Quora Question Pair Similarity

Bindiya Roy Sumanth Balabhadruni Harsha Bezawada

13th August, 2021

1 Introduction

The Quora dataset is composed of pairs of questions, and the task is to determine if the two questions are duplicates of each other, that is, that they have the same meaning.
Quora is a question answering website where users ask questions and other users respond and the best answers are up-voted. Duplicate questions on this site poses an issue because, if treated independently, these may prevent a user from seeing a high quality response that already exists and responders are unlikely to answer the same question twice.
Our task requires Natural Language Processing (NLP), whose task is to understand human language. The challenging part of the problem is to represent sentences as numerical inputs so the learning algorithms can work on it. Our aim is to use some statistical methods such as Bag-of-Words, Tf-Idf vectorizer etc. to get a perfect vocabulary, then to use various ML models and observe their performance on the dataset.

2 DATA

2.1 Exploration

The Quora Question Pairs dataset consists of a 1,00,000-question pair. So, we are having only the Training Set and we don't have the Test Set. We will split the training set into Training Set, Validation and Test Set.

Each Sample Point is comprising of
- id: unique ID of each pair
- qid1: ID of first question
- qid2: ID of second question
- question1: text of first question
- question2: text of second question
- is duplicate: are the questions duplicates of each other (0 indicates not duplicate, 1 indicates duplicate).

Of the 1,00,000 question pairs, 62879 (62.879 %) have a negative (0) label and 37121 (37.121 %) have a positive (1) label.

2.2 Text Preprocessing

Here we will be pre-processing the data in order to make the learning process smoother.

Step-1: Removing all the Special characters
• These characters are most often found in comments, references, currency numbers etc.
• These characters add no value to text-understanding and induce noise into algorithms.

Step-2: Convert the data into standard form either into lower case or uppercase
• Because if we won't change it into standard form then the Bag Of Words Increases and the Dimensionality also increases and it will lead us to decline in the accuracy
• And there is no change in the meaning if we change it to a Standard Form

Step-3: Removing the Stop Words
• Stops words are the words which are used for sentence formation
• If we keep the Stop Words then again, the size of document term matrix which leads to decrease in the Accuracy
• After Removing the stop words, the overall meaning of the resulting sentence is remains same (like 1 OR 0)

Step-4: Stemming or Lemmatization
• It is used to convert the words into root words.

Step-5: Learn the vocabulary from the Corpus
• Each question in the dataset is called as Document and set of all the documents in the dataset is known as Corpus.

Step-6: Convert Corpus to Document Term Matrix
• A corpus of documents and a dictionary of terms contain all the words that appear in the documents. The term-document matrix then is a two-dimensional matrix whose rows are the terms and columns are the documents, so each entry (i, j) represents the frequency of term i in document j.

This Bag of words won't preserve the order. So, in order to avoid that we use n-grams.

3 Model Building

To get the cleaned, tokenized, vectorized and pre-processed data, we need to make a few different Statistical data models during data preparation. When we have our data ready, then we can go ahead and make some classification. As this is a labelled data and a binary classification problem, we would like to go with the Linear models - Logistic Regression, Linear SVM and the Tree based Model - Decision Tree, Random Forest Classifier.
Quora currently uses a random forest with hand engineered features on this problem. For this reason we are interested in applying these tree based models to understand how they performed and how they differed from linear models. As we have a large dataset,

- Logistic Regression works great with large dimensions but SVM will not be a good approach in this case as it will take a long training time for that and also choosing a good 'kernel' is not so easy.

- A Decision tree model is very intuitive and required less effort for data preparation but we can end up with data overfitting. On the other hand, Random Forest is a very good approach for classification; it reduces the overfitting in Decision Trees and helps to improve the accuracy.

Hence among the Tree Based Models, Random Forest Classifier and among the Linear Models, Logistic Regression will be the better choices. But still will perform all these algorithms to observe all the results and to explore more.

3.1 Hyper Parameter

Now whatever algorithm we use, our major task is the Hyper Parameter Tuning to get the optimal hyper parameters for having a better model. **Logistic Regression**: L1 or L2 regularization will be used to reduce the overfitting – underfiiting problem, controlled by *alpha*. **Decision Tree**: max depth, num of leaf nodes and **Random Forest**: max depth, num of leaf nodes, num of estimators are the hyper parameters needs to be tuned.

In each approach, we must need to analyze the errors in train & test data; predict on the test data; then evaluate the model using evaluation metrics for classification such as Accuracy, Confusion Matrix, Precision & Recall, F1-score, ROC-AUC, Log-loss to see which Model performs best.