Quora Duplicate Question Pair Project

Name: Bhavika Balani
Roll no: OPD202227

Purpose

- Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions.

- Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question.

- The goal of this project is to create a model that allows users to input two questions and get predictions about whether they are duplicate or not.

Data Overview

	id	qid1	qid2	question1	question2	is_duplicate
398782	398782	496695	532029	What is the best marketing automation tool for...	What is the best marketing automation tool for...	1
115086	115086	187729	187730	I am poor but I want to invest. What should I do?	I am quite poor and I want to be very rich. Wh...	0
327711	327711	454161	454162	I am from India and live abroad. I met a guy f...	T.I.E.T to Thapar University to Thapar Univers...	0
367788	367788	498109	491396	Why do so many people in the U.S. hate the sou...	My boyfriend doesnt feel guilty when he hurts ...	0
151235	151235	237843	50930	Consequences of Bhopal gas tragedy?	What was the reason behind the Bhopal gas trag...	0

id - the id of a training set question pair

qid1, qid2 - unique ids of each question

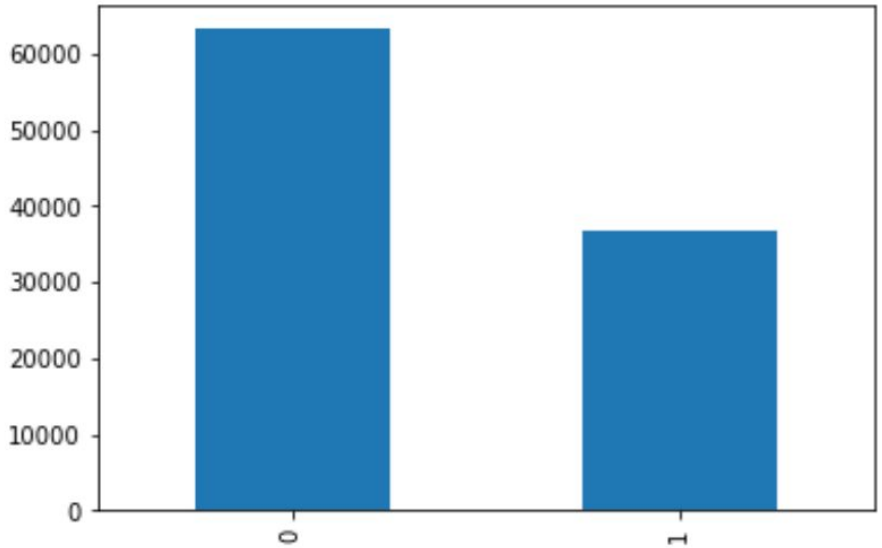
question1, question2 - the full text of each question

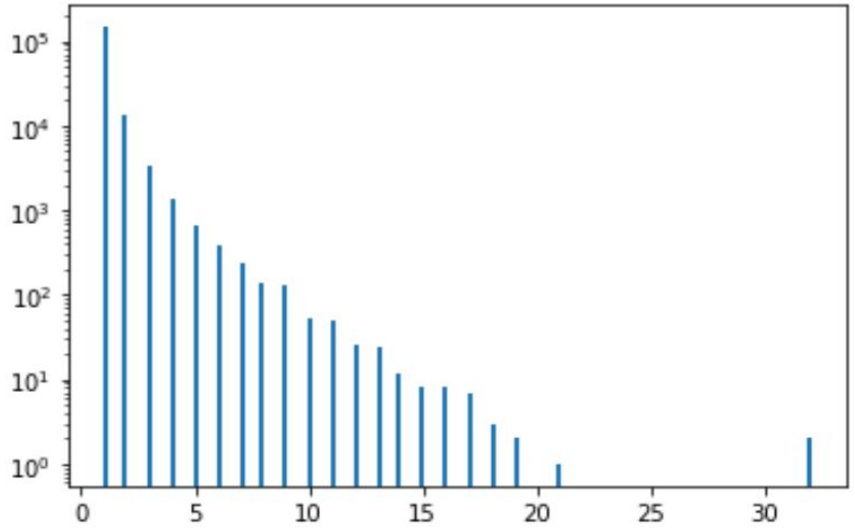
is_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

Type of Problem

It is a binary classification problem, for a given pair of questions we need to predict if they are duplicate or not.

EDA


60000
50000
40000
30000
20000
10000
0
0
1


10^5
10^4
10^3
10^2
10^1
10^0
0
5
10
15
20
25
30

Data Preprocessing

- Check for null values
- Replacing special characters
- Decontracting words

Basic Feature Engineering

- q1_len: The length of the first question in terms of the number of words.
- q2_len: The length of the second question in terms of the number of words.
- q1_words: A list of words in the first question after tokenization.
- q2_words: A list of words in the second question after tokenization.
- words_common: The number of words that are common between the two questions.
- words_total: The total number of unique words in both questions combined.
- word_share: The ratio of words_common to words_total, indicating the extent of word overlap between the questions.

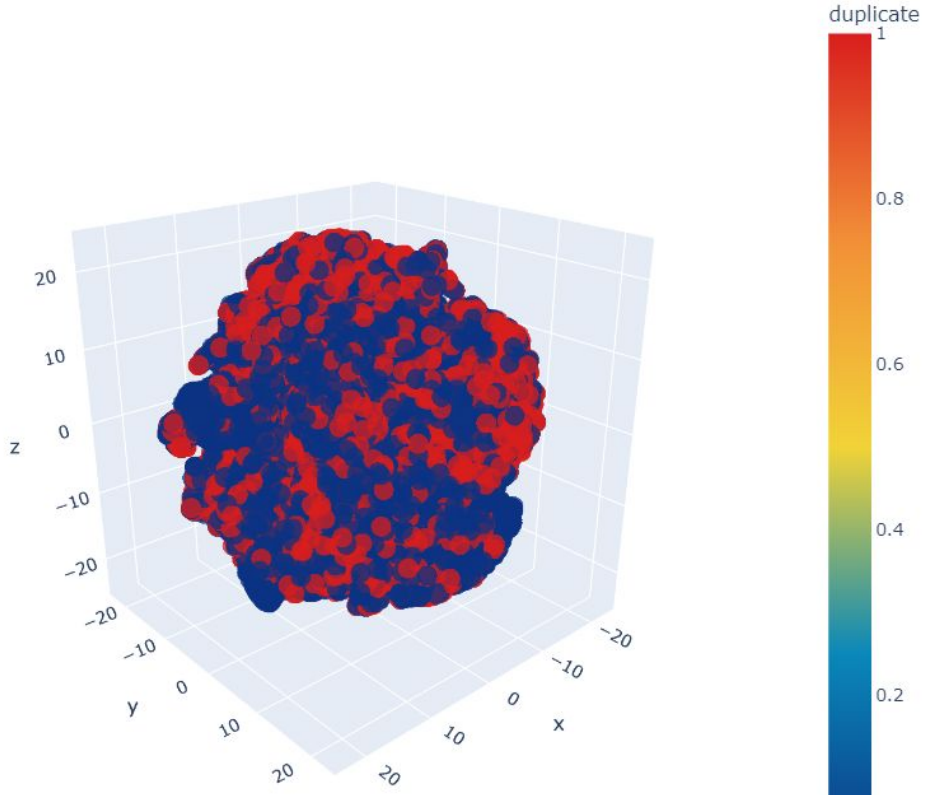
Advanced Feature Engineering

1. **Token Features:**
- cwc_min: Ratio of common words to the length of the smaller question.
- cwc_max: Ratio of common words to the length of the larger question.
- csc_min: Ratio of common stop words to the smaller stop word count among the two questions.
- csc_max: Ratio of common stop words to the larger stop word count among the two questions.
- ctc_min: Ratio of common tokens to the smaller token count among the two questions.
- ctc_max: Ratio of common tokens to the larger token count among the two questions.
- last_word_eq: 1 if the last word in the two questions is the same, 0 otherwise.
- first_word_eq: 1 if the first word in the two questions is the same, 0 otherwise.

2. **Length Based Features:**
- mean_len: Mean length of the two questions (number of words).
- abs_len_diff: Absolute difference between the length of the two questions (number of words).
- longest_substr_ratio: Ratio of the length of the longest common substring among the two questions to the length of the smaller question.

3. **Fuzzy Features:**
- fuzz_ratio: fuzz_ratio score from fuzzywuzzy.
- fuzz_partial_ratio: fuzz_partial_ratio from fuzzywuzzy.
- token_sort_ratio: token_sort_ratio from fuzzywuzzy.
- token_set_ratio: token_set_ratio from fuzzywuzzy.

Dimensionality Reduction Using t-SNE



TF-IDF Features

Machine Learning Algorithms

Random Forest Classifier

In [52]:
```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
rf = RandomForestClassifier()
rf.fit(X_train,y_train)
y_pred = rf.predict(X_test)
accuracy_score(y_test,y_pred)
```

Out[52]:
```
0.8026
```

```
# for random forest model
confusion_matrix(y_test,y_pred)
```

```
array([[10704,  1868],
       [ 2080,  5348]], dtype=int64)
```

Machine Learning Algorithms

XGBoost Classifier

```
from xgboost import XGBClassifier
xgb = XGBClassifier()
xgb.fit(X_train,y_train)
y_pred1 = xgb.predict(X_test)
accuracy_score(y_test,y_pred1)
```

3]: 0.8004

```
# for xgboost model
confusion_matrix(y_test,y_pred1)
```

```
array([[10622,  1950],
       [ 2042,  5386]], dtype=int64)
```

Prediction

In [68]:
```
q1 = 'Where is the capital of India?'
q2 = 'What is the current capital of India?'
q3 = 'What is the current capital of Pakistan?'

query_features = query_point_creator(q1, q2)
print(query_features)
```

```
[[29. 36.  6. ...  0.  0.  0.]]
```

In [69]:
```
rf.predict(query_point_creator(q1,q2))
```

Out[69]:
```
array([1], dtype=int64)
```

In [70]:
```
rf.predict(query_point_creator(q1,q3))
```

Out[70]:
```
array([0], dtype=int64)
```

Prediction

Duplicate Question Pairs

Enter question 1

What is the idea behind democracy?

Enter question 2

What is the core idea behind democracy?

Find

Duplicate

Prediction


Duplicate Question Pairs
Enter question 1
What is the idea behind democracy?
Enter question 2
What is the core idea behind socialism?
Find
Not Duplicate