Quora Question Answering Chatbot

Cheong Sik Feng[1], Lim Jing[2], Chieu Hai Leong[2], Wong Jialiang Joshua[2]

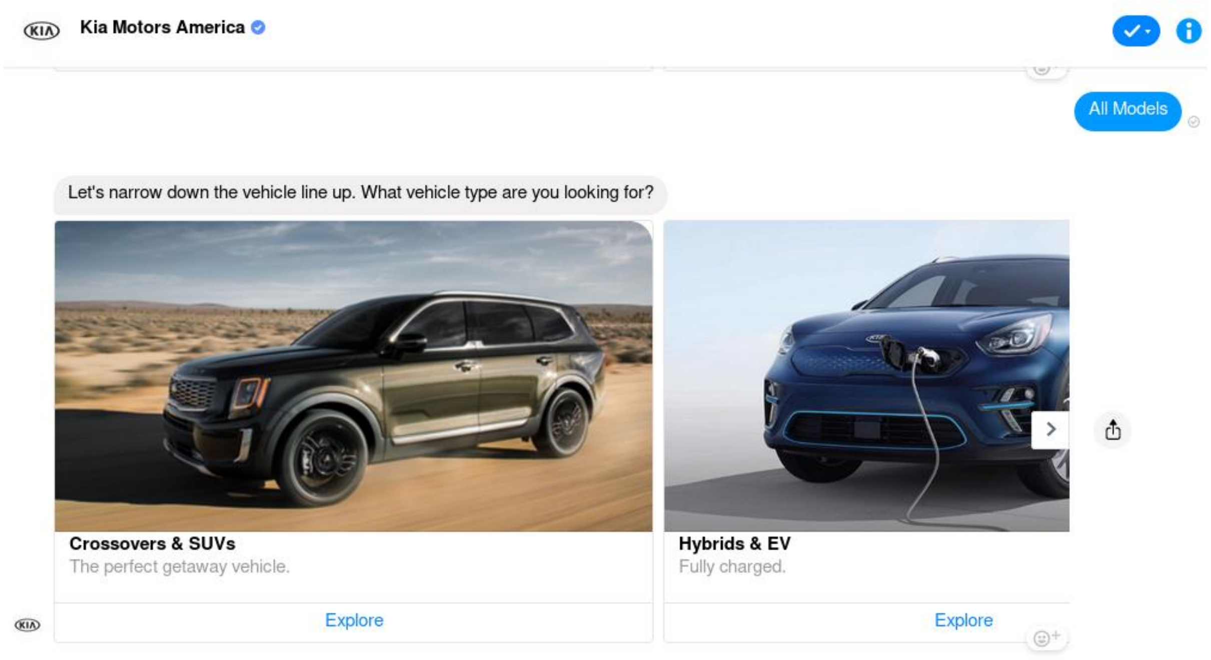
[1]NUS High School of Mathematics and Science, 20 Clementi Avenue 1 Singapore 129957
[2]DSO National Laboratories, 12 Science Park Drive Singapore 118225

ABSTRACT

Chatbots have become increasingly common in our everyday lives. It is a tool for increasing productivity through automation. Most chatbots use natural language processing (NLP) to process user inputs and allows users to have a conversation with them. In this project, we built a chatbot using term-frequency-inverse-document-frequency (tf-idf) and cosine similarity. Interaction with the chatbot is through the Telegram messaging service and allows users to ask questions. It will in turn return a set of similar questions to the user and the user can choose any query from the question set. The answer to that query will be retrieved from Quora and will be presented to the user.

INTRODUCTION

Chatbots have become increasingly common in our everyday lives. Most chatbots use NLP to process user input and converse with them. For example, KIA-Motors had made an interactive chatbot NiroBot. NiroBot is capable of answering a consumer's questions on their new electric vehicle Niro, as well as scheduling test drives. [1]



Let's narrow down the vehicle line up. What vehicle type are you looking for?

An advantage of chatbots over traditional FAQ pages is that users do not have to look through all the questions in the FAQ section to select the most similar question. The chatbot can do that for them in a much shorter amount of time.

Such a chatbot could be used in many cases. For example, an organization could find it useful when people often send questions to the support team. Many of the questions would have been asked before, and take lots of time for real people to read through and reply. The questions can be compiled together, with the answers for each question included and fed to the chatbot. This allows people to ask the chatbot directly and get the most similar questions that have already been asked. If users do not find a satisfactory answer, they can then send an email and ask the helpdesk. The new question and answer can be added to the dataset as well, for future users to refer to. This greatly reduces both the time taken for a user to get an answer, as well as manpower needed to answer questions.

Goal

The goal of this project is to write an information retrieval chatbot that will match the user's input with the most similar questions in the chatbot's dataset, and present them to the user.

We have also implemented other features such as highlighting of important terms in the questions returned by the chatbot, so users know what the key terms are; simple syntax which allows users to specify terms that must appear in the resulting queries; as well as retrieving the selected question's answer from Quora.

We evaluated tf-idf and cosine similarity on its ability to detect duplicate questions in the dataset. We also evaluated an alternative model known as doc2vec and compared it to tf-idf.

MATERIALS AND METHODS

Libraries

- Python 3.7
- Gensim
- Rasa
- Selenium
- UDA

Python is a general-purpose programming language. There have been many libraries developed for doing NLP in Python.

Gensim is a NLP library which contains many functions and models to do NLP. Gensim aims to be efficient and scalable. [2] Gensim is used to find the cosine similarity of the tf-idf vectors and vectors from the doc2vec model.

Rasa is an open source library used to build AI assistants. We have used Rasa NLU in our chatbot, which is based on Facebook's StarSpace to match the user's message to his intent. [3]

Selenium is a web automation framework. It starts a web browser and is able to do any task that can be done typically on the web. We used Selenium to fetch answers on the user selected questions from Quora online.

Unsupervised Data Augmentation (UDA) is a semi-supervised learning method which achieves state-of-the-art results on a wide variety of language and vision tasks. [4] UDA is able to generate more questions from the original question dataset through back-translation to augment the original dataset. The augmentation will generate a larger dataset than the original one. This will allow the user's query to have a higher probability of matching to one of the queries in the larger dataset.

We have integrated the chatbot to Telegram through the HTTP Telegram Bot API. This allows users to interact with the chatbot through the Telegram chat client. Telegram allows setting up of bot accounts without the need to use a phone number for them, unlike WhatsApp which requires all accounts to use a unique phone number. Telegram also allows the bot to use manual polling to receive updates from Telegram, unlike Facebook Messenger which only allows the bot to receive updates through a HTTPS webhook.

Dataset

The set of questions are obtained from the Quora Question Pair competition on Kaggle. [5] There are 2 datasets given, *train.csv* and *test.csv*.

train.csv provides 537,933 genuine questions from Quora with additional information on whether some question pairs are duplicates.

test.csv contains approximately 4 million pairs of questions. Many of them are not genuine questions from Quora because it is to prevent competitors from manually classifying the questions during the competition.

Hence *train.csv* would be more suitable. The augmented and original versions of the questions in *train.csv* are used as the corpus (set of all documents) for the chatbot, with each question as a single document.

Methods

Step 1: We preprocess each document in the corpus by tokenizing the documents and stemming the tokens. Stemming words is the process of removing suffixes from words. This allows different words with the same root to be recognized to have the same meaning when generating tf-idf vectors. For example, the 2 words “dodge” and “dodging” are different, but have the same meaning. Stemming would reduce both words to “dodg”, so they will be treated as the same word when generating tf-idf vectors. We also remove extra whitespace and punctuations as they usually are unimportant to the meaning of the document.

Step 2: We generate a bag-of-words model (BoW), which is a simplified way of representing a document by disregarding the order of the terms and only keeping track of the number of occurrences and the terms in the document.

Step 3: We then generate tf-idf vectors from the BoW. Tf-idf is a numerical statistic that shows how important a word is to a document in a corpus. Tf–idf is the product of two statistics, term frequency (tf) and inverse document frequency (idf).

$$\text{weight} = \text{tf} * \text{idf}$$

Often, tf is defined as the number of times term t appears in the document d, and idf is the logarithmically scaled inverse fraction of the documents that contain the word.

$$\text{tf} = \text{frequency of } t \text{in } d$$

$$\text{idf} = log \frac{|D|}{|\{d: t \in d, d \in D\}|}$$

The weight of each term represents a unit vector in n-dimensional space of the tf-idf vector, where n is the number of distinct terms in the whole corpus.

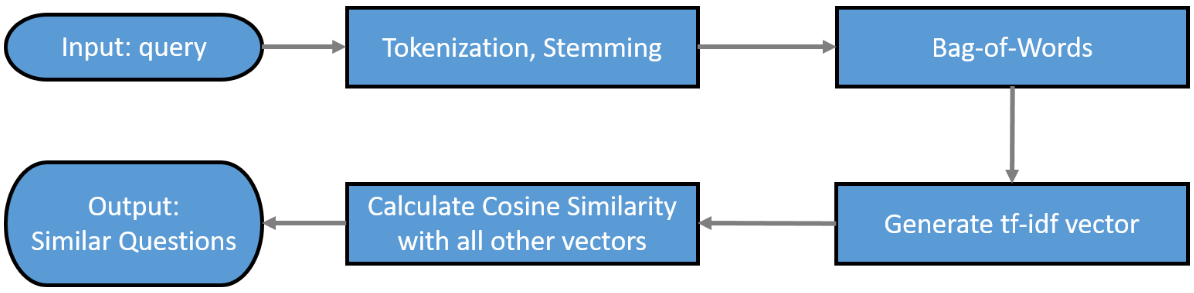
However, since the documents in the corpus are usually short, we do not consider the term frequency, and only consider the presence of the term in the document, and let

$$\text{tf} = \begin{cases} 1 & \text{if } t \in d \\ 0 & \text{otherwise} \end{cases}$$

Step 4: We use cosine similarity to calculate the similarity of the tf-idf vectors. Cosine similarity finds the angle between the 2 tf-idf vectors. 2 vectors with a smaller angle will be more similar. Since the weight of each term in any document will always be positive, the largest angle between 2 tf-idf vectors will be $\frac{\pi}{2}$, hence angle θ between 2 tf-idf vectors ranges between $\left[0, \frac{\pi}{2}\right]$. $cos\ \theta$ is strictly decreasing for $\theta \in \left[0, \frac{\pi}{2}\right]$, hence a cosine similarity close to 1 means that the 2 documents are very similar, and a cosine similarity close to 0 means the 2 documents are very different.

$$\text{cosine similarity} = cos\ \theta = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}}$$

We have preprocessed and generated the tf-idf vectors of all the sentences in the dataset to speed up retrieval time. When the user sends a request, the bot only needs to calculate the tf-idf vector of the user's input and compare it to the tf-idf vectors of the sentences in the dataset.

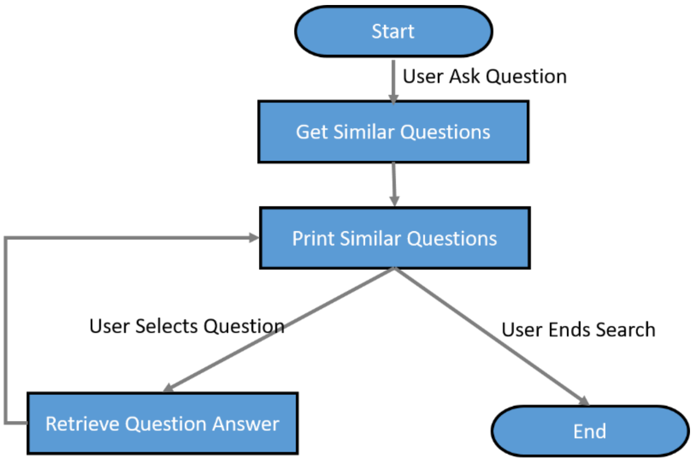

Input: query
Tokenization, Stemming
Bag-of-Words
Generate tf-idf vector
Calculate Cosine Similarity with all other vectors
Output: Similar Questions

Flowchart of how the chatbot gets processes and retrieves similar questions.

For key term highlighting, we select the 2 terms with the highest tf-idf values and highlight them unless they are inside our list of stop words. Stop words are the most common English words. They are very common so they would not be of much importance.

We have also explored the feasibility of employing the doc2vec model [7], which is an alternative to tf-idf model. Both doc2vec and tf-idf models generate a vector from a document for cosine similarity calculation. A doc2vec model is a 2 layered neural network that is trained to recognize relations between words in a document. The doc2vec model was trained with a continuous BoW (CBoW) window of 5, and trained for 20 epochs. The doc2vec model will return a vector of 100 dimensions, which we can then use cosine similarity to determine its similarity.

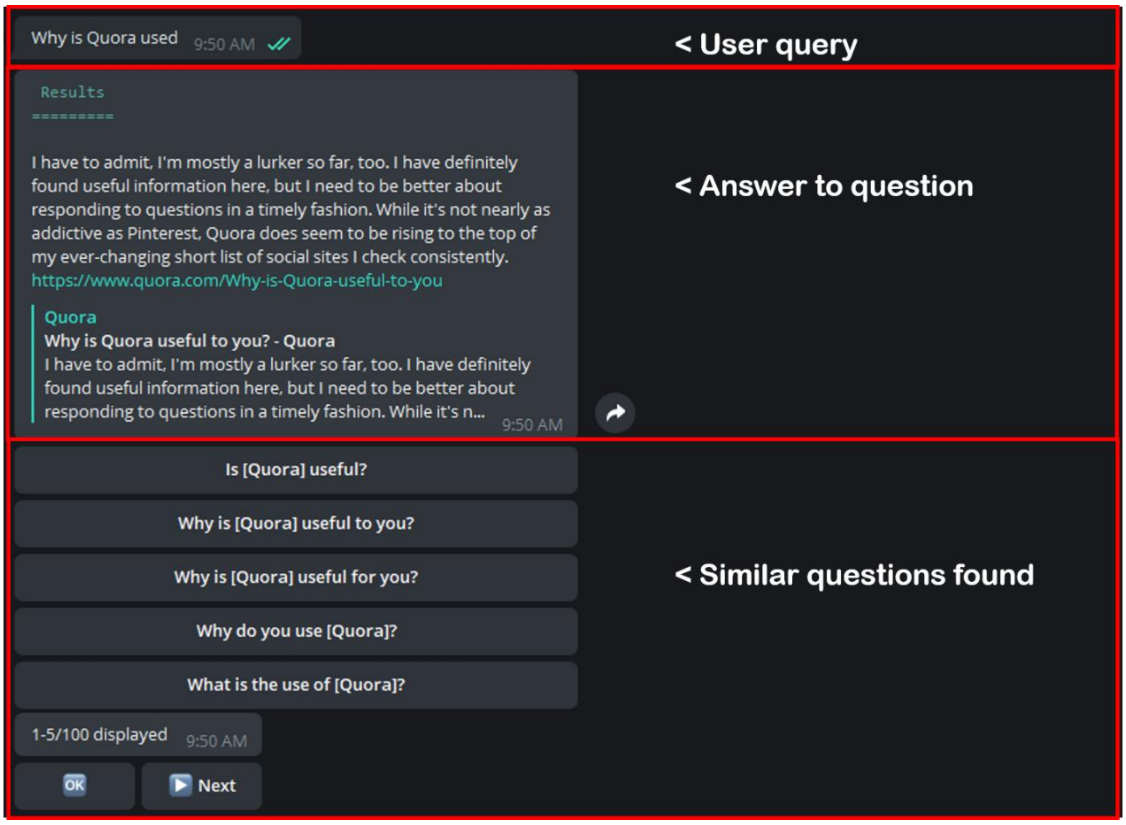
RESULTS



The diagram on the left shows the flowchart diagram of the bot.

The chatbot is able to take a user's input and match it with the most similar questions using cosine similarity and tf-idf vectors.

The user is able to then select a question and the bot will scrape Quora to get the answer to the question.



Why is Quora used 9:50 AM

< User query

```
Results
=========
```

I have to admit, I'm mostly a lurker so far, too. I have definitely found useful information here, but I need to be better about responding to questions in a timely fashion. While it's not nearly as addictive as Pinterest, Quora does seem to be rising to the top of my ever-changing short list of social sites I check consistently.
https://www.quora.com/Why-is-Quora-useful-to-you

Quora
Why is Quora useful to you? - Quora
I have to admit, I'm mostly a lurker so far, too. I have definitely found useful information here, but I need to be better about responding to questions in a timely fashion. While it's n... 9:50 AM

< Answer to question

Is [Quora] useful?
Why is [Quora] useful to you?
Why is [Quora] useful for you?
Why do you use [Quora]?
What is the use of [Quora]?
1-5/100 displayed 9:50 AM

< Similar questions found

Figure 1. Screenshot of user interface

In the screenshot, we can see that the term "Quora" is surrounded by square brackets, since it has the highest tf-idf value. The terms with the next highest tf-idf value are "useful" and "use", which are both recognized as "use" after stemming. However, they are inside the list of stop-words provided by Gensim, hence are not highlighted with square brackets.

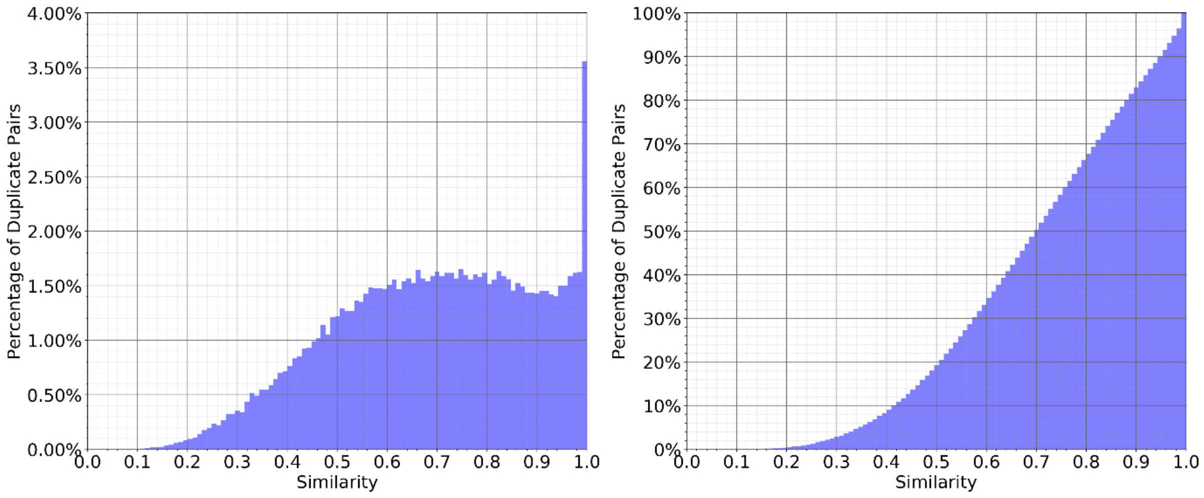


Figure 2. Probability mass plot (left) and cumulative plot (right) of similarity of duplicate question pairs using tf-idf and cosine similarity.

From the cumulative plot in Figure 2 we can see that approximately 50% of the duplicate question pairs have a similarity above 0.7, and only 20% of duplicate question pairs have similarity below 50%. This shows that using tf-idf with cosine similarity can give a good estimate of whether a question pair is duplicate.

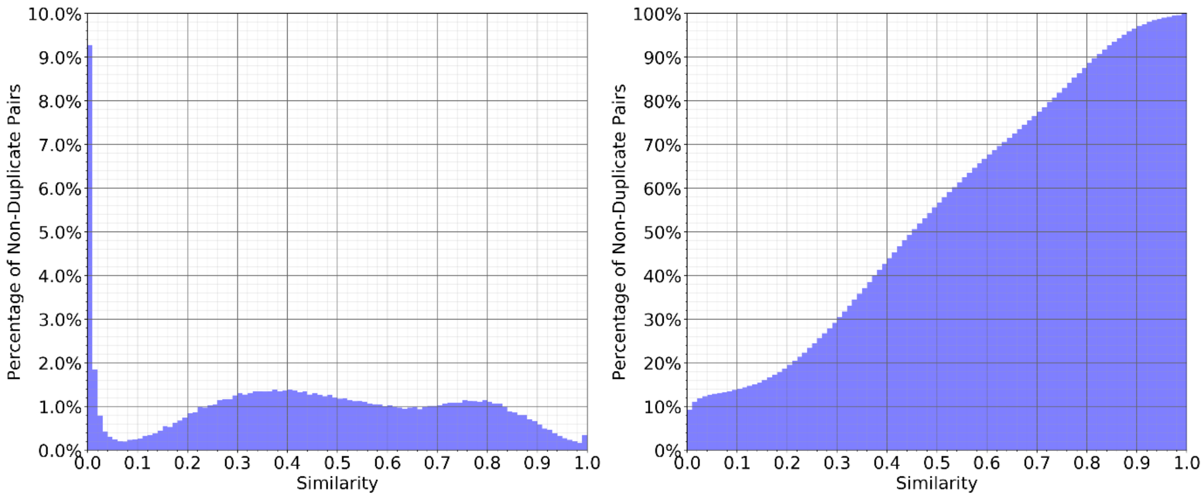


Figure 3. Probability mass plot (left) and cumulative plot (right) of similarity of non-duplicate question pairs using tf-idf and cosine similarity.

From the cumulative plot in Figure 3 we can see that approximately 25% of the non-duplicate question pairs have similarity above 0.7, and approximately 55% of non-duplicate question pairs have similarity below 0.5.

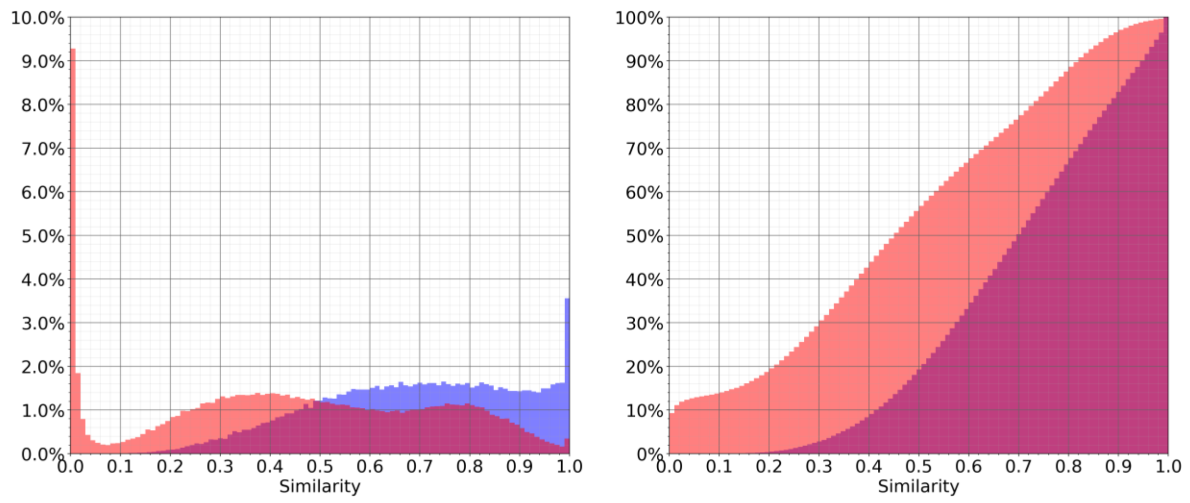


Figure 4. Superimposed plots of Figure 2 and Figure 3. Probability mass plot (left) and cumulative plot (right). Duplicate (blue), non-duplicate (red).

By comparing the probability mass plot in Figure 4, we can see that for similarity greater than 0.5, there is a higher percentage of duplicate questions than non-duplicate questions; for similarity smaller than 0.5 there is higher percentage of non-duplicate questions than duplicate questions. Hence we conclude that tf-idf with cosine similarity is a good method of measuring similarity.

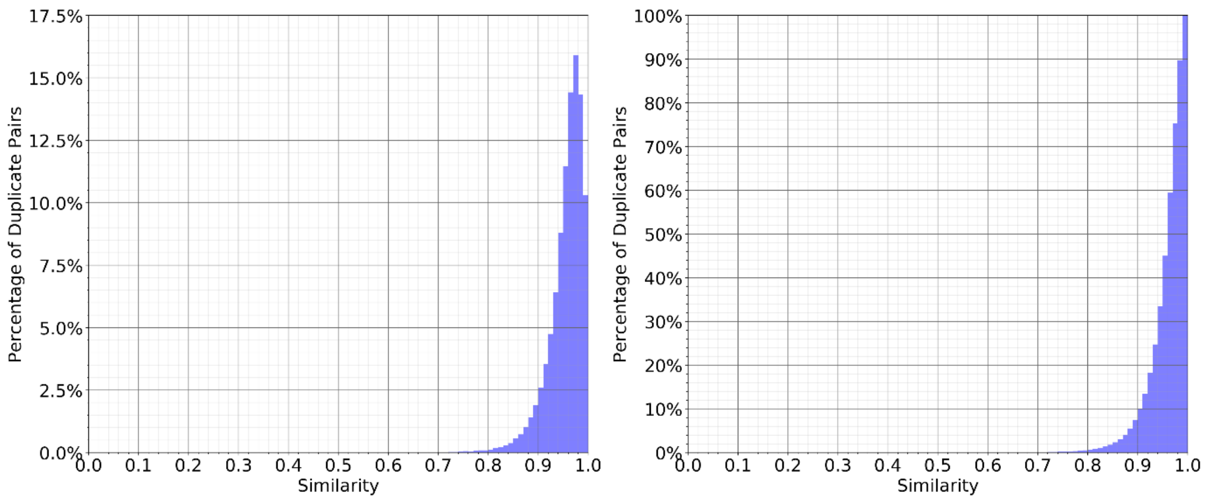


Figure 5. Probability mass plot (left) and cumulative plot (right) of similarity of duplicate question pairs using doc2vec model and cosine similarity.

From Figure 5, we see that the doc2vec model is able to identify duplicate question pairs extremely well, with more than 90% having cosine similarity greater than 0.9.

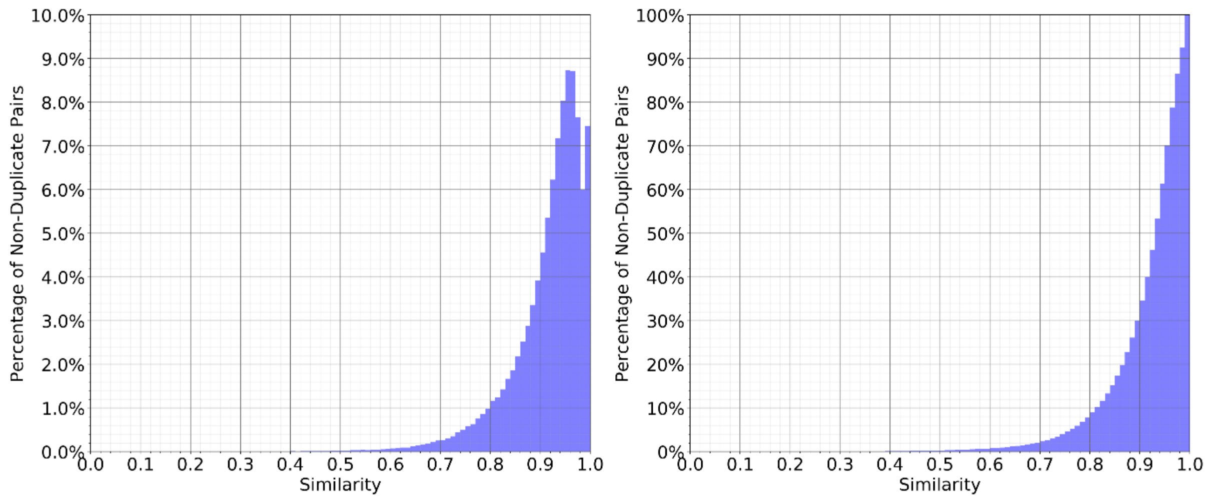


Figure 6. Probability mass plot (left) and cumulative plot (right) of similarity of non-duplicate question pairs using doc2vec model and cosine similarity.

However, from Figure 6, we also see that it does not identify non-duplicate question pairs well, since 90% of the non-duplicate question pairs have similarity greater than 0.8.

An example of a pair of non-duplicate questions having high similarity is "Why is the number for Skype at 1-855-425-3768 always busy?" and "How could I get Skype to work on an android 4.1.1 phone?" which has a similarity of 0.93.

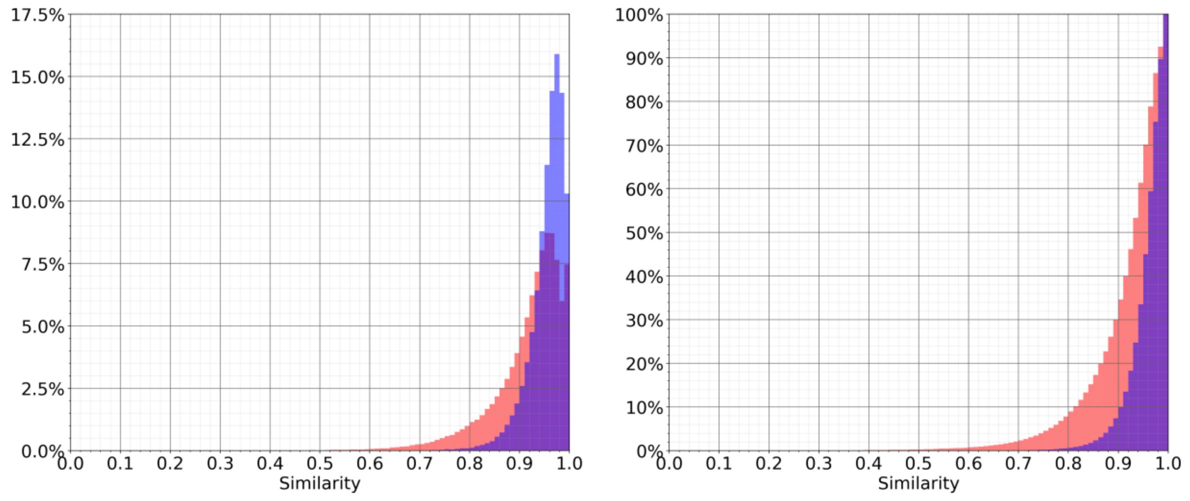


Figure 7. Superimposed plots of Figure 5 and Figure 6. Probability mass plot (left) and cumulative plot (right). Duplicate (blue), non-duplicate (red).

From Figure 7, we can also observe that only for question pairs with similarity greater than 0.94 there is a higher percentage of duplicate pairs than non-duplicate pairs. Hence this doc2vec model is not a good method to identify duplicate question pairs.

DISCUSSION

1. The first limitation of using cosine similarity and tf-idf vectors is that tf-idf does not consider the relations between words or the location of the words in sentences.
2. The second limitation of using tf-idf is that another sentence is only similar if the same words are used. However, there are usually different words that have the same meanings.
3. The third limitation of using tf-idf is that some words have multiple meanings. Tf-idf is unable to differentiate which meaning is used in the context of the sentence.
4. Our chatbot tries to retrieve answers from Quora for the question the user selected. Sometimes, the script is unable to retrieve the right answer. This is because it is using Quora's search function to get the question URL, which may not be the same exact question. This can be prevented by having an answer dataset that contains the answers for each question, which would additionally reduce time taken to retrieve the answer.
5. Using a webhook is the recommended way of receiving updates from Telegram, but I could not use it because DSO servers are behind a firewall. As a result, I used long polling to request for messages sent to the chatbot from Telegram. Using this method, no ports on the server will need to be exposed, with the only disadvantage being some additional delay to fetch the updates.
6. Even though doc2vec is able to learn relations between words from the documents during training, it was unable to differentiate duplicate and non-duplicate question pairs well.

FUTURE WORK

Our application of cosine similarity on the vectors from the doc2vec model has produced unsatisfactory results. For future work, we can improve it by using machine learning to improve the matching results from the output vectors to better determine duplicate question pairs.

We can also explore the use of Bidirectional Encoder Representations from Transformers (BERT) [8] to find duplicate question pairs. BERT is a recent method that achieves state of the art results on many NLP benchmarks.

ACKNOWLEDGEMENTS

I would like to thank my project mentors, Mr. Lim Jing, Dr. Chieu Hai Leong and Mr. Wong Jialiang Joshua for their guidance throughout the process of the research.

BIBLIOGRAPHY

[1]. Kia NiroBot - First Chatbot developed for an automotive brand in the US. (27 June, 2017). Retrieved from https://www.ansibleww.com/case-study/kia-nirobot/

[2]. Radim Řehůřek, & Petr Sojka (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). ELRA.

[3]. Tom Bocklisch and Joey Faulkner and Nick Pawlowski and Alan Nichol (2017). Rasa: Open Source Language Understanding and Dialogue Management *CoRR*, *abs/1712.05181*.

[4]. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., & Le, Q. (2019). Unsupervised Data Augmentation for Consistency Training arXiv preprint arXiv:1904.12848.

[5]. Quora. (2017). Quora Question Pairs. Available from Question Pairs Dataset | Kaggle. Retrieved from https://www.kaggle.com/quora/question-pairs-dataset

[6]. Maali Mnasri (2019) Recent advances in conversational NLP : Towards the standardization of Chatbot building *CoRR*, *abs/1903.09025*

[7]. Quoc V. Le and Tomas Mikolov (2014). Distributed Representations of Sentences and Documents *CoRR*, *abs/1405.4053*.

[8]. Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding *arXiv preprint arXiv:1810.04805*.

APPENDIX

All code written for the chatbot is available at https://gitlab.com/sikfeng/quora-qa-chatbot