Quora Question Pairs

Problem

Given the question pairs, classify whether question pairs have same intent or not

Importance of Problem

Statistics

- 190 million monthly users
- 100 million unique monthly users
- Around 4,00,000 different topics of discussion

Quora is one of the largest knowledge sharing platform when users ask query and other users answers them on one common platform. With millions of user active there is high chance that questions asked are of same intent and repeated unintentionally.

We require a good classifier for these questions of same intent so they don’t appear repeatedly and user answer them only once rather than answers same question many times. This way each question will have best answer to it.

Clustering can be used to address this problem and users get the most appropriate answers to their questions from hundreds of users willing to answer. Hence improve the experience overall.

Input and desired output

In our model Input is a training data comprising of question pairs with assigned values from {0,1}, corresponding to each question pairs. Value is 1 if it is duplicate and 0 if it is not a duplicate.

We are also provided with the test data of question pairs. So the desired output is to label these pairs efficiently whether they have the same intent or not.

What is challenging about it

One challenging thing about this problem was that the questions included lowercase and unchanged, punctuation replaced in different ways, stop words included and excluded, stemmed and not stemmed, etc. -- and to build features from all of these different representations.

What are the existing solutions

Top rankers used NLP, nn and graphical methods to extract features.

Most of these solutions used models like lightGBM, XGB, LSTM,LGB, LR.

Our solution

1. Preprocessing the training data
    a. Text- cleaning
    b. Word-stemming
    c. Features extraction
    d. Word share
2. Getting better insights
3. Training and building model
4. Train-Validate-Test-Split
5. Training using XGBoost

Text Processing

Tokenizing

Breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens

A **RegexpTokenizer** splits a string into substrings using a regular expression.

Lemmatizing

Grouping together the inflected forms of a word so they can be analysed as a single item

Stopword Filtering

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

Doc2Vec

An unsupervised algorithm to generate vectors for sentence/paragraphs/documents

Stemming (with/without stopword filtering)

A sort of normalizing method

Many variations of words carry the same meaning, other than when tense is involved.

Key learning point

This problem helped us to learn about building features and models that can be used to extract information from English texts.