which gave a general look of the entire corpus by using a specific query, including the state of the text.
2. Word Sketch had the shortcut stored in it of each word's grammar and a word's consistency (Kim et al,2015).

2.6.7 Xaira Tool

Xaira was a tool designed to be used instead of SARA, it's primary purpose was to be an application used for text analysis on the British National Corpus. Xaira is no longer tied to the one corpus, and thus available for more general use. It took advantage of Unicode and XML technologies to help achieve that (Bernard and Dodd, 2003). It can do complicated search, as well as filter the text with the results depending on the XML annotations (Roberts et al,2005). However, it is not easy to use as some work must be done to the text prior to Xaira application.

2.6.8 The XML Technology

XML is a portable, widely supported and open technology for data storage and exchange. The markup language defined in XML is known as applications.xml and can be written by hand or generated by a computer. Also, it is a meta-mark-up language. The tags in XML are not predefined but rather are user defined. XML is

used to structure and describe information, and because XML can give a large amount of helpful support, it will be everywhere and will be able to be used by anyone. Therefore, we decided to describe the corpus using XML structure (Fawcett et al,2012) (Najeeb et al,2015).In Table 1 below, we had mentioned some of the techniques discussed above advantages and disadvantages.

2.6.9 The Sklearn

Sklearn is a python module ,it is free tool in the internet using for machine learning and data mining analysis working with python and nltk tool to help of preprocessing text ,stemming ,extraction ,tokenization and it has a lot of packages and classes all working together help linguistics and computing working with the text data (Pedregosa et al,2011). Also allow an implementations of many well known machine learning algorithms like SVM(Support Vector Machine) ,K-means and many more.

2.7 Search Tools for Hadith

In this part, we review some website have the ability to search for Hadith. On the other hand, this website builds to make access and search Hadith available on the internet, not for the research purpose.For our comparative reason to our tool we select each tool for one languages one for Arabic, English, French and Russian.

2.7.1 Search Truth Tool

This tool allows the user to search in one of the four books of Hadith (Sahih Al-Bukhari, Sahih Muslim, Sunan Abu-Dawud, Malik's Muwatta).In this tool user can search by English keyword there is three way to search the word : the first user can search for any word means any word similar to the selected word for example (if we search for the word “**real**” the search result had contained (real, realize, really..etc) and same result will get if we select the second option All word. But when we select the third option exact word we had found only the selected word no more.Therefore we think there is no difference between the first and second option. Also, we can search for more than one word for example (the two festival) is one of the concept names.

Table 2.1: The advantages and disadvantages for the tools

Tool	Advantage	Disadvantage
Wordsmith	Works with the plain text of Arabic	It is paid software
aConCorde	Full Arabic support, word frequency , open source, analysis	Ignores mark-up annotation within a corpus
Xaria	Full Arabic support	Text must be annotated with XML first
Sketch Engine	Easy to use, powerful , a variety of query types	It is not FreeSource
XML	very powerful, easy to implement, well formed	Not all browsers support XML

2.7.2 AL Muhaddith Search Engine

AL Muhaddith is website allow users to search in many number of books some of them are Hadith books .The search can be done by Arabic or English languages ,user can decide the search only in specific book or books, search can be by one or two words depends on the user needs.

2.7.3 Dourous search tool

In this search tool, the search had been for only one word, space not allow so the user can search only by one word. For example, if we search for the word(ablutions) the result will be 41 results but if we write(fewer ablutions) the result will come zero. besides, the sometimes very strange result will come with full of wrong text display(it may be browser problems) like see Figure 2.4.

Table 2.2: The finding and limitation for the related works

Investigator	Research	Finding	Limitation
Abdelbaset Powder and Anne De Roeck(2001)	Assessment of a Significant Arabic Corpus	Sundry newspaper Text	Only one newspaper include no the balance of data
Tressy Arts, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff and Vit Suchomel (2012)	arTenTen:Arabic Corpus and Word Sketches	Create web with billion of Arabic words	Only work with sketch engine
Motaz K.Saad and Wesam Ashour (2010)	Open Source Arabic Corpora	Arabic corpus contain 10 categories	There is more than 10 categories have to be included
Abdulla Alfaifi and Eric Atwell,2013	Arabic Learner Corpus	Discover mistake made by Arabic learner	Only for males they ignore females
Kais Dukes, Eric Atwell and Abdul-Baquee M.Shareef,2010	Quranic Arabic Corpus	Grammar, syntax, semantic for each word in Quran	Only for Quran
Maha ALrabiah, AbdulMalik Al-Salman and Eric Atwell,2013	KSUCCA	Morphology of the Quran words	Text not tagged or Annotated
Sameh Alansary , Magdy Nagi and Noha Adly,2013	ICA	Consider Arabic from all Arab world	Classical Arabic only



Figure 2.1:A snapshot of the Search Truth tool



Figure 2.2:A snapshot of AL Muhaddith Searches Engine

2.7.4 Hadith Encyclopedia

In this website contain Hadith in Russian language from different book of Hadith including Sahih Al-Bukhari.user can search by one single keyword like ("веры"), and can search by more than one word, for example, we search for ("Начало откровений") the result was correct we get (7) Ahadith for the first Russian concept.



Figure 2.3:A snapshot of the Dourous Tool

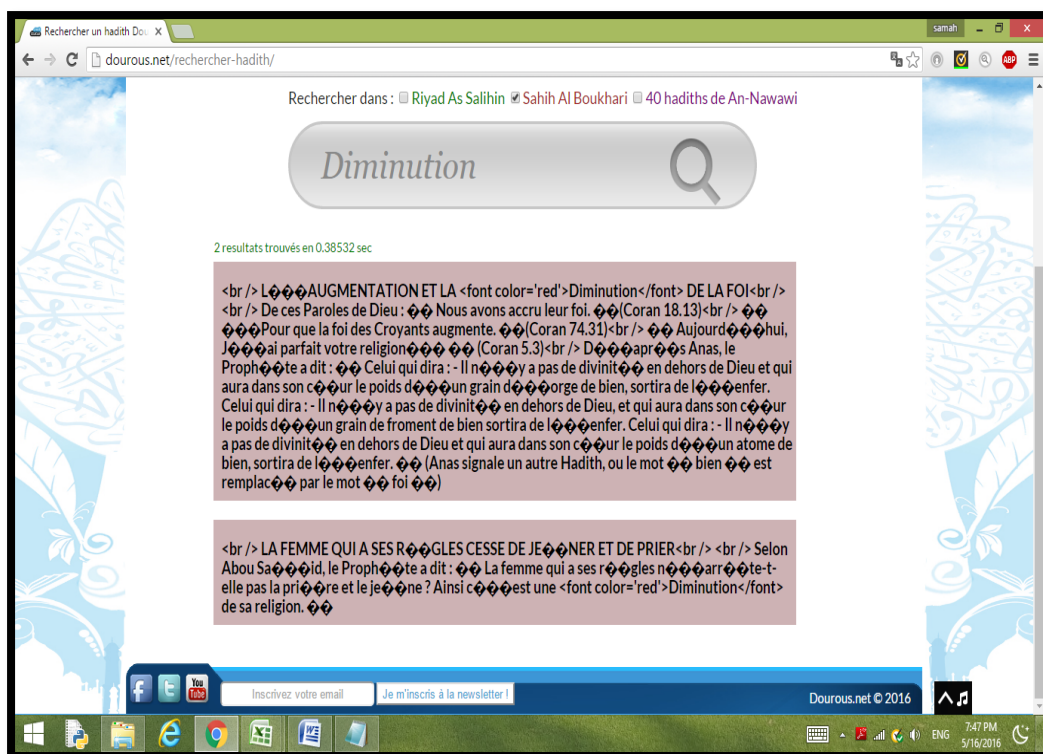


Figure 2.4: A snapshot for the Error result in Dourous Tool

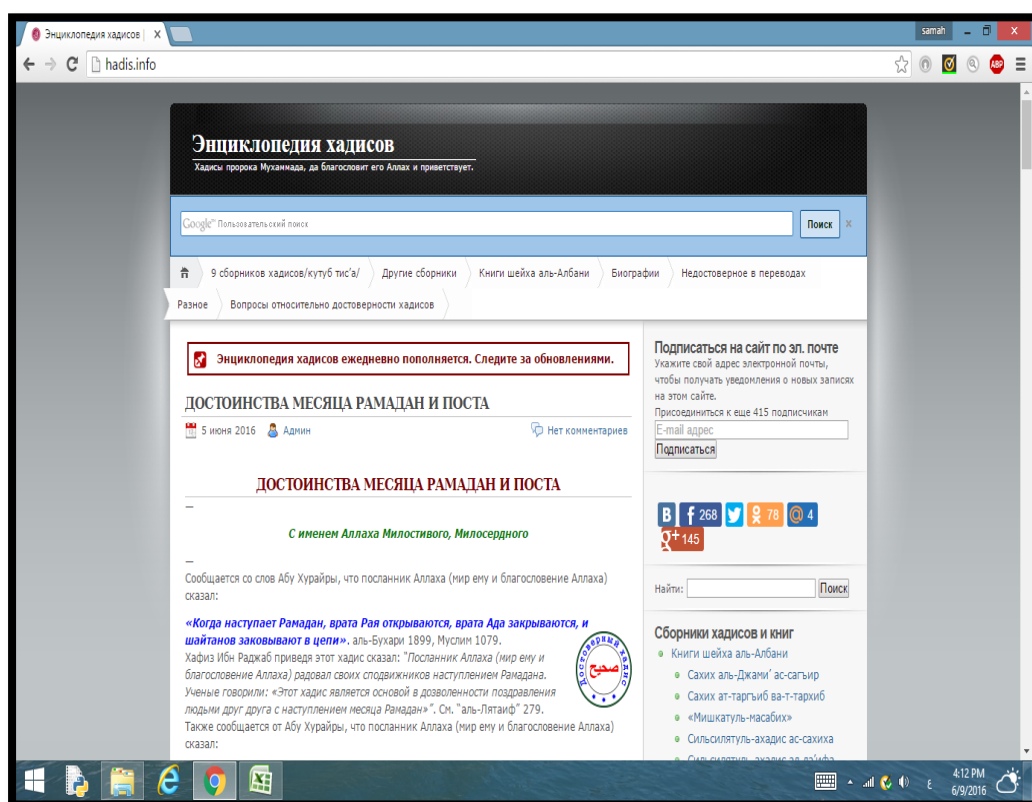


Figure 2.5: A snapshot of Hadith Encyclopedia in Russian

2.8 Key ideas from the Hadith search tools

Hadith search tools allow users to search for English keyword/s. Some allow users to search for Arabic keyword/s. Some allow users to search for English/Arabic keyword/s. Some allow users to search for French keyword. Some search for exact same word/s. Some search for morphemes word/s.

Table 2.3: The features in Hadith search tools

Hadith Tools Features	AL Muhaddith	Search Truth	Dourous	Hadith Encyclopedia
English keyword/s search	N	Y	N	N
Arabic keyword/s search	Y	N	N	N
French keyword search	N	N	Y	N
Russian keyword search	N	N	N	Y
Search for exact match word/s	Y	Y	Y	Y
Search for semantically word/s	Y	N	N	N
Topics index	N	Y	N	N

2.9 Summary

In this chapter, we discussed what the information retrieval systems are also we discussed some methods of text processing like tokenization, removing stop words, stemming and normalization all these method can be apply or some of them only. Beside that we discussed the definition of parallel corpora in the respect of the linguistics sight of view. Finally we review some of the search tools for Hadith which had been found in the internet and we realize that none of these systems are able to satisfy the objectives that were established for this research.

CHAPTER III

METHODOLOGY

3.1 Introduction

This chapter provides full details about the research methodology of this thesis the work was done in five phases. In the first phase intends to discuss the requirements of users who are willing to be familiarized with Hadith. The said questionnaire contains (Appendix A) The first set of questions aimed to personal information. The remaining eight questions about Hadith started by identifying why users reading Hadith also requested them to declare their opinion regarding finding Hadith explanation and word meaning in Arabic, moreover Finding moral for each Hadith, And Hadith explanation, word meaning, and moral in one website. Muslim are finding Hadith in different languages, along with the source and classification of each Hadith respectively. Finally, the users had been requested to indicate their search methods preferences whether like to read Hadith from websites or from books. In the second phase had explained how the data had been collected, what method we use to perform the pre-processing operation. In phase three had showed how the MHC had been structured to be uploaded in the sketch engine. So that researcher has showed how to compile the MHC in one of the most linguistic analysis tools in the internet sketch engine. In phase four the researcher had explained how we had applied vector space model into our corpus, calculating the TF-IDF weights and finally find the cosine similarity between the relevant

documents. Finally, in phase five using the Flask and Python technologies (Grinberg,2014) to develop the website to hold our corpus and the search engine, in the following will give full explain what we are doing in the methodology section.

3.2 The Methodology

In the following sections we had explained in full details how the work had been done:

3.2.1 Phase one: Design Requirements for MHC

This section discusses the requirements of users who are willing to be familiarized with Hadith. The first set of questions aimed to personal information. The remaining eight questions about Hadith started by identifying why users reading Hadiths, also requested them to declare their opinion regarding finding Hadith explanation and word meaning in Arabic, moreover Finding moral for each Hadith, And Hadith explanation, word meaning, and moral in one website. Finding Hadith in different languages, along with the main source and classification of each Hadith respectively. Finally, the users had been requested to indicate their search methods preferences whether like to read Hadith from websites or from books.

3.2.1.1 Important of Hadith

Since the main idea is to build Hadith corpus to offer fully supported and more trusted Hadith source for the benefit of the users, it is planned to identify the requirements of Hadith users. An online survey has been prepared and conducted, which consists of two parts. The first part is relating to personal information of the users; it is about their nationalities, religions, genders, age, mother tongue languages and occupations with avoidance of their private information such as name, telephone etc. The second part is composed of 7 questions, which summarizes the main features of the intended Hadith corpus. Two versions have been created, one in the Arabic language while the other is in The English language, to allow easy collecting

potential users' requirements. In the 22nd of May 2015, the survey was launched and distributed over the electronic emails, Facebook, and other social media. Within two days the number of the replies was a Figure up to 154 from different countries, nationalities, and occupations, as expected. The received information has been analyzed, discussed and thereafter many decisions have been taken as illustrated in section 4 hereunder.

3.2.1.2 Survey result

In the following section had been explained the results and discussions depending on the gathered data:

A. Religion

One of the most important pointers of the questionnaire findings is the Religion of participants, which Shows that 99% are Muslim (see Figure 3.1). It is Obvious that potential users will most probably be Muslims. But other users with different religions will be targeted to use the intended MHC for their purposes i.e. educational and / or religious. Therefore, the MHC will be built to serve both Muslims and non-Muslims.

B. Gender

1540 Persons has participated in the survey, with 708 (46%) females 832 (54%) are males. This indicates that the MHC is appreciated by both genders. Therefore Muslims and non-Muslims, regardless their gender, can benefit from the intended MHC as illustrated in (see Figure 3.2).

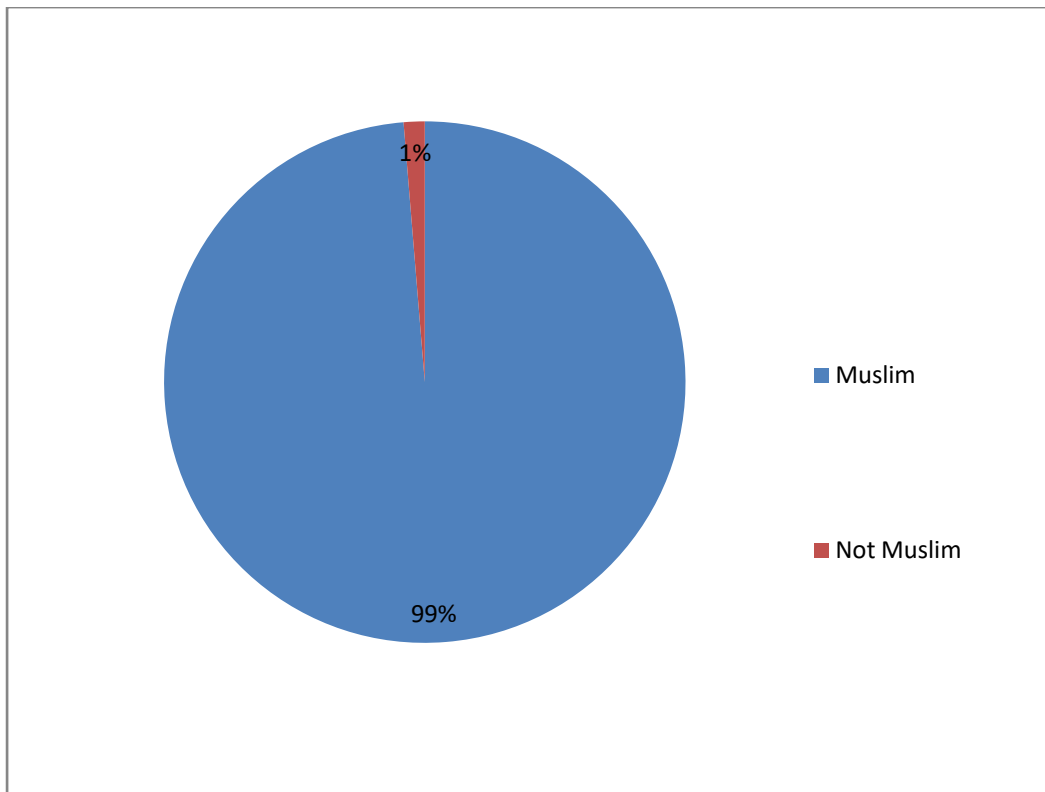


Figure 3.1: The MHC Muslims & non-Muslims Participants

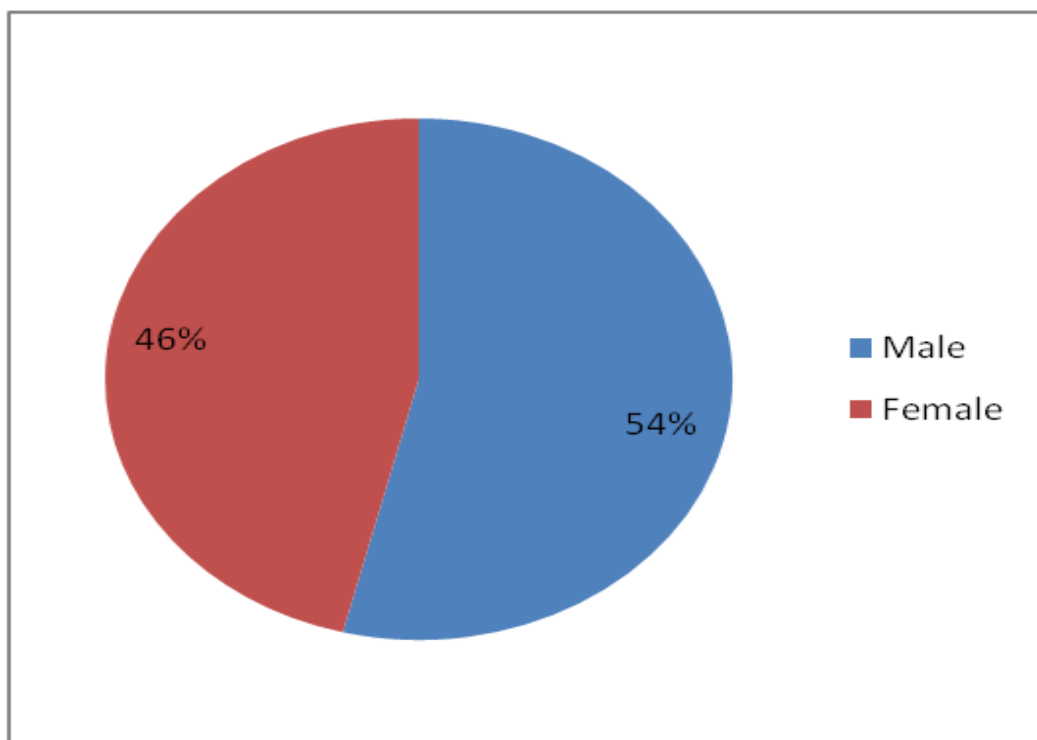


Figure 3.2: The MHC participation of both gender

C. Age

Age is usually used to compare different groups of people used to meet the different interests and way to get different views of different ages. In our case this will lead to discover what the user needs to read and find about Hadith and this will give us about what to consider when we design these requirements. The survey results show that 19% of the contributors are 15-25 years old, and 19% are 41-60 years old while 0% is above 60 years old. The remaining 62% is between 26-40 years old. This indicates that the people above 60 are not included in our survey as some report separated them into two parts: First users are those interested in this technology, but they have hesitation about the use of the fear of extreme complexity or high costs. This category needs to step on the feet of the use of the network to some support and practical help and encouragement from the social environment, second type who do not care about the Internet and find internet not useful to them[25]. So that people in middle age and youth users who are already familiarized with using the new technologies will constitute the main target group for the anticipated MHC as they prefer to exploit the latest internet application. Figure 3.3 is showing the contribution of target users from different ages.

D. Nationality

Potential users from different corners around the world demonstrated their desire to have such IHC that will satisfy their requirements. They are Saudis, Sudanese, Egyptians, Jordanians, Russians, French, Yemenis, Syrians, Pakistanis, Indians and British as illustrated below (see Figure 3.4). Users from other nationalities that have not participated in this survey due to time constraints can also be considered as possible users for the intended MHC.

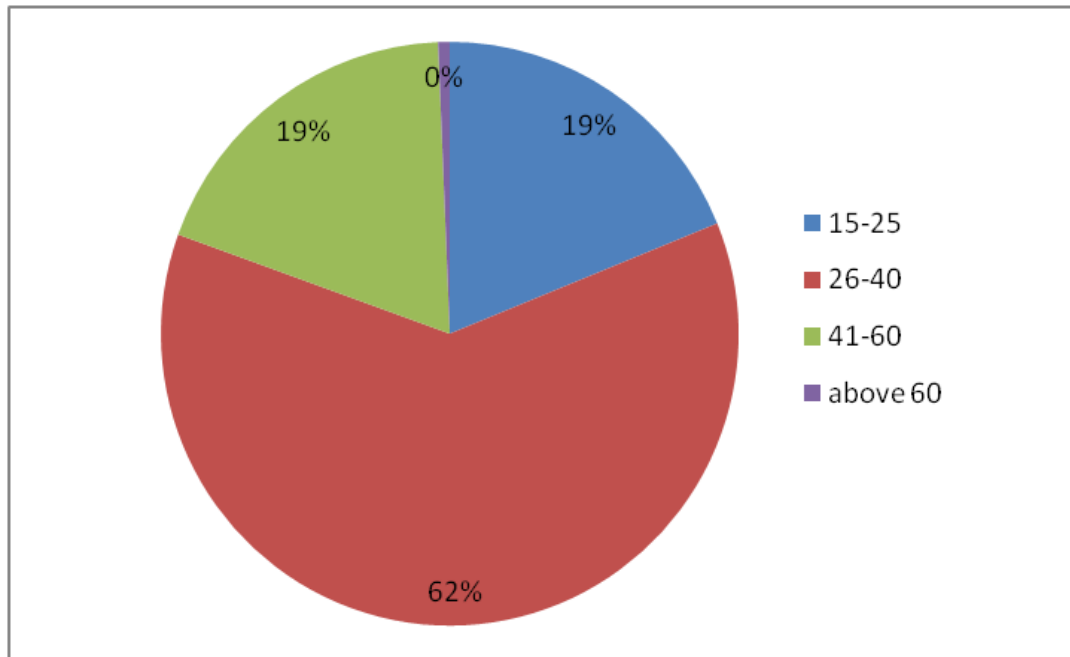


Figure 3.3: The MHC participants of different age

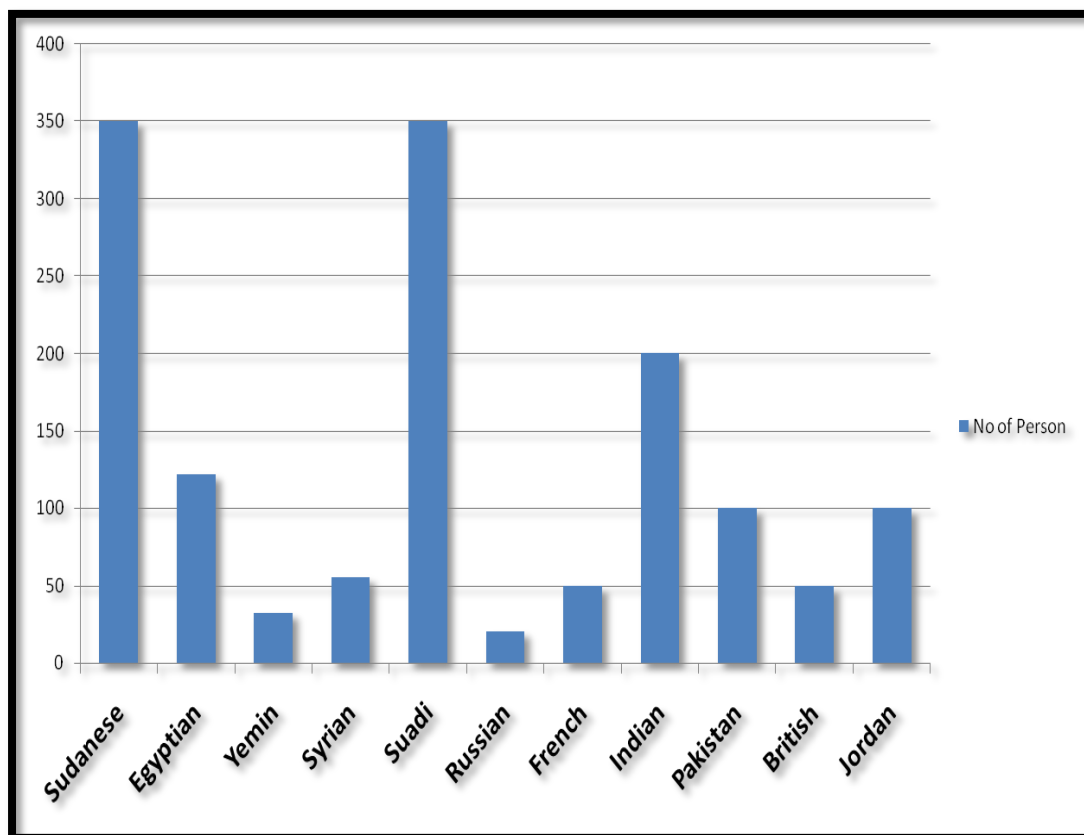


Figure 3.4: The MHC participants from different nationalities

E. Religious and educations purposes users

The MHC is projected to serve many purposes, the most important are the religious one. There is a need for an all-inclusive website that contains Prophet Mohammed (peace be upon him) Hadiths, meaning and moral of each Hadith. The survey's question are about why the users wish to study about Hadith in MHC shows that 69% assuring that they do it for religious reasons while 31% required to use this MHC for both motives religious as well as educational motives (see Figure 3.5). Hadith is a very important for Muslims because the MHC will enable them to read more about prophet Mohammed (peace be upon him), about his doctrine his private life and his companions. Also Muslims effort their best to learn from Prophet Mohammed (peace be upon him) and from his Hadith to turn into good Muslims and to amplify their faith to obey Allah instructions.

F. Hadith explanation in Arabic

This category helps us to see if the Arabic speaker needs the explanation for each Hadith. The questionnaire's participants respond to the question regarding preferring to have in each Hadith explanation in Arabic was Figure up to 75% believe that it is very useful and 21% assume it is useful while only 4% consider it is not useful (see Figure 3.6). Therefore, from this result realized that explanation in Arabic is very important to be there in the Hadith corpus.

G. Word meaning

Similar percentages had been observed regarding the response of the participants to survey's question concerning their desire to have for each Hadith meaning of each word in Arabic. 71% of them seems it is very useful and 26% believes it is useful while 3% only consider it as not useful (see Figure 3.7). Therefore knowing the word meaning for each Hadith is useful as per the decision of 97% participants, the thing that confirms their interest in having the word meaning of each

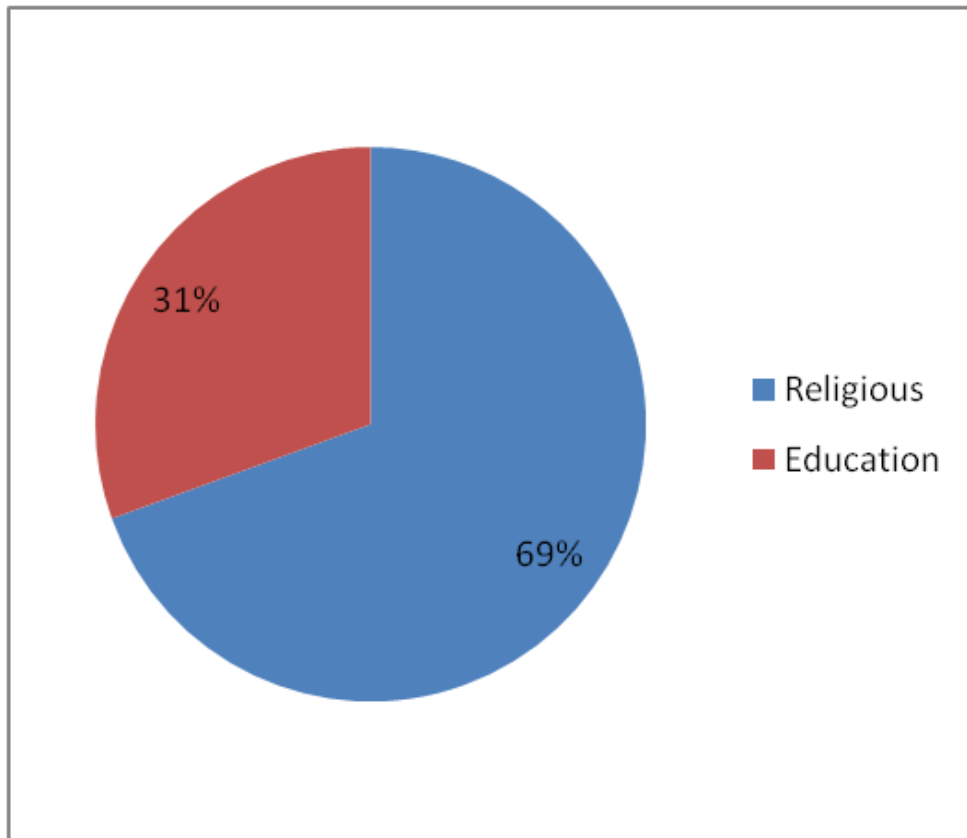


Figure 3.5: The MHC religious and educational motives

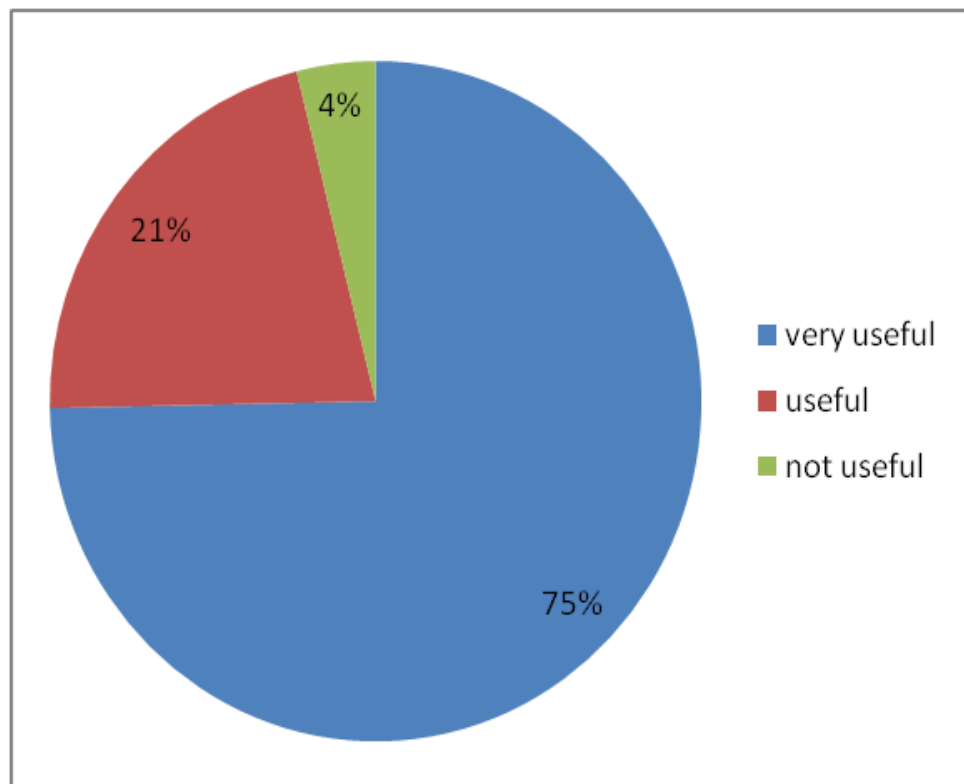


Figure 3.6 : The MHC Hadith explanation in Arabic

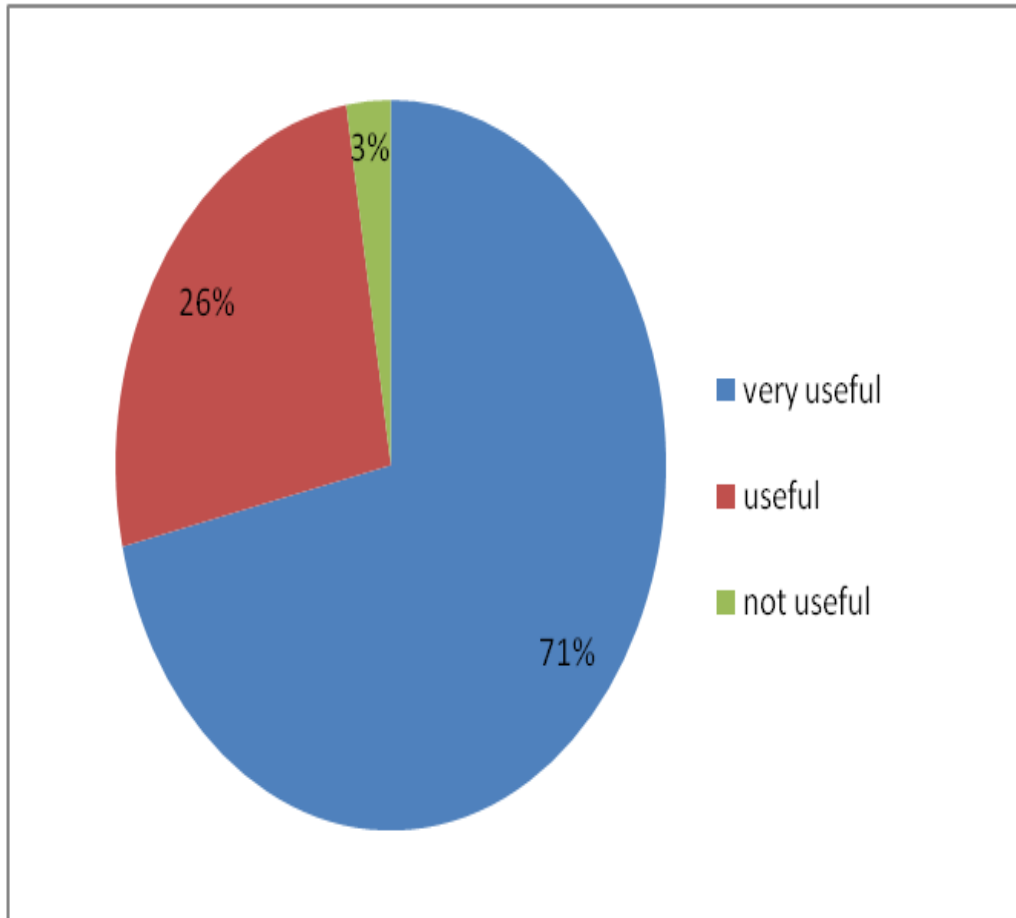


Figure 3.7: MHC respond to having the Meaning of words in Arabic

Hadith and this is the very high percentage in which the purposed MHC expectations can depend on.

H. Moral and Lesson learn

One of the most significant constructive usages of the intentional MHC is the inclusion of the moral (lesson learn) from each Hadith. 70% of the participants consider the inclusion of the moral from each Hadith in the MHC is very useful and

29% suppose it is useful while 1% had selected the not useful answer (see Figure 3.8). As a conclusion, 99% of all participants voted for very useful and useful for the inclusion of the moral of each Hadith in the MHC, which will facilitate Muslim to Figure out the right directions for their religious life and personal life as well.

1. One website

Searching on the internet is somehow time-consuming, so in the survey one question had been dedicated to discovering the users opinion of having Hadith explanation, word meaning and moral(lesson learn) on one website. The questionnaire outcome proved the participant willing to have Hadith explanation, words meaning and moral from each Hadith in one location is very useful as 77% of them decided that it will be very useful and 22% of them thought it well is useful i.e. 99 % are willing to get their information from one location on the internet (see Figure 3.9). This is, of course, will save the user time and will be worth significance.

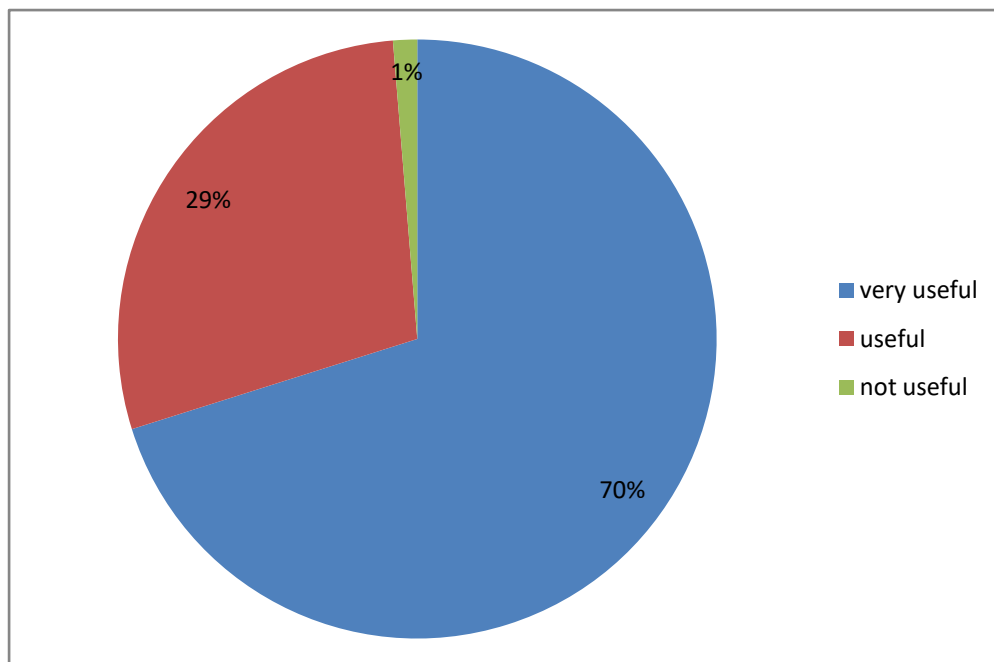


Figure 3.8: The MHC moral gained from Hadith

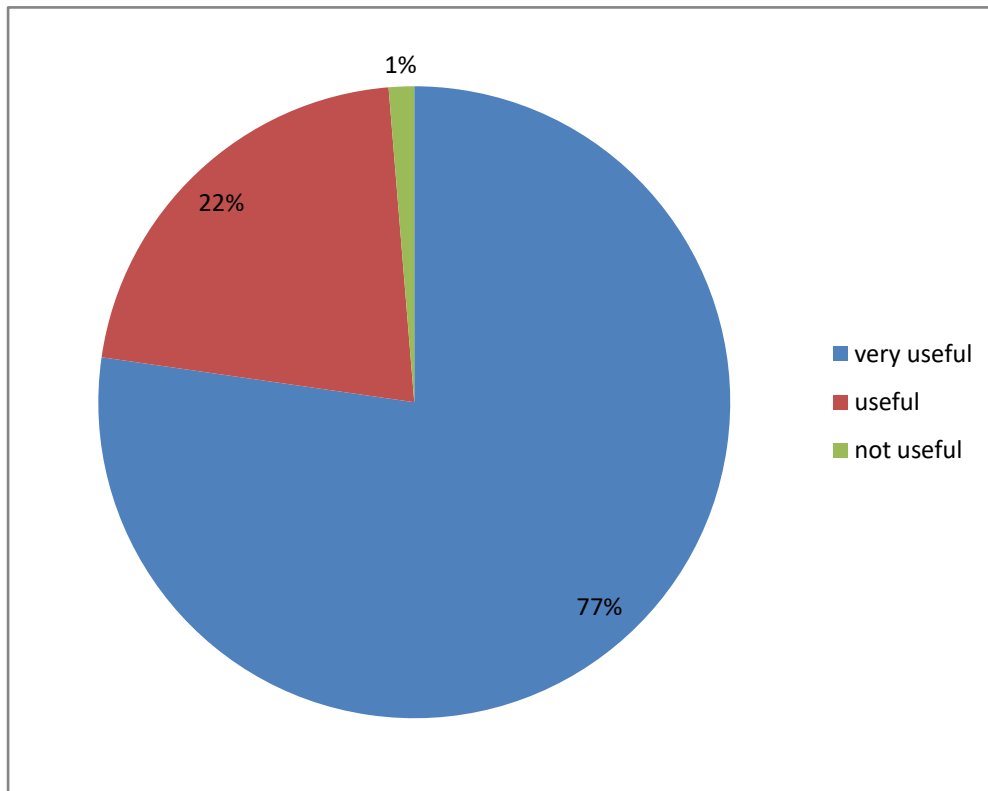


Figure 3.9: The MHC one website feature

J. Hadith in different languages

The original language for Hadith is Arabic, However for Muslims learning of Arabic language will be added value. To measure the importance of Hadith translations in different languages in the survey participants had been requested to answer the question if they find Hadith in different translation is very useful, useful or not useful ?. The answer to the question had been showed that 65% of the participant consider it as very useful, 25% of them think it is useful and 10% of the participants believe that the matter is not useful(see Figure 3.10). Therefore, users think that if they find Hadith in their original language that will allow them to understand Hadith very well, moreover this will make the Hadith statements so clear for them, normally when if someone want to study something it better to study it in his mother language for better understanding and implementation of gained knowledge.

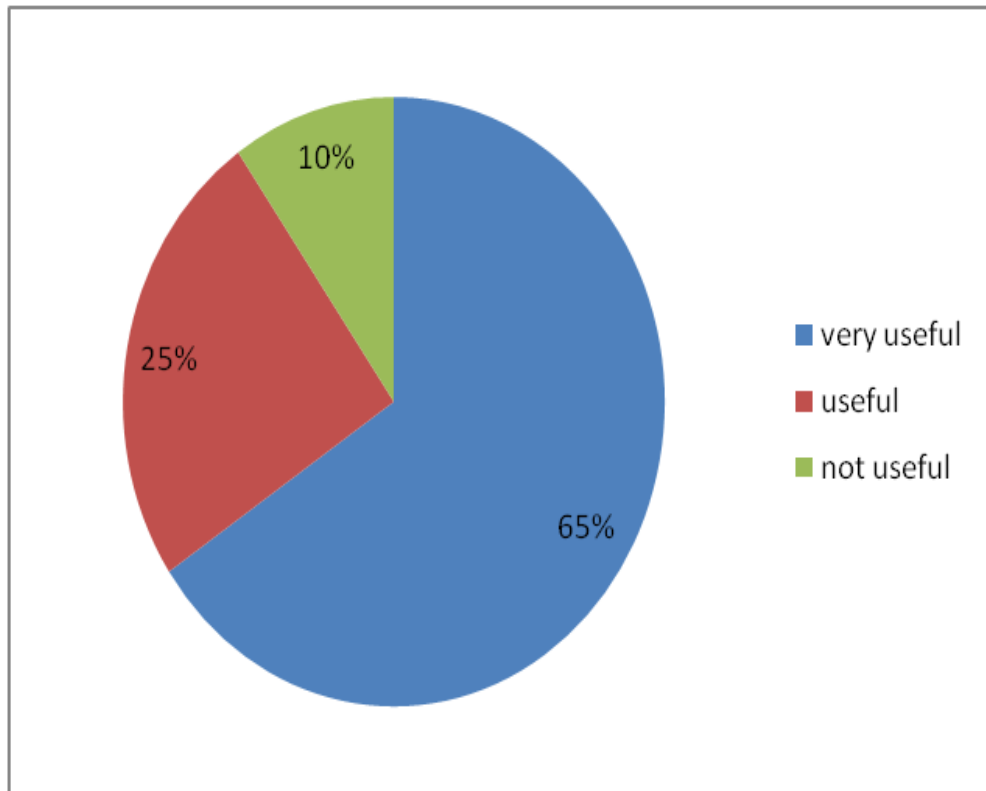


Figure 3.10: The MHC Hadith in different languages

K. Hadith source

The most recognized Hadith sources are Sahih Al- Bukhari, Sahih Muslim, Riyadh Al-Saliheen and many other books. The questionnaire outcomes confirmed that 64% of the participants prefer and consider the information about from which source the Hadiths are copied in the intended MHC is very useful, 33% believe it is useful and 3% declare that is not useful (see Figure 3.11).

L. Hadith classification

Reading Hadith is very important for Muslim life so when Muslim find any Hadith they prefer to check if this Hadith is strong(correct(Sahih) or weak(Daeef) or good{(Hassan). In the survey, the participants respond to the question regarding their

willingness to have the classification of each Hadith include in the intended MHC e.g. correct(Sahih) or weak(Daeef) or good (Hassan).The final responds to this question were 86% of participants agreed that is worthy and very useful and 13% of them decided that is useful and the remaining 1% are not agreed and declared that is not useful (see Figure 3.12).

M. Hadith Search

Regarding the searching facilities in the intended MHC, the survey inquired the prospective users whether they prefer to search about Hadith from books, from one website or from different websites.The perspective users had responded to this inquiry as follow, 33% of the participants prefer to have searching facilities for all information within one MHC,22% of them like better to obtain the required information from different websites ,23% would like to depend on books only for collecting the desired information ,22% of them think it is worthy to look for the information on the websites as well as in the books (see Figure 3.13).From the above it is obviously that users divided into two major groups, the first group prefer to utilize the new technologies, such as websites over the internet and the second is traditional users who used the hard copies books as have lag of trusting in website information, whoever by developing authenticated and authorized MHC to serve the user's needs in searching for the required information and data, then the users from the second group can be converted to a backup target group for the intended the MHC

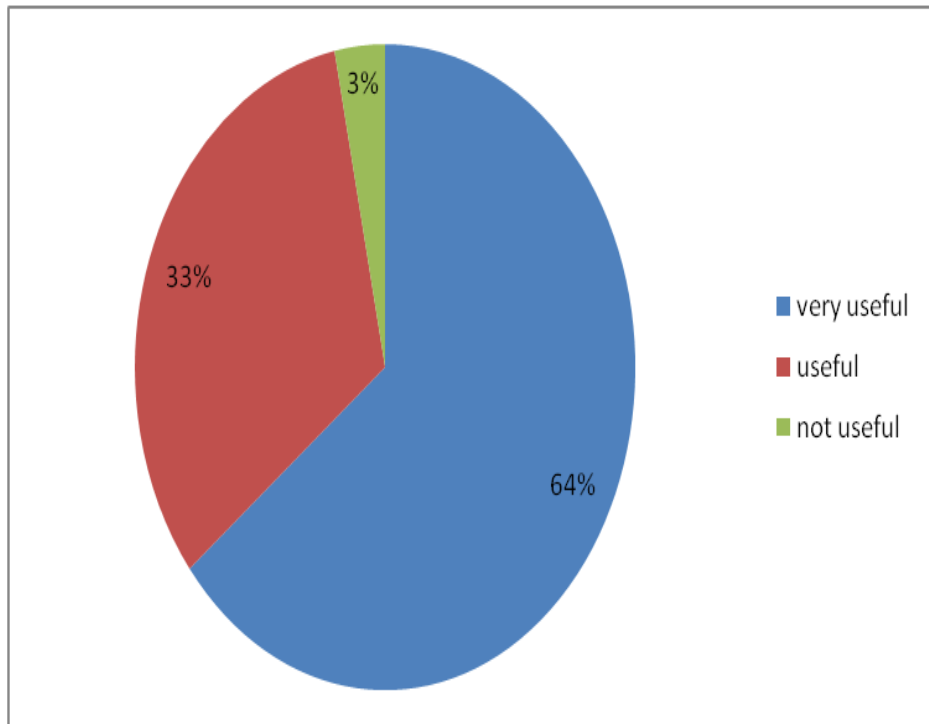


Figure 3.11: The MHC Hadith Sources

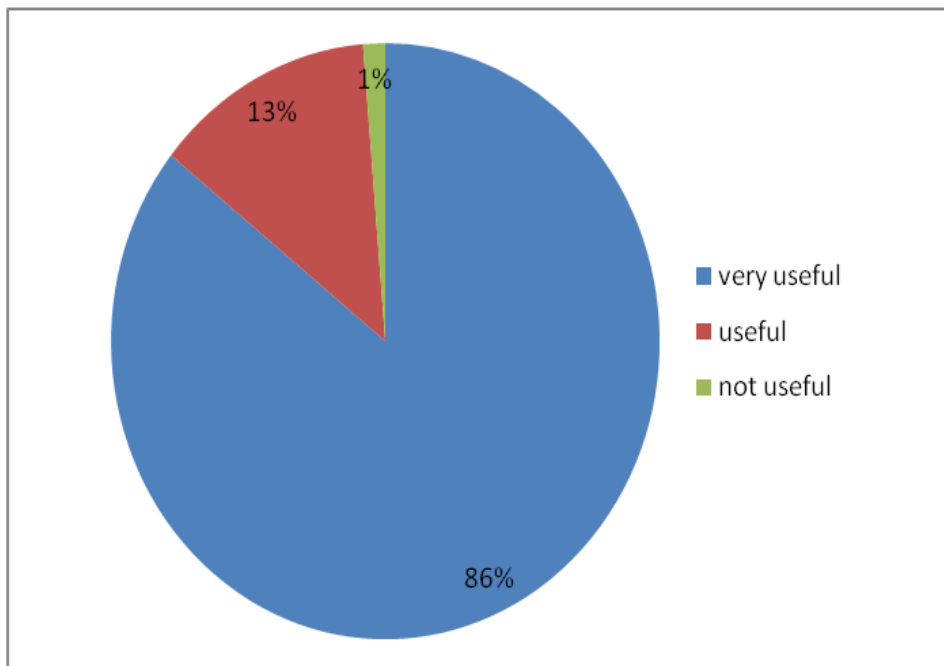


Figure 3.12: The MHC Hadith Classification

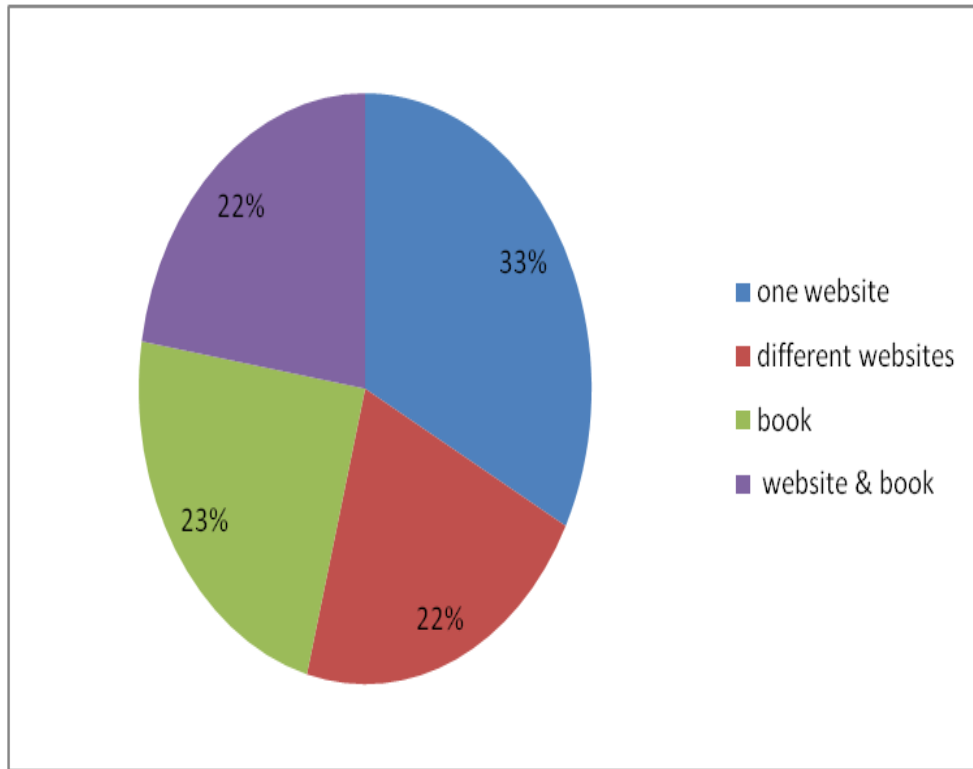


Figure 3.13 :The MHC Searching Facilities

3.2.2 Phase two: Collect the Data

First, in this section will start by explaining in details how we collect the data, pre-processing data, data annotation, and files generation.

3.2.2.1 Data collection

- Searching the internet, considering it as the biggest source of corpora, from which we can derive our texts.
- Selecting and organizing the texts, which will include written texts in Arabic, English, French and Russian. Text collection was done manually because the text was found in several different formats like (.doc,.pdf).
- The process for the (.doc) file was straightforward copy and paste. On the other hand, the (.pdf) files had to convert first to editable text then we can copy from this; for that job we used Nitro Pro 10 converter. This software works perfectly with the English text but makes noticeable mistakes for Arabic and French, so we have to run manually through the text for correction

of pdf-to-text conversion errors.

3.2.2.2 Data Cleaning

The cleaning means removal of data that is irrelevant, as well as the data that we do not want. Cleaning the data is a time-consuming job; we tried two ways to do this:

- Manually: to remove individual unwanted words or characters, or to correct some wrong word, or to replace one word with another word.

- Semi-Automatically: by using the find and replace tool available in MS-Excel to:

Replace all cases of a wrong word with the corrected one

Replace all cases of a character with another character

Remove the diacritics

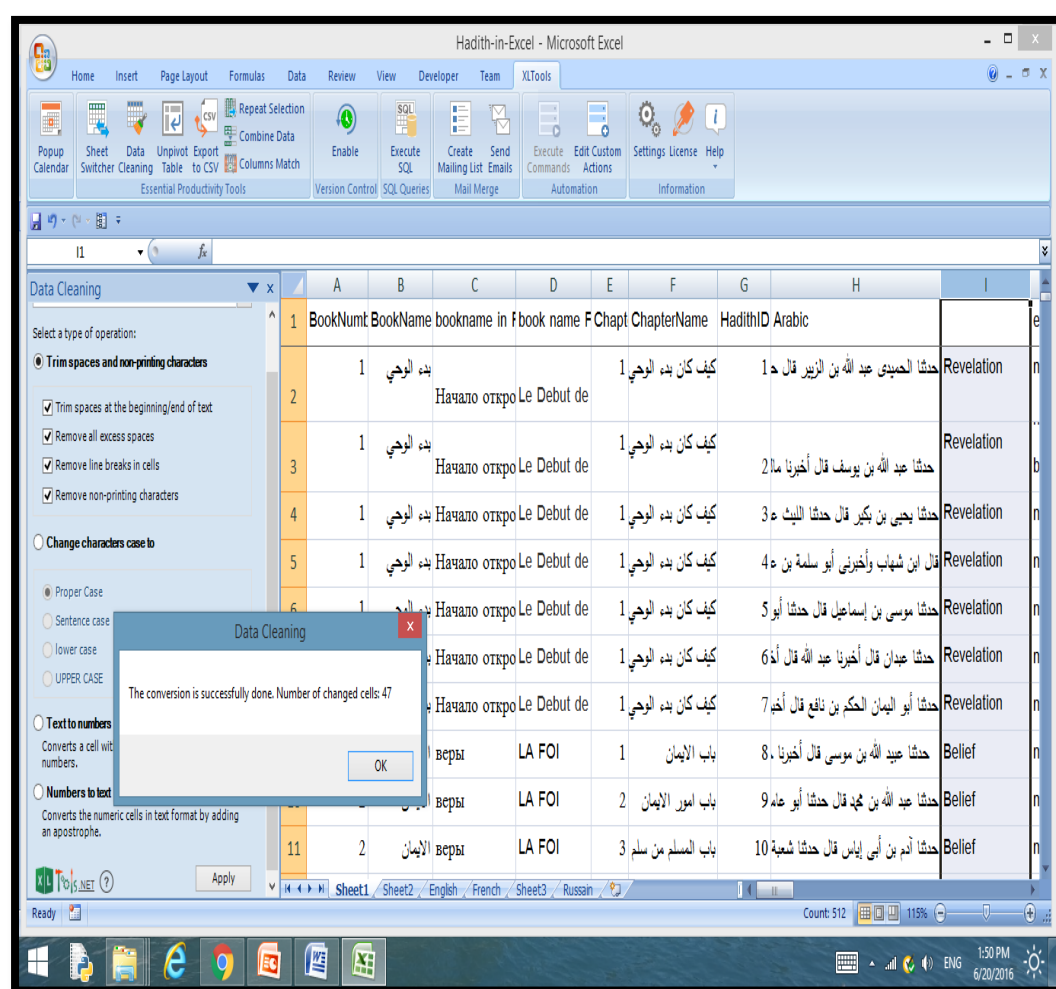


Figure 3.14: Snapshot of the XLTools

Automatically by XLTools(Figure 3.14) After we organize the data in Excel file we use the XLTools to ,remove spaces from the beginning/end of the text ,Remove all spaces between text ,Remove line breaks in cells ,Remove non-printing characters And change character case to lower/upper case.

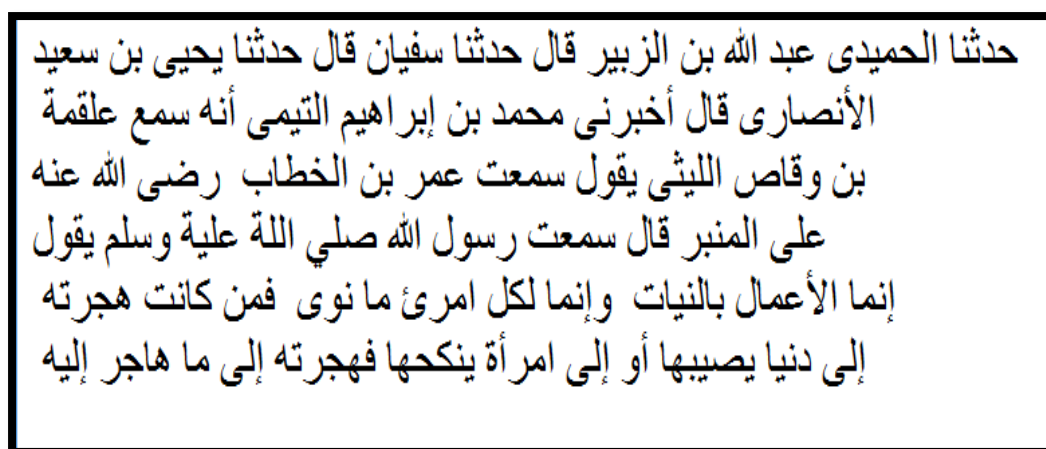
3.2.2.3 Data Preprocessing

We save the English text in the .csv file as raw document consider each raw is one document(Hadith), the English text contain 2030 documents

Tokenization: take the entire English document put in the way of terms using the built-in function `word_tokenize()` from the `nlTK` package built in the NLP(Natural language Processing) tool (Chen,2010), so when we run it we get 92,577,5 separated tokens from the English text.

3.2.2.4 File Generation

Copying a text from the Internet and pasting it into Microsoft Word, every text was encoded using Unicode UTF-8.Finally, we have clean data in the plain text format saved into four separated files:Arabic.txt see Figure 3.15, English.txt see Figure 3.16,French.txt see Figure 3.17 and Russian.txt see Figure 3.18.



حدثنا الحميدى عبد الله بن الزبير قال حدثنا سفيان قال حدثنا يحيى بن سعيد
الأنصارى قال أخبرنى محمد بن إبراهيم التيمى أنه سمع علقمة
بن وقاص الليثى يقول سمعت عمر بن الخطاب رضى الله عنه
على المنبر قال سمعت رسول الله صلى الله عليه وسلم يقول
إنما الأعمال بالنيات وإنما لكل امرئ ما نوى فمن كانت هجرته
إلى دنيا يصيبها أو إلى امرأة ينكحها فهجرته إلى ما هاجر إليه

Figure 3.15: Plain text for example Arabic Hadith

narrated umar bin alkhattab i heard allahs apostle saying the reward of deeds depends upon the intentions and every person will get the reward according to what he has intended so whoever emigrated for worldly benefits or for a woman to marry his emigration was for what he emigrated for

Figure 3.16: Plain text for example English Hadith

selon alqama ben ouaqas el laïti alors quil était sur le minbar omar ben el khattab prononça les paroles suivantes jai entendu lenvoyé de dieu dire les actions ne valent que selon les intentions pour chaque homme les intentions sont déterminantes ainsi celui qui émigrera pour les biens de ce monde ou pour chercher une épouse ne sera rétribué que pour lobjectif quil sétait fixé

Figure 3.17: Plain text for example French Hadith

Начало откровений
 Глава: О том, как откровения начали ниспосы сланнику Аллаха, д
 Сообщается, чт бин аль-Хаттаб, да будет доволен им Аллах, сказа
 «Я слы посланник Аллаха, да благословит его Аллах и приветству
 Передают со слов ‘Аиши, да будет доволен ею Аллах, что (одна
 «О посланник Алла приходят к тебе откровения?» Посланник Ал
 ‘Аиша, да будет доволен ею Аллах, сказала:
 «И м съ видеть, как в очень холодные дни ему ниспосылались от
 Сообщается, что мат верных ‘Аиша, да будет доволен ею Аллах,
 «Ниспослание откровений посланнику Аллаха, да благословит е

Figure3.18: Plain text for example Russian Hadith

3.2.2.5 Data annotation

So far, we had collected a file for each language, contain all Hadith found on the WWW in that language. The next step is to create a parallel multilingual corpus, with an XML file for each Hadith with parallel translations and metadata. To create the XML file for each Hadith we need to map each Arabic Hadith to the translation in the English, French, and Russian. This work had to be done manually to recheck that each Arabic Hadith matches the correct translation in the other languages because Hadith is very sensitive text; we used bilingual volunteers to check the English, French, and Russian translation texts.

We built one MS-Excel file to store all the mapped Hadith translations and metadata correctly. We notice from (Table-2) that we did not find the same number of Hadith from all the Languages. For that reason, we would considered only the intersection of all four Hadith collections, and the number of Hadith in common to all four is limited to the smallest language file, which is English with 2030 Hadith. We mapped it with all the 3 languages (Elanwar, R.I.M, 2012). To generate XML schema see (Table 3.1), for the data and apply it to the data in the MS-Excel file for each

row,each row will have the 4 text of Hadith from the different languages plus some metadata or information about each Hadith like Hadith number, book name, book number, see (Table 3.2). Finally, save the file as (H_1_E_A_R_F.xml)to indicate H for Hadith,1 was the Hadith number, E for English, A for Arabic text, F for French text and R for Russian text, see Appendix (sample-1). Consequently, we have 2030 XML files.

Table 3.1: The XML tags for Multilingual Hadith Corpus

Tag	means
<Data>	Tag for element root
<Hadith_Arabic>	Tag Arabic Hadith
< Hadith_ English >	Tag English Hadith
<Hadith_French>	Tag French Hadith
<Hadith_Russian>	Tag Russian Hadith
<Hadith_Explian>	Tag for Hadith meaning in Arabic
<Hadith_Source>	Tag for the source book name
<notice>	Tag for extra information
<Hadith_Narrated_by>	Tag for who narrated the Hadith

Table 3.2: The Column name in the MS-Excel file for the MHC

اسم الكتاب	رقم الكتاب	اسم الباب	رقم الباب	رقم الحديث	نص الحديث عربي	نص الحديث انجليزي	نص الحديث فرنسي	نص الحديث روسي
Book name	Book id	Chapter name	Chapter id	Hadith id	Arabic Hadith	English Hadith	French Hadith	Russian Hadith

Finally, The result of this part was that we have a Multilingual Hadith Corpus with around 2Million words of Hadith from the four languages: Arabic, English, French, Russian. Besides we have one document contain Arabic Hadith along with the three language translations in one XML file as you see in Appendix (sample-1).All these

Hadith come from SAHAI ALBUKHARI only. Table 3.3 shows that number of Hadith and number of words in each language file. The sizes differ because not all the Hadith are translated into all three other languages. So we were not able to find the translations for all the Arabic Hadith online.

Table 3.3: The number of Ahadith by words in the Corpus

Language	Number of Hadith	Number of words
Arabic	2030	66,808,1
English	2030	92,583,7
French	2030	48,943,1
Russian	2030	38,756,4
Total	8120	2,470,913

3.2.3 Phase Three :Vector space model

Retrieval methodologies assign a measure of similarity between a query and a document. These methodologies are based on the common notion that the more often terms are found in both the document and the query, the more "relevant" the document is assumed to be to the query.

Additionally, a retrieval methodology is an algorithm that takes a query Q and a set of documents D_1, D_2, \dots, D_n and identifies the cosine similarity $\cos(\theta)$ for each of the documents $1 \leq i \leq n$. Our model is based on the idea of each document can be represented into vector and by the same way represent the query into vector then select method to measure the closeness of any two documents. Therefore we had converted the entire corpus documents into vectors consider that each document (Hadith) represent one vector (v) in the space (see Figure 3.19), and to avoid the different length

of each document had transformed all of them into same length of vectors using the tfidf representation in the sklearn (Pedregosa et al, 2011). so that will be accept the

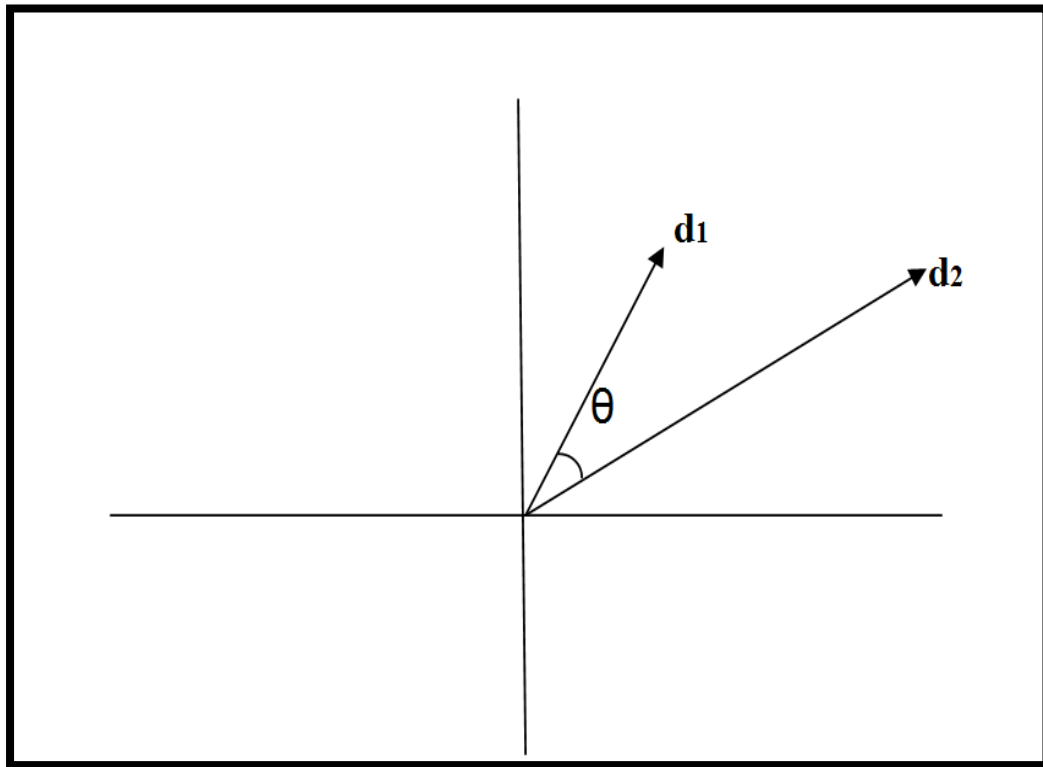


Figure 3.19: The angle(θ) between two vector documents

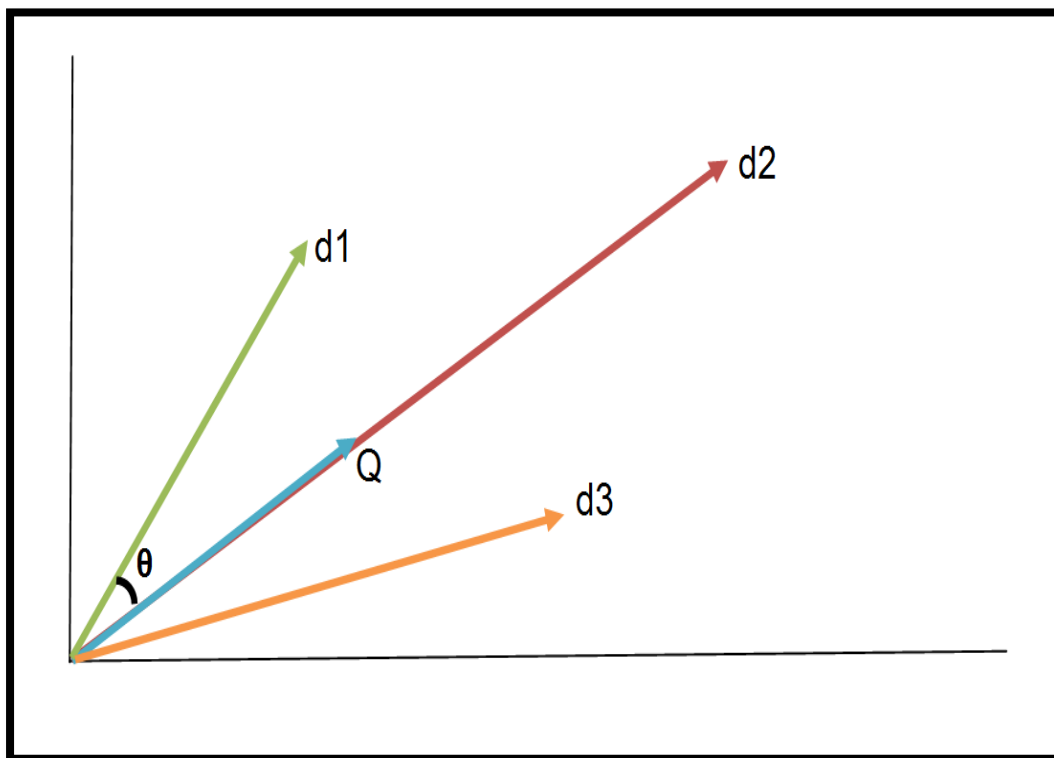


Figure 3.20 : The similarity between Q and d_2 in the vector space

Query from the user converted to the vector by the same size of the vector documents then apply the cosine similarity to find the most relevant documents to the user query. For in Figure 3.20 show that there are 3 documents (d1,d2,d3) and Q in the vector space model each document represent as vector and we notice that d2 and Q are similar because the vector between the two vectors are 0 (Zero) or in another word the cosine(0)=1 and that means there are 100% similarity.

For the formal definition of and declare the use of weights based on the collection frequency. Weight is computed using the TF_IDF and for that to construct a vector that represent each document ,we had used the following formula from (Grossman et al,2012):

$$idf_j = \log \left(N / df_j \right) \quad (3.1)$$

Where:

df_j : number of document which contain t_j

N : is the total number of documents.

Calculation of the weighting factor(tfidf) see Equation(3.2) for a term in a document is defined as a combination of term frequency(tf),and inverse document frequency(idf) see Equation (3.1) .To compute the value of the j th entry in the vector corresponding to document i , The following equation is used:

$$W_{ij} = tf_{ij} \times idf_j \quad (3.2)$$

Where:

tf_{ij} : Is number of occurrences of the term t_j in document D_i

W_{ij} : Is the weight of the term (i) in specific document (j).

we represent our corpus as a group of documents ($D1, D2, \dots, Dn$) When a document retrieval system is used to query a collection of documents with t terms, the system computes a vector D ($d_{i1}, d_{i2}, \dots, d_{in}$) of size n for each document. The vectors are filled with term weights as described above. Similarly, a vector Q ($W_{q1}, W_{q2}, \dots, W_{qn}$) is constructed for the terms found in the query.

The vector is a representation of the mathematical deal with the numbers only so when we want to convert the texts to the vectors must use numbers in this study we decided to use a statistical measure tfidf in order to be able to represent each document in the figure of the vectors so that the application can find the angle between two vectors by calculation of the cosine measure of the angle between the two vectors will lead to the similarity value. However, the cosine value is between $[1,0]$ if the cosine is 0 that means the two vector are orthogonally and there is no similarity between them, on the other hand If the cosine is 1 that means the two vector are similar or in the same direction to each other and the angle between them are 0, in general the cosine value near 1 means small angle and more similar and the value closer to 0 means large angle with less probability of similarity. A cosine similarity (CS) between a query Q and a document D_i is defined by the product of the two vectors. Since a query vector is similar in length(n :vector size) to a document vector, this same measure is often used to compute the similarity between two documents:

$$\text{Cosine}(D_i, Q) = \frac{\sum_{j=1}^n d_{ij} \cdot q_j}{\sqrt{\sum_{j=1}^n d_{ij}^2 \cdot \sum_{j=1}^n q_j^2}} \quad (3.3)$$

Where:

d_{ij} : is the document vector.

q_j : is the query vector.

To speed up our scanning process instead of scanning the entire collections of documents we created inverted Indexes for the distinct terms mapping each term to the specific posting list which contain the document numbers for all the documents contains that term (Goker , Davies, 2009).In order to build the inverted indexes for our collection we use the python dictionary which is consist of two important things keys and values and for our purpose the terms represent the dictionary “keys” which is unique that mean each term represent one “key” ,on the other hand the values can be a list of number which in our case represent the documents numbers Table 3.4 show the sample from the inverted index of English terms.

Table 3.4:The Inverted Indexes
for the English corpus

Index (Key)	Posting List (Value)
Stress	1167
Buyer	[532, 535, 539, 769, 1315]
Pain	[459, 769, 1278]
Struck	[319, 458, 459, 750]
Sinner	[392, 437, 610, 611, 699, 868, 1011, 1126]

Table 3.5: The Inverted Indexes
for the Arabic corpus

Index (Key)	Posting List (Value)
ر هط	1217 ,1004 ,603 ,453 ,320 ,319
تكبير	,1780 ,1779 ,1741 ,1695 ,1690 ,1662 ,1291 , 1931 ,1797
باسط	1920,1939
صيام	,933 ,906 ,894 ,885 ,883 ,882 ,881 ,880 ,879 , ,968 ,966 ,961 ,943
وعظ	1988 ,1938 ,1936 ,1818 ,436 ,423 ,404 ,213

Table 3.6: The Inverted Indexes
for the French corpus

Index (Key)	Posting List (Value)
Fassion	1559
gardiennag	691
Convert	562, 646, 740, 741, 1304, 1420, 1450, 1704
descent	958, 1015, 1016
vient	157, 162, 163, 571, 1451

Table 3.7: The Inverted Indexes
for the Russian corpus

Index (Key)	Posting List (Value)
присутств	45, 95, 98, 667, 758, 1576, 1631, 1932
деся	6, 680, 894, 934, 1397
локт	304, 305, 315, 638, 915, 1172, 1742, 1823
Состоян	90, 133, 357, 706, 1035, 1304, 1344, 1347, 1359, 1360, 1383
локт	304, 305, 315, 638, 915, 1172, 1742, 1823

3.2.3.1 English Corpus

Convert the collection of the Ahadith from the (englishhadith.csv) shown in Figure 3.21 to the list of documents to do that a proposed feeding algorithm is used for generate the English corpus and feed the corpus into TfidfVectorizer as shown in

Figure 3.22. The python programming is used to implement these algorithm. After run the code only 2030 documents had been generated. Therefore the TfidfVectorizer will compute the TF_IDF weight for each token in the entire corpus. Consequently a sparse matrix of 2030x5313 of type '<class 'numpy.float64'>' with 60379 stored elements in Compressed Sparse Row format. So that mean will end up with only 7,196 distinct terms from the English Hadith text out of 92,583,7 terms.

In order to be able to store such a matrix in memory but also to speed up algebraic operations matrix / vector, implementations will typically use a sparse representation such as the implementations available in the (scipy.sparse) package. SciPy (pronounced “Sigh Pie”) is a collection of mathematical algorithms and functions built work with Python language.

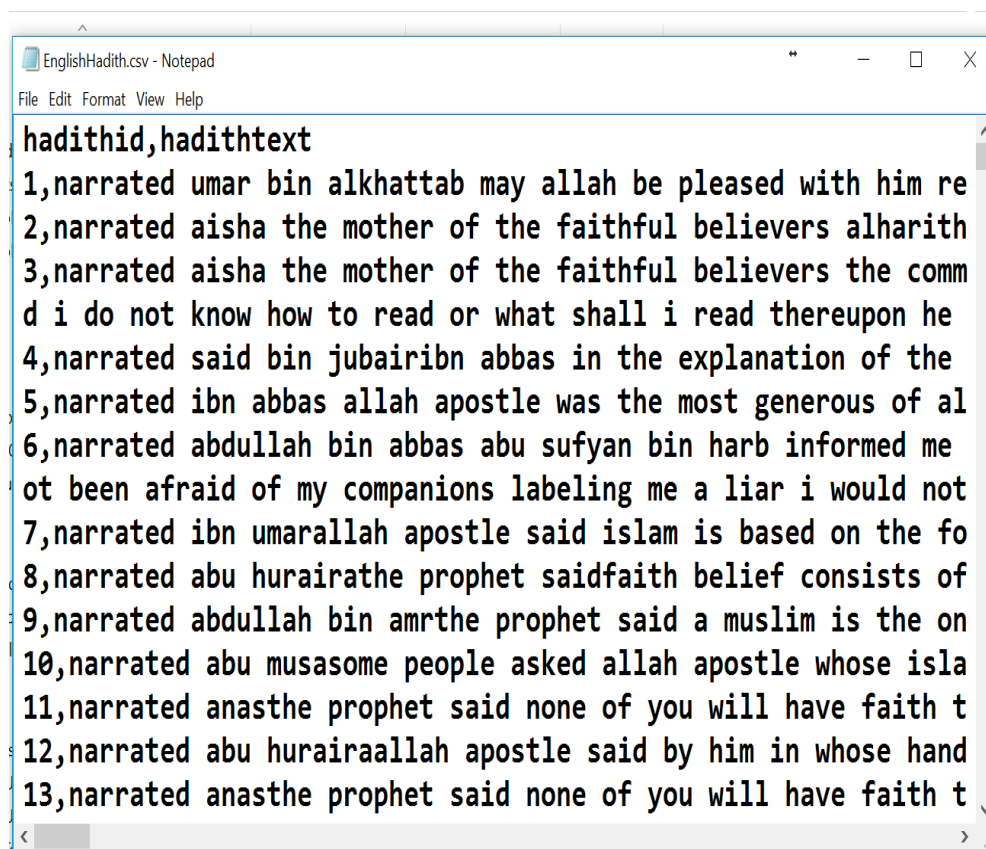


Figure 3.21 :The English documents in the csv file


```
1.variables filedata as file,line as string,  
   newlist as list  
2.read method:  
   for line in filedata  
       add line to newlist  
3.feed method:  
   import the TF-IDF  
   create object from the TF-IDF class  
   feed the newlist to the object
```

Figure 3.22 :A proposed feeding algorithm

3.2.3.2 Arabic Corpus

Convert the collection of the Ahadith from the (arabichadith.csv) shown in Figure 3.23 to the list of documents to do that a proposed feeding algorithm is used for generate the Arabic corpus and feed the corpus into TfidfVectorizer as shown in Figure 3.22. The python programming is used to implement these algorithm. After run the code only 2030 documents had been generated. Therefore the TfidfVectorizer will compute the TF-IDF weight for each token in the entire corpus. Consequently a sparse matrix of 2030x12851 of type '<class 'numpy.float64'>' with 66761 stored elements in Compressed Sparse Row format. So that mean will end up with only 12,853 distinct terms from the Arabic Hadith text out of 66,808,1 terms.

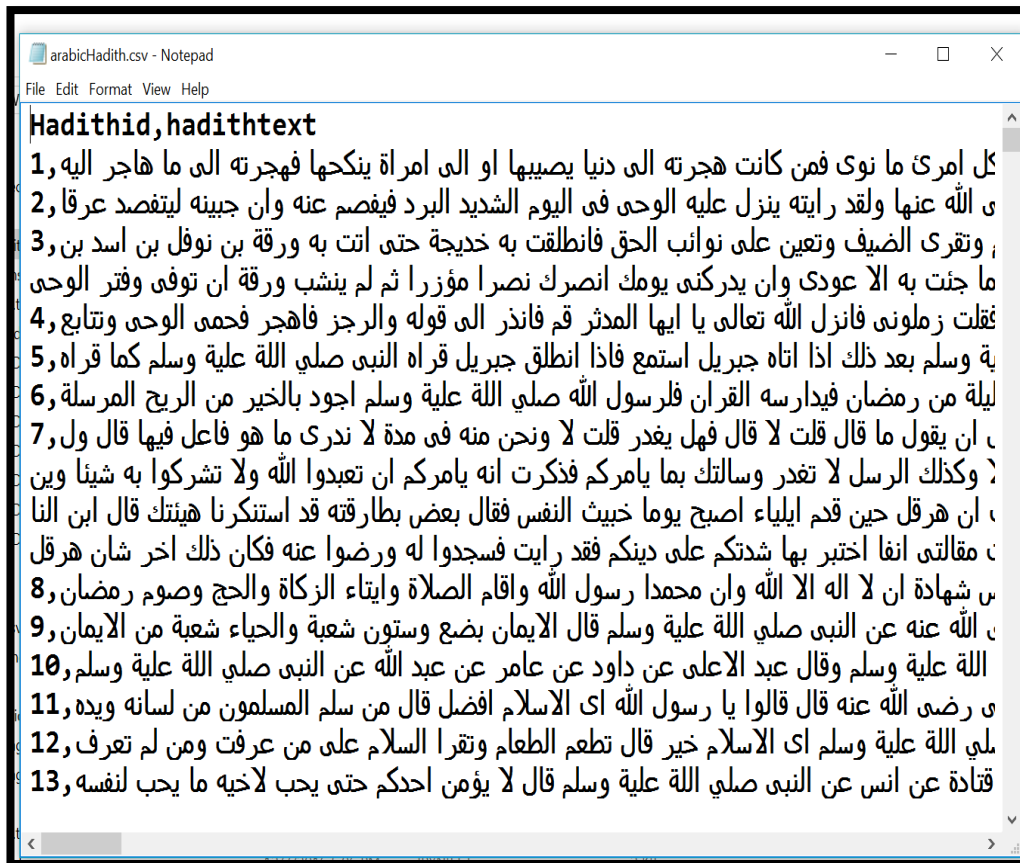


Figure 3.23 :The Arabic documents in the csv file

3.2.3.3 French Corpus

Convert the collection of the Ahadith from the (Frenchhadith.csv) shown in Figure 3.24 to the list of documents to do that a proposed feeding algorithm is used for generate the Arabic corpus and feed the corpus into TfidfVectorizer as shown in Figure 3.22. The python programming is used to implement these algorithm. After run the code only 2030 documents had been generated. Therefore the TfidfVectorizer will compute the TF-IDF weight for each token in the entire corpus. Consequently a sparse matrix of 2030x8279 of type '<class 'numpy.float64'>' with 43786 stored elements in Compressed Sparse Row format. So that mean will end up with only 8,267 distinct terms from the Arabic Hadith text out of 48,943 terms.

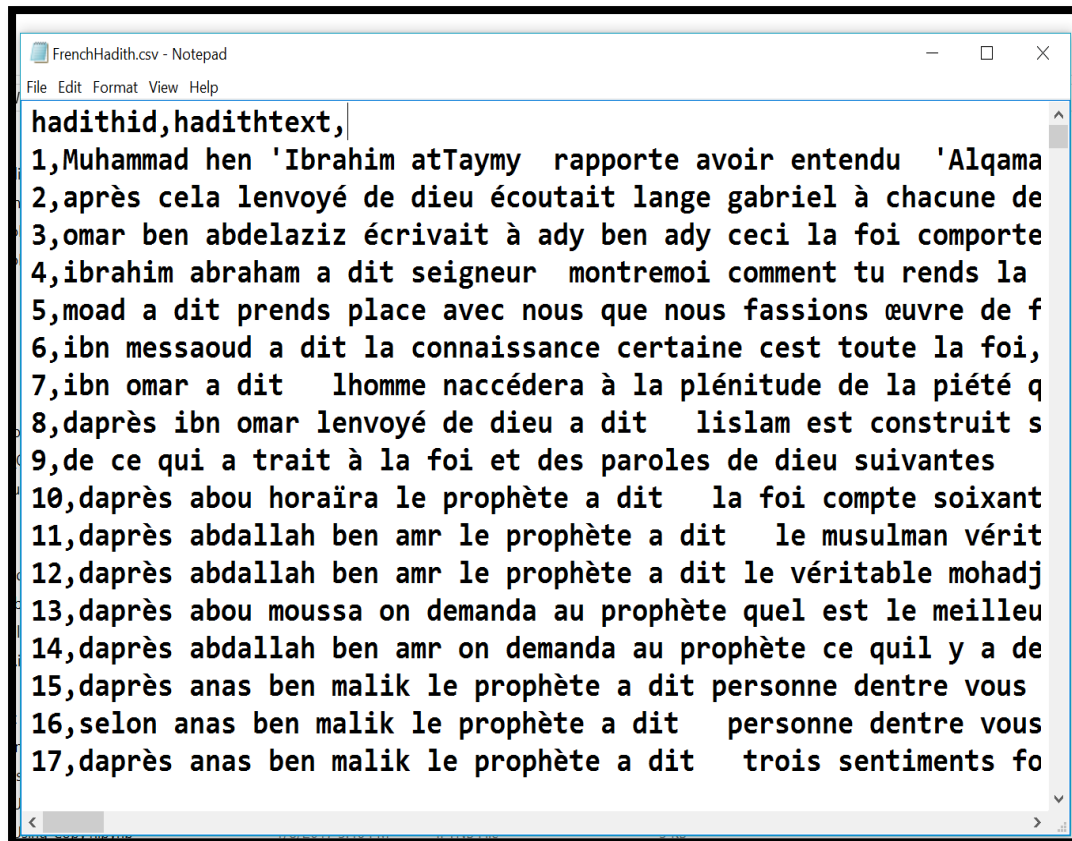


Figure 3.24 :The French documents in the csv file

3.2.3.4 Russian Corpus

Convert the collection of the Ahadith from the (Russianhadith.csv) shown in Figure 3.25 to the list of documents to do that a proposed feeding algorithm is used for generate the Arabic corpus and feed the corpus into TfidfVectorizer as shown in Figure 3.22. The python programming is used to implement these algorithm. After run the code only 2030 documents had been generated. Therefore the TfidfVectorizer will compute the TF-IDF weight for each token in the entire corpus. Consequently a sparse matrix of 2030x11702 of type '<class 'numpy.float64'>' with 39486 stored elements in Compressed Sparse Row format. So that mean will end up with only 11,355 distinct terms from the Arabic Hadith text out of 38,756,4 terms.

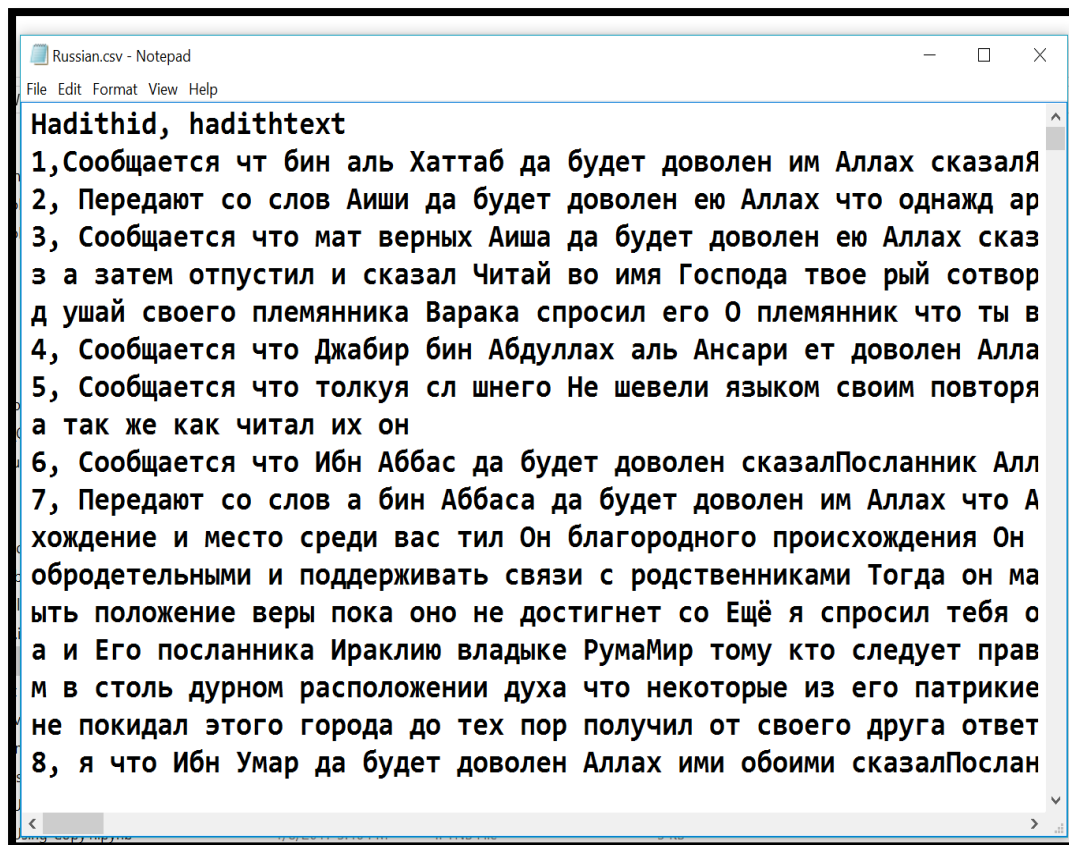


Figure 3.25: The Russian documents in the csv file

3.2.4 Phase Four: MHC in Sketch Engine

Sample data from the Multilingual Hadith Corpus (MHC) has been uploaded into SketchEngine Figure 3.27 in preparation for further inquiry and analysis. This involved several steps as follows. Because our corpus is multilingual we cannot create one corpus including multiple languages but in SketchEngine there an option for this type of corpus called "parallel corpora". Parallel corpora in SketchEngine work as independent corpora; to mark that two (or more) corpora are aligned (work as a parallel corpus), a special structure is required in each of the corpora. The name of this structure is defined as ALIGNSTRUCT corpus attributes and the default is aligned. The alignment has to be strictly 1:1 and the name of the aligning structure are fixed to align (Kilgarriff, 2014)(Van Zaanen, 2004).

The Arabic language has a difficult structure and is different from English, French, and Russian. Most text analytics tools assume English text, and may not adapt readily to Arabic, Russian and French. One challenge is that Arabic text is written in Arabic alphabet and script and displayed right-to-left, but XML and another markup in the file and the English and French parallel texts are normally written in Latin alphabet and script and displayed left-to-right, and the Russian text is written in Russian Cyrillic alphabet and script, also left-to-right; so compatible storage and display poses extra challenges. Also, Hadith texts are originally in printed books, and online sources are mainly scanned PDF, so we need to use OCR to extract the text data; but OCR systems are best for English and give more errors when applied to Arabic, French and Russian. Also, for later analysis of our corpus, we can use existing tools like Part-of-Speech tagger for English, but PoS-taggers for Arabic-Russian and French are not as accurate.

Corpora: Recent **My own** Featured Parallel All













Language	Name	Words	
Arabic	HadithArabic	3,524	  
English	HadithEnglish	4,323	  
French	HadithFrench	5,082	  
Russian	HadithRussian	5,301	  

Figure 3.27: Snapshot of parallel Corpora in the Sketch Engine

3.2.4.1 Steps for building parallel corpora

1. Create four different files for the four languages we had in our corpus; we had kept one file for Arabic Hadith, one for English Hadith, one for French Hadith and one for the Russian Hadith.
2. The data must be markup with ALIGNSTRUCT, using the tag <align>, each statement must be included between open and closed align <align>...</align>. For a sample of the aligned structure for each corpus see Figure 3.28 for Arabic corpus, Figure 3.29 for English corpus, Figure 3.30 for French corpus and Figure 3.31 for Russian corpus.
3. The four corpus files must have the same number of align tags before we uploaded them into SketchEngine; then they can work as parallel corpora.
4. Create an account in the SketchEngine website; after that login to that account; then only you can start uploading and searching corpora.
5. The four corpora were uploaded each in a separate corpus file, to create the HadithArabic corpus for the Arabic text, HadithEnglish for the English text, HadithFrench for the French text and the HadithRussian for the Russian text.
6. After uploading the files ,it was necessary to go to the configure corpus tab for each corpus in the sidebar and use the ALIGNED field to select the corpora which we want to align with. For our case each one corpus must align with the other three; if we start with HadithArabic corpus, then we go to the ALIGNED field and select the HadithEnglish, HadithFrench, and HadithRussian corpora to align with and save. We had to repeat the same process for the other three corpora; in the end, we have each corpus aligned with the other three.
7. The corresponding aligns segments in data from all corpora will be automatically connected .
8. Next, it is necessary to recompile four corpora HadithArabic, HadithEnglish, HadithFrench, and HadithRussian.
9. After the parallel corpora are connected, we can use features like generate a simple query using the concordance function, and the other features available in SketchEngine.

<align>

عن أبي عبد الرحمن عبد الله بن عمر بن الخطاب رضي الله عنهما
قال "سمعت رسول الله صلى الله عليه وسلم يقول
بني الإسلام على خمس شهادة أن لا إله إلا الله وأن محمدا رسول الله
 وإقام الصلاة وإيتاء الزكاة وحج البيت، وصوم رمضان
</align>

Figure 3.28: Example of ALIGNSTRUCT for the HadithArabic Corpus

<align>

On the authority of Tamim Al-Dari
that the prophet said: "Religion is
sincerity". We said: "To whom?" He said:
"To Allah and His Book, and His messenger,
and to the leaders of the Muslims and
their common folk". narrated by Muslim
</align>

Figure 3.29: Example of ALIGNSTRUCT for the HadithEnglish Corpus


```
<align>
Selon la Mère des Croyants, Oumm Abdallâh
Aïcha (que Dieu soit satisfait d'elle),
l'Envoyé de Dieu, alla Allah u alihi WA sallam ,
(à lui, bénédiction et salut) a dit:« Quiconque
apporte à notre religion une nouveauté qui n'en
provient pas, celui-là est à repousser ».
</align>
```

Figure 3.30 : Example of ALIGNSTRUCT for the HadithFrench Corpus

```
<align>
По свидетельству Матери Правоверных 1,
(1 Титул жен Пророка.) Умм Абдуллах Аиши (да будет
Аллах милостив к ней), которая сказала: Посланник
Аллаха (да благословит его Аллах и да ниспошлет ему
мир)сказал: Если кто-нибудь введет новшество в наше
дело, где ему не место, оно будет отвергнуто.
</align>
```

Figure 3.31 : Example of ALIGNSTRUCT for the HadithRussian Corpus

3.2.5 Phase Five:Web application Design

The Hadith contains many of the concepts .On the way to help Muslims to find and understand these concepts we decided to design Hadith Corpus Search Engine(HCSE), which is a search engine for MHC .The purpose of Hadith Searches to enhance precision and accuracy when searching for specific concepts as well as abstract concepts in Hadith. Besides, it is the only tool that allows users to search by concept overall hierarchy of topics or abstract concept in the Hadith, using the imported knowledge from the book of Sahih Bukhari as one of the most important specialized books in Hadith and the correct one based Muslim scholars (Bader Alden,2000) .

The Hadith Corpus Search Engine (HCSE) is a tool for the study of Hadith in different languages on the web. The corpus below were built by gathering from the web and extracting textual content from web pages. Searches can be performed to find words or phrases. Results are given as concordance lines in HTML format. Using of The MVC (Model View Controller)new architecture has allowed us to enhance the search performance.

3.2.5.1 Software Requirements

To build the web application we use one of the most recent development web technologies called Flask is a framework provides a solid core with the basic services, with the ability to work with python and natural language processing tool (nltk) for these reason Flask was consider as our best option to build our website (Grinberg,2014) .Beside, we use Microsoft Visual Studio 2012 is an integrated development environment (IDE) from Microsoft. It is used to develop our web application to contain the search engine Figure 3.32 show the interface of the website which contain the entire corpus of Hadith.for the dataset we are using the Multilingual Hadith Corpus which we already descried it in phase one process.

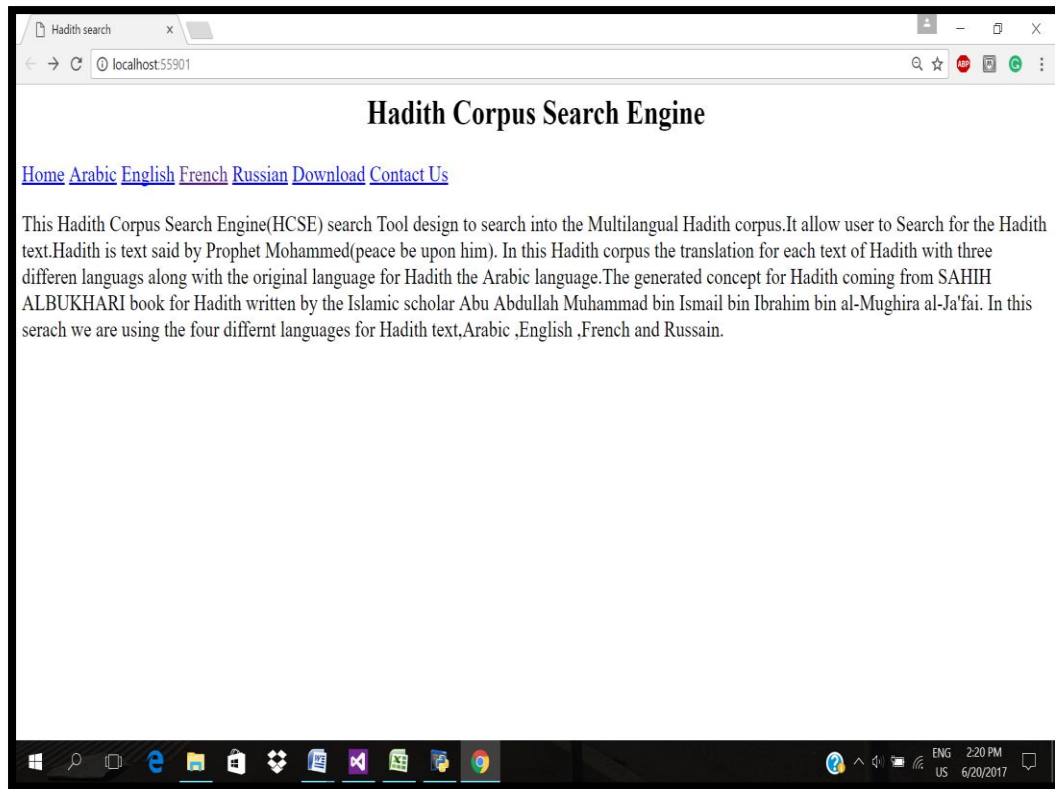


Figure 3.32: The search Tool interface

3.2.5.2 Search Algorithm

Search based on a user query(sometimes called ad hoc search because the range of possible queries is huge and not prespecified) is not the only text-based task that is studied in information retrieval (Yang,2010).A proposed search Algorithm has been designed see Figure 3.33 to generate the search engine using the benefit of the TF-IDF ,coefficient similarity and data mining techniques like stop words ,tokenization and stemming in python and .To implement our algorithm we use Visual studio 2012 ,Flask ,Python and nltk

1. Accept query “q” from user
2. System call remove stop words function
3. System call tokenization function
4. System call Tfidfvectorizer to convert q to vector “qv”
5. For each term “t” in q :
 - 5.1. If t in inverted Index list:
 - 5.2. Return posting list “pl”
6. Find similarity method
 - 6.1. For each document “d” in pl:
 - 6.1.1. Calculate Cosine simillarity (qv,d)
 - 6.1.2. New list have all relevant documents with CS
7. Returen maximum 20 of CS from Newlist to the user

Figure 3.33 : A propsed search Algorithm

3.2.5.3 Search by Arabic concept

Search by keyword will return the specific word in our case the specific Ahadith related to the specific concept. but when we search using the Arabic word we get problem of getting zero results when we know we have data like when we search for the concept "الإيمان" will get no result and that because the concept word in the database is written "الايمان" with no (tashkel"تشكيل") so the word will not match, and to solve this problem we generate a python function to take the Arabic word replace each ("أ", "إ", "آ") by the character ("ا") and each("ة")by("ة") see Figure 3.34.

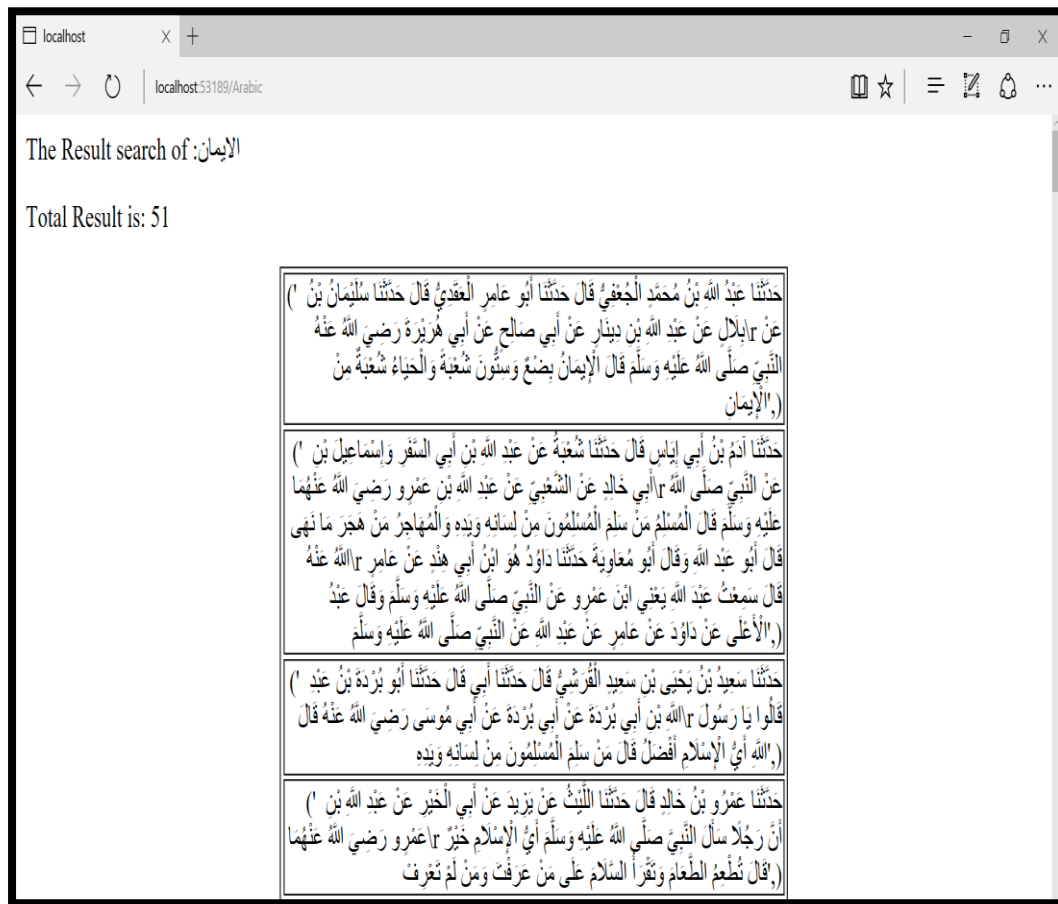


Figure 3.34: Snapshot for Arabic search concept “الإيمان”

3.2.5.4 Search by English Concept

English text is written in Latin alphabet and the display of English text from left to right most of the data analysis tools ,data mining method all support English text for example there is no problem if the user write in upper or lower case if user search for the concept “knowledge” or “KNOWLEDGE” the result will be the same in our search see Figure 3.35.

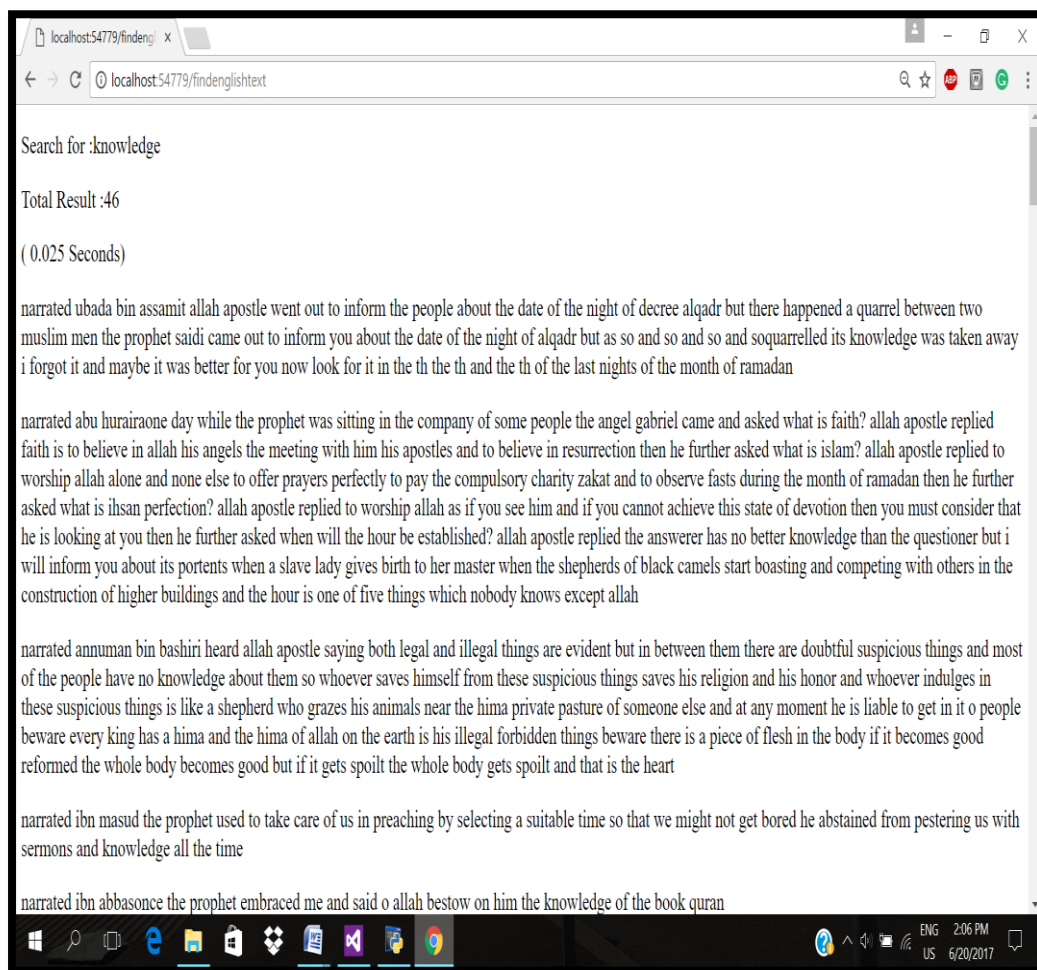


Figure 3.35: Snapshot for English search concept “knowledge”

3.2.5.5 Search by Russian Concept

Russian text is written by Russian Cyrillic alphabet and script, writing is left-to-right. For the Cyrillic alphabet, we need compatible storage and display poses extra challenges .for example to save and display the Russian character correctly in text file the text must be save as Unicode (UTF-8) and also when retrieve the text same process will be use python have package called “codecs” use to make the Russian text display correctly. For example if the user search for the concept “омовения” meaning (“الوضوء”)

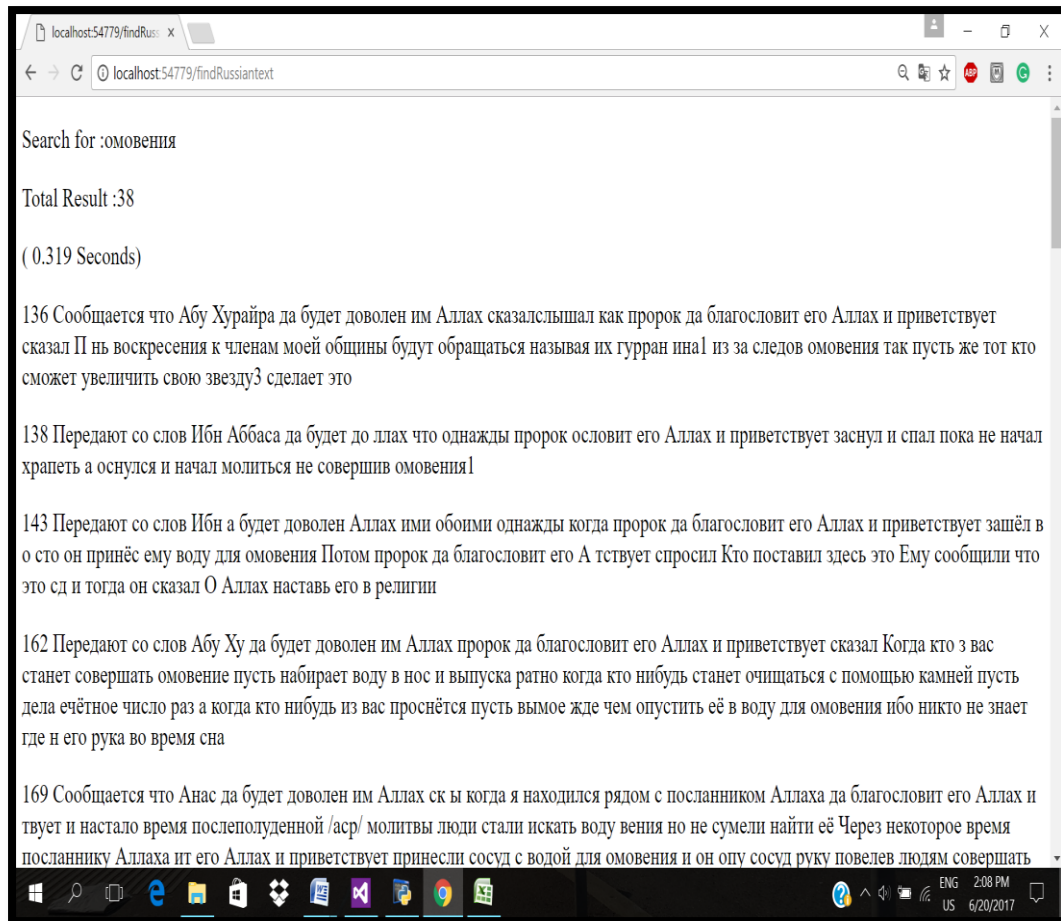


Figure 3.36 : Snapshot for Russian search concept “омовения”

3.2.5.6 Search by French keyword

English and French parallel texts are normally written in Latin alphabet and displayed left-to-right. So the search using the French concept is working fine as the English concept. For example if user search for the concept “foi” or “FOI” the result will be the same as in Figure 3.37.

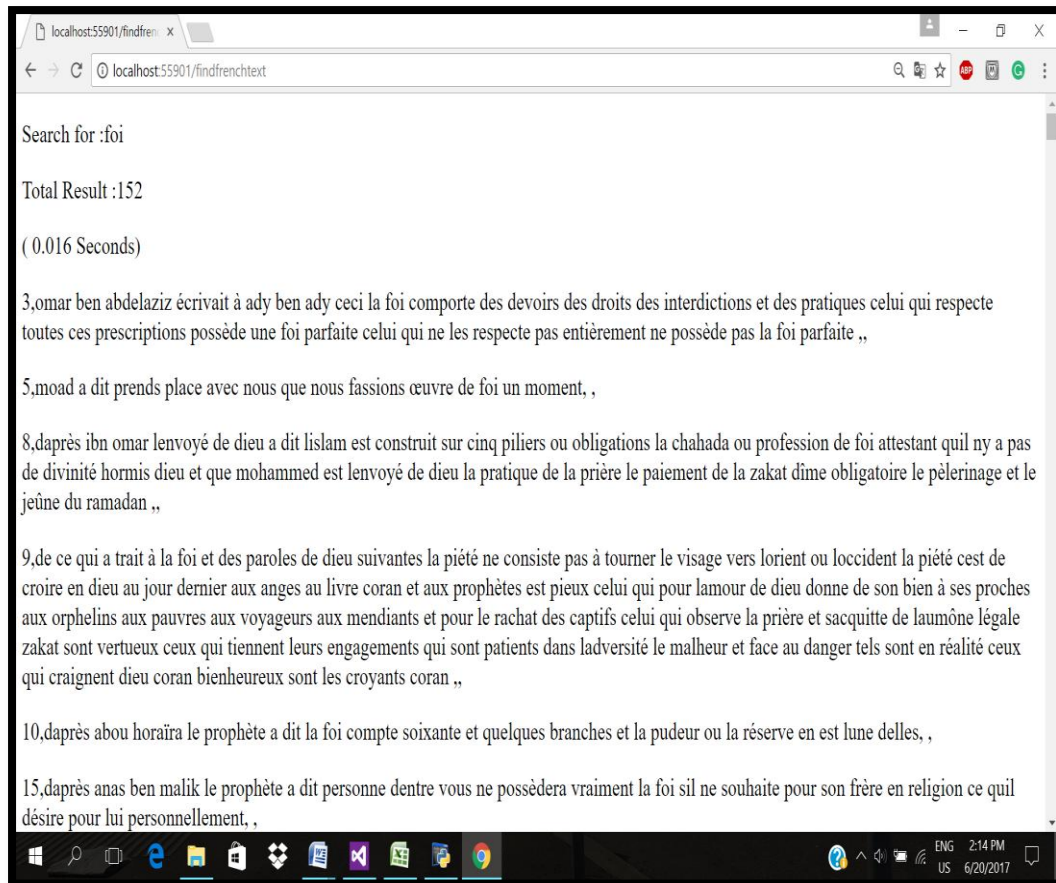


Figure 3.37: Snapshot for French search concept “foi”

3.2.5.7 Download

In download page allow user to download the MHC text in different format plain text, xml file format and HTML format. Beside user can select the language of the file and if the user wants to download the text in one file or in separated file for each text see Figure 3.38.

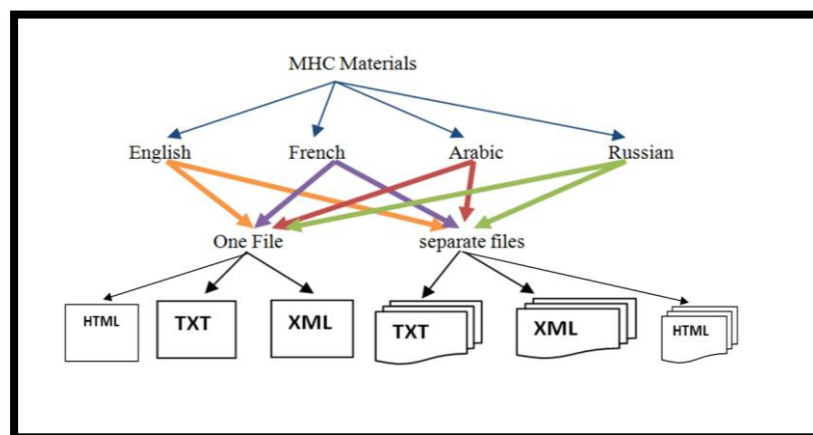


Figure 3.38: The different options of download files in the MHC

3.2.5.8 Contact Us

In contact us page we allow user to contact us regarding any matter related to our corpus or if the user have any question or want to send any message for us we asked the user to write their Email so we can get back to them see Figure 3.39.

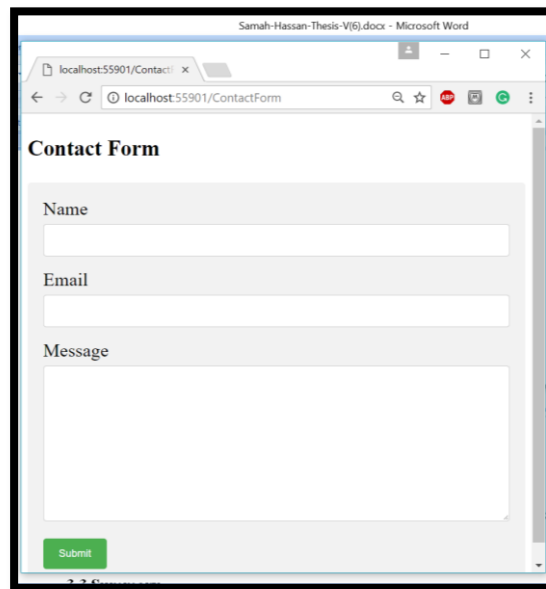
A screenshot of a web browser window displaying a 'Contact Form'. The browser's address bar shows 'localhost:55901/ContactForm'. The form itself is titled 'Contact Form' and contains three input fields: 'Name', 'Email', and 'Message'. The 'Message' field is a larger text area. At the bottom of the form is a green 'Submit' button. The browser window also shows a tab for 'localhost:55901/Contact' and a document titled 'Samah-Hassan-Thesis-V16.docx - Microsoft Word' in the background.

Figure 3.39: Snapshot of the contact form

3.3 Summary

In this investigation, the aim was to explain in full details what we are doing to accomplish our goal started by showing how we gathering our data , cleaning and organization of them make them can be feed directly to our TF-IDF and coefficient similarity merged algorithm which we design it to enhance the performance of the information retrieval system while user searching in Hadith text. Finally we show how we build a web application to hold the entire MHC and search engine.

CHAPTER IV

RESULTS AND DISCUSSION

4.1 Introduction

There is persuasive evidence that the Hadith plays an important role in regulating Muslim life (Al Imam, 2000). There has been a lot of work done in the creation of Arabic corpora, with many of them focusing on the Quran (Dukes et al, 2010). This is very beneficial, with many of the Quranic corpus resources being excellent, but mentions within these corpus resources of the Hadith, the words of Prophet Mohammed (Peace be upon him), is rare. The Hadith for Muslims is second in importance only to the Quran. In the Islamic Rules (Shree Al-Islamia), the Hadith is considered the second source of religious knowledge for Muslims, as in the Hadith you will find teachings on all areas of life of Muslims mentioned in the prophet Mohammed's (Peace be upon him) words. The Hadith guides on how to be good a Muslim: the prophet Mohammed (Peace be upon him) explains everything necessary for Muslims to live their life: how to eat, how to drink, how to sleep, how to deal with other people, how to pray, how to obey Allah, and how to do everything else be it minor or major. Therefore, a multilingual Hadith corpus would be useful for Muslims all around the world as it will allow them to know what each word is, what it means, and what it teaches us about our religion.

4.2 Arabic Stemming

After reading the Arabic text before feeding to the **TFIDFvectorize** consider the stemming process for the entire Arabic text in the research we use **snowballstemmer** stemmer from nltk because it contain Arabic stemmer use with python program. After data stemming only 6,882 terms are left and can be

considered as distinct terms. Table 4.1 shows the comparison between the number of terms before and after data stemming. Figure 4.1 illustrate the change in the number of terms .

Table 4.1: Number of the terms for Arabic text before and after stemming process

Arabic Text	Number of distinct terms
Before	12,853
After	6,882
Percentage in Reduction	46.46 %

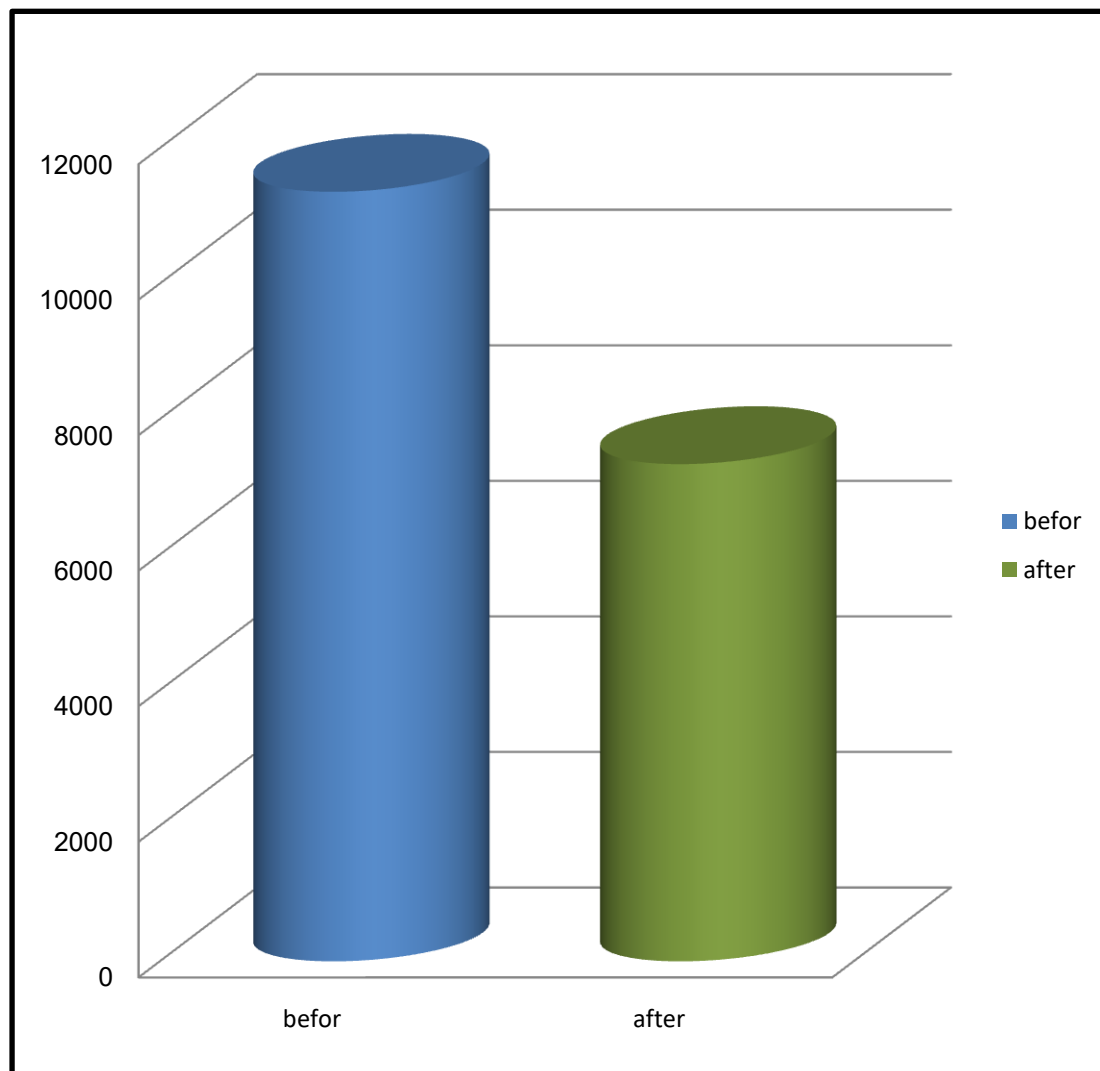


Figure 4.1: The change of number of terms for Arabic before and after stemming process

4.3 English Stemming

After reading the Arabic text before feeding to the TFIDFvectorize consider the stemming process for the entire English corpus . After data stemming only 5,313 terms are left and can be considered as distinct terms. Table 4.2 shows the comparison between the number of terms before and after data stemming. Figure 4.2 illustrate the change in the number of terms .

Table 4.2: Number of the terms for English text before and after stemming process

English Text	Number of distinct terms
Before	7196
After	5313
Percentage in Reduction	26.17 %

4.4 French Stemming

After reading the Arabic text before feeding to the TFIDFvectorize consider the stemming process for the entire French corpus . After data stemming only 5,181 terms are left and can be considered as distinct terms. Table 4.3 shows the comparison between the number of terms before and after data stemming. Figure 4.3 illustrate the change in the number of terms .

4.5 Russian Stemming

After reading the Arabic text before feeding to the TFIDFvectorize consider the stemming process for the entire Russian corpus. After data stemming only 7,337 terms are left and can be considered as distinct terms. Table 4.4 shows the comparison between the number of terms before and after data stemming. Figure 4.4 illustrate the change in the number of terms.

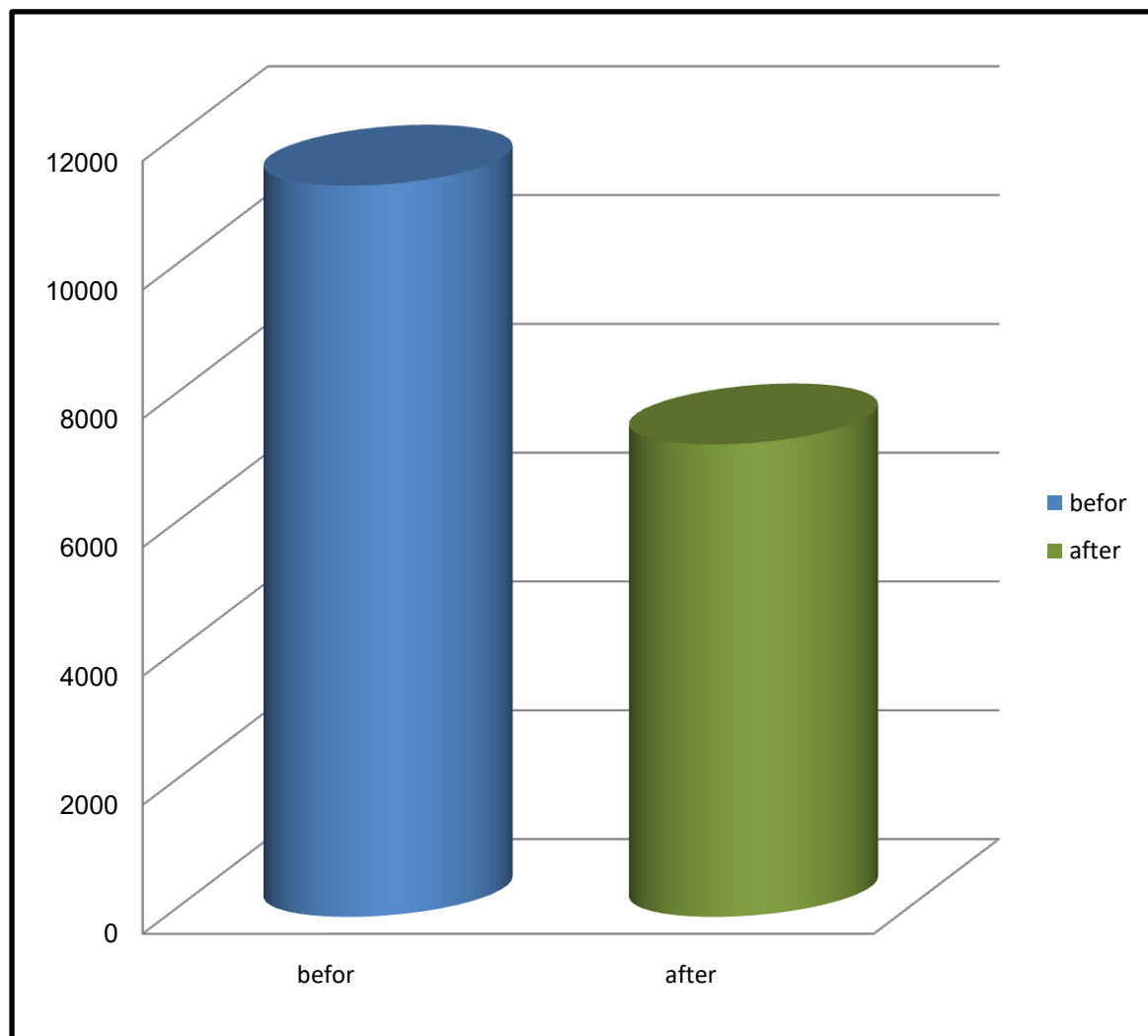


Figure 4.2: The change of number of terms for English before and after stemming process

Table 4.3 : Number of the terms for French text before and after stemming process

French Text	Number of distinct terms
Before	8267
After	5181
Percentage in Reduction	37.33%

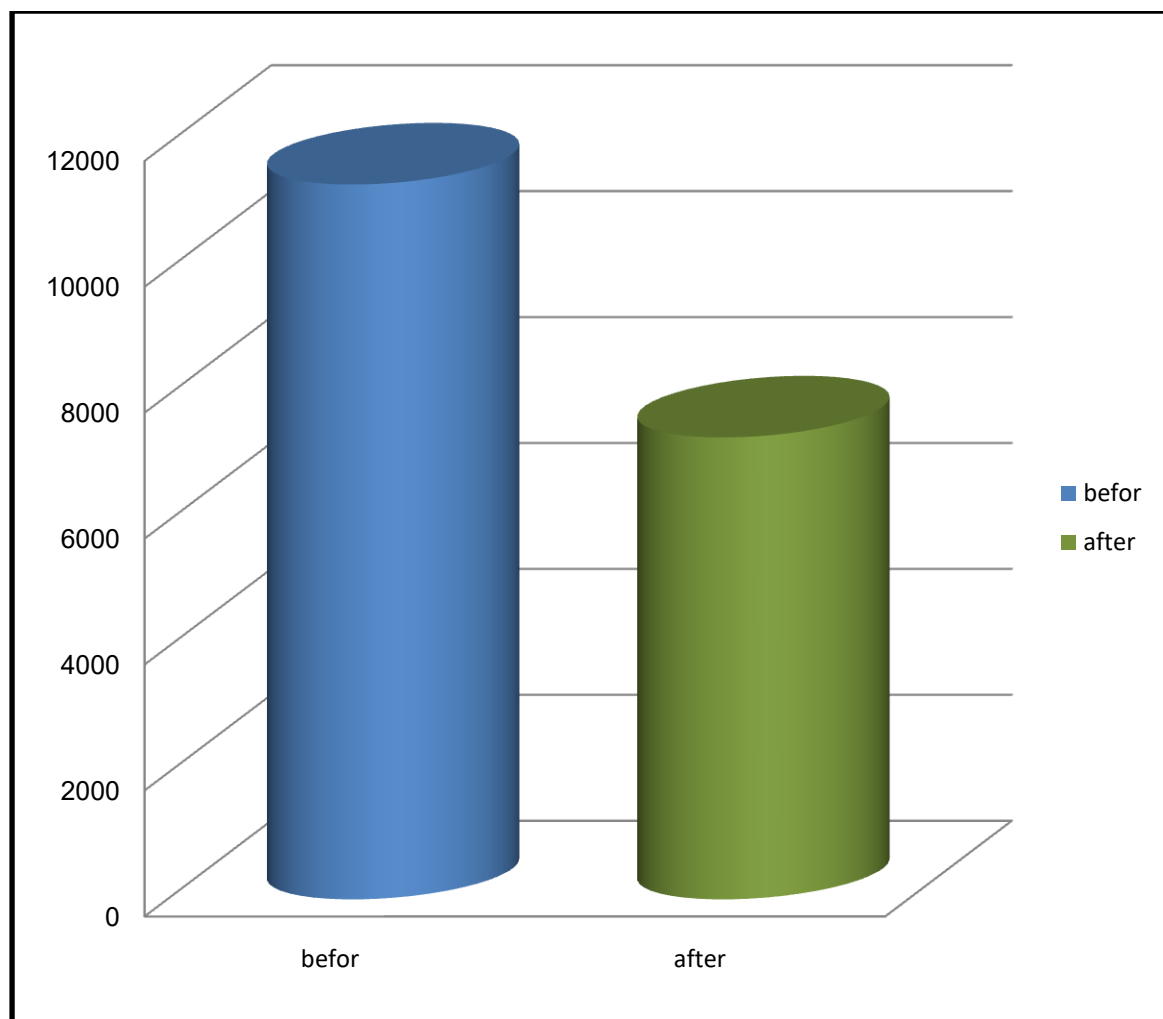


Figure 4.3: The change of number of terms for French before and after stemming process

Table 4.4: Number of the terms for Russian text before and after stemming process

Russian Text	Number of distinct terms
Before	11,355
After	7,337
Percentage in Reduction	35,39 %

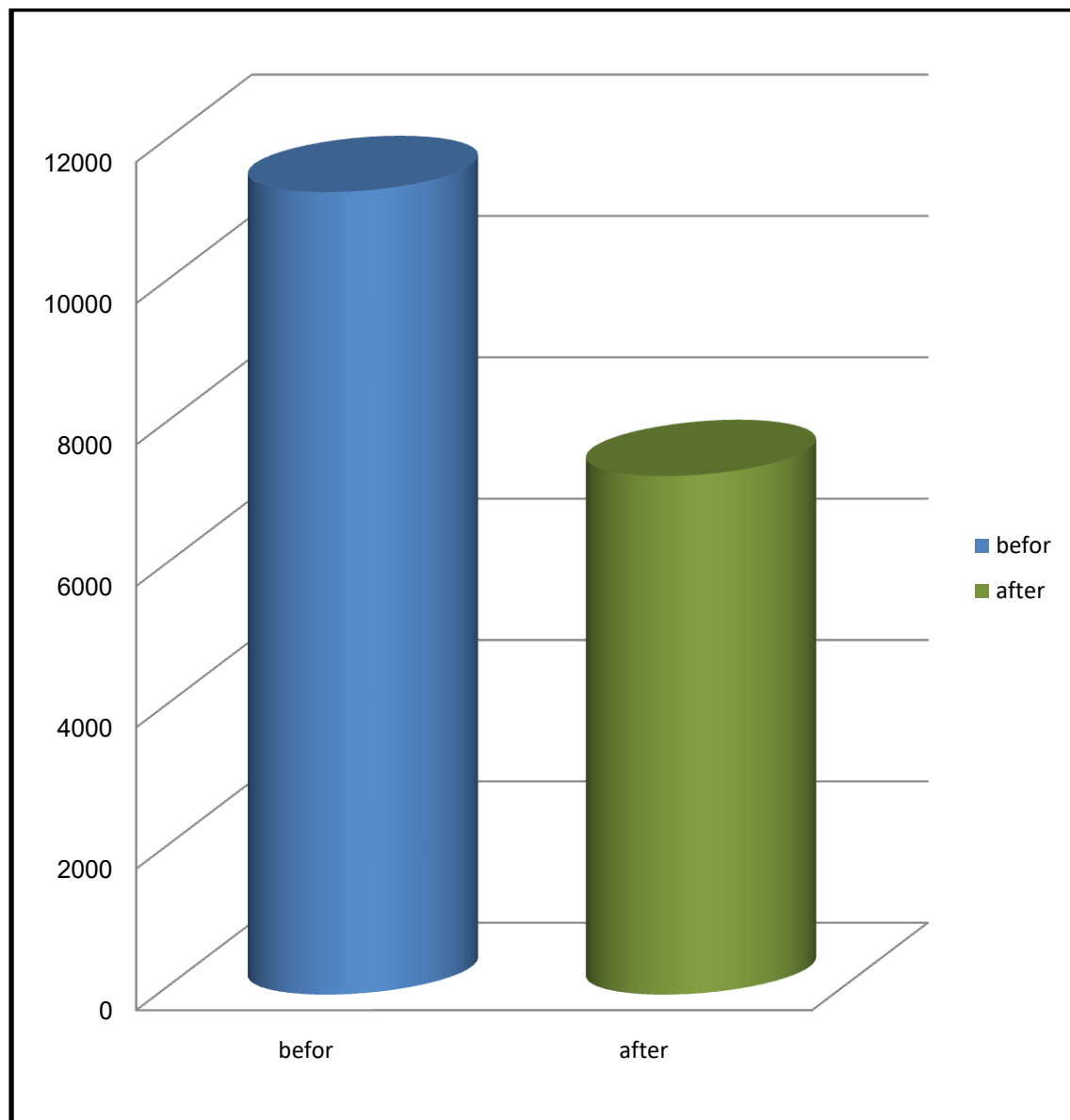


Figure 4.4: The change of number of terms for Russian before and after stemming process

4.6 Information retrieval system evaluation

Recent developments in IR have highlighted the need for measure the efficiency of IR system data by the standard way there are three major aspects must be taken into account (Christopher et al,2008):

- A document collection
- A test group of information needs, expressible as queries

- A set of correct relevance judgments that relevant for each query

4.7 Quantitative Evaluation

There is always an urgent need to improve the performance of Information retrieval system and seek to make it more efficient and effective, and to assess this improvements researcher uses the measurement recall and precision as an indicator of improved performance if the rank rate high for both recall see Equation (4.1) and precision see Equation (4.2) mean a clear indication of the efficiency and effectiveness of the system ” The best know IR measures are precision and recall ” (Bar-Ilan, 2002) ,so that what will we use to assess our IR system for Hadith (Harrage et al,2011). Besides, evaluation is very critical and tedious task in information retrieval system (Zuva,Zuva,2012).That is why to evaluate our search engine we decide to select a set contain 10 concepts from our corpus called as our gold standard to evaluate and calculate the precision and recall for our system for each languages started with the English, Arabic, French and Russian. In Table 4.5 we declare our gold standard for the four languages (Manning et al,2008).

Table 4.5: The gold standard for our Search

Russian	French	English	Arabic
переселится	expatriation	emigration	الهجرة
благословит	intention	Intention	النيات
дела	Actes	Deed	الاعمال
откровение	r��velation	Revelation	الوحي
мирского	monde	Worldly	الدنيا
он хотел жен	epouser	Marry	ينكحها
веры	foi	Faith	الايمان
наука	Connaissance	Knowledge	العلم
лицемерие	hypocrisie	Hypocrisy	النفاق
омовения	Ablutions	Ablution	الوضوء

$$\text{Recall (R)} = \frac{\text{Number of correct answer given by the system}}{\text{total number of the correct answers exist in the text}} \quad (4.1)$$

$$\text{Precision(P)} = \frac{\text{Number of the correct answer given by the system}}{\text{Total number of answers given by the system}} \quad (4.2)$$

The formulas for the precision and the recall values were obtained from (Martin, Jurafsky, 2000). The F-measure formula obtained from (Christopher et al, 2008) see Equation(4.3).

$$\text{F – Measure} = 2 \frac{R*P}{R+P} \quad (4.3)$$

Table4.6 : The correct relevant documents for
the selected English gold standard

Concept	Ahadith
emigration	[0, 299, 348, 507, 804, 805, 1476]
Intention	[0, 640, 788, 924, 1220, 1260, 1400, 1427, 1476, 1487, 1572, 1592, 1724]
Deeds	[0, 69, 195, 196, 281, 299, 348, 373, 446, 477, 507, 510, 564, 578, 618, 641, 661, 666, 676, 717, 730, 770, 794, 806, 818, 826, 861, 882, 924, 933, 974, 987, 1010, 1011, 1020, 1028, 1079, 1091, 1117, 1134, 1136, 1139, 1187, 1198, 1268, 1331, 1342, 1353, 1463, 1464, 1469, 1471, 1475, 1476, 1496, 1538, 1545, 1592, 1646, 1757, 1842, 1875, 1922, 1934, 2006, 2019]
Revelation	[48, 189, 431, 454, 1128, 1364]
world	[0, 196, 263, 299, 507, 593, 665, 769, 1476]
marry	[0, 12, 299, 507, 514, 519, 522, 524, 526, 1476]
faith	[112, 223, 276, 277, 334, 427, 445, 485, 497, 556, 605, 643, 676, 1000, 1011, 1087, 1088, 1112, 1154, 1190, 1265, 1287, 1294, 1298, 1320, 1365, 1398, 1431, 1442, 1808]
Knowledge	[82, 190, 224, 257, 272, 290, 426, 566, 567, 571, 587, 591, 619, 620, 622, 676, 697, 728, 791, 830, 858, 869, 880, 919, 930, 948, 959, 1013, 1030, 1037, 1103, 1109, 1420, 1431, 1453, 1631, 1709, 1753, 1764, 1775, 1786, 1820, 1964, 1986, 1990, 1993]
hypocrisy	[556, 669, 736, 1254]
Ablution	[59, 63, 66, 71, 111, 174, 175, 176, 335, 338, 360, 368, 379, 401, 412, 423, 457, 504, 508, 584, 585, 612, 623, 634, 645, 656, 668, 679, 690, 701, 712, 734, 823, 834, 845, 853, 856, 867, 872, 878, 901, 912, 923, 934, 945, 956, 978, 989, 1012, 1026, 1034, 1045, 1047, 1056, 1078, 1113, 1124, 1135, 1146, 1148, 1152, 1153, 1164, 1169, 1190, 1191, 1192, 1210, 1214, 1215, 1223, 1224, 1225, 1233, 1238, 1239, 1240, 1241, 1243, 1245, 1280, 1283, 1289, 1290, 1310, 1325, 1337, 1338, 1427, 1440, 1443, 1539, 1540, 1542, 1584, 1592, 1641, 1671, 1811, 1829, 1834, 1900, 1942, 1945]

Table 4.7: The calculation of precision & recall
for English gold standard

concept	Precision (%)	Recall (%)
emigration	100	30.4
intention	100	15.3
deeds	100	56.3
Revelation	100	85.7
worldly	81.8	20.9
marry	100	38.4
faith	100	100
Knowledge	100	67.6
hypocrisy	100	100
Ablution	12.5	7

Table 4.8: The precision & Recall & F-measure
for English gold standard

Arabic	Precision (%)	Recall (%)	F- measure
With stop words & diacritics & no stemming	0	0	0
Remove stop words & no stemming	0	0	0
Remove stop words & diacritics & no stemming	89.4	52	66
Remove stop words & diacritics & apply stemming	98.4	90	94.2

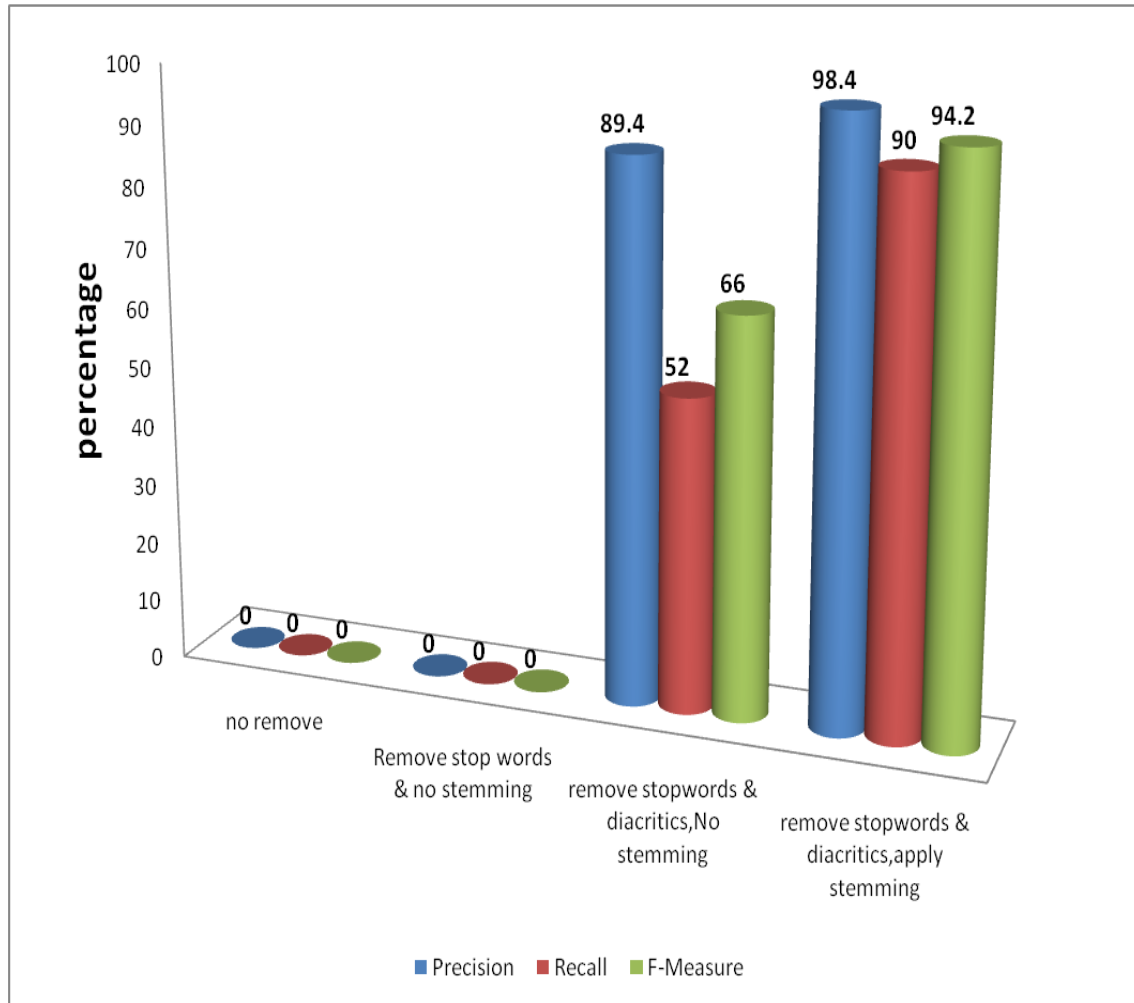


Figure 4.5: The improvement of precision & recall and F-measure for English gold standard

Table 4.9: The correct relevant documents for Arabic gold standard

Concept	Ahadith
الهجرة	[426,0]
النيات	[0]
الاعمال	[0,112,496,1108,1506]
الوحي	[514,773,833,1107,1249,1360,1888]
الدنيا	,887 ,754 ,738 ,737 ,616 ,448 ,439 ,380 ,240 ,239 ,166 ,103 ,82 ,6] 1297 ,960,1758,1699,1423,1297

النكاح	[0]
الايمان	,1683 ,1495 ,1474 ,1473 ,1472 ,1184 ,1151 ,862 ,776 ,665 , ,1867, 364,1899
العلم	[,1845 ,1818 ,1812 ,1801 ,1790 ,1683 ,623 ,459 ,277 ,244 2000 ,1938, 2]
النفاق	[776,1294]
الوضوء	[59, 63, 66, 71, 111, 174, 175, 176, 335, 338, 360, 368, 379, 401, 412, 423, 457, 504, 508, 584, 585, 612, 623, 634, 645, 656, 668, 679, 690, 701, 712, 734, 823, 834, 845, 853, 856, 867, 872, 878, 901, 912, 923, 934, 945, 956, 978, 989, 1012, 1026, 1034, 1045, 1047, 1056, 1078, 1113, 1124, 1135, 1146, 1148, 1152, 1153, 1164, 1169, 1190, 1191, 1192, 1210, 1214, 1215, 1223, 1224, 1225, 1233, 1238, 1239, 1240, 1241, 1243, 1245, 1280, 1283, 1289, 1290, 1310, 1325, 1337, 1338, 1427, 1440, 1443, 1539, 1540, 1542, 1584, 1592, 1641, 1671, 1811, 1829, 1834, 1900, 1942, 1945]

Table 4.10 : Result of precision

&recall for Arabic

Precision (%)	Recall(%)	Arabic
100	14	الهجرة
100	100	النيات
80	67	الاعمال
70	64	الوحى
53	40	الدنيا
50	50	ينكحها
30	45	الايمان
93	33	العلم
100	100	النفاق
71	80	الوضوء

Table 4.11: Result of precision & Recall
for Arabic gold standard

Arabic	Precision (%)	Recall (%)	F- measure
With stop words & diacritics & no stemming	0	0	0
Remove stop words & no stemming	0	0	0
Remove stop words & diacritics & no stemming	75	59	66
Remove stop words & diacritics & apply stemming	96.5	82	88.8

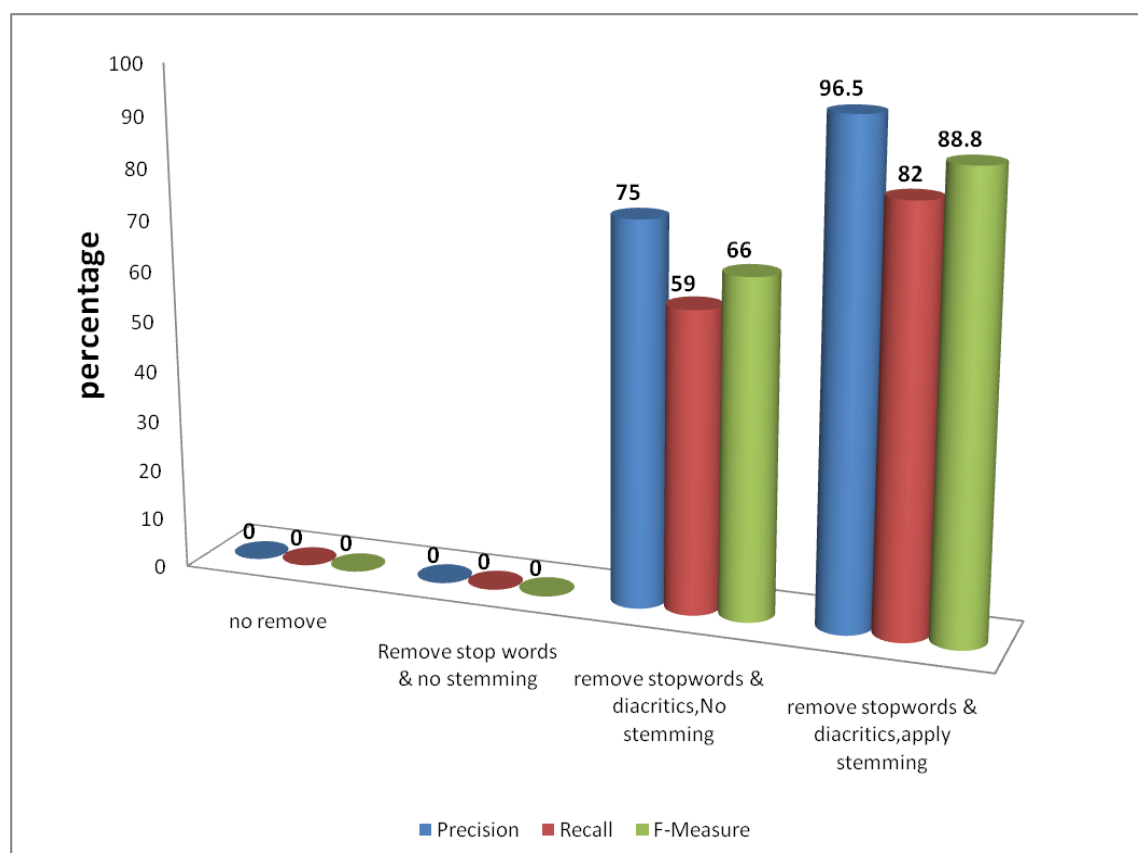


Figure 4.6: The improvement of precision & recall & F-Measure for Arabic gold standard

Table 4.12: The correct relevant documents

for the gold standard of French

Term	Ahadith
expatriation	[0]
intention	[0, 718, 1464, 1493, 1748]
Actes	[333, 336, 579, 584, 812]
révélation	[148, 344, 457, 673, 796, 1051, 1065, 1148, 1355, 1449, 1571, 1993]
monde	[0, 368, 448, 603, 796, 878, 1000, 1218, 1316, 1367, 1454, 1539, 2085]
epouser	[0]
foi	[1, 101, 373, 374, 375, 395, 490, 556, 574, 576, 646, 658, 659, 778, 884, 889, 1013, 1091, 1092, 1122, 1125, 1126, 1237, 1271, 1293, 1326, 1337, 1338, 1360, 1371, 1393, 1449, 1460, 1471, 1482, 1487, 1493, 1504, 1559, 1560, 1593, 1602, 1615, 1670, 1693, 1704, 1726, 1892, 1922, 1925, 1945, 2003, 2050]
Connaissance	[120, 246, 301, 1100, 1109, 1223, 1670, 1693, 1746, 1757, 1758, 1915, 2112]
hypocrisie	[889, 1460, 1615, 1626]
Ablutions	[55, 56, 263, 274, 284, 411, 413, 519, 534, 579, 601, 623, 645, 690, 756, 767, 779, 801, 812, 823, 834, 867, 945, 956, 997, 1012, 1100, 1113, 1135, 1179, 1201, 1246, 1262, 1268, 1278, 1279, 1286, 1317, 1323, 1327, 1363, 1410, 1464, 1569, 1573, 1643, 1666, 1724, 1726, 1793, 1975, 2087]

Table 4.13:Result of precision &recall for French

Term	Precision (%)	Recall(%)
expatriation	100	100
intention	50	90.9
Actes	100	31.2
révélation	46.6	82.3
monde	76.9	71.4
epouser	50	100
foi	62.5	65.3
Connaissance	100	90
hypocrisie	100	100
Ablutions	69	63.4

Table 4.14: The precision & Recall & F-Measure for French gold standard

Arabic	Precision (%)	Recall (%)	F-Measure
With stop words & diacritics & no stemming	0	0	0
Remove stop words & no stemming	0	0	0
Remove stop words & diacritics & no stemming	75.5	80	78
Remove stop words & diacritics & apply stemming	97.5	91.7	95

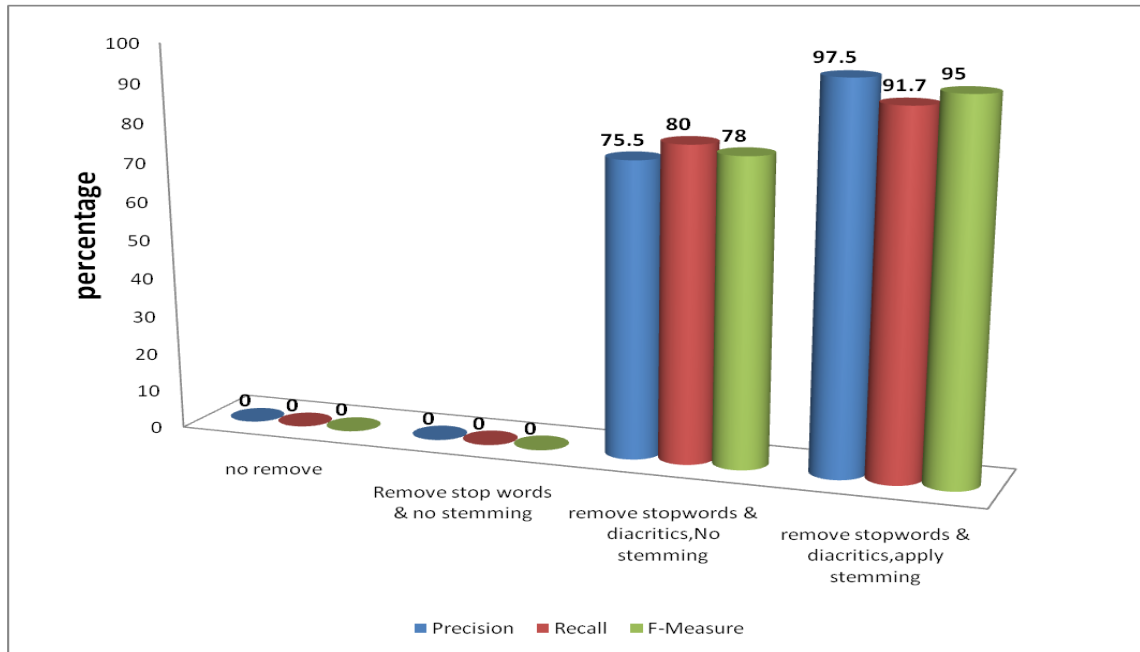


Figure 4.7: The improvement of precision & recall and F-Measure for French gold standard

Table 4.15: The correct relevant documents for the gold standard of Russian

Term	Ahadith
Переселится	[0]
Намерения	[0, 1667]
Дела	[0, 34, 152, 195, 294, 300, 671, 738, 1313, 1423, 1434, 1804, 1827, 1852]
Откровение	[500]
Мир	[137, 320, 324, 329, 728, 736, 977, 1018, 1086, 1499, 1781, 1786]
Брак	[0]
Веры	[649, 755, 1115, 1236, 1391, 1609, 1826]
Знание	[2, 240, 1706, 1740, 1772]
Лицемерие	[776, 1294]
Омовения	[150, 373, 390, 396, 467, 664, 671, 744, 883, 936, 1098, 1113, 1116, 1135, 1144, 1145, 1173, 1175, 1182, 1185, 1205, 1209, 1218, 1221, 1276, 1308, 1309, 1310, 1312, 1317, 1341, 1353, 1395, 1411, 1454, 1550, 1750]

Table 4.16: The calculation of precision & recall for Russian

Term	Precision (%)	Recall(%)
Переселится	100	25
Намерения	100	33.3
Дела	100	100
Откровение	100	14.2
Мир	77.5	100
Брак	100	100
Веры	45	25
Знание	100	100
Лицемерие	33.3	14.2
Омовения	30	22.5

Table 4.17: The precision & Recall for Russian gold standard

Arabic	Precision (%)	Recall (%)	F-Measure
With stop words & diacritics & no stemming	0	0	0
Remove stop words & no stemming	0	0	0
Remove stop words & diacritics & no stemming	79	54	64
Remove stop words & diacritics & apply stemming	98	91	94

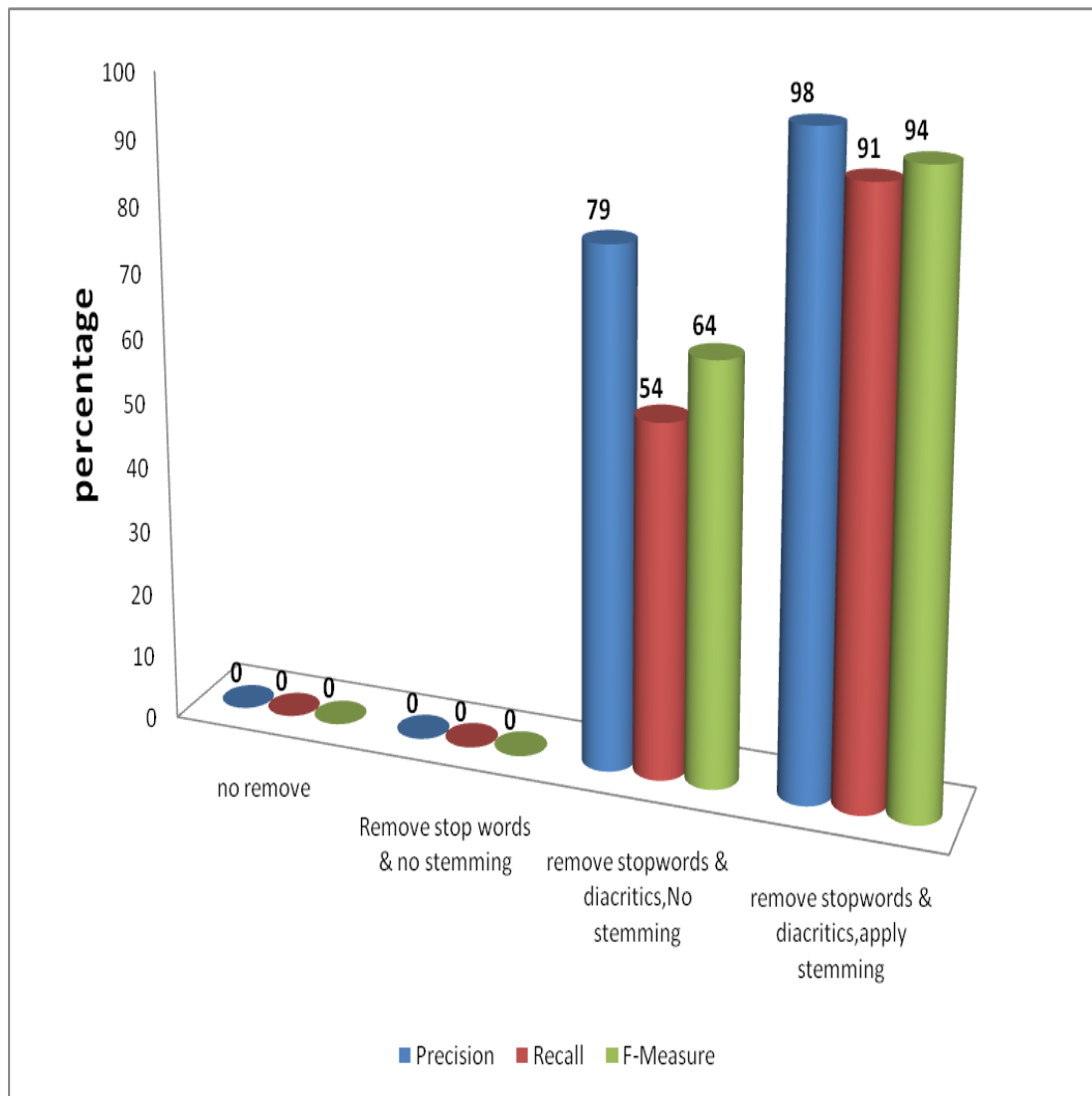


Figure 4.8: The improvement of precision & recall and F-Measure for Russian gold standard

Table 4.18: The differences in Recall, Precision and F-Measure between the languages

	Arabic %	English %	French %	Russian %
Precision	96.5	98.4	97.5	98
Recall	82	90	91.7	91
F-Measure	88.8	94.2	95	94

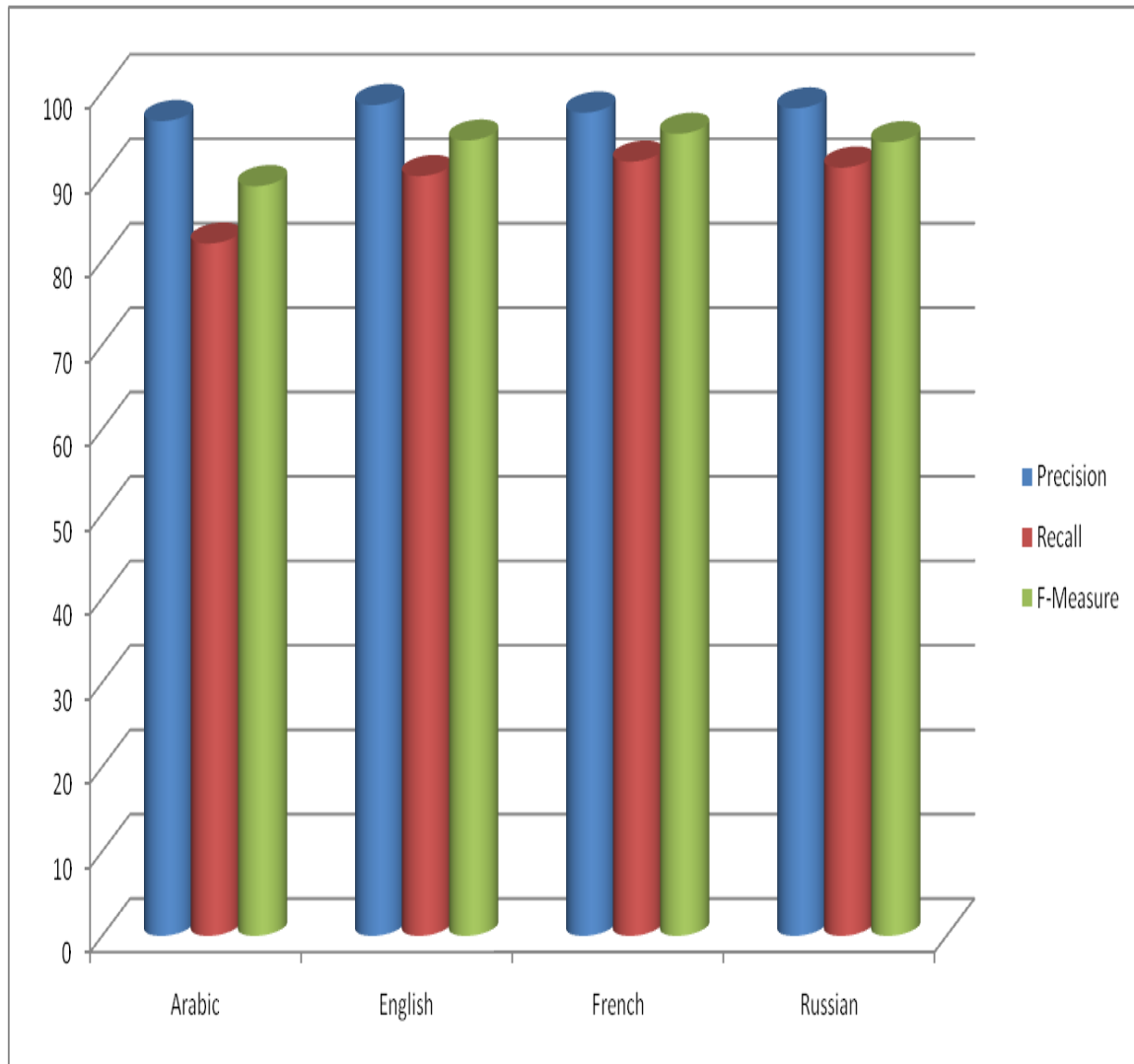


Figure 4.9: The differences in Recall, Precision and F-Measure between the languages

4.8 Qualitative Evaluation

Alqahtani, Atwell (2017) mention 13 evaluation criteria to compare the search tools for Quran so that we want to following their steps we had taken 8 criteria out of the 13 to compare the computational Hadith search tools with the HCSE (Table 4.19). On the other hand we select the search tools from the world wide web depend on the search languages, one site for each language because we did not find one search engine have the four languages in the same time. The evaluated Hadith search tools with the HCSE tool in the Table 4.20.

Table 4.19: The evaluation criteria for computational Hadith search tools

Comparison criteria	Possible values for criteria
1. Search techniques	A. keywords search B. Returns Hadith that contains any query words
2. Query Analyzer	A. stem of the query word B. spelling checker C. Non of A Or B
3. Database size	A. All Hadith in Sahih Bukhari B. Part of Hadith in Sahih Bukhari
4. Query types	A. one word B. Two words C. Sentence
5. Availability	Is this tool available to be used by others: A. Available B. Not available
6. Result ranking	How are the retrieved results ordered: A. Ranked result B. Not ranked
7. User categories	The target users for this application: A. Public B. Education C. Islamic scholar D. Linguistics scholar
8. language of input query	A. Arabic language B. Non-Arabic language

4.9 Issue to discuss

Our plan was to working in the same number of the concept in the four languages but we found the full concept in the Arabic language because it is the original language of Hadith and regarding the other languages may will find more the one word regarding one concept like “الايمن” we find in the translation as “faith” or “belief” in English language, and concept like “العلم” it may came as “science” or “Knowledge” in English. Because Hadith is very sensitive and important text we can not do the translation must be done by the expertise in Islamic for that reason we consider only the concept which we find it in the translated book by the expertise and work on it .

This study has raised important questions about the nature of Hadith concept. First, the concept must be one or two words but for the Hadith, we notice that the concept may be one sentence, second why not all the text of Hadith not translated.

Table 4.20 : The evaluated Hadith search tools with the HCSE

Hadith search tool Comparison Criteria	AL Muhaddith	Search Truth	Dourous	Hadith Encyclopedia	HCSE
1.Search techniques	A	A	A	A	A,B
2.Query Analyzer	C	C	C	C	A
3. Database size	A	A	B	B	B
4. Query types	A,B	A,B	A	A	A,B,C
5. Availability	A	A	A	A	A
6.Result ranking	B	B	B	B	A
7.User categories	A	A	A	A	A,B,C,D
8.language of input query	A	B	B	B	A,B

4.10 Summary

The result show that while we are searching in the Hadith using the MHC implementing the TF-IDF to determine the most important terms for each document(Hadith) , calculate the coefficient similarity between the query and all the documents to find the most relevant documents to the query on these aspect before apply stemming the result of precision and recall for the Arabic are 75 % and 59 % ,result of precision and recall for English are 89.4 % and 52 % , result of precision and recall for French are 75.5 % and 79.4 % and for Russian precision and recall are

78.5 % and 53.4 respectively. Beside that we notice that there is more improvement when we apply the stemming process which all the result show improvement regards all the languages starting with precision and recall for Arabic 96.5 % and 82 %, for English are 98.4 % and 90 % ,for French are 97.5 and 91.7 and for Russian are 98 % and 91 % respectively.

CHAPTER V

CONCLUSION AND FUTURE WORKS

5.1 Introduction

In this chapter we had given the final conclusion for our work and suggestion of some recommendation of the future works.

5.2 Conclusion

We conclude that developing of a MHC would be a significant and worthy project. The proposed MHC has the potential to become a wonderfully useful source for the most Muslims around the world, as well as for other researchers from other religious backgrounds who are seeking information regarding the Hadith in different languages.

Moreover the research prove that we can use the MHC to enhance the search result of the Hadith in the information retrieval system. As our result in chapter four show that the precision and recall for the Arabic language are 98.4 % and 90 %, for the English the precision and recall are 96.5 % and 82 % , for French are 97.5 % and 91.7 % and for Russian are 98 % and 91 % respectively.

For study of the Hadith using the corpus query tool SketchEngine. By building parallel corpora from our MHC data we show that we can make it available in the SketchEngine website and use the features available in their website like concordance and word list. Linguistics and lexicography researchers can undertake further study about Hadith, like data consistency ,data definition and data frequency as examples. on the other hand the CSTH show that a significant implications for the understanding of how the concept of Hadith work and the important of translating all

the concept in all the language to make the Hadith text is understandable and easy to find and search by Arabic and by all other languages.

Finally as our best knowledge the MHC is the first source for Hadith in multiple languages for the existing corpora.

5.3 Future Works

For the future work the researcher recommends the following the MHC can be extended to have Hadith in other languages, Hadith Explanation in other languages, and Hadith classification can be done.

For the future work, we plan to annotation the MHC with XML metadata tagging, and Part-of-Speech or lemma tagging to make it more useful for lexical research, and adding more languages like Chinese and Urdu.

This research will serve as a base for future studies to extend the search concept tool with more language like Chinese ,Urdu, Turkey and more languages.

5.4 Main Contributions

Our main contribution can be descried shortly as following:

1. Developing a Multilingual Hadith Corpus based on the survey design to focuses on the nine aspects of user requirements.
- 2.To enhance the search for Hadith in the information retrieval system .
- 3.Develop website to hold the entire MHC as open source ,access file in three different formats (.txt , .html ,xml).
- 4.Ability to use the MHC with corpora analysis tools like SketchEngine.

Bibliography

Al-Saif, A. and Markert, K., 2010, May. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *LREC*.

Atwell, E.S., Howarth, P.A. and Souter, D.C., 2003. The ISLE corpus: Italian and German spoken learner's English. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, 27, pp.5-18.

Atwell, E. and Hardie, A., Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.

Atwell, E. and Hardie, A., Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.

Al-Sulaiti, L. and Atwell, E.S., 2006. The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, 11(2), pp.135-171.

Al-Saif, A. and Markert, K., 2010, May. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *LREC*.

Abu Shawar, B.A., 2005. *A corpus based approach to generalise a chatbot system* (Doctoral dissertation, University of Leeds).

Alrehaili, S.M. and Atwell, E., 2014. Computational ontologies for semantic tagging of the Quran: A survey of past approaches. In *LREC 2014 Proceedings*. European Language Resources Association.

Habash, N., Soudi, A. and Buckwalter, T., 2007. On arabic transliteration. In *Arabic computational morphology* (pp. 15-22). Springer Netherlands.

Roberts, A., Al-Sulaiti, L. and Atwell, E., 2006. aConCorde: Towards an open-source, extendable concordancer for Arabic. *Corpora*, 1(1), pp.39-60.

Sawalha, M. and Atwell, E., 2011. Accelerating the processing of large corpora: using grid computing technologies for lemmatizing 176 million words Arabic internet corpus. *Advanced Research Computing Open Event*.

Sharaf, A.B.M. and Atwell, E., 2012. QurSim: A corpus for evaluation of relatedness in short texts. In *LREC* (pp. 2295-2302).

Reference

Abbas, N.H., 2009. *Quran's search for a concept tool and website* (Doctoral dissertation, University of Leeds (School of Computing)).

Alfaifi, A.Y.G. and Atwell, E.S., 2013. Arabic learner corpus v1: A new resource for arabic language research.

AL Imam Majd al-Din Abu Saadat Al Mubarak bin Mohammed Al-Jazari bin Al-Athir., 2000. *Alnehia in a strange word for Hadith and Al-Athr*. Dar Ibn Al-Jawzia. First edition.

Alrabiah, M., Al-Salman, A. and Atwell, E.S., 2013. The design and construction of the 50 million words KSUCCA. In *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics* (pp. 5-8). The University of Leeds.

Alansary, S., Nagi, M. and Adly, N., 2007, December. Building an international corpus of Arabic (ICA): Progress of compilation stage. In *7th international conference on language engineering, Cairo, Egypt* (pp. 5-6).

Al-Sulaiti, L. and Atwell, E.S., 2006. The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, 11(2), pp.135-171.

Al-Shalabi, R., Kanaan, G. and Gharaibeh, M., 2006, April. Arabic text categorization using KNN algorithm. In *Proceedings of The 4th International Multiconference on Computer Science and Information Technology* (Vol. 4, pp. 5-7).

Alsaleem, S., 2011. Automated Arabic Text Categorization Using SVM and NB. *Int. Arab J. e-Technol.*, 2(2), pp.124-128.

Al-Thubaity, A.O., 2015. A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, 49(3), pp.721-751.

Ahmed, R.O.M., Supervised, A.A.M.A. and Ali, S.A.A.M., 2015. *Design of Arabic Dialects Information Retrieval Model for Solving Regional Variation Problem* (Doctoral dissertation, Sudan University of Science and Technology).

AL Imam Majd al-Din Abu Saadat Al Mubarak bin Mohammed Al-Jazari bin Al-Athir.(2000) .Alnehia in a strange word for Hadith and Al-Athr. Dar Ibn Al-Jawzia. First edition.

Arts, T., Belinkov, Y., Habash, N., Kilgarriff, A. and Suchomel, V., 2014. arTenTen: Arabic corpus and word sketches. *Journal of King Saud University-Computer and Information Sciences*, 26(4), pp.357-371.

Bader Alden Abu Mohammed AlAene,2000.Omdet Alqari shrih Albukhari. Published by:Dar Alktub Alalmea.

Bar-Ilan, J., 2002. Criteria for evaluating information retrieval systems in highly dynamic environments. In *Proceedings of the 2nd International Workshop on Web Dynamics, Honolulu, Hawaii, USA*.

Bennett, B., 2005. Modes of concept definition and varieties of vagueness.*Applied Ontology*, 1(1), pp.17-26.

Bilal, K. and Mohsin, S., 2012, December. Muhadith: A cloud based distributed expert system for classification of ahadith. In *Frontiers of Information Technology (FIT), 2012 10th International Conference on* (pp. 73-78). IEEE.

Barzilay, R. and McKeown, K.R., 2001, July. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting on Association for Computational Linguistics* (pp. 50-57). Association for Computational Linguistics.

- Büttcher, S., Clarke, C.L. and Cormack, G.V., 2016. *Information retrieval: Implementing and evaluating search engines*. Mit Press.
- Chen, Y., 2010, August. Natural Language Processing in Web data mining. In *2010 IEEE 2nd Symposium on Web Society* (pp. 388-391). IEEE.
- Christopher, D.M., Prabhakar, R. and Hinrich, S.C.H.Ü.T.Z.E., 2008. Introduction to information retrieval. *An Introduction To Information Retrieval*, 151, p.177.
- Croft, W.B., Metzler, D. and Strohmann, T., 2010. *Search engines*. Pearson Education.
- Dukes, K., Atwell, E. and Sharaf, A.B.M., 2010, May. Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. In *LREC*.
- Elanwar, R.I.M., 2012. A semi-automatic annotation tool for Arabic online handwritten text. *CU Theses*.
- Fawcett, J., Ayers, D. and Quin, L.R., 2012. *Beginning XML*. John Wiley & Sons.
- Goweder, A. and De Roeck, A., 2001. Assessment of a significant Arabic corpus. In *Arabic NLP Workshop at ACL/EACL*.
- Giunchiglia, F., Kharkevich, U. and Zaihrayeu, I., 2008. Concept search: Semantics enabled syntactic search.
- Goker, A. and Davies, J. eds., 2009. *Information retrieval: Searching in the 21st century*. John Wiley & Sons.
- Grinberg, M., 2014. *Flask Web Development: Developing Web Applications with Python*. " O'Reilly Media, Inc."
- Grossman, D.A. and Frieder, O., 2012. *Information retrieval: Algorithms and heuristics* (Vol. 15). Springer Science & Business Media.

Harrag, F., El-Qawasmah, E. and Al-Salman, A.M.S., 2011, April. Stemming as a feature reduction technique for Arabic text categorization. In *Programming and Systems (ISPS), 2011 10th International Symposium on* (pp. 128-133). IEEE.

Héja, E., 2010, May. The Role of Parallel Corpora in Bilingual Lexicography. In *LREC*.

Jbara, K., 2010. Knowledge discovery in Al-Hadith using text classification algorithm. *Journal of American Science*, 6(11), pp.409-19.

Kim, J.D., Cohen, K.B. and Kim, J.J., 2015, August. PubAnnotation-query: a search tool for corpora with multi-layers of annotation. In *BMC Proceedings*(Vol. 9, No. 5, p. 1). BioMed Central.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D., 2004. Itri-04-08 the sketch engine. *Information Technology*, 105, p.116.

Levene, M., 2011. *An introduction to search engines and web navigation*. John Wiley & Sons.

McEnery, T. and Xiao, R., 2007. Parallel and comparable corpora: The state of play. *Corpus-based perspectives in linguistics*, pp.131-145.

Martin, J.H. and Jurafsky, D., 2000. Speech and language processing. *International Edition*.

Manning, C.D., Raghavan, P. and Schütze, H., 2008. *Introduction to information retrieval* (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press.

Najeeb, M., Abdelkader, A., Al-Zghoul, M. and Osman, A., 2015. A lexicon for hadith science based on a corpus. *Int. J. Comput. Sci. Inf. Technol*, 6, pp.1336-1340.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,

- Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), pp.2825-2830.
- Perkins, J., 2014. *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd.
- Porter, M.F., 1980. An algorithm for suffix stripping. *Program*, 14(3), pp.130-137.
- Ramos, J., 2003, December. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.
- Rapp, R., 1999, June. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 519-526). Association for Computational Linguistics.
- Roberts, A., Al-Sulaiti, L. and Atwell, E., 2005, July. aConCorde: Towards a proper concordance of Arabic. In *Proceedings of the Corpus Linguistics 2005 Conference, University of Birmingham, UK*.
- Roberts, A., 2009. Grammatical Inference and Corpus Linguistics.
- Saad, M.K. and Ashour, W., 2010. Arabic morphological tools for text mining. *Corpora*, 18, p.19.
- Salton, G. and Buckley, C., 1990. *Flexible text matching for information retrieval*. Cornell University.
- Salton, G., 1968. *Automatic information organization and retrieval*. New York: McGraw-Hill.
- Shanahan, J.G., Qu, Y. and Wiebe, J. eds., 2006. *Computing attitude and affect in text: theory and applications* (Vol. 20). Dordrecht, the Netherlands: Springer.

Tiedemann, J. and Nygaard, L., 2004. The OPUS Corpus-Parallel and Free: <http://logos.uio.no/opus>. In *LREC*.

Van Zaanen, M., Roberts, A. and Atwell, E.S., 2004. A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation. In *Proceedings of LREC'04 Workshop on The Amazing Utility of Parallel and Comparable Corpora* (pp. 58-61). European Language Resources Association.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L. and Trón, V., 2007. Parallel corpora for medium density languages. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 292, p.247.

Yahya, A., 1989. On the complexity of the initial stages of Arabic Text Processing. *Birzeit University, Birzeit, West Bank*.

Yang, C.C., 2010. Search engines information retrieval in practice.

Zuva, K. and Zuva, T., 2012. Evaluation of information retrieval systems. *International Journal of Computer Science & Information Technology (IJCSIT)*, 4(3).

Wikipedia (2016).Hadith[online].Available from - <https://en.wikipedia.org/wiki/Hadith>. [Accessed:9th March 2016]

United States (2016).Official-Languages [online].Available from - <http://www.un.org/en/sections/about-un/official-languages/index.html>. [Accessed:9th Jan 2015]

Appendix A

MHC survey

Hadith User Requirements Survey 20.05.2015

By Samah Mohammed Osman, PhD Student, Sudan University of Science and Technology, Khartoum

THANK YOU IN ADVANC FOR YOUR HELP WITH THIS PROJECT

Part -1:

1. Religion:

2. Gender: male\ female

3. What is your age ?

a- (20-30)

b- (31-40)

c- (41-50)

d- (51-60)

e- Above 61

4. What is your job?

a- Computer scientist

b- Programmer

c- Teacher

d- Engineering

e -Other.....

5. Why do you want to know about Hadith ? ..(تعبد)...

a- Education

b-Religious

Part -2:

If you want to read and understand Hadith ,would you like to have:-

1.Hadith explanation in Arabic.

- a- very useful
- b- useful
- c-not useful

2.Meaning of each word in Hadith by Arabic.

- a- very useful
- b- useful
- c- not useful

3.Find the Hadith, word meaning, grammar and benefits in one website.

- a- very useful
- b- useful
- c- not useful

4.Find Hadith translated in different languages.

- a- very useful
- b- useful
- c- not useful

5.For each Hadith you find the source from which book the Hadith is copied.

- a- very useful
- b- useful
- c- not useful

6- for each Hadith you want to know the classification (correct(Sahih),weak(Daeef),good(Hassan).

- a- very useful
- b- useful
- c- not useful

7. When you want to read about Hadith ,would you prefer to read from:

a- books

b- one website

c- different websites

8. Which is easier for you .If you want to search about any Hadith words :

a- from different websites

b- from one Website

Thank You for giving me your time and completing this Survey

Samah Mohammed Osman

Phd Student

Sudan University of Science and Technology

Appendix B

Examples of MHC files format

A.1 XML file format:

```
<?xml version="1.0" encoding="UTF-8"?>
<!--XML database-->
<Data>
<Hadith_Source> الحديث الاول </Hadith_Source>
<Notice>الطبعة الثالثة</Notice>
<Hadith_Arabic> عَنْ أَمِيرِ الْمُؤْمِنِينَ أَبِي حَفْصٍ عُمَرَ بْنِ الْخَطَّابِ قَالَ: سَمِعْتُ رَسُولَ اللَّهِ يَقُولُ: " إِنَّمَا
الأَعْمَالُ بِالنِّيَّاتِ، وَإِنَّمَا لِكُلِّ امْرِئٍ مَا نَوَى، فَمَنْ كَانَتْ هِجْرَتُهُ إِلَى اللَّهِ وَرَسُولِهِ فَهَاجَرَتْهُ إِلَى اللَّهِ وَرَسُولِهِ، وَمَنْ
كَانَتْ هِجْرَتُهُ لِدُنْيَا يُصِيبُهَا أَوْ امْرَأَةٍ يَنْكِحُهَا فَهَاجَرَتْهُ إِلَى مَا هَاجَرَ إِلَيْهِ
</Hadith_Arabic>
<Hadith_Explian>انما الاعمال بالنيات اي صحة مايقع من المكلف من قول او فعل او كماله وترتب الثواب
عليه لا يكون الاحسب ماينوية والنيات هي جمع نية وهي القصد وعزم القلب علي امر من الامور وهجرة في
اللغة الخروج من ارض الي ارض ومفارقة الوطن والاهل مشتقة من الهجر وهو ضد الوصل وشرعا هي
مفارقة دار الكفر الي دار الاسلام والمراد بها هنا الخروج من مكة الي غيرها الي مدينة رسول الله صلي الله
عليه وسلم ويصيبها اي يحصلها وينكحها اي يتزوجها وفهجرة الي ماهاجر اليه اي جزاء عملة الغرض الدنيوي
</Hadith_Explian>
<Hadith_Narrated_by>رواة مسلم والبخاري</Hadith_Narrated_by>
<Hadith_English>Narrated 'Umar bin Al-Khattab: I heard Allah's Apostle saying,
"The reward of deeds depends upon the intentions and every person will get the
reward according to what he has intended. So whoever emigrated for worldly benefits
or for a woman to marry, his emigration was for what he emigrated for
</Hadith_English>
<Hadith_French>Le Commandeur des Croyants, Aboû Hafç Omar ben El-Kattâb
(que Dieu soit satisfait de lui) a dit: J'ai entendu l' Envoyé de Dieu, salla Allah u alihi
wa sallam , (à lui, bénédiction et salut) dire: « Les actions ne valent que par leurs
intentions ". Leurs Niyates: « Chacun ne recevra la récompense qu'il mérite que selon
ce qu'il a entendu faire. A celui qui a accompli l'hégire pour plaire à Allah et à Son
Envoyé, son hégire lui sera comptée, comme accomplie en vue de Dieu et de Son
Envoyé. Celui qui l'a accomplie pour obtenir quelque bien en ce bas monde, ou pour
```

épouser une femme, son hégire lui sera comptée selon ce qu'il recherchait

alors</Hadith_French>

<Hadith_Russian>Сообщается, что 'Умар бин аль-Хаттаб, да будет доволен им Аллах, сказал: - Я слышал, как посланник Аллаха, , ска- зал: «Поистине, дела (оцениваются) только по намерениям и, поистине, каждому человеку (достанется) лишь то, что он намеревался (об- рести), и (поэтому) переселявшийся ради чего- нибудь мирского или ради женщины, на кото- рой он хотел жениться , переселится (лишь) к тому, к чему он переселялся</Hadith_Russian>

</Data>

A.2 HTML file format

<!DOCTYPE html><html>

<body>

<p>

</p><p>الحديث الاول في الاربعين النووية

<p>

عَنْ أَمِيرِ الْمُؤْمِنِينَ أَبِي حَفْصٍ عُمَرَ بْنِ الْخَطَّابِ قَالَ: سَمِعْتُ رَسُولَ اللَّهِ يَقُولُ: " إِنَّمَا الْأَعْمَالُ بِالنِّيَّاتِ، وَإِنَّمَا لِكُلِّ امْرِئٍ مَا نَوَى، فَمَنْ كَانَتْ هِجْرَتُهُ إِلَى اللَّهِ وَرَسُولِهِ فَهَجْرَتُهُ إِلَى اللَّهِ وَرَسُولِهِ، وَمَنْ كَانَتْ هِجْرَتُهُ لِدُنْيَا يُصِيبُهَا أَوْ امْرَأَةٍ يَنْكِحُهَا فَهَجْرَتُهُ إِلَى مَا هَاجَرَ إِلَيْهِ </p>

<p>

</p><p>رواة مسلم البخاري

<p>

انما الاعمال بالنيات اي صحة مايقع من المكلف من قول او فعل او كماله وترتب الثواب عليه لا يكون الاحسب ماينوية والنيات هي جمع نية وهي القصد وعزم القلب علي امر من الامور وهجرة في اللغة الخروج من ارض الي ارض ومفارقة الوطن والاهل مشتقة من الهجر وهو ضد الوصل وشرعا هي مفارقة دار الكفر الي دار الاسلام والمراد بها هنا الخروج من مكة الي غيرها الي مدينة رسول الله صلي الله عليه وسلم ويصيبها اي يحصلها وينكحها اي يتزوجها وهجرة الي ماهاجر اليه اي جزء عملة الغرض الدنيوي الذي قصده ان حصله </p><p>والافلا شئ له

<p>

Narrated 'Umar bin Al-Khattab:

I heard Allah's Apostle saying, "The reward of deeds depends upon the intentions and

every person will get the reward according to what he has intended. So whoever emigrated for worldly benefits or for a woman to marry, his emigration was for what he emigrated for</p>

<p>

Le Commandeur des Croyants, Aboû Hafç Omar ben El-Kattâb (que Dieu soit satisfait de lui) a dit: J'ai entendu l' Envoyé de Dieu, salla Allah u alihi wa sallam , (à lui, bénédiction et salut) dire:

« Les actions ne valent que par leurs intentions ". Leurs Niyates:

« Chacun ne recevra la récompense qu'il mérite que selon ce qu'il a entendu faire. A celui qui a accompli l'hégire pour plaire à Allah et à Son Envoyé, son hégire lui sera comptée, comme accomplie en vue de Dieu et de Son Envoyé. Celui qui l'a accomplie pour obtenir quelque bien en ce bas monde, ou pour épouser une femme, son hégire lui sera comptée selon ce qu'il recherchait alors</p>

<p>

Сообщается, что 'Умар бин аль-Хаттаб, да будет доволен им Аллах, сказал:

- Я слышал, как посланник Аллаха, , сказал: «Поистине, дела (оцениваются) только по намерениям и, поистине, каждому человеку (достанется) лишь то, что он намеревался (обрести), и (поэтому) переселявшийся ради чего-нибудь мирского или ради женщины, на которой он хотел жениться , переселится (лишь) к тому, к чему он переселялся</p>

</body>

</html>

A.3 Plain Text for Arabic Hadith with no Metadata

حدثنا الحميدى عبد الله بن الزبير قال حدثنا سفيان قال حدثنا يحيى بن سعيد
الأنصارى قال أخبرنى محمد بن إبراهيم التيمى أنه سمع علقمة
بن وقاص الليثى يقول سمعت عمر بن الخطاب رضى الله عنه
على المنبر قال سمعت رسول الله صلى الله عليه وسلم يقول
إنما الأعمال بالنيات وإنما لكل امرئ ما نوى فمن كانت هجرته
إلى دنيا يصيبها أو إلى امرأة ينكحها فهجرته إلى ما هاجر إليه

A.4 Plain text for Arabic Hadith with Metadata

صحيح البخاري
الطبعة الخامسة
1993
1
بدء الوحي
حدثنا الحميدى عبد الله بن
الزبير قال حدثنا سفيان قال حدثنا يحيى بن سعيد الأنصارى قال
أخبرنى محمد بن إبراهيم التيمى أنه سمع علقمة بن وقاص الليثى
يقول سمعت عمر بن الخطاب رضى الله عنه على المنبر
قال سمعت رسول الله صلى الله عليه وسلم يقول
إنما الأعمال بالنيات وإنما لكل امرئ ما نوى فمن كانت هجرته إلى
دنيا يصيبها أو إلى امرأة ينكحها فهجرته إلى ما هاجر إليه

A.5 Plain text for English Hadith with no Metadata

narrated umar bin alkhatab i heard allahs apostle saying the reward of deeds depends upon the intentions and every person will get the reward according to what he has intended so whoever emigrated for worldly benefits or for a woman to marry his emigration was for what he emigrated for

A.6 Plain text for English Hadith with Metadata

Sahih Bukhari

Fifth Edition

1993

1

Revelation

Narrated Said bin Jubair: Ibn 'Abbas in the explanation of the Statement of Allah. 'Move not your tongue concerning (the Quran) to make haste therewith.'

Said "Allah's Apostle used to bear the revelation with great trouble and used to move his lips (quickly) with the Inspiration." Ibn 'Abbas moved his lips saying, "I am moving my lips in front of you as Allah's Apostle used to move his." Said move lips saying: "I am moving my lips, as I saw Ibn 'Abbas moving his. " Ibn 'Abbas added, "So Allah revealed 'Move not your tongue conc

A.7 Plain text for French Hadith with no Metadata

selon alqama ben ouaqas el laïti alors qu'il était sur le minbar omar ben el khattab prononça les paroles suivantes j'ai entendu l'envoyé de dieu dire les actions ne valent que selon les intentions pour chaque homme les intentions sont déterminantes ainsi celui qui émigrera pour les biens de ce monde ou pour chercher une épouse ne sera rétribué que pour l'objectif qu'il s'était fixé

A.9 Plain text for French Hadith with Metadata

Sahih D'el Bokhari

1

la revelation

Selon Aïcha, la mère des Croyants, Harit Ben Hicham demanda au Prophète Ô Envoyé de Dieu comment se manifeste à toi la Révélation Parfois répondit celui ci je ressens comme le timbre d'une clochette et c'est le plus éprouvant pour moi Puis quand l'épreuve se termine alors seulement je comprends le sens du message En d'autres occasions l'ange prend une apparence humaine et je retiens les Paroles qu'il me communique Et Aïcha d'ajouter Certains jours de grand froid le Prophète recevait la Révélation, et à la fin, je voyais son front ruisseler de sueur

A.9 Plain text for Russian Hadith with no Metadata

Начало откровений

Глава: О том, как откровения начали ниспосы сланнику Аллаха, д
Сообщается, чт бин аль-Хаттаб, да будет доволен им Аллах, сказ:
«Я слы посланник Аллаха, да благословит его Аллах и приветству
Передают со слов ‘Аиши, да будет доволен ею Аллах, что (одна
«О посланник Алла приходят к тебе откровения?» Посланник Ал
‘Аиша, да будет доволен ею Аллах, сказала:
«И м сь видеть, как в очень холодные дни ему ниспосылались от
Сообщается, что мат верных ‘Аиша, да будет доволен ею Аллах,
«Ниспослание откровений посланнику Аллаха, да благословит е

A.10 Plain text for Russian Hadith with Metadata

Appendix C

The programming code

view.py file

```

from flask import Flask, url_for, request, render_template;
from MHCAApp import app;
from MHCAApp.models.sqlclient import Client;
import nltk
from MHCAApp.models.classfile import classfile;
from MHCAApp.models.MHCClass import MHCClass;
import re,math;
import nltk;
import timeit;

## server/

@app.route('/')
def second():
    createLink7 = "<p><a href='" + url_for('findarabictext') + "'>Home</a>";
    createLink1 = "<a href='" + url_for('findarabictext') + "'>Arabic</a>";
    createLink2 = "<a href='" + url_for('findenglishtext') + "'>English</a>";
    createLink3 = "<a href='" + url_for('findfrenchtext') + "'>French</a>";
    createLink4 = "<a href='" + url_for('findRussiantext') + "'>Russian</a>";
    createLink5 = "<a href='" + url_for('Downloadf') + "'>Download</a>";
    createLink6 = "<a href='" + url_for('ContactForm') + "'>Contact
Us</a></P>";

    return """<html>
        <head>
            <title>Hadith search</title>
        </head>
        <body>
            <h2 align="center"> Hadith Corpus Search Engine</h2>
            <br\>
            <br\>
            """ + createLink7 + """
            """ + createLink1 + """
            """ + createLink2 + """
            """ + createLink3 + """
            """ + createLink4 + """
            """ + createLink5 + """
            """ + createLink6 + """

            <p>
This Hadith Corpus Search Engine(HCSE) search Tool design to search into the
Multilingual Hadith corpus.It allow user to Search for the Hadith text.Hadith
is text said by Prophet Mohammed(peace be upon him).
In this Hadith corpus the translation for each text of Hadith with three
differen languags along with the original language for Hadith the Arabic
language.The generated concept for Hadith coming from SAHIH ALBUKHARI book for
Hadith written by the Islamic scholar Abu Abdullah Muhammad bin Ismail bin
Ibrahim bin al-Mughira al-Ja'fai.
In this serach we are using the four differnt languages for Hadith text,Arabic
,English ,French and Russain.</p>

```

```

        </body>
    </html>""";

@app.route('/Downloadf', methods=['GET', 'POST'])
def Downloadf():
    if request.method == 'GET':
        return render_template('Downloadfiles.html');
    elif request.method == 'POST':

        return render_template('CreatedQuestion.html')
    else:
        return "<h2>Invalid request</h2>";

@app.route('/ContactForm', methods=['GET', 'POST'])
def ContactForm():
    if request.method == 'GET':
        return render_template('contactForm.html');
    elif request.method == 'POST':

        return render_template('CreatedQuestion.html')
    else:
        return "<h2>Invalid request</h2>";

@app.route('/show', methods=['GET', 'POST'])
def show():
    if request.method == 'GET':

        #return render_template('showhadith.html');
        return render_template('fileopen.html');
    elif request.method == 'POST':

        bookname = request.form['bookname'];
        # meaning = request.form['meaning'];

        client = Client();
        allhadith= client.getAllhadith(bookname);
        # hadith= client.getHadith(word);

        return render_template('allhadith.html',
allhadith=allhadith,bookname=bookname)
    else:
        return "<h2>Invalid request</h2>";

    #This Function it will return the meaning of the word from the table wordbu
@app.route('/create', methods=['GET', 'POST'])
def create():
    if request.method == 'GET':
        return render_template('searchbyword.html');
    elif request.method == 'POST':
        word = request.form['word'];
        client = Client();
        meaning= client.getAnswer(word);
        return render_template('CreatedQuestion.html', word =
word,meaning=meaning)
    else:
        return "<h2>Invalid request</h2>";

@app.route('/searchhadith', methods=['GET', 'POST'])

```

```

def searchhadith():
    if request.method == 'GET':
        return render_template('searchbyword.html');
    elif request.method == 'POST':
        word = request.form['word'];
        client = Client();
        meaning= client.getwordbu(word);
        hadith= client.getHadith(word);
        return
render_template('Correct.html',word=word,meaning=meaning,hadith=hadith)
    else:
        return "<h2>Invalid request</h2>";

@app.route('/findenglishtext', methods=['GET', 'POST'])
def findenglishtext():
    if request.method == 'GET':

        return render_template('searchbyword.html');

    elif request.method == 'POST':
        start = timeit.default_timer();
        Classfile=classfile();
        wordsearch=request.form['word'];
        #This part take the word change it to lower case before search in
the text
        allcontent,total=Classfile.getEnglishText(wordsearch.lower());
        #allcontent=Classfile.getconcordance(wordsearch);

        stop = timeit.default_timer();
        tt1=stop - start;
        tt=round(tt1,3)
        return
render_template('Textdemo.html',wordsearch=wordsearch,allcontent=allcontent,tot
al=total,tt=tt);

    else:
        return "<h2>Invalid text</h2>";

@app.route('/findfrenchtext', methods=['GET', 'POST'])
def findfrenchtext():
    if request.method == 'GET':

        return render_template('searchbyword.html');

    elif request.method == 'POST':
        start = timeit.default_timer();
        Classfile=classfile();
        wordsearch=request.form['word'];
        #This part take the word change it to lower case before search in
the text
        allcontent,total=Classfile.getFrenchText(wordsearch.lower());
        stop = timeit.default_timer();
        tt1=stop - start;
        tt=round(tt1,3)
return render_template('Textdemo.html',allcontent=allcontent,total=total,tt=tt);

    else:
        return "<h2>Invalid request</h2>";

@app.route('/findarabictext', methods=['GET', 'POST'])
def findarabictext():

```

```

    if request.method == 'GET':

        return render_template('searchbyword.html');

    elif request.method == 'POST':
        Classfile=classfile();
        wordsearch=request.form['word'];
        #This part take the word change it to lower case before search in
the text
        allcontent,total=Classfile.getArabicText(wordsearch);

        return render_template('Textdemo.html',allcontent=allcontent,total=total);

    else:
        return "<h2>Invalid request</h2>";

@app.route('/findRussiantext', methods=['GET', 'POST'])
def findRussiantext():
    if request.method == 'GET':

        return render_template('searchbyword.html');

    elif request.method == 'POST':
        start = timeit.default_timer();
        Classfile=classfile();
        wordsearch=request.form['word'];
        #This part take the word change it to lower case before search in
the text
        allcontent,total=Classfile.getRussianText(wordsearch.lower());
        stop = timeit.default_timer();
        tt1=stop - start;
        tt=round(tt1,3)
        return
render_template('Textdemo.html',wordsearch=wordsearch,allcontent=allcontent,tot
al=total,tt=tt);

    else:
        return "<h2>Invalid request</h2>";

@app.route('/Arabic', methods=['GET', 'POST'])
def Arabic():
    if request.method == 'GET':

        return render_template('ConceptArabic.html');
    elif request.method == 'POST':

        text = request.form['concept'];
        # meaning = request.form['meaning'];
        concept=convertArabic(text);
        client = Client();
        result= client.getconcept(concept);
        # hadith= client.getHadith(word);
        NO1=client.getNoFrawsAr(concept);
        NO=str(NO1).strip('[(,)]');
Return render_template('arabicResult.html',NO=NO,concept=concept,result=result)
    else:
        return "<h2>Invalid request</h2>";

@app.route('/English', methods=['GET', 'POST'])
def English():
    if request.method == 'GET':

```

```

        return render_template('ConceptArabic.html');
    elif request.method == 'POST':
        concept = request.form['concept'];
        client = Client();
        result= client.getconceptEnglish(concept);
        NO1=client.getNoFraws(concept);
        NO=str(NO1).strip('[(,)]');

    return render_template('EnglishResult.html',NO=NO,
concept=concept,result=result);

    else:
        return "<h2>Invalid request</h2>";

@app.route('/French', methods=['GET', 'POST'])
def French():
    if request.method == 'GET':

        return render_template('ConceptArabic.html');
    elif request.method == 'POST':
        concept = request.form['concept'];
        client = Client();
        result= client.getconceptFrench(concept);
        NO1=client.getNoFrawsFr(concept);
        NO=str(NO1).strip('[(,)]');
        return render_template('EnglishResult.html',NO=NO,
concept=concept,result=result);
    else:
        return "<h2>Invalid request</h2>";

@app.route('/Russ', methods=['GET', 'POST'])
def Russ():
    if request.method == 'GET':

        return render_template('ConceptArabic.html');
    elif request.method == 'POST':
        concept = request.form['concept'];
        client = Client();
        result= client.getconceptRussian(concept);
        NO1=client.getNoFrawsRu(concept);
        NO=str(NO1).strip('[(,)]');
        return
render_template('RussainResult.html',NO=NO,concept=concept,result=result);

    else:
        return "<h2>Invalid request</h2>";
#This function to correct the arabic word
def convertArabic(text):
    text = re.sub("[\u0600-\u06FF]", "|", text)
    text = re.sub("ى", "ي", text)
    text = re.sub("ة", "ه", text)
    text = re.sub("س", "س", text)
    return(text)

@app.route('/highlight', methods=['GET', 'POST'])
def highlight(query, text):
    def span_matches(match):
        html = '<span class="query">{0}</span>'
        return html.format(match.group(0))
    return re.sub(query, span_matches, text, flags=re.I)

@app.route('/allmatch', methods=['GET', 'POST'])

```

```

def allmatch():
    if request.method == 'GET':

        return render_template('ConceptArabic.html');
    elif request.method == 'POST':
        concept = request.form['concept'];
        client = Client();
        allcontent= client.getAllmatchEnglish(concept);
        return render_template('Textdemo.html',allcontent=allcontent);

    else:
        return "<h2>Invalid request</h2>";

@app.route('/display', methods=['GET', 'POST'])
def display():
    if request.method == 'GET':

        return render_template('searchbyword.html');
    elif request.method == 'POST':
        mhcclass=MHCCClass();
        wordsearch=request.form['word'];
        result=mhcclass.dispalyXmlData(wordsearch);
        return render_template('XmlDisplay.html',result=result);
    else:
        return "<h2>Invalid request</h2>";

@app.route('/mytoken', methods=['GET', 'POST'])
def mytoken():
    if request.method == 'GET':

        return render_template('ForTokenizer.html');
    elif request.method == 'POST':
        token = request.form['token'];
        client = classfile();
        result= client.getToken(token);
        return render_template('incorrect.html',result=result);
    else:
        return "<h2>Invalid request</h2>";
    #the follwoing function to return the corpus
@app.route('/allHadith', methods=['GET', 'POST'])
def allHadith():
    if request.method == 'GET':

        return render_template('H1.html');
    elif request.method == 'POST':
        concept = request.form['concept'];

        return render_template('Textdemo.html',allcontent=allcontent);

    else:
        return "<h2>Invalid request</h2>";

```

Models Class

```
#I build this class to handle the File search
#All the text file will be handle from these class functions
import nltk;
import csv;
from nltk.tokenize import word_tokenize,sent_tokenize
import codecs

class classfile(object):
    """description of class"""
    def getEnglishText(self,word):
        #This function will return the apperance of the word in the English
        Text of Hadith
        try:
            myfile=open("C:\ipython\Book1.txt","r");
            mylist=[];
            text=word;
            count=0;
            for line in myfile:
                if text in line:
                    mylist.append(line);
                    count=count+1;
            total=count;
            return mylist,total;
            myfile.close();
        except:
            return err;

    def getArabicText(self,word):
        try:
            myfile=codecs.open("C:\ipython\Arabictext.txt","r","utf-8");
            mylist=[];
            text=word;
            count=0;
            for line in myfile:
                if text in line:
                    mylist.append(line);
                    count=count+1;
            total=count;
            return mylist,total;
            myfile.close();
        except:
            return err;

    def getFrenchText(self,word):
        try:
            myfile=open("C:\ipython\HF.csv","r");
            mylist=[];
            text=word;
            count=0;
            for line in myfile:
                if text in line:
                    mylist.append(line);
                    count=count+1;
            total=count;
            return mylist,total;
            myfile.close();
        except:
            return err;

    def getRussianText(self,word):
```



```

        try:
myfile=codecs.open("C:\ipython\RH.csv","r","utf-8");
mylist=[];
text=word;
count=0;
for line in myfile:
    if text in line:
        mylist.append(line);
        count=count+1;
total=count;
return mylist,total;
myfile.close();
except:
    return err;

```

searchByword.html

```

<html>

    <body>
        <br />
        <br />

        <h2 align="center"> Hadith Corpus Search Engine</h2>
<form method="post">

    <div align="center">

        <input name="word" type="text" style="width: 377px" />
    </div>
    <div align="center">
        <br />
        <button type="submit">Search</button>
    </div>
</form>

    </body>
</html>

```

Appendix D

Hadith source books

اسم المترجم	العام	دار النشر	الطبعة	اللغة	اسم المؤلف	اسم الكتاب
	2002	دار ابن كثير	الاولي	العربية	محمد بن اسماعيل ابو عبد الله البخاري الجعفي	صحيح البخاري
M.Muhsin Khan	2009		First	ENGLISH	Abu Abdalla Muhammad bin Ismail bin Ibrahim bin Al- Mughira Al jafai	SAHIH BUKHARI
				Russian		Программа по изучению хадисов (منهج (الحديث النبوي
Владимир (Абдулла) Михайлович Нирша, кандидат философских наук				Russian	Имам Мухаммад бин Исма'ил Абу 'Абдуллах аль-Джу'фи аль-Бухари	САХИХ АЛЬ- БУХАРИ МУХТАСАР
Harkat Ahmed	2003	المكتبة العصرية - صيدا بيروت	الثالثة	French	l'Imam Abu Abdalla Muhammad ben Ismail	Le Sahih d' al-Bukha'ry

List of Publication:

Conference paper:

Hassan, S. M. O. and Atwell, E., 2016. 'Compilation of an Islamic Hadith Corpus', in R Saeed, A El Faki Proceedings, tenth *International Computing Conference in Arabic*, Khartoum, Sudan 2016, Phillips Publishing, Phillipsburg, NJ USA.

Journal Paper:

Hassan, S. M. O. and Atwell, E., 2016. Design Requirements for Multilingual Hadith Corpus, *International Journal of Science and Research (IJSR)*, <https://www.ijsr.net/archive/v5i4/v5i4.php>, Volume 5 Issue 4, April 2016, 494 - 498

Hassan, S. M. O. and Atwell, E., 2016. Concept Search Tool for Multilingual Hadith Corpus, *International Journal of Science and Research (IJSR)*, <https://www.ijsr.net/archive/v5i4/v5i4.php>, Volume 5 Issue 4, April 2016, 1326 - 1328

Hassan, S. M. O. and Atwell, E., 2016. Design and Implementing of Multilingual Hadith Corpus, *International Journal of Recent Research in Social science and Humanities (IJRRSSH)*, <http://www.paperpublications.org/>, Volume 3 Issue 2, April 2016, pp: (100-104).