Project: Identifying Quora Questions with Similar Intent

Murad Bozik, Brian T. Cook, Yuchi Zhang

Information Retrieval and Text Analytics

Leiden University

21 May, 2020

1 Introduction

Quora is one of the most popular websites on the interest, wherein users can ask questions that often receive answers from other users who purport to be knowledgeable about that particular topic. For example, the question ”What should I expect in a Software Engineer interview at Google and how should I prepare?” received a thoughtful answer from Gayle McDowell, a former Google hiring manager. Popular subjects on the website include (but are not limited to) sports, travel, movies, and technology.

As one can imagine, however, questions are often repeated. This can affect the utility of the site; a user interested in positing a particular question might have several if not hundreds of already question/answer pairs to sift through. With an Alexa rank of 239 as of May 2020 and a recent valuation of $2 Billion USD [1], it is in the best interest of the Quora operators to pursue improvements to the effectiveness of the site. We propose participating in a (closed) Kaggle competition whose aim is to detect duplicate questions on Quora.

2 Description of dataset and problem

In essence, we are trying to identify whether or not two questions posited by Quora users are duplicates. Finding verbatim duplicated (i.e., exact same words and formatting) is a fairly straightforward process; identifying that “Who was the scientist with the big hair?” and “Who invented $E = mc^2$?” are both about Albert Einstein is remarkably more complicated.

The Kaggle dataset[2] we will be using provides testing and training data so that machine learning techniques can be readily applied to this problem. The training set is comprised of actual Quora

[1] https://www.vox.com/recode/2019/5/16/18627157/quora-value-billion-question-answer

[2] https://www.kaggle.com/c/quora-question-pairs/data

questions, while the testing set is computer-generated and thus impervious to using publicly available Quora information to supplement the classifier. Data is stored in form of query IDs, queries stored as strings (e.g., “Who was the scientist with the big hair?” and “Who invented $E = mc^2$?”), and a variable indicating agreement (1 for similar intent, 0 for dissimilar intent). Reasonable people may disagree about the similarity between two queries, so the provided labels are an honest attempt at providing consensus-driven answers to whether or not question pairs are similar in intent.

In order to make this sort of inference between two Quora questions, we must clean the data in the following ways:

1. Convert text to lower case.
2. Unify non-standard expressions in the text, e.g. 1000 g → 1 kg.
3. Stemming, e.g. consulting and consultant → consult.

These are common techniques for standardizing text data and are not unique to our proposed vector-based approach of delineating similar queries from dissimilar ones. Once the data has been cleaned, we can begin applying different classification techniques in order to determine which Quora questions are similar in intent.

3 Methods

3.1 Features Engineering

Co-occurrence, weighted co-occurrence words, special co-occurrence words.

3.2 NLP vectorlize and features

Talk about our proposed vector-formulation (presumably a high-dimensional vector space comprised of common words in the English language). Similarity is $\frac{1}{2}(1 + \cos\theta)$ (so as to be $\in [0, 1]$), where θ is the angle between the query vectors?

4 Results

5 Comparison to other approaches

Can give one paragraph descriptions of these and then compare our vector-based results to these approaches.

5.1 Fuzzy Wuzzy

5.2 XGBOOST

5.3 CNN

5.4 LSTM-RNN

5.5 Adaptive Multi-Attention Network

6 Discussion

7 Conclusion