Duplicate Question-Pair Classification In Quora Questions Dataset

ZAHIDUZZAMAN BISWAS
INSTRUCTOR: JAN ZIKES

GOAL

Train a model that takes two questions as input and classify them either as

- 1 (duplicate) or
- 0 (not duplicate

Mathematical representation :

f(question1, question2) = 1 or 0

WHY IS IT NEEDED?

For Quora

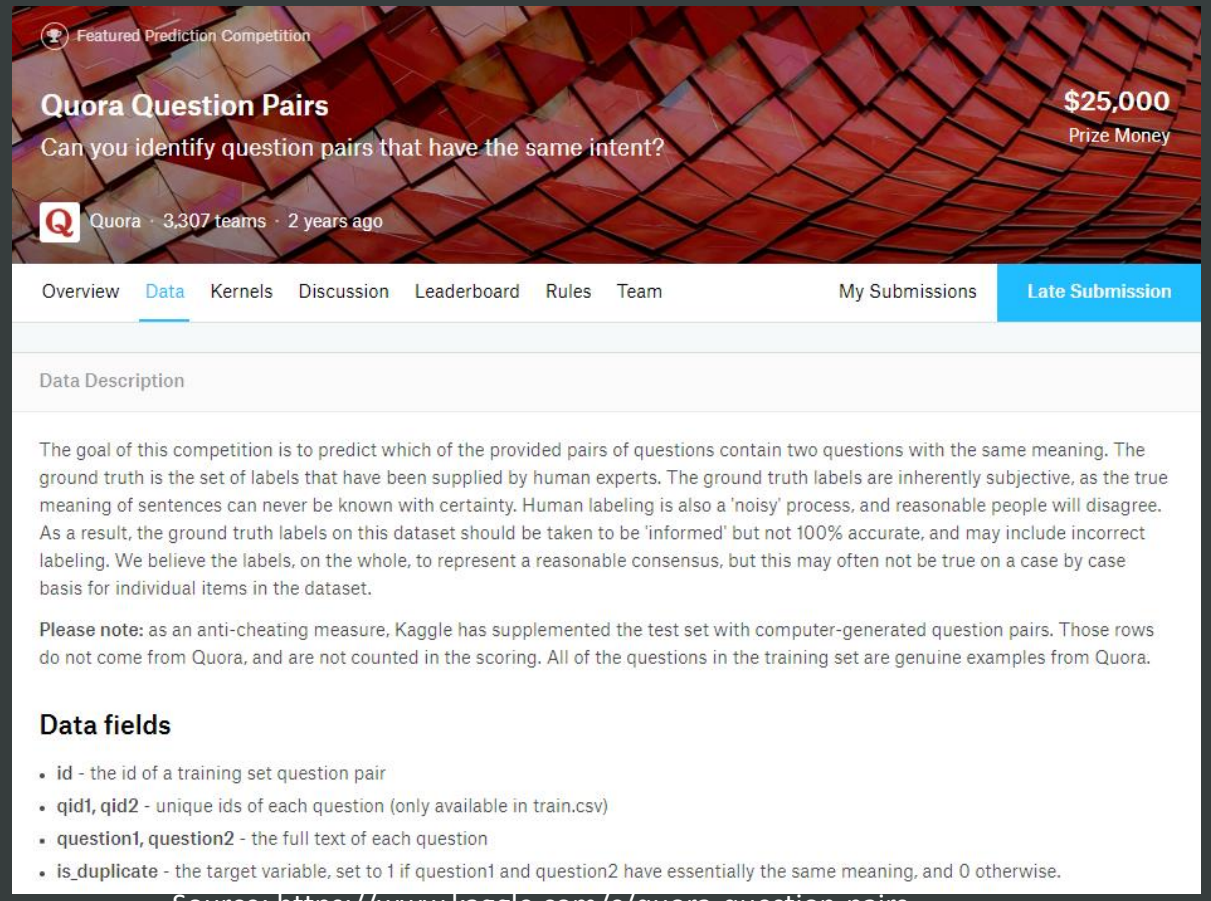
- To remove duplicates and organize answers in one thread
- To provide best user experience
- To reduce data storage usage

For Quora Users

- To avoid the maze of answers
- To avoid asking duplicate questions
- To be more efficient

Motivation

- First official release on January 24, 2017
- http://qim.fs.quoracdn.net/quora_duplicate_questions.tsv
- Competition was hosted on Kaggle.com



Featured Prediction Competition

Quora Question Pairs

Can you identify question pairs that have the same intent?

$25,000
Prize Money

Quora · 3,307 teams · 2 years ago

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | My Submissions | **Late Submission**

Data Description

The goal of this competition is to predict which of the provided pairs of questions contain two questions with the same meaning. The ground truth is the set of labels that have been supplied by human experts. The ground truth labels are inherently subjective, as the true meaning of sentences can never be known with certainty. Human labeling is also a 'noisy' process, and reasonable people will disagree. As a result, the ground truth labels on this dataset should be taken to be 'informed' but not 100% accurate, and may include incorrect labeling. We believe the labels, on the whole, to represent a reasonable consensus, but this may often not be true on a case by case basis for individual items in the dataset.

Please note: as an anti-cheating measure, Kaggle has supplemented the test set with computer-generated question pairs. Those rows do not come from Quora, and are not counted in the scoring. All of the questions in the training set are genuine examples from Quora.

Data fields

- **id** - the id of a training set question pair
- **qid1, qid2** - unique ids of each question (only available in train.csv)
- **question1, question2** - the full text of each question
- **is_duplicate** - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

Source: https://www.kaggle.com/c/quora-question-pairs

DATA EXPLORATION

Training Data: train.csv Rows: 404290 Column: 6 Memory: 64MB

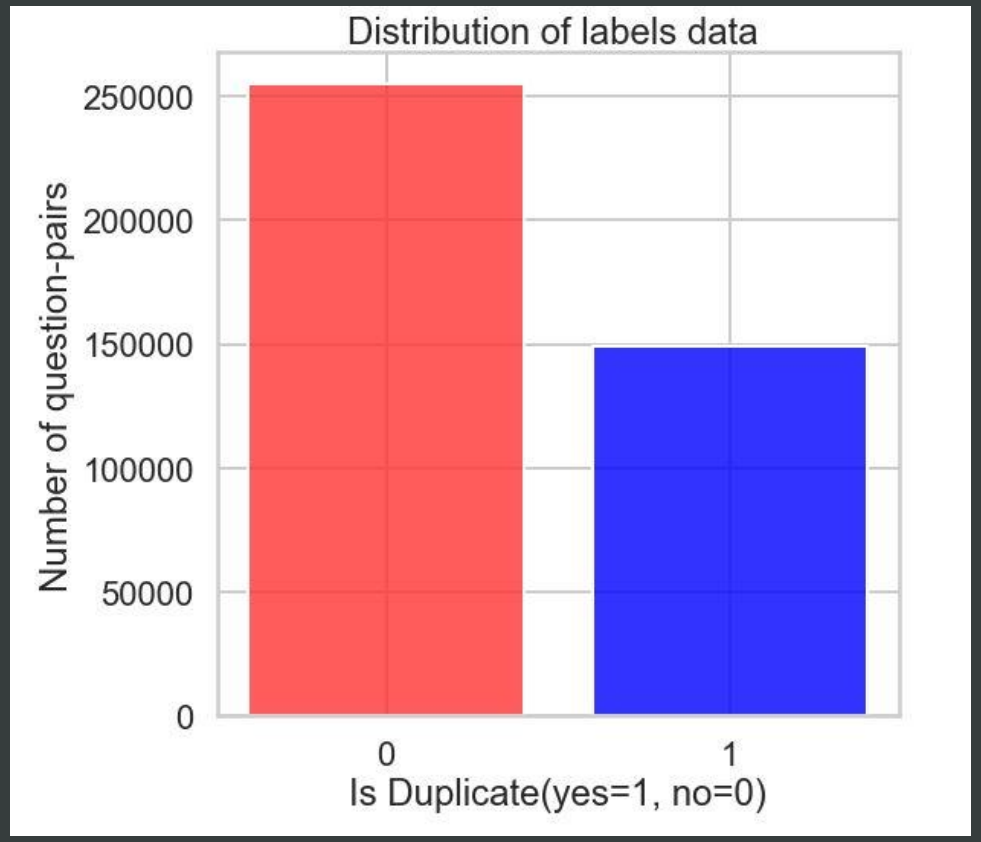
id	qid1	qid2	question1	question2	Is_duplicate
0	1	2	What is the step by step guide…	What is the step by step guide to invest	0
1	3	4	What is the story of (Koh-i-Noor) Diam…	What would happen if Indian Government..	0

Test Data: test.csv Rows: ~2.3 Million (3.5Million+) Column: 3 Memory: 314MB

test_id	question1	question2
0	How does the Surface Pro himself 4 compare wit..	Why did Microsoft choose core m3 and not core ..
1	Should I have a hair transplant at age 24? How...	How much cost does hair transplant require?

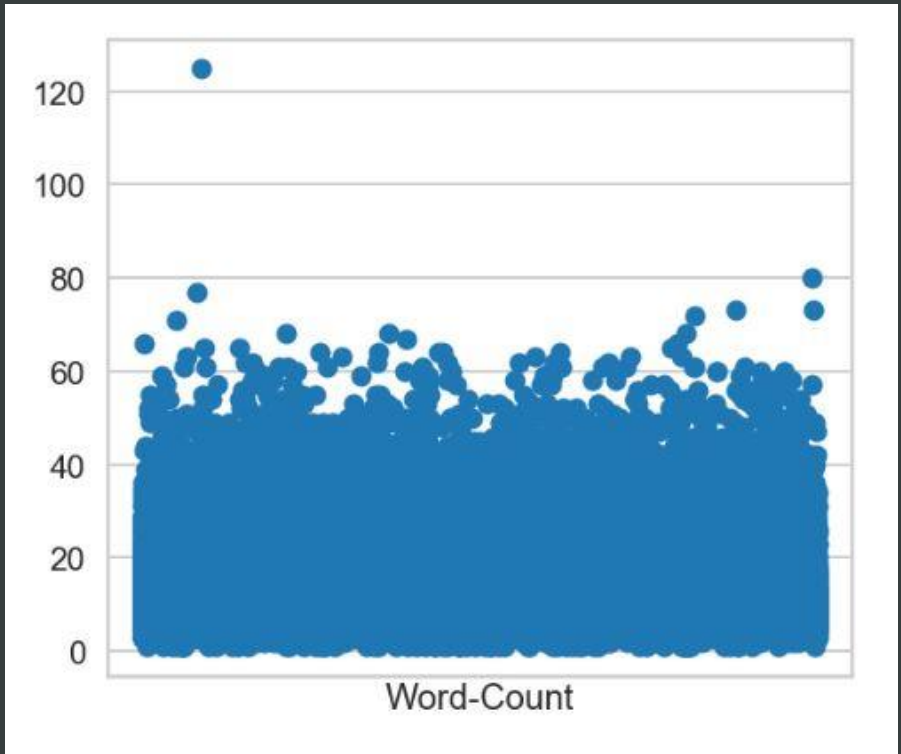
TRAINING DATA EXPLORATION

- Missing Values : 3
- Duplicate questions pair : 40%
- Real questions from Quora
- Labels are not 100% accurate
- Note: Test data is computer Generated



EDA of ‘question1’ Column

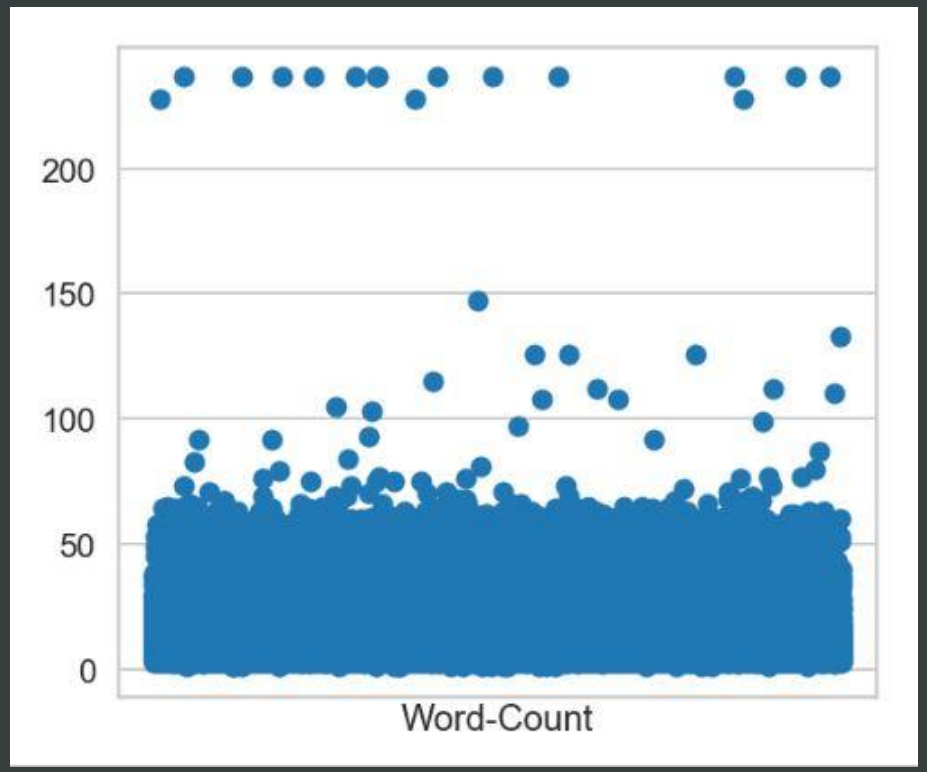
MAX WORD COUNT : 125



MIN WORD COUNT : 1

EDA of ‘question2’ Column

MAX WORD COUNT : 237



MIN WORD COUNT : 1

Insights

- Most of the questions are about India, US politics, economics etc.
- Questions are asked from different perspective
- Intent of the questions are subjective
- Is it really possible accurately identify the intent of a question?
- Challenge: can our model predict the questions with same intent?

What is the step by step guidance to invest in share market?
What is the step by step guidance to invest in share market in India?

How is the new Harry Potter book 'Harry Potter an the Cursed Child'?
How bad is the new book by J. k Rowling?

Our Approach

Step 1 Text Preprocessing

- Remove Clutters
- Stem and corrections

Step 2 Vectorizations

- Tf-idf
- doc2vec

Step 3 Feature Selections and Modeling

- Add features (word-count, character-count, cosine distance, Euclidean distance, doc2vec)
- Naïve Bayes

Add a Slide Title - 2

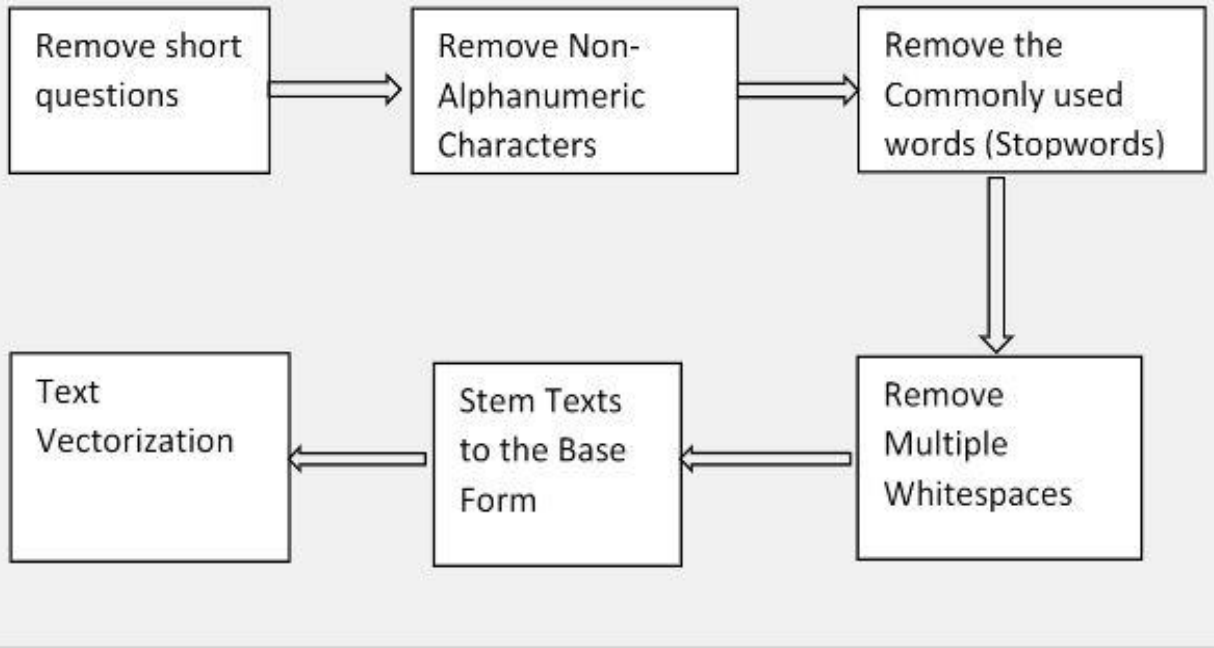


Figure 5: Flow chart of the overall steps to the text data processing

Naïve Bayes Algorithm

- Features are independent
- Order of word does not matter

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Evaluation Metrics

- Log loss
- Prediction Probability

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)]$$

Experiments and Results

Model No.	Model Specs.	Parameters	Public Score	Private Score
3	- Tf-idf vectorization - Doc2vec Trained on Questions Corpus - Multinomial Naïve Bayes	MultinomilaNB : α = 0.1 TfidfVectorizer: min_df = 1 , ngram = (1,2) Doc2Vec: Window = 1, epochs = 1, min_count = 1, vector_size = 5	0.398	0.399
4	- Tf-idf vectorization - Doc2vec Trained on Questions Corpus - Multinomial Naïve Bayes	MultinomilaNB : α = 0.1 TfidfVectorizer: min_df = 1 , ngram = (1,2) Doc2Vec: Window = 7, epochs = 10, min_count = 2, vector_size =100	0.390	0.398

Experiments and Results (cont.)

Model No.	Model Specs.	Parameters	Public Score	Private Score
5	- Tf-idf vectorization - Doc2vec Trained on Wikipedia Corpus - Multinomial Naïve Bayes	MultinomilaNB : α = 0.1 TfidfVectorizer: min_df = 1 , ngram = (1,2) Doc2Vec: epochs = 50, vector_size =300	0.390	0.398

Classifier	Training Time	Public Score	Private Score
Naïve Bayes	2 hours 13 minutes	0.560	0.561
Logistic Regression	Approximately 30 minutes	0.542	0.542
SVM	Over 8 hours (On 50% of training data)	0.550	0.552

Recommendations For Future Experiments

- Correct misspelled words, replace abbreviation with the complete words
- Add special characters count, word share ratio, capitalization count as features
- Besides cosine and Euclidean distances, other distance features can be added
- Deep learning implementation such as LSTM

References

- [1] Lili Jiang, Shuo Chang, and Nikhil Dandekar. Engineering at Quora: Semantic Question Matching with Deep Learning. https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning. 2017
- [2] Budanitsky and G. Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics, 32(1):13–47.
- [3] Wang, Zhiguo, Wael Hamza, and Radu Florian. "Bilateral Multi-Perspective Matching for Natural Language Sentences." arXiv preprint arXiv:1702.03814 (2017).
- [4] Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 5528–5531. IEEE, 2011.
- [5] Tim Rocktaschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, Phil Blunsom. Reasoning about entailment with neural attention. In ICLR 2016.
- [6] Tomas Mikolov, Iliya Sutskever, Kai Chen, Greg Corrado, and Jeffery Dan. Distributed Representation of Words and Phrases and . In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 5528–5531. IEEE, 2011.