IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 10, Issue 2, February 2020, ||Series -II|| PP 55-58

Detection of Duplicate Question in Quora Using R Programming

Dhanamma Jagli[1], Anikesh S M[2], Adhitya M Nair[3], Balraj Gowda[4]
[1] Assistant Professor, [2, 3, 4] MCAFinal Year Students,
VESIT, Mumbai University, Mumbai, India.
Received 26 February 2020; Accepted 09 March 2020

ABSTRACT: Where else but Quora can physicist help chef with math problem and get cooking tips in return? "Quora" is a place to gain and share knowledge—about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world. Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. Quora values canonical questions because they provide a better experience to active seekers and writers, and offer more value to both of these groups in the long term. Doing so will make it easier to find high quality answers to questions resulting in an improved experience for Quora writers, seekers, and readers.
KEYWORDS: Quora, R programming, Keras, Tensor Flow.

I. INTRODUCTION

In order to build a high-quality knowledge base, it's important that we ensure each unique question exists on Quora only once. Writers shouldn't have to write the same answer to multiple versions of the same question, and readers should be able to find a single canonical page with the question they're looking for. For example, we'd consider questions like "What are the best ways to lose weight?", "How can a person reduce weight?", and "What are effective weight loss plans?" To be duplicate questions because they all have the same intent. To prevent duplicate questions from existing on Quora, we are trying to make a system which automatically identifies when questions with the same intent have been asked multiple times.

Quora recently released a public dataset of duplicate questions that can be used to train duplicate question detection models like the one they use at Quora. In this project, we'll give you a sense of what's possible with our duplicate question dataset by outlining a few deep learning explorations.

II. THE PROBLEM STATEMENT

Understanding semantic relatedness of sentences would allow understanding of much of the user-generated content on the internet, such as on Quora. In this paper, we address the problem of actual duplicate or exact semantic coincidence between questions. Solving this problem would be useful in helping Quora organize and duplicate their knowledge base. The pairs of questions in our problem had been already similar in that they have many low document-frequency words in common; however, the negative examples will have subtle semantic differences. Some of the differences are due to different scopes of each question. For example, a question asking why something happens is different from a question asking whether that thing happens. Conversely, our model needs to recognize when two questions use different words and phrases with the same semantic meaning, or the users' intended questions are the same and would elicit the same answers. Our model tries to learn these patterns.

III. PROPOSED MODEL ARCHITECTURE

The proposed model is useful to find the duplicate quora questions and answers so that searching accuracy is improved. The architecture is shown as in the below.

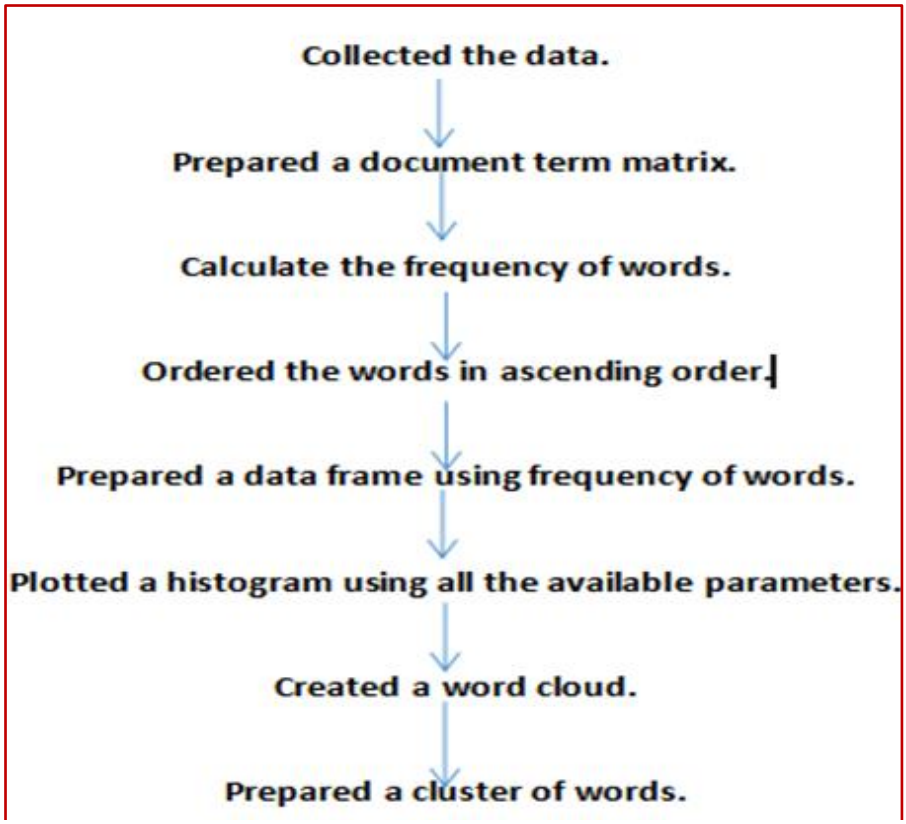

Figure 1: Workflow of proposed model

IV. PROPOSED ALGORITHM

Document-term matrix: A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. There are various schemes for determining the value that each entry in the matrix should take. One such scheme is the TF - IDF.

Word cloud: Text mining methods allow us to highlight the most frequently used keywords in a paragraph of texts. One can create a word cloud, also referred as a text cloud or tag cloud, which is a visual representation of text data.

The procedure of creating word clouds is very simple in R if you know the different steps to execute. The text mining package (tm) and the word cloud generator package (word cloud) are available in R for helping us to analyze texts and to quickly visualize the keywords as a word cloud.

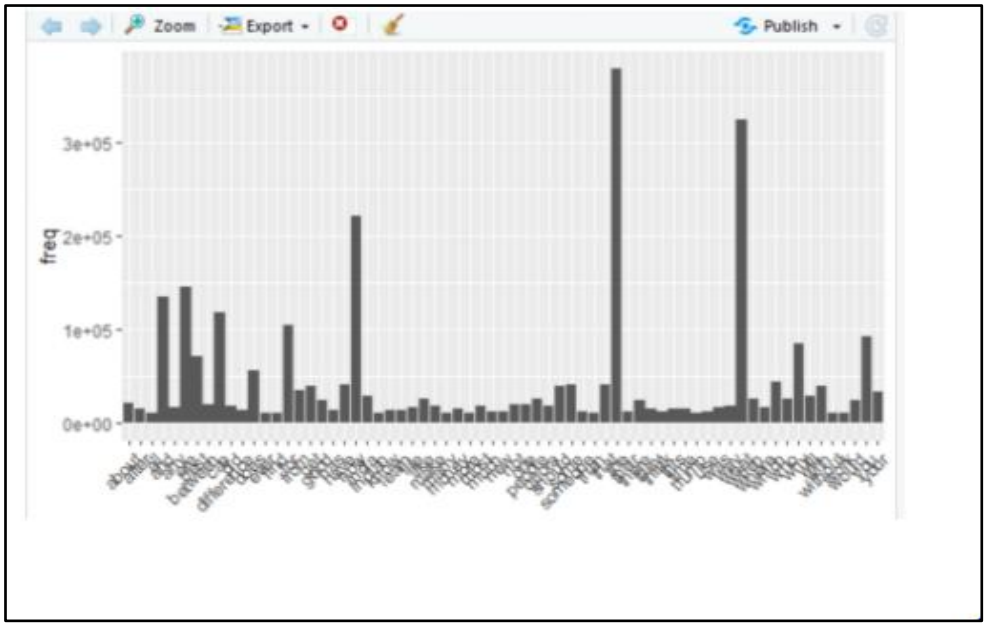

Figure 2: Histogram of Repeated Questions

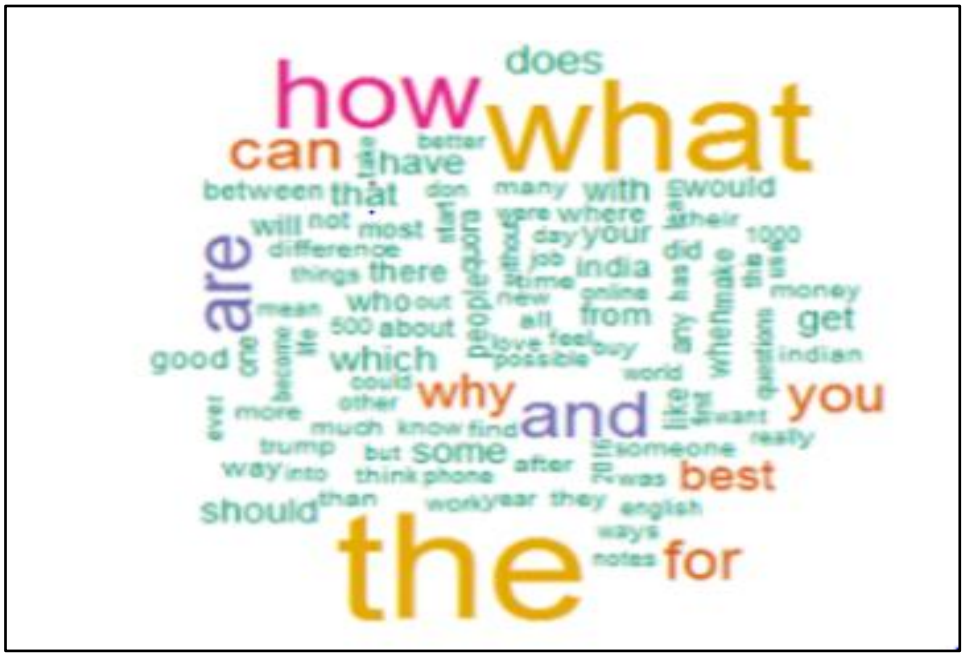

Figure 3: Word Cloud of Frequent Words

R program: R is a programming language and software environment for statistical analysis, graphic representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac. This programming language was named R, based on the first letter of the first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language S. With the help of Machine learning, deep mining and deep learning software like R programming and with the help of various readily available libraries like queries, tensorflow, code toolkit we are able to implement the project.

V. IMPLEMENTATION

In order to build a high-quality knowledge base, it's important that we ensure each unique question exists on Quora only once. Writers shouldn't have to write the same answer to multiple versions of the same question, and readers should be able to find a single canonical page with the question they're looking for. For example, we'd consider questions like "What are the best ways to lose weight?", "How can a person reduce weight?", and "What are effective weight loss plans?" To be duplicate questions because they all have the same intent. To prevent duplicate questions from existing on Quora, we are trying to make a system which automatically identifies when questions with the same intent have been asked multiple times.

Quora recently released a public dataset of duplicate questions that can be used to train duplicate question detection models like the one they use at Quora. In this project, we'll give you a sense of what's possible with our duplicate question dataset by outlining a few deep learning explorations.

This work can be taken in more detail and more work can be done on the project in order to bring modifications and additional features. The current implementation does not produce accurate results as the complexity of hardware required are too high. The current version lacks implementation of complex algorithms, which can be added in future. We can include algorithms like GloVec, Siamese LSTM, CNN so that we can get more accurate results in future.

VI. CONCLUSION

In this research, the results on the Quora duplicate dataset problem using simplest algorithms available was produced. We are still actively developing our methods and look forward to improving our performance further including complex Machine Learning and Deep Learning Algorithms which can lead to more accurate results.

REFERENCES

[1]. V. Aswini and S. K. Lavanya, "Pattern discovery for text mining," 2014 Int. Conf. Comput. Power, Energy, Inf. Commun., pp. 412–416, 2014.
[2]. L. Ge and T. S. Moh, "Improving text classification with word embedding," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018-January, pp. 1796–1805, 2017.
[3]. Z. Imtiaz, M. Umer, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, "Duplicate Questions Pair Detection Using Siamese MaLSTM," IEEE Access, vol. 8, pp. 21932–21942, 2020.
[4]. D. Li and J. Qian, "Text sentiment analysis based on long short-term memory," 2016 1st IEEE Int. Conf. Comput. Commun. Internet, ICCCI 2016, pp. 471–475, 2016.
[5]. C. Saedi, J. Rodrigues, J. Silva, A. Branco, and V. Maraev, "Learning profiles in duplicate question detection," Proc. - 2017 IEEE Int. Conf. Inf. Reuse Integr. IRI 2017, vol. 2017-January, pp. 544–550, 2017.
[6]. X. Song, X. Wang, and X. Hu, "Semantic pattern mining for text mining," Proc. - 2016 IEEE Int. Conf. Big Data, Big Data 2016, pp. 150–155, 2016.
[7]. https://blogs.rstudio.com/tensorflow/posts/2018-01-09-keras-duplicate-questions-quora/
[8]. [https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning
[9]. https://www.kaggle.com/quora/question-pairs-dataset
[10]. https://www.linkedin.com/pulse/duplicate-quora-question-abhishek-thakur/