**Duplicate Question Pairs**

***Group Members:***

***Abhijit Shaw***

***Asansol Engineering College***

***10800220081***

***Amit Kumar Singh***

***Asansol Engineering College***

***10800220078***

***Nisha Jha***

***Asansol Engineering College***

***10800220070***

***J Srikanth Chetty***

***Asansol Engineering College***

***10800221141***

***Agnishwar Mukherjee***

***Asansol Engineering College***

***10800220043***

# Table of Contents

1. **Acknowledgement**
2. **Project Objective**
3. **Project Scope**
4. **Data Description**
5. **Data cleaning**
6. **Model Building**
7. **Code**
8. **Future Scope of Improvements**
9. **Project Certificate**

# Acknowledgement

I take this opportunity to express my profound gratitude and deep regards to my faculty **Prof. Arnab Chakraborty** for his exemplary guidance, monitoring, and constant encouragement throughout the course of this project. The blessing, help and guidance given by him/her time to time shall carry me a long way in the journey of life on which I am about to embark.

I am obliged to my project team members for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment.

***Abhijit Shaw***

***Amit Kumar Singh***

***Nisha Jha***

***J Srikanth Chetty***

***Agnishwar Mukherjee***

Project Objective

Duplicate question pairs refer to pairs of questions that are very similar in their meaning or intent, often differing only in their wording or phrasing. These pairs may arise when multiple users ask the same or similar questions in online forums, search engines, or question-answering systems. Identifying and managing duplicate question pairs can be important in improving the quality and efficiency of such systems, as it allows users to find relevant information more easily and avoids duplication of effort in answering the same questions repeatedly.

The objective for this project is to develop a machine learning model that can accurately identify whether a pair of questions are duplicates or not. This task can be important in various applications such as question-answering systems, search engines, and online forums where users may ask similar questions repeatedly.

To achieve this objective, we are takin the following steps:

1. Collect a large dataset of question pairs, with labels indicating whether each pair is a duplicate or not.
2. Preprocess the data by cleaning and normalizing the questions and transforming them into a format that can be fed into the machine learning model.
3. Extract relevant features from the questions, such as word embeddings or other semantic representations, to capture their meaning and similarity.
4. Train a machine learning model on the labeled data using appropriate algorithms such as logistic regression, decision trees, etc.
5. Evaluate the model's performance on a held-out test set, using metrics such as accuracy and precision.
6. Tune the model's hyperparameters and experiment with different feature representations and model architectures to improve its performance.
7. Deploy the model in a real-world application, such as an API or web service, that can take in a pair of questions and output a prediction of whether they are duplicates or not.

Overall, the objective of this project is to develop a robust and accurate machine learning model that can automatically identify duplicate question pairs and improve the efficiency and effectiveness of various applications that rely on question-answering and search functionality.

# Project Scope

The broad scope of this project includes a variety of applications and techniques that can be used to improve natural language processing (NLP) and information retrieval.

Some of these are given below:

* This project can be used in Q&A forums and online communities, where identifying and removing duplicate questions can improve the user experience and reduce the workload of moderators and administrators. ML models can be trained to identify duplicate questions based on the features of the questions, such as the words used, the length of the question, and the structure of the sentence.
* This project can also be used in search engines, where ML models can be used to rank search results based on their similarity to the user's query. This can improve the relevance and accuracy of search results, making it easier for users to find the information they need.
* This project can also be used for text classification, where ML models can classify text based on the topic or intent of the text. For example, an ML model could be trained to classify customer support tickets based on the issue the customer is experiencing.
* In addition, this project can be used for recommendation systems, where ML models can recommend products, services, or content based on the similarity of the user's preferences to other users or items in the dataset.

The techniques and methods developed in this project can contribute to the broader field of NLP and information retrieval, improving the accuracy and efficiency of various applications and systems.

Data Description

Source of data:

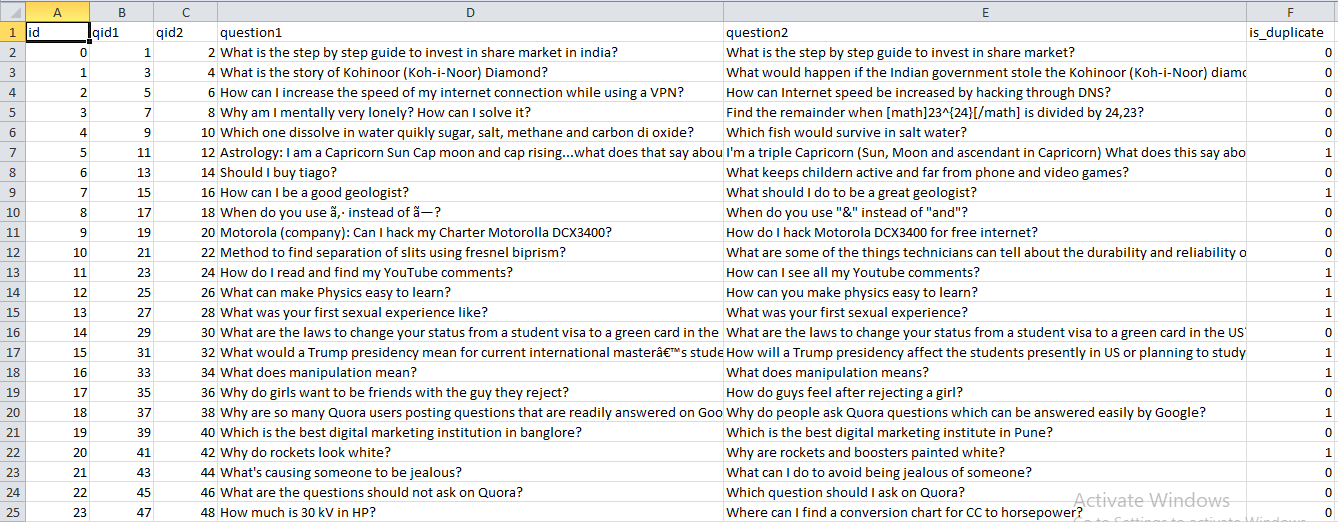
The data used in this project has been taken from Kaggle. The data used was originally posted by Quora on Kaggle.

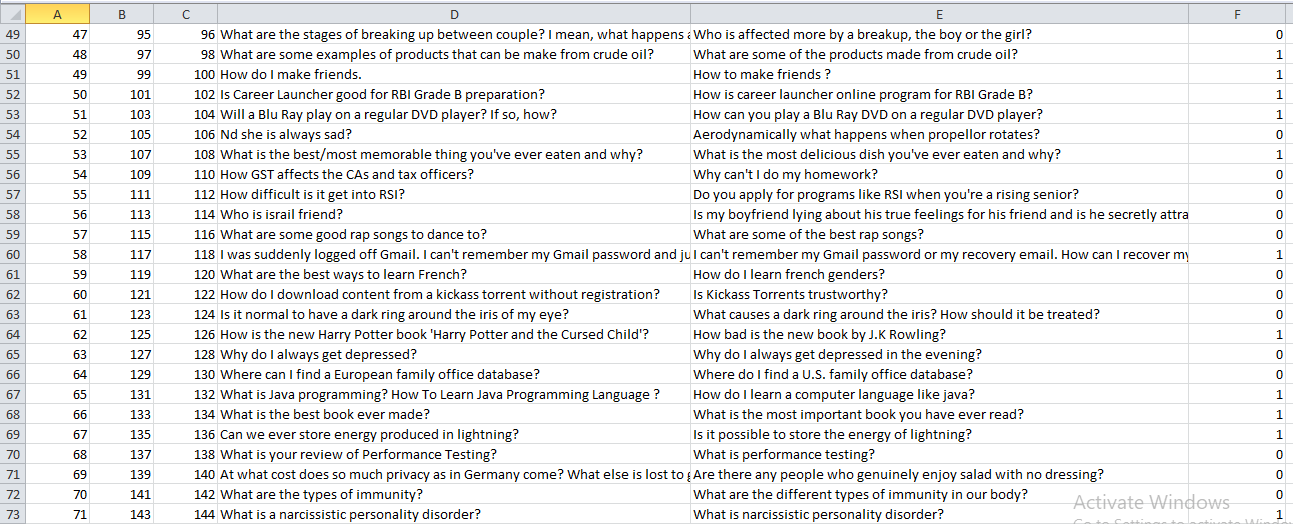
Data Description:

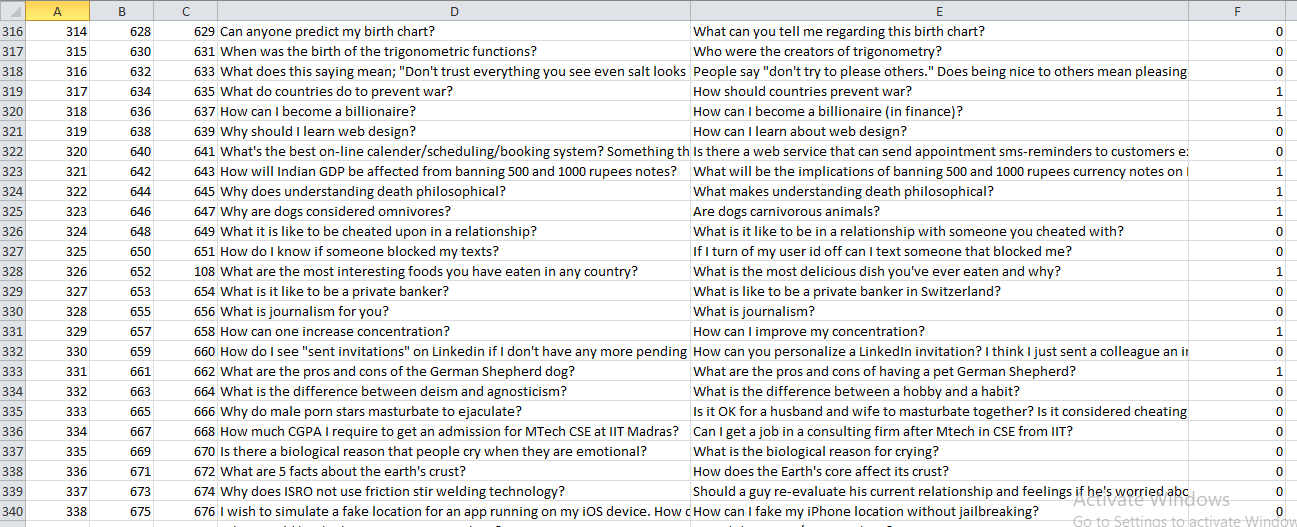
* Total no. of rows in training data is **404290**
* Total no. of columns is **6**
* Available Columns:
  + **id**: A unique id for the question pair
  + **qid1**: id of the first question.
  + **qid2**: id of the second question
  + **question1**: the first question
  + **question2**: second question
  + **is\_duplicate**: Whether both are duplicate or not.
* No. of non-duplicate data points is **255027**
* No. of duplicate data points is **149263**
* There are 3 rows that have missing values (in question1 and question2).
* **36.92%** of question pairs are duplicates and **63.08%** of questions pair non-duplicate.
* Out of **808574** total questions (including both question1 and question2), **537929** are unique.
* Most of the questions are repeated very few times. Only a few of them are repeated multiple times.
* One question is repeated **157** times which is the max number of repetitions.

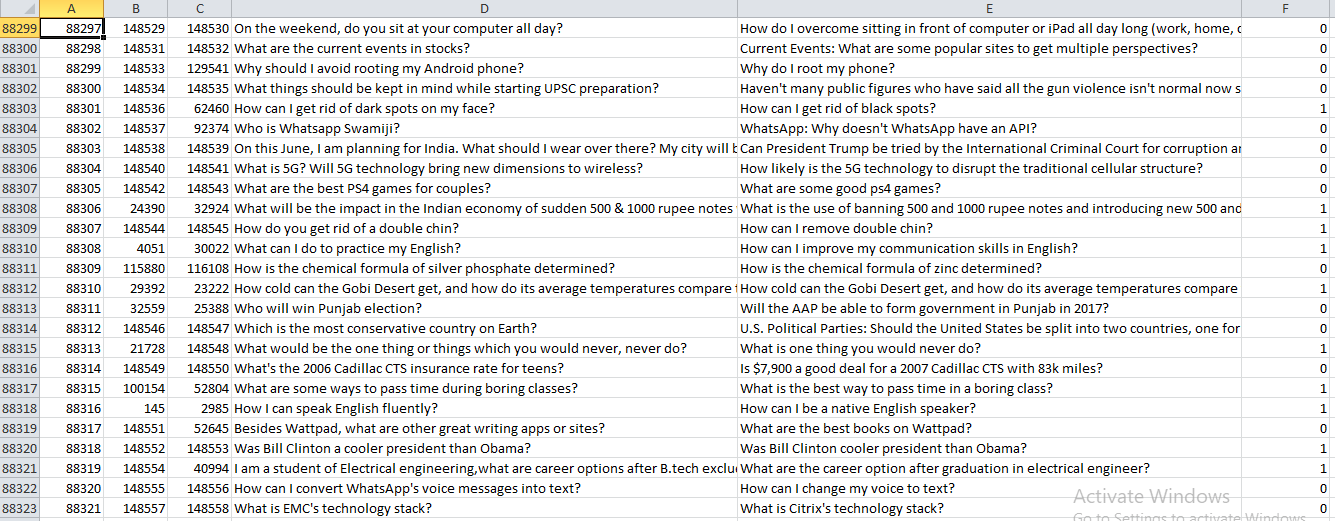
There are some questions with very few characters, which do not make sense. It will be taken care of later with Data Cleaning.

Data Overview:









**Data cleaning**

* We have converted everything to lower case.
* We have removed contractions.
* We have replaced currency symbols with currency names.
* We have removed non-alphanumeric characters.
* We have removed inflections with the word lemmatize.
* We have also removed HTML tags

# Feature Extraction

We have created **23** features from the questions.

**Token features:**

* id: unique identifier for the question pair (unused)
* qid1: unique identifier for the first question (unused)
* qid2: unique identifier for the second question (unused)
* question1: full unicode text of the first question
* question2: full unicode text of the second question
* is\_duplicate: label 1 if questions are duplicates, 0 otherwise
* Cwc\_min : This is the ratio of no. of common words to the length of smaller questions.
* Cwc\_max: This is the ratio of no. of common words to the length of larger questions.
* Csc\_min: This is the ratio of no. of common stopwords to the smaller stopword count among the 2 questions.
* Csc\_max: This is the ratio of no. of common stopwords to the larger stopword count among the 2 questions.
* Ctc\_min: This is the ratio of the number of common tokens to the smaller token count among the two tokens.
* Ctc\_max:This is the ratio of the number of common tokens to the larger token count among the two tokens.
* Last\_word\_eq: 1. if the last word in the two questions is same ,0 otherwise
* First\_word\_eq: 1. if the first word in the two questions is same ,0 otherwise

**Length Based Features:**

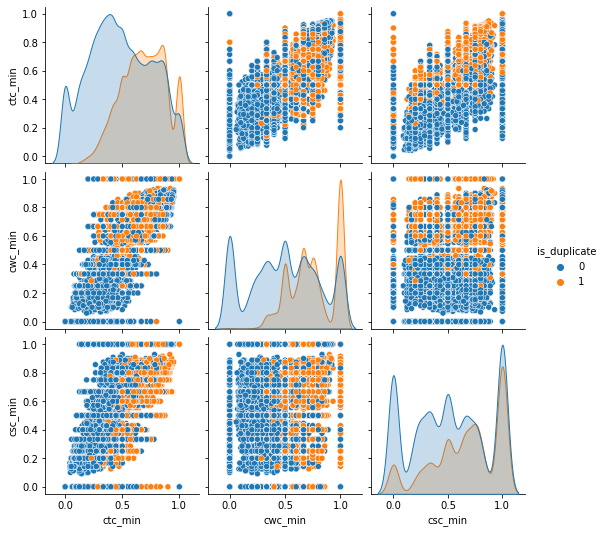
* Mean\_len: Mean of the length of two questions (no. of words)
* abs\_len\_diff: Absolute difference between the length of two questions (no. of words)
* Longest\_substr\_ratio: Ratio of the length of the longest substring among the two questions to the length of the smaller question.

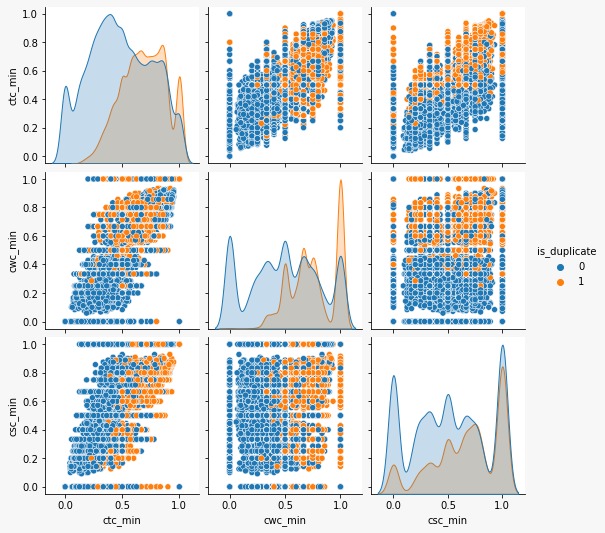
**Fuzzy Features:**

* 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

# EDA with Features

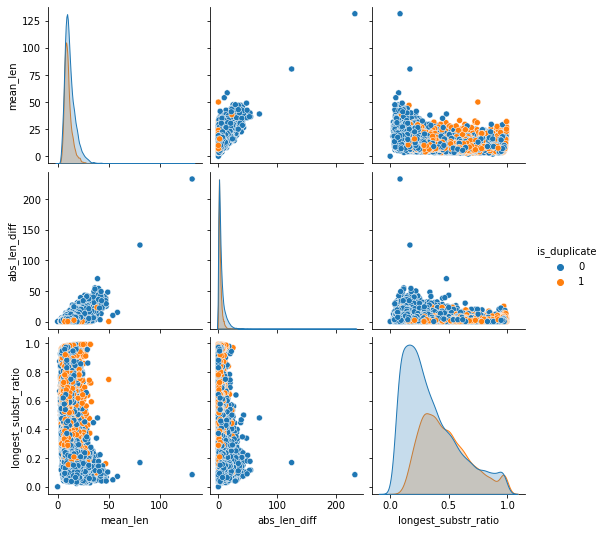
**Graph 0 :**



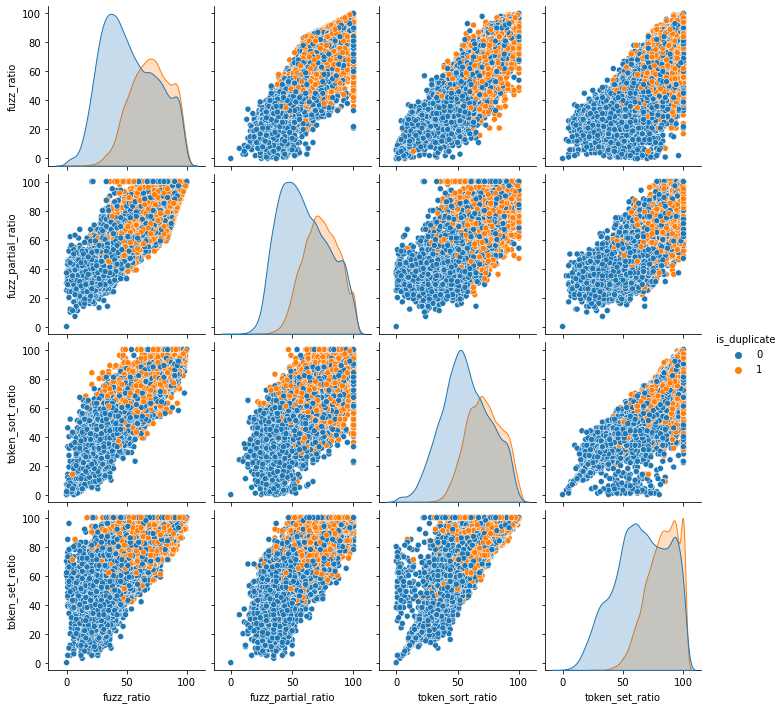


****

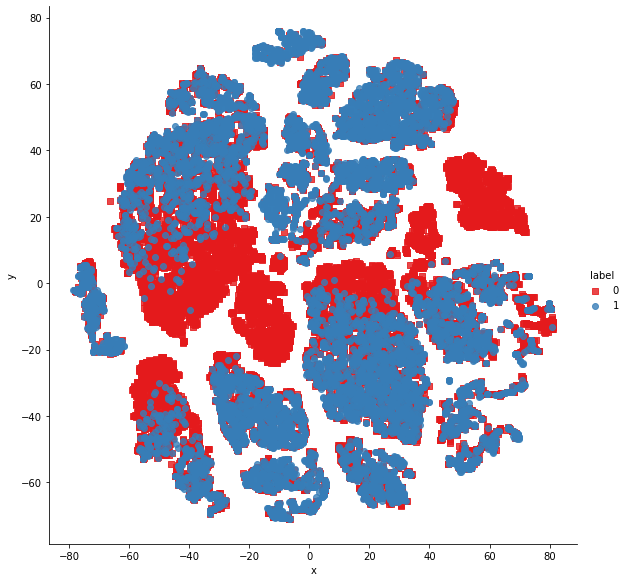
**Graph 3 :**



**Graph 4 :**



**Graph 5 :**



**Model Building**

## Random Forest:

A random forest model is a type of machine learning algorithm that is used for classification, regression, and other predictive tasks. It is an ensemble learning method that combines multiple decision trees, called a forest, and aggregates their results to produce a final prediction.

The random forest algorithm creates decision trees by randomly selecting subsets of the training data and features at each split, which helps to reduce overfitting and increase the accuracy of the model. The individual decision trees are constructed using a subset of the available features and training samples, and each tree makes a prediction based on its own set of rules.

Once the trees are constructed, the random forest algorithm aggregates their predictions to generate the final output. For classification tasks, the model uses majority voting, where the predicted class with the most votes across all the trees is selected. For regression tasks, the model takes the average of the predicted values from all the trees.

Random forests have several advantages over individual decision trees, including better accuracy, reduced overfitting, and increased robustness to outliers and noisy data. They are widely used in various domains such as finance, healthcare, and marketing for tasks such as fraud detection, disease diagnosis, and customer segmentation.

The accuracy achieved using Random Forest model was found out to be 0.7846666666666666

**XGBoost:**

XGBoost (Extreme Gradient Boosting) is a machine learning algorithm that uses a gradient boosting framework to train decision trees. It is an ensemble learning method that combines multiple weak learners, in this case decision trees, to create a more accurate and powerful model.

XGBoost is a supervised learning algorithm that can be used for both regression and classification tasks. The algorithm is iterative and works by training decision trees in a series of rounds, called boosting rounds. In each boosting round, the algorithm adds a new decision tree that corrects the errors of the previous trees. The process continues until the desired level of accuracy is achieved, or until a specified number of rounds is reached.

One of the key features of XGBoost is its ability to handle missing values and outliers in the data, making it more robust and accurate than other algorithms. It also includes a number of regularization techniques, such as L1 and L2 regularization, to prevent overfitting and improve the generalization of the model.

Another important feature of XGBoost is its ability to handle large datasets with high-dimensional features. It uses a parallel processing approach to speed up the training process and reduce computation time.

XGBoost has been used in various domains such as finance, healthcare, and e-commerce for tasks such as fraud detection, image classification, and customer churn prediction. It is considered one of the most powerful and popular machine learning algorithms and has won numerous competitions in the field of data science.

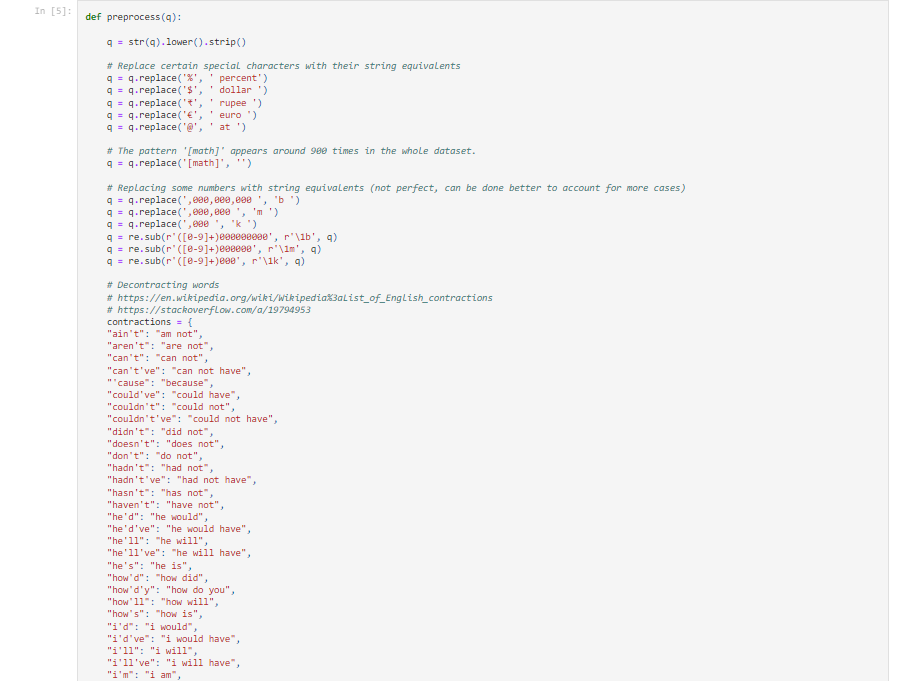
The accuracy achieved using XGBoost model was found out to be 0.7926666666666666

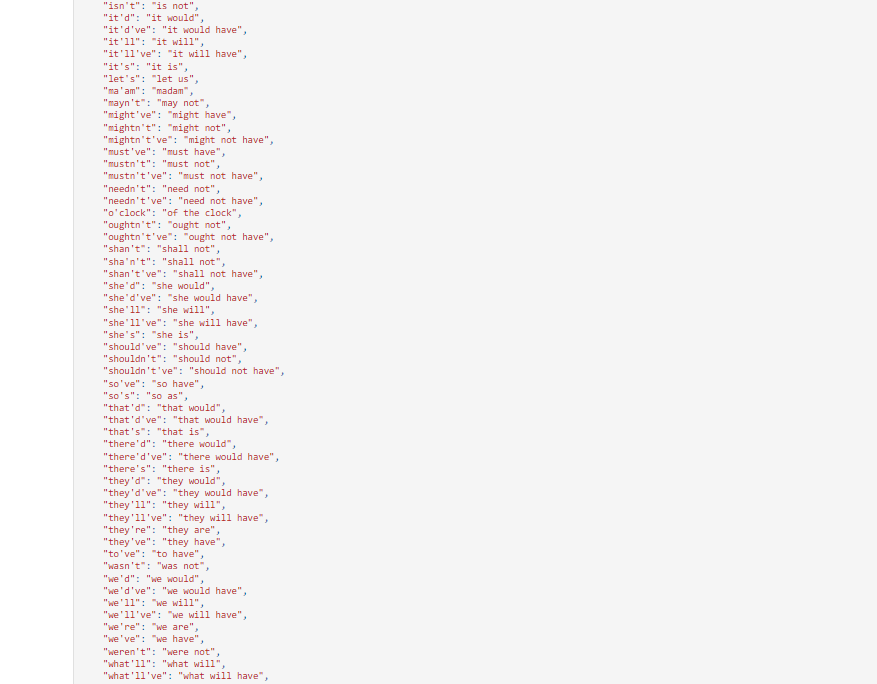
On comparing both the models we see that XGBoost model has more accuracy than Random Forest model but when we see the outcome of Confusion Matrix, we find that the probability of getting the outcome as duplicate for a non-duplicate question is more in case of XGBoost model.

Hence the accepted model is Random Forest Model.

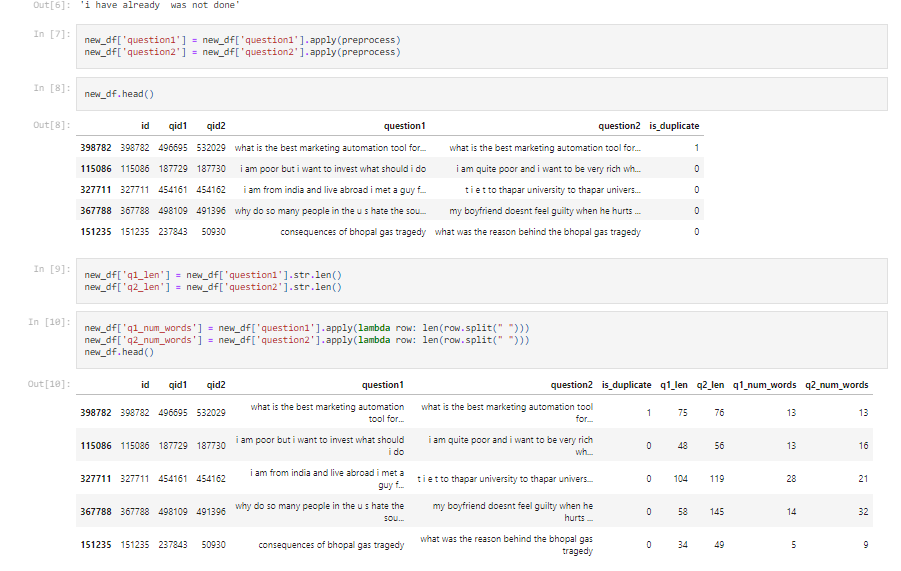
**Code**

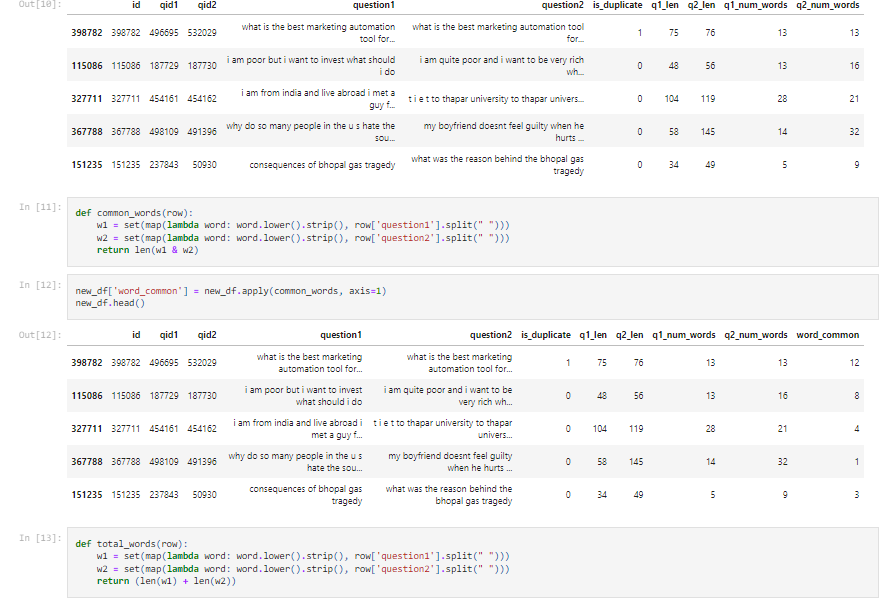


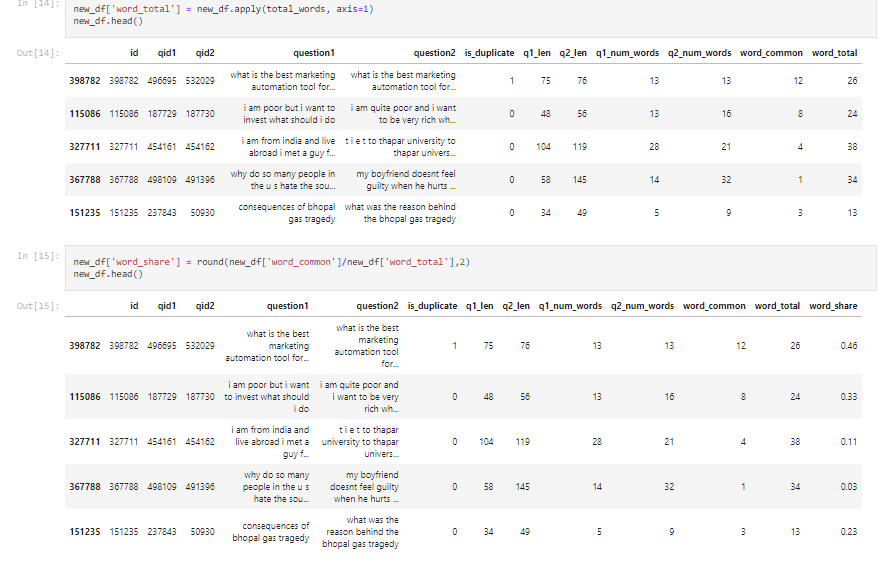


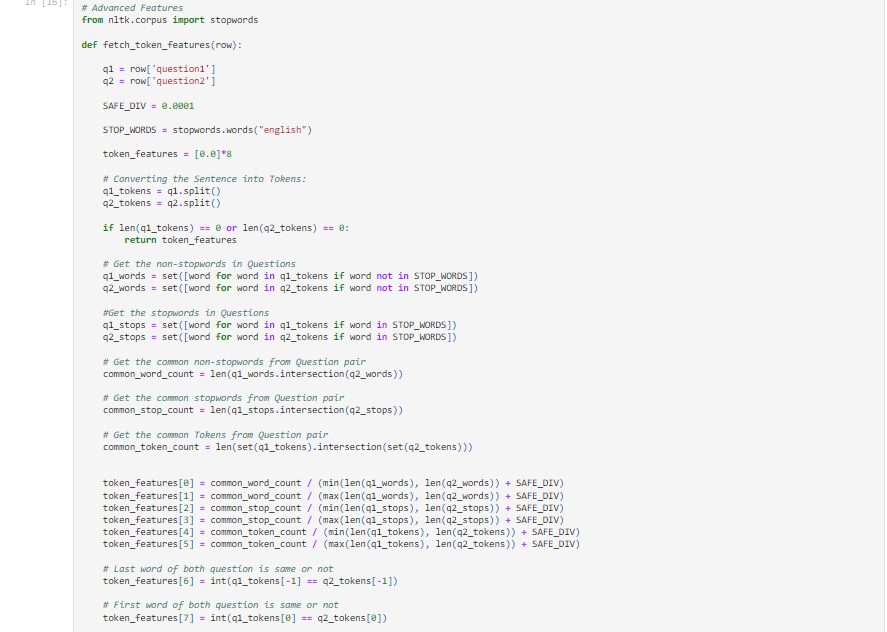






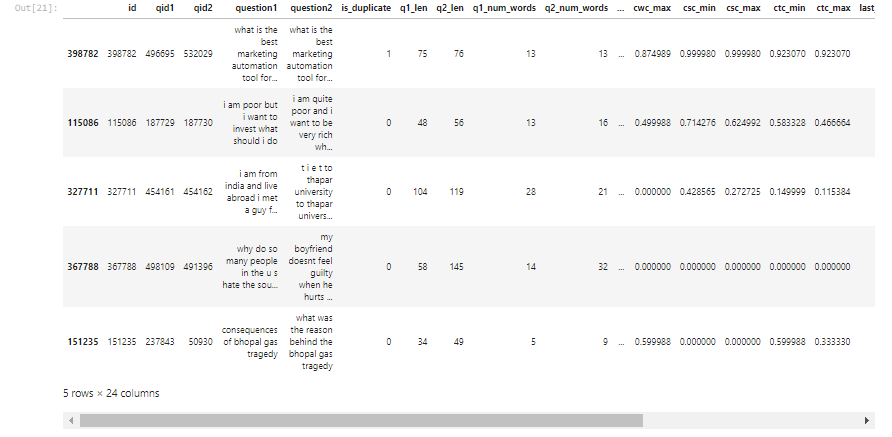




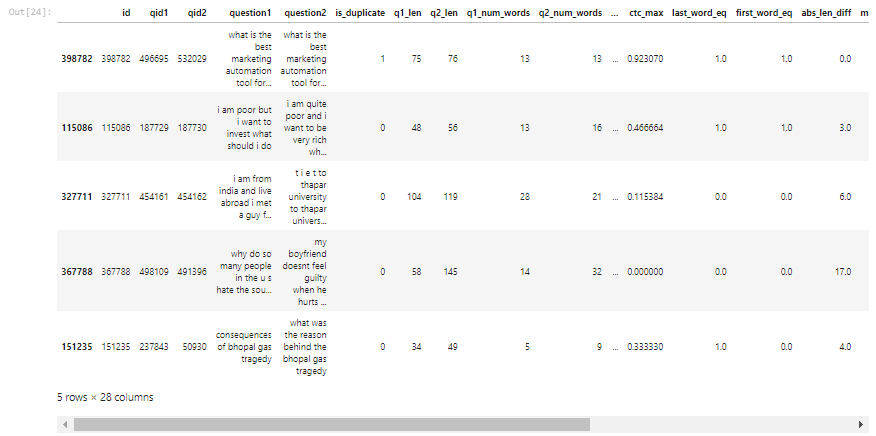






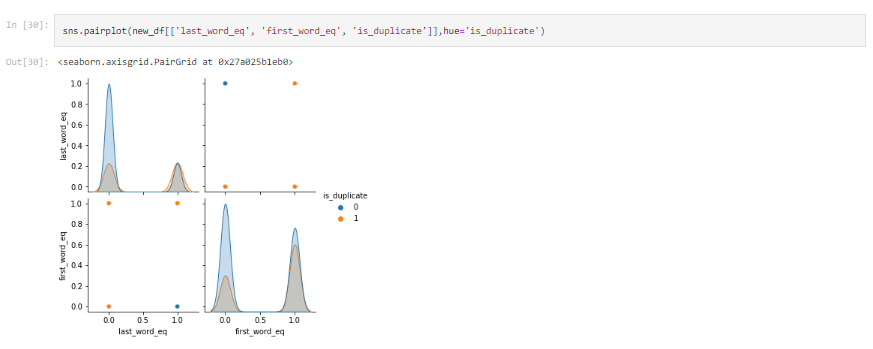


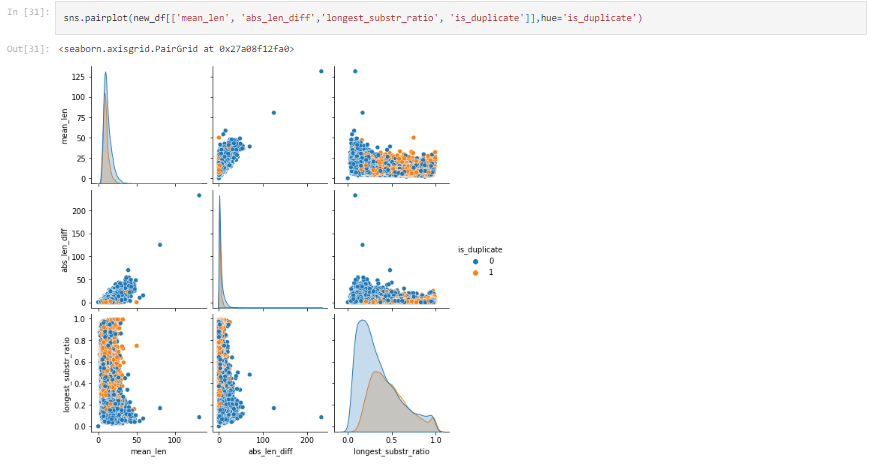


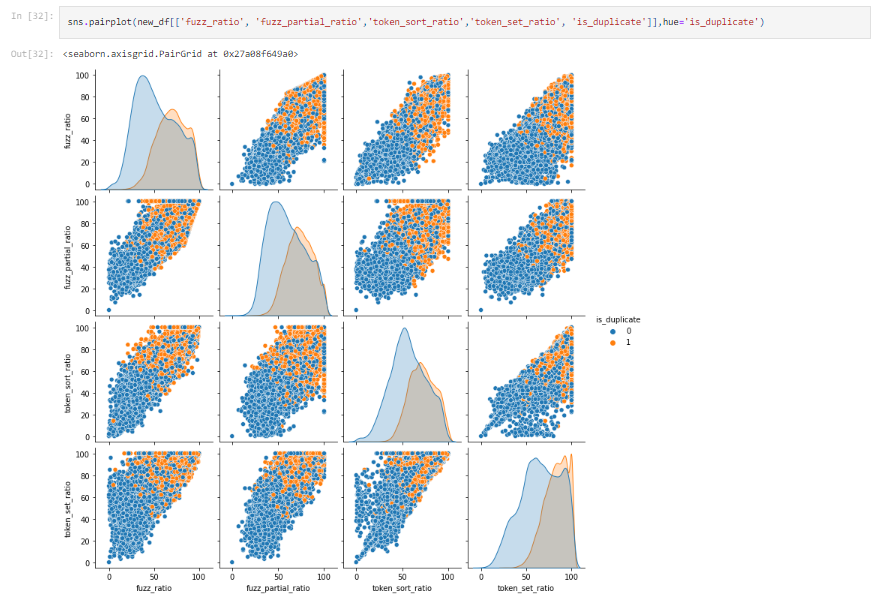




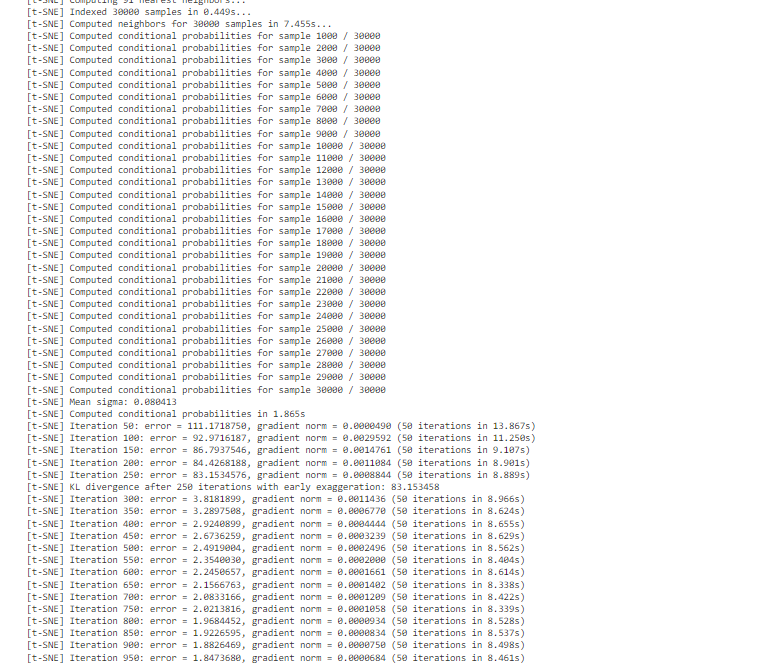


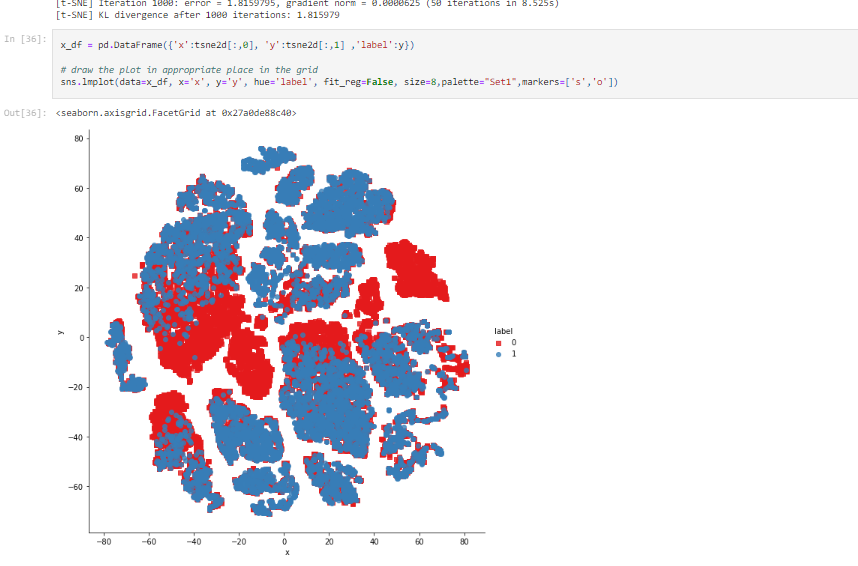


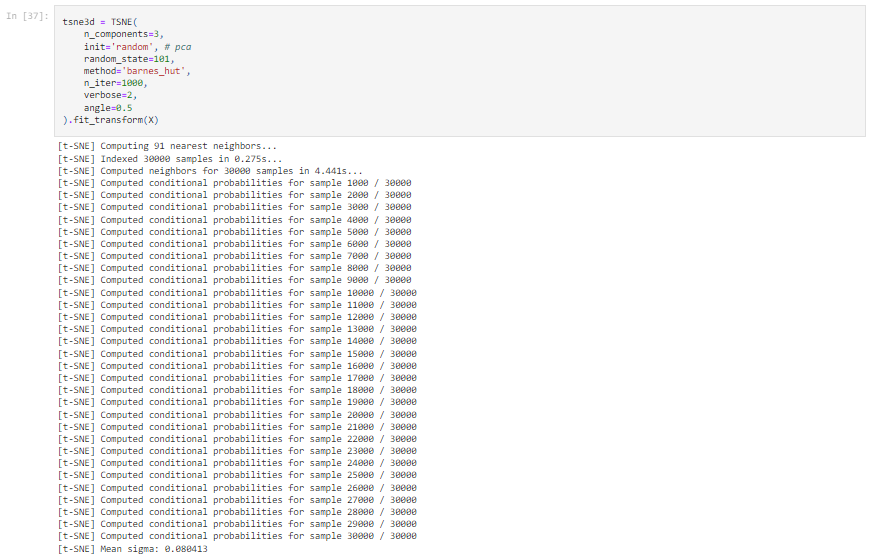


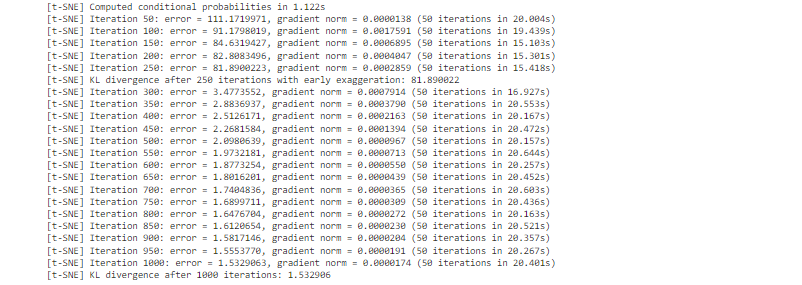




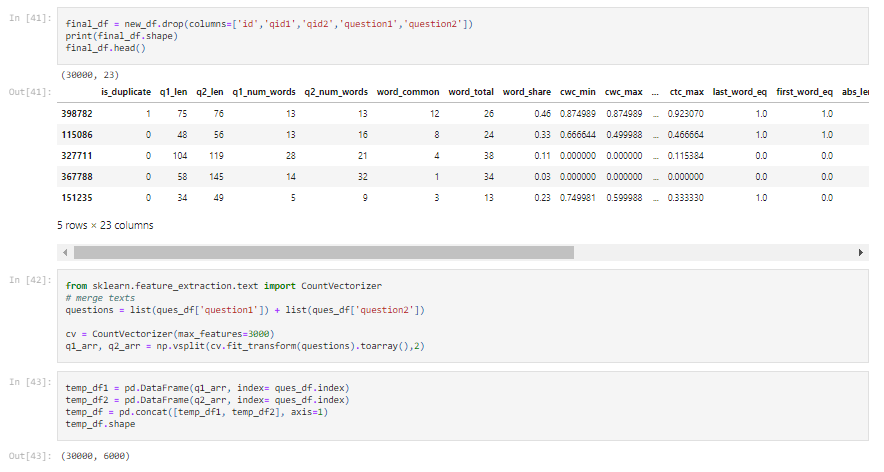


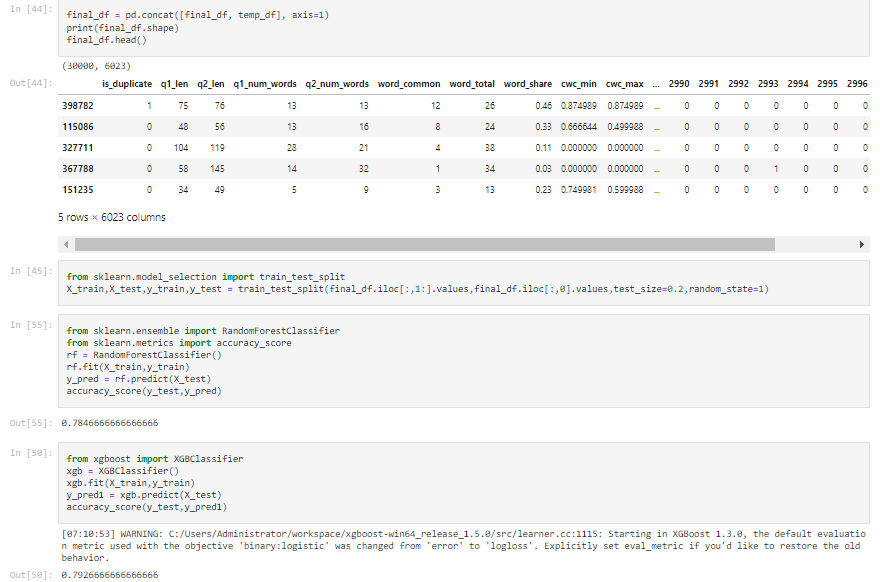


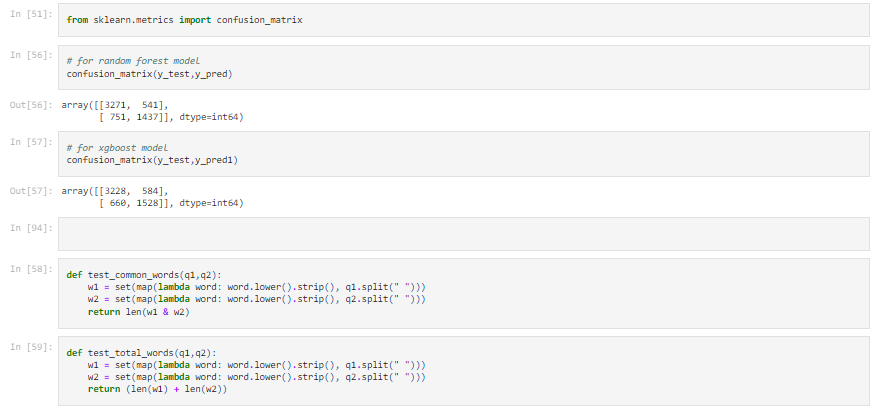


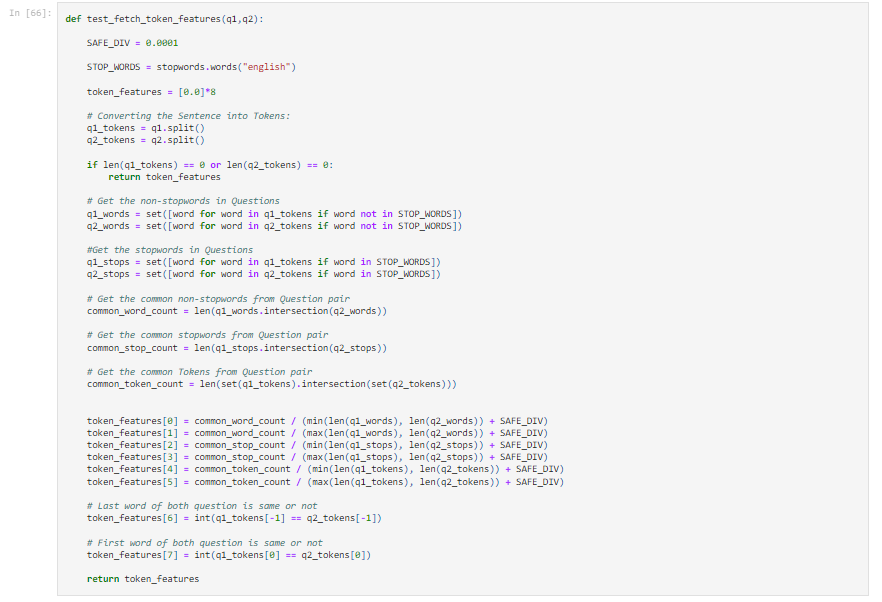


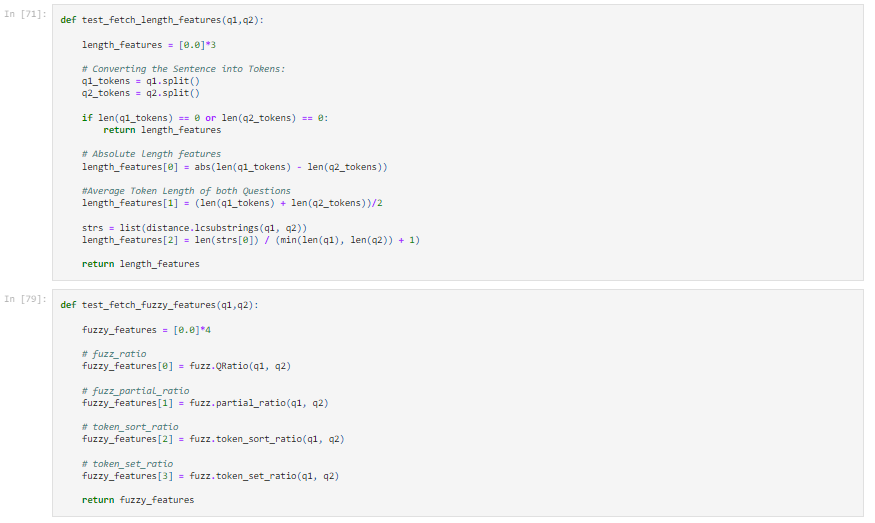
















# Future Scope of Improvements

* We are using BOW (bag of words) for sampling of our dataset which gives us a lack in semantic
* In spite of that we can use w2v (Word2Vec). The purpose and usefulness of Word2vec is to group the vectors of similar words together in vectorspace. Which will increase semantic meaning of our model

Our model is based on accuracy giving binary result as 0 and 1

1 – Stands for Duplicacy

0 – Non-Duplicacy

* Instead of that we can use **metric logloss** to return fractional result
* Every time we are getting 0 and 1 from our model to get result, we can also set a threshold (between 0 and 1) to increase the speed and accuracy.
* Later on, we will make a webpage on this approach to get a better GUI of our model and easy to access for users.

# Certificate

This is to certify that Mr. Abhijit Shaw of Asansol Engineering College, University roll number: 10800220081, has successfully completed a project on Duplicate Question Pairs using Machine Learning under the guidance of Prof. Arnab Chakraborty.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Prof. Arnab Chakraborty***

**Asansol Engineering College**

# Certificate

This is to certify that Mr. Amit Kumar Singh of Asansol Engineering College, University roll number: 10800220078, has successfully completed a project on Duplicate Question Pairs using Machine Learning under the guidance of Prof. Arnab Chakraborty.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Prof. Arnab Chakraborty***

**Asansol Engineering College**

# Certificate

This is to certify that Ms. Nisha Jha of Asansol Engineering College, University roll number: 10800220070, has successfully completed a project on Duplicate Question Pairs using Machine Learning under the guidance of Prof. Arnab Chakraborty.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Prof. Arnab Chakraborty***

**Asansol Engineering College**

# Certificate

This is to certify that Mr. J Srikanth Chetty of Asansol Engineering College, University roll number: 10800221141, has successfully completed a project on Duplicate Question Pairs using Machine Learning under the guidance of Prof. Arnab Chakraborty.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Prof. Arnab Chakraborty***

**Asansol Engineering College**

# Certificate

This is to certify that Mr. Agnishwar Mukherjee of Asansol Engineering College, University roll number: 10800220043, has successfully completed a project on Duplicate Question Pairs using Machine Learning under the guidance of Prof. Arnab Chakraborty.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Prof. Arnab Chakraborty***

**Asansol Engineering College**