Extracting Topics from the Holy Quran Using Generative Models

Mohammad Alhawarat
Department of Computer Science,
Prince Sattam Bin Abdulaziz University,
Alkharj, Saudi Arabi

***Abstract*—The holy Quran is one of the Holy Books of God. It is considered one of the main references for an estimated 1.6 billion of Muslims around the world. The Holy Quran language is Arabic. Specialized as well as non-specialized people in religion need to search and lookup certain information from the Holy Quran. Most research projects concentrate on the translation of the holy Quran in different languages. Nevertheless, few research projects pay attention to original text of the holy Quran in Arabic language. Keyword search is one of the Information Retrieval (IR) methods but will retrieve what is called exact search. Semantic search aims at finding deeper meanings of a text, and it is a hot field of study in Natural Language Processing (NLP). In this paper topic modeling techniques are explored to setup a framework for semantic search in the holy Quran. As the Holy Quran is the word of God, its meanings are unlimited. In this paper the words of chapter Joseph (Peace Be Upon Him (PBUH)) from the Holy Quran is analyzed based on topic modeling techniques as a case study. Latent Dirichlet Allocation (LDA) topic modeling technique has been applied in this paper into two structures (Hizb Quarters and verses) of Joseph chapter as: words, roots and stems. The log-Likelihood has been calculated for the two structures of the chapter. Results show that the best structure to use is verses, which gives the least energy for data. Some of the results of the attained topics are shown. These results suggest that topic modeling techniques failed to capture in an accurate manner the coherent topics of the chapter.**

Keywords*—*Statistical models; Latent Dirichlet Analysis (LDA); Holy Quran; Unsupervised Learning

I. Introduction

The holy Quran is considered an essential reference for Muslims where they read in a regular basis. They usually need to search it and retrieve relevant information based on more than just simple keyword search techniques.

Dealing with the holy Quran is different from dealing with regular Arabic corpora that is usually extracted from Newspapers and speeches, and hence is the word of human. The holy Quran is the word of God and the meanings of its words are unlimited. The sequence of text is different from human words. For example, one topic could repeat in different places in the holy Quran with different details and sometimes in different contexts. Also, one chapter usually has many topics. While one topic might be started in one verse, another topic may starts immediately in the next verse. Also, one verse may have different topics. Moreover, there are different authentic interpretations for the verses of the holy Quran; therefore it is very hard for a computer to manage them in the way scholars do especially in situations where meanings are seem opposite to each other. Finally, there is much relevant information that is found in prophet Mohammad (PBUH) sayings (Hadith) that interpret many verses of the holy Quran. For all of these reasons, it sometimes hard to resolve a disambiguation if a word has many synonyms and different senses.

Research in Arabic NLP still young and have many challenges [1]. This is because that Arabic language is different from many other natural languages [2], [3]. Words in Arabic language have many derivations and have also complex Diglossia (modern and colloquial) [4]. Also, Arabic letters appear in different shapes according to their position in the word. Another characteristic of the Arabic language is the diacritic. Some of these diacritical marks are usually not written, but is understood by Arabic readers. Therefore, two exact written words without diacritical marks have totally different meanings. All of these and other characteristics of the Arabic language should be taken in consideration when processing Arabic text.

The holy Quran can be considered as a ”Golden Text” to use in Text mining and NLP fields. This might be true for different reasons: it’s the word of God, it’s limited in terms of text size and it has many translations and many interpretations. These all together encourage building a semantic comprehensive source for the holy Quran that will allow advanced semantic search and knowledge extraction.

Searching in the holy Quran is an essential task for Muslims as well as non-Muslims who study it. Many applications have been built to allow search in the holy Quran. Most of these search engines allow simple search techniques where some of them are mentioned in [5]. However, few research projects are concerned with advanced search in the holy Quran using some NLP techniques such as the papers presented in The holy Quran and new technology workshop that held by King Fahad Complex for printing the holy Quran in Al-Madinah Al-Munawwarah, Saudi Arabia in 2008. The workshop participants discussed different issues related to the holy Quran including searching techniques. Also more papers are presented in another event in Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences that held in Al-Madinah Al-

Munawwarah, Saudi Arabia in 2013. The presented papers are related to a wide range of topics concerning the holy Quran including natural processing issues, security, education and many more.

There are different approaches to model and cluster topics in text documents such as LDA, Latent Semantic Analysis (LSA) and traditional clustering techniques such as K-means. In this research LDA is used for several reasons including accuracy, scalability and comprehension [6], [7], [8].

LDA has been developed to extract topics from text using statistical methods [9], [10], [11], [12], [13]. LDA is one of the techniques that belongs to a large family called probabilistic modeling. The basic intuition behind LDA is that a text document has multiple topics where each topic is defined as a distribution over a set of words. There are many flavors of the LDA model; a thorough review of the LDA topic modeling techniques can be found in [14]. Topic modeling has been applied to many field of study such as Information Retrieval IR, geographical IR, computational linguistics and NLP [15], [16], [17], [18], [19], [20].

This paper aims to build up the first stage in a framework that will allow possible semantic search in the holy Quran. This is done by applying LDA topic modeling to chapter Joseph of the holy Quran as a case study. This chapter has been chosen because it includes relative topics regarding story of the prophet Joseph (PBUH). The LDA topic modeling has been applied to words, roots and stems of that chapter. Next stages might include: studying the topics of the whole holy Quran, linking the text of the holy Quran to both authenticate interpretation of the holy Quran and the related Sayings of the prophet Mohammad (PBUH). These might be achieved using machine learning, text mining as well as NLP techniques.

It should be stated explicitly here that this research is not a religious study; rather it is a statistical study that might result in information that would guide specialized religious people to understand more about the word of God.

The paper is organized as follows: in section II related work is presented, in section III topic modeling is introduced, in section IV the methodology as well as preparation of the Data Set is explained, in section V experimental setups are explained, section VI includes discussion of the results attained in the paper and finally section VII contains conclusion.

II. Related Work

Shoaib et. al. [5] have proposed a simple WordNet for the English translation of the second chapter of the holy Quran (Al-Baqrah). They have created topic-synonym relations between the words in that chapter with different priorities. They have defined different relations that are used in traditional WordNet such as: synonymy, polysemy, hyperonymy, hyponymy, holonymy and meronymy. Then they developed a semantic search algorithm that will fetch all verses that contains the query word and its synonyms with high priority. It is not clear how the authors build their simple WordNet. In similar studies, usually authentic religion references should be used such as interpretation of the holy Quran or meanings of the words of the holy Quran. However, the results show that the developed semantic search outperform simple search algorithms.

Similar work has been carried out to extract verses from the holy Quran using an expert system that use Web Ontology Language (OWL) [21]. Again the work use English translation of the holy Quran and not Arabic language.

Another work explored the structure of a simple domain Quran ontology for birds and animals that are mentioned in the holy Quran [22]. The authors propose a framework for semantic search in the holy Quran using their domain ontology and they have evaluated it using SPARQL query language. This work uses English translation of the holy Quran.

Data mining techniques such as SVM and nave Bayesian classifiers are used cluster chapters of the holy Quran based on Major Phases of Prophet Mohammads (PBUH) Messengership [23]. This work classifies chapters of the holy Quran rather than verses or words of the holy Quran.

LDA topic modeling technique has been used to extract topics from an Arabic corpora composed of Newspapers [24]. The authors have developed a preprocessing lemma-based stemming algorithm and then applied the LDA technique on Arabic processed text.

In [25] author has used clustering techniques in machine learning to extract topics of the holy Quran. The extraction of topics was based on a corpus that is composed of the verses of the holy Quran using nonnegative matrix factorization. The author used Buckwalter code for Arabic letters [3]. Topics are visualized and related verses for each topic are shown for selected topics based on the topic main keywords. One of the shortcoming of his work is that verses are dealt with separately as each as a document. The author claims that he has extracted and identified the underlying topics of the holy Quran. However, this claim is far from reality as no one could identify the underlying topics of the holy Quran even well-known scholars of Quran studies. Also, the it is totally unclear how he has linked the keywords of each topic with the related verses that correspond to topic keywords. Nevertheless, the findings are promising and might help in revealing deeper meanings of the holy Quran by specialized people in Quranic studies.

LDA technique has been compared LDA with K-means clustering technique [8]. The authors have applied both LDA and K-means technique on a set of Arabic documents from OSAC (Open Source Arabic Corpora). The results show that LDA outperforms K-means in most instances.

III. Topic Modeling

Topic modeling is a hot field of study in both machine learning and NLP. Topic models are generative models that are based on probability distributions of multiple topics in a document over a set of words. Such models basically depend on term-frequencies in a document. One of these models is LDA. As mentioned previously, LDA is better than other models such as LSA for several reasons[6], [7], [8]. LDA outperforms LSA in many applications including semantic representation [12] and have been used in different fields in the last decade or so including NLP [15], [16], [17]. It is used by researchers to extract important and hot topics; usually from large corpora.

The basic intuition behind LDA is that a set of words of documents are randomly pre-assigned with probability distributions that would represent multiple-topic latent structure on those documents. After that, latent structure of the topics of documents is inferred statistically in a reverse-engineering manner.

Initially, a number of topics *T* should be specified. Then, a term distribution ϕ over a parameter β is chosen for each topic. After that, ratios θ of topic distribution for document *d* are specified. Then, a topic z_i is chosen and after that a word is chosen conditioned on that topic over a parameter α. Both ϕ and θ are Dirichlet distributions.

The probability of the *i*th word in a specific document is given by:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \tag{1}$$

where z_i represents a latent variable that designates the topic for the drawn *i*th word. $P(w_i|z_i = j)$ represents the probability of the word w_i under topic *j*. $P(z_i = j)$ represents the probability of a word from topic *j* of a document.

Note that $P(w|z)$ can be represented by a multinomial distributions ϕ over a term distribution such that $P(w|z = j) = \phi_w^{(j)}$ and $P(z)$ can be represented by a multinomial distributions θ over a topic distribution over *D* documents such that $P(z = j) = \theta_j^{(d)}$.

Then an estimation method is used to infer the latent structure of the topics of documents. Different estimation methods can be used in this context including: Variational Expectation-Maximization (VEM) method and Gibbs sampling. For more information about details of these methods please refer to [10], [11], [13], [26].

Besides LDA, Correlated Topic Model (CTM) can be used to extract correlated topics from documents. CTM is an extension of LDA. LDA usually uses Gibbs sampling for model estimation.

IV. Data Set Preparation and Methodology

The text of chapter Joseph in the format of `CP1256` has been taken from [27] in the shape of two structures: Hizb quarters and verses, all without diacritic. The frequency details of these selected structures are shown in table I. For more information about the text structure of the holy Quran please refer to [27].

TABLE I: The number of documents for the Joseph chapter based on different structures and for words, roots and stems after applying tf-idf measure on DTMs

	No. of Hizb quarters/terms	No. of Verses/terms
Original No. (TF)	6/721	111/721
With TF-IDF (words)	6/299	89/323
With TF-IDF (roots)	6/327	108/163
With TF-IDF (stems)	6/398	103/193

These two structures will be used in the topic modeling process in three shapes: words, roots and stems. Because the text of the holy Quran is the word of God, there is no margin for errors in the process of extracting both roots and stems. Therefore, the roots and stems have been constructed manually; based on two web sites [28], [29] and verified by the authors according to their experience in Arabic language and as native speakers.

These data sets will be used as the input for the implantation of the LDA to reveal the main topics for the text of the chapter of Joseph (PBUH). Different experimental setups are prepared to compute the topic models for the text of that chapter based on the aforementioned structures.

V. Experiments

Both packages `tm` and `topicmodels` of *R* are used in experiments (a practical guide for `topicmodels` can be found in [30]). First, the `tm` package will be used for text preparation and processing as building the corpus, removing stop words and building the Document Term matrix (DTM). Second, the `topicmodels` package will be used to build and fit LDA model for all structures of the text with the three shapes of word.

The text with two structures has been processed where the stop words are removed. Then, three DTMs have been built for text as: words, roots and stems. The content of the DTM is basically calculated using Term Frequencies (TF) measure. After that, the tf-idf measure has been applied on each DTM to remove frequent terms that appears on most documents, and hence are not recognized as important terms. This has been done by calculating the median and choosing high-frequent terms with frequency more than the calculated median.

TABLE II: The number of topics along with the log-Likelihood for the fitted topic models for the Joseph chapter estimated by Gibbs sampling with 10-fold cross-validation

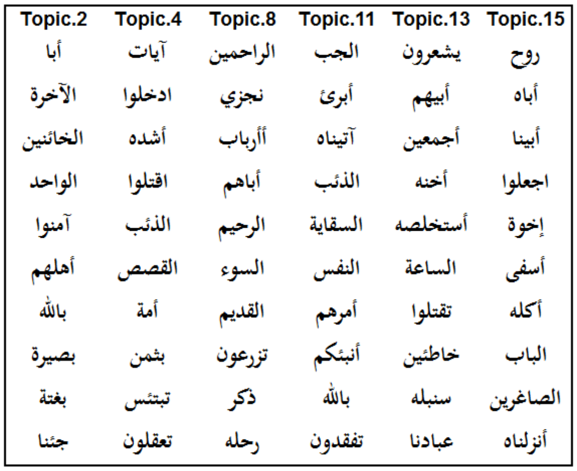
	Hizb quarters Topics No./Log-Likelihood	Verses Topics No./Log-Likelihood
Word	17/-1258	8/-250
Root	44/-827	5/-169
Stem	27/-860	19/-172

After that, different experimental setups are prepared to find the main topics in the chapter of Joseph (PBUH). These are found first using TF measure and then using different estimation techniques for LDA besides Correlated Topics Model (CTM)-where CTM can use VEM only:

- VEM.
- VEM with fixed α.
- Gibbs

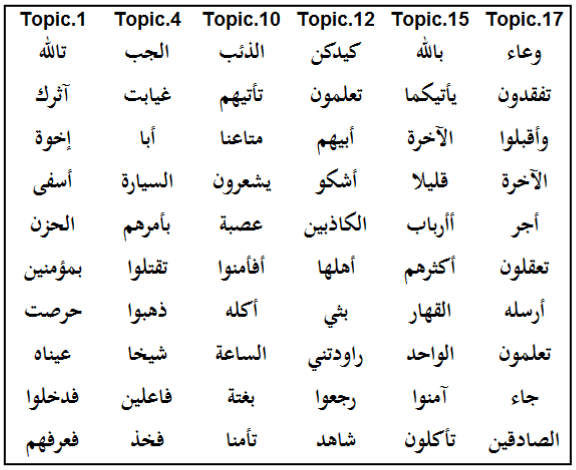
Then, a validation technique that is based on the log-Likelihood of the data set is calculated. This is performed to find the best number of topics for each structure of that chapter. The best number of the topics is calculated using 10-fold cross-validation technique for the two structures with the three term shapes, results of log-likelihood and number of topics are shown in table II. Then the topics are recorded for all cases using the best topic numbers that are calculated according to the aforementioned technique. In some cases different topics number is chosen because the energy-based topics number is large. The main parameters are set as suggested by [11] where $\alpha = 50/k$ (where k is the number of topics) and $\beta = 0.1$. In many of the experiment setups, the seed parameter of the LDA and CTM models are set to the number of terms according to table I.

Samples of the results of the topics are shown in figures 1 - 13 for the two structures with three shapes of the terms: words, roots and stems. Figures 1 - 9 represent the *Verses* structure where figures 1 - 3 are for words, figures 4 - 6 are for stems and figures 1 - 3 are for roots. Figures 10 - 13 represent the *Hizb Quarters* structure for words, roots and stems.



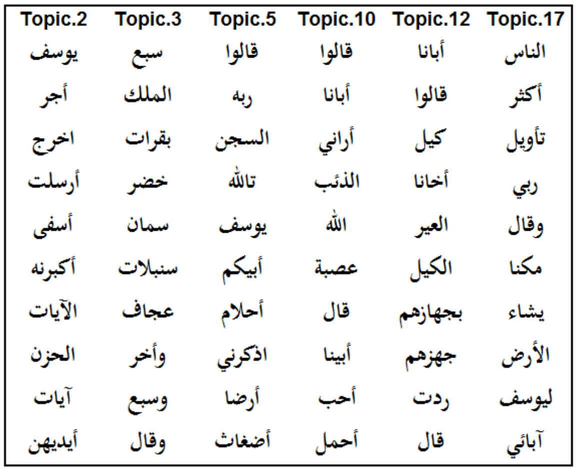
Topic.2	Topic.4	Topic.8	Topic.11	Topic.13	Topic.15
أبا	آيات	الراحمين	الجب	يشعرون	روح
الآخرة	ادخلوا	نجزي	أبرئ	أبيهم	أباه
الخائنين	أشده	أأرباب	آتيناه	أجمعين	أبينا
الواحد	اقتلوا	أباهم	الذئب	أخنه	اجعلوا
آمنوا	الذئب	الرحيم	السقاية	أستخلصه	إخوة
أهلهم	القصص	السوء	النفس	الساعة	أسفى
بالله	أمة	القديم	أمرهم	تقتلوا	أكله
بصيرة	بثمن	تزرعون	أنبئكم	خاطئين	الباب
بغتة	تبتئس	ذكر	بالله	سنبله	الصاغرين
جئنا	تعقلون	رحله	تفقدون	عبادنا	أنزلناه

Fig. 1: Sample of topics for words based on **Verses** where **Gibbs sampling** is used (Topics Number is 17)



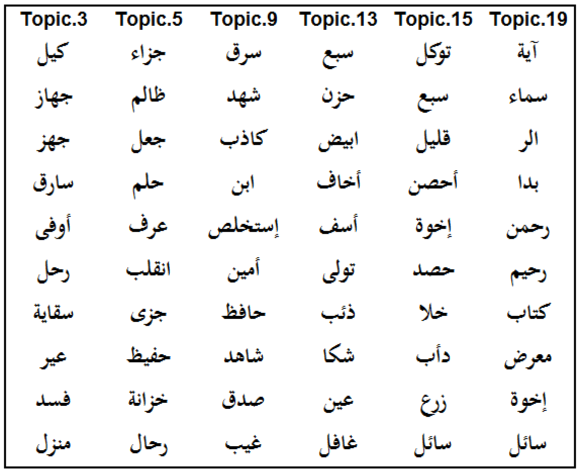
Topic.1	Topic.4	Topic.10	Topic.12	Topic.15	Topic.17
تالله	الجب	الذئب	كيدكن	بالله	وعاء
آثرك	غيابت	تأتيهم	تعلمون	يأتيكما	تفقدون
إخوة	أبا	متاعنا	أبيهم	الآخرة	وأقبلوا
أسفى	السيارة	يشعرون	أشكو	قليلا	الآخرة
الحزن	بأمرهم	عصبة	الكاذبين	أأرباب	أجر
بمؤمنين	تقتلوا	أفأمنوا	أهلها	أكثرهم	تعقلون
حرصت	ذهبوا	أكله	بثي	القهار	أرسله
عيناه	شيخا	الساعة	راودتني	الواحد	تعلمون
فدخلوا	فاعلين	بغتة	رجعوا	آمنوا	جاء
فعرفهم	فخذ	تأمنا	شاهد	تأكلون	الصادقين

Fig. 2: Sample of topics for **words** based on Verses where **CTM** is used (Topics Number is 17)



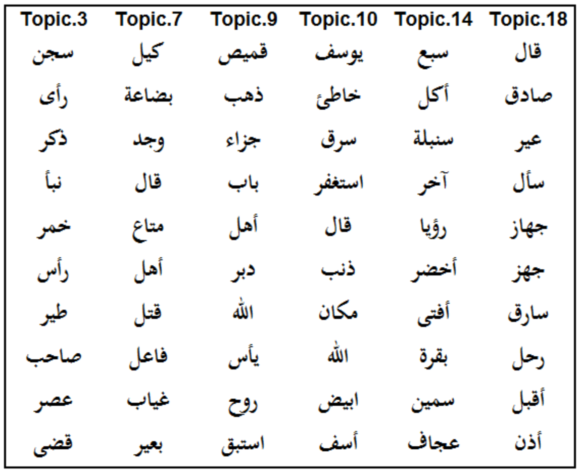
Topic.2	Topic.3	Topic.5	Topic.10	Topic.12	Topic.17
يوسف	سبع	قالوا	قالوا	أبانا	الناس
أجر	الملك	ربه	أبانا	قالوا	أكثر
اخرج	بقرات	السجن	أراني	كيل	تأويل
أرسلت	خضر	تالله	الذئب	أخانا	ربي
أسفى	سمان	يوسف	الله	العير	وقال
أكبرنه	سنبلات	أبيكم	عصبة	الكيل	مكنا
الآيات	عجاف	أحلام	قال	بجهازهم	يشاء
الحزن	وأخر	اذكرني	أبينا	جهزهم	الأرض
آيات	وسبع	أرضا	أحب	ردت	ليوسف
أيديهن	وقال	أضغاث	أحمل	قال	آبائي

Fig. 3: Sample of topics for **words** based on Verses where **TF** is used (Topics Number is 17)



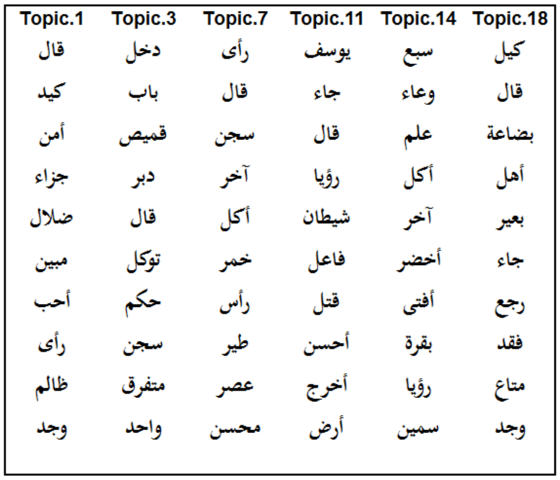
Topic.3	Topic.5	Topic.9	Topic.13	Topic.15	Topic.19
كيل	جزاء	سرق	سبع	توكل	آية
جهاز	ظالم	شهد	حزن	سبع	سماء
جهز	جعل	كاذب	ابيض	قليل	الر
سارق	حلم	ابن	أخاف	أحصن	بدا
أوفى	عرف	إستخلص	أسف	إخوة	رحمن
رحل	انقلب	أمين	تولى	حصد	رحيم
سقاية	جزى	حافظ	ذئب	خلا	كتاب
عير	حفيظ	شاهد	شكا	دأب	معرض
فسد	خزانة	صدق	عين	زرع	إخوة
منزل	رحال	غيب	غافل	سائل	سائل

Fig. 4: Sample of topics for **stems** based on Verses where **VEM** is used (Topics Number is 19)



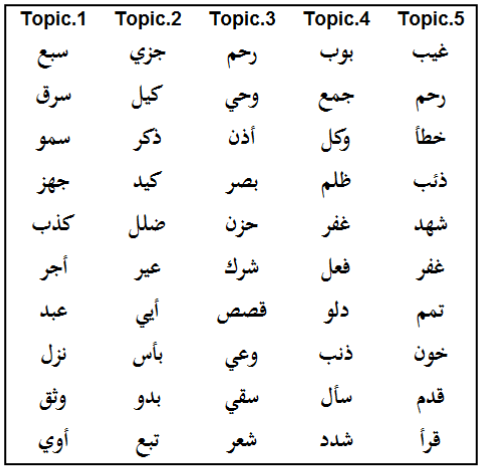
Topic.3	Topic.7	Topic.9	Topic.10	Topic.14	Topic.18
سجن	كيل	قميص	يوسف	سبع	قال
رأى	بضاعة	ذهب	خاطئ	أكل	صادق
ذكر	وجد	جزاء	سرق	سنبلة	عير
نبأ	قال	باب	استغفر	آخر	سأل
خمر	متاع	أهل	قال	رؤيا	جهاز
رأس	أهل	دبر	ذنب	أخضر	جهز
طير	قتل	الله	مكان	أفتى	سارق
صاحب	فاعل	يأس	الله	بقرة	رحل
عصر	غياب	روح	ابيض	سمين	أقبل
قضى	بعير	استبق	أسف	عجاف	أذن

Fig. 5: Sample of topics for **stems** based on Verses where **VEM with fixed** α is used (Topics Number is 19)



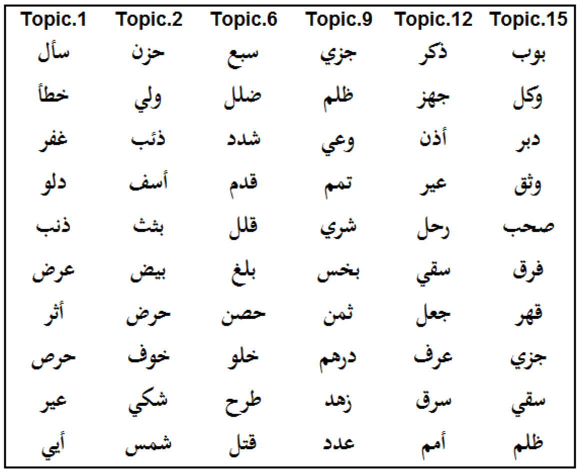
Topic.1	Topic.3	Topic.7	Topic.11	Topic.14	Topic.18
قال	دخل	رأى	يوسف	سبع	كيل
كيد	باب	قال	جاء	وعاء	قال
أمن	قميص	سجن	قال	علم	بضاعة
جزاء	دبر	آخر	رؤيا	أكل	أهل
ضلال	قال	أكل	شيطان	آخر	بعير
مبين	توكل	خمر	فاعل	أخضر	جاء
أحب	حكم	رأس	قتل	أفتى	رجع
رأى	سجن	طير	أحسن	بقرة	فقد
ظالم	متفرق	عصر	أخرج	رؤيا	متاع
وجد	واحد	محسن	أرض	سمين	وجد

Fig. 6: Sample of topics for **stems** based on Verses where **TF** is used (Topics Number is 19)



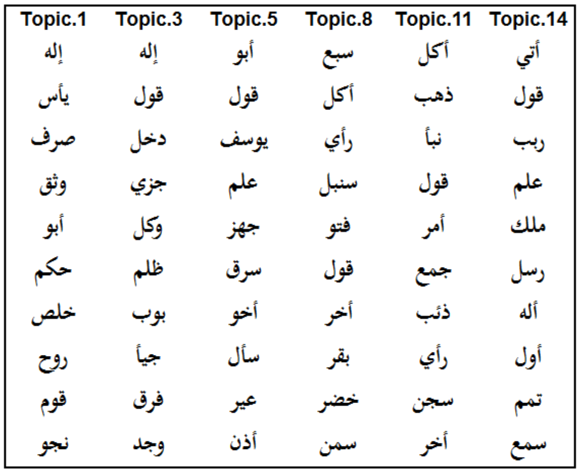
Topic.1	Topic.2	Topic.3	Topic.4	Topic.5
سبع	جزي	رحم	بوب	غيب
سرق	كيل	وحي	جمع	رحم
سمو	ذكر	أذن	وكل	خطأ
جهز	كيد	بصر	ظلم	ذئب
كذب	ضلل	حزن	غفر	شهد
أجر	عير	شرك	فعل	غفر
عبد	أبي	قصص	دلو	تمم
نزل	بأس	وعي	ذنب	خون
وثق	بدو	سقي	سأل	قدم
أوي	تبع	شعر	شدد	قرأ

Fig. 7: Sample of topics for **roots** based on Verses where **Gibbs sampling** is used (Topics Number is 5)



Topic.1	Topic.2	Topic.6	Topic.9	Topic.12	Topic.15
سأل	حزن	سبع	جزي	ذكر	بوب
خطأ	ولي	ضلل	ظلم	جهز	وكل
غفر	ذئب	شدد	وعي	أذن	دبر
دلو	أسف	قدم	تمم	عير	وثق
ذنب	بثث	قلل	شري	رحل	صحب
عرض	بيض	بلغ	بخس	سقي	فرق
أثر	حرض	حصن	ثمن	جعل	قهر
حرص	خوف	خلو	درهم	عرف	جزي
عير	شكي	طرح	زهد	سرق	سقي
أبي	شمس	قتل	عدد	أمم	ظلم

Fig. 8: Sample of topics for **roots** based on Verses where **VEM** is used (Topics Number is 15)



Topic.1	Topic.3	Topic.5	Topic.8	Topic.11	Topic.14
إله	إله	أبو	سبع	أكل	أتي
يأس	قول	قول	أكل	ذهب	قول
صرف	دخل	يوسف	رأي	نبأ	ربب
وثق	جزي	علم	سنبل	قول	علم
أبو	وكل	جهز	فتو	أمر	ملك
حكم	ظلم	سرق	قول	جمع	رسل
خلص	بوب	أخو	أخر	ذئب	أله
روح	جيأ	سأل	بقر	رأي	أول
قوم	فرق	عير	خضر	سجن	تمم
نجو	وجد	أذن	سمن	أخر	سمع

Fig. 9: Sample of topics for **roots** based on **Verses** where TF is used (Topics Number is 15)

VI. Results and Discussion

The number of documents and terms of the chapter of Joseph (PBUH) is shown in table I. Both TF and TF-IDF measures are used and then the number of documents and terms are recorded for words, roots and stems. The results of applying LDA model to the text of the chapter with Gibbs sampling technique is shown in table II. Note that the term with low energy are roots and stems compared with high energy for words.

Many experiment setups have been carried out with different parameter settings apart from the aforementioned setups in section V. Sample of the results are shown in figures 1 - 13. All the results of all the experiments show that most of the resulted topics are a mix of more than one topic. However, very few topics form one coherent topic such as topic number three of figure 3 and topic number three and fourteen of figure 5.

Some topics include a mix of two to may be five topics. In some cases all of the terms of the topic are coherent except one or two words such as topic number 12 of figure 10.

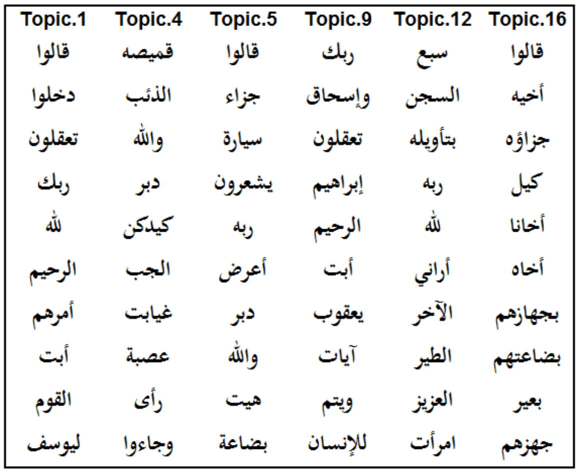
Regarding the shapes of the word; on one hand the roots are considered problematic as there are many shared words between topics such as the topics that appear in figure 12. One of the reasons behind this is that there are some different words in meaning but their root in Arabic language is the same. On the other hand, both words and stems show better results as it appear in most of the figures. For words it is obvious that each word has usually its own semantic in one context. For stems, although there is more than a word with the same stem but they have the same semantic in similar contexts.

The estimation methods that are used in this study show different ”percentage of successful” with different shapes of words. For example, TF measure gives better results than TF-IDF measure in certain cases. On another occasion, CTM gives better results. The same is true for VEM, VEM with fixed α and Gibbs sampling.

Also, it is important to mention that all of the numerical results including best number of topics as well as log-Likelihood of the data are based on the `seed` parameter for LDA and CTM models. However, many experiments are executed with different values for `seed` parameter without affecting the quality of the resulted topics.

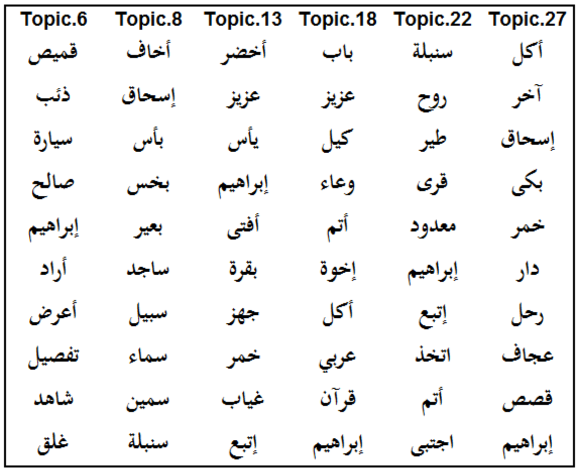
In other set of experiments, the parameter `alpha` is set to smaller numbers than that suggested by [11] where $\alpha = 50/k$ (k is the number of topics). When α is set to $1/k$, the results show topics with slightly better quality.

Although the topic modeling techniques used in this study failed to extract coherent topics, still the results are promising as some topics are coherent even that they are very few.



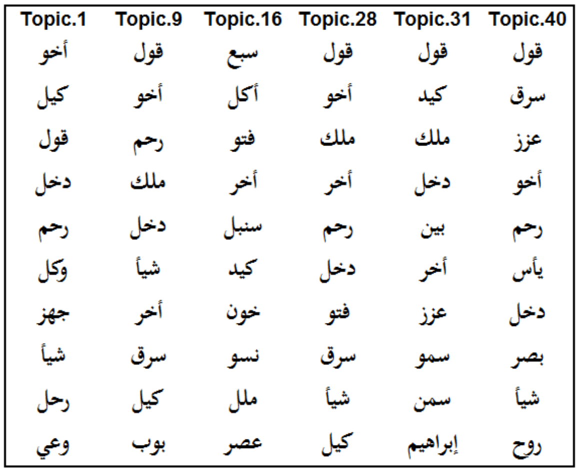
Topic.1	Topic.4	Topic.5	Topic.9	Topic.12	Topic.16
قالوا	قميصه	قالوا	ربك	سبع	قالوا
دخلوا	الذئب	جزاء	وإسحاق	السجن	أخيه
تعقلون	والله	سيارة	تعقلون	بتأويله	جزاؤه
ربك	دبر	يشعرون	إبراهيم	ربه	كيل
لله	كيدكن	ربه	الرحيم	لله	أخانا
الرحيم	الجب	أعرض	أبت	أراني	أخاه
أمرهم	غيابت	دبر	يعقوب	الآخر	بجهازهم
أبت	عصبة	والله	آيات	الطير	بضاعتهم
القوم	رأى	هيت	ويتم	العزيز	بعير
ليوسف	وجاءوا	بضاعة	للإنسان	امرأت	جهزهم

Fig. 10: Sample of topics for **words** based on **Hizb Quarters** where VEM is used (Topics Number is 17)



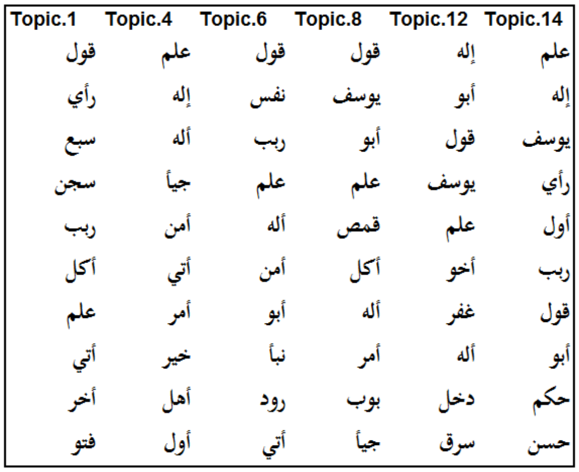
Topic.6	Topic.8	Topic.13	Topic.18	Topic.22	Topic.27
قميص	أخاف	أخضر	باب	سنبلة	أكل
ذئب	إسحاق	عزيز	عزيز	روح	آخر
سيارة	بأس	يأس	كيل	طير	إسحاق
صالح	بخس	إبراهيم	وعاء	قرى	بكى
إبراهيم	بعير	أفتى	أتم	معدود	خمر
أراد	ساجد	بقرة	إخوة	إبراهيم	دار
أعرض	سبيل	جهز	أكل	إتبع	رحل
تفصيل	سماء	خمر	عربي	اتخذ	عجاف
شاهد	سمين	غياب	قرآن	أتم	قصص
غلق	سنبلة	إتبع	إبراهيم	اجتبى	إبراهيم

Fig. 11: Sample of topics for **stems** based on **Hizb Quarters** where Gibbs sampling is used (Topics Number is 26)



Topic.1	Topic.9	Topic.16	Topic.28	Topic.31	Topic.40
أخو	قول	سبع	قول	قول	قول
كيل	أخو	أكل	أخو	كيد	سرق
قول	رحم	فتو	ملك	ملك	عزز
دخل	ملك	أخر	أخر	دخل	أخو
رحم	دخل	سنبل	رحم	بين	رحم
وكل	شيأ	كيد	دخل	أخر	يأس
جهز	أخر	خون	فتو	عزز	دخل
شيأ	سرق	نسو	سرق	سمو	بصر
رحل	كيل	ملل	شيأ	سمن	شيأ
وعي	بوب	عصر	كيل	إبراهيم	روح

Fig. 12: Sample of topics for **roots** based on **Hizb Quarters** where VEM is used (Topics Number is 44)



Topic.1	Topic.4	Topic.6	Topic.8	Topic.12	Topic.14
قول	علم	قول	قول	إله	علم
رأي	إله	نفس	يوسف	أبو	إله
سبع	أله	ربب	أبو	قول	يوسف
سجن	جيأ	علم	علم	يوسف	رأي
ربب	أمن	أله	قمص	علم	أول
أكل	أتي	أمن	أكل	أخو	ربب
علم	أمر	أبو	أله	غفر	قول
أتي	خير	نبأ	أمر	أله	أبو
أخر	أهل	رود	بوب	دخل	حكم
فتو	أول	أتي	جيأ	سرق	حسن

Fig. 13: Sample of topics for **roots** based on **Hizb Quarters** where TF is used (Topics Number is 15)

VII. Conclusion

The `topicmodels` *R* package has been used to analyse the underlying topics of the chapter Joseph (PBUH). First the best number of topics for the two structures have been calculated for the three shapes of words and the results are shown in table I. After that, several experiment setups are executed for both of the document structures with three term shapes: word, root and stem. Then, results are recorded and samples of the result are shown in figures 1 - 13. The results are evaluated based on understanding of the meanings and interpretation of the chapter of Joseph (PBUH). The results suggest that verses structure is better than Hizb quarters one in forming more coherent topics. Most of the resulted topics include a mix of more than one topic out of the main topics of the chapter of Joseph (PBUH). However, few of the resulted topics contain one coherent topic.

Semantic search in the holy Quran can be supported by finding accurate coherent topics which helps in finding contextual terms related to the user search terms. The holy Quran contains hundreds of topics if not thousands. While one verse may contain multiple topics, another set of verses may comprise one topic. Also, one topic may repeat in several contexts and in more than one chapter. If the results are enhanced by combining LDA with another technique then they can be then used together to search for relevant words according to the distribution of topics over words.

The results of this study strongly suggests that while statistical methods succeeded in extracting important topics from text corpora of humans -as many studies show, it failed to achieve the same results with the word of God. This is obvious because the words of God are unlimited in meaning and are one of the attributes/characters of God.

Future work may include exploring more statistical methods and/or combining the methods used in this study with other data mining techniques. Also, if the text of the holy Quran would be linked to one of its authentic interpretations, then topic modeling might find coherent topics because interpretations are the word of human.

REFERENCES

[1] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," vol. 8, no. 4, pp. 14:1–14:22, Dec. 2009. [Online]. Available: http://doi.acm.org/10.1145/1644879.1644881

[2] M. Saad and W. Ashour, "Arabic morphological tools for text mining," in *6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, 2010*, 2010, p. 112117.

[3] N. Y. Habash, *Introduction to Arabic Natural Language Processing*, G. Hirst, Ed. Morgan and Claypool Publishers, 2010.

[4] M. DIAB and N. HABASH, "Arabic dialect tutorial," in *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL07)*, 2007, pp. 29–34.

[5] M. Shoaib, M. Nadeem Yasin, U. Hikmat, M. Saeed, and M. Khiyal, "Relational wordnet model for semantic search in holy quran," in *International Conference on Emerging Technologies, 2009. ICET 2009.*, Oct 2009, pp. 29–34.

[6] I. Biro, "Document classification with latent dirichlet allocation," Ph.D. dissertation, Eötvös Loránd University, 2009.

[7] P. Crossno, A. Wilson, T. Shead, and D. Dunlavy, "Topicview: Visually comparing topic models of text collections," in *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*, Nov 2011, pp. 936–943.

[8] A. Kelaiaia and H. Merouani, "Clustering with probabilistic topic models on arabic texts," in *Modeling Approaches and Algorithms for Advanced Computer Applications*, ser. Studies in Computational Intelligence, A. Amine, A. M. Otmane, and L. Bellatreche, Eds. Springer International Publishing, 2013, vol. 488, pp. 65–74.

[9] D. Blei and J. Lafferty, "Topic models," in *Text Mining: Theory and Applications*, Srivastava and M. Sahami, Eds. Taylor and Francis, 2006.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[11] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, Apr. 2004.

[12] T. L. Griffiths, J. B. Tenenbaum, and M. Steyvers, "Topics in semantic representation," *Psychological Review*, vol. 114, p. 2007, 2007.

[13] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent Semantic Analysis: A Road to Meaning.*, T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Eds. Laurence Erlbaum, 2006.

[14] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[15] G. K. Gerber, R. D. Dowell, T. Jaakkola, and D. K. Gifford, "Automated discovery of functional generality of human gene expression programs." *PLoS Computational Biology*, vol. 3, no. 8, 2007.

[16] J. Boyd-Graber, D. M. Blei, and X. Zhu, "A topic model for word sense disambiguation," in *Empirical Methods in Natural Language Processing*, 2007.

[17] S. Gerrish and D. M. Blei, "Predicting legislative roll calls from text." in *ICML*, L. Getoor and T. Scheffer, Eds. Omnipress, 2011, pp. 489–496.

[18] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2006, pp. 178–185.

[19] Z. Li, C. Wang, X. Xie, X. Wang, and W.-Y. Ma, "Exploring lda-based document model for geographic information retrieval," in *Advances in Multilingual and Multimodal Information Retrieval*, ser. Lecture Notes in Computer Science, C. Peters, V. Jijkoun, T. Mandl, H. Mller, D. Oard, A. Peas, V. Petras, and D. Santos, Eds. Springer Berlin Heidelberg, 2008, vol. 5152, pp. 842–849.

[20] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Association for Computational Linguistics, 2008, pp. 363–371.

[21] A. A. Aliyu Rufai Yauri, Rabiah Abdul Kadir and M. A. A. Murad, "Quranic verse extraction base on concepts using owl-dl ontology." vol. 6, no. 23, pp. 4492–4498, 2013.

[22] M. S. Hikmat Ullah Khan, Syed Muhammad Saqlain and M. Sher, "Ontology-based semantic search in holy quran," vol. 2, no. 6, pp. 562–566, 2013.

[23] M. Nassourou, "Using machine learning algorithms for categorizing quranic chapters by major phases of prophet mohammads messengership," vol. 2, no. 11, pp. 863–871, 2012.

[24] A. Brahmi, A. Ech-Cherif, and A. Benyettou, "An arabic lemma-based stemmer for latent topic modeling," *Int. Arab J. Inf. Technol.*, vol. 10, no. 2, pp. 160–168, 2013.

[25] M. H. Panju, "Statistical extraction and visualization of topics in the qur'an corpus," Master's thesis, University of Waterloo, 2014.

[26] W. M. Darling, "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 642–647.

[27] M. Alhawarat, M. Hegazi, and A. Hilal, "Processing the text of the holy quran: a text mining study," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 6, no. 2, pp. 262–267, February 2015.

[28] Mushafqatar.com. (2015) Mushaf qatar. [Online]. Available: http://www.mushafqatar.com/index.php?group=gather

[29] Almaany.com. (2015) Almaany. [Online]. Available: http://www.almaany.com/quran-b/

[30] B. Grn, J. Kepler, U. Linz, K. Hornik, and W. W. Wien, "topicmodels: An r package for fitting topic models," *Journal of Statistical Software*, vol. 3, no. 8, 2011.