It should be mentioned that the two translations contain two different word counts, namely, translation(A) contains~400000 and the number is ~60000 for the translation (B). To divide the translations to the same number of documents different normalizations were used. We will refer to translation (A) as (A), and the translation (B) as (B).

…

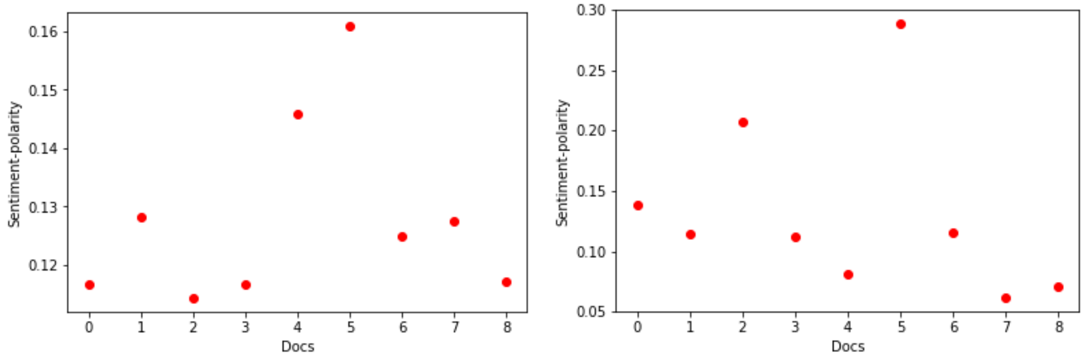


Figure 1. a) Sentiment analysis of Translation (A) b) Sentiment analysis of translation (B)

Translation (B) generally show more positive sentiments than translation (A) with a greater variation between positive and negative sentiments.

To further explore the word distribution of two translations, the-most-common words for (A), and (B) are mapped in figure 2. There are many overlaps between the 50 most-common words of (A) and (B).

For example, the words “King”, “god”,” night”, ”love” are observed in both (A) and (B).

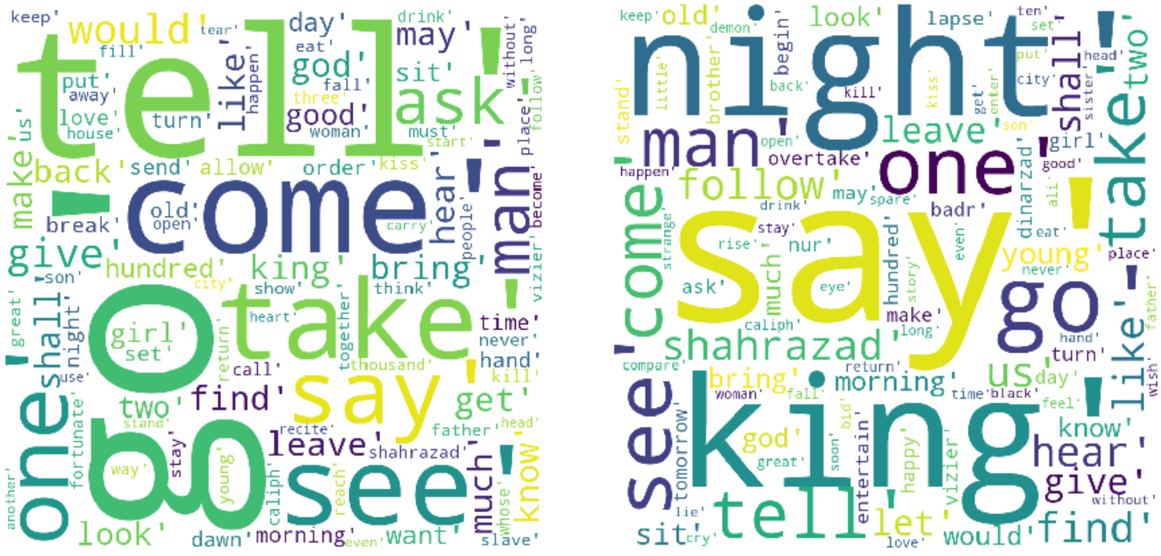


Figure 2. The 50 most common word map of a) Translation (A), b) Translation (B)

To determine the origins of the Arabian Nights, we need to assign meaningful vectors (mathematical representation) to individual words in the tale. The vectors will represent a word and will hypothetically point to a spatial location in a higher-dimensional space. Vectors can be added or subtracted and meaningful information can be extracted by conducting cosine similarities between vectors. Similar words will point to the same location in the higher-dimensional space. The process of assigning vectors to each word and giving similar words the same mathematical representation called "word embedding". Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing where words or phrases from the vocabulary are mapped to vectors of real numbers.

In 2012, Thomas Mikolov summarized the meaning of words in a modest number of vector dimensions. Mikolov trained a neural network to predict word occurrences near each target word. In 2013, Mikolov and his teammates released the software for creating these word vectors and called it Word2vec.

Word2vec learns the meaning of words merely by processing a large corpus of unlabeled text. No one has to label the words in the Word2vec vocabulary which means it is a form of unsupervised learning. For instance, we may not directly inform the model that soccer is a sport. The model comes to learns this fact by considering how much the two words of "sport" and "soccer" appear in the same context (e.g. Sportiveness). To take advantage of this feature, in this paper, we created two different topics, namely "Persianness" and "Arabness". Persianness consists of two words (Persian) and (Persians); "Arabness", however, consists of the words (Arab) and (Arabs). The same analogy was used to create the topic "Indianness", but our model concluded a weak correlation between Indianness and the Arabian Nights and therefore suggests that Indian culture plays an insignificant role in influencing the overall narrative.

Resulting word vectors will be able to identify synonyms, antonyms, or words that just belong to the same category, such as people, feelings, places, actions, names, or concepts. This is also possible with statistical methods, for example with latent semantic analysis (LSA), but the word2vec method implements tighter limits. Tighter limits on a word's neighborhood will be reflected in higher accuracy of the word vectors. Statistical approaches such as Latent semantic analysis of words and n-grams did not capture all the literal meanings of a word, much less the implied or hidden meanings. Some of the connotations of a word are lost with LSA's oversized bags of words. They also, artificially, give the most-frequent words higher statistical significance which would introduce noise to the results.

The neural language model, word2vec, was trained using two different Arabian Nights translations in which first translation (A) contains 400000 tokens and translation (B) contains 70000 tokens. Before training the Word2vec, The robust preprocessing algorithms will reduce noise and will exclude unimportant words (i.e. stopwords). Figure 3 demonstrates a flowchart that has been used to train a word2vec model.

The input includes the corpus of Arabian Night's reliable translations which enters the first filter called Tokenizer. Tokenization is to give each word its own identity and isolated them from the larger text data. The tokenized corpus will then enter the lemmatization filter in which words with the same roots will merge together. For example, “driver”, “driving”, “drove” are all imply the same meaning which is rooted in a root word “drive”.

Stopwords are the words that are not adding any meaning to the text. For example, article words such as “the”, “an”, etc. should be excluded fro the text for dimensionality/noise reduction purposes.

After preprocessing, the data is prepared for a training session. Training a word2vec algorithm is a form of unsupervised learning in which the model.

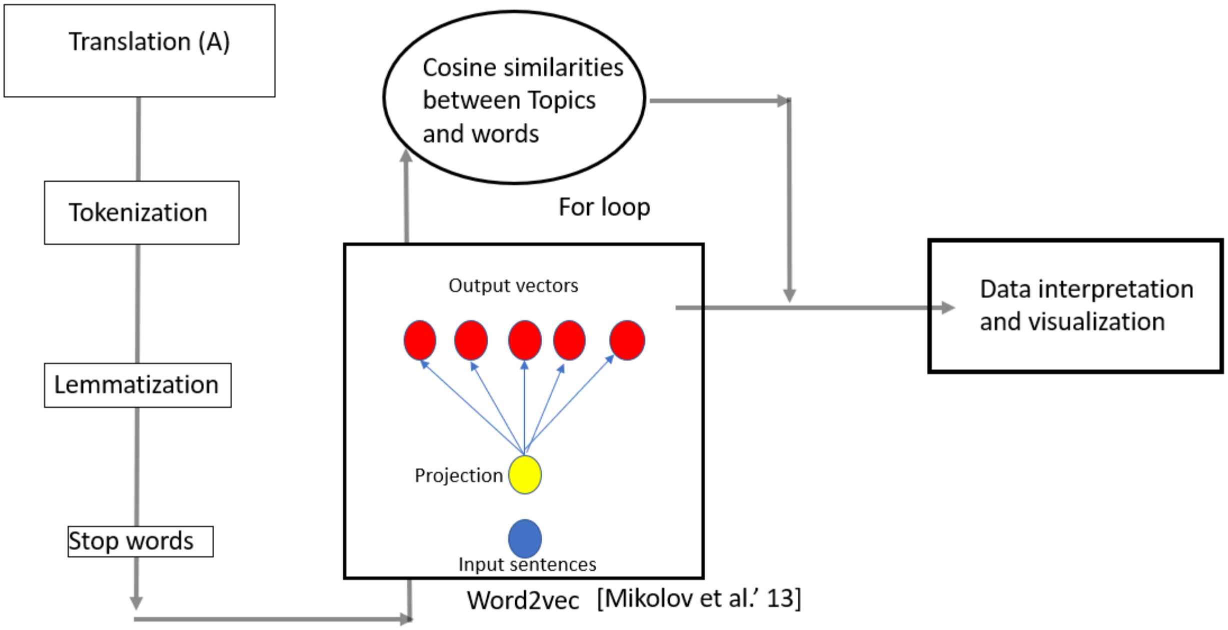

Translation (A)
Tokenization
Lemmatization
Stop words
Cosine similarities between Topics and words
For loop
Output vectors
Projection
Input sentences
Word2vec [Mikolov et al.’ 13]
Data interpretation and visualization

Figure 3. Flowchart of neural language (word2vec) algorithm to calculate cosine similarity between the topics and Arabian Nights word vectors

The word “king” is one of the-most-frequent words in the Arabian Nights. It occurred more than thousands of times in the tale, and it often represents “authority” and “power”.

Figure 4 shows the correlation and similarity between the “king” and our assigned topics, “Persianness” and “Arabness”. In general, there is a higher correlation between “Persianness” and the-most-frequent-words, and therefore we may conclude that Persian culture had a greater influence in shaping the narrative.

The similarity of the word “King” is around 5 times higher to “Persianness” in comparison to “Arabness”. Latter suggests that Persian culture is more present to shape authoritarian figures of the tale.

“Arabness”, is generally higher for Text B suggesting that the brief version is more focused on Arab culture. Persianness, however, is still around two times higher when compare the correlation between the “king” and the topics.

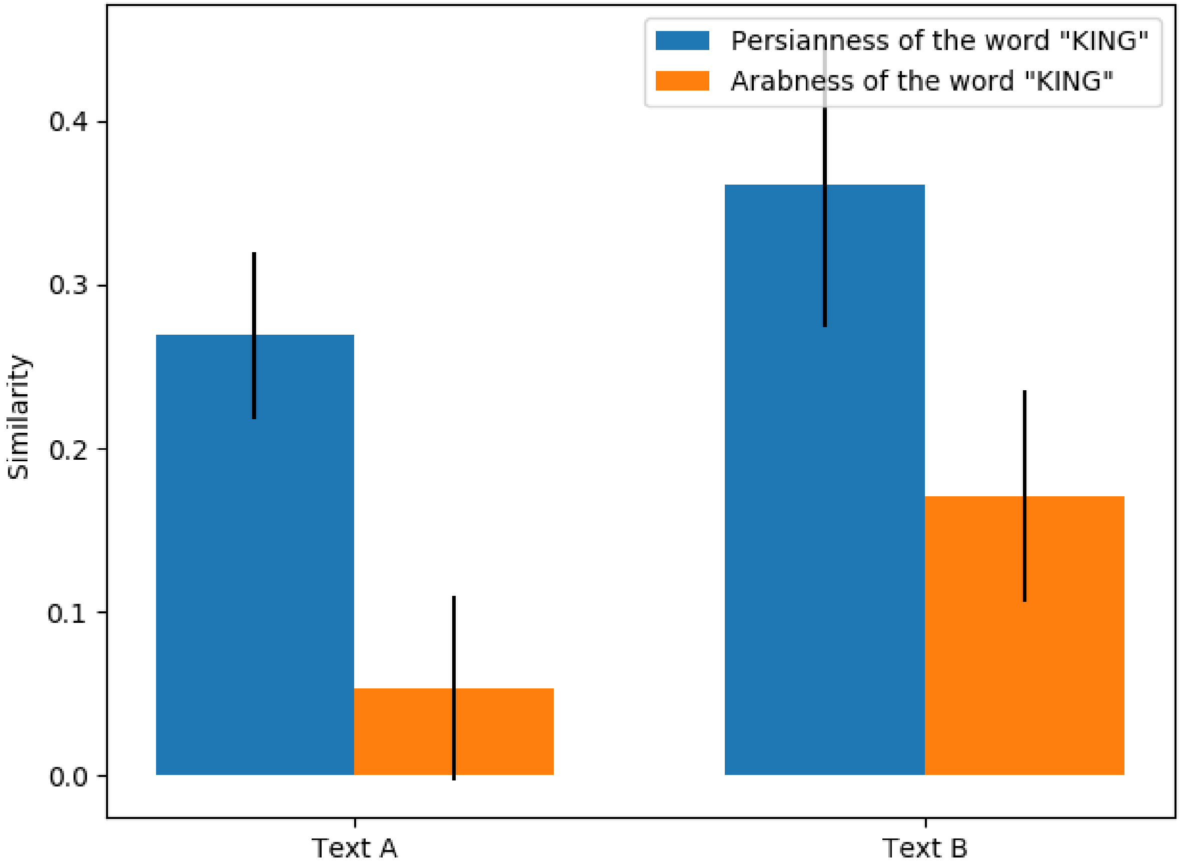


Figure 4. The similarity between the topics, “Persianness”, and “Arabness” and the word “KING” in a) Translation (A) b) Translation (B)

Figure 4 shows the “Persianness” and “Arabness” of the word, “Love”. “Love” would represent the softness and romanticism in translations and it occurred more than 800 times and it is one of the-frequent-words.

Persianness of “Love” is more than 8 times its Arabness in the text (A). Text (B) show a higher correlation between “love” and topics in which the Persianness of “Love” is close to twice of its “Arabness”.

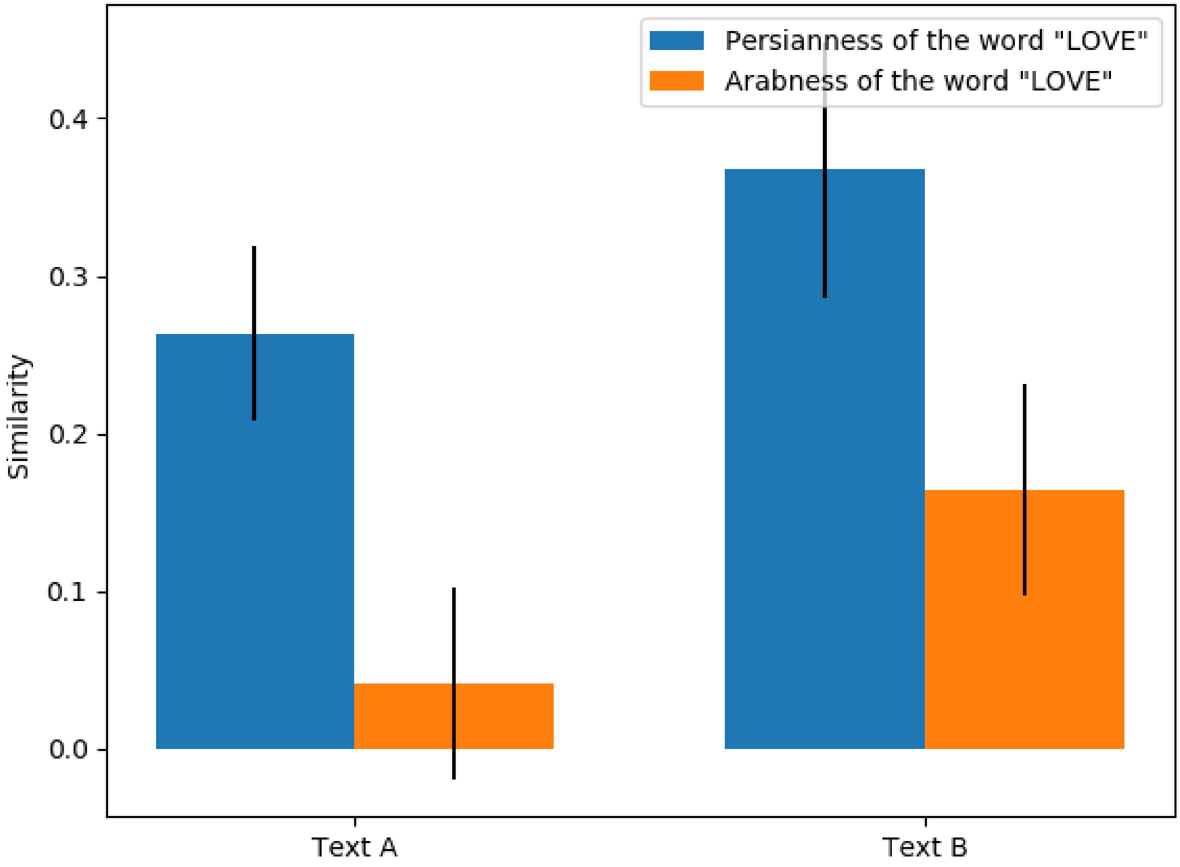
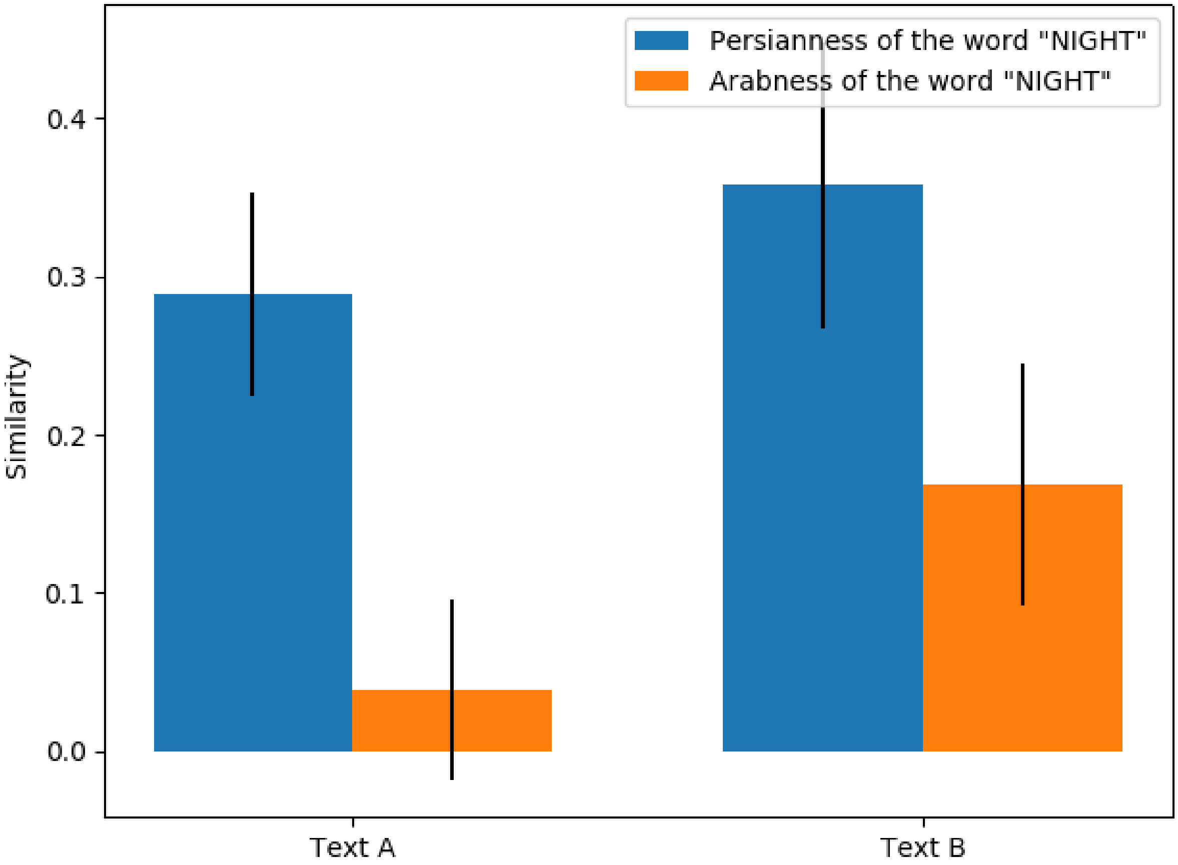
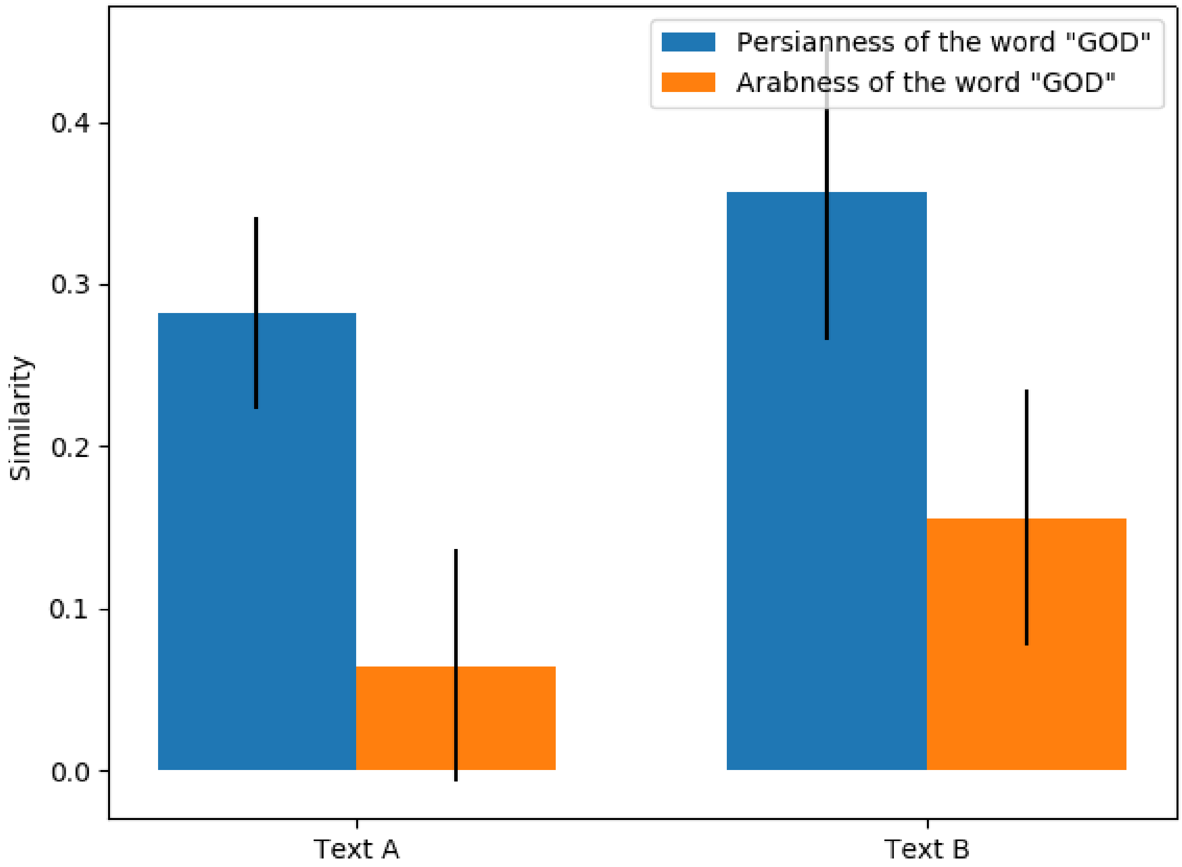
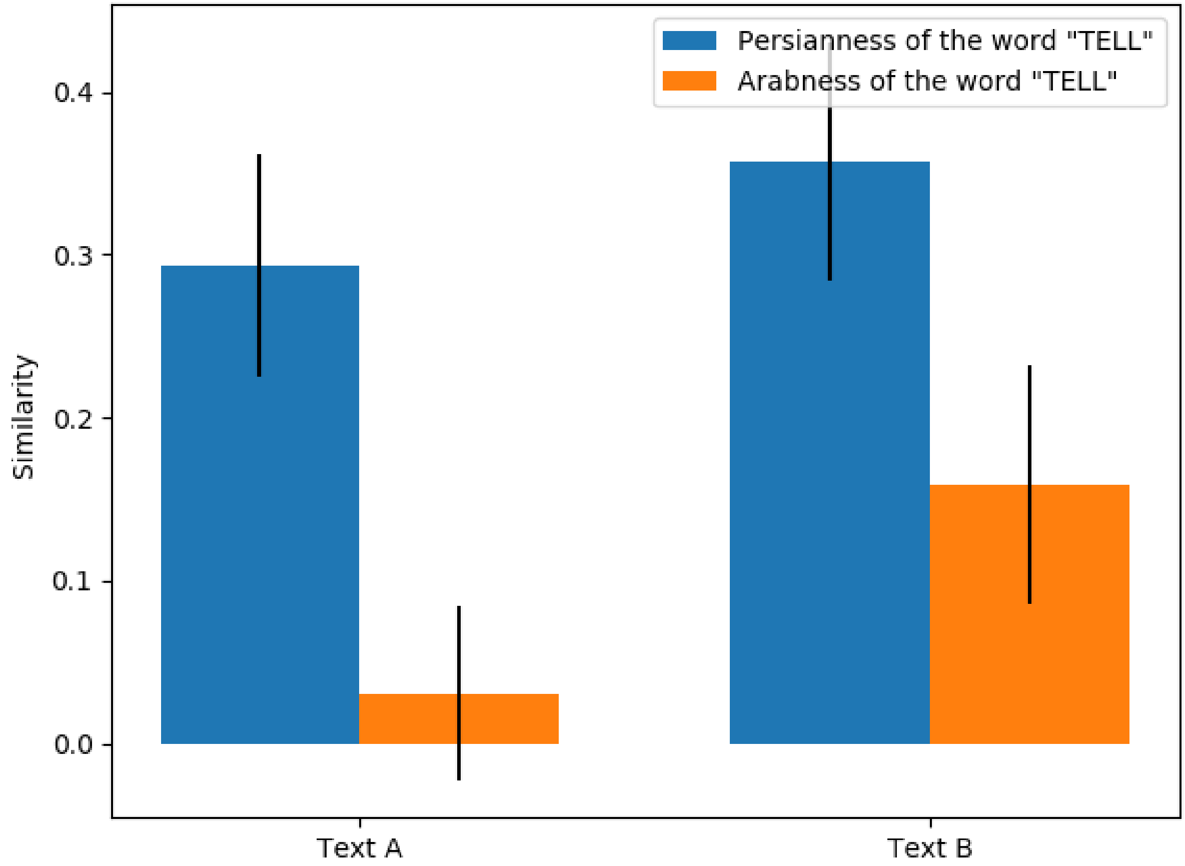
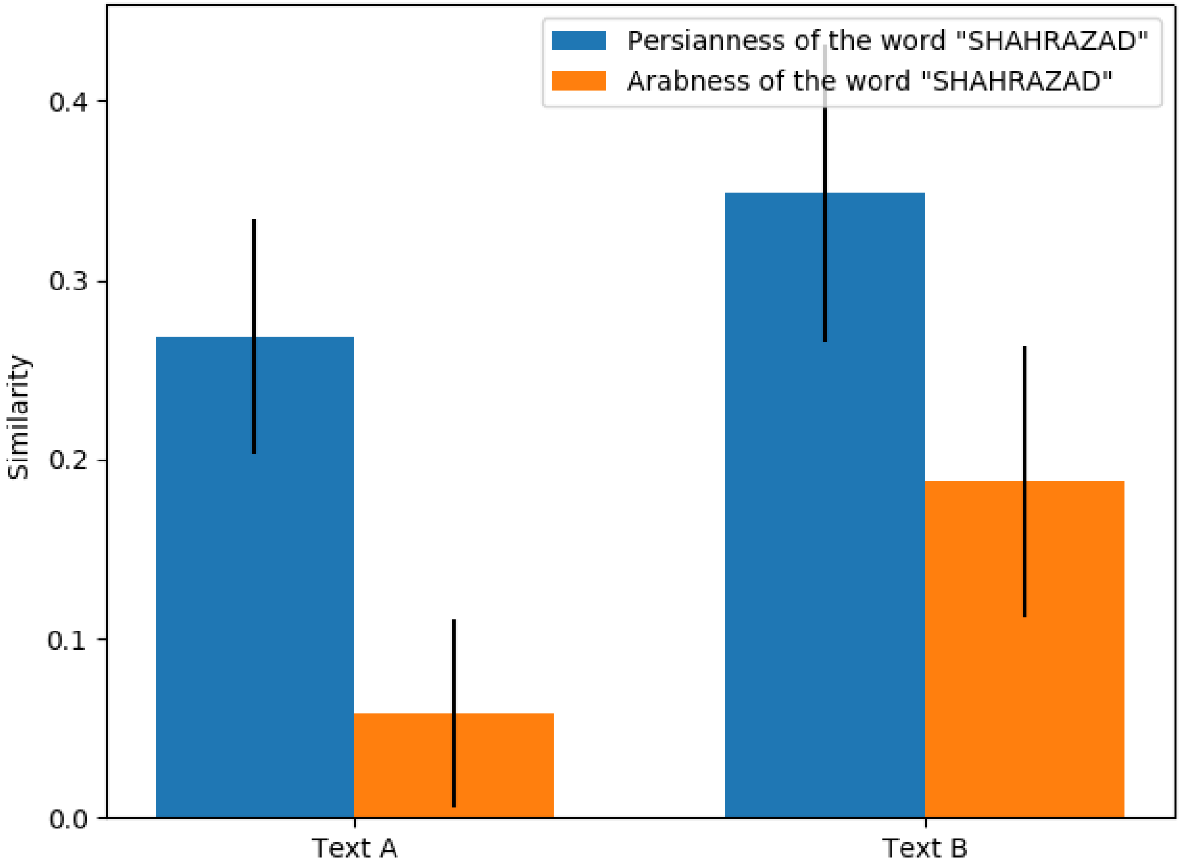


Figure 5. The similarity between the topics, “Persianness”, and “Arabness” and the word “LOVE” in a) Translation (A) b) Translation (B)


🟦 Persianness of the word "NIGHT"
🟧 Arabness of the word "NIGHT"
Similarity
0.4
0.3
0.2
0.1
0.0
Text A
Text B


Persianness of the word "GOD"
Arabness of the word "GOD"
Similarity
0.4
0.3
0.2
0.1
0.0
Text A
Text B


Persianness of the word "TELL"
Arabness of the word "TELL"
Similarity
0.4
0.3
0.2
0.1
0.0
Text A
Text B


Persianness of the word "SHAHRAZAD"
Arabness of the word "SHAHRAZAD"
Similarity
0.0
0.1
0.2
0.3
0.4
Text A
Text B