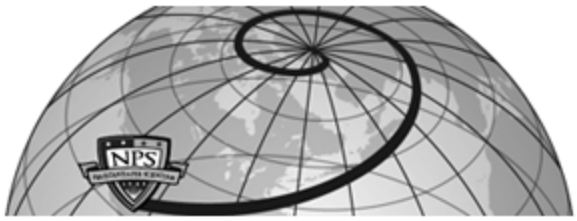

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

ENHANCING A WEB CRAWLER WITH ARABIC SEARCH CAPABILITY

by

Qui V. Nguyen

September 2010

Thesis Advisor:	Weilian Su
Second Reader:	John C. McEachen

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** September 2010	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE Enhancing a Web Crawler with Arabic Search Capability			**5. FUNDING NUMBERS**
6. AUTHOR(S)			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number _____N.A._____.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited			**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

Many advantages of the Internet—ease of access, limited regulation, vast potential audience, and fast flow of information—have turned it into the most popular way to communicate and exchange ideas. Criminal and terrorist groups also use these advantages to turn the Internet into their new play/battle fields to conduct their illegal/terror activities. There are millions of Web sites in different languages on the Internet, but the lack of foreign language search engines makes it impossible to analyze foreign language Web sites efficiently. This thesis will enhance an open source Web crawler with Arabic search capability, thus improving an existing social networking tool to perform page correlation and analysis of Arabic Web sites. A social networking tool with Arabic search capabilities could become a valuable tool for the intelligence community. Its page correlation and analysis results could be used to collect open source intelligence and build a network of Web sites that are related to terrorist or criminal activities.

14. SUBJECT TERMS Nutch, Lucene, Web Crawler, Information Retrieval in Arabic, Stemming in Arabic			**15. NUMBER OF PAGES** 121
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

ENHANCING A WEB CRAWLER WITH ARABIC SEARCH CAPABILITY

Qui V. Nguyen
Lieutenant, United States Navy
B.S., University of Texas at Dallas, 1999

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

from the

NAVAL POSTGRADUATE SCHOOL
September 2010

Author: Qui V. Nguyen

Approved by: Weilian Su
Thesis Advisor

John C. McEachen
Second Reader

R. Clark Robertson
Chairman, Department of Electrical and Computer Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Many advantages of the Internet—ease of access, limited regulation, vast potential audience, and fast flow of information—have turned it into the most popular way to communicate and exchange ideas. Criminal and terrorist groups also use these advantages to turn the Internet into their new play/battle fields to conduct their illegal/terror activities. There are millions of Web sites in different languages on the Internet, but the lack of foreign language search engines makes it impossible to analyze foreign language Web sites efficiently. This thesis will enhance an open source Web crawler with Arabic search capability, thus improving an existing social networking tool to perform page correlation and analysis of Arabic Web sites. A social networking tool with Arabic search capabilities could become a valuable tool for the intelligence community. Its page correlation and analysis results could be used to collect open source intelligence and build a network of Web sites that are related to terrorist or criminal activities.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. INTRODUCTION 1
A. MOTIVATION 1
B. RESEARCH OBJECTIVES 1
C. THESIS ORGANIZATION 2

II. ARABIC INFORMATION RETRIEVAL 3
A. INFORMATION RETRIEVAL 3
B. THE CHALLENGES OF ARABIC IR 4
C. RESEARCH IN ARABIC IR 5
D. LIGHT STEMMER ALGORITHM 6
1. Introduction 6
2. The Algorithm 6
a. Normalization *6*
b. Light Stemmers *7*
3. Results 8
E. CHAPTER SUMMARY 9

III. LUCENE AND NUTCH 11
A. INTRODUCTION 11
B. LUCENE 11
1. Overview 11
2. Indexing Process 11
3. Analyzer 13
C. NUTCH 13
1. Architecture Overview 13
2. Plug-In Architecture 14
D. CHAPTER SUMMARY 15

IV. ARABICANALYZER PLUG-IN DEVELOPMENT 17
A. INTRODUCTION 17
B. REQUIREMENT 18
C. DEVELOPMENT PROCESS 18
1. Implementation of the Light Stemmer Algorithm 18
2. Development of *ArabicAnalyzer* Plug-in 18
3. Creating Arabic Ngram profile 19
D. CHAPTER SUMMARY 20

V. EXPERIMENTAL SETUP 21
A. PROBLEM STATEMENT 21
B. HARDWARE AND SOFTWARE CONFIGURATION 21
C. METHODOLOGY 22
D. RESULTS AND DISCUSSION 23
1. Terms Count 23
2. Crawl Time 24

3. Search Results 25
E. CHAPTER SUMMARY 32
VI. CONCLUSION 33
A. SUMMARY 33
B. FUTURE WORK 33
APPENDIX A 35
APPENDIX B 47
APPENDIX C 57
APPENDIX D 69
APPENDIX E 81
APPENDIX F 91
LIST OF REFERENCES 101
INITIAL DISTRIBUTION LIST 103

LIST OF FIGURES

Figure 1. String removed by light stemming. From [14] 7
Figure 2. Monolingual 11-point precision results. From [14].....8
Figure 3. Lucene indexing architecture. From [17].....12
Figure 4. Nutch's architecture. From [18].....14
Figure 5. The process of indexing a Web site 17
Figure 6. *ArabicAnalyzer* plug-in architecture. From [19].....19

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1. The uniterpolated average precision. From [14] 9
Table 2. The number of terms counts 23
Table 3. Average crawl time of www.america.gov/ar/ 24
Table 4. Average crawl time of www.bbc.co.uk/arabic/ 24
Table 5. Average crawl time of www.addustour.com 24
Table 6. Average crawl time of www.aawsat.com 24
Table 7. Search terms 25
Table 8. Search results of term “Economy” using ArabicAnalyzer 26
Table 9. Search results of term “Economy” using NutchDocumentAnalyzer 27
Table 10. Search results of term “The United States” using ArabicAnalyzer 28
Table 11. Search results of term “The United States” using NutchDocumentAnalyzer 29
Table 12. Search results of term “Democratic” using ArabicAnalyzer 30
Table 13. Search results of term “Democratic” using NutchDocumentAnalyzer 31

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

After more than eight years of the War on Terrorism, Improvised Explosive Devices (IEDs) have become the weapon of choice for the terrorist in Iraq and Afghanistan. IEDs accounted for the majority of causalities of Allied forces and civilians. One of the reasons for the proliferation of IEDs is the ease of access to training material available on the Internet. The Internet is a cheap, convenient, yet powerful tool to access a vast reservoir of information and knowledge. Unfortunately, the Internet also empowers technology-savvy terror networks and extremist groups to create IED education networks and distribute the IED know-how to their operatives and supporters quickly and efficiently.

One solution to counter this problem is a social networking tool that applies networking theory and social network analysis to identify terrorist IED education networks quickly. This tool would utilize an open source web crawler that could index Arabic websites into a searchable database for analyzing and querying to collect more actionable intelligence.

The Nutch project was selected as the search engine of choice for this social networking tool. Its transparency ranking information allows the users the ability to tailor the ranking to meet the user's specific requirements. Its versatile plug-in architecture provides extensibility, flexibility and maintainability.

To enable Nutch indexing of Arabic websites, an Arabic language analyzer needs to be added into Nutch's library. Multiple experiments were used to test the performance of the Arabic language analyzer with moderate results.

Overall, Nutch with an added Arabic analyzer would be a valuable tool improving an existing social networking tool to perform page correlation and analysis of Arabic websites. Its results could be used to identify IED education networks and to collect open source intelligence.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I want to thank my family, Andrea and Rosie, for supporting me. Without their support and encouragement, I would not finish this thesis.

I also would like to thank Dr. Su for his guidance in the past year. His advice was valuable to the completion of this thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. MOTIVATION

Since its invention, the Internet has revolutionized communication. It enables people to exchange ideas and share information rapidly and cheaply. Unfortunately, its lack of regulation and pervasive communication also has turned it into the new tool for the tech-savvy terrorists: "Today, almost without exception, all major (and many minor) terrorist and insurgent groups have web sites" [1]. Many terror organizations such as Al-Qaeda actively use the Internet to recruit new members, solicit donations from sympathizers, and spread propaganda.

They also turn the Internet into their virtual training grounds, offering tutorials on building IEDs and planning attacks. These training materials are easily accessible to anyone with an Internet connection. This is the main contribution to the explosion of IED attacks in Iraq and Afghanistan. To counter the proliferation of IED technology, these IED education networks need to be identified, monitored and referred to sovereign authorities for further action as necessary.

One possible solution for this problem is a social networking tool that applies network science to identify the IED education network via the World Wide Web. In [2], network science is defined as the study of networks which "contrasts, compares, and integrates techniques and algorithms developed in disciplines as diverse as mathematics, statistics, physics, social network analysis, information science and computer science." The social network tool would incorporate an open source web crawler that could index Arabic websites into a searchable database for analyzing and querying.

B. RESEARCH OBJECTIVES

The research objectives of this thesis were to enhance a Web crawler engine with Arabic search capability that could index Arabic language websites proficiently, thus improving an existing social networking tool to perform page correlation and analysis of

Arabic websites. The newly enhanced Web crawler could help speed up the analytical process of the social networking tool to effectively identify IED education networks via the World Wide Web.

C. THESIS ORGANIZATION

This thesis consists of six chapters. An overview of the motivation, objectives and thesis organization is provided in Chapter I. A brief discussion about information retrieval, a description of Arabic information retrieval challenges, stemming in Arabic and the light stemmer algorithm is contained in Chapter II. Lucene—a scalable Information Retrieval (IR) library; Nutch—an open source search engine; and Nutch's plug-in architecture are introduced in Chapter III. The implementation process of the light stemmer algorithm into Lucene's analyzer database, and development of the *ArabicAnalyer* plug-in are discussed in Chapter IV. The performance of *ArabicAnalyzer* and *NutchDocumentAnalyzer* are compared in Chapter V. The summary of the thesis and future research recommendation are discussed in Chapter VI.

II. ARABIC INFORMATION RETRIEVAL

A. INFORMATION RETRIEVAL

The fast growth of the Internet accompanied with the explosion of data available via the World Wide Web has made the finding of useful information a tedious and difficult task. These difficulties have attracted renewed interested in Information Retrieval and its techniques.

Information Retrieval (IR) is the science of locating relevant documents in a large collection of documents. The retrieval process is influenced by queries supplied by the user's input, the indexing process and the natural language that is being indexing [3].

In [4], some popular IR classic strategies are the Vector Space Model, Probabilistic Retrieval, Language Model, and Inference Networks.

The Vector Space Model is a widely used retrieval strategy. In this model, both the query and each document are represented as vector in terms of space. A measure of similarity between the two vectors is computed.

In the Probabilistic Retrieval model, a probability based on the likelihood that a term will appear in a relevant document is computed for each term in the collection. For terms that match between a query and a document, the similarity measure is computed as the combination of the probabilities of each of the matching terms [4].

In the Language Model, a language model is inferred for each document; then the probability of generating the query according to each of these models is computed. Documents are then ranked according to these probabilities [5].

Inference Networks, also known as Bayesian networks, are used to model documents, the documents' contents and the query. It then uses this information to derive —"infer"—other relationships. The strength of this inference is then used as the similarity coefficient [4].

B. THE CHALLENGES OF ARABIC IR

According to [6], there are over 200 million native Arabic speakers in the world and over 20 million people speaking it as a second language. There is also an exponential growth of Internet in speaking countries. From [7], the numbers of Internet users in Middle East countries alone have grown from 3 million in 2000 to 58 million in 2009. So, there is increasingly a demand for an Arabic IR as well, but Arabic poses many challenges for IR

First, Arabic has a very complex morphology system. In [8], the authors observed:

> Arabic has two genders, feminine and masculine; three numbers, singular, dual and plural; and three grammatical cases, nominative, genitive, and accusative. A noun has the nominative case when it is a subject, accusative when it is the object of a verb, and genitive when it is the object of a preposition.

This would compound the complexity of any Arabic IR to deal with this morphology system.

Second, there are a lot of ambiguities in Arabic. One of the major contributions to this phenomenon is that orthographic variations are widespread in Arabic [9]. The authors gave an example that sometimes in combining HAMZA with ALEF (أ) or MADDA with ALEF (آ), the HAMZA (ء) or MADDA (~) is dropped, rendering it ambiguous to whether the HAMZA (ء) or MADDA (~) is present. Another contribution to the higher level of ambiguity is that sometimes vowels (diacritics) are omitted in written Arabic, which may change the meaning of the words. This uncertainty would affect the precision and recall of any Arabic IR.

Finally, the plural form of irregular nouns, broken plurals, is common in Arabic. A broken plural's form does not resemble its initial singular form. It does not obey normal morphological rules. Because of that, it is very difficult to design an algorithm to transform this kind of plural to singular form [9].

C. RESEARCH IN ARABIC IR

Research on Arabic IR has focused on using word roots and stems as index terms. A stem is the remainder of the word after removing prefixes and suffixes. On the other hand, the root is the origin of the word that remains after removing nonessential characters, prefixes and suffixes. When using word roots as index terms, a linguistic knowledge and an understanding of the languages' morphology are needed. On the other hand, prior knowledge of the language is not required when using stems as index terms. In [10], the authors recognized that "stemming is one of many tools besides normalization that is used in information retrieval to combat the vocabulary mismatch problem." As discussed in section 2b, Arabic is very difficult to stem, therefore, there were only a few available Arabic stemmers.

One of the earliest stemmers was the root-based stemmer proposed by Khoja and Garside. This stemmer removed all the stopwords, punctuation, and numbers. Then it peeled away prefixes and suffixes. After that, it matched the result against a list of patterns to extract the root. Finally, it matched the extracted root against a list of known "valid" roots. There are a few weaknesses in the Khoja stemmer. First, it can provide wrong solutions when removing prefixes and suffixes. It also can generate wrong roots for words that contain *EBDAL* [10], [11], [12].

Buckwalter's morphological analyzer is another useful stemmer. First, this stemmer converts the Arabic word into English letters. Then, it segments it into all probabilities of prefixes, stems, and suffixes. After that, it checks every probability with its build-in lexicon libraries (prefixes dictionary, stems dictionary and suffixes dictionary). If all the word elements (prefix, stem, suffix) are found in their respective libraries, three truth tables indicating their legal combination (prefixes-suffixes, prefixes-stems, and stems-suffixes) are used to determine whether they are compatible. If the word elements pass all three truth tables, the probability is valid. This stemmer provides highly reliable results, but its performance is slow [13].

The light stemmer is another approach for Arabic IR. Most light stemmers in [8], [14] are based on the same idea: extract stems by deleting the most frequent prefixes and

suffixes. These stemmers are not interested in producing the Arabic root. This thesis applies the light stemmer algorithm in [14] to enable the Web crawler with an Arabic search capability. A more detailed discussion is in the next section.

D. LIGHT STEMMER ALGORITHM

1. Introduction

The light stemmer allows for good information retrieval results without providing the correct morphological analyses [10]. Anyone can employ the light stemmer algorithm without the required language skills.

2. The Algorithm

The stemmer has two parts: Normalization and Stemmer. The Normalization process is used to normalize the orthography—the writing system—of the queries and corpus. The stemmer removes suffixes using the light stemmer algorithm to extract the stems [14].

a. Normalization

In [14], before stemming, corpus and queries are normalized as follows:

(1). Convert to Windows Arabic encoding (CP12560).

(2). Remove punctuation.

(3). Remove diacritics (primary weak vowels).

(4). Remove non letters.

(5). Replace آ (ALEF with MADDA above), أ (ALEF with HAMZA above), and إ (ALEF with HAZA below) with ا (ALEF)

(6). Replace final ى (ALEF MAKSURA) with ي (YEH)

(7). Replace final ة (TEH MARBUTA) with ه (HEH)

b. Light Stemmers

After the corpus and queries are normalized, the stemmer is applied as follows:

(1). Remove و (WAW) if the remainder of the word is three or more characters long.

(2). Remove any of the definite articles if this leaves two or more characters.

(3) Go through the list of suffixes once in the right to left order indicated in Figure 1, removing any that are found at the end of the word, if this leaves two or more characters.

	Remove from front	Remove Suffixes
Light1	ال، وال، بال، كال، فال	none
Light2	ال، وال، بال، كال، فال، و	none
Light3	"	ه، ة
Light8	"	ها، ان، ات، ون، ين، يه، ية، ه، ة، ي

Figure 1. String removed by light stemming. From [14]

Light1, Light2, Light3 and Light8 apply the same algorithm in the stemming process. The difference between them is the number of prefixes and suffixes that are removed in step 3 of the light stemmer's algorithm. In Light1, the Light Stemmer algorithm only removes five prefixes and no suffixes. In Light2, the Light Stemmer algorithm removes six prefixes and no suffixes. In Light3, the Light Stemmer algorithm removes six prefixes and two suffixes. In Light8, the Light Stemmer algorithm removes six prefixes and 10 suffixes.

3. Results

The authors in [14] compared the retrieval effectiveness of the light stemmer algorithm (Light8) and of a morphological analyzer (Khoja stemmer). Raw in Figure 2 means no normalization and stemming. From Figure 2, we see that the light stemmer outperforms Khoja stemmer and raw retrieval. From Table 1, we see that light stemmer improved over 90% in average precision from raw retrieval.

The authors concluded that stemming is very effective on Arabic IR. For monolingual retrieval, the light stemmer has demonstrated improvement of around 100% in average precision due to stemming and related processes.

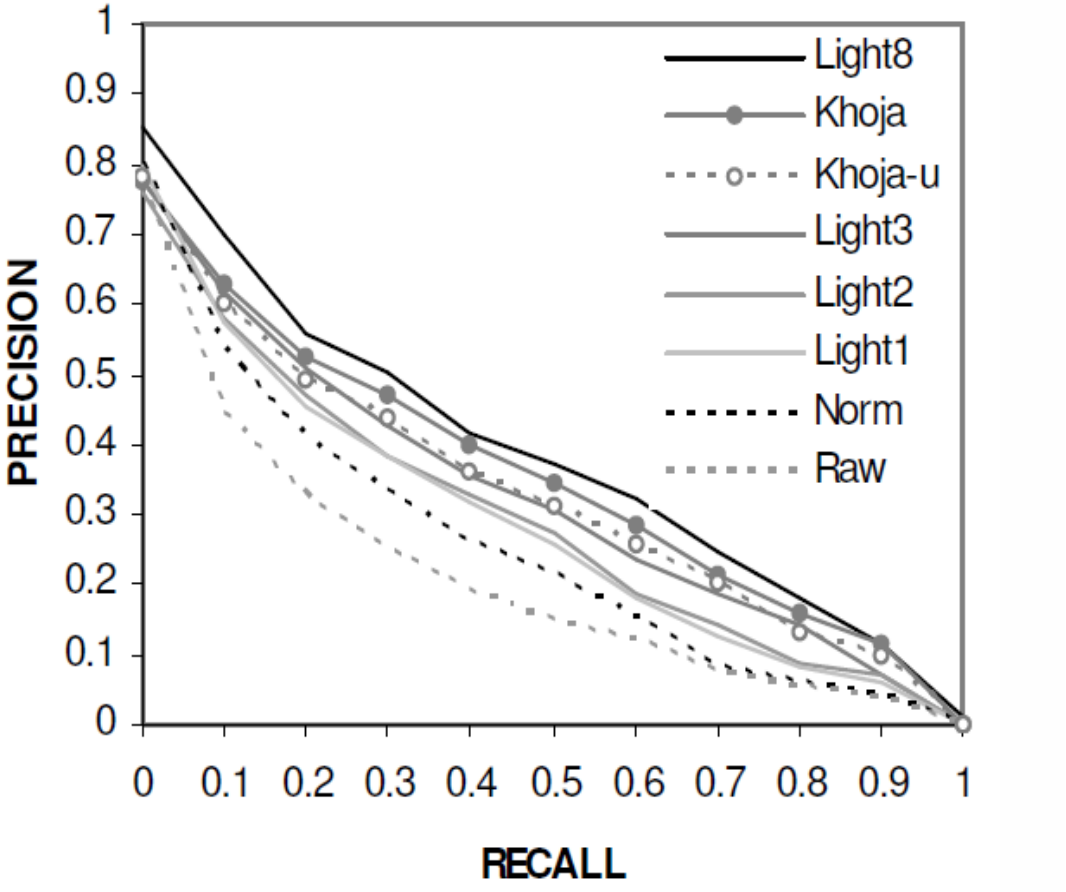

Figure 2. Monolingual 11-point precision results. From [14]

Table 1. The uniterpolated average precision. From [14]

Stemmer	raw	khoja-u	khoja	light8
Av. Precision	.194	.313	.341	.376
Pct. Change		61.7	76.2	94.3

E. CHAPTER SUMMARY

In this chapter, the challenges of Arabic IR and past Arabic IR research were covered. Also discussed was the effectiveness of light stemmer in Arabic IR. In the next chapter, Lucene, Nutch and Nutch's plug-in architecture are introduced.

THIS PAGE INTENTIONALLY LEFT BLANK

III. LUCENE AND NUTCH

A. INTRODUCTION

Lucene and Nutch, created by Doug Cutting, are two open-source software projects. According to [15], Lucene is a high performance, scalable Information Retrieval (IR) library that provides Java-based indexing and searching technology and advanced analysis/tokenization capabilities. On the other hand, Nutch is a search engine that was built on top of Lucene. Together, they can make a full-featured search engine that offers transparency into how Web sites are ranked, and an understanding of how a large search engine works [16].

B. LUCENE

1. Overview

Lucene is a software library that enables users to add indexing and searching capabilities to their application. Lucene can index and search any type of data as long as it can be converted into a text format. This means Lucene can be used to search Web pages, pdf files, and Microsoft® Word files because textual information can be extracted from them. With this feature, Lucene is the best toolkit for a search engine.

2. Indexing Process

Indexing is the process of converting text into an index, a data structure that improves the speed of data retrieval operations. The index is the fundamental component of Lucene.

From [16], to index data with Lucene, the data must be converted into a stream of plain text tokens, a format that Lucene can process. After that, Lucene prepares the data for indexing by breaking the stream of plain text into chunks or tokens and performing a number of operations on them. For instance, the tokens could be lowercase before

indexing, to make the search case-insensitive. This step is called analysis. After the input has been analyzed, it is ready to be added into the index. The Indexing process is illustrated in Figure 3.

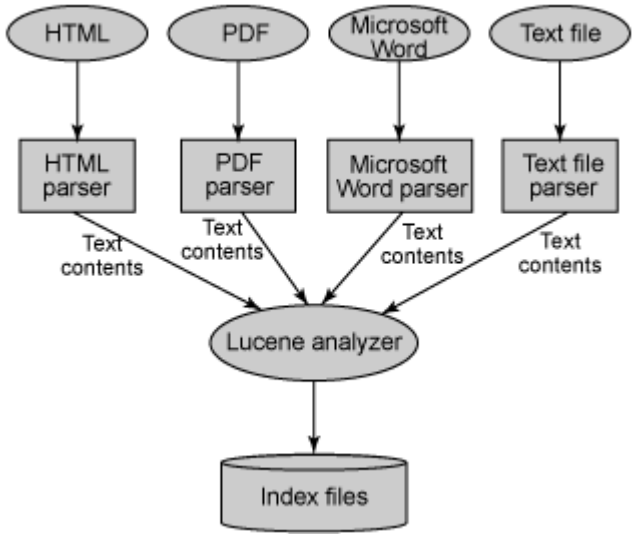

Figure 3. Lucene indexing architecture. From [17]

Lucene implements an innovative approach to maintaining the index—rather than maintaining a single index, Lucene builds multiple index segments and merges them periodically. Using segments allows a quick way to add new documents to the index by adding them to the newly created index segments and only periodically merging them with other existing segments. This process makes additions efficient because it minimizes physical index modifications.

Some IR libraries need to index the whole corpus again when new data is added to their index; Lucene does not need to do that because it supports incremental indexing. This means Lucene allows the contents of newly added documents be searchable immediately without indexing the whole corpus again [15].

3. Analyzer

As discussed above, analysis is a very important step in the indexing process. It converts a field of text into the most fundamental indexed representation, terms. These terms are used to determine what documents match a query during searches.

An analyzer is an encapsulation of the analysis process. The analyzer's job is to process strings of text into a stream of tokens by performing any number of operations on them. Lucene includes several built-in analyzers that do a good job at analyzing English-based text. For analyzing non-English languages, specific language analyzers are needed. Lucene's core API provides building blocks to create custom language analyzers.

C. NUTCH

1. Architecture Overview

Nutch is a complete open-source Web search engine that can operate at one of three scales: local file system, intranet, or whole Web [15]. Nutch can be divided into two parts: the crawler and the searcher.

From [18], components of the crawler are WebDB, the fetch list, fetchers and updates. WebDB is a custom database that tracks every known page and relevant link. It maintains a small set of facts about each page, such as the last crawled date. Fetch lists are generated from WebDB. These lists contain the URLs that users want to download. The fetchers consume the fetch lists to produce the WebDB updates and the Web contents. The updates tell which page has changed since the last crawl. The contents are used to search. The WebDB-fetch cycle is designed to repeat forever, maintaining an up-to-date image of the Web.

Once the Web content is produced, Nutch can get ready to process queries using the searchers. First, the indexer processes the Web content of all terms and pages into an inverted index. The document set is divided into a set of index segments, each of which is fed into a single searcher process. Each searcher also draws upon the Web content

from earlier to provide a cached copy of any Web page. Finally, a pool of Web servers handles the interaction with users and contact searcher for results. A generic overview of Nutch's architecture is shown in Figure 4.

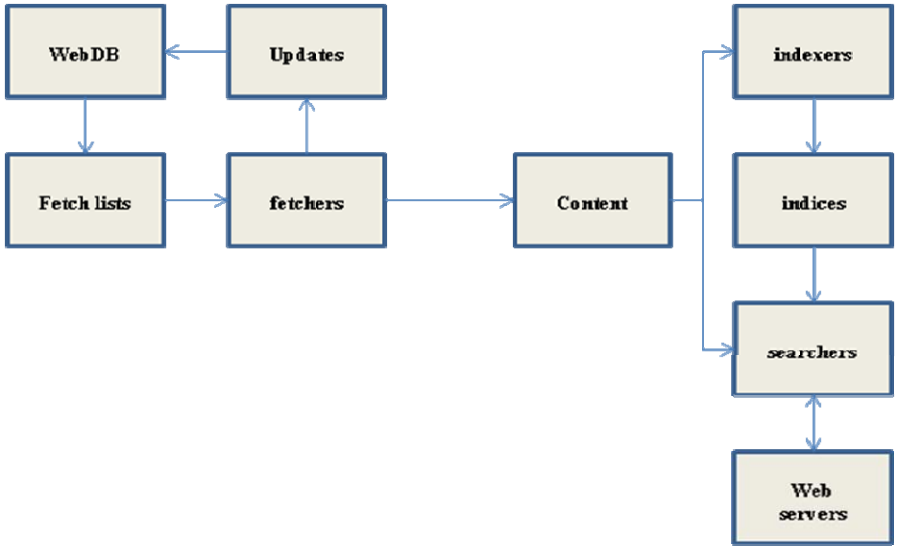

Figure 4. Nutch's architecture. From [18]

2. Plug-In Architecture

Nutch's plug-in system is based on the Eclipse 2.0 plug-in architecture. It provides a core service for controlling a set of tools working together to support programming tasks. After reviewing Eclipse's architecture from [19] and applying it to Nutch's plug-in system, we observe that the three most important components of Nutch's plug-in system are Extension, ExtensionPoints and Plug-in. The Extension class provides a way to add some new functions to a plug-in. It is defined by a plug-in that wants to extend its functionality to another plug-in. ExtensionPoints define an interface that must be implemented by the Extension. A plug-in, pluggable component, defines a number of extension-points that may allow it to be augmented by different kinds of extension.

This system is the mechanism of Nutch's extensibility. Users can contribute to the Nutch platform by wrapping their tools in plug-ins. The new plug-ins can add new processing elements to existing plug-ins, and Nutch provides a set of core plug-ins to assist the process.

D. CHAPTER SUMMARY

In this chapter, the overview of Lucene's indexing process and analyzer were examined. The overview of Nutch's architecture and its plug-in system were also studied. In the next chapter, the implementation process of the light stemmer algorithm into Nutch is discussed.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ARABICANALYZER PLUG-IN DEVELOPMENT

A. INTRODUCTION

When Nutch finishes fetching a segment of Web sites, the *language-identifier* plug-in is called to identify the language of the Web sites and attach a language code to those Web sites. After that, the *Analyzerfactory* instantiates the *NutchAnalyzer* interface, which defines an extension point that associates with the specific language code. The *NutchAnalyzer* extension point is an abstract class that extends the Lucene Analyzer class, so that Lucene analyzers can be easily integrated as *NutchAnalyzer* plug-ins. The policy of the *Analyzerfactory* for finding the *NutchAnalyzer* extension to use is to return the first one that matches a specified language code. If none is found, then the default *NutchDocumentAnalyzer* is used. After *Analyzerfactory* identifies the right analyzer basing on the language code, the *NutchAnalyzer* calls the correct analyzer, in this case *ArabicAnalyzer,* from the Lucene analyzer library to index the Web site. The process of indexing a Web site is shown in Figure 5.

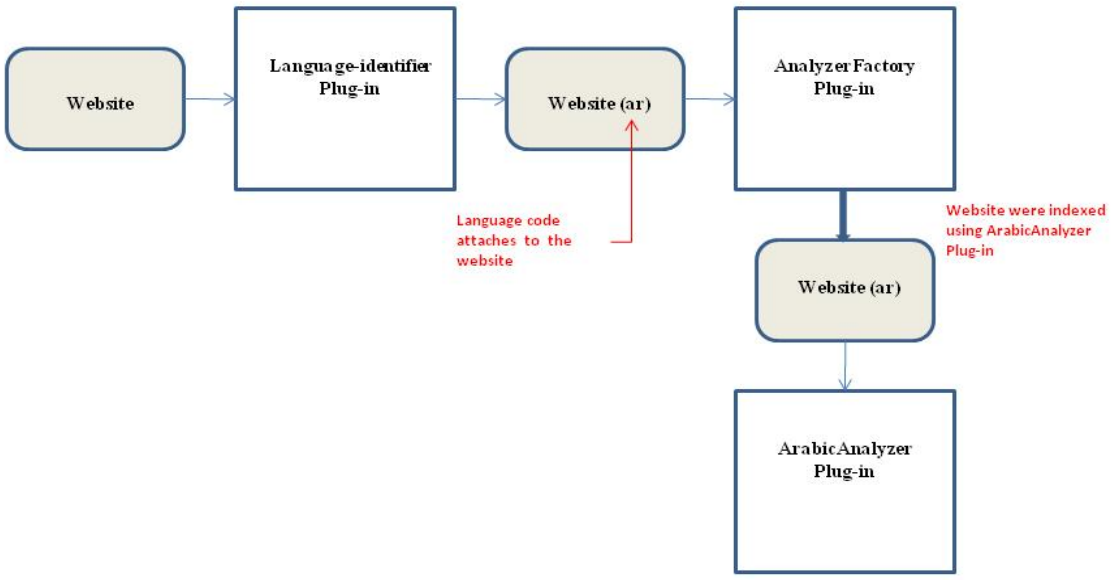

Figure 5. The process of indexing a Web site

B. REQUIREMENT

To enable Nutch with Arabic-search capability, there are several tasks that need to be completed. First, the Lucene analysis library needs to be updated with the *ArabicAnalyzer* that implemented the light stemming algorithm. Secondly, an *ArabicAnalyzer* plug-in is needed for Nutch to be able to access the Lucene analysis library. Finally, an Arabic Ngram profile is needed to train Nutch how to recognize Arabic text.

C. DEVELOPMENT PROCESS

1. Implementation of the Light Stemmer Algorithm

As stated above, the Lucene analysis library needs to be updated with the *ArabicAnalyzer*, which implements the light stemmer algorithm. The analysis package contains three primary files: *ArabicAnalyzer, ArabicNormalizationFilter*, and *ArabicStemFilter.*

The *ArabicAnalyzer* first creates a list of Arabic stop words that is based on the stoplist from *http://members.unine.ch/jacques.savoy/clef/index.html.* It uses the standard Stopfilter to filter out all the stop words from the token stream. The result is then fed into *ArabicNormalizationFilter*, which normalizes the orthography. The final result is then fed into the *ArabicStemFilter,* which stems the token stream using the light stemmer algorithm.

2. Development of *ArabicAnalyzer* Plug-in

The host plug-in is the *ArabicAnalyzer* class in Nutch. The *NutchAnalyzer*, a Nutch built-in extension point, defines the interface that must be implemented by the Nutch's *ArabicAnalyzer.* The extender plug-in is the *ArabicAnalyzer* from Lucene's analysis library that extends the functions of the Nutch's *ArabicAnalyzer*; in this case, the Lucene's *ArabicAnalyzer* enables the Nutch's *ArabicAnalyzer* to index Arabic text.

Basically, the Nutch's *ArabicAnalyzer* plug-in is a wrapper that sets the stages and makes it possible to run Lucene's *ArabicAnalyzer*. The *ArabicAnalyzer* plug-in architecture that was derived from [19] is shown in Figure 6.

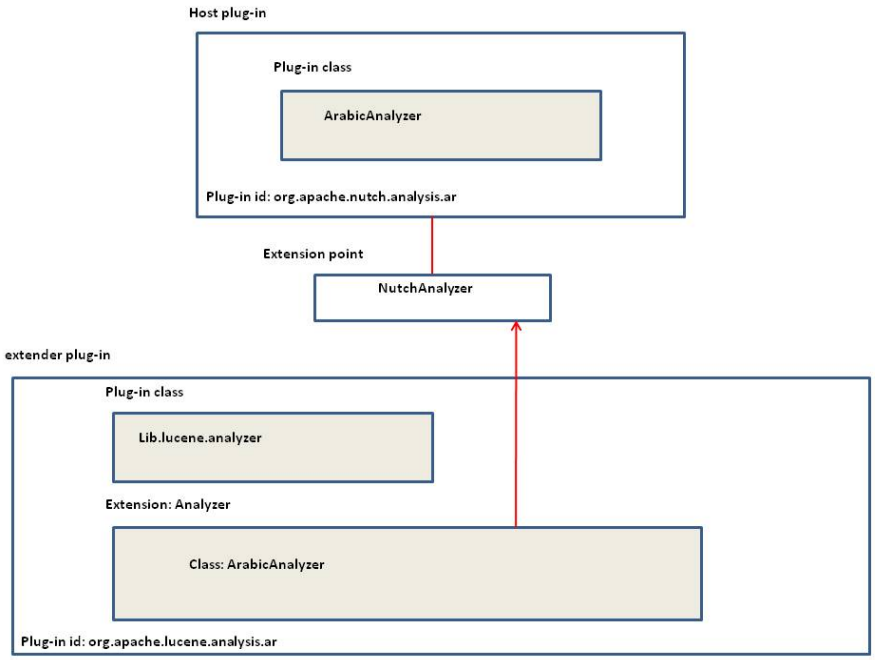

Figure 6. *ArabicAnalyzer* plug-in architecture. From [19]

3. Creating Arabic Ngram profile

Nutch uses the *language-identifier* plug-in in standard Nutch's library to create an Arabic profile based on the "1000 most frequent words" by Jacques Savoy from the Web site http://members.unine.ch/jacques.savoy/clef/index.html. This trains Nutch to "recognize" Arabic Web sites so that it can invoke the right analyzer to index the Web sites.

D. CHAPTER SUMMARY

In this chapter, the ArabicAnalyzer plugin development process is discussed. The Lucene's analyzer library is enhanced with the *ArabicAnalyzer* that implements the light stemming algorithm. The Nutch's plug-in architecture is utilized to create the *ArabicAnalyzer* plug-in. The plug-in enables the Nutch search engine to index Arabic-language Web sites using the *ArabicAnalyzer* in the Lucene's analyzer library. In the next chapter, the performance of *ArabicAnalyzer* and *NutchDocumentAnalyzer* are compared.

V. EXPERIMENTAL SETUP

A. PROBLEM STATEMENT

These experiments will compare the result of Nutch when it used the default *NutchDocumentAnalyzer* with *ArabicAnalyzer* to analyze the same Web site.

The *NutchDocumentAnalyzer* separates the stream of tokens into individual terms without applying any filter. For example, the token stream "hello world" becomes "hello" "world" after *NutchDocumentAnalyzer* processes it. This study uses *NutchDocumentAnalyzer*'s index result as a baseline, because no term is discarded during indexing when using *NutchDocumentAnalyzer* [15].

On the other hand, the *ArabicAnalyzer* applies several filters when analyzing the stream of tokens. First, the token stream goes to *StopFilter,* which removes all the stop words in the custom-built stop words list. The result is then filtered again using *ArabicNormalizationFilter* to normalize the orthography. After that, the result again is filtered using *ArabicStemFilter,* which applies the light stemming algorithm. The final result is then stored into the index database.

B. HARDWARE AND SOFTWARE CONFIGURATION

The platform used to conduct the experiments was a single Dell XPS M1530 laptop personal computer. This machine had an Intel Core 2 Duo CPU T9300 at 2.5 GHz with 4 GB of RAM and a 185 GB hard disk. The operating system used was Microsoft Windows Vista Home Premium with Service Pack 2.

Nutch 1.0 and Lucene 2.4.0 were used to implement the light stemmer algorithm and for all the experiments.

C. METHODOLOGY

There were three experiments to collect data. The first experiment used Nutch to crawl eight Web sites with the depth of five and topN of 50. TopN determines the maximum number of pages that are retrieved at each level up to the depth. The Web sites are *alarabiya.net*, *aljazeera.net*, *alriyadh.com*, *addustour.com*, *aawsat.com*, *bbc.co.uk/Arabic/*, *arabic.cnn.com* and *america.gov/ar/*. Nutch only indexes the Web pages within these sites using *ArabicAnalyzer* and *NutchDocumentAnalyzer*.

The second experiment computes the average crawl time and its standard deviation. The crawler was set to crawl four out of the eight Web sites above 25 times each.

The third experiment compares the ranking of the top 10 pages after using the two algorithms to search for three different Arabic terms.

To disable *ArabicAnalyzer*, the following code was added into the property block of nutch-site.xml file in the conf folder so that *AnalyzerFactory* is forced to use *NutchDocumentAnalyzer* to index these sites by not specifying any analyzer:

<property>
<name>plugin.includes</name>
<value>protocol-http|urlfilter-regex|parse-(text|html|js)|index-(basic|anchor)|query-(basic|site|url)|response-(json|xml)|summary-basic|scoring-opic|language-identifier</value>
<description>Regular expression naming plugin directory names to include. Any plugin not matching this expression is excluded.
</description>
</property>

To enable *ArabicAnalyzer*, the following code replaces the above code within the nutch-site.xml file. With *ArabicAnalyzer* on, the *AnalyzerFactory* uses it to index these sites:

<property>
<name>plugin.includes</name>

<value>protocol-http|urlfilter-regex|parse-(text|html|js)|index-(basic|anchor)|query-(basic|site|url)|response-(json|xml)|summary-basic|scoring-opic|language-identifier|analysis-ar</value>

<description>Regular expression naming plugin directory names to include. Any plugin not matching this expression is excluded.

</description>

</property>

D. RESULTS AND DISCUSSION

1. Terms Count

The first experiment shows that Nutch needs 20% to 37% fewer terms to index the same number of documents from the same Web site when it uses *ArabicAnalyzer*. The result also means that the *ArabicAnalyzer* plug-in is more efficient when searching its index database, because it searches fewer terms to locate the relevant documents. See Table 2 for the detailed breakdown of each Web site.

Table 2. The number of terms counts

Web sites	**NutchDocumentAnalyzer (Terms count)**	**ArabicAnalyzer (Terms count)**
arabic.cnn.com	24776	15827
alarabiya.net	21140	15806
alriyadh.com	20898	13163
aljazeera.net	18096	13658
bbc.co.uk/arabic/	16061	9957
america.gov/ar/	11435	7958
addustour.com	2888	2075
aawsat.com	1050	847

2. Crawl Time

The second experiment shows that Nutch takes longer to index the same Web sites when it uses *ArabicAnalyzer*. This result is expected, because there are more filters in *ArabicAnalyzer*: thus, it requires more processing power and time to index Web sites.

The results, as illustrated in Tables 3 to 6, also show that the crawl times fluctuated more when Nutch used *ArabicAnalyzer*.

Table 3. Average crawl time of www.america.gov/ar/

	Average Crawl time (sec)	**Standard Deviation (sec)**
NutchDocumentAnalyzer	362.92	15.7
ArabicAnalyzer	375.2	25.32

Table 4. Average crawl time of www.bbc.co.uk/arabic/

	Average Crawl time (sec)	**Standard Deviation (sec)**
NutchDocumentAnalyzer	482.76	5.95
ArabicAnalyzer	546.64	37.05

Table 5. Average crawl time of www.addustour.com

	Average Crawl time (sec)	**Standard Deviation (sec)**
NutchDocumentAnalyzer	104.56	1.67
ArabicAnalyzer	105.12	2.38

Table 6. Average crawl time of www.aawsat.com

	Average Crawl time (sec)	**Standard Deviation (sec)**
NutchDocumentAnalyzer	69.56	2.52
ArabicAnalyzer	70.2	2.84

3. Search Results

For the third experiment, the index database of the Web site www.america.gov/ar/ is used to collect search results data. The terms shown in Table 7 are used for the search.

Table 7. Search terms

Normal Form	Light Stemmer Form	Meaning
الاقتصاد	اقتصاد	Economy
أميركا	اميركا	The United States
الديمقراطية	ديمقراط	Democratic

The Light Stemmer forms are searched using the *ArabicAnalyzer*'s index database and the Normal forms are searched using the *NuthDocumentAnalyzer*'s index database.

When comparing the top 10 pages of the search term "economy," the top seven pages are the same; for the search term "The United States," all top 10 pages are the same; and for the search term "Democratic," six pages are the same but with the ranking different. In all three cases, the search results from NutchDocumentAnalyzer have better ranking scores than the search results from ArabicAnalyzer.

By the title of the search results, one can conclude that their contents are related to the search terms. The two algorithms also hit a high mark on relevance of information that relates to the search terms. See Tables 8 through 13 for the breakdown.

Table 8. Search results of term “Economy” using ArabicAnalyzer

Top 10 pages using ArabicAnalyzer	**Score for Query**
www.america.gov/ar/econ.html	0.3486507
www.america.gov/ar/publications/books/outline-of-the-us-economy.html	0.12422927
www.america.gov/ar/econ/business.html	0.09217107
www.america.gov/ar/reviving_trade_ar.html	0.033118278
www.america.gov/ar/publications/books.html#outline_economy	0.016127191
http://www.america.gov/ar/	0.003058498
http://www.america.gov/ar/multimedia/photogallery.html	6.69E-04
www.america.gov/ar/publications/books.html	6.47E-04
www.america.gov/ar/publications/ejournalusa/1209.html	5.85E-04
www.america.gov/ar/index.html	5.73E-04

Table 9. Search results of term “Economy” using NutchDocumentAnalyzer

Top 10 pages using NutchDocumentAnalyzer	**Score for Query**
www.america.gov/ar/econ.html	0.38501537
www.america.gov/ar/publications/books/outline-of-the-us-economy.html	0.13663794
www.america.gov/ar/econ/business.html	0.09989148
www.america.gov/ar/publications/books.html#outline_economy	0.01747075
www.america.gov/ar/	0.002472951
www.america.gov/ar/multimedia/photogallery.html	5.41E-04
www.america.gov/ar/index.html	4.64E-04
www.america.gov/ar/world/europe.html	4.64E-04
www.america.gov/ar/world/mideast.html	4.64E-04
www.america.gov/ar/world/scasia.html	4.02E-04

Table 10. Search results of term “The United States” using ArabicAnalyzer

Top 10 pages using ArabicAnalyzer	**Score for Query**
www.america.gov/ar/pages/footer/local/about-us.html	0.1196895
www.america.gov/ar/publications/books-content/musliminamerica.html	0.11078926
www.america.gov/ar/amlife.html	0.105654
www.america.gov/ar/services/mobile.html	0.042377986
www.america.gov/ar/multimedia/photogallery.html#/4110/mosques_ar/	0.022628564
www.america.gov/ar/publications/books.html#beingmuslim	0.015091554
www.america.gov/ar/multimedia/photogallery.html#/4110/religious_freedom_ar/	0.01136245
www.america.gov/ar/publications/books.html#governed	0.01132718
www.america.gov/ar/multimedia/photogallery.html#/4110/islam_ar/	0.011314282
www.america.gov/ar/	0.003082759

Table 11. Search results of term “The United States” using NutchDocumentAnalyzer

Top 10 pages using NutchDocumentAnalyzer	**Score for Query**
www.america.gov/ar/pages/footer/local/about-us.html	0.11997691
www.america.gov/ar/publications/books-content/musliminamerica.html	0.11105819
www.america.gov/ar/amlife.html	0.10594571
www.america.gov/ar/services/mobile.html	0.042462345
www.america.gov/ar/multimedia/photogallery.html#/4110/mosques_ar/	0.022691099
www.america.gov/ar/publications/books.html#beingmuslim	0.01513323
www.america.gov/ar/multimedia/photogallery.html#/4110/religious_freedom_ar/	0.011393607
www.america.gov/ar/publications/books.html#governed	0.011358418
www.america.gov/ar/multimedia/photogallery.html#/4110/islam_ar/	0.01134555
www.america.gov/ar/	0.002807639

Table 12. Search results of term “Democratic” using ArabicAnalyzer

Top 10 pages using NutchDocumentAnalyzer	**Score for Query**
www.america.gov/ar/global/democracy.html	0.2665834
www.america.gov/ar/global.html	0.16062789
www.america.gov/ar/publications/ejournalusa/608.html	0.033635326
www.america.gov/ar/publications/ejournalusa/0110.html	0.030587077
www.america.gov/ar/democracy/global/index.html	0.027611194
www.america.gov/ar/	0.002160408
www.america.gov/ar/multimedia/podcast.html	6.10E-04
www.america.gov/ar/publications/books.html	5.85E-04
www.america.gov/ar/amlife.html	5.46E-04
www.america.gov/ar/publications/ejournalusa.html	5.40E-04

Table 13. Search results of term “Democratic” using NutchDocumentAnalyzer

Top 10 pages using NutchDocumentAnalyzer	**Score for Query**
www.america.gov/ar/global/democracy.html	0.29354417
www.america.gov/ar/global.html	0.17680001
www.america.gov/ar/publications/ejournalusa/0110.html	0.033559922
www.america.gov/ar/democracy/global/index.html	0.030395675
www.america.gov/ar/	0.002139257
www.america.gov/ar/multimedia/podcast.html	6.04E-04
http://www.america.gov/ar/amlife.html	5.40E-04
www.america.gov/ar/amlife/people.html	4.68E-04
www.america.gov/ar/econ.html	4.68E-04
www.america.gov/ar/multimedia.html	4.68E-04

A more detailed breakdown of the score of the top 10 pages using *ArabicAnalyzer* and *NutchDocumentAnalyzer* is shown in Appendices A through F.

E. CHAPTER SUMMARY

In this chapter, the results of several experiments to compare the performance of *ArabicAnalyzer* and *NutchDocumentAnalyzer* were described. In the next chapter, the thesis summary and future work recommendations are discussed.

VI. CONCLUSION

A. SUMMARY

Arabic IR is a challenging problem because of the complexity of Arabic languages. Even though the light stemmer algorithm was not a perfect solution for Arabic IR problem, it showed improvement over other popular methods. The ArabicAnalyzer plug-in inherited the same strengths and weaknesses of the light stemmer algorithm. It also was not perfect, but it did show great promise in saving storage overhead.

The experiments completed in this thesis showed that there are advantages and disadvantages when implementing the *ArabicAnalyzer* plug-in. It is clear by looking at the data that, in general, the *ArabicAnalyzer* plug-in performed as well as the default setting. The query results were relevant to the search terms. It was observed that the plug-in ran slower than the default setting, but the speed issue could be overlooked since the data that this research was trying to gather did not have to be in real time. On the other hand, the *ArabicAnalyzer* plug-in would require at least 20% less memory for its index database, compared with the default setting: the savings in storage could become a major plus when indexing the Internet.

B. FUTURE WORK

For future research, the plug-in needs to be integrated into the social networking tool and experiments need to be conducted to determine the recall, precision and relevance of the plug-in in the integration environment. The experiments should also help determine the strengths and weaknesses of the plug-in in such environments and recommend improvement.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A

This is the detail score for query of top 10 pages using *ArabicAnalyzer*.

Search Term: **اقتصاد** (economy)

Page 1:

- boost = 0.22821301
- digest = 767d250a62c827c2bd330c0674546358
- lang = ar
- segment = 20100305180909
- title = الاقتصاد - الاقتصاد - America.gov
- tstamp = 20100305230954510
- url = http://www.america.gov/ar/econ.html

score for query: اقتصاد

- 0.3486507 = (MATCH) sum of:
  - 0.18338637 = (MATCH) weight(anchor:اقتصاد^2.0 in 15), product of:
    - 0.2879631 = queryWeight(anchor:اقتصاد^2.0), product of:
      - 2.0 = boost
      - 4.075775 = idf(docFreq=5, numDocs=130)
      - 0.035326175 = queryNorm
    - 0.63683987 = (MATCH) fieldWeight(anchor:اقتصاد in 15), product of:
      - 1.0 = tf(termFreq(anchor:اقتصاد)=1)
      - 4.075775 = idf(docFreq=5, numDocs=130)
      - 0.15625 = fieldNorm(field=anchor, doc=15)
  - 6.6904654E-4 = (MATCH) weight(content:اقتصاد in 15), product of:
    - 0.03756986 = queryWeight(content:اقتصاد), product of:
      - 1.0635134 = idf(docFreq=121, numDocs=130)

        - 0.035326175 = queryNorm
    - 0.017808065 = (MATCH) fieldWeight(content:اقتصاد in 15), product of:
        - 2.4494898 = tf(termFreq(content:اقتصاد)=6)
        - 1.0635134 = idf(docFreq=121, numDocs=130)
        - 0.0068359375 = fieldNorm(field=content, doc=15)
- 0.16459529 = (MATCH) weight(title:اقتصاد^1.5 in 15), product of:
    - 0.23745762 = queryWeight(title:اقتصاد^1.5), product of:
        - 1.5 = boost
        - 4.4812403 = idf(docFreq=3, numDocs=130)
        - 0.035326175 = queryNorm
    - 0.6931565 = (MATCH) fieldWeight(title:اقتصاد in 15), product of:
        - 1.4142135 = tf(termFreq(title:اقتصاد)=2)
        - 4.4812403 = idf(docFreq=3, numDocs=130)
        - 0.109375 = fieldNorm(field=title, doc=15)

*****

Page 2:

- boost = 0.16124225
- digest = fdaa17fd08dfde3bb91a83a6d98afa04
- lang = ar
- segment = 20100305180909
- title = موجز اقتصاد امريكا - Outline of the U.S. Economy - America.gov
- tstamp = 20100305230918398
- url = http://www.america.gov/ar/publications/books/outline-of-the-us-economy.html

score for query: اقتصاد

- 0.12422927 = (MATCH) sum of:

- 0.07335455 = (MATCH) weight(anchor:اقتصاد^2.0 in 84), product of:
  - 0.2879631 = queryWeight(anchor:اقتصاد^2.0), product of:
    - 2.0 = boost
    - 4.075775 = idf(docFreq=5, numDocs=130)
    - 0.035326175 = queryNorm
  - 0.25473595 = (MATCH) fieldWeight(anchor:اقتصاد in 84), product of:
    - 1.0 = tf(termFreq(anchor:اقتصاد)=1)
    - 4.075775 = idf(docFreq=5, numDocs=130)
    - 0.0625 = fieldNorm(field=anchor, doc=84)
- 9.948079E-4 = (MATCH) weight(content:اقتصاد in 84), product of:
  - 0.03756986 = queryWeight(content:اقتصاد), product of:
    - 1.0635134 = idf(docFreq=121, numDocs=130)
    - 0.035326175 = queryNorm
  - 0.026478883 = (MATCH) fieldWeight(content:اقتصاد in 84), product of:
    - 5.0990195 = tf(termFreq(content:اقتصاد)=26)
    - 1.0635134 = idf(docFreq=121, numDocs=130)
    - 0.0048828125 = fieldNorm(field=content, doc=84)
- 0.049879905 = (MATCH) weight(title:اقتصاد^1.5 in 84), product of:
  - 0.23745762 = queryWeight(title:اقتصاد^1.5), product of:
    - 1.5 = boost
    - 4.4812403 = idf(docFreq=3, numDocs=130)
    - 0.035326175 = queryNorm
  - 0.21005814 = (MATCH) fieldWeight(title:اقتصاد in 84), product of:
    - 1.0 = tf(termFreq(title:اقتصاد)=1)
    - 4.4812403 = idf(docFreq=3, numDocs=130)

- 0.046875 = fieldNorm(field=title, doc=84)

*****

Page 3:

boost = 0.16781548

- digest = b4649130898e202ca38ef61b6b22b917
- lang = ar
- segment = 20100305180909
- title = الأعمال والتجارة - الأعمال والتجارة - America.gov
- tstamp = 20100305231006880
- url = http://www.america.gov/ar/econ/business.html

score for query: اقتصاد

- 0.09217107 = (MATCH) sum of:
  - 0.091693185 = (MATCH) weight(anchor:اقتصاد^2.0 in 16), product of:
    - 0.2879631 = queryWeight(anchor:اقتصاد^2.0), product of:
      - 2.0 = boost
      - 4.075775 = idf(docFreq=5, numDocs=130)
      - 0.035326175 = queryNorm
    - 0.31841993 = (MATCH) fieldWeight(anchor:اقتصاد in 16), product of:
      - 1.0 = tf(termFreq(anchor:اقتصاد)=1)
      - 4.075775 = idf(docFreq=5, numDocs=130)
      - 0.078125 = fieldNorm(field=anchor, doc=16)
  - 4.7789037E-4 = (MATCH) weight(content:اقتصاد in 16), product of:
    - 0.03756986 = queryWeight(content:اقتصاد), product of:
      - 1.0635134 = idf(docFreq=121, numDocs=130)
      - 0.035326175 = queryNorm

        - 0.012720046 = (MATCH) fieldWeight(content:اقتصاد in 16), product of:
            - 2.4494898 = tf(termFreq(content:اقتصاد)=6)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.0048828125 = fieldNorm(field=content, doc=16)

*****

Page 4:

- boost = 0.030659562
- digest = 4304d87a1d51187c1c1d0b2b4d1597a8
- lang = ar
- segment = 20100305181031
- title = إنعاش اقتصاد التجارة - إنعاش اقتصاد التجارة - America.gov
- tstamp = 20100305231127141
- url = http://www.america.gov/ar/reviving_trade_ar.html

score for query: اقتصاد

- 0.033118278 = (MATCH) sum of:
    - 0.018338637 = (MATCH) weight(anchor:اقتصاد^2.0 in 104), product of:
        - 0.2879631 = queryWeight(anchor:اقتصاد^2.0), product of:
            - 2.0 = boost
            - 4.075775 = idf(docFreq=5, numDocs=130)
            - 0.035326175 = queryNorm
        - 0.06368399 = (MATCH) fieldWeight(anchor:اقتصاد in 104), product of:
            - 1.0 = tf(termFreq(anchor:اقتصاد)=1)
            - 4.075775 = idf(docFreq=5, numDocs=130)
            - 0.015625 = fieldNorm(field=anchor, doc=104)
    - 8.363082E-5 = (MATCH) weight(content:اقتصاد in 104), product of:

        - 0.03756986 = queryWeight(content:اقتصاد), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035326175 = queryNorm
        - 0.002226008 = (MATCH) fieldWeight(content:اقتصاد in 104), product of:
            - 2.4494898 = tf(termFreq(content:اقتصاد)=6)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 8.544922E-4 = fieldNorm(field=content, doc=104)
    - 0.014696008 = (MATCH) weight(title:اقتصاد^1.5 in 104), product of:
        - 0.23745762 = queryWeight(title:اقتصاد^1.5), product of:
            - 1.5 = boost
            - 4.4812403 = idf(docFreq=3, numDocs=130)
            - 0.035326175 = queryNorm
        - 0.06188897 = (MATCH) fieldWeight(title:اقتصاد in 104), product of:
            - 1.4142135 = tf(termFreq(title:اقتصاد)=2)
            - 4.4812403 = idf(docFreq=3, numDocs=130)
            - 0.009765625 = fieldNorm(field=title, doc=104)

*****

Page 5:

- boost = 0.02675021
- digest = aaf055c1e690c63cf69285f8ab04f499
- lang = ar
- segment = 20100305181330
- title = بتک - بتک - America.gov
- tstamp = 20100305231345369
- url = http://www.america.gov/ar/publications/books.html#outline_economy

score for query: اقتصاد

- 0.016127191 = (MATCH) sum of:
    - 0.016046308 = (MATCH) weight(anchor:اقتصاد^2.0 in 80), product of:
        - 0.2879631 = queryWeight(anchor:اقتصاد^2.0), product of:
            - 2.0 = boost
            - 4.075775 = idf(docFreq=5, numDocs=130)
            - 0.035326175 = queryNorm
        - 0.05572349 = (MATCH) fieldWeight(anchor:اقتصاد in 80), product of:
            - 1.0 = tf(termFreq(anchor:اقتصاد)=1)
            - 4.075775 = idf(docFreq=5, numDocs=130)
            - 0.013671875 = fieldNorm(field=anchor, doc=80)
    - 8.088332E-5 = (MATCH) weight(content:اقتصاد in 80), product of:
        - 0.03756986 = queryWeight(content:اقتصاد), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035326175 = queryNorm
        - 0.0021528779 = (MATCH) fieldWeight(content:اقتصاد in 80), product of:
            - 3.3166249 = tf(termFreq(content:اقتصاد)=11)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 6.1035156E-4 = fieldNorm(field=content, doc=80)

*****

Page 6:

- boost = 1.0000145
- digest = 0d5b023c802941ddb358071073a98833
- lang = ar
- segment = 20100305180856

- title = الصفحة الألولى - الصفحة الألولى - America.gov
- tstamp = 20100305230902835
- url = http://www.america.gov/ar/

score for query: اقتصاد

- 0.0030584983 = (MATCH) sum of:
    - 0.0030584983 = (MATCH) weight(content:اقتصاد in 0), product of:
        - 0.03756986 = queryWeight(content:اقتصاد), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035326175 = queryNorm
        - 0.08140829 = (MATCH) fieldWeight(content:اقتصاد in 0), product of:
            - 2.4494898 = tf(termFreq(content:اقتصاد)=6)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.03125 = fieldNorm(field=content, doc=0)

*****

Page 7:

- boost = 0.23009512
- digest = 15d9ca5e7382f701cd03fb542ae3ab22
- lang = ar
- segment = 20100305180909
- title = أبلوبوم روص - الحاكية ألامريكية بالروص - America.gov
- tstamp = 20100305230915350
- url = http://www.america.gov/ar/multimedia/photogallery.html

score for query: اقتصاد

- 6.6904654E-4 = (MATCH) sum of:
    - 6.6904654E-4 = (MATCH) weight(content:اقتصاد in 38), product of:
        - 0.03756986 = queryWeight(content:اقتصاد), product of:

        - 1.0635134 = idf(docFreq=121, numDocs=130)
        - 0.035326175 = queryNorm
    - 0.017808065 = (MATCH) fieldWeight(content:اقتصاد in 38), product of:
        - 2.4494898 = tf(termFreq(content:اقتصاد)=6)
        - 1.0635134 = idf(docFreq=121, numDocs=130)
        - 0.0068359375 = fieldNorm(field=content, doc=38)

*****

Page 8:

- boost = 0.22996004
- digest = a0130240b4348578aa8a83e59187dfb3
- lang = ar
- segment = 20100305180909
- title = بتک - بتک - America.gov
- tstamp = 20100305231001279
- url = http://www.america.gov/ar/publications/books.html

score for query: اقتصاد

- 6.4706657E-4 = (MATCH) sum of:
    - 6.4706657E-4 = (MATCH) weight(content:اقتصاد in 73), product of:
        - 0.03756986 = queryWeight(content:اقتصاد), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035326175 = queryNorm
        - 0.017223023 = (MATCH) fieldWeight(content:اقتصاد in 73), product of:
            - 3.3166249 = tf(termFreq(content:اقتصاد)=11)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.0048828125 = fieldNorm(field=content, doc=73)

*****

Page 9:

- boost = 0.16516872
- digest = bc202e5e0f508e4291bb897eec7814dc
- lang = ar
- segment = 20100305180909
- title = الحكم والنمو - 1209 - America.gov
- tstamp = 20100305230941328
- url = http://www.america.gov/ar/publications/ejournalusa/1209.html

score for query: اقتصاد

- 5.8529375E-4 = (MATCH) sum of:
    - 5.8529375E-4 = (MATCH) weight(content:اقتصاد in 97), product of:
        - 0.03756986 = queryWeight(content:اقتصاد), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035326175 = queryNorm
        - 0.01557881 = (MATCH) fieldWeight(content:اقتصاد in 97), product of:
            - 3.0 = tf(termFreq(content:اقتصاد)=9)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.0048828125 = fieldNorm(field=content, doc=97)

*****

Page 10:

- boost = 0.19712433
- digest = c25a22a11ab6bec420c26625155ced62
- lang = ar
- segment = 20100305180909
- title = الصفحة الأولى - الصفحة الأولى - America.gov

- tstamp = 20100305230929458
- url = http://www.america.gov/ar/index.html

score for query: اقتصاد

- 5.7346845E-4 = (MATCH) sum of:
    - 5.7346845E-4 = (MATCH) weight(content:اقتصاد in 30), product of:
        - 0.03756986 = queryWeight(content:اقتصاد), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035326175 = queryNorm
        - 0.015264055 = (MATCH) fieldWeight(content:اقتصاد in 30), product of:
            - 2.4494898 = tf(termFreq(content:اقتصاد)=6)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.005859375 = fieldNorm(field=content, doc=30)

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B

This is the detail score for query of top 10 pages using *NutchDocumentAnalyzer*.

Search Term: اقتصاد (ecomony)

Page 1:

- boost = 0.22826105
- digest = c33a5dc3f7d8475491bfafcf91c8b283
- lang = ar
- segment = 20100307101102
- title = اقتصاد - اقتصاد - America.gov
- tstamp = 20100307151153574
- url = http://www.america.gov/ar/econ.html

score for query: اقتصاد

- 0.38501537 = (MATCH) sum of:
  - 0.1991137 = (MATCH) weight(anchor:اقتصاد^2.0 in 16), product of:
    - 0.29873407 = queryWeight(anchor:اقتصاد^2.0), product of:
      - 2.0 = boost
      - 4.2657595 = idf(docFreq=4, numDocs=131)
      - 0.035015345 = queryNorm
    - 0.6665249 = (MATCH) fieldWeight(anchor:اقتصاد in 16), product of:
      - 1.0 = tf(termFreq(anchor:اقتصاد)=1)
      - 4.2657595 = idf(docFreq=4, numDocs=131)
      - 0.15625 = fieldNorm(field=anchor, doc=16)
  - 5.40958E-4 = (MATCH) weight(content:اقتصاد in 16), product of:
    - 0.037221763 = queryWeight(content:اقتصاد), product of:
      - 1.063013 = idf(docFreq=122, numDocs=131)

        - 0.035015345 = queryNorm
    - 0.01453338 = (MATCH) fieldWeight(content:اقتصاد in 16), product of:
        - 2.0 = tf(termFreq(content:اقتصاد)=4)
        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.0068359375 = fieldNorm(field=content, doc=16)
- 0.18536073 = (MATCH) weight(title:اقتصاد^1.5 in 16), product of:
    - 0.25088066 = queryWeight(title:اقتصاد^1.5), product of:
        - 1.5 = boost
        - 4.776585 = idf(docFreq=2, numDocs=131)
        - 0.035015345 = queryNorm
    - 0.7388403 = (MATCH) fieldWeight(title:اقتصاد in 16), product of:
        - 1.4142135 = tf(termFreq(title:اقتصاد)=2)
        - 4.776585 = idf(docFreq=2, numDocs=131)
        - 0.109375 = fieldNorm(field=title, doc=16)

*****

Page 2:

- boost = 0.16124225
- digest = 6120d6b7e6584b6a71b7d9990a68b952
- lang = ar
- segment = 20100307101102
- title = موجز اقتصاد امريكي - Outline of the U.S. Economy - America.gov
- tstamp = 20100307151112494
- url = http://www.america.gov/ar/publications/books/outline-of-the-us-economy.html

score for query: اقتصاد

- 0.13663794 = (MATCH) sum of:
    - 0.07964548 = (MATCH) weight(anchor:اقتصاد^2.0 in 85), product of:

- 0.29873407 = queryWeight(anchor:اقتصاد^2.0), product of:
  - 2.0 = boost
  - 4.2657595 = idf(docFreq=4, numDocs=131)
  - 0.035015345 = queryNorm
- 0.26660997 = (MATCH) fieldWeight(anchor:اقتصاد in 85), product of:
  - 1.0 = tf(termFreq(anchor:اقتصاد)=1)
  - 4.2657595 = idf(docFreq=4, numDocs=131)
  - 0.0625 = fieldNorm(field=anchor, doc=85)

o 8.196751E-4 = (MATCH) weight(content:اقتصاد in 85), product of:

- 0.037221763 = queryWeight(content:اقتصاد), product of:
  - 1.063013 = idf(docFreq=122, numDocs=131)
  - 0.035015345 = queryNorm
- 0.022021394 = (MATCH) fieldWeight(content:اقتصاد in 85), product of:
  - 4.2426405 = tf(termFreq(content:اقتصاد)=18)
  - 1.063013 = idf(docFreq=122, numDocs=131)
  - 0.0048828125 = fieldNorm(field=content, doc=85)

o 0.056172792 = (MATCH) weight(title:اقتصاد^1.5 in 85), product of:

- 0.25088066 = queryWeight(title:اقتصاد^1.5), product of:
  - 1.5 = boost
  - 4.776585 = idf(docFreq=2, numDocs=131)
  - 0.035015345 = queryNorm
- 0.22390243 = (MATCH) fieldWeight(title:اقتصاد in 85), product of:
  - 1.0 = tf(termFreq(title:اقتصاد)=1)
  - 4.776585 = idf(docFreq=2, numDocs=131)
  - 0.046875 = fieldNorm(field=title, doc=85)

*****

Page 3:

- boost = 0.16784814
- digest = 2e923bcfb9409e9be88aad90198063bc
- lang = ar
- segment = 20100307101102
- title = الأعمال والتجارة - الأعمال والتجارة - America.gov
- tstamp = 20100307151205095
- url = http://www.america.gov/ar/econ/business.html

score for query: الاقتصاد

- 0.09989148 = (MATCH) sum of:
  - 0.09955685 = (MATCH) weight(anchor:الاقتصاد^2.0 in 17), product of:
    - 0.29873407 = queryWeight(anchor:الاقتصاد^2.0), product of:
      - 2.0 = boost
      - 4.2657595 = idf(docFreq=4, numDocs=131)
      - 0.035015345 = queryNorm
    - 0.33326244 = (MATCH) fieldWeight(anchor:الاقتصاد in 17), product of:
      - 1.0 = tf(termFreq(anchor:الاقتصاد)=1)
      - 4.2657595 = idf(docFreq=4, numDocs=131)
      - 0.078125 = fieldNorm(field=anchor, doc=17)
  - 3.34631E-4 = (MATCH) weight(content:الاقتصاد in 17), product of:
    - 0.037221763 = queryWeight(content:الاقتصاد), product of:
      - 1.063013 = idf(docFreq=122, numDocs=131)
      - 0.035015345 = queryNorm
    - 0.008990197 = (MATCH) fieldWeight(content:الاقتصاد in 17), product of:
      - 1.7320508 = tf(termFreq(content:الاقتصاد)=3)

        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.0048828125 = fieldNorm(field=content, doc=17)

*****

Page 4:

- boost = 0.02675021
- digest = 4eb9183dbdc405b0d40ef3c92da5ed66
- lang = ar
- segment = 20100307101458
- title = بتک - بتک - America.gov
- tstamp = 20100307151513307
- url = http://www.america.gov/ar/publications/books.html#outline_economy

score for query: الاقتصاد

- 0.01747075 = (MATCH) sum of:
    - 0.017422449 = (MATCH) weight(anchor:الاقتصاد^2.0 in 81), product of:
        - 0.29873407 = queryWeight(anchor:الاقتصاد^2.0), product of:
            - 2.0 = boost
            - 4.2657595 = idf(docFreq=4, numDocs=131)
            - 0.035015345 = queryNorm
        - 0.058320932 = (MATCH) fieldWeight(anchor:الاقتصاد in 81), product of:
            - 1.0 = tf(termFreq(anchor:الاقتصاد)=1)
            - 4.2657595 = idf(docFreq=4, numDocs=131)
            - 0.013671875 = fieldNorm(field=anchor, doc=81)
    - 4.8299826E-5 = (MATCH) weight(content:الاقتصاد in 81), product of:
        - 0.037221763 = queryWeight(content:الاقتصاد), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.035015345 = queryNorm

        - 0.0012976233 = (MATCH) fieldWeight(content:اقتصاد in 81), product of:
            - 2.0 = tf(termFreq(content:اقتصاد)=4)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 6.1035156E-4 = fieldNorm(field=content, doc=81)

*****

Page 5:

- boost = 1.0000145
- digest = eed4dd9817b50ffda0aef158be6e4c12
- lang = ar
- segment = 20100307101052
- title = الصفحة الأولى - الصفحة الأولى - America.gov
- tstamp = 20100307151057483
- url = http://www.america.gov/ar/

score for query: اقتصاد

- 0.002472951 = (MATCH) sum of:
    - 0.002472951 = (MATCH) weight(content:اقتصاد in 0), product of:
        - 0.037221763 = queryWeight(content:اقتصاد), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.035015345 = queryNorm
        - 0.06643831 = (MATCH) fieldWeight(content:اقتصاد in 0), product of:
            - 2.0 = tf(termFreq(content:اقتصاد)=4)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.03125 = fieldNorm(field=content, doc=0)

*****

Page 6:

- boost = 0.23014276
- digest = 5f50883579dcc0acb85ff3052764f758

- lang = ar
- segment = 20100307101102
- title = أبلوبم روص - الحاكيات الأمريكيات باب روص - America.gov
- tstamp = 20100307151109851
- url = http://www.america.gov/ar/multimedia/photogallery.html

score for query: الاقتصاد

- 5.40958E-4 = (MATCH) sum of:
  - 5.40958E-4 = (MATCH) weight(content:الاقتصاد in 39), product of:
    - 0.037221763 = queryWeight(content:الاقتصاد), product of:
      - 1.063013 = idf(docFreq=122, numDocs=131)
      - 0.035015345 = queryNorm
    - 0.01453338 = (MATCH) fieldWeight(content:الاقتصاد in 39), product of:
      - 2.0 = tf(termFreq(content:الاقتصاد)=4)
      - 1.063013 = idf(docFreq=122, numDocs=131)
      - 0.0068359375 = fieldNorm(field=content, doc=39)

*****

Page 7:

- boost = 0.19715214
- digest = e84ec632a6d47466f40d6beacbcfbdf7
- lang = ar
- segment = 20100307101102
- title = الصفحة الأولى - الصفحة الأولى - America.gov
- tstamp = 20100307151123237
- url = http://www.america.gov/ar/index.html

score for query: الاقتصاد

- 4.636783E-4 = (MATCH) sum of:
  - 4.636783E-4 = (MATCH) weight(content:الاقتصاد in 31), product of:

        - 0.037221763 = queryWeight(content:داصتقالا), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.035015345 = queryNorm
        - 0.012457183 = (MATCH) fieldWeight(content:داصتقالا in 31), product of:
            - 2.0 = tf(termFreq(content:داصتقالا)=4)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.005859375 = fieldNorm(field=content, doc=31)

*****

Page 8:

- boost = 0.20153543
- digest = 67da72f899f80475f1ae62770921f1bd
- lang = ar
- segment = 20100307101102
- title = ايسارواو ابوروأ - ايسارواو ابوروأ - America.gov
- tstamp = 20100307151127260
- url = http://www.america.gov/ar/world/europe.html

score for query: داصتقالا

- 4.636783E-4 = (MATCH) sum of:
    - 4.636783E-4 = (MATCH) weight(content:داصتقالا in 128), product of:
        - 0.037221763 = queryWeight(content:داصتقالا), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.035015345 = queryNorm
        - 0.012457183 = (MATCH) fieldWeight(content:داصتقالا in 128), product of:
            - 2.0 = tf(termFreq(content:داصتقالا)=4)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.005859375 = fieldNorm(field=content, doc=128)

*****

Page 9:

- boost = 0.20091416
- digest = 9256cf74ef9b595d81f726d5f347898a
- lang = ar
- segment = 20100307101102
- title = الشرق الأوسط وشمال أفريقيا - الشرق الأوسط وشمال أفريقيا - America.gov
- tstamp = 20100307151148461
- url = http://www.america.gov/ar/world/mideast.html

score for query: اقتصاد

- 4.636783E-4 = (MATCH) sum of:
    - 4.636783E-4 = (MATCH) weight(content:اقتصاد in 129), product of:
        - 0.037221763 = queryWeight(content:اقتصاد), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.035015345 = queryNorm
        - 0.012457183 = (MATCH) fieldWeight(content:اقتصاد in 129), product of:
            - 2.0 = tf(termFreq(content:اقتصاد)=4)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.005859375 = fieldNorm(field=content, doc=129)

*****

Page 10:

- boost = 0.20039715
- digest = adbc4a97340b57bcc256c62131041c4c
- lang = ar
- segment = 20100307101102
- title = بونج وشرق آسيا - بونج وشرق آسيا - America.gov
- tstamp = 20100307151119356

- url = http://www.america.gov/ar/world/scasia.html

score for query: داصتقالا

- 4.015572E-4 = (MATCH) sum of:
    - 4.015572E-4 = (MATCH) weight(content:داصتقالا in 130), product of:
        - 0.037221763 = queryWeight(content:داصتقالا), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.035015345 = queryNorm
        - 0.010788237 = (MATCH) fieldWeight(content:داصتقالا in 130), product of:
            - 1.7320508 = tf(termFreq(content:داصتقالا)=3)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.005859375 = fieldNorm(field=content, doc=130)

APPENDIX C

This is the detail score for query of top 10 pages using *ArabicAnalyzer*.

Search Term: **الكريما** (The United States)

Page 1:

- boost = 0.15805063
- digest = 6b6baa67bd29d99a3cf293efeb2bc3e1
- lang = ar
- segment = 20100305181031
- title = عقوم لوح - فوغ تود الكريمأ عقوم لوح - فوغ تود الكريمأ - America.gov
- tstamp = 20100305231050378
- url = http://www.america.gov/ar/pages/footer/local/about-us.html

score for query: الكريما

- 0.1196895 = (MATCH) sum of:
  - 0.059957497 = (MATCH) weight(anchor:الكريما^2.0 in 68), product of:
    - 0.261373 = queryWeight(anchor:الكريما^2.0), product of:
      - 2.0 = boost
      - 3.6703098 = idf(docFreq=8, numDocs=130)
      - 0.035606395 = queryNorm
    - 0.22939436 = (MATCH) fieldWeight(anchor:الكريما in 68), product of:
      - 1.0 = tf(termFreq(anchor:الكريما)=1)
      - 3.6703098 = idf(docFreq=8, numDocs=130)
      - 0.0625 = fieldNorm(field=anchor, doc=68)
  - 4.8168114E-4 = (MATCH) weight(content:الكريما in 68), product of:
    - 0.037867878 = queryWeight(content:الكريما), product of:
      - 1.0635134 = idf(docFreq=121, numDocs=130)

        - 0.035606395 = queryNorm
    - 0.012720046 = (MATCH) fieldWeight(content:الكريما in 68), product of:
        - 2.4494898 = tf(termFreq(content:الكريما)=6)
        - 1.0635134 = idf(docFreq=121, numDocs=130)
        - 0.0048828125 = fieldNorm(field=content, doc=68)
  - 0.059250325 = (MATCH) weight(title:الكريما^1.5 in 68), product of:
    - 0.23934121 = queryWeight(title:الكريما^1.5), product of:
        - 1.5 = boost
        - 4.4812403 = idf(docFreq=3, numDocs=130)
        - 0.035606395 = queryNorm
    - 0.24755588 = (MATCH) fieldWeight(title:الكريما in 68), product of:
        - 1.4142135 = tf(termFreq(title:الكريما)=2)
        - 4.4812403 = idf(docFreq=3, numDocs=130)
        - 0.0390625 = fieldNorm(field=title, doc=68)

Page 2:

- boost = 0.16184442
- digest = 0f454ab63865ae2e08003bb23896bfad
- lang = ar
- segment = 20100305180909
- title = المسلمون في أمريكا - Being Muslim in America - America.gov
- tstamp = 20100305231009575
- url = http://www.america.gov/ar/publications/books-content/musliminamerica.html

score for query: الكريما

- 0.11078926 = (MATCH) sum of:
  - 0.059957497 = (MATCH) weight(anchor:الكريما^2.0 in 72), product of:

- 0.261373 = queryWeight(anchor:الكريم^2.0), product of:
  - 2.0 = boost
  - 3.6703098 = idf(docFreq=8, numDocs=130)
  - 0.035606395 = queryNorm
- 0.22939436 = (MATCH) fieldWeight(anchor:الكريم in 72), product of:
  - 1.0 = tf(termFreq(anchor:الكريم)=1)
  - 3.6703098 = idf(docFreq=8, numDocs=130)
  - 0.0625 = fieldNorm(field=anchor, doc=72)

o 5.5619737E-4 = (MATCH) weight(content:الكريم in 72), product of:

- 0.037867878 = queryWeight(content:الكريم), product of:
  - 1.0635134 = idf(docFreq=121, numDocs=130)
  - 0.035606395 = queryNorm
- 0.014687842 = (MATCH) fieldWeight(content:الكريم in 72), product of:
  - 2.828427 = tf(termFreq(content:الكريم)=8)
  - 1.0635134 = idf(docFreq=121, numDocs=130)
  - 0.0048828125 = fieldNorm(field=content, doc=72)

o 0.050275568 = (MATCH) weight(title:الكريم^1.5 in 72), product of:

- 0.23934121 = queryWeight(title:الكريم^1.5), product of:
  - 1.5 = boost
  - 4.4812403 = idf(docFreq=3, numDocs=130)
  - 0.035606395 = queryNorm
- 0.21005814 = (MATCH) fieldWeight(title:الكريم in 72), product of:
  - 1.0 = tf(termFreq(title:الكريم)=1)
  - 4.4812403 = idf(docFreq=3, numDocs=130)
  - 0.046875 = fieldNorm(field=title, doc=72)

*****

Page 3:

- boost = 0.23032264
- digest = ce4a12d589c1a56e886d5b6848609391
- lang = ar
- segment = 20100305180909
- title = الحياة الأمريكية - الحياة الأمريكية - America.gov
- tstamp = 20100305230939904
- url = http://www.america.gov/ar/amlife.html

score for query: الكريما

- 0.105654 = (MATCH) sum of:
  - 0.10492562 = (MATCH) weight(anchor:الكريما^2.0 in 3), product of:
    - 0.261373 = queryWeight(anchor:الكريما^2.0), product of:
      - 2.0 = boost
      - 3.6703098 = idf(docFreq=8, numDocs=130)
      - 0.035606395 = queryNorm
    - 0.40144014 = (MATCH) fieldWeight(anchor:الكريما in 3), product of:
      - 1.0 = tf(termFreq(anchor:الكريما)=1)
      - 3.6703098 = idf(docFreq=8, numDocs=130)
      - 0.109375 = fieldNorm(field=anchor, doc=3)
  - 7.283851E-4 = (MATCH) weight(content:الكريما in 3), product of:
    - 0.037867878 = queryWeight(content:الكريما), product of:
      - 1.0635134 = idf(docFreq=121, numDocs=130)
      - 0.035606395 = queryNorm
    - 0.019234907 = (MATCH) fieldWeight(content:الكريما in 3), product of:
      - 2.6457512 = tf(termFreq(content:الكريما)=7)

        - 1.0635134 = idf(docFreq=121, numDocs=130)
        - 0.0068359375 = fieldNorm(field=content, doc=3)

*****

Page 4:

- boost = 0.15872316
- digest = dcfeb490d3db633d16bfb0588d67076d
- lang = ar
- segment = 20100305180909
- title = ربع نشيديإ لي‌ابوم فوغ تود الكريمأ لاوجلا ةخسن - نشيديإ لي‌ابوم ةمدخ ةزهجأ PDA - America.gov
- tstamp = 20100305230942596
- url = http://www.america.gov/ar/services/mobile.html

score for query: الكريما

- 0.042377986 = (MATCH) sum of:
    - 4.8168114E-4 = (MATCH) weight(content:الكريما in 114), product of:
        - 0.037867878 = queryWeight(content:الكريما), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035606395 = queryNorm
        - 0.012720046 = (MATCH) fieldWeight(content:الكريما in 114), product of:
            - 2.4494898 = tf(termFreq(content:الكريما)=6)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.0048828125 = fieldNorm(field=content, doc=114)
    - 0.041896306 = (MATCH) weight(title:الكريما^1.5 in 114), product of:
        - 0.23934121 = queryWeight(title:الكريما^1.5), product of:
            - 1.5 = boost
            - 4.4812403 = idf(docFreq=3, numDocs=130)
            - 0.035606395 = queryNorm

        - 0.17504844 = (MATCH) fieldWeight(title:الكريما in 114), product of:
            - 1.0 = tf(termFreq(title:الكريما)=1)
            - 4.4812403 = idf(docFreq=3, numDocs=130)
            - 0.0390625 = fieldNorm(field=title, doc=114)

*****

Page 5:

- boost = 0.04832446
- digest = 87c8a44e7bc9cb3221f6823da385f8dd
- lang = ar
- segment = 20100305181031
- title = أبلوبما روص - الحاكيةالأمريكيةالباب روص - America.gov
- tstamp = 20100305231143209
- url = http://www.america.gov/ar/multimedia/photogallery.html#/4110/mosques_ar/

score for query: الكريما

- 0.022628564 = (MATCH) sum of:
    - 0.02248406 = (MATCH) weight(anchor:الكريما^2.0 in 50), product of:
        - 0.261373 = queryWeight(anchor:الكريما^2.0), product of:
            - 2.0 = boost
            - 3.6703098 = idf(docFreq=8, numDocs=130)
            - 0.035606395 = queryNorm
        - 0.08602288 = (MATCH) fieldWeight(anchor:الكريما in 50), product of:
            - 1.0 = tf(termFreq(anchor:الكريما)=1)
            - 3.6703098 = idf(docFreq=8, numDocs=130)
            - 0.0234375 = fieldNorm(field=anchor, doc=50)
    - 1.4450435E-4 = (MATCH) weight(content:الكريما in 50), product of:

        - 0.037867878 = queryWeight(content:الكريما), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035606395 = queryNorm
        - 0.0038160137 = (MATCH) fieldWeight(content:الكريما in 50), product of:
            - 2.4494898 = tf(termFreq(content:الكريما)=6)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.0014648438 = fieldNorm(field=content, doc=50)

*****

Page 6:

- boost = 0.033444975
- digest = 7212084a79cd19adbfc07dc50d3c0ea4
- lang = ar
- segment = 20100305181031
- title = بتک - بتک - America.gov
- tstamp = 20100305231138931
- url = http://www.america.gov/ar/publications/books.html#beingmuslim

score for query: الكريما

- 0.015091554 = (MATCH) sum of:
    - 0.014989374 = (MATCH) weight(anchor:الكريما^2.0 in 75), product of:
        - 0.261373 = queryWeight(anchor:الكريما^2.0), product of:
            - 2.0 = boost
            - 3.6703098 = idf(docFreq=8, numDocs=130)
            - 0.035606395 = queryNorm
        - 0.05734859 = (MATCH) fieldWeight(anchor:الكريما in 75), product of:
            - 1.0 = tf(termFreq(anchor:الكريما)=1)
            - 3.6703098 = idf(docFreq=8, numDocs=130)

        - 0.015625 = fieldNorm(field=anchor, doc=75)
    - 1.0217999E-4 = (MATCH) weight(content:الكريما in 75), product of:
        - 0.037867878 = queryWeight(content:الكريما), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035606395 = queryNorm
        - 0.002698329 = (MATCH) fieldWeight(content:الكريما in 75), product of:
            - 3.4641016 = tf(termFreq(content:الكريما)=12)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 7.324219E-4 = fieldNorm(field=content, doc=75)

*****

Page 7:

- boost = 0.0420541
- digest = 3354b6239b6eb27b9d241073f88fc34e
- lang = ar
- segment = 20100305181031
- title = أبل موبوص - الحاكية ألمريكية اب لوصل - America.gov
- tstamp = 20100305231110244
- url = http://www.america.gov/ar/multimedia/photogallery.html#/4110/religious_freedom_ar/

score for query: الكريما

- 0.01136245 = (MATCH) sum of:
    - 0.01124203 = (MATCH) weight(anchor:الكريما^2.0 in 52), product of:
        - 0.261373 = queryWeight(anchor:الكريما^2.0), product of:
            - 2.0 = boost
            - 3.6703098 = idf(docFreq=8, numDocs=130)
            - 0.035606395 = queryNorm

        - 0.04301144 = (MATCH) fieldWeight(anchor:الكريما in 52), product of:
            - 1.0 = tf(termFreq(anchor:الكريما)=1)
            - 3.6703098 = idf(docFreq=8, numDocs=130)
            - 0.01171875 = fieldNorm(field=anchor, doc=52)
    - 1.20420285E-4 = (MATCH) weight(content:الكريما in 52), product of:
        - 0.037867878 = queryWeight(content:الكريما), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035606395 = queryNorm
        - 0.0031800114 = (MATCH) fieldWeight(content:الكريما in 52), product of:
            - 2.4494898 = tf(termFreq(content:الكريما)=6)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.0012207031 = fieldNorm(field=content, doc=52)

*****

Page 8:

- boost = 0.02675021
- digest = d4493509fb1e3146c2003310c9b70cbd
- lang = ar
- segment = 20100305181330
- title = بتک - بتک - America.gov
- tstamp = 20100305231409034
- url = http://www.america.gov/ar/publications/books.html#governed

score for query: الكريما

- 0.01132718 = (MATCH) sum of:
    - 0.01124203 = (MATCH) weight(anchor:الكريما^2.0 in 77), product of:
        - 0.261373 = queryWeight(anchor:الكريما^2.0), product of:
            - 2.0 = boost

        - 3.6703098 = idf(docFreq=8, numDocs=130)
        - 0.035606395 = queryNorm
      - 0.04301144 = (MATCH) fieldWeight(anchor:الكريما in 77), product of:
        - 1.0 = tf(termFreq(anchor:الكريما)=1)
        - 3.6703098 = idf(docFreq=8, numDocs=130)
        - 0.01171875 = fieldNorm(field=anchor, doc=77)
  - 8.514999E-5 = (MATCH) weight(content:الكريما in 77), product of:
    - 0.037867878 = queryWeight(content:الكريما), product of:
      - 1.0635134 = idf(docFreq=121, numDocs=130)
      - 0.035606395 = queryNorm
    - 0.0022486076 = (MATCH) fieldWeight(content:الكريما in 77), product of:
      - 3.4641016 = tf(termFreq(content:الكريما)=12)
      - 1.0635134 = idf(docFreq=121, numDocs=130)
      - 6.1035156E-4 = fieldNorm(field=content, doc=77)

*****

Page 9:

- boost = 0.025411258
- digest = 6b7361561b7255632af783ca69a88410
- lang = ar
- segment = 20100305181330
- title = أبلوم روص - الحاكية الأمريكية اب روصل - America.gov
- tstamp = 20100305231411435
- url = http://www.america.gov/ar/multimedia/photogallery.html#/4110/islam_ar/

score for query: الكريما

- 0.011314282 = (MATCH) sum of:
  - 0.01124203 = (MATCH) weight(anchor:الكريما^2.0 in 49), product of:

        - 0.261373 = queryWeight(anchor:الكريما^2.0), product of:
            - 2.0 = boost
            - 3.6703098 = idf(docFreq=8, numDocs=130)
            - 0.035606395 = queryNorm
        - 0.04301144 = (MATCH) fieldWeight(anchor:الكريما in 49), product of:
            - 1.0 = tf(termFreq(anchor:الكريما)=1)
            - 3.6703098 = idf(docFreq=8, numDocs=130)
            - 0.01171875 = fieldNorm(field=anchor, doc=49)
    - 7.225217E-5 = (MATCH) weight(content:الكريما in 49), product of:
        - 0.037867878 = queryWeight(content:الكريما), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035606395 = queryNorm
        - 0.0019080068 = (MATCH) fieldWeight(content:الكريما in 49), product of:
            - 2.4494898 = tf(termFreq(content:الكريما)=6)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 7.324219E-4 = fieldNorm(field=content, doc=49)

*****

Page 10:

- boost = 1.0000145
- digest = 0d5b023c802941ddb358071073a98833
- lang = ar
- segment = 20100305180856
- title = الصفحة الأولى - الصفحة الأولى - America.gov
- tstamp = 20100305230902835
- url = http://www.america.gov/ar/

score for query: الكريما

- 0.0030827592 = (MATCH) sum of:
    - 0.0030827592 = (MATCH) weight(content:الكريما in 0), product of:
        - 0.037867878 = queryWeight(content:الكريما), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035606395 = queryNorm
        - 0.08140829 = (MATCH) fieldWeight(content:الكريما in 0), product of:
            - 2.4494898 = tf(termFreq(content:الكريما)=6)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.03125 = fieldNorm(field=content, doc=0)

APPENDIX D

This is the detail score for query of top 10 pages using *NutchDocumentAnalyzer*.

Search Term: **أمريكا** (America)

Page 1:

- boost = 0.1580853
- digest = 65d01f780ed747de9fd07241fb39df44
- lang = ar
- segment = 20100307101231
- title = حول عقوم أمريكا تود توغ فـ - فـوغ تـود الكـريمأ عقوم لـوح - America.gov
- tstamp = 20100307151249455
- url = http://www.america.gov/ar/pages/footer/local/about-us.html

score for query: أمريكا

- 0.11997691 = (MATCH) sum of:
  - 0.060125146 = (MATCH) weight(anchor:أمريكا^2.0 in 69), product of:
    - 0.26155776 = queryWeight(anchor:أمريكا^2.0), product of:
      - 2.0 = boost
      - 3.6779728 = idf(docFreq=8, numDocs=131)
      - 0.035557326 = queryNorm
    - 0.2298733 = (MATCH) fieldWeight(anchor:أمريكا in 69), product of:
      - 1.0 = tf(termFreq(anchor:أمريكا)=1)
      - 3.6779728 = idf(docFreq=8, numDocs=131)
      - 0.0625 = fieldNorm(field=anchor, doc=69)
  - 4.8056475E-4 = (MATCH) weight(content:أمريكا in 69), product of:
    - 0.037797898 = queryWeight(content:أمريكا), product of:
      - 1.063013 = idf(docFreq=122, numDocs=131)

        - 0.035557326 = queryNorm
    - 0.01271406 = (MATCH) fieldWeight(content:أمريكا in 69), product of:
        - 2.4494898 = tf(termFreq(content:أمريكا)=6)
        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.0048828125 = fieldNorm(field=content, doc=69)
- 0.0593712 = (MATCH) weight(title:أمريكا^1.5 in 69), product of:
    - 0.23942009 = queryWeight(title:أمريكا^1.5), product of:
        - 1.5 = boost
        - 4.488903 = idf(docFreq=3, numDocs=131)
        - 0.035557326 = queryNorm
    - 0.2479792 = (MATCH) fieldWeight(title:أمريكا in 69), product of:
        - 1.4142135 = tf(termFreq(title:أمريكا)=2)
        - 4.488903 = idf(docFreq=3, numDocs=131)
        - 0.0390625 = fieldNorm(field=title, doc=69)

*****

Page 2:

- boost = 0.16184442
- digest = be96f39b462a546d99cbfa50ba70c710
- lang = ar
- segment = 20100307101102
- title = المسلمون في أمريكا - Being Muslim in America - America.gov
- tstamp = 20100307151207698
- url = http://www.america.gov/ar/publications/books-content/musliminamerica.html

score for query: أمريكا

- 0.11105819 = (MATCH) sum of:
    - 0.060125146 = (MATCH) weight(anchor:أمريكا^2.0 in 73), product of:

- 0.26155776 = queryWeight(anchor:الكريمأ^2.0), product of:
  - 2.0 = boost
  - 3.6779728 = idf(docFreq=8, numDocs=131)
  - 0.035557326 = queryNorm
- 0.2298733 = (MATCH) fieldWeight(anchor:الكريمأ in 73), product of:
  - 1.0 = tf(termFreq(anchor:الكريمأ)=1)
  - 3.6779728 = idf(docFreq=8, numDocs=131)
  - 0.0625 = fieldNorm(field=anchor, doc=73)

o 5.5490836E-4 = (MATCH) weight(content:الكريمأ in 73), product of:

- 0.037797898 = queryWeight(content:الكريمأ), product of:
  - 1.063013 = idf(docFreq=122, numDocs=131)
  - 0.035557326 = queryNorm
- 0.014680931 = (MATCH) fieldWeight(content:الكريمأ in 73), product of:
  - 2.828427 = tf(termFreq(content:الكريمأ)=8)
  - 1.063013 = idf(docFreq=122, numDocs=131)
  - 0.0048828125 = fieldNorm(field=content, doc=73)

o 0.050378136 = (MATCH) weight(title:الكريمأ^1.5 in 73), product of:

- 0.23942009 = queryWeight(title:الكريمأ^1.5), product of:
  - 1.5 = boost
  - 4.488903 = idf(docFreq=3, numDocs=131)
  - 0.035557326 = queryNorm
- 0.21041733 = (MATCH) fieldWeight(title:الكريمأ in 73), product of:
  - 1.0 = tf(termFreq(title:الكريمأ)=1)
  - 4.488903 = idf(docFreq=3, numDocs=131)
  - 0.046875 = fieldNorm(field=title, doc=73)

*****

Page 3:

- boost = 0.23039404
- digest = 8ed8fcd743ff1ce5d4c42db83fc549af
- lang = ar
- segment = 20100307101102
- title = الحياة الأميركية - الحياة الأميركية - America.gov
- tstamp = 20100307151136760
- url = http://www.america.gov/ar/amlife.html

score for query: أميركا

- 0.10594571 = (MATCH) sum of:
  - 0.10521901 = (MATCH) weight(anchor:أميركا^2.0 in 3), product of:
    - 0.26155776 = queryWeight(anchor:أميركا^2.0), product of:
      - 2.0 = boost
      - 3.6779728 = idf(docFreq=8, numDocs=131)
      - 0.035557326 = queryNorm
    - 0.40227827 = (MATCH) fieldWeight(anchor:أميركا in 3), product of:
      - 1.0 = tf(termFreq(anchor:أميركا)=1)
      - 3.6779728 = idf(docFreq=8, numDocs=131)
      - 0.109375 = fieldNorm(field=anchor, doc=3)
  - 7.2669686E-4 = (MATCH) weight(content:أميركا in 3), product of:
    - 0.037797898 = queryWeight(content:أميركا), product of:
      - 1.063013 = idf(docFreq=122, numDocs=131)
      - 0.035557326 = queryNorm
    - 0.019225854 = (MATCH) fieldWeight(content:أميركا in 3), product of:

        - 2.6457512 = tf(termFreq(content:أكريم)=7)
        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.0068359375 = fieldNorm(field=content, doc=3)

*****

Page 4:

- boost = 0.1587577
- digest = a5795145f4a839cf52528d1b49e03bd1
- lang = ar
- segment = 20100307101102
- title = ربع نشيديإ لي ابوم فوغ تود الكريمأ لاوجلا ةخسن - نشيديإ لي ابوم قمدخ أزهج PDA - America.gov
- tstamp = 20100307151139253
- url = http://www.america.gov/ar/services/mobile.html

score for query: الكريمأ

- 0.042462345 = (MATCH) sum of:
    - 4.8056475E-4 = (MATCH) weight(content:الكريمأ in 115), product of:
        - 0.037797898 = queryWeight(content:الكريمأ), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.035557326 = queryNorm
        - 0.01271406 = (MATCH) fieldWeight(content:الكريمأ in 115), product of:
            - 2.4494898 = tf(termFreq(content:الكريمأ)=6)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.0048828125 = fieldNorm(field=content, doc=115)
    - 0.04198178 = (MATCH) weight(title:الكريمأ^1.5 in 115), product of:
        - 0.23942009 = queryWeight(title:الكريمأ^1.5), product of:
            - 1.5 = boost
            - 4.488903 = idf(docFreq=3, numDocs=131)

        - 0.035557326 = queryNorm
    - 0.17534778 = (MATCH) fieldWeight(title:الكريمأ in 115), product of:
        - 1.0 = tf(termFreq(title:الكريمأ)=1)
        - 4.488903 = idf(docFreq=3, numDocs=131)
        - 0.0390625 = fieldNorm(field=title, doc=115)

*****

Page 5:

- boost = 0.04832446
- digest = 22e534f031e9c7ac8682fcd4f86523e4
- lang = ar
- segment = 20100307101231
- title = أبلوبم روص - الحاكيةة ألامريكيةة باب روص - America.gov
- tstamp = 20100307151334977
- url = http://www.america.gov/ar/multimedia/photogallery.html#/4110/mosques_ar/

score for query: الكريمأ

- 0.022691099 = (MATCH) sum of:
    - 0.02254693 = (MATCH) weight(anchor:الكريمأ^2.0 in 51), product of:
        - 0.26155776 = queryWeight(anchor:الكريمأ^2.0), product of:
            - 2.0 = boost
            - 3.6779728 = idf(docFreq=8, numDocs=131)
            - 0.035557326 = queryNorm
        - 0.08620249 = (MATCH) fieldWeight(anchor:الكريمأ in 51), product of:
            - 1.0 = tf(termFreq(anchor:الكريمأ)=1)
            - 3.6779728 = idf(docFreq=8, numDocs=131)
            - 0.0234375 = fieldNorm(field=anchor, doc=51)

- 1.4416942E-4 = (MATCH) weight(content:الكريمأ in 51), product of:
  - 0.037797898 = queryWeight(content:الكريمأ), product of:
    - 1.063013 = idf(docFreq=122, numDocs=131)
    - 0.035557326 = queryNorm
  - 0.0038142179 = (MATCH) fieldWeight(content:الكريمأ in 51), product of:
    - 2.4494898 = tf(termFreq(content:الكريمأ)=6)
    - 1.063013 = idf(docFreq=122, numDocs=131)
    - 0.0014648438 = fieldNorm(field=content, doc=51)

*****

Page 6:

- boost = 0.033444975
- digest = 80c97402726fad635131db1bb29555be
- lang = ar
- segment = 20100307101231
- title = بتک - بتک - America.gov
- tstamp = 20100307151330985
- url = http://www.america.gov/ar/publications/books.html#beingmuslim

score for query: الكريمأ

- 0.01513323 = (MATCH) sum of:
  - 0.0150312865 = (MATCH) weight(anchor:الكريمأ^2.0 in 76), product of:
    - 0.26155776 = queryWeight(anchor:الكريمأ^2.0), product of:
      - 2.0 = boost
      - 3.6779728 = idf(docFreq=8, numDocs=131)
      - 0.035557326 = queryNorm
    - 0.057468325 = (MATCH) fieldWeight(anchor:الكريمأ in 76), product of:
      - 1.0 = tf(termFreq(anchor:الكريمأ)=1)

        - 3.6779728 = idf(docFreq=8, numDocs=131)
        - 0.015625 = fieldNorm(field=anchor, doc=76)
    - 1.01943166E-4 = (MATCH) weight(content:أكريم in 76), product of:
        - 0.037797898 = queryWeight(content:أكريم), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.035557326 = queryNorm
        - 0.0026970592 = (MATCH) fieldWeight(content:أكريم in 76), product of:
            - 3.4641016 = tf(termFreq(content:أكريم)=12)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 7.324219E-4 = fieldNorm(field=content, doc=76)

*****

Page 7:

- boost = 0.0420541
- digest = 1e1bc6ad9ffbfcdb82ea012b44610bed
- lang = ar
- segment = 20100307101231
- title = أبل هوص - الحاكية ألمريكية باب لوص - America.gov
- tstamp = 20100307151307072
- url = http://www.america.gov/ar/multimedia/photogallery.html#/4110/religious_freedom_ar/

score for query: أكريم

- 0.011393607 = (MATCH) sum of:
    - 0.011273465 = (MATCH) weight(anchor:أكريم^2.0 in 53), product of:
        - 0.26155776 = queryWeight(anchor:أكريم^2.0), product of:
            - 2.0 = boost
            - 3.6779728 = idf(docFreq=8, numDocs=131)

        - 0.035557326 = queryNorm
    - 0.043101244 = (MATCH) fieldWeight(anchor:الكريمأ in 53), product of:
        - 1.0 = tf(termFreq(anchor:الكريمأ)=1)
        - 3.6779728 = idf(docFreq=8, numDocs=131)
        - 0.01171875 = fieldNorm(field=anchor, doc=53)
- 1.2014119E-4 = (MATCH) weight(content:الكريمأ in 53), product of:
    - 0.037797898 = queryWeight(content:الكريمأ), product of:
        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.035557326 = queryNorm
    - 0.003178515 = (MATCH) fieldWeight(content:الكريمأ in 53), product of:
        - 2.4494898 = tf(termFreq(content:الكريمأ)=6)
        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.0012207031 = fieldNorm(field=content, doc=53)

*****

Page 8:

- boost = 0.02675021
- digest = df2eeaef879a60aaaddf7c8403cba7fa
- lang = ar
- segment = 20100307101458
- title = بتک - بتک - America.gov
- tstamp = 20100307151541037
- url = http://www.america.gov/ar/publications/books.html#governed

score for query: الكريمأ

- 0.011358418 = (MATCH) sum of:
    - 0.011273465 = (MATCH) weight(anchor:الكريمأ^2.0 in 78), product of:
        - 0.26155776 = queryWeight(anchor:الكريمأ^2.0), product of:

        - 2.0 = boost
        - 3.6779728 = idf(docFreq=8, numDocs=131)
        - 0.035557326 = queryNorm
    - 0.043101244 = (MATCH) fieldWeight(anchor:الكريمأ in 78), product of:
        - 1.0 = tf(termFreq(anchor:الكريمأ)=1)
        - 3.6779728 = idf(docFreq=8, numDocs=131)
        - 0.01171875 = fieldNorm(field=anchor, doc=78)
- 8.495264E-5 = (MATCH) weight(content:الكريمأ in 78), product of:
    - 0.037797898 = queryWeight(content:الكريمأ), product of:
        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.035557326 = queryNorm
    - 0.0022475494 = (MATCH) fieldWeight(content:الكريمأ in 78), product of:
        - 3.4641016 = tf(termFreq(content:الكريمأ)=12)
        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 6.1035156E-4 = fieldNorm(field=content, doc=78)

*****

Page 9:

- boost = 0.025411258
- digest = 295971814b3454a9d44144054b5c194a
- lang = ar
- segment = 20100307101458
- title = ألبوم روص - الحاكية ألمريكية باب روصلا - America.gov
- tstamp = 20100307151543423
- url = http://www.america.gov/ar/multimedia/photogallery.html#/4110/islam_ar/

score for query: الكريمأ

- 0.0113455495 = (MATCH) sum of:

- 0.011273465 = (MATCH) weight(anchor:أمريكا^2.0 in 50), product of:
  - 0.26155776 = queryWeight(anchor:أمريكا^2.0), product of:
    - 2.0 = boost
    - 3.6779728 = idf(docFreq=8, numDocs=131)
    - 0.035557326 = queryNorm
  - 0.043101244 = (MATCH) fieldWeight(anchor:أمريكا in 50), product of:
    - 1.0 = tf(termFreq(anchor:أمريكا)=1)
    - 3.6779728 = idf(docFreq=8, numDocs=131)
    - 0.01171875 = fieldNorm(field=anchor, doc=50)
- 7.208471E-5 = (MATCH) weight(content:أمريكا in 50), product of:
  - 0.037797898 = queryWeight(content:أمريكا), product of:
    - 1.063013 = idf(docFreq=122, numDocs=131)
    - 0.035557326 = queryNorm
  - 0.0019071089 = (MATCH) fieldWeight(content:أمريكا in 50), product of:
    - 2.4494898 = tf(termFreq(content:أمريكا)=6)
    - 1.063013 = idf(docFreq=122, numDocs=131)
    - 7.324219E-4 = fieldNorm(field=content, doc=50)

*****

Page 10:

- boost = 1.0000145
- digest = eed4dd9817b50ffda0aef158be6e4c12
- lang = ar
- segment = 20100307101052
- title = الصفحة الأولى - الصفحة الأولى - America.gov
- tstamp = 20100307151057483
- url = http://www.america.gov/ar/

score for query: الكريمأ

- 0.0028076388 = (MATCH) sum of:
    - 0.0028076388 = (MATCH) weight(content:الكريمأ in 0), product of:
        - 0.037797898 = queryWeight(content:الكريمأ), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.035557326 = queryNorm
        - 0.07428029 = (MATCH) fieldWeight(content:الكريمأ in 0), product of:
            - 2.236068 = tf(termFreq(content:الكريمأ)=5)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.03125 = fieldNorm(field=content, doc=0)

APPENDIX E

This is the detail score for query of top 10 pages using *ArabicAnalyzer*.

Search Term: **طارقمي د** (Democratic)

Page 1:

- boost = 0.16689056
- digest = 6e1b0463970c5b60bb75636a698cf1b3
- lang = ar
- segment = 20100305180909
- title = الديمقراطية - الديمقراطية - America.gov
- tstamp = 20100305230951886
- url = http://www.america.gov/ar/global/democracy.html

score for query: ط ا ر ق م ي د

- 0.2665834 = (MATCH) sum of:
  - 0.15995954 = (MATCH) weight(anchor:طارقميد^2.0 in 23), product of:
    - 0.30052778 = queryWeight(anchor:طارقميد^2.0), product of:
      - 2.0 = boost
      - 4.2580967 = idf(docFreq=4, numDocs=130)
      - 0.035288982 = queryNorm
    - 0.5322621 = (MATCH) fieldWeight(anchor:طارقميد in 23), product of:
      - 1.0 = tf(termFreq(anchor:طارقميد)=1)
      - 4.2580967 = idf(docFreq=4, numDocs=130)
      - 0.125 = fieldNorm(field=anchor, doc=23)
  - 5.846775E-4 = (MATCH) weight(content:طارقميد in 23), product of:
    - 0.037530307 = queryWeight(content:طارقميد), product of:
      - 1.0635134 = idf(docFreq=121, numDocs=130)

        - 0.035288982 = queryNorm
    - 0.01557881 = (MATCH) fieldWeight(content:طارقمي in 23), product of:
        - 3.0 = tf(termFreq(content:طارقمي)=9)
        - 1.0635134 = idf(docFreq=121, numDocs=130)
        - 0.0048828125 = fieldNorm(field=content, doc=23)
- 0.1060392 = (MATCH) weight(title:طارقمي^1.5 in 23), product of:
    - 0.22539584 = queryWeight(title:طارقمي^1.5), product of:
        - 1.5 = boost
        - 4.2580967 = idf(docFreq=4, numDocs=130)
        - 0.035288982 = queryNorm
    - 0.47045767 = (MATCH) fieldWeight(title:طارقمي in 23), product of:
        - 1.4142135 = tf(termFreq(title:طارقمي)=2)
        - 4.2580967 = idf(docFreq=4, numDocs=130)
        - 0.078125 = fieldNorm(field=title, doc=23)

*****

Page 2:

- boost = 0.23113073
- digest = 5285dc46473be73851750b409de012a5
- lang = ar
- segment = 20100305180909
- title = التحدي العالمي - التحدي العالمي - America.gov
- tstamp = 20100305230921085
- url = http://www.america.gov/ar/global.html

score for query: طارقمي

- 0.16062789 = (MATCH) sum of:
    - 0.15995954 = (MATCH) weight(anchor:طارقمي^2.0 in 22), product of:

- 0.30052778 = queryWeight(anchor:طارقميد^2.0), product of:
  - 2.0 = boost
  - 4.2580967 = idf(docFreq=4, numDocs=130)
  - 0.035288982 = queryNorm
- 0.5322621 = (MATCH) fieldWeight(anchor:طارقميد in 22), product of:
  - 1.0 = tf(termFreq(anchor:طارقميد)=1)
  - 4.2580967 = idf(docFreq=4, numDocs=130)
  - 0.125 = fieldNorm(field=anchor, doc=22)

o 6.683421E-4 = (MATCH) weight(content:طارقميد in 22), product of:

- 0.037530307 = queryWeight(content:طارقميد), product of:
  - 1.0635134 = idf(docFreq=121, numDocs=130)
  - 0.035288982 = queryNorm
- 0.017808065 = (MATCH) fieldWeight(content:طارقميد in 22), product of:
  - 2.4494898 = tf(termFreq(content:طارقميد)=6)
  - 1.0635134 = idf(docFreq=121, numDocs=130)
  - 0.0068359375 = fieldNorm(field=content, doc=22)

*****

Page 3:

- boost = 0.031816483
- digest = bba906c38386b2e71f42a4f7d365e8cb
- lang = ar
- segment = 20100305181031
- title = ةيطارقميدلا وقاوسألا - ةيطارقميدلا وقاوسألا - America.gov
- tstamp = 20100305231058196
- url = http://www.america.gov/ar/publications/ejournalusa/608.html

score for query: طارقميد

- 0.033635326 = (MATCH) sum of:

- 0.017495574 = (MATCH) weight(anchor:طارقميد^2.0 in 98), product of:
  - 0.30052778 = queryWeight(anchor:طارقميد^2.0), product of:
    - 2.0 = boost
    - 4.2580967 = idf(docFreq=4, numDocs=130)
    - 0.035288982 = queryNorm
  - 0.058216166 = (MATCH) fieldWeight(anchor:طارقميد in 98), product of:
    - 1.0 = tf(termFreq(anchor:طارقميد)=1)
    - 4.2580967 = idf(docFreq=4, numDocs=130)
    - 0.013671875 = fieldNorm(field=anchor, doc=98)
- 2.33871E-4 = (MATCH) weight(content:طارقميد in 98), product of:
  - 0.037530307 = queryWeight(content:طارقميد), product of:
    - 1.0635134 = idf(docFreq=121, numDocs=130)
    - 0.035288982 = queryNorm
  - 0.006231524 = (MATCH) fieldWeight(content:طارقميد in 98), product of:
    - 6.0 = tf(termFreq(content:طارقميد)=36)
    - 1.0635134 = idf(docFreq=121, numDocs=130)
    - 9.765625E-4 = fieldNorm(field=content, doc=98)
- 0.015905881 = (MATCH) weight(title:طارقميد^1.5 in 98), product of:
  - 0.22539584 = queryWeight(title:طارقميد^1.5), product of:
    - 1.5 = boost
    - 4.2580967 = idf(docFreq=4, numDocs=130)
    - 0.035288982 = queryNorm
  - 0.07056865 = (MATCH) fieldWeight(title:طارقميد in 98), product of:
    - 1.4142135 = tf(termFreq(title:طارقميد)=2)
    - 4.2580967 = idf(docFreq=4, numDocs=130)

- 0.01171875 = fieldNorm(field=title, doc=98)

*****

Page 4:

- boost = 0.11378951
- digest = b8c157220365a4bf104bc045832885be
- lang = ar
- segment = 20100305180909
- title = أكثر من انتخابات: كيف ينتقل الحكم في النظم الديمقراطية - 0110 - America.gov
- tstamp = 20100305231013594
- url = http://www.america.gov/ar/publications/ejournalusa/0110.html

score for query: د ي م ق ر ا ط

- 0.030587077 = (MATCH) sum of:
  - 5.9466175E-4 = (MATCH) weight(content:طارقميد in 88), product of:
    - 0.037530307 = queryWeight(content:طارقميد), product of:
      - 1.0635134 = idf(docFreq=121, numDocs=130)
      - 0.035288982 = queryNorm
    - 0.01584484 = (MATCH) fieldWeight(content:طارقميد in 88), product of:
      - 4.358899 = tf(termFreq(content:طارقميد)=19)
      - 1.0635134 = idf(docFreq=121, numDocs=130)
      - 0.0034179688 = fieldNorm(field=content, doc=88)
  - 0.029992415 = (MATCH) weight(title:طارقميد^1.5 in 88), product of:
    - 0.22539584 = queryWeight(title:طارقميد^1.5), product of:
      - 1.5 = boost
      - 4.2580967 = idf(docFreq=4, numDocs=130)
      - 0.035288982 = queryNorm
    - 0.13306552 = (MATCH) fieldWeight(title:طارقميد in 88), product of:

        - 1.0 = tf(termFreq(title:ديمقراطي)=1)
        - 4.2580967 = idf(docFreq=4, numDocs=130)
        - 0.03125 = fieldNorm(field=title, doc=88)

*****

Page 5:

- boost = 0.028445216
- digest = c04e43d37fb6f380a397373427882a1e
- lang = ar
- segment = 20100305181151
- title = مقدمة - الديمقراطية في العالم | مواطنون لهم حرية التعبير - America.gov
- tstamp = 20100305231252420
- url = http://www.america.gov/ar/democracy/global/index.html

score for query: ديمقراطي

- 0.027611194 = (MATCH) sum of:
    - 0.019994942 = (MATCH) weight(anchor:ديمقراطي^2.0 in 14), product of:
        - 0.30052778 = queryWeight(anchor:ديمقراطي^2.0), product of:
            - 2.0 = boost
            - 4.2580967 = idf(docFreq=4, numDocs=130)
            - 0.035288982 = queryNorm
        - 0.06653276 = (MATCH) fieldWeight(anchor:ديمقراطي in 14), product of:
            - 1.0 = tf(termFreq(anchor:ديمقراطي)=1)
            - 4.2580967 = idf(docFreq=4, numDocs=130)
            - 0.015625 = fieldNorm(field=anchor, doc=14)
    - 1.181473E-4 = (MATCH) weight(content:ديمقراطي in 14), product of:
        - 0.037530307 = queryWeight(content:ديمقراطي), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035288982 = queryNorm

        - 0.0031480505 = (MATCH) fieldWeight(content:طارقمي in 14), product of:
            - 3.4641016 = tf(termFreq(content:طارقمي)=12)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 8.544922E-4 = fieldNorm(field=content, doc=14)
    - 0.0074981037 = (MATCH) weight(title:طارقمي^1.5 in 14), product of:
        - 0.22539584 = queryWeight(title:طارقمي^1.5), product of:
            - 1.5 = boost
            - 4.2580967 = idf(docFreq=4, numDocs=130)
            - 0.035288982 = queryNorm
        - 0.03326638 = (MATCH) fieldWeight(title:طارقمي in 14), product of:
            - 1.0 = tf(termFreq(title:طارقمي)=1)
            - 4.2580967 = idf(docFreq=4, numDocs=130)
            - 0.0078125 = fieldNorm(field=title, doc=14)

*****

Page 6:

- boost = 1.0000145
- digest = 0d5b023c802941ddb358071073a98833
- lang = ar
- segment = 20100305180856
- title = الصفحة الأولى - الصفحة الأولى - America.gov
- tstamp = 20100305230902835
- url = http://www.america.gov/ar/

score for query: طارقمي

- 0.0021604078 = (MATCH) sum of:
    - 0.0021604078 = (MATCH) weight(content:طارقمي in 0), product of:
        - 0.037530307 = queryWeight(content:طارقمي), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)

        - 0.035288982 = queryNorm
    - 0.05756435 = (MATCH) fieldWeight(content:طارقمي in 0), product of:
        - 1.7320508 = tf(termFreq(content:طارقمي)=3)
        - 1.0635134 = idf(docFreq=121, numDocs=130)
        - 0.03125 = fieldNorm(field=content, doc=0)

*****

Page 7:

- boost = 0.22860475
- digest = 5a62cd3a20d5393ff5806fd92af1edef
- lang = ar
- segment = 20100305180909
- title = تسالكدوب - تسالكدوب - America.gov
- tstamp = 20100305230934690
- url = http://www.america.gov/ar/multimedia/podcast.html

score for query: طارقمي د

- 6.1011006E-4 = (MATCH) sum of:
    - 6.1011006E-4 = (MATCH) weight(content:طارقمي in 60), product of:
        - 0.037530307 = queryWeight(content:طارقمي), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035288982 = queryNorm
        - 0.016256463 = (MATCH) fieldWeight(content:طارقمي in 60), product of:
            - 2.236068 = tf(termFreq(content:طارقمي)=5)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.0068359375 = fieldNorm(field=content, doc=60)

*****

Page 8:

- boost = 0.22996004
- digest = a0130240b4348578aa8a83e59187dfb3

- lang = ar
- segment = 20100305180909
- title = بتک - بتک - America.gov
- tstamp = 20100305231001279
- url = http://www.america.gov/ar/publications/books.html

score for query: طارقمید

- 5.846775E-4 = (MATCH) sum of:
    - 5.846775E-4 = (MATCH) weight(content:طارقميد in 73), product of:
        - 0.037530307 = queryWeight(content:طارقميد), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035288982 = queryNorm
        - 0.01557881 = (MATCH) fieldWeight(content:طارقميد in 73), product of:
            - 3.0 = tf(termFreq(content:طارقميد)=9)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.0048828125 = fieldNorm(field=content, doc=73)

*****

Page 9:

- boost = 0.23032264
- digest = ce4a12d589c1a56e886d5b6848609391
- lang = ar
- segment = 20100305180909
- title = الحياة الأميركية - الحياة الأميركية - America.gov
- tstamp = 20100305230939904
- url = http://www.america.gov/ar/amlife.html

score for query: طارقمید

- 5.4569903E-4 = (MATCH) sum of:
    - 5.4569903E-4 = (MATCH) weight(content:طارقميد in 3), product of:
        - 0.037530307 = queryWeight(content:طارقميد), product of:

        - 1.0635134 = idf(docFreq=121, numDocs=130)
        - 0.035288982 = queryNorm
    - 0.0145402225 = (MATCH) fieldWeight(content:طارقميد in 3), product of:
        - 2.0 = tf(termFreq(content:طارقميد)=4)
        - 1.0635134 = idf(docFreq=121, numDocs=130)
        - 0.0068359375 = fieldNorm(field=content, doc=3)

*****

Page 10:

- boost = 0.2296042
- digest = bc9c562d0a61b335f5a8730f14412dcb
- lang = ar
- segment = 20100305180909
- title = إي لانروج وي أس آي ه - إي لانروج وي أس آي ه - America.gov
- tstamp = 20100305230924025
- url = http://www.america.gov/ar/publications/ejournalusa.html

score for query: طارقميد

- 5.4010196E-4 = (MATCH) sum of:
    - 5.4010196E-4 = (MATCH) weight(content:طارقميد in 86), product of:
        - 0.037530307 = queryWeight(content:طارقميد), product of:
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.035288982 = queryNorm
        - 0.014391088 = (MATCH) fieldWeight(content:طارقميد in 86), product of:
            - 3.4641016 = tf(termFreq(content:طارقميد)=12)
            - 1.0635134 = idf(docFreq=121, numDocs=130)
            - 0.00390625 = fieldNorm(field=content, doc=86)

APPENDIX F

This is the detail score for query of top 10 pages using *NutchDocumentAnalyzer*.

Search Term: **الديمقراطية** (Democratic)

Page 1:

- boost = 0.16692342
- digest = c49d13e1fa4eb518258862a27800f398
- lang = ar
- segment = 20100307101102
- title = الديمقراطية - الديمقراطية - America.gov
- tstamp = 20100307151151020
- url = http://www.america.gov/ar/global/democracy.html

score for query: ا د ل م ق ر ا ط ي ة

- 0.29354417 = (MATCH) sum of:
  - 0.17619587 = (MATCH) weight(anchor:الديمقراطية^2.0 in 24), product of:
    - 0.31401145 = queryWeight(anchor:الديمقراطية^2.0), product of:
      - 2.0 = boost
      - 4.488903 = idf(docFreq=3, numDocs=131)
      - 0.03497641 = queryNorm
    - 0.5611129 = (MATCH) fieldWeight(anchor:الديمقراطية in 24), product of:
      - 1.0 = tf(termFreq(anchor:الديمقراطية)=1)
      - 4.488903 = idf(docFreq=3, numDocs=131)
      - 0.125 = fieldNorm(field=anchor, doc=24)
  - 5.458426E-4 = (MATCH) weight(content:الديمقراطية in 24), product of:
    - 0.03718038 = queryWeight(content:الديمقراطية), product of:

        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.03497641 = queryNorm
      - 0.014680931 = (MATCH) fieldWeight(content:الديمقراطية in 24), product of:
        - 2.828427 = tf(termFreq(content:الديمقراطية)=8)
        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.0048828125 = fieldNorm(field=content, doc=24)
  - 0.116802454 = (MATCH) weight(title:الديمقراطية^1.5 in 24), product of:
    - 0.23550858 = queryWeight(title:الديمقراطية^1.5), product of:
      - 1.5 = boost
      - 4.488903 = idf(docFreq=3, numDocs=131)
      - 0.03497641 = queryNorm
    - 0.4959584 = (MATCH) fieldWeight(title:الديمقراطية in 24), product of:
      - 1.4142135 = tf(termFreq(title:الديمقراطية)=2)
      - 4.488903 = idf(docFreq=3, numDocs=131)
      - 0.078125 = fieldNorm(field=title, doc=24)

*****

Page 2:

- boost = 0.23117816
- digest = 6f317cffadc06ecf85513b0eb565f1b8
- lang = ar
- segment = 20100307101102
- title = التحدي العالمي - التحدي العالمي - America.gov
- tstamp = 20100307151115329
- url = http://www.america.gov/ar/global.html

score for query: الديمقراطية

- 0.17680001 = (MATCH) sum of:

- 0.17619587 = (MATCH) weight(anchor:الديمقراطية^2.0 in 23), product of:
  - 0.31401145 = queryWeight(anchor:الديمقراطية^2.0), product of:
    - 2.0 = boost
    - 4.488903 = idf(docFreq=3, numDocs=131)
    - 0.03497641 = queryNorm
  - 0.5611129 = (MATCH) fieldWeight(anchor:الديمقراطية in 23), product of:
    - 1.0 = tf(termFreq(anchor:الديمقراطية)=1)
    - 4.488903 = idf(docFreq=3, numDocs=131)
    - 0.125 = fieldNorm(field=anchor, doc=23)
- 6.041371E-4 = (MATCH) weight(content:الديمقراطية in 23), product of:
  - 0.03718038 = queryWeight(content:الديمقراطية), product of:
    - 1.063013 = idf(docFreq=122, numDocs=131)
    - 0.03497641 = queryNorm
  - 0.016248815 = (MATCH) fieldWeight(content:الديمقراطية in 23), product of:
    - 2.236068 = tf(termFreq(content:الديمقراطية)=5)
    - 1.063013 = idf(docFreq=122, numDocs=131)
    - 0.0068359375 = fieldNorm(field=content, doc=23)

*****

Page 3:

- boost = 0.11378951
- digest = ab333ad468abf764c43637fe53b7e4f7
- lang = ar
- segment = 20100307101102
- title = أكثر من انتخابات: كيف ينتقل الحكم في النظم الديمقراطية - 0110 - America.gov
- tstamp = 20100307151214969

- url = http://www.america.gov/ar/publications/ejournalusa/0110.html

score for query: ا ديمقراطية

- 0.033559922 = (MATCH) sum of:
    - 5.2319805E-4 = (MATCH) weight(content:الديمقراطية in 89), product of:
        - 0.03718038 = queryWeight(content:الديمقراطية), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.03497641 = queryNorm
        - 0.0140718855 = (MATCH) fieldWeight(content:الديمقراطية in 89), product of:
            - 3.8729835 = tf(termFreq(content:الديمقراطية)=15)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.0034179688 = fieldNorm(field=content, doc=89)
    - 0.033036724 = (MATCH) weight(title:الديمقراطية^1.5 in 89), product of:
        - 0.23550858 = queryWeight(title:الديمقراطية^1.5), product of:
            - 1.5 = boost
            - 4.488903 = idf(docFreq=3, numDocs=131)
            - 0.03497641 = queryNorm
        - 0.14027822 = (MATCH) fieldWeight(title:الديمقراطية in 89), product of:
            - 1.0 = tf(termFreq(title:الديمقراطية)=1)
            - 4.488903 = idf(docFreq=3, numDocs=131)
            - 0.03125 = fieldNorm(field=title, doc=89)

*****

Page 4:

- boost = 0.028445216
- digest = 9212154ec8740ad77458648f74aa149c
- lang = ar
- segment = 20100307101343
- title = مقدمة - الديمقراطية في العالم | مواطنون لهم حرية التعبير - America.gov

- tstamp = 20100307151423606
- url = http://www.america.gov/ar/democracy/global/index.html

score for query: ةيطارقميدلا

- 0.030395675 = (MATCH) sum of:
    - 0.022024484 = (MATCH) weight(anchor:الديمقراطية^2.0 in 15), product of:
        - 0.31401145 = queryWeight(anchor:الديمقراطية^2.0), product of:
            - 2.0 = boost
            - 4.488903 = idf(docFreq=3, numDocs=131)
            - 0.03497641 = queryNorm
        - 0.07013911 = (MATCH) fieldWeight(anchor:الديمقراطية in 15), product of:
            - 1.0 = tf(termFreq(anchor:الديمقراطية)=1)
            - 4.488903 = idf(docFreq=3, numDocs=131)
            - 0.015625 = fieldNorm(field=anchor, doc=15)
    - 1.12010006E-4 = (MATCH) weight(content:الديمقراطية in 15), product of:
        - 0.03718038 = queryWeight(content:الديمقراطية), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.03497641 = queryNorm
        - 0.0030126106 = (MATCH) fieldWeight(content:الديمقراطية in 15), product of:
            - 3.3166249 = tf(termFreq(content:الديمقراطية)=11)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 8.544922E-4 = fieldNorm(field=content, doc=15)
    - 0.008259181 = (MATCH) weight(title:الديمقراطية^1.5 in 15), product of:
        - 0.23550858 = queryWeight(title:الديمقراطية^1.5), product of:
            - 1.5 = boost
            - 4.488903 = idf(docFreq=3, numDocs=131)

        - 0.03497641 = queryNorm
    - 0.035069555 = (MATCH) fieldWeight(title:الديمقراطية in 15), product of:
        - 1.0 = tf(termFreq(title:الديمقراطية)=1)
        - 4.488903 = idf(docFreq=3, numDocs=131)
        - 0.0078125 = fieldNorm(field=title, doc=15)

*****

Page 5:

- boost = 1.0000145
- digest = eed4dd9817b50ffda0aef158be6e4c12
- lang = ar
- segment = 20100307101052
- title = الصفحة الأولى - الصفحة الأولى - America.gov
- tstamp = 20100307151057483
- url = http://www.america.gov/ar/

score for query: الديمقراطية

- 0.0021392573 = (MATCH) sum of:
    - 0.0021392573 = (MATCH) weight(content:الديمقراطية in 0), product of:
        - 0.03718038 = queryWeight(content:الديمقراطية), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.03497641 = queryNorm
        - 0.05753726 = (MATCH) fieldWeight(content:الديمقراطية in 0), product of:
            - 1.7320508 = tf(termFreq(content:الديمقراطية)=3)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.03125 = fieldNorm(field=content, doc=0)

*****

Page 6:

- boost = 0.22865272
- digest = dc127d214554a59575782c318462f4e8

- lang = ar
- segment = 20100307101102
- title = تسالكدوب - تسالكدوب - America.gov
- tstamp = 20100307151128611
- url = http://www.america.gov/ar/multimedia/podcast.html

score for query: ةيطارقميدلا

- 6.041371E-4 = (MATCH) sum of:
    - 6.041371E-4 = (MATCH) weight(content:الديمقراطية in 61), product of:
        - 0.03718038 = queryWeight(content:الديمقراطية), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.03497641 = queryNorm
        - 0.016248815 = (MATCH) fieldWeight(content:الديمقراطية in 61), product of:
            - 2.236068 = tf(termFreq(content:الديمقراطية)=5)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.0068359375 = fieldNorm(field=content, doc=61)

*****

Page 7:

- boost = 0.23039404
- digest = 8ed8fcd743ff1ce5d4c42db83fc549af
- lang = ar
- segment = 20100307101102
- title = الحياة الأمريكية - الحياة الأمريكية - America.gov
- tstamp = 20100307151136760
- url = http://www.america.gov/ar/amlife.html

score for query: ةيطارقميدلا

- 5.4035656E-4 = (MATCH) sum of:
    - 5.4035656E-4 = (MATCH) weight(content:الديمقراطية in 3), product of:
        - 0.03718038 = queryWeight(content:الديمقراطية), product of:

        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.03497641 = queryNorm
    - 0.01453338 = (MATCH) fieldWeight(content:الديمقراطية in 3), product of:
        - 2.0 = tf(termFreq(content:الديمقراطية)=4)
        - 1.063013 = idf(docFreq=122, numDocs=131)
        - 0.0068359375 = fieldNorm(field=content, doc=3)

*****

Page 8:

- boost = 0.22700267
- digest = c639ba79e6601f1242cee32b3ba640f4
- lang = ar
- segment = 20100307101102
- title = الانسان والأماكن - الانسان والأماكن - America.gov
- tstamp = 20100307151120668
- url = http://www.america.gov/ar/amlife/people.html

score for query: الديمقراطية

- 4.6796253E-4 = (MATCH) sum of:
    - 4.6796253E-4 = (MATCH) weight(content:الديمقراطية in 7), product of:
        - 0.03718038 = queryWeight(content:الديمقراطية), product of:
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.03497641 = queryNorm
        - 0.012586276 = (MATCH) fieldWeight(content:الديمقراطية in 7), product of:
            - 1.7320508 = tf(termFreq(content:الديمقراطية)=3)
            - 1.063013 = idf(docFreq=122, numDocs=131)
            - 0.0068359375 = fieldNorm(field=content, doc=7)

*****

Page 9:

- boost = 0.22826105

- digest = c33a5dc3f7d8475491bfafcf91c8b283
- lang = ar
- segment = 20100307101102
- title = الاقتصاد - الاقتصاد - America.gov
- tstamp = 20100307151153574
- url = http://www.america.gov/ar/econ.html

score for query: الديمقراطية

- 4.6796253E-4 = (MATCH) sum of:
  - 4.6796253E-4 = (MATCH) weight(content:الديمقراطية in 16), product of:
    - 0.03718038 = queryWeight(content:الديمقراطية), product of:
      - 1.063013 = idf(docFreq=122, numDocs=131)
      - 0.03497641 = queryNorm
    - 0.012586276 = (MATCH) fieldWeight(content:الديمقراطية in 16), product of:
      - 1.7320508 = tf(termFreq(content:الديمقراطية)=3)
      - 1.063013 = idf(docFreq=122, numDocs=131)
      - 0.0068359375 = fieldNorm(field=content, doc=16)

*****

Page 10:

- boost = 0.23110132
- digest = 3c7f5c1dc4d604ef275f72043bf8cfc1
- lang = ar
- segment = 20100307101102
- title = secondary Multimedia - وسائل إعلامية - America.gov
- tstamp = 20100307151158462
- url = http://www.america.gov/ar/multimedia.html

score for query: الديمقراطية

- 4.6796253E-4 = (MATCH) sum of:
  - 4.6796253E-4 = (MATCH) weight(content:الديمقراطية in 38), product of:

- 0.03718038 = queryWeight(content:الديمقراطية), product of:
  - 1.063013 = idf(docFreq=122, numDocs=131)
  - 0.03497641 = queryNorm
- 0.012586276 = (MATCH) fieldWeight(content:الديمقراطية in 38), product of:
  - 1.7320508 = tf(termFreq(content:الديمقراطية)=3)
  - 1.063013 = idf(docFreq=122, numDocs=131)
  - 0.0068359375 = fieldNorm(field=content, doc=38)

LIST OF REFERENCES

[1] B. Hoffman, "The use of the Internet by Islamic Extremists," in Testimony presented to the House Permanent Select Committee on Intelligence, p. 4, May 4, 2006.

[2] K. Börner, S. Sanyal, and A. Vespignani, "Network science," in *Annual Review of Information Science & Technology*, Vol. 41, B. Cronin, ed., pp. 537–607, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ, 2007.

[3] H. M. Harmanani, W.T. Keirouz, and S. Raheel, "A rule based extensive stemmer for information retrieval with application to Arabic," *Int. International Arab J. of Inform. Tech.,* vol. 3, no. 3, pp. 265–272, July 2006.

[4] D. A. Grossman and O. Frieder, *Information Retrieval Algorithm and Heuristics*, 2nd Ed. Springer, Norwell, MA, 2004.

[5] W. B. Croft, Ed., *Advances in Information Retrieval: recent research from the Center for Intelligent Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, 2003.

[6] R. El Gamal, "Arabic speakers, a dying breed in the Arab world?" *Kuwait Times*, December 27, 2007.[Online]. Available: http://www.kuwaittimes.net/read_news.php?newsid=MzA5NTkzOTI1 (accessed: May 17, 2009).

[7] "Internet Usage Statistics The Internet Big Picture World Internet Users and Population Stats." [Online]. Available: http://www.internetworldstats.com/stats.htm. (accessed Feb 10, 2009).

[8] A. Chen and F. Gey, "Building an Arabic Stemmer for Information Retrieval," in *Proc. 11th Text Retrieval Conf.,* 2002, pp. 631–640.

[9] J. Xu, A. Fraser and R. Weischedel, "Empirical Studies in Strategies for Arabic Retrieval," in *Proc. 25th Annual International Conf. Research and Development in Information Retrieval(SIGIR 2002),* 2002, pp. 269–274.

[10] L. S. Larkey and M.E Connell, "Arabic Information Retrieval at UMass in TREC-10," in *Proc. 10th Text Retrieval Conf.,* 2001, pp. 562–570.

[11] R. Sonbol and N. Ghneim, "Arabic Morphological Analysis: a New Approach," in *Information and Communication Technologies: From Theory to Applications,* pp. 1–6, 2008.

[12] K. Taghva, R. Elkhoury and J. Coombs, “Arabic Stemming Without a Root Dictionary,” Information Science Research Institute (ISRI), 2005. [Online]. Available: http://www.isri.unlv.edu/publications/isripub/Taghva2005b.ps. (accessed: May 20, 2009).

[13] T. BuckWalter, “Qamus: Arabic lexicography,” 2003. [Online]. Available: http://www.qamus.org/ (accessed: March 10, 2009).

[14] L. S. Larkey, L. Ballesteros and M.E. Connell, “Improving stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis,” in *SIGIR 2002*, pp. 275–282, 2002.

[15] O. Gospodnetic and E. Hatcher, *Lucene In Action*. Manning Publications Co., Greenwich, CT, 2005.

[16] T. White, “Introduction to Nutch, Part 1: Crawling,” January 10, 2006. [Online]. Available: http://today.java.net/pub/a/today/2006/01/10/introduction-to-nutch-1.html (accessed: May 10, 2009).

[17] D. P. Zhou, “Delve inside the Lucene indexing mechanism,” June 27, 2006. [Online]. Available: http://www.ibm.com/developerworks/library/wa-lucene/ (accessed: May 10, 2009).

[18] M. Cafarella and D. Cutting, “Building Nutch: Open Source Search,” May 5, 2004. [Online]. Available: http://queue.acm.org/detail.cfm?id=988408 (accessed: May 17, 2009)

[19] A. Bolour, “Notes on the Eclipse Plug-in Architecture,” July 3, 2003. [Online]. Available: http://www.eclipse.org/articles/Article-Plug-in-architecture/plugin_architecture.html (accessed: May 17, 2009)

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Chairman, Code EC
   Department of Electrical and Computer Engineering
   Naval Postgraduate School
   Monterey, California

4. Dr. Weilian Su
   Department of Electrical and Computer Engineering
   Naval Postgraduate School
   Monterey, California

5. Prof. John C. McEachen
   Department of Electrical and Computer Engineering
   Naval Postgraduate School
   Monterey, California

6. Qui Nguyen
   Lieutenant, United States Navy
   Naval Postgradute School
   Monterey, California