Detecting Duplicate Questions on Quora

Ali Bahrami

January 07, 2022

Agenda

1. Introduction
2. Project Overview
3. Exploratory Data Analysis
4. Preprocessing and Feature Extraction
5. Models & Evaluation
6. Going Forward

Introduction

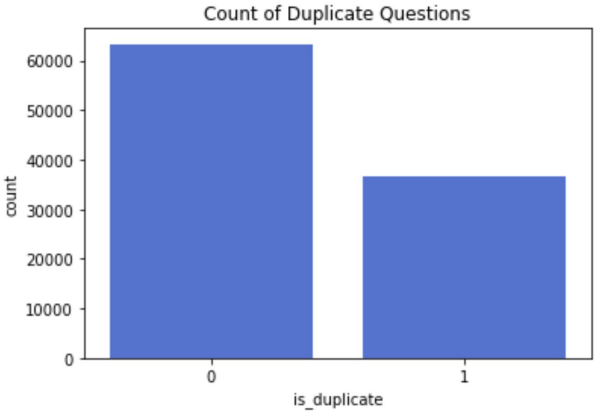
- The popularity of knowledge-sharing platforms, like Quora, Stack Overflow and Reddit, has exploded in the recent years
- With so many questions asked everyday, it would be ideal to suggest previously asked questions to user instead of them having to ask a new question
- This would reduce wait time for the user and help them see more relevant content instead of seeing the same question worded differently
- It would also reduce the storage and processing burden on the system
- However, it is difficult to find similar questions because search engines have limitations
- Enter: Natural language processing

Project Overview

- In this project, the task is to compare two question pairs and determine if they are duplicates or not
- As simple as that, or so I thought …

EDA

- The dataset has 404,290 question pairs provided by Quora
- Data distribution:
  - Non-duplicate: 63%
  - Duplicates: 37%
- For this task, I have taken a sample of 100,000 data points
- Data Sample:



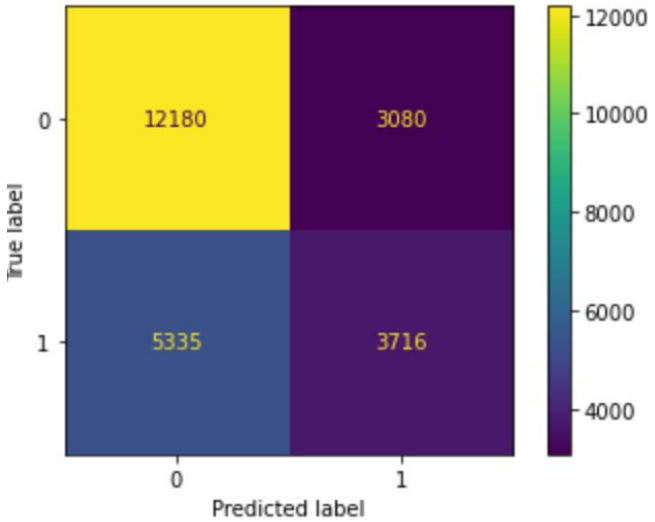
	id	qid1	qid2	question1	question2	is_duplicate
342588	342588	470580	470581	How do I backup exchange contacts from an iPho...	I have many important contacts on my iPhone an...	0
122868	122868	25284	198774	Did Einstein say "Everybody is a genius but if...	Did Einstein consider himself a genius?	0
117273	117273	190842	190843	What are the natural remedies for damage skin?	What are some natural remedies for swelling?	0
110272	110272	114131	14381	How too fall out of love?	Is it normal to fall out of love?	1
146109	146109	1772	6614	How can you increase your height?	How should I increase my height?	1

Preprocessing & Feature Extraction

- **Data Cleaning:**
  - Deal with missing values
  - Tokenize
  - Remove stopwords
  - Remove punctuations
  - Normalized case
  - Stem the words
- **Feature Extraction:**
  - **Basic Features:**
    - Length of each question and the difference between the lengths
    - Number of characters(without spaces) in each question
    - Number of words in each question
    - Number of common words between the question pairs
      Word2Vec: pre-trained from gensim:
  - **Word2Vec:**
    - **Word Mover's Distance** (*wmdistance*): a measure of dissimilarity between two text documents, as the minimum amount of distance that the embedded words of one document need to "travel" to reach the embedded words of another document.

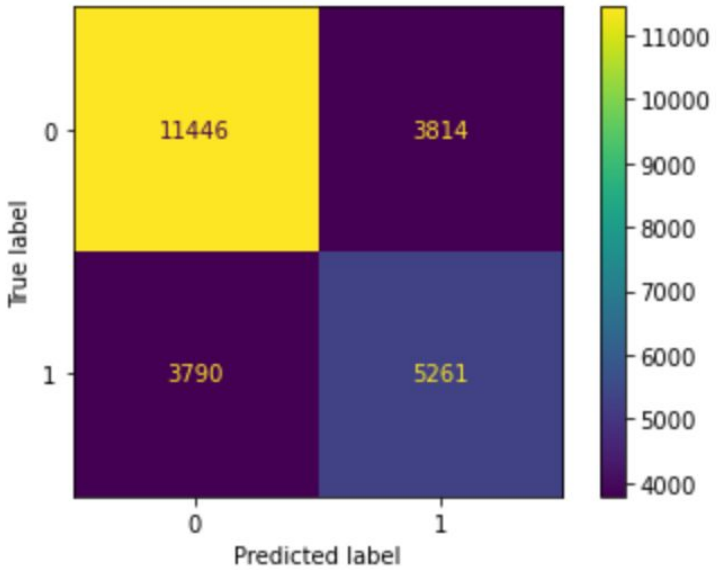
Logistic Regression

	precision	recall	f1-score	support
0	0.70	0.80	0.74	15260
1	0.55	0.41	0.47	9051
accuracy			0.65	24311
macro avg	0.62	0.60	0.61	24311
weighted avg	0.64	0.65	0.64	24311



Random Forest Classifier

	precision	recall	f1-score	support
0	0.75	0.75	0.75	15260
1	0.58	0.58	0.58	9051
accuracy			0.69	24311
macro avg	0.67	0.67	0.67	24311
weighted avg	0.69	0.69	0.69	24311



Note: Both models perform relatively well (~75% of the times) at predicting non-duplicates, but performs poorly on identifying duplicates with 58% accuracy.

Going forward

- Extract more features using other similarity techniques
- Try different feature sets
- Model with other classifier and ensemble models, such as XGBoost
- Deep learning

Thank you!