Statistical Extraction and Visualization of Topics in the Qur'an Corpus

by

Maysum H. Panju

A research paper
presented to the University of Waterloo
in partial fulfillment of the
requirement for the degree of
Master of Mathematics
in
Computational Mathematics

Supervisor: Prof. Ali Ghodsi

Waterloo, Ontario, Canada, 2014

I hereby declare that I am the sole author of this report. This is a true copy of the report, including any required final revisions, as accepted by my examiners.

I understand that my report may be made electronically available to the public.

Abstract

Unsupervised machine learning techniques are described and applied on the mildly preprocessed Arabic text of the Holy Qur'an, with promising results. Topic modelling based on nonnegative matrix factorization was used to successfully extract meaningful topics underlying the set of 6236 verses in the corpus. Data visualization using t-SNE dimensionality reduction correctly grouped verses of the Holy Qur'an into clusters based on theme and word usage. This accessible paper begins with an introductory view of machine learning, and includes motivating descriptions of the implemented techniques before presenting a summary of findings. A graphical display combining the results of topic modelling and data visualization demonstrates the consistency of the studied models.

Acknowledgements

I would like to thank my supervisor, Dr. Ali Ghodsi, for accepting me as a student and guiding me towards and along a path of study I feel passionate about.

This work could not have been possible without the resources available from the Qur'an Corpus project, under the leadership of Kais Dukes of Leeds University. I am grateful to have such a wonderful framework to build upon.

Thanks also goes to all of my family, friends, roommates, and dear ones for supporting me during this exciting year of graduate school.

Finally, a special thanks goes to my beautiful mother, who made sure to instil a love of the Holy Qur'an when I was younger.

Dedication

This small work is dedicated to
the Living Companion of the Holy Qur'an

May God hasten his advent

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

When there is too much data, it can be difficult to obtain a "big picture" understanding of what the data represents. This is true in many scenarios, but particularly true when trying to summarize general ideas from large collections of text documents. Although each individual document may contain particular merit, sometimes the general summary of the entire collection provides a greater insight on the underlying message.

However, extracting meaningful summaries of text collections can be difficult, especially for massive datasets. With the rise of machine learning in recent times, this task became much more feasible, as it became possible to identify trends and patterns in a corpus of text algorithmically. Methods to uncover intrinsic patterns and structure underlying data were developed, and techniques for massively reducing the complexity of enormous datasets were devised. Unsupervised learning algorithms made it possible to model hidden topics and visualize incomprehensible data in ways that would not have been possible using traditional methods.

The focus of this work is to apply such techniques on the text of the Holy Qur'an, the book of the religion Islam. As a sacred text revered for its linguistic and spiritual value, the Holy Qur'an has been the subject of numerous studies and analyses performed manually. This work intends to further the progress in this field by performing methods of unsupervised machine learning on the Arabic text. Applying statistical and computational methods on this corpus, free of human intervention, have the potential of revealing intrinsic structure behind the ancient text that could not have been uncovered by any manual investigation.

In particular, two main approaches of study will be taken. First, topic modelling will be applied to the Holy Qur'an in order to discover what topics exist within the text, as dictated by the words of the book itself, and not influenced by any human annotator. Second, a data visualization technique will allow the verses of the Holy Qur'an to be displayed in a unique representation that demonstrates their relationship with each other based on content, rather than by chapter. Finally, these two approaches can be combined to present a stunning graphical presentation of the verses of the Qur'an organized by topical structure. The success of this combined image is suggestive of the cooperative potential of the independent methods of topic modelling and data visualization.

The remainder of this essay is laid out as follows. Chapter 2 provides some background on machine learning, the nature of the Holy Qur'an, and how others have approached this study in the past. Chapter 3 outlines the formation of the fundamental term-document matrix, as well as descriptions of the topic modelling and data visualization problems. Section 4 demonstrates some of the results of these methods, illustrating the ability to automatically identify underlying structure of the Holy Qur'an. Finally, concluding remarks and next steps are given in Chapter 5.

It should be clarified here that although the nature of this research delves very closely into examining the content of the Holy Qur'an, this work is meant to only be an analysis of statistical procedures and not a study of religious material. The literal value of words used in the Holy Qur'an will be handled extensively, but inferences beyond that are not within the scope of this project.

Chapter 2

Background

We begin by presenting some background information on the concept of unsupervised learning, the nature of the Holy Qur'an, and a glimpse at some of the previous work in computational study of the Holy Qur'an.

2.1 Unsupervised Learning on Text

Machine learning is the practice of identifying and discovering patterns in large sets of data. There are many frameworks in which this may take place, depending upon the situation and the learning environment. One of the most important distinctions between techniques is that of supervision, or whether or not there is a "teacher" to aid in the learning process.

In supervised learning [13], the machine[1] is given a sample of data consisting of information about objects and the correct way that they should be treated, according to some supervising "teacher". This "treatment" may refer to the attachment of a label, or assigning of a value, or a similar operation, depending on the context. The machine's goal is to learn what the pattern is so that if it is given a new object in the future, it will be able to determine the correct treatment without assistance.

Supervised learning has found great use in the task of classifying text documents into several predefined, discrete categories [20]. A teacher manually assigns categories to several documents reserved for the purpose of training the machine. Once the machine examines these documents and how they were labelled, it is able to classify new documents in a

[1] This may be any agent implementing a learning algorithm; it need not be an actual physical "machine".

similar way automatically. Such an approach to classification is highly effective since it is not required to explicitly define the specification of each category, but only to supply a set of examples instead.

Unsupervised learning [10] is when patterns are to be found in data, but with no teacher guiding the process. The machine is only provided with raw data and is expected to identify patterns on its own. Since there is no notion of "correctness", it is harder to control what kind of patterns may be found. Typically, the result of such a learning algorithm is a clustering system that identifies groups of input units that seem likely to be associated with each other. Generally, if the dataset is large enough and the varieties of input are distinct enough, an unsupervised learning algorithm should be able to divide input units into classes in a very structured and reasonable way.

One application of unsupervised learning is to uncover patterns and similarities within a collection of text documents, without external influence or annotation [8, 6]. Although relationships between words and topics in documents may be observable by inspection, statistical methods that are unbiased by human tendencies may cause hidden patterns to emerge. These patterns allow a deeper understanding of the structure that is instrinsic within the text itself, independent of the researcher performing the study. It is this type of approach that will be taken in this work, in an attempt to examine the patterns and structure inherent within a corpus of enormous literary, historical, and religious significance: the Holy Qur'an.

2.2 What is the Holy Qur'an?

The Holy Qur'an is the sacred text of the religion of Islam, and is accepted by 1.6 billion Muslims around the world to be the divine word of God as revealed to Muhammad, the Prophet of Islam [23]. The book was intermittently revealed in portions across a period of 23 years in the early seventh century, and is widely accepted as intact and unchanged since its original compilation over 1400 years ago. The Arabic text has been considered one of the greatest masterpieces in classical Arabic literature, and remains the object of careful study by both religious scholars and linguists until today.

The Holy Qur'an is composed of 6236 verses, divided up into 114 chapters of vastly varying sizes. Although it has been translated dozens of times into hundreds of languages, the original text in pure, classical Arabic has been carefully preserved, and is the only form that is accepted as authentic. The text is written in various styles of rhythmic poetry, gradually transitioning in form and tone as the revelation progressed. The content and theme within the verses also adapted as time went on.

Unlike most books, the Holy Qur'an is not structured in a linear format. There is no consistent sequence underlying the ordering of the chapters, and even within a single chapter, many distinct topics may be addressed with little or no transition in between. A concept that is discussed briefly within a few verses is often readdressed later in a different passage, many times in a different chapter entirely. This seemingly scattered arrangement, a by-product of the sporadic nature of the revelation and compilation process, serves as a reminder that the Holy Qur'an is meant to be a book of guidance and a reference, rather than a sequenced story or ordered instruction manual. As a result, the reader of the Holy Qur'an is presented with an assortment of topics within a small passages, and the entire text must be read and taken into account in order to consider all verses that deal with a single topic.

With its intriguing poetic structure as well as its enormous religious significance, the Holy Qur'an has attracted the scholarly attention of linguists and clerics alike. Its verses and chapters have been carefully memorized and studied ever since time of its revelation, and have continued until today. While traditional linguistic study of the Holy Qur'an was limited to manual inspection and observation, the advance of technology made advanced statistical and computational analysis possible. The richness of the text and the deep layers of its meaning offer enormous potential for further study that is far from being exhausted.

To summarize, here are some of the reasons for why it would be worthwhile to apply unsupervised learning techniques on the Holy Qur'an:

- Themes and patterns in the Holy Qur'an, previously sought through slow and painstaking manual labours, may be discovered rapidly with the aid of technology.
- As a collection of over 6000 verses spanning a variety of scattered, unordered, and repetitive concepts, the Holy Qur'an is well suited for unsupervised learning methods such as topic modelling and verse clustering.
- The statistically generated results of machine learning algorithms may produce new insights into Qur'anic themes that could not have been observed using traditional methods.
- Every opportunity to further understand Qur'anic content would be of interest to religious scholars and followers of Islam.
- The development of learning algorithms suited for the Holy Qur'an might produce results that have applications in other contexts as well, such as on other collections of sentence-length documents, or on other sets of Arabic text.

2.3 Previous Work

For the reasons discussed above, the Holy Qur'an has enjoyed extensive scholarly attention for centuries. Older works have already placed attempts at matching up verses of the Holy Qur'an by topic in order to extract deeper meaning from connected themes. In more recent years, the use of technology has greatly sped up and broadened the task of computationally analyzing the Holy Qur'an.

One interesting application of machine learning on the Holy Qur'an is the idea of building a classifier to determine the geographic location of revelation of each verse, based on its textual content. This idea has been explored by researchers such as Sharaf [21] and Bin Dost et.al. [5]. A further step was taken by Nassourou, who attempted to classify a full chronological order of the revelation [17].

In addition to the countless historical works that attempt to categorize the topics of the Holy Qur'an, more recent work in mapping the ontology of the text has been considered in a survey by Ahmed et. al. [1]. An explicit attempt at mapping the concept structure of the Holy Qur'an has been presented by Dukes in the Qur'an Corpus [9]. This latter source has been of great avail to computational study of the Holy Qur'an in a variety of projects due to its easily accessible database of Qur'anic text and morphological analysis. Some work by Sharaf [22] excellently depicts verse similarity based on related structure, and Thabet [27] discusses the idea comparing chapters of the Holy Qur'an on a thematic basis.

This work is independent of previous studies in the field, since it focuses on finding structured relationships at a verse level within the Holy Qur'an, with no annotation or guidance from human supervision.

Chapter 3

Procedures

We now present the methods of machine learning that were applied to the text of the Holy Qur'an. There were two main techniques that were implemented in this work, tackling the problems of topic modelling and data visualization, respectively. The significance of each of these problems, as well as the algorithms used to solve them, are outlined below. Both methods build upon a shared framework that is the common starting point in text analysis, known as the term-document matrix.

3.1 The Term-Document Matrix

The term-document matrix effectively encodes textual information from a corpus into a mathematical object that can be used for computations. Building one up with suitable encoding and dimensionality reduction requires a few steps, beginning with the representation of text using vectors.

3.1.1 Vector Representation of Text

The first and most important step in preparing a term-document matrix is to represent each document as a vector in a very large vector space. There is no obvious best way in how this representation should be done, but it is evident that the vector space must be massively multi-dimensional in order to capture the huge variability of documents within the given corpus.

A common approach is to use the "bag-of-words model", which has the advantage of being extremely simple and easy to implement. In this system, the entire vocabulary of the language is laid out as an ordered set of distinct words. The underlying vector space has as many dimensions as there are words in the language, with each dimension corresponding to a distinct word in the ordered vocabulary. Each entry in a document vector stores the frequency of the associated word in the document. Hence, vectors will be very long, sparse columns with only a few nonzero entries, all of which will typically be very small positive numbers.

The bag-of-words model has obvious deficiencies. Word order is completely lost, so documents like "*the man ate the chicken*" and "*the chicken ate the man*" would be indistinguishable under this vector representation scheme. This model also encounters difficulty when dealing with words that have multiple meanings. For example, given the documents "*He sat at the bank of the river*" and "*He withdrew some cash from the bank*", there is no distinction given to the differences in meaning to the word "bank" in each context.

Alternatives to the bag-of-words model do exist. Techniques using n-grams [7] allow some ability for encoding word order, at the cost of a considerable increase in vector dimensionality. Neural-network based models are able to provide very powerful vector representations at a word level [2, 16], and although the idea has been extended to document-level representation [14], an implementation is not yet known to exist. The bag-of-words model, however, works well enough, despite its shortcomings, and is a favoured approach due to its simplicity.

Once documents are represented as vectors, a term-document matrix is constructed by stacking the document column vectors into a matrix. With each row of the matrix representing a term and each column representing a document, the (i, j)-th entry in the matrix represents the frequency of term i in document j. This enormous matrix is non-negative and incredibly sparse, but fully describes the term presence in every document of the corpus.

3.1.2 Adapting to the Arabic Language

One important feature about the Holy Qur'an is that the original text is entirely written in pure, classical Arabic. Numerous translations exist in many different languages, but only the original Arabic text is regarded as the authentic and accepted word of God. The Arabic language has many differences from Western languages, some of which will need to be taken into consideration when applying machine learning techniques.

The Arabic language is an ancient Semitic language that has evolved to encompass numerous different dialects around the world today. The written script is based on an alphabet of 28 characters, and is written from right to left in an elegant, cursive-like flow that is renowned in calligraphic artwork. One particular characteristic about written Arabic is that generally, only the consonants and long vowels are written as letters. Short vowels are represented by optional diacritical marks that are often omitted in ordinary text. In formal and authentic works, however, such as the text of the Holy Qur'an, diacritic marks are almost always explicitly displayed.

Grammatically, the Arabic language is based on a root system, where small root words are used to generate a variety of related words using numerous different combinations of conjugations and affixes. Arabic words undergo inflections depending upon their position in the sentence as well as other factors, such as definitude and grammatical role. These inflections are portrayed using different diacritic markers which affect the spelling of the words.

To facilitate the use of Western-based machine learning software on the Holy Qur'an, all Arabic symbols were first converted into ASCII characters that software can work with more easily. A one-to-one mapping that can translate every Arabic letter and diacritic mark into a distinct English letter or punctuation symbol guarantees a conversion that is unambiguous and reversible. The mapping system chosen for this purpose was the Buckwalter transliteration scheme [11], introduced in the 1990s and accepted as a standard format in similar applications.

3.1.3 Vector Preprocessing and Dimensionality Reduction

Using transliteration, the Arabic text of the Holy Qur'an corpus can be converted into strings of English symbols. It is possible to therefore construct the term-document matrix at this stage. However, it is worthwhile to apply more textual preprocessing before the matrix is assembled. In particular, stemming the corpus will help produce a smaller matrix and generate more meaningful results.

A single Arabic word can carry different spellings based on its grammatical role in a sentence, sometimes differing only in diacritic placement. When building a document vector for use in the term-document matrix, each of these distinct spellings would constitute different words. These multiply-counted terms increase the dimension of the vector space, and also conceal the relationship between words differing only in diacritic placement. A more intelligent scheme would identify words that carry the same literal meaning, despite differences in spelling, and categorize them together as a single term during frequency counting and matrix building.

Stemming algorithms are designed to take this a step further and collect common roots of entire families of related words. As an example in the English language, the words “swimmer”, “swim”, “swam”, and “swimming” all relate to the same idea of swimming, but have different meanings and grammatical functions. A stemmer might classify all or some of these words under the same category, depending on how aggressive the stemming is. In Arabic, a single root word often generates a massive pool of words that may include words of radically unrelated meaning. There has correspondingly been a variety in scholarly opinion on how stemming should be implemented and to what extent words should be classified together. In particular, the case of Qur’anic Arabic has received a lot of scholarly attention; examples include root-based stemming [31], light stemming [26], and different forms of morphological extraction [19, 24].

In this work, the lemma-based approach supported by the Qur’an corpus is used [9]. Unlike the linguistic stem, the lemma of a word retains the same part of speech and essential meaning, while stripping away adjustments such as plurals, definite markers, and possessive indicators. Rather than rely on stemming techniques that may group too many or too few words together, the lemma solution provides a middle-route and ensures that the meaning of the word is preserved.

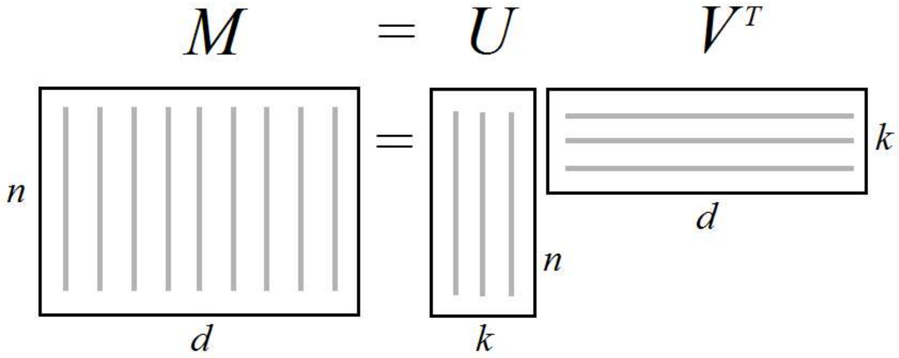
Once a term-document matrix is constructed, the corpus is ready for statistical analysis, using techniques such as topic modelling and data visualization.

3.2 Topic Modelling

Topic modelling is an unsupervised learning method based on the idea that a large group of documents may be accurately classified into a small number of topics. The learning method tries to discover what these topics are, and how they contribute to each document in the corpus, without having any external input indicating what the topics may be. The unsupervised nature of this procedure enables discovery of topics that may reveal hidden structure within the corpus, some of which may be neither visible nor anticipated through manual inspection. This type of machine learning is thus useful for identifying new insights and patterns within corpora that have already been exposed to repeated and extensive study.

Some popular topic modelling techniques involve reducing the dimensionality of the term-document matrix, such as by using the singular value decomposition. By keeping

Figure 3.1: A nonnegative matrix factorization, $M = UV^T$.



only the most significant dimensions of a matrix, a low rank factorization that encompasses the main topics of the document collection is produced. Latent semantic analysis methods generally fall within this type of procedure [8], and probabilistic variations such as latent Dirichlet allocation have also been developed [6]. The technique used in this work, however, is based on nonnegative matrix factorization, due to its ability to classify documents as nonnegative combinations of multiple topics. The technique and its details are presented below.

3.2.1 Nonnegative Matrix Factorization

Consider a collection of d text documents that use a total of n distinct terms (after stemming). The term-document matrix then takes the form of a sparse, nonnegative $n \times d$ matrix, which we may call M.

Suppose that the matrix M can be factorized into the form $M = UV^T$, where U is $n \times k$ and V is $d \times k$, for some small value of k. Since M is nonnegative, it is possible that U and V may also contain only nonnegative entries. If k is significantly smaller than n and d, then M has been factored into a product of two low-rank, nonnegative matrices. This setup is illustrated in Figure 3.1.

Recall that each of the d columns of M represents a term frequency distribution for one of the documents of the corpus. As each of the k column vectors of U is also a nonnegative vector of dimension n, they may also be interpreted as term frequency distributions. Furthermore, each of the actual documents, represented by the columns of M, may be written as a linear combination of the k columns of U. The coefficients in these combinations are given by the nonnegative entries in the d rows of V. Indeed, this is exactly what is meant by the factorization $M = UV^T$.

In other words, the essence of the entire corpus can be represented by just k separate groups of words in fixed proportions, taken in different combinations. Hence, this simple factorization provides great insight into the underlying structure of the word distributions in the documents of the corpus. If M is a term-document matrix representing how commonly each term appears in each document, then U is a term-topic matrix indicating how commonly each term appears in each topic, and V is a document-topic matrix which shows how strongly each topic appears in each document.

This is a very simple idea, but the consequences are powerful. The k topics found to summarize the corpus were generated completely automatically, without any human influence. Thus the topics are not given by titles like "Politics" or "Mathematics", but are instead defined by the lists of words that strongly represent them, as given by the term-topic matrix U. (Alternatively, topics are defined by the lists of documents that include them in high proportion, as given by the document-topic-matrix V.) It is up to human viewers to examine the resulting word (or document) lists and determine what labels would make suitable topic titles.

The results may not always correspond to what a human might consider a reasonable assortment of well-defined topics, however; the factorization only produces lists of words that tend to appear with each other within the documents of the corpus, or lists of documents that tend to use many words in common. These lists may correspond to what a human would perceive as reasonable topics, but it is not always the case. It is not rare for words to come up in these lists that seem topically unrelated, but actually appear together often when they are used in the corpus. In this way, topic modelling enables previously ignored patterns in word usage and document similarity to be identified.

3.2.2 Achieving such a Factorization

It is surprising enough that a corpus of many documents can be simplified to compositions of only a few topics. The possibility that these hidden topics can be discovered, as well as their impact on each document in the corpus, using only the frequencies of words within each document, is astounding. But topic modelling as described here depends upon the existence of a nonnegative matrix factorization. The number of anticipated topics, k, must also be supplied or estimated.

Unfortunately, computing a nonnegative matrix factorization of a nonnegative matrix like M, or even a close approximation to such a factorization, is in general an NP-hard problem. It is not certain that such a factorization will exist. If the factorization does exist, it may require an exponential amount of calculation to produce. The best approach is to aim for approximations that may be good enough; that is, to determine nonnegative matrices U and V such that some objective function like

$$\|M - UV^T\|$$

is minimized. This would correspond to a factorization that multiplies to a product as "close" as possible to the actual term-document matrix, hence implying that the factorization is a reasonable approximation.

In practice, there are many different algorithms available for producing approximate nonnegative-matrix factorizations [15, 3]. The method chosen in this work is the Rank-1 Downdate (R1D) algorithm, which was chosen for its speed, accuracy, and simplicity. This algorithm also has demonstrated its effectiveness in identifying correct hidden topics when applied to the topic modelling problem. Details on the implementation of this algorithm are omitted from this work, but may be found in the original paper that introduced the technique [4].

Comments regarding the application of this topic modelling procedure on the Holy Qur'an will be discussed in the forthcoming chapter on Results. Before that, a second method of unsupervised learning applicable to text documents will be introduced.

3.3 Data Visualization

Patterns become much easier to understand and comprehend when they are presented in a visual display. Text documents, however, are difficult to treat as visual objects. Thus, one task pertaining to machine learning is to present a text corpus in a graphic manner that demonstrates meaningful relationships between documents in a clear and accurate way.

3.3.1 Dimensionality Reduction and Manifold Embedding

It has already been shown above that the documents in a corpus may be represented using a term-document matrix, where every document is encoded as a vector that describes the frequencies of words within that document. Although vectors may be visualized as points in a vector space, it is impossible to physically represent the vector space fully if it has more than three dimensions. As the vectors in the document space typically have thousands of dimensions (corresponding to all of the possible terms appearing in the entire corpus), a direct visual representation is certainly out of the question.

Thus, dimensionality reduction is necessary when visualizing high dimensional data. In a familiar setting, a three-dimensional object can be projected onto a two-dimensional surface in the form of a photograph, although losing the sense of physical depth. This concept can be generalized as higher dimensional entities are projected onto spaces of fewer dimensions, necessarily sacrificing some information in the process. The key is to find a way to retain as much significant information as possible, while faithfully representing the original object.

A lower-dimensionality representation, or manifold embedding, preserves structure if points near each other in the original space are mapped to points near each other in the resulting space, and if points that are distant to begin with remain distant after the projection as well. In terms of our intended application, this means that verses of the Qu'ran with similar word usage, and hence similar document vectors, will be the ones that are plotted near each other in the two-dimensional visualization. Another desirable characteristic is that the mapping should affect all points well, so that a single visualization effectively depicts the entire corpus.

Some techniques, like traditional Principal Component Analysis, are linear operations that effectively reduce the dimensionality of the dataset, but focus on keeping dissimilar points apart and cannot always keep similar points close together. Some common nonlinear mappings, such as Local Linear Embedding [18], Isomap [25], and Maximum Variance Unfolding [30], may perform better in this regard, but are unable to satisfy the requirement for global applicability.

In 2008, a new dimensionality reduction technique designed for data visualization was introduced, with a focus on keeping similar points near each other in the projected map. Called the t-SNE algorithm, this method has shown strong performance in generating two-dimensional representations of large vector spaces in a way that very effectively preserves relationships between points in the original dataset [29]. This is the method that was used to visually organize the verses of the Holy Qur'an in a two dimensional map, and observations on this representation will be presented in the following chapter.

Chapter 4

Results

A term-document matrix was prepared using the Buckwalter transliteration of the Holy Qur'an, filtered to include only the lemma-form of each Arabic word instead of their original spellings. As discussed in Section 3.1.3, this form of reducing words to their grammatical roots is effective both for dimensionality reduction and to add emphasis on repeated content. Further, only lemmas for parts of speech believed to be of topical significance, such as nouns, proper nouns, adjectives, and verbs, were retained. The transliteration and morphological annotation were obtained through the freely available Qur'an Corpus [9], and machine learning software Weka [12] was used to generate word frequencies for the term-document matrix.

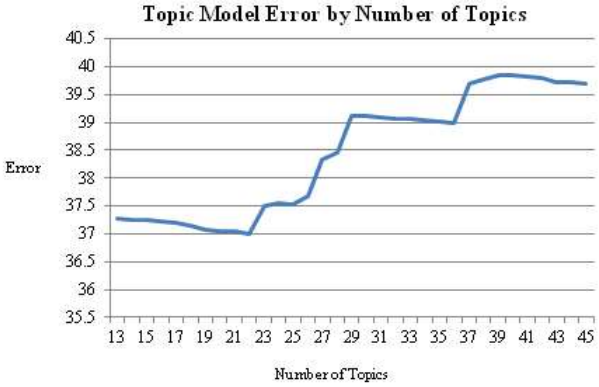
Verses of the Holy Qur'an vary tremendously in length. A normalization step on the term-document matrices was necessary to remove the bias towards longer verses. Each document vector was scaled down by the length of the verse, so that the sum of the entries in each column of the term-document matrix would be 1.

The resulting term-document matrix M is a sparse, nonnegative matrix with 4755 rows corresponding to distinct key lemmas, and 6236 columns corresponding to each verse of the Holy Qur'an. The values in the matrix range from 0 to 10 with a total of 46430 nonzero entries, making up about 0.16% of the matrix.

4.1 Topic Modelling Results

Nonnegative factorizations of the term-document matrix M were taken using values of k ranging from 13 to 45. Factorizations were obtained using the R1D method as described

Figure 4.1: Factorization error ranging by number of topics in topic model on the four base corpora.



in Section 3.2.2. For each factorization $M = UV^T$, the error $\|M - UV^T\|$ was measured using the Frobenius norm. The resulting error values are plotted in Figure 4.1.

The error takes a minimum value at $k = 22$, so analysis proceeded with the assumption that there are 22 prevalent topics in the Holy Qur'an corpus. A word list was obtained for each of the 22 topics by identifying the rows carrying the largest positive entries in each of the 22 columns of U. Similarly, verses with strong tendencies toward each of these topics were found by identifying the rows with largest entries out of the 22 columns of V.

Table 4.1 shows a summary of five of these 22 topics, highlighting ten of the most impactful words and eight of the most prominent verses for each topic. Labels were manually assigned to each category to reflect their contents. It should be noted that although the text in the tables are displayed in English to aid understanding, all of the computation was done only on Arabic words, preserving the original vocabulary of the Holy Qur'an for accuracy.

As hoped, a variety of interesting insights are revealed through this unsupervised classification of Qur'anic themes. Although topics like *Paradise* and *Worship* might have been detected without the aid of machine learning, it is reassuring that the results confirm expectations. Furthermore, the lists shed light on relations that may not have been as clear at first glance. For example, Topic 15 largely concerns the believing servants of God, but among the terms that characterize it are the seemingly unrelated words "hand" and "heart". Another remarkable observation is that topics 18 and 19, roughly describing God's message and worship, respectively, are characterized by extremely similar groups of words, despite the difference in content and in corresponding verses. Reasoning behind

Table 4.1: Abridged summaries of five of the topics identified in the Holy Qur’an.

Topic 3	**“The Lord”**
Key words:	*Lord, he was, he denied, he said, bounty, he believed, universe, people, merciful, which*
Key verses:	53:42. And that to your Lord is the goal- 23:59. And those who do not associate (aught) with their Lord, 43:14. And surely to our Lord we must return. 53:49. And that He is the Lord of the Sirius; 44:20. And surely I take refuge with my Lord and your Lord that you should stone me to death: 37:180. Glory be to your Lord, the Lord of Honor, above what they describe. 37:5. The Lord of the heavens and the earth and what is between them, and Lord of the easts. 26:26. He said: Your Lord and the Lord of your fathers of old.
Topic 12	**“Paradise**”
Key words:	*garden, he made, fountain, he believed, he performed, earth, every, he created, bliss, the pious ones*
Key verses:	26:134. And gardens and fountains; 26:147. In gardens and fountains, 44:52. In gardens and springs; 37:43. In gardens of pleasure, 53:15. Near which is the garden, the place to be resorted to. 15:45. Surely those who guard (against evil) shall be in the midst of gardens and fountains: 51:15. Surely those who guard (against evil) shall be in gardens and fountains. 44:25. How many of the gardens and fountains have they left!
Topic 15	**“The Believers”**
Key words:	*believer, people, God, he helped, hand, heart, he killed, punishment, he took, he healed*
Key verses:	26:114. And I am not going to drive away the believers; 23:1. Successful indeed are the believers, 37:81. Surely he was of Our believing servants. 37:122. Surely they were both of Our believing servants. 37:132. Surely he was one of Our believing servants. 15:77. Most surely there is a sign in this for the believers. 27:77. And most surely it is a guidance and a mercy for the believers. 26:102. But if we could but once return, we would be of the believers.
Topic 18	**“God’s Message”**
Key words:	*utterance, he reminded, reminder, beneficient, thing, he oppressed, he provided benefit, he took, instead of, my son*
Key verses:	37:155. Will you not then mind? 28:51. And certainly We have made the word to reach them so that they may be mindful. 51:8. Most surely you are at variance with each other in what you say, 20:28. (That) they may understand my word; 20:44. Then speak to him a gentle word haply he may mind or fear. 50:18. He utters not a word but there is by him a watcher at hand. 37:31. So the sentence of our Lord has come to pass against us: (now) we shall surely taste; 50:29. My word shall not be changed, nor am I in the least unjust to the servants.
Topic 19	**“Worship”**
Key words:	*he worshipped, instead, thing, beneficient, my son, he oppressed, he provided benefit, he brought harm, ownership, he took*
Key verses:	37:161. So surely you and what you worship, 1:5. Thee do we serve and Thee do we beseech for help. 36:61. And that you should serve Me; this is the right way. 36:22. And what reason have I that I should not serve Him Who brought me into existence? And to Him you shall be brought back; 37:95. Said he: What! do you worship what you hew out? 26:92. And it shall be said to them: Where are those that you used to worship; 51:56. And I have not created the jinn and the men except that they should serve Me. 15:99. And serve your Lord until there comes to you that which is certain.

phenomena like these will not be speculated upon in this work.

4.2 Data Visualization Results

A direct application of the t-SNE algorithm on the Holy Qur'an corpus would require an unreasonable amount of computing time and memory. Instead, the term-document matrix M was first subject to a dimensionality reduction using Principal Component Analysis, reducing its rank to 50. The t-SNE algorithm was then applied for 1000 iterations using a perplexity parameter of 100. More information on the technical details of t-SNE, which are omitted in this work, may be found in the original paper by the algorithm's inventors [29]. The working implementation of the t-SNE algorithm used for this research was also provided by these authors [28].

By applying a new preprocessing step on the matrix M, the clustering ability of the t-SNE map could be improved. A Term-Frequency, Inverse-Document-Frequency (TF-IDF) transformation on the term-document matrix replaces each entry $f_{i,j}$ of M with

$$f_{i,j} \log_2 \left(\frac{n_{docs}}{n_{docs}(i)} \right),$$

where n_{docs} represents the total number of documents in the corpus (in this case, the number of verses of the Holy Qur'an is 6236) and $n_{docs}(i)$ represents the number of documents in the corpus that contain term i. This transformation has the effect of reducing the relative frequency of terms that appear in many of the verses of the Qur'an, as these terms are probably generic and not subject-related.

The resulting two-dimensional maps is shown in Figure 4.2. In these graph, each point represents one of the verses of the Holy Qur'an, and placement of these points is an attempt to indicate structure behind verse content. Verses that are similar would be projected near each other onto this map. The resulting clusters of points therefore suggest groups of closely related verses based on word usage, possibly influenced by topic. The first plot shows the clustering based on the original term-frequency matrix, and the second demonstrates the effect of the TF-IDF adjustment. As hoped, the clusters in the second map are more well-defined.

It should be noted that in the t-SNE generated map, there is no meaning associated to the coordinate axes. The x and y coordinates of the points are irrelevant, except in how they group points together in clusters. The attraction of points towards the centre of the

Figure 4.2: Plots showing point representations of verses of the Holy Qur'an, as generated by the t-SNE algorithm. The first plot is based on the original term-frequency matrix, and the second takes into account TF-IDF adjustments. Clusters of points represent groups of similar verses.

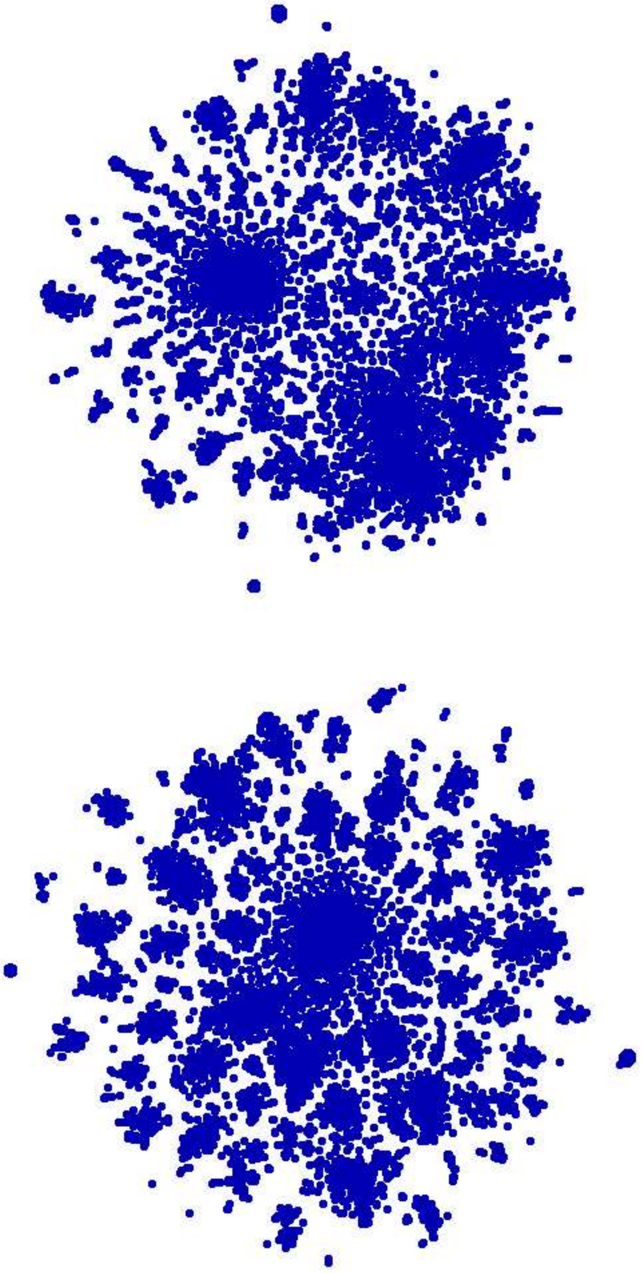


Figure 4.3: The clusters of the t-SNE plot with verses examined in detail in Table 4.2.

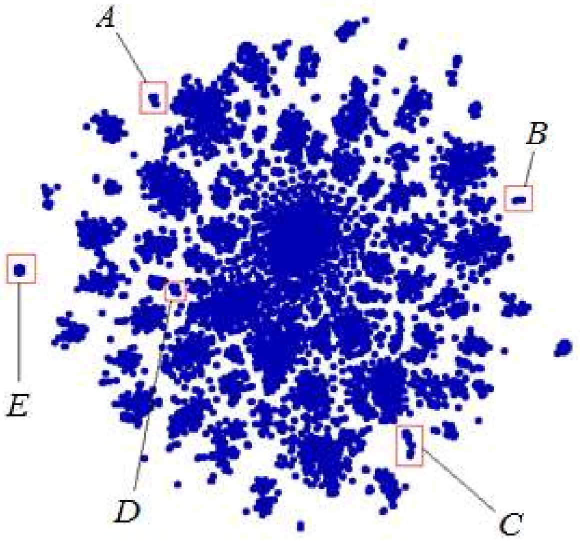


chart is a natural result of the t-SNE algorithm; verses that are not as clearly classifiable into distinct clusters tend to aggregate in the middle of the plot.

It is difficult to measure the effectiveness of the t-SNE map at retaining the structure behind the verses of the Holy Qur'an. One method is to examine the verses associated with points found in clusters, and to manually determine if the verses are related in theme or word usage. Some lists of verses corresponding to clustered points in the TF-IDF-adjusted map of Figure 4.2 are found in Table 4.2. Each cluster was manually assigned a label based on the content of the verses found within it. As the verse contents seem to surround common themes, it seems that the clustering method by the t-SNE algorithm is effective.

4.3 Topic Modelling and Data Visualization

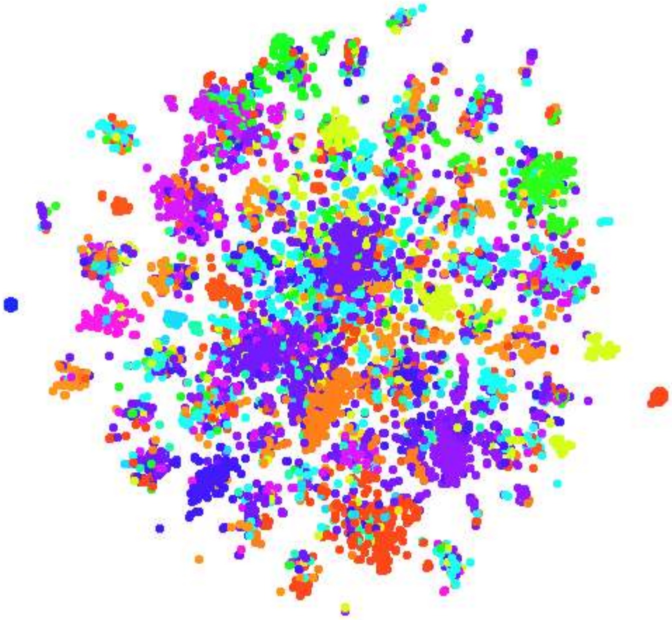
Thus far, topic modelling and data visualization have taken two different approaches on identifying and representing hidden linguistic structure within the verses of the Holy Qur'an. However, at this stage, they may be combined to present a truly remarkable graphic representation of the corpus.

Each verse of the Holy Qur'an was classified into a linear combination of 22 topic vectors in Section 4.1. For each verse, 22 nonnegative coefficients were assigned to determine how

Table 4.2: Verses in five of the clusters of the t-SNE map, highlighted in Figure 4.2.

Cluster A Verses:	***“The Glorified Lord of the Heavens and Earth”*** **30:27.** And He it is Who originates the creation, then reproduces it, and it is easy to Him; and His are the most exalted attributes in the heavens and the earth, and He is the Mighty, the Wise. **38:66.** The Lord of the heavens and the earth and what is between them, the Mighty, the most Forgiving. **45:37.** And to Him belongs greatness in the heavens and the earth, and He is the Mighty, the Wise. **48:7.** And God’s are the hosts of the heavens and the earth; and God is Mighty, Wise. **57:1.** Whatever is in the heavens and the earth declares the glory of God, and He is the Mighty, the Wise. **59:1.** Whatever is in the heavens and whatever is in the earth declares the glory of God, and He is the Mighty, the Wise. **59:24.** He is God the Creator, the Maker, the Fashioner; His are the most excellent names; whatever is in the heavens and the earth declares His glory; and He is the Mighty, the Wise. **61:1.** Whatever is in the heavens and whatever is in the earth declares the glory of God; and He is the Mighty, the Wise. **62:1.** Whatever is in the heavens and whatever is in the earth declares the glory of God, the King, the Holy, the Mighty, the Wise.
Cluster B Verses:	***“The Later Generations”*** **37:78.** And We perpetuated to him (praise) among the later generations. **37:108.** And We perpetuated (praise) to him among the later generations. **37:119.** And We perpetuated (praise) to them among the later generations. **37:129.** And We perpetuated to him (praise) among the later generations. **56:14.** And a few from among the latter. **56:40.** And a numerous company from among the last. **74:53.** Nay! but they do not fear the hereafter. **75:21.** And neglect the hereafter. **77:17.** Then did We follow them up with later ones.
Cluster C Verses:	***“Resurrection after Death; Greatness”*** **4:73.** And if grace from God come to you, he would certainly cry out, as if there had not been any friendship between you and him: Would that I had been with them, then I should have attained a mighty good fortune. **17:49.** And they say: What! when we shall have become bones and decayed particles, shall we then certainly be raised up, being a new creation? **23:35.** What! does he threaten you that when you are dead and become dust and bones that you shall then be brought forth? **23:82.** They say: What! When we are dead and become dust and bones, shall we then be raised? **24:16.** And why did you not, when you heard it, say: It does not beseem us that we should talk of it; glory be to Thee! this is a great calumny? **33:42.** And glorify Him morning and evening. **37:16.** What! when we are dead and have become dust and bones, shall we then certainly be raised, **37:53.** What! when we are dead and have become dust and bones, shall we then be certainly brought to judgment? **37:60.** Most surely this is the mighty achievement. **38:67.** Say: It is a message of importance, **44:57.** A grace from your Lord; this is the great achievement. **56:46.** And they persisted in the great violation. **56:47.** And they used to say: What! when we die and have become dust and bones, shall we then indeed be raised? **56:74.** Therefore glorify the name of your Lord, the Great. **56:96.** Therefore glorify the name of your Lord, the Great. **68:4.** And most surely you conform (yourself) to sublime morality. **69:33.** Surely he did not believe in God, the Great, **69:52.** Therefore-glorify the name of your Lord, the Great. **78:2.** About the great event, **79:11.** What! when we are rotten bones?
Cluster D Verses:	***“Calling Upon Others”*** **4:117.** They do not call besides Him on anything but idols, and they do not call on anything but a rebellious Shaitan. **6:41.** Nay, Him you call upon, so He clears away that for which you pray if He pleases and you forget what you set up (with Him). **7:193.** And if you invite them to guidance, they will not follow you; it is the same to you whether you invite them or you are silent. **7:197.** And those whom you call upon besides Him are not able to help you, nor can they help themselves. **17:110.** Say: Call upon God or call upon, the Beneficent God; whichever you call upon, He has the best names; and do not utter your prayer with a very raised voice nor be silent with regard to it, and seek a way between these. **71:8.** Then surely I called to them aloud: **84:11.** He shall call for perdition, **96:17.** Then let him summon his council, **96:18.** We too would summon the braves of the army.
Cluster E Verses:	***“Initial Letters”*** **2:1.** Alif Lam Mim. **3:1.** Alif Lam Mim. **7:1.** Alif Lam Mim Suad. **19:1.** Kaaf Ha Ya Ayn Saad. **20:1.** Ta Ha. **26:1.** Ta Sin Mim. **28:1.** Ta sin Mim. **29:1.** Alif Lam Mim. **30:1.** Alif Lam Mim. **31:1.** Alif Lam Mim. **32:1.** Alif Lam Mim. **36:1.** Ya Seen. **40:1.** Ha Mim. **41:1.** Ha Mim! **42:1.** Ha Mim. **42:2.** Ain Sin Qaf. **43:1.** Ha Mim. **44:1.** Ha Mim! **45:1.** Ha Mim. **46:1.** Ha Mim.

Figure 4.4: Topic modelling on display: a t-SNE plot of the Holy Qur'an with verses coloured by topic.



closely the verse aligned with each of the topics. Verses may therefore be classified into one of the 22 topics by looking at which of the coefficients has the largest value for each verse, hence shaping the content of the verse most prominently.

In this way, every verse may be assigned one of the 22 discovered topics of the Holy Qur'an. Figure 4.3 shows the TF-IDF-adjusted t-SNE plot of Figure 4.2 with each point now coloured according to its topic, so that verses of the same topic are painted using the same colour. The fact that so many of the clusters prominently feature a single colour elegantly demonstrates that topic assignment discovered by R1D matrix factorization aligns closely with the natural clusters formed by t-SNE dimensionality reduction.

Chapter 5

Conclusion

The underlying topics of the Holy Qur'an were effectively extracted and identified, and verses were meaningfully clustered on a rich visual representation of the corpus. Based on the generally successful results of this research, the Holy Qur'an was proven to be a suitable subject for unsupervised machine learning techniques on text.

There are, of course, next steps that may lead to improved performance. The bag-of-words model, used to generate the fundamental term-document matrix, is known to be inherently deficient. A more effective vector representation scheme might substantially improve all of the present results. Another shortcoming comes from the treatment of every verse independent of all others. This assumption prevents pronominal resolution, which would assign a proper entity to each pronoun mentioned in the corpus. The Holy Qur'an is replete with pronouns that, if replaced with their true reference when counting word frequencies, might demonstrate verse relationships that are impossible to recognize in the current model. Despite these failings, the overall outcomes in this research were promising, and indicate that the future of computational analysis on the Holy Qur'an is bright.

The original intention of this work was to uncover deeper meaning behind the content of the Holy Qur'an using statistical and computational analysis of the words themselves. The abridged results presented in this essay represent just a small fraction of the generated models during the process of research. Perhaps deeper investigation into the full collection of results will reveal more structure and patterns behind the text, bringing humanity another step forward in uncovering the hidden mysteries within the word of God.

References

[1] Omar Ahmad, Irfan Hyder, Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, Nurfadhlina Mohd Sharef, and Muhammad Mansoor. A survey of searching and information extraction on a classical text using ontology-based semantics modeling: A Case of Quran. *Life Science Journal*, 10(4), 2013.

[2] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

[3] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.

[4] Michael Biggs, Ali Ghodsi, and Stephen Vavasis. Nonnegative matrix factorization via rank-one downdate. In *Proceedings of the 25th international conference on Machine learning*, pages 64–71. ACM, 2008.

[5] Muhammad Khyzer Bin Dost and Munir Ahmad. Statistical profile of Holy Quran and symmetry of Makki and Madni surras. *Pakistan Journal of Commerce and Social Sciences*, January 2008.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[7] Daniel de Kok and Harm Brouwer. *Natural Language Processing for the Working Programmer*, chapter 3. N-grams. 2010, 2011.

[8] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[9] Kais Dukes. The Qur'anic Arabic Corpus, 2011.

[10] Nizar Grira, Michel Crucianu, and Nozha Boujemaa. Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content, Report of the MUSCLE European Network of Excellence (FP6)*, 2004.

[11] Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. On arabic transliteration. In *Arabic computational morphology*, pages 15–22. Springer, 2007.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian Witten. *The WEKA Data MiningSoftware: An Update.* SIGKDD Explorations, 11, 1 edition, 2009.

[13] Sotiris B Kotsiantis, ID Zaharakis, and PE Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[14] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[15] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[17] Mohamadou Nassourou. A knowledge-based hybrid statistical classifier for reconstructing the chronology of the quran. *WEBIST/WTM*, 2011.

[18] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[19] Motaz K Saad and Wesam Ashour. Arabic morphological tools for text mining. *Corpora*, 18:19, 2010.

[20] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[21] Abdul-Baquee M. Sharaf. *The Qur'an Annotation for Text Mining.* PhD thesis, University of Leeds, School of Computing, December 2009.

[22] Abdul-Baquee M Sharaf and Eric Atwell. Qursim: A corpus for evaluation of relatedness in short texts. In *LREC*, pages 2295–2302, 2012.

[23] Allamah Muhammad Husayn Tabatabai. *The Qur'an in Islam.* Zahra Publications.

[24] Rafi Talmon and Shuly Wintner. Morphological tagging of the quran. In *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, an EACL03 Workshop*, pages 67–74. Citeseer, 2003.

[25] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[26] Naglaa Thabet. Stemming the Qur'an. In *Semitic '04 Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, pages 85–88, 2004.

[27] Naglaa Thabet. Understanding the thematic structure of the qur'an: an exploratory multivariate approach. In *Proceedings of the ACL Student Research Workshop*, pages 7–12. Association for Computational Linguistics, 2005.

[28] Laurens van der Maaten and Geoffrey Hinton. Users guide for t-sne software.

[29] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[30] Kilian Q Weinberger and Lawrence K Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI*, volume 6, pages 1683–1686, 2006.

[31] Raja Yusof, Roziati Zainuddin, M Baba, and Z Yusof. Qur'anic words stemming. *Arabian Journal for Science and Engineering (AJSE)*, 35(2):37–49, 2010.