Analyzing duplicate questions on the Quora Dataset

Category: Natural Language

Team Members: Rohit Prakash Apte (rapte)

Motivation

Sites like Quora and stackoverflow rely on the user community helping each other out by answering questions. One of the main challenges these sites face is having questions with the same intent framed differently and asked multiple times. This can lead to multiple answer threads, inconsistent knowledge on the topic, and make it difficult for product experts to effectively reach a wider audience.

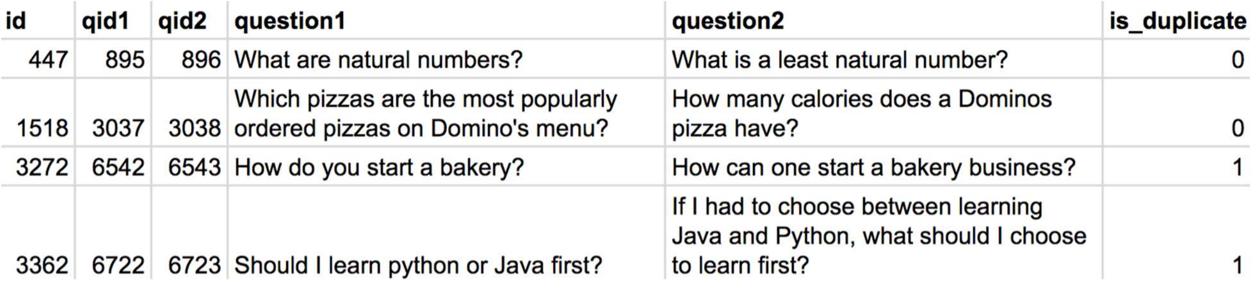
These sites have policies on handling this issue. This is Quora's policy on how questions should be phrased:

- Questions should be written as simply as possible.
- When a question can be phrased in different ways, the best version is the one that is simplest and most commonly asked. We encourage people to edit questions to get them into this state.
- In general, questions should be phrased so that they are unique and reusable. Simple and reusable questions make it easier for Quora to minimize the number of new questions added which are slight variants of existing questions.

Multiple questions can be merged into a common thread. These are mainly handled by the moderators or what are considered the senior users (by the platform's metrics). But this means remembering if a question was asked before (and linking/merging to it) as well as having the relevant people continue to stay engaged in this process.

For my project I want to see if we can use Machine Learning to automatically classify duplicate questions.

Quora has published a dataset of duplicate questions. The dataset consists of 404k question pairs and a label for whether the question was classified as a duplicate or not. There is also a competition hosted by Kaggle for this dataset. This is what the sample data looks like.



id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

Method

There have been some significant breakthroughs in Natural Language Processing in recent years. Recurrent Neural Networks, character level CNNs and attention mechanisms have showed improvements in Machine Translation, Question Answering and other tasks. I will see if we can apply these methods to our problem.

I will create a baseline using a simple classifier like Logistic Regression or Random Forest and then apply more sophisticated algorithms like LSTMs, CNNs and attention to see if we improve our accuracy.

Experiments

This is a binary classification problem. I will vectorize each question into words using different techniques (bag of words, word2vec, glove, etc.) and then train the model on that dataset to predict whether the questions are duplicates or not.

I will look at different approaches to training the model (using Logistic Regression, LSTMs and CNNs) and see if stacking models can give better results.

Since this is an old Kaggle competition, we can test our results on the leaderboard against a hidden test set.