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

Introduction

Chatbots have become increasingly common in our everyday lives. Most chatbots use NLP to process user input and converse with them.

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

Methods

- Tf-idf and cosine similarity to get most similar questions using Gensim[1]
- Unsupervised Data Augmentation (UDA)[2] to generate more questions from the dataset
- Rasa[3] to match user message to intent
- Selenium to retrieve answer from Quora
- Telegram HTTP BOT API for user interface

The dataset we use is obtained from the Kaggle Quora Question Pairs Competition[4]. We use train.csv for the chatbot.

Similarity

Tf-idf is a statistical model that shows how important a term is to a document in a corpus.

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

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

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. Cosine similarity is used to compare the similarity of 2 tf-idf vectors.

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

Cosine similarity finds the angle between the vectors. The weight of each term in a document is always positive, $\theta \leq \frac{1}{2}\pi$. $\cos\theta$ is strictly decreasing for $0 \leq \theta \leq \frac{1}{2}\pi$, $\therefore$ higher cosine similarity $\implies$ more similar.

Limitations

- Tf-idf does not consider relations between words
- Question pairs are similar only if they use the same words. However, they may be different words with the same meaning.
- Some words have multiple meanings depending on the words around it. Tf-idf is unable to differentiate which meaning is used
- The chatbot may not retrieve the right answer from Quora as some questions in the dataset had been deleted.
- Recommended method of receiving user messages from Telegram is through webhooks. However, we use long polling to do so instead, which does not require us to expose ports, but introduces some delay to receiving updates.

References

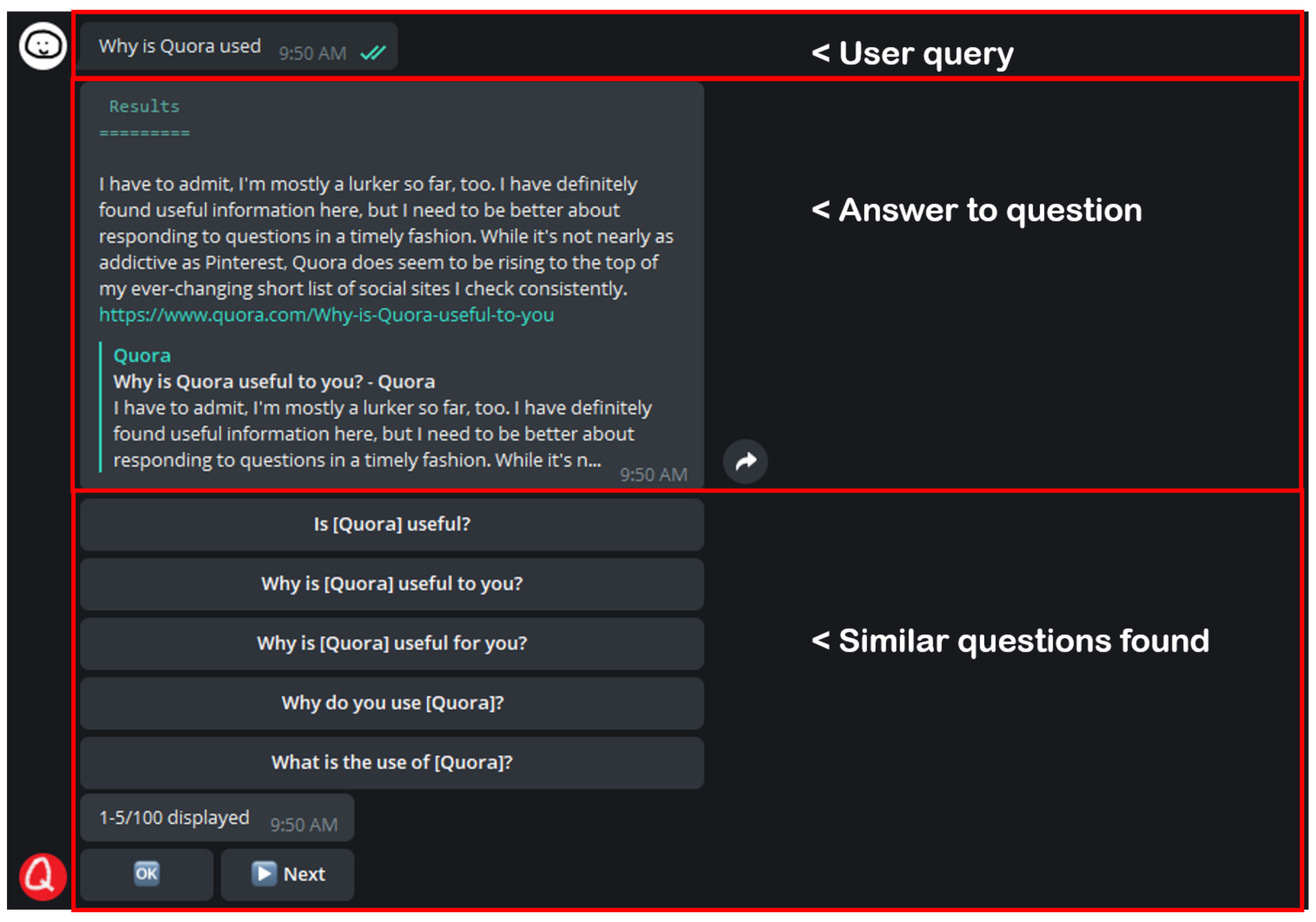
[1] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[2] Qizhe Xie et al. "Unsupervised Data Augmentation for Consistency Training". In: *arXiv preprint arXiv:1904.12848* (2019).

[3] Tom Bocklisch et al. "Rasa: Open Source Language Understanding and Dialogue Management". In: *CoRR* abs/1712.05181 (2017). arXiv: 1712.05181. URL: http://arxiv.org/abs/1712.05181.

[4] Quora. *Quora Question Pairs*. https://www.kaggle.com/c/quora-question-pairs/data. 2017.

Results



Why is Quora used 9:50 AM

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

1-5/100 displayed 9:50 AM

< User query
< Answer to question
< Similar questions found
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]?
OK
Next

Figure 1: Screnshot of chatbot and user interaction

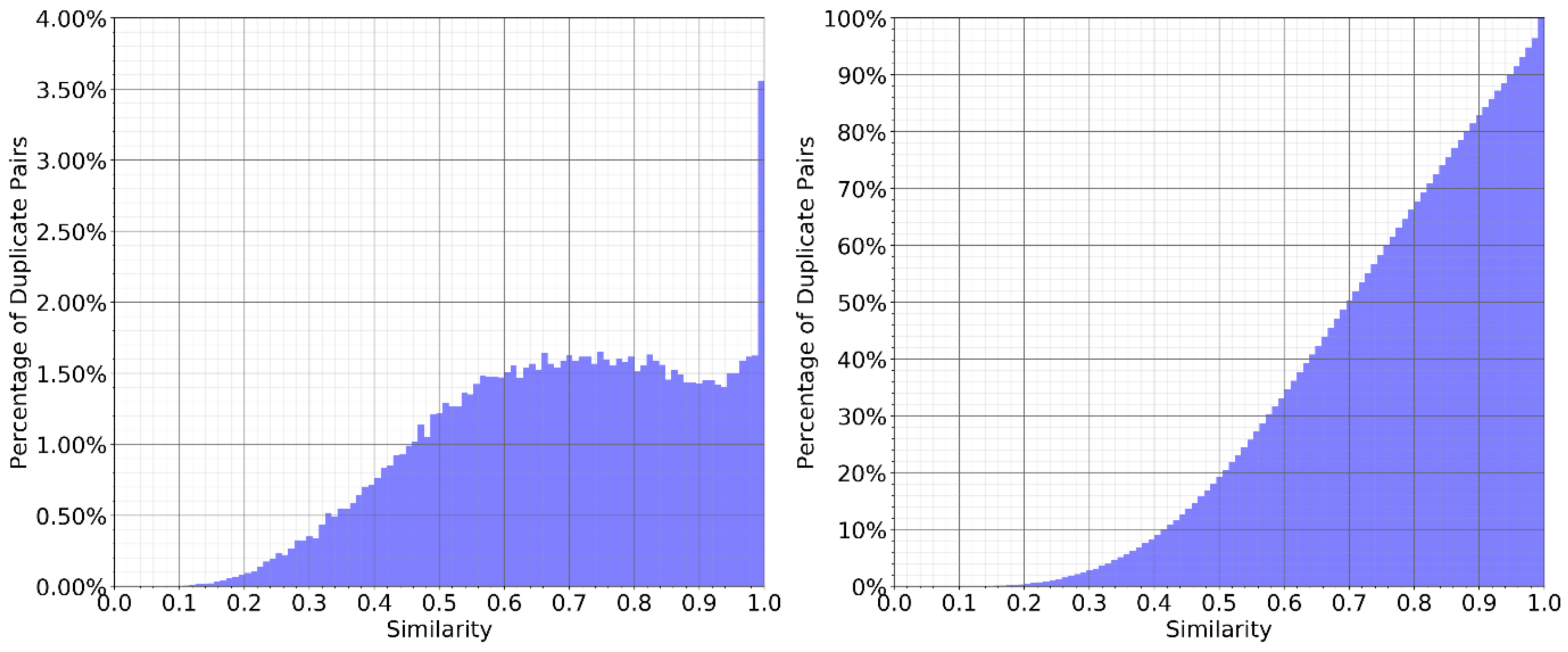


Figure 2: Probability mass (left) and cumulative (right) histogram of similarity of duplicate question pairs

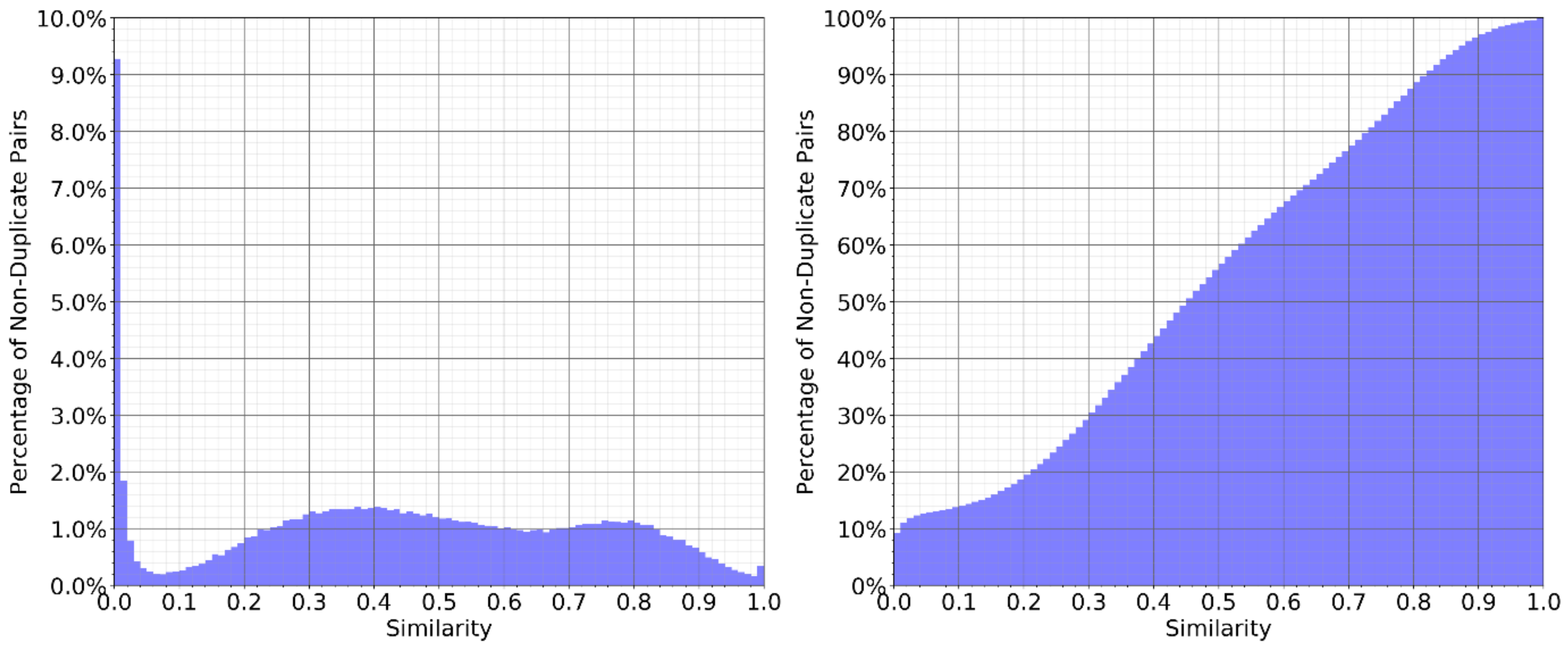


Figure 3: Probability mass (left) and cumulative (right) histogram of similarity of non-duplicate question pairs

An Initiative By: YDSP Young Defence Scientists Programme
DSTA Defence Science & Technology Agency
DSO National Laboratories