Spring 1-1-2011

Topic Modeling as an Analysis Tool to Understand the Impact of the Iraq War on the Iraqi Blogosphere

Mossaab Bagdouri
University of Colorado at Boulder, mossaab.bagdouri@colorado.edu

Follow this and additional works at: https://scholar.colorado.edu/csci_gradetds

Part of the Computer Sciences Commons

Recommended Citation
https://scholar.colorado.edu/csci_gradetds/22

This Thesis is brought to you for free and open access by Computer Science at CU Scholar. It has been accepted for inclusion in Computer Science Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
Topic Modeling as an Analysis Tool to Understand the Impact of the Iraq War on the Iraqi Blogosphere

by

Mossaab Bagdouri

State Engineer, ENSIAS, Morocco, 2007

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment of the requirements for the degree of
Master of Science
Department of Computer Science
2011
This thesis entitled:
Topic Modeling as an Analysis Tool to Understand
the Impact of the Iraq War on the Iraqi Blogosphere
written by Mossaab Bagdouri
has been approved for the Department of Computer Science

___________________________________________
Leysia Palen (chair)

___________________________________________
Kenneth Anderson

___________________________________________
Gloria Mark

___________________________________________
James Martin

Date _____________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
Social media provide a democratized platform for expressing one’s opinion or viewpoint. The everyday discussions that people have with their families, friends and colleagues became available through blogging services. The emergence of the blogging activity made the classic ethnographic approaches more difficult to deploy. The use of these methods becomes even more problematic when the available data contain a wide variety of languages. This thesis proposes the use of topic modeling as a method for quantitatively analyzing crawled blogs that were created by Iraqi citizens and active over a period of 8 years since the beginning of the Iraq war in 2003. This document presents how data were collected, the way dominant languages were separated in different datasets, the limits of using classification and clustering techniques, the benefits of employing topic modeling, and the evaluation of this technique.
To mom and dad,
Khadija, Chahid and Jannat,
Amine, Abderrahman and Hiba.
Acknowledgements

I would like to express my sincere gratitude to my thesis advisor Dr. Leysia Palen for her academic guidance and personal support; and my graduate advisor Dr. James Martin for introducing me to the Natural Language Processing world and providing the motivation I needed for this project.

I would like to thank Dr. Gloria Mark and Dr. Kenneth Anderson for agreeing to be on my thesis committee. Dr. Mark also helped me with correlation computation.

My special thanks go to Dr. Ban Al-Ani for our discussions about the history of blogging in Iraq and for spending a time to verify the output of my work.

A special thank to my co-authors of an ICWSM submission on the Iraqi blogs for agreeing on the reuse of a part of the paper toward this thesis.

I would like to thank Dr Martha Palmer, Dr. Aleksandra Sarcevic and the students Ali Alzabarah, William Corvey, Hansu GU, Sophia Liu, Casey McTaggart, Christopher Schenk, Aaron Schram, Catharine Starbird, Sarah Vieweg and Joanne White for their thoughts that gave me a better understanding of Crisis Informatics.
# Table of Contents

Table of Contents .................................................................................................................. vi

List of Tables.......................................................................................................................... viii

List of Figures........................................................................................................................... ix

Chapter 1  Introduction ......................................................................................................... 1

1.1. Iraqi War and Social Media ......................................................................................... 1
1.2. Thesis Background and Contribution ......................................................................... 3

Chapter 2  Data Gathering .................................................................................................. 4

2.1. Blogger ......................................................................................................................... 4
2.1. Blogger API .................................................................................................................. 5
2.2. Jericho ......................................................................................................................... 7
2.3. Crawling ....................................................................................................................... 8
2.4. Preprocessing .............................................................................................................. 14
2.4.1. Morphological Analysis ........................................................................... 14
2.4.2. Porter stemmer ...................................................................................... 15
2.4.3. Buckwalter morphological analyzer ..................................................... 15
2.4.4. Indices properties .................................................................................. 16

Chapter 3 Initial Analysis Attempts: Classification and Clustering .................. 18

3.1. Classification ............................................................................................ 18
3.2. Clustering .................................................................................................. 19

3.2.1. Definition .............................................................................................. 19
3.2.2. K-Means .............................................................................................. 19
3.2.3. Implementation ..................................................................................... 20
3.2.4. Results .................................................................................................. 20

Chapter 4 Topic Modeling ............................................................................... 26

4.1. Definition .................................................................................................. 26
4.2. Related Work ............................................................................................ 27
4.3. Implementation ......................................................................................... 28
4.4. Results ...................................................................................................... 33

Chapter 5 Evaluation ....................................................................................... 38

Chapter 6 Conclusion and Future Work .......................................................... 42

Bibliography ..................................................................................................... 44
List of Tables

Table 1: Examples of Usage of Porter’s Stemmer ................................................................. 15
Table 2: Indices Properties .................................................................................................. 16
Table 3: Most Frequent Terms in a Partitioning of 10 English Clusters ......................... 21
Table 4: Most Frequent Terms in a Partitioning of 20 English Clusters ......................... 24
Table 5 (a, b): Structure of Input Files for MATLAB Toolbox ....................................... 29
Table 6: Matrix of Word/Topic Distribution ........................................................................ 30
Table 7: Matrix of Document/Topic Distribution .............................................................. 30
Table 8: Most Frequent Terms in the English Posts in a Partitioning of 20 Topics .......... 32
Table 9: Most Frequent Terms in the English Posts in a Partitioning of 5 Topics ............. 32
Table 10: Most Frequent Terms in the English Posts by Topic ......................................... 33
Table 11: Translations of the Most Frequent Terms in the Arabic Posts by Topic .......... 35
Table 12: Distribution of Tokens over Topics for the Post in Figure 18 ......................... 37
List of Figures

Figure 1: Growth of Internet users as a percentage of population ........................................ 2
Figure 2: Simplified Class Diagram of the Blogger Service .................................................. 5
Figure 3: Example of Blogger Feed ....................................................................................... 6
Figure 4: Code Sample for Extracting the Content of a Tag from an HTML Page .............. 7
Figure 5 (a, b): Style Differences Between Two Blogs ...................................................... 8
Figure 6: Contradiction Between the Declared and the Used Languages ......................... 9
Figure 7: Language Detection Algorithm ........................................................................... 10
Figure 8: Arabic Content with its Translation in the Same Post ...................................... 11
Figure 9: Arabic Text Written in Latin Characters, with its Translation ....................... 12
Figure 10: Number of Topics per Month .......................................................................... 13
Figure 11: Number of Active Authors per Month .............................................................. 14
Figure 12: Example of Usage of Buckwalter’s Morphological Analyzer ....................... 16
Figure 13 - K-means Algorithm ................................................................. 20
Figure 14: Distribution of Documents Over a Partitioning of 10 English Clusters .......... 22
Figure 15: Distribution of Documents Over a Partitioning of 20 English Clusters .......... 24
Figure 16: Distribution of English Topics ...................................................... 34
Figure 17: Distribution of Arabic Topics ......................................................... 35
Figure 18: Example of a Post ........................................................................ 36
Figure 19: Distribution of Topics for the Post in Figure 18 .................................... 36
Figure 20: Body Count ................................................................................. 39
Figure 21: Timeline of War Topics in English .................................................. 40
Figure 22: Timeline of War Topics in Arabic .................................................... 41
1.1. Iraqi War and Social Media

It took less than 4 decades to transform the 1960's ARPANET military project to the widely used platform for research, education, entertainment and social communication that the Internet is today. This rapid emergence of Internet access and services provided, however, was not experienced uniformly around the world. The progress of the Internet in Iraq was interrupted by 2 wars, an embargo and state censorship. The Regional Documentation Center, hosted in Al-Mustansiriyyah University in Baghdad, had about 1000 dial-up users in 1989, and benefited from connections to Western database vendors [4]. These connections were cut the day Iraq invaded Kuwait on August 2, 1990 [7][8]. In 1998, the government established The General Company of Internet and Information Services, which allowed Internet access to Iraqi citizens 2 years later [2]. State monitoring didn't encourage Internet expansion, and by 2002, there were only 45,000 Internet users within a population of 24 million people [2]. The US invasion in 2003 “destroyed what remained of Iraq’s brittle infrastructure” [3]. But a few months later, engineers from the
State Company for Internet Services succeeded in establishing an Internet café with a satellite transceiver, 50 computers and a diesel generator [4]. There were 150 Internet cafés in Baghdad in 2004, and 2000 nationwide in 2007 [4]. In 2008, about 1% of the population had Internet access. Even if this value represents an important increase in Internet availability compared to previous years, the accessibility remains very low compared to the region and in the world (Figure 1).

![Figure 1: Growth of Internet users as a percentage of population](image)

“Why have persians taken to blogging so easily than arabs? why isn't there a single arabic weblog?” This post published in December 2002 was one of the earliest posts that marked the birth of the Iraqi blogosphere. The author, Salam Pax, is known for being the first Iraqi blogger [15]. In his blog *Where is Raed?*² and later *Shut up you fat whiner,*³ he started to report and

---

¹ The World Bank: http://data.worldbank.org/indicator/IT.NET.USER.P2/countries/IQ-1A-1W?display=graph
² http://dear_raed.blogspot.com
comment on various topics such as Saddam’s regime, finance, the Iraqi war, but also daily life and Internet accessibility challenges. Other Iraqi citizens have begun blogging since that time within Iraq and abroad. They expanded the coverage of topics to other interests including but not limited to art, culture and religion.

1.2. Thesis Background and Contribution

While information was for a long time monopolized by official sources and traditional media, the emergence of blogs and social media platforms provided a rich source of information that represent an interesting addition, if not an alternative to the existing ones. Project EPIC (Empowering the Public with Information in Crisis)\(^4\) is a “research effort to support the information needs by members of the public during times of mass emergency.” This research covers a wide range of disciplines from human behavior studies to policy issues to microblogging during natural hazards and blogging during times of political conflict.

As part of the EPIC project, the work behind this thesis focuses on the use of blogs during the Iraqi war by Iraqi citizens. In this dissertation, I start in chapter 2 by presenting my method for gathering and preprocessing the blogs. Then I discuss, in chapter 3, the challenges limiting the analysis of this data by some classical NLP approaches. Chapter 4 is dedicated to the choice of topic models as a promising alternative. In chapter 5 I evaluate the output of topic modeling with an external measure. I conclude in chapter 6 and provide an overview of future work.

\(^3\) http://justzipit.blogspot.com
\(^4\) http://epic.cs.colorado.edu
Chapter 2  Data Gathering

Collecting enough data is a first order step to allow quantitative analysis of any phenomenon. In this chapter I introduce the Blogger platform. I then present the crawling of blogs based on Blogger API. I describe the size of the dataset before discussing its preprocessing.

2.1. Blogger

Blogger is a blog-publishing platform that allows users to create blogs, publish posts and interact by comments. A blog is generally hosted at [subdomain].blogspot.com, and can be set to have public or private access. Figure 2 captures the main elements that compose the structure of Blogger service. A profile is a user account and can have multiple blogs. Each blog can have multiple posts. Each post can have multiple comments. A comment can be anonymous, that is its author is unknown, or owned by a profile, i.e. related to a blogger user account.
2.1. Blogger API

Blogger API is a framework that allows developers to access and edit blog contents through an ATOM feed based on Google Data API. The interaction between a client application and Blogger API can be performed via the HTTP requests GET, POST, PUT, and DELETE. As in this research I am interested in retrieving contents from Iraqi blogs, I will need to use only the request GET. Figure 3 illustrates an excerpt of the returned response of the request GET /feeds/12108309/posts/default. This is a call to get the posts of the blog of id 12108309. The response contains informations about the blog such as its title and owner (author). The response provides also a list of posts. A post is presented as an "entry" with an id, publication and update date, title and content (I didn't show all of the content because of space limitations).
The requests used in the crawler were

- **Profile request**: Retrieval of author information and the list of blogs owned by a given profile (GET /feeds/[profile-id]/blogs).

- **Blog request**: Retrieval of blog information and the list of posts for a given blog (GET /feeds/[blog-id]/posts/default).

- **Post request**: Retrieval of post information and the list of comments for a given post (GET /feeds/[blog-id]/[post-id]/comments/default).
In many cases, the owner of a blog may want to increase the traffic to his blog. He can, thus, set the blog feed to show only a summary of the post entry instead of the whole content. In this case, the Blogger API will not be sufficient for getting a post's full content, and I will have to get this content directly from the post's page.

### 2.2. Jericho

Jericho HTML parser\(^5\) is an open source java library that enables the parsing of HTML documents. One of the main reasons for choosing this API is its simplicity. Figure 4 shows a piece of code to get the content of the 5\(^{th}\) \(<P>\) tag from a page hosted at [URL].

![figure](image)

**Figure 4: Code Sample for Extracting the Content of a Tag from an HTML Page**

The main challenge, though, was that the design differs widely from one blog to another. Figure 5 shows different uses of the HTML tag \(<H3>\). In (a) it is used to format the date, while it is used in (b) to show the title. I found a solution to this problem by accessing the page that allows emailing, to a friend, the link to a post. This page, unlike the original post page, is style independent, and can be easily parsed by my crawler.

---

\(^5\) [http://jericho.htmlparser.net](http://jericho.htmlparser.net)
2.3. Crawling

I developed an application to crawl a list of Iraqi blogs indexed at iraqblogcount.blogspot.com and iraqiblogindex.blogspot.com to create the dataset. I chose to limit this crawler to the blogs hosted on blogspot.com because of the comprehensiveness and access to the Blogger API 2.0 (as described in section 2.1). These 2 indices contained 366 blogs; however 94 of them were not publicly available. While trying to access them, blogspot.com returns an error message stating that a blog either does not exist, or is not publicly available. A non-existent blog can be the result of either an indexation error from the original 2 indices, a removal by its author or a violation of terms of service. An author can also restrict the access to his blog to a predefined list of readers. I captured the remaining 272 in their entirety. Since many authors write multiple blogs, a scan of the 272 publicly available blogs led to the discovery of 185 additional blogs written by the same authors. I captured these blogs as well, leading to a final data set of 457 blogs containing 46,828 individual posts and 188,011 comments covering

6 http://glimpseofiraq.blogspot.com/2006/10/iraq-war-parties.html
7 http://iraqithoughts.blogspot.com/2006/12/end-of-tryant.html
8 http://www.blogger.com/terms.g
the period 1 May 2002 to 3 February 2011. For each post and comment, I extracted the title, author, content and publication and update date. I stored this data in a MySQL\textsuperscript{9} database.

The posts were then preprocessed to detect the language used to enable comparison of Arabic versus English posts. I found that the actual language of the content was not always equal to that indicated in the metadata of the pages. While the language mentioned in the metadata of Figure 6 is British English, it is clear that the content of the corresponding post is instead written in Arabic.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image.png}
\caption{Contradiction Between the Declared and the Used Languages}
\end{figure}

This contradiction led to the design the algorithm described in Figure 7. A post that had more than 50\% of the characters (after excluding email addresses and hyperlinks) as Arabic letters was classified as an Arabic post. Similarly, a post that had more than 50\% of the characters (after excluding email addresses and hyperlinks) as Latin letters was classified as an English post. This algorithm was based on the assumption that posts having dominant Arabic characters are Arabic posts even if there are other languages that are written in Arabic characters such as Persian and Urdu. I assumed also that posts having dominant Latin characters are English posts, even if these characters can be used in other European languages.

\textsuperscript{9}\url{http://www.mysql.com}
This scheme returned 11,668 Arabic posts and 31,246 English posts. In the remaining undetermined 3,914 posts, the dominant characters were neither Arabic nor Latin, and were excluded from the datasets.

In the example of Figure 8, the length of the post is 3087 characters. The count of Arabic characters is 1172 (i.e. 38% of the length of the post), and the count of English characters is 1304 (i.e. 42% of the length of the post). As none of these percentages exceeds 50%, the algorithm succeeded in excluding this post from the dataset.

**Function getLanguage(post):**

- \(\text{count(Arabic)} \leftarrow 0\)
- \(\text{count(Latin)} \leftarrow 0\)
- \(\text{count(Other)} \leftarrow 0\)
- Remove all URLs and email addresses.
- For each character \(c\) in the post:
  - If \(c\) is an Arabic character, then \(\text{count(Arabic)} \leftarrow \text{count(Arabic)} + 1\)
  - Else if \(c\) is a Latin character, then \(\text{count(Latin)} \leftarrow \text{count(Latin)} + 1\)
  - Else (i.e. \(c\) is a space, punctuation, special character or a character from any other alphabet), then \(\text{count(Other)} \leftarrow \text{count(Other)} + 1\)
- If \(\text{count(Arabic)} > \text{count(Latin)} + \text{count(Other)}\), then return "Arabic"
- Else if \(\text{count(Latin)} > \text{count(Arabic)} + \text{count(Other)}\), then return "English"
- Else return "Undetermined"
Every time I pass near it my eyes keep looking at the open windows showing the rooms and walkways on the upper floors. The quick glimpses and the rest that was filled out by my imagination made me have such a strong wish to spend there a couple of days when I get married. Even if I was given the choice to spend some time on the moon I would have chosen this special place with a special person.

It was one of the very few beautiful places in Basra. The big spaces, the beautifully designed wooden façade and the gorgeous gardens. The big festivities halls. And the many cafes where you can have a cappuccino and enjoy the view of the river which is right in front of it. But unfortunately it has been destroyed by the hands of people like many other things which if they have survived destruction thru military attacks were destroyed by people. How strange, how could someone not only have the nerve steal and loot but to let the fire eat these places for days. I often thought that these people were not sane.

A couple of days ago I went by the lace and saw a couple of people still trying to rip things from its walls, things which are surly worthless otherwise they would have been looted a long time ago. They were doing what they were doing very slowly with no hurry and wearing work overalls, they wouldn’t want their clothes to get dirty, would they? I realized they were not only insane but they totally believe in the legitimacy of what they are doing....... I turned my face away in sadness and shame although the Basra Sheraton is not the last thing to be destroyed in Basra nor the first thing whose destruction hurts me.
To verify the assumption, I manually coded a random sample of 1% of the posts that were classified as either Arabic or English posts. In all of the coded 117 Arabic posts, the dominant language was Arabic. For the coded 313 English posts, 311 of them were verified to be English. In the other 2 posts, one\textsuperscript{11} was all about code written in Java, and the other\textsuperscript{12} had only the word \textit{dEsIgN70s}. This verification demonstrates the reliability of this algorithm (100\% for Arabic and 99\% for English), and thus make the character-to-language assumption justifiable.

One case I encountered that shows a limit of this algorithm is illustrated in Figure 9. In this post, the author chose to use Latin characters to write Arabic words. The algorithm classified it as an English post though one would like to have it returned as undetermined. However, I didn't see a similar case in the random sample. Thus, I assume that this represents a relatively rare case.

\textbf{Figure 9: Arabic Text Written in Latin Characters,}\textsuperscript{13} with its Translation

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure9.png}
\caption{Figure 9: Arabic Text Written in Latin Characters,\textsuperscript{13} with its Translation}
\end{figure}

\begin{itemize}
\item \textit{Two years!}
\item Baghdad…
\item The city of peace and love, the city of science and art, is no more what it was.
\item Baghdad, where everyone is proud, and everyone is smiling, doesn't exist anymore!
\item An ugly occupation, lines of unemployed, scarcity in oil and gas, power outage and death transmission, is all what baghdad is about today.
\end{itemize}

\footnotesize
\textsuperscript{11} http://ukitech.blogspot.com/2008/07/hibernate-disjunction.html
\textsuperscript{12} http://shlonkombakazay.blogspot.com/2005/02/design70s.html
\textsuperscript{13} http://allahkareem.blogspot.com/2005/04/two-years.html
Figure 10 shows the evolution of publication activity in the crawled period for Arabic and English posts. Figure 11 presents the growth of active authors during the same period. An author is active in a month for a determined language if she has published at least one post in this month and language. Thus, an author might be active for both languages in the same month if she published 2 posts in 2 different languages, or might be inactive if she didn't have any publication in that month.

In both figures 12 and 13, I notice that English content shows a rapid increase, while the Arabic posts didn't emerge until the beginning of 2009.
2.4. Preprocessing

2.4.1. Morphological Analysis

Posts were tokenized and stemmed. The Porter stemmer [16] and Buckwalter morphological analyzer [7] were used respectively for English and Arabic posts. My preliminary work showed that the use of stemming was of some benefit for English, and critical for Arabic. In fact, it minimizes the loss of shared meaning of different forms of the same root. Table 1 shows for instance that the lemma countri was retrieved for its different forms: countries and country. This computation was more interesting for Arabic because of the morphological complexity of the language, as words can have a large number of forms when they get concatenated to articles and prepositions.
2.4.2. Porter stemmer

Porter's stemmer is considered to be the most common English stemmer. Porter's algorithm is 5 sequential steps that consist of the selection and application of a rule from a set of rules. Because of space considerations, I limit the description of the whole algorithm to some illustrations of rules' usage:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSES</td>
<td>caresses → caress</td>
</tr>
<tr>
<td>IES</td>
<td>countries → countri</td>
</tr>
<tr>
<td>Y → I</td>
<td>country → countri</td>
</tr>
<tr>
<td>SS → SS</td>
<td>caress → caress</td>
</tr>
<tr>
<td>S →</td>
<td>cats → cat</td>
</tr>
</tbody>
</table>

Table 1: Examples of Usage of Porter's Stemmer

2.4.3. Buckwalter morphological analyzer

Tim Buckwalter developed a morphological analyzer for the Arabic language. This analyzer is based on 3 lexicons: prefix, stem and suffix dictionaries. The parser tries to match the input to possible concatenation combinations and returns a list of possible fragments. Figure 12 shows the output provided by this morphological analyzer of an Arabic word. In this analysis, I kept only the first solution if many possibilities were returned.

Without stemming, words like "بلدي" (my country), "بلدك" (your country), and "بلدنا" (our country) would be considered by the algorithms presented in the next sections as independent terms even if linguistically the lemma "بلد" (country) is shared among them.
2.4.4. Indices properties

Table 2 summarizes the main properties of the indices. For each language, documents count is the count of posts. Vocabulary size is the count of unique words after stemming. Terms count is the count of all occurrences of all the words in the index.

<table>
<thead>
<tr>
<th></th>
<th>Arabic Index</th>
<th>English Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents count</td>
<td>11,668</td>
<td>31,246</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>73,250</td>
<td>126,100</td>
</tr>
<tr>
<td>Terms count</td>
<td>4,783,395</td>
<td>5,934,079</td>
</tr>
</tbody>
</table>

Table 2: Indices Properties
In an imaginary dataset of 2 documents *one fish two fish*, and *red fish blue fish*, the vocabulary size would be 5, and the terms count would be 8.

Now that I have the data gathered and preprocessed, I can start the quantitative analysis. In next chapter I present my first attempts with classification and clustering.
Chapter 3  Initial Analysis Attempts: Classification and Clustering

This chapter presents 2 different approaches for the analysis of crawled posts. The former is a supervised technique (classification), and the latter is an unsupervised one (clustering). I discuss the limits of both techniques to the special case that the crawled data represent.

3.1. Classification

Classification is a supervised machine learning group of methods that aim to assign documents to a predefined set of classes. Classes can be defined by rules provided by domain experts, or based on previously classified elements that can be used as a training data for the upcoming documents. The families of algorithms used most for text classification are Bayesian classification, vector space classification and support vector machines.

The only classification effort related to this research that I am aware of was performed by Al-Ani, Mark and Semaan [1]. The methodology as explained in this paper consisted of the
analysis of the first 5 posts of a sample of 27 Iraqi blogs. Each blog was then assigned to one of the four categories art, diary personal, diary war and journalism. This classification approach has two main issues; first the classification concerns blogs instead of posts, which is supposing that all of the content of a blog is strictly about one unique category. Second, in my preliminary investigation of the posts, I found some topics, for instance about religion and electronics, which cannot fit in these four categories. I decided then to categorize the data with an unsupervised approach to see if I could infer more knowledge.

3.2. Clustering

3.2.1. Definition

Clustering is an unsupervised machine learning set of algorithms that group documents into clusters based on their similarity. Documents from the same cluster should be as similar as possible, while documents from different clusters should be as dissimilar as possible. Similarity can be computed by a distance metric in which similarity between two documents is inversely proportional to the distance that separates them.

3.2.2. K-Means

K-means is a popular clustering algorithm that aims to minimize the average Euclidian distance between documents of the same cluster and its center or mean. This algorithm starts by selecting $K$ random documents as $K$ centers for the future clusters. Each document in the dataset will be assigned to the cluster of the nearest center. At the end of this iteration, a new center is computed for each cluster. These steps can be repeated until convergence is achieved or a fixed number of iterations is reached. The final result is a subdivision of the dataset into $K$ distinct clusters.
3.2.3. Implementation

I implemented the K-Means algorithm as described in Figure 13 on the English and Arabic posts separately. I based the implementation on Lucene\textsuperscript{14}. I used the cosine similarity to compute the similarity $\text{sim}(d, d')$ between two documents $d$ and $d'$.

1. Create $k$ empty clusters.
2. Select $k$ random posts from the dataset, and assign each as a centroid to a cluster.
3. While computation has not converged yet
   a. For each document in the dataset different than the centroids
      i. Compute the distance between this document and each centroid.
      ii. Assign the document to the cluster that minimizes this distance, i.e. The cluster that maximizes the similarity among its centroid in the document.
   b. For each cluster
      i. For each document $d$: $s(d) \leftarrow 0$
         • For each document $d'$ different than $d$: $s(d) \leftarrow s(d) + \text{sim}(d, d')$
      ii. Choose the document $d$ that maximizes $s(d)$ and set it as a new centroid.
4. Return $k$ clusters

Figure 13 - K-means Algorithm

3.2.4. Results

I ran my implementation first to retrieve 10 clusters. Table 3 presents the most frequent terms in each cluster. The reader can easily identify the dominant topics of many clusters. Cluster CE02 contains probably religious posts, as we find terms like god, bible, spirit, jesus,

\textsuperscript{14} http://lucene.apache.org
faith. CE03 can be a cluster of French posts not excluded from the dataset with the language detection algorithm presented in section Figure 7. CE05 may be about online entertainment. However, 9 of the 10 clusters represent less than 8% of the number of English posts (Figure 14). All of the other 92% are clustered together in CE09. Unfortunately, it is difficult to infer one dominant topic in this cluster. One may instead retrieve a daily life topic (time, people, work, call, friend), a political topic (government, american) and a violence topic (war, kill).

<table>
<thead>
<tr>
<th>ID</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE01</td>
<td>upload origin nofril tokyo janes photo ル flickr カ の info html net seesaa blog articl ・て 片倉 japan いかり augus らと view read は</td>
</tr>
<tr>
<td>CE02</td>
<td>god day bibl spirit jesus road faith life post thing love morn time holi light live yesterday avila philippin talk motiv call peac better mel save peopl alarilla christ tag</td>
</tr>
<tr>
<td>CE03</td>
<td>de la en le les il se à dan est des qui une du pour au ce sur tout je comm ne par pas mai avec son san bien leur</td>
</tr>
<tr>
<td>CE04</td>
<td>pictur site ann post googl cat hey i'm time you'r nice coulter check search read well day link pic year cool laugh good trip click peopl top video hope photo</td>
</tr>
<tr>
<td>CE05</td>
<td>site news googl click internet web check onlin user start blog work provid busi compani armenian servic file develop websit imag market game video free network access find armenia creat</td>
</tr>
<tr>
<td>CE06</td>
<td>american soldier museum armenian nativ video news アラブ戦争 genocid idol read muslim william barb street mosque kid pride washington agress file jew movi sculptor recogn ot op wink turk heartless</td>
</tr>
<tr>
<td>CE07</td>
<td>japan tokyo upload nofril origin tower water の augus 水 給 estat click 塔 jutaku asagaya view octob 阿北宅地 に 住あ佐 residenci morn ケ imag</td>
</tr>
<tr>
<td>CE08</td>
<td>love time good life post day blog friend happi i'm peopl read thing best learn year feel live long help heart find true nice go well care lot better thought</td>
</tr>
<tr>
<td>CE09</td>
<td>time peopl iraq iraki day year thing well good work call live countri al go start baghdad today govern post life read war american place long kill friend talk feel</td>
</tr>
<tr>
<td>CE10</td>
<td>hello post baghdad amman cairo pictur storey shop morn view today center commerci garden hous small beauti roman flower ladi wear museum good ceram yesterday room egyptian north mosqu piec</td>
</tr>
</tbody>
</table>

Table 3: Most Frequent Terms in a Partitioning of 10 English Clusters (CE = Cluster in English)
I then tried another value of $k$. For $k = 20$, I made the same observation (Table 4, Figure 15). There was one dominant topic (CE’17) with more than 88% of the posts. And the frequent terms are almost the same encountered in CE09, and thus one can't conclude what is the distribution of the topics covered in this cluster.

<table>
<thead>
<tr>
<th>ID</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE’01</td>
<td>test sourc damascus t cat amman click receiv cute view ic mosqu comput snowboardbrill المرا حل logitech kinder chicago lernspielzeug webcam larger streetscen desk flower amawi class fuck big laubsaug code</td>
</tr>
<tr>
<td>CE’02</td>
<td>blog post day time blogger iraqi read start baghdad link write iraq comment check today i’m good pictur arab year find interest thing hello hope site al updat friend email</td>
</tr>
<tr>
<td>CE’03</td>
<td>god day bibl spirit road faith life post thing love morn time jesus holi light live yesterday avila philippin talk motiv call peac better mel save peopl alarilla christ tag</td>
</tr>
<tr>
<td>CE'04</td>
<td>fayrouz hancock mark txa pictur enjoy visit love photo display day photograph time camera newseagl life food festiv garden usual place dalla beauti beaumont cover work i'm famili nice good</td>
</tr>
<tr>
<td>CE'05</td>
<td>de la en le les il se à dan est des qui une du pour au ce sur tout je comm ne par pas mai avec son san bien leur</td>
</tr>
<tr>
<td>CE'06</td>
<td>dead fish boat iraqi left fire govern port ingredi persian cultur sea investig equip georg lorna appear vartan relax bfcdb chocolati chocol languag serviceman french desir spot defens free milk</td>
</tr>
<tr>
<td>CE'07</td>
<td>pictur ann site googl i'm hey search you'r cat well coulter time read check trip click pic nice year laugh peopl post video go doctor hope alright celebr cool news</td>
</tr>
<tr>
<td>CE'08</td>
<td>translat note ra html common librari post apach bit valid april help file consid mix gwt javascript review write site code size input hope て の 4 い inconveni し</td>
</tr>
<tr>
<td>CE'09</td>
<td>site googl click internet web user onlin work compani check busi provid servic develop start imag market game file websit video access free network find time good open set includ</td>
</tr>
<tr>
<td>CE'10</td>
<td>news armenian armenia asbarez turkey panorama visit turkish azerbaijan msnbc break economi hyemedia polit presid facebook hot aysor net panarmenian aliyev time azerbaijani link wikileaks church expert attack hold good</td>
</tr>
<tr>
<td>CE'11</td>
<td>al link allah prophet jesus محمد ha moham islam symbol creat hay el plural testament text الإبراهيمية القياسية code var azim vav creator chapter ﷺ لام</td>
</tr>
<tr>
<td>CE'12</td>
<td>photo read origin build upload nofril flickr japanes articl net html blog seesaa ルビ tokyo 京の目り septemb hurri pictur info demolish 大offic 片倉</td>
</tr>
<tr>
<td>CE'13</td>
<td>japan upload tokyo nofril origin tower water の auguts 水 給 estat click 塔 jutaku asagaya view octob 阿北宅地に住あげ佐 residenci morn 住 imag</td>
</tr>
<tr>
<td>CE'14</td>
<td>upload origin nofril tokyo august japan view wall march cherri tv close larg roppongi photo ル mark screen tuma pictur day の paint nagai 晩 dawn report wiki kenji yellow</td>
</tr>
<tr>
<td>CE'15</td>
<td>waghorn eduardo learn chile copyright nada hace la escribir friend para hug por te comienzo industri dolor pleno medley seguiré medio se tren año doquier record una memoria sentir hay</td>
</tr>
</tbody>
</table>
| CE'16                  | hello post amman baghdad cairo shop morn view hous garden small today beauti
|                       | roman ladi wear flower museum ceram good yesterday egyptian piec build downtown
|                       | kitten ruin insid apart plant |
| CE'17                 | time peopl iraq iraqi day year thing well good work call live countri al go start
|                       | baghdad govern today life war read american post kill place long friend talk feel |
| CE'18                 | armenian genocid news armenia call recogn mp asbarez resolut turkey net
|                       | panarmenian turkish eu court karabakh azerbaijan european hous pelosi polit vote
|                       | download panorama recognit time anca res azerbaijani hyemedia |
| CE'19                 | net panarmenian upload 片倉 nofril seesaa articl html info ル・ビ blog
|                       | origin のかてにいりら t 口またすっは |
| CE'20                 | upload nofril origin photo blue descript english janeses short flickr info box ていが
|                       | のる内所都某はにうかしたりれも |

Table 4: Most Frequent Terms in a Partitioning of 20 English Clusters

Figure 15: Distribution of Documents Over a Partitioning of 20 English Clusters
I argue that these results are due to the size of the posts (few paragraphs), which allows bloggers, unlike in micro-blogging platforms like Twitter, to cover multiple topics in a single post. This led me to try another alternative: topic modeling.
Chapter 4  Topic Modeling

After experiencing the limits of classification and clustering in the previous chapter, I investigate, in this one, another unsupervised technique: topic modeling. Once I define this technique, I discuss some of the related work in the literature. I next present my implementation and show my findings.

4.1. Definition

Topic models are probabilistic models for discovering latent topics of a set of documents. The idea behind these models is that words presented in a document are generated from a hidden process based on their association to some topics. In other terms, words like cats and dogs are more likely to appear together in a document rather than showing up with stock market. Thus, the process tries to find the mixture of topics that generated the documents in the corpus, while a topic itself is a probability distribution of correlated words over the vocabulary [5]. Latent
Dirichlet Allocation (LDA) is the most common use of topic models. The LDA model assumes the following generative process for each document $d$ in a corpus $D$:

1. Choose the topic distribution $\theta_d$ from a uniform Dirichlet prior with parameter $\alpha$.
   \[ \theta_d \sim \text{Dir}(\alpha) \]

2. For each word $w_d$ in the document $d$:
   a. Choose a topic $z_{w,d}$ from a multinomial distribution with parameter $\theta_d$.
   \[ z_{w,d} \sim \text{Multinomial}(\theta_d) \]
   b. Choose a word $w_d$ from a multinomial probability with parameter $\beta$ conditioned on the topic $z_{w,d}$.
   \[ w_d \sim \text{Multinomial}(\beta_{Z_{w,d}}) \]

4.2. Related Work

Since the introduction of LDA in 2003, it has been applied to a wide variety of domains and applications, many of which include correlations with, and in some cases predictions of, events external to the model.\textsuperscript{15} Yano, Cohen and Smith applied two variants of LDA to predict response to political blogs [19]. The first, LinkLDA, models the users most likely to react to a post. The second, CommentLDA, guesses the content of these comments.

Ramage, Dumais and Liebling tried to take the study of microblogging services "beyond their traditional roles as social networks" by implementing a Labeled LDA to translate the content of tweets into multidimensional characteristics [17]. This partially unsupervised learning

\textsuperscript{15} Parts of this section were taken from [3].
model categorizes the learned dimensions into four categories: substance (events and ideas), social communication, status (personal updates) and language style.

Polylingual Topic Models were introduced by Mimno et al. to discover parallel topics across multilingual corpora that are either official translations (EuroParl) or corresponding articles from different languages (Wikipedia) [14]. These models were also used to compute the divergence between pairs of documents and infer the score of disagreement between languages.

Topic models were also presented as an approach to track the drift of topics over time in research fields by Hall, Jurafsky and Manning [10] as well as in weblogs related to "Toyota" and "iPhone" by Knights, Mozer and Nicolov [13]. In a preliminary examination, Kireyev, Palen and Anderson applied topic models to crisis-related microblog posts [12].

Some work was done to validate the analysis of social media from external facts. Bollen, Mao and Zeng investigated the correlation between the Dow Jones Industrial Average and Twitter feeds to find a causality relation between mood dimensions and stock market trends [6].

4.3. Implementation

I applied the Matlab Topic Modeling Toolbox [9] to the crawled collection. This toolbox needs 2 files as input. One (Table 5a) is a list of the unique words in the index. For each language, the size of this list corresponds to the vocabulary size presented in Table 2. The other (Table 5b) is a file containing the assignment of each word to its document. That is, in each line there are the ID of the document, the ID of the word and the number of occurrences of this word in that document. Thus, if a word $w$ doesn’t exist in a document $d$, there is no entry in the file for the tuple $(w, d)$. If this word was mentioned $n$ times in that document, the file should have a corresponding entry of 3 values: $w, d$ and $n$. 
The toolbox uses Gibbs sampling to discover topics. The Gibbs sampler is a Markov chain Monte Carlo algorithm that can be used to infer the distribution that generated a set of observations. In the case of LDA, the observations are the documents with their words, and the distribution is the mixture of topics. This iterative algorithm starts by a random assignment of the values, which means that, at the beginning, each term gets a random topic. Then Gibbs is run over a fixed number of iterations or until convergence is reached. During each iteration, all of the terms are picked one by one randomly, and a topic is sampled based on the topic assignment of the previous terms of the same iteration. The reader should see [11] for more detailed explanation of the use of Gibbs sampler for the approximate inference of LDA.

This computation results in a matrix (M₁) that contains the assignment of each term to a topic. For the Iraqi crawled posts, this means that each of the 5,934,079 (4,783,395) terms in the English (Arabic) dataset has an assigned topic. This matrix is then used to generate two others. The first (M₂, Table 6) contains how many times a word of the vocabulary was assigned to each topic. The sum of each line is, thus, the count of times that word appears in the dataset. The second (M₃, Table 7) encloses the count of times the terms of a document were assigned to each topic. The sum of each line is, hence, the word counts of the corresponding document.

Table 5 (a, b): Structure of Input Files for MATLAB Toolbox

<table>
<thead>
<tr>
<th>Line number</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>47757</td>
<td>iraq</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>105524</td>
<td>war</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Document ID</th>
<th>Word ID</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>47757</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>47757</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>47757</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>47757</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Topic ID</td>
<td>Topic 1</td>
<td>Topic 2</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>47757</td>
<td>1</td>
<td>38895</td>
</tr>
<tr>
<td>47758</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Table 6: Matrix of Word/Topic Distribution**

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>...</th>
<th>Topic T</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>1</td>
<td>...</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>0</td>
<td>...</td>
<td>13</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Table 7: Matrix of Document/Topic Distribution**

M₂ was then used to get an order list of the most relevant words to each topic. In Table 6, the word of ID 47758 is more relevant to Topic 1 than the word of ID 47757.

I used M₃ to get the distribution of topics for each document as a percentage after normalizing the counts.

As the toolbox’s developers suggest, I set the hyper-parameters Alpha and Beta to 50/T and 200/V where T is the number of topics and V is the size of the vocabulary. I ran 100 iterations. I tried various values of T. Table 8 shows the most frequent terms for the English posts in a portioning of 20 topics. I argue that the data set was “over-partitioned” in this case. In fact, I found many topics sharing the same meaning based on the most frequent terms. For instance topics TE’02 and TE’04 both have terms related to international politics: Lebanon, Israel, Arab, Iran, Middle, East, America and Turkey. TE’07 and TE’13 talk about daily life and share the words life, time and talk. The reader can also find similarities between topics TE’09
and TE’14, and between topics TE’10 and TE’19. In addition, I couldn’t make sense of the correlation, in topic TE’01, between music terms (*song* and *Fayrouz*\(^{16}\)), French words (*de*, *la*, *en*, *le*, *il* and *les*) and Arabic words (*مان*، *في*، *فم*).  

<table>
<thead>
<tr>
<th>ID</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE’01</td>
<td>de la hubbi Google en el le song file من في <em>Fayrouz</em> il Hancock les</td>
</tr>
<tr>
<td>TE’02</td>
<td>report prison Lebanon offic israel intern torture human court year author right release investigate charge</td>
</tr>
<tr>
<td>TE’03</td>
<td>book university year work play music team art student film school great study city</td>
</tr>
<tr>
<td>TE’04</td>
<td>Arab Iran Country Middle politic Iranian Israel east power support region regime nation America Turkey</td>
</tr>
<tr>
<td>TE’05</td>
<td>Iraqi election govern party Iraq vote politic Kurdish Kurd minister nation member constitution Maliki Kurdistan</td>
</tr>
<tr>
<td>TE’06</td>
<td>hand eye head man face turn left black stand walk street open door light move</td>
</tr>
<tr>
<td>TE’07</td>
<td>love time day feel I’m guy life remember happy ask smile call talk dream listen</td>
</tr>
<tr>
<td>TE’08</td>
<td>oil time year model t rate figure product nature number system term field discovery</td>
</tr>
<tr>
<td>TE’09</td>
<td>kill baghdad police bomb attack soldier Iraqi force wound army report military city</td>
</tr>
<tr>
<td>TE’10</td>
<td>news picture video tv site photo media busy Internet report image journalist watch</td>
</tr>
<tr>
<td>TE’11</td>
<td>muslim women God Islam Arab christian religion men man woman Allah live peace people life</td>
</tr>
<tr>
<td>TE’12</td>
<td>fact point case question matter discuss issue view interest public social person object reason article</td>
</tr>
<tr>
<td>TE’13</td>
<td>thing people good time go well I’m bad better life lot person talk thought change</td>
</tr>
<tr>
<td>TE’14</td>
<td>Iraqi people Iraq country Saddam live kill American war happen govern year help life</td>
</tr>
<tr>
<td>TE’15</td>
<td>receive radio turn high circuit connect control power signal design low set work</td>
</tr>
</tbody>
</table>

\(^{16}\) Fayrouz is the name of a famous Arab singer.
On the opposite side, Table 9 shows that using only 5 topics caused the merging of separable themes. For example, Topic TE’’03 could be separated into 2 different topics, one for local politics (Iraqi, Iraq, politic, Saddam and govern), and the other for international politics (Arab, politic, muslim and American). Similarly, topic TE’’04 could be split into 3 topics: economy (oil and work), local politics (vote, election and Kurdish) and electronics (high, receive, power and turn).

I decided at the end to work with $T = 10$ topics. This doesn’t necessary imply that this is the perfect choice, but I believe it is a good one, specially that it was validated by human analysis. I used the same number of topics for both datasets (Arabic and English).
An alternative was provided by Griffiths and Steyvers to select $T$ [9]. They proposed to estimate the likelihood $P(w|T)$ for different values of $T$. This likelihood, as a function of $T$, increases until reaching a maximum $T_0$, and then decreases. They suggest that this $T_0$ is the optimum number of topics.

### 4.4. Results

The ten topics discovered by the English and Arabic models are presented in Tables 5 and 6, respectively. The topics are presented in descending order of frequency, and each topic is illustrated with the words most relevant to the topic.

<table>
<thead>
<tr>
<th>ID</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE01</td>
<td>People Iraqi Iraq Country American war Saddam live year kill America happen terrorist government</td>
</tr>
<tr>
<td>TE02</td>
<td>year work company busy develop tea market project system service money provide case program</td>
</tr>
<tr>
<td>TE03</td>
<td>day time t friend thing I'm feel start house talk go work good told car</td>
</tr>
<tr>
<td>TE04</td>
<td>Iraq Baghdad kill Iraqi force war attack military police bomb report American soldier bush army</td>
</tr>
<tr>
<td>TE05</td>
<td>Iraq al Iraqi govern politic elect Sunni party Iran Arab nation vote leader minister Kurdish</td>
</tr>
<tr>
<td>TE06</td>
<td>god women Muslim love life man Islam men heart person religion woman Christian face eye</td>
</tr>
<tr>
<td>TE07</td>
<td>oil de time receive model turn high t power radio la circuit set work point</td>
</tr>
<tr>
<td>TE08</td>
<td>al news year day today call Baghdad ago Abu watch TV video family play story</td>
</tr>
<tr>
<td>TE09</td>
<td>water picture city place black build white small food long game green open red photo</td>
</tr>
<tr>
<td>TE10</td>
<td>post read blog write Arab book comment university time blogger music link interest student publish</td>
</tr>
</tbody>
</table>

Table 10: Most Frequent Terms in the English Posts by Topic
<table>
<thead>
<tr>
<th>ID</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA01</td>
<td>One want elapse days same go people day man saying listen know say was</td>
</tr>
<tr>
<td>TA02</td>
<td>Life times poetry love amiability eye heart death night art soul beautiful spirit face sound</td>
</tr>
<tr>
<td>TA03</td>
<td>Allah son say peace the Imam pray Islam Mohammed saying kin father worshipper Muslim prophet</td>
</tr>
<tr>
<td>TA04</td>
<td>Iraq the force government president say council security governorate operation Baghdad visit America elections Al-Maliki</td>
</tr>
<tr>
<td>TA05</td>
<td>Other many book enable write period something person same one do picture time site more</td>
</tr>
<tr>
<td>TA06</td>
<td>The scholar religion history other Arab saying knowledge idea first origin culture earth own soul</td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>TA07</td>
<td>Iraq America Arab people occupation Saddam party war Ba'ath Iran nation force Israel resistant</td>
</tr>
<tr>
<td>TA08</td>
<td>The Baghdad one expand day logic hour children poison ten three big kill car</td>
</tr>
<tr>
<td>TA09</td>
<td>United country politic operation united state no the government force meeting general society new other rights</td>
</tr>
<tr>
<td>TA10</td>
<td>Work year media company information office studying network university service program newspapers director system programs</td>
</tr>
</tbody>
</table>

Table 11: Translations of the Most Frequent Terms in the Arabic Posts by Topic (TA = Topic in Arabic)
To give an example, a single post (Figure 18) was analyzed and produced the distribution of topics present in this post (Figure 19), along with topic assignment to each token (Table 12).

Figure 18: Example of a Post

Figure 19: Distribution of Topics for the Post in Figure 18

http://thoughtsfrombaghdad.blogspot.com/2008/03/all-roads-lead-to-baghdad.html
I presented in this chapter the topics discovered in Arabic and English datasets. The next chapter will be dedicated to evaluating these results against an external source.
Chapter 5  Evaluation

The results of any unsupervised technique become more interesting if they can be evaluated against known facts from an external resource. In this chapter I demonstrate the success of the use of topic models by showing the high correlation of the timeline of war related topics with the Iraqi Body Count website.

In order to evaluate the output of the topic models, I looked for an external measure that captures the ongoing events during the war. Iraq Body Count\textsuperscript{18} (IBC) is an independent nonprofit organization that made public its database of the "deaths caused by US-led coalition forces and paramilitary or criminal attacks by others" since 2003. Unlike other sources that provide estimates, IBC numbers are based on recorded and documented violent deaths from various sources such as journalists, hospitals, morgues, NGOs, and official statistics. Figure 20 illustrates

\textsuperscript{18}http://www.iraqbodycount.org
the timeline of the body count for a period of 8 years as retrieved from IBC on February 16, 2011. The website notes that: “IBC’s figures are constantly updated and revised as new data comes in, and frequent consultation is advised.”

The main topics that can be related to the body count are the ones that have more weight for war and violence terms. In TE01 and TE04 the reader can see, among the most frequent terms, the words war, kill, terrorist, attack, military, bomb, report, soldier and army. I merged these topics in the timeline presented by Figure 21. Similarly, I considered TA07 and TA08 as war topics for Arabic posts because of the appearance of the terms occupation, war, force, resistant, poison and kill. Their timeline is presented in Figure 22.
I found a highly significant positive correlation between the war topics and the body count. Lagging the body count after the war topics in English by two months I got a correlation coefficient $R = 0.8551$ and a level of significance $p < 0.001$ for a number of pairs $N = 84$. I limited the computation to the period 2004 – 2010, as in 2003 only a few posts were published (Figure 10). The high attention to violence themes by the bloggers in 2006 could be expected during a catastrophic condition described by the UN as “a civil war-like situation” [18]. This said, I can’t explain why the blogs were predicting the body count. I hypothesized that bloggers were updating their posts after external events occurred, and I used posts’ update-date instead of publication-date. I got a very similar correlation coefficient ($R = 0.8591$) for the same advance of two months. Further research is needed to clarify this behavior of relationship.
The correlation of war topics in Arabic was less significant. This could be expected given that the blogging activity in Arabic didn’t emerge until 2009 (Figure 10). I claim that English posts are most likely to be generated by Iraqi bloggers living outside Iraq, and thus they don't face the lack of Internet accessibility that Arabic bloggers faced.

The high correlation of discovered topics with the body count confirmed the validity of the choice of topic models as a method for analyzing the big dataset of the blogs that I crawled. In the next chapter I review the problematic of the thesis and conclude with future work.
Chapter 6  Conclusion and Future Work

This thesis presented a method for the analysis of overabundant social data related to political crises. I showed series of steps and decisions that I took over the conduct of this project. I first introduced a Blogger-based crawler. I then proposed an algorithm for distinguishing Arabic and English posts in two independent datasets. In a preprocessing phase, I chose to integrate a morphological analyzer into each of them. Next I discussed the limitation of classification and demonstrated the inadequacy of clustering as a technique for investigating this specific data. Afterward I presented an implementation of topic models and showed its efficiency by finding a high correlation between the topics discovered and an external measure (the Iraqi body count). Overall, topic modeling can provide a rapid insight into a big dataset that we don’t have any a priori knowledge about. The alternative of manually analyzing the data is challenging because finding a random representative sample may not be possible if the number of documents exceeds several thousands.

Some improvement to this effort can be done in future work. In any particular dataset that is similar to the one presented in this thesis, in which there is a prior knowledge of the dominant
languages coming from different alphabets, the algorithm presented in Figure 7 can be adapted such that the number of possible outputs will be equal to the number of distinct alphabets. Nevertheless, this algorithm will fail in a mixture of documents, for instance, from various European languages. As section 3.2.4 shows, clustering returned a group of documents in which the most frequent terms are French words. Clustering can be used, then, for language detection instead of a character-count-based algorithm. A machine translation service can also be used to avoid eliminating content that is available in languages that a researcher doesn’t understand.

In addition, I plan to verify if the same adopted approach and choices will work for other datasets. I think of using this work toward the analysis of data being collected these days about the political changes in the Middle East, as part of the efforts being made in the EPIC project. One challenge could be the adaptation of the algorithms to microblogging based data, known for the challenges presented by length limits, versus the relatively long posts that I have collected from Iraqi blogs.

The analysis of the comments can enrich this work. They can be used to get a better understanding of the interaction between the bloggers or between authors and their followers. In a preliminary trial, I found that topic models can be used for spam detection on the comments. In a run of 10 topics, I got two clear spam topics: pornography and gambling. After eliminating the comments that have most contribution to these topics, I may get cleaner data. I can also use comments for running topic modeling on all of the data instead of having two separate runs (one for English and another Arabic). I want to verify if comments help to correlate Arabic and English posts. In fact, comments are not necessarily published with the same language of their posts, and thus topic modeling might discover correlated words from different languages.


