TE MINI-PROJECT REPORT ON

DUPLICATE TEXT-PAIR DETECTION MECHANISM

Submitted in partial fulfillment of the requirements

of the degree of bachelor's in engineering

by

MIHIR CHHEDA	TE-5 63
NIDHI DAULAT	TE-5 64
MISHKAT SHAIKH	TE-6 39
SARVESH SHARMA	TE-6 42

Under the guidance of

Ms. Pranali Wagh

Mr. Santosh Rathod

DEPARTMENT OF INFORMATION TECHNOLOGY

SHAH & ANCHOR KUTCHHI ENGINEERING COLLEGE

CHEMBUR, MUMBAI-400088.

2021-2022

Mahavir Education Trust's

SHAH & ANCHOR KUTCHHI ENGINEERINGCOLLEGE

Mahavir Education Trust Chowk, W.T. Patil Marg, Chembur, Mumbai 400 088

Affiliated to University of Mumbai, Approved by D.T.E. & A.I.C.T.E.

# UG Programs accredited by NBA FOR 3years w.e.f.1st July, 2019

Certificate

This is to certify that the report of the mini project entitled

"DUPLICATE TEXT-PAIR DETECTION MECHANISM"

is a bonafide work of

MIHIR CHHEDA	TE-5 63
NIDHI DAULAT	TE-5 64
MISHKAT SHAIKH	TE-6 39
SARVESH SHARMA	TE-6 42

submitted to the

UNIVERSITY OF MUMBAI

during semester VI in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING

in

INFORMATION TECHNOLOGY

(Ms.Pranali Wagh)
Guide

(Mr. Santosh Rathod)
Co-Guide

(Ms. Swati Nadkarni)
I/c Head of Department

(Dr. Bhavesh Patel)
Principal

Approval for Mini Project Report for T. E. semester VI

This project report entitled "**Duplicate Text-Pair Detection Mechanism"** by Mihir Chheda, Nidhi Daulat, Mishkat Shaikh and Sarvesh Sharma is approved for semester VI in partial fulfilment of the requirement for the award of the degree of Bachelor of Engineering.

Guide:

Ms. Pranali Wagh

Co-Guide:

Mr. Santosh Rathod

Examiners:

1._____________________________

2._____________________________

Date: 5th May 2022.

Place: Mumbai

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Mihir Mahesh Chheda TE5-63	Mihir
Nidhi Nevil Daulat TE5-64	Nidhi
Mishkat Anis Ahmed Shaikh TE6-39	
Sarvesh Ramesh Sharma TE6-42	

(Name and Roll no) (Signature)

Date: 5th May 2022.

Table of Contents

Description	**Page No.**
Abstract	i
List of Figures	ii
List of Tables	iii
Chapter 1 Introduction	7
Chapter 2 Review of Literature (Comparative Analysis)	8
Chapter 3 Proposed System	11
Chapter 4 Methodology	12
Chapter 5 Summary	20
Chapter 6 References	22
Acknowledgement	23

ABSTRACT

This paper presents the results of systematic and comparative evaluations of a wide range of automated retrieval methods when applied to larger data sets continuously, thus allowing you to study the study profiles of this work under these different methods and assess their relevance. This study was made possible by turning to the latest releases for research purposes, with the online query query engine, a new database containing more than 400,000 pairs that have been labeled for their duplicate question segments. Automatic detection of equally mathematical questions is a very important function of the question-and-answer system. The Quora Database, released in the Quora Question Pairs competition organized by Kaggle, has now been used extensively to train the system in solving the task of identifying duplicate questions. However, the basic truth labels on this database are not 100% accurate and may include incorrect labeling. In this paper, we focus on improving the quality of the Quora database. A model has been created that aims to provide the result of whether a pair of included questions is duplicate or not.

Chapter 1

INTRODUCTION

DQD detection is a recent natural language processing project (NLP) that has been the subject of a number of practical studies, in which two segments of a query are considered mathematically equal, and then repeat, if they can find the same answer. Among the many applications for natural language processing, the questionnaire is a hot and tempting research environment with a wide range of commercial potential. With the advent of the Web, answering questions is a good indication of the problem of overload. The past decade has seen the emergence and rapid growth of public forums that answer questions such as Quora and Stack Overflow. Over the years, they have collected a large number of related questions and answers. Not surprisingly, many people ask similar questions. As a result, there is a need to find the same queries as users in the existing database to answer the database, so that the system can retrieve the answer by receiving answers from the same queries. As a long-term problem in understanding natural language, the automatic acquisition of mathematically equitable questions is now a very important function in the question-and-answer system. Much of the motivation for this research article comes from the use of DQD to support online questions that answer community forums, and forums, in general. For example, if used in the first case, DQD can be used to automatically detect whether a new user query in a forum was previously requested in that forum, and to help mark and eventually remove it as a duplicate query, reducing the increase in duplicate queries which is a major obstacle. And when embedded in a discussion area, DQD can be used to compare a newly added question with a pair of website answers to previous questions and if the same question is found, to respond by submitting a corresponding response, thus avoiding turning to the human driver.

Chapter 2

REVIEW OF LITERATURE

A **literature survey** or **literature review** is a type of review article. A literature review is a scholarly paper that presents the current knowledge including substantive findings as well as theoretical and methodological contributions to a particular topic. Literature reviews are secondary sources and do not report new or original experimental work. Literature reviews are a basis for research in nearly every academic field. A narrow-scope literature review may be included as part of a peer-reviewed journal article presenting new research, serving to situate the current study within the body of the relevant literature and to provide context for the reader.

Comparative Analysis

Sr. No.	Title	Author	Publisher	Date	Technology Used	Advantages	Disadvantages	Features to be Implemented
1.	On Application of Natural Language Processing in Machine Translation	Zhaorong Zong, Changchun Hong	IEEE	16 Sept. 2018	Neural machine translation & NLP in python	This article compares machine translation with NLP	It requires more computing power for machine translation	To understand NLP and its use case in real world.
2.	Translation of natural language queries to structured data sources	Ruslan Posevkin, Igor Bessmertny	IEEE	30 Nov 2015	SQL, Machine learning	It converts natural language query to SQL query.	Being in SQL data format it limits the scope of treating inconsistent data.	Querying text-pair questions into data format was exercised.

3.	Information Processing and Retrieval from CSV File by Natural Language	Chalermpol Tapsai	IEEE	30 Dec. 2018	Python in machine learning	Converting raw data of csv into NLP format for direct processing purpose.	Searching or retrieval of data from csv formats is quite a limited scope.	The project uses quora dataset of train.csv file for processing of data.
4.	Research on Data Preprocessing and 3D Matrix Model	Liu Hongling, Wan Di	IEEE	25 Oct. 2020	Python in machine learning	For overcoming the big data problems this paper offers a 3-D matrix solution.	It is difficult to analyze and comprehend mathematical solutions and logics on a big data.	A large dataset of 8,00,000 question pairs is simplified and then used.
5.	Pointer-Generator Abstractive Text Summarization Model with Part of Speech Features	Shuxia Ren, Zheming Zhang	IEEE	20 Oct. 2019	Logical & computational mathematics and python.	This paper focuses on abstract text summarization using OOV problems.	It is a lengthy process and requires a lot of features to be included in dataset.	Instead of using OOV our model is implemented using BOW to make it simpler to use.
6.	ATC: An Automatic Text Comparison Tool Based on Diff Algorithm	Li Lixun, Wang Gaoshan, Dou Zengjie, Feng Yan	IEEE	20 Mar 2020	ATC model	This article focuses on use of ATC model for combining multiple features and using it to analyze.	Multiple features can lead to a large data storage which is not scalable.	Use of advanced feature engineering is done in detection mechanism.
7.	Improve Quora Question Pair Dataset	Huong T. Le, Dung T. Cao	IEEE	1st Sept 2021	Python in machine learning	This paper focuses on analyzing of quora text-	It can be enhanced more by taking more	The idea of duplicate text-pair detection model

	for Question Similarity Task					pairs to detect similarity.	features and a larger dataset.	was considered and experimented.
8.	Learning Profiles in Duplicate Question Detection	Chakaveh Saedi, Jo˜ao Rodrigues, Jo˜ao Silva, Ant´onio Branco	IEEE	6 Aug. 2017	Python in machine learning.	Analyzing different question pairs semantically and finding their duplication.	A lot of pre-processing and analyzing of data has to be done to reach the result.	Identification of duplicate pairs was understood and used in our project mechanism.
9.	An Effective Crop Prediction Using Random Forest Algorithm	Dr. V.Geetha A. Punitha M. Abarna M. Akshaya	IEEE	4 July 2020	Random forest algorithm was used in python.	Crop based accuracy prediction was done using random forest.	The accuracy of random forest is less in comparison to other algorithms.	Random forest algorithm was used to predict the accuracy of text and data and then build a model according to it.
10.	Bengali Words Classification by Its Prefix Using Machine Learning Classifiers	K.M. Shahriar Islam, Sharun Akter Khushbu	IEEE	8 July 2021	Algorithms like random forest, decision tree.	The article focused to classifying Bengali text words using various algorithms.	Words classified were only for Bengali purpose and not English which is our use case.	English grammar and its words were analysed and applied algorithms by taking this paper as a reference.

Chapter 3

PROPOSED SYSTEM

An identification mechanism of duplicate text-pairs is developed. Quora dataset from kaggle is used. Data pre-processing and exploratory data analysis is performed over it to obtain a good accuracy score. Random forest classifier and XGB classifier algorithms are used to compare the accuracy percentage. A prediction model is built using the stream-lit module of python that finds out whether an entered pair of questions are a duplicate of each other or not. This use-case helps websites like Quora and StackOverflow to effectively eliminate duplicate questions from their site, and provide a clean User Interface.

Technologies Used:

1. **Front End:** Stream-lit module of python
2. **Backend:** Python Programming
3. **Dataset:** Quora dataset from kaggle

Chapter 4

METHODOLOGY

1) Algorithms

a) Random Forest Classifier Algorithm

Random Forest could be a popular machine learning algorithm that's a part of a supervised learning strategy may be used for both Classification and Regression problems in ML. it's supported the concept of integrated learning, which is that the process of integrating multiple dividers to resolve complex problems and improve model performance.
As the name suggests, "The Random Forest could be a subdivision that contains variety of decision trees for the assorted datasets set and takes measurement to boost the prediction accuracy of that database." rather than wishing on one decision tree, the random forest takes a prediction from each tree and is predicated on multiple predictable votes, and predicts the ultimate outcome.
The large number of trees within the forest results in high accuracy and prevents the matter of overcrowding.

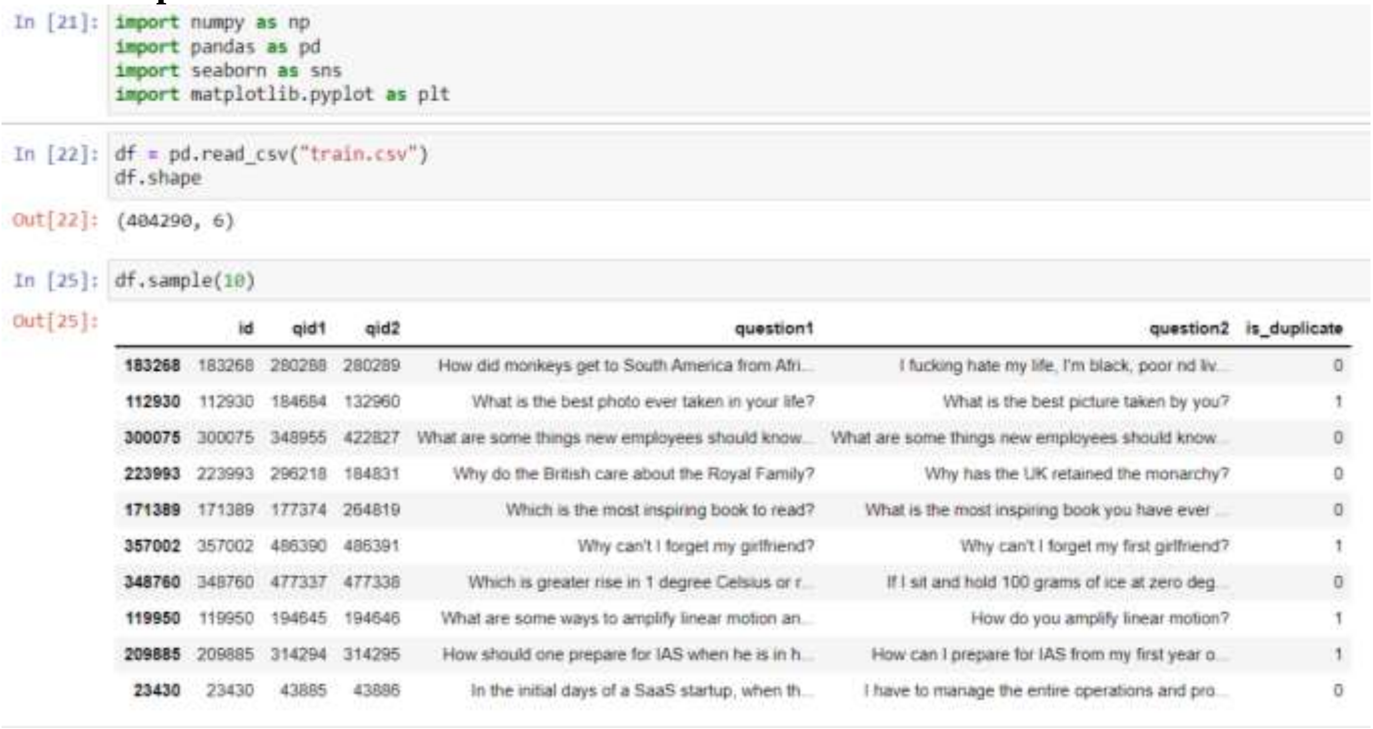
b) XGB Classifier Algorithm

XGBoost is an implementation of advanced Gradient decision trees. XGBoost models dominate most Kaggle tournaments.
In this algorithm, decision trees are created in sequence. Weight plays a very important role in XGBoost. Weights are given to any or all independent variants which are then incorporated into the choice tree predicting results. the load of the variables predicted that the tree is wrong increases and these variables are then fed to the second decision tree. These individual variables / predictions are then compiled to produce a more robust and accurate model. It can work on retransmission, classification, level, and guessing problems defined by the user.

2) EXPERIMENTATION AND RESULTS

1. Initial Exploratory Data Analysis

a. Sample rows of dataset



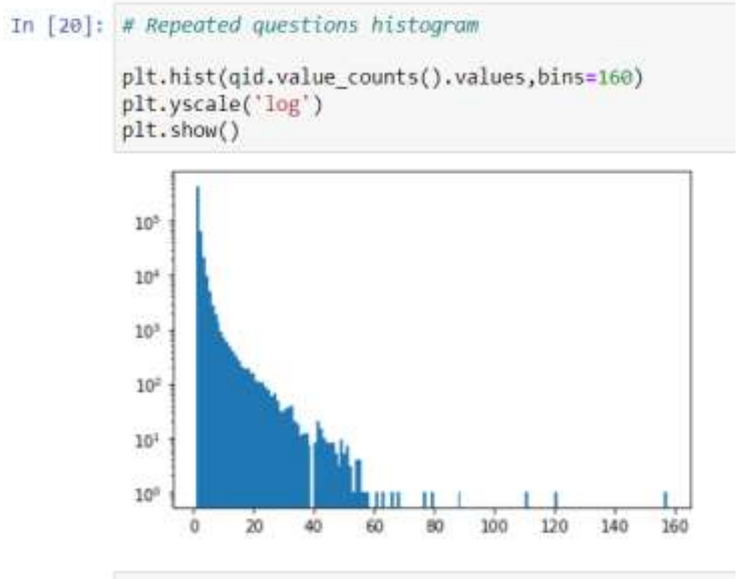
```
In [21]: import numpy as np
         import pandas as pd
         import seaborn as sns
         import matplotlib.pyplot as plt
```

```
In [22]: df = pd.read_csv("train.csv")
         df.shape

Out[22]: (404290, 6)
```

```
In [25]: df.sample(10)

Out[25]:
```

	id	qid1	qid2	question1	question2	is_duplicate
183268	183268	280288	280289	How did monkeys get to South America from Afri…	I fucking hate my life, I'm black, poor nd liv…	0
112930	112930	184684	132960	What is the best photo ever taken in your life?	What is the best picture taken by you?	1
300075	300075	348955	422827	What are some things new employees should know…	What are some things new employees should know…	0
223993	223993	296218	184831	Why do the British care about the Royal Family?	Why has the UK retained the monarchy?	0
171389	171389	177374	264819	Which is the most inspiring book to read?	What is the most inspiring book you have ever …	0
357002	357002	486390	486391	Why can't I forget my girlfriend?	Why can't I forget my first girlfriend?	1
348760	348760	477337	477338	Which is greater rise in 1 degree Celsius or r…	If I sit and hold 100 grams of ice at zero deg…	0
119950	119950	194645	194646	What are some ways to amplify linear motion an…	How do you amplify linear motion?	1
209885	209885	314294	314295	How should one prepare for IAS when he is in h…	How can I prepare for IAS from my first year o…	1
23430	23430	43885	43886	In the initial days of a SaaS startup, when th…	I have to manage the entire operations and pro…	0

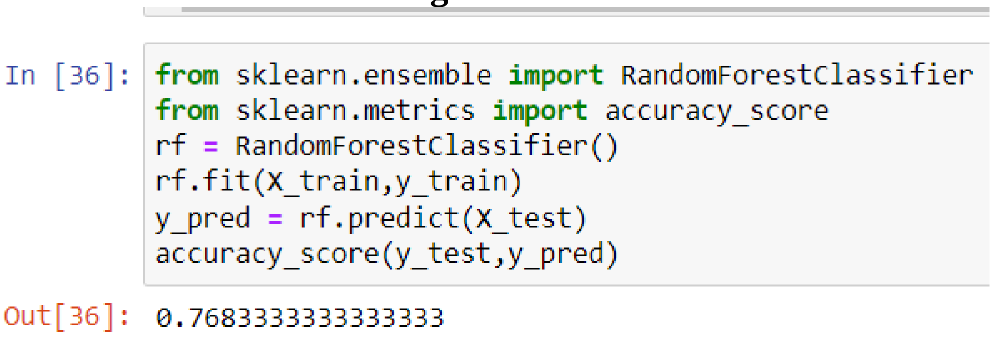
b. Histogram of repeated questions



```
In [20]: # Repeated questions histogram

         plt.hist(qid.value_counts().values,bins=160)
         plt.yscale('log')
         plt.show()
```

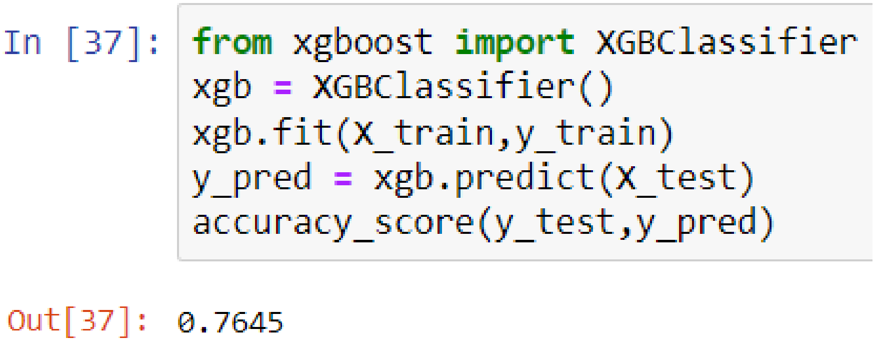
c. Initial accuracy

I. Random Forest Classifier algorithm



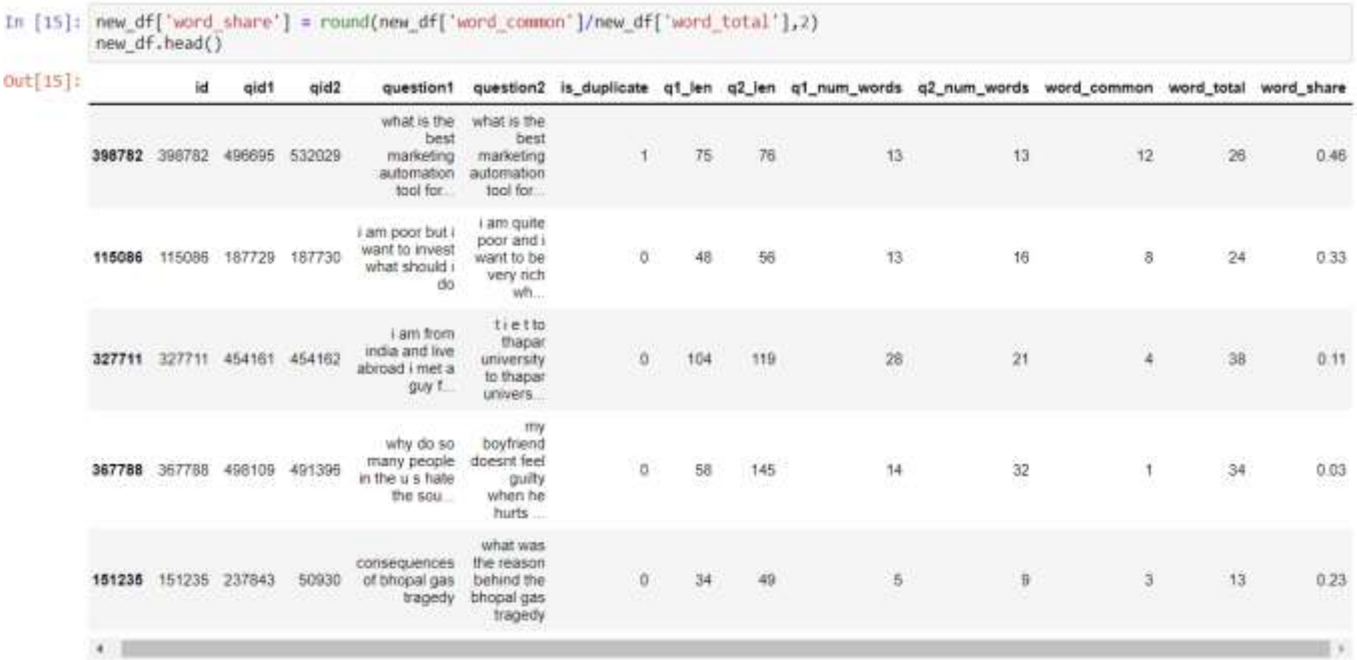
```
In [36]: from sklearn.ensemble import RandomForestClassifier
         from sklearn.metrics import accuracy_score
         rf = RandomForestClassifier()
         rf.fit(X_train,y_train)
         y_pred = rf.predict(X_test)
         accuracy_score(y_test,y_pred)

Out[36]: 0.7683333333333333
```

II. XGB Classifier algorithm



```
In [37]: from xgboost import XGBClassifier
         xgb = XGBClassifier()
         xgb.fit(X_train,y_train)
         y_pred = xgb.predict(X_test)
         accuracy_score(y_test,y_pred)

Out[37]: 0.7645
```

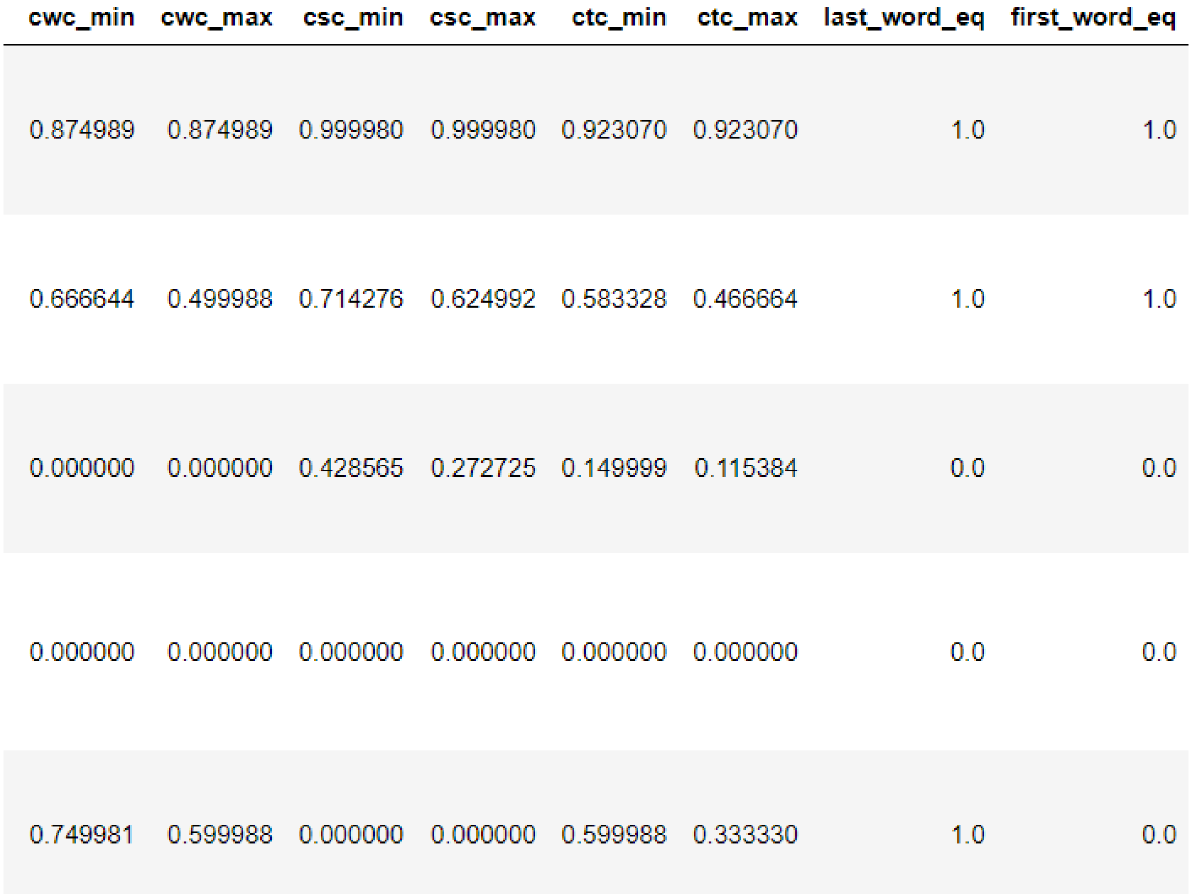
2. Advanced Features



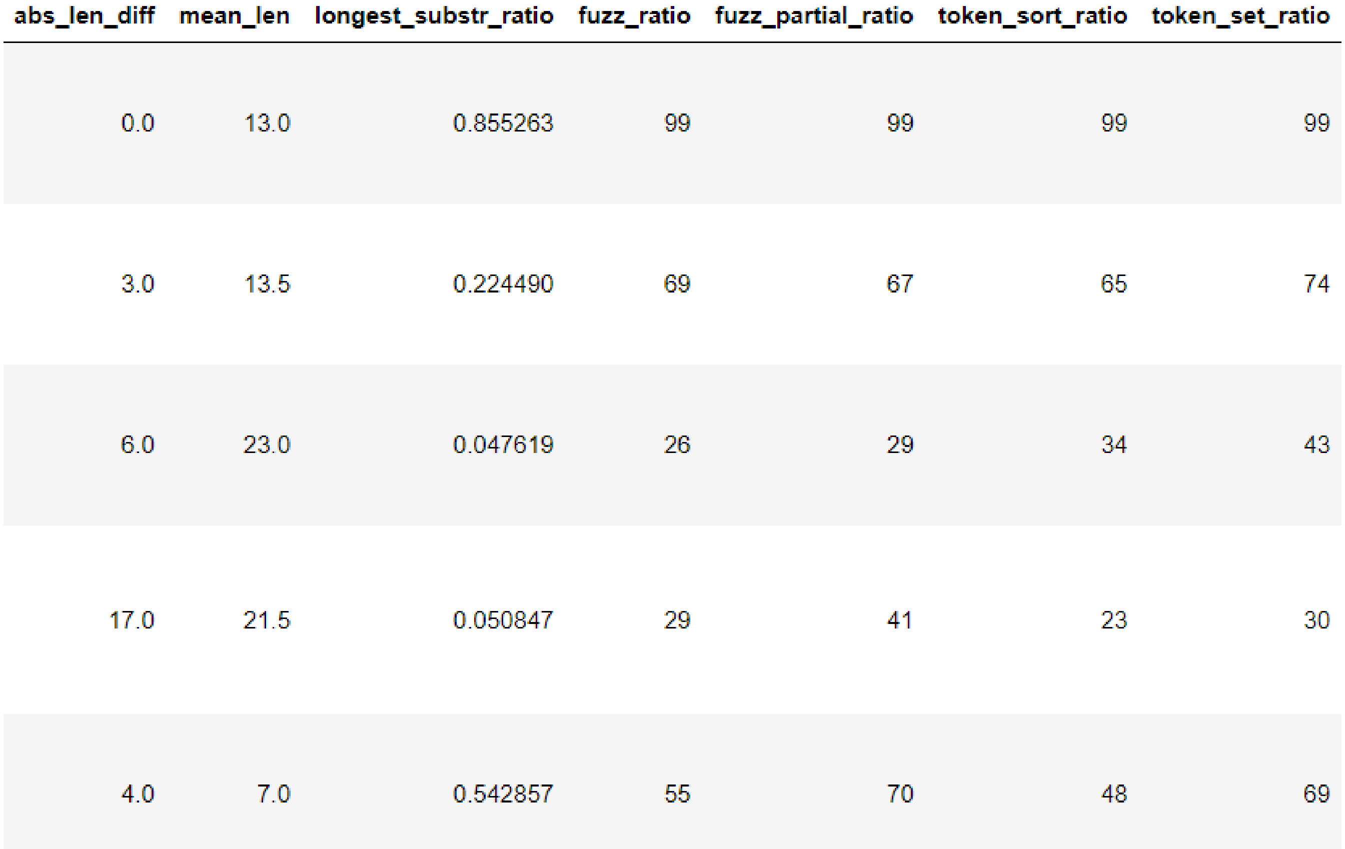
```
In [15]: new_df['word_share'] = round(new_df['word_common']/new_df['word_total'],2)
         new_df.head()
```

Out[15]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_num_words	q2_num_words	word_common	word_total	word_share
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1	75	76	13	13	12	26	0.46
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	48	56	13	16	8	24	0.33
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	104	119	28	21	4	38	0.11
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	58	145	14	32	1	34	0.03
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0	34	49	5	9	3	13	0.23



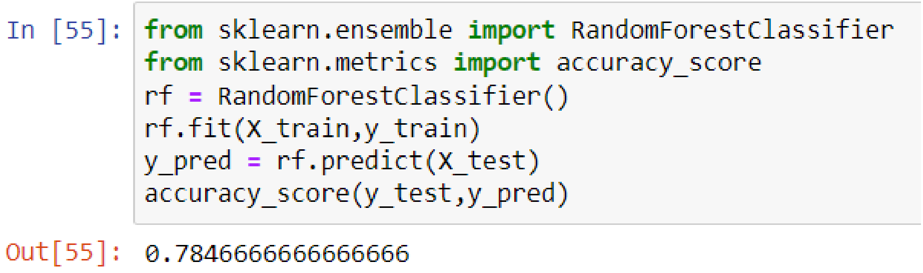
cwc_min	cwc_max	csc_min	csc_max	ctc_min	ctc_max	last_word_eq	first_word_eq
0.874989	0.874989	0.999980	0.999980	0.923070	0.923070	1.0	1.0
0.666644	0.499988	0.714276	0.624992	0.583328	0.466664	1.0	1.0
0.000000	0.000000	0.428565	0.272725	0.149999	0.115384	0.0	0.0
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0
0.749981	0.599988	0.000000	0.000000	0.599988	0.333330	1.0	0.0



abs_len_diff	mean_len	longest_substr_ratio	fuzz_ratio	fuzz_partial_ratio	token_sort_ratio	token_set_ratio
0.0	13.0	0.855263	99	99	99	99
3.0	13.5	0.224490	69	67	65	74
6.0	23.0	0.047619	26	29	34	43
17.0	21.5	0.050847	29	41	23	30
4.0	7.0	0.542857	55	70	48	69

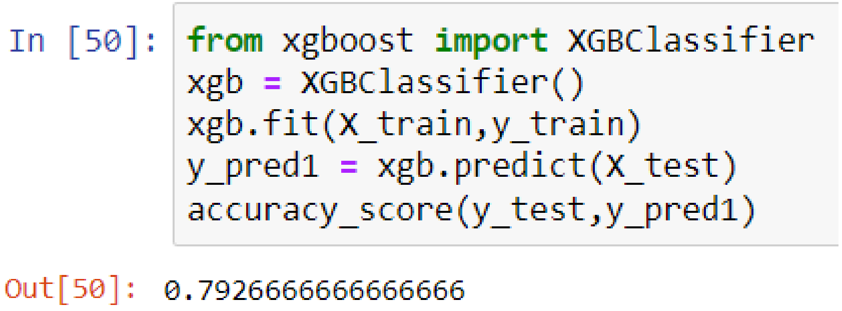
3. Final Accuracy

i. Random Forest Classifier Algorithm



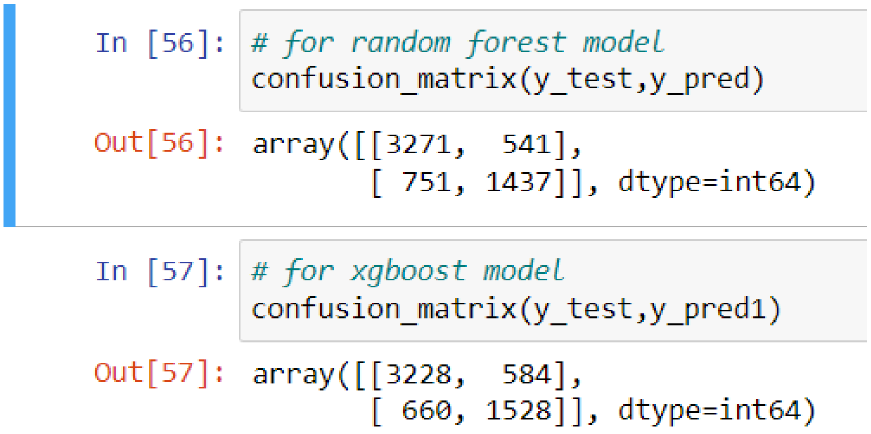
```
In [55]: from sklearn.ensemble import RandomForestClassifier
         from sklearn.metrics import accuracy_score
         rf = RandomForestClassifier()
         rf.fit(X_train,y_train)
         y_pred = rf.predict(X_test)
         accuracy_score(y_test,y_pred)

Out[55]: 0.7846666666666666
```

ii. XGB Classifier Algorithm



```
In [50]: from xgboost import XGBClassifier
         xgb = XGBClassifier()
         xgb.fit(X_train,y_train)
         y_pred1 = xgb.predict(X_test)
         accuracy_score(y_test,y_pred1)

Out[50]: 0.7926666666666666
```

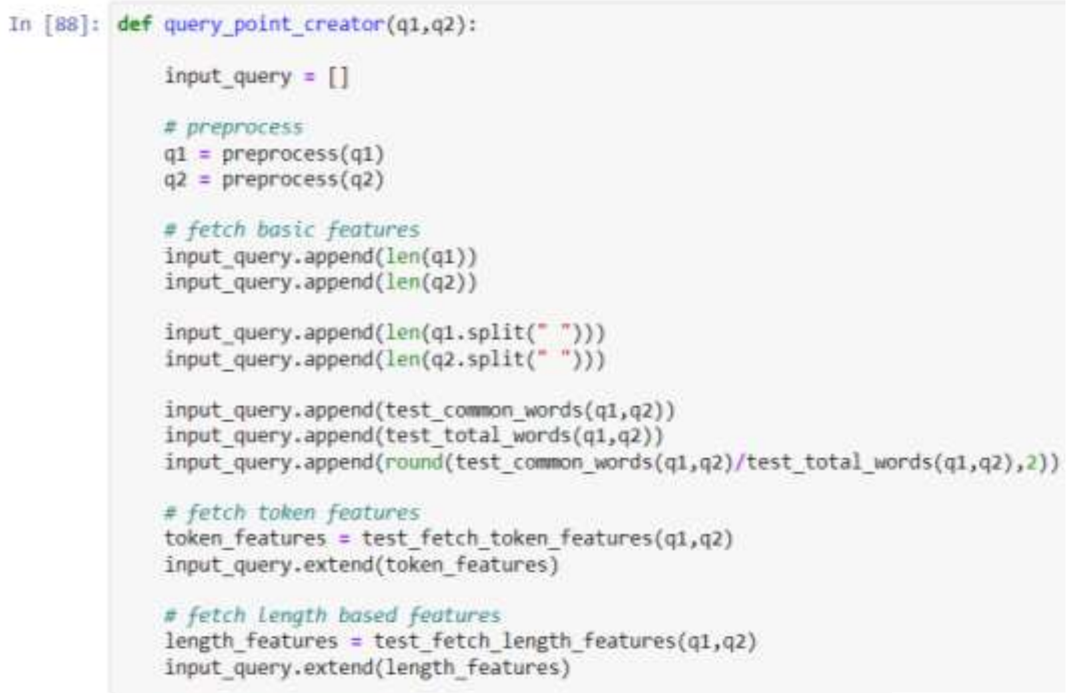
iii. Confusion Matrix



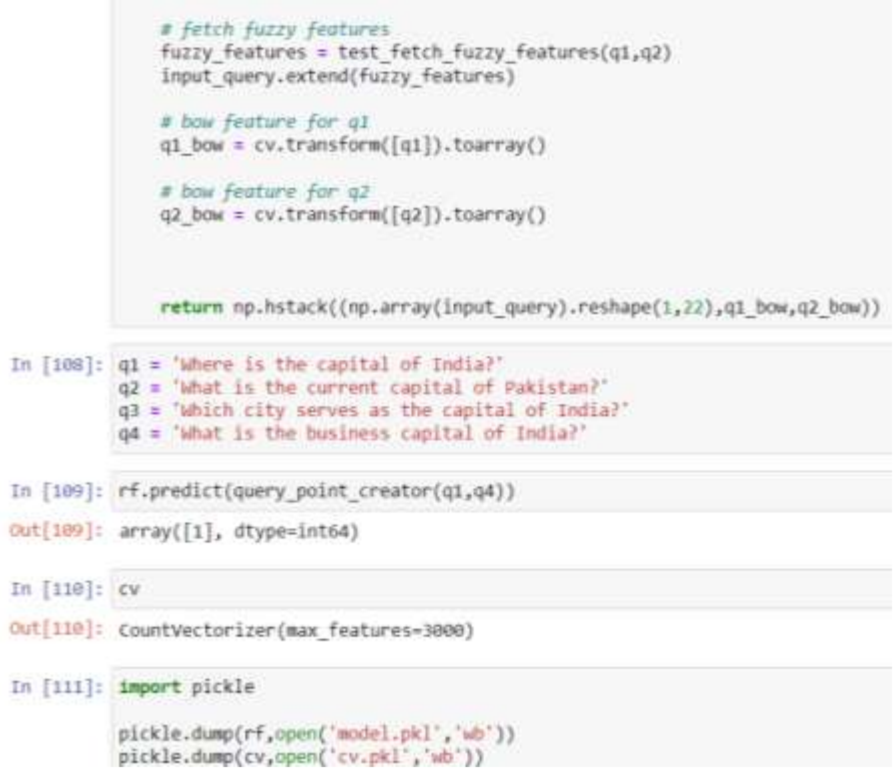
```
In [56]: # for random forest model
         confusion_matrix(y_test,y_pred)

Out[56]: array([[3271,  541],
                [ 751, 1437]], dtype=int64)

In [57]: # for xgboost model
         confusion_matrix(y_test,y_pred1)

Out[57]: array([[3228,  584],
                [ 660, 1528]], dtype=int64)
```

4. Prediction Model

i. Jupyter Notebook Code



```
In [88]: def query_point_creator(q1,q2):

             input_query = []

             # preprocess
             q1 = preprocess(q1)
             q2 = preprocess(q2)

             # fetch basic features
             input_query.append(len(q1))
             input_query.append(len(q2))

             input_query.append(len(q1.split(" ")))
             input_query.append(len(q2.split(" ")))

             input_query.append(test_common_words(q1,q2))
             input_query.append(test_total_words(q1,q2))
             input_query.append(round(test_common_words(q1,q2)/test_total_words(q1,q2),2))

             # fetch token features
             token_features = test_fetch_token_features(q1,q2)
             input_query.extend(token_features)

             # fetch length based features
             length_features = test_fetch_length_features(q1,q2)
             input_query.extend(length_features)
```

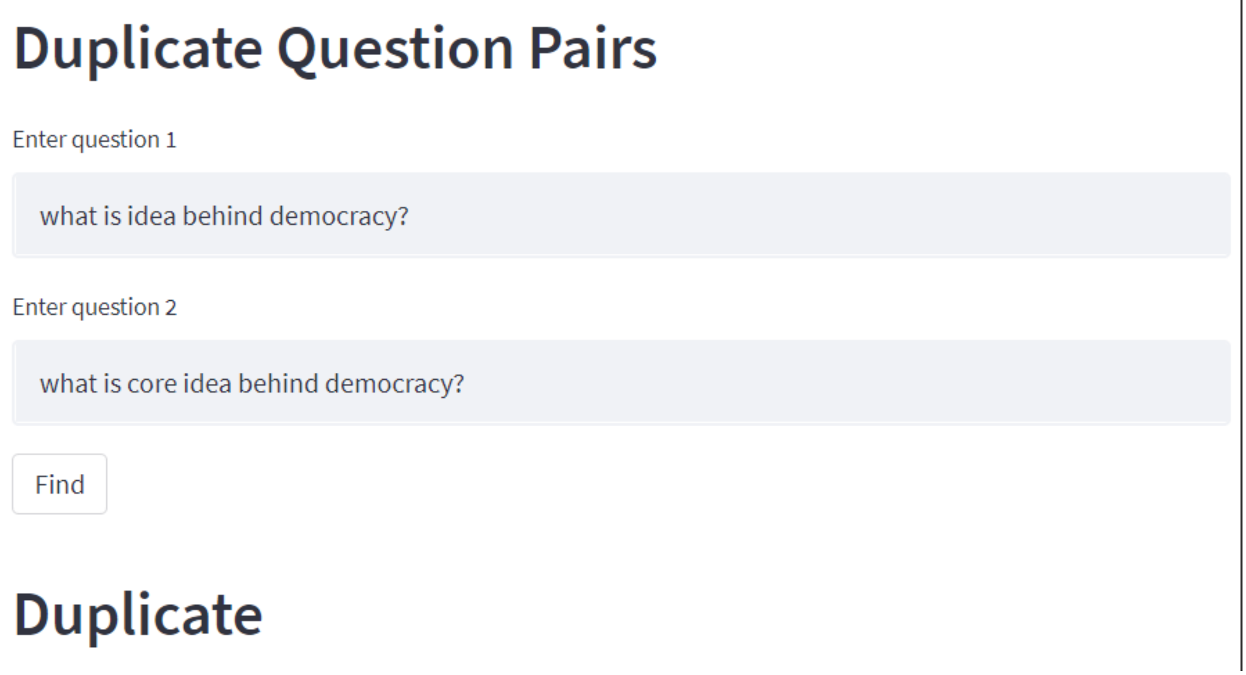


```
             # fetch fuzzy features
             fuzzy_features = test_fetch_fuzzy_features(q1,q2)
             input_query.extend(fuzzy_features)

             # bow feature for q1
             q1_bow = cv.transform([q1]).toarray()

             # bow feature for q2
             q2_bow = cv.transform([q2]).toarray()



             return np.hstack((np.array(input_query).reshape(1,22),q1_bow,q2_bow))

In [108]: q1 = 'Where is the capital of India?'
          q2 = 'What is the current capital of Pakistan?'
          q3 = 'Which city serves as the capital of India?'
          q4 = 'What is the business capital of India?'

In [109]: rf.predict(query_point_creator(q1,q4))

Out[109]: array([1], dtype=int64)

In [110]: cv

Out[110]: CountVectorizer(max_features=3000)

In [111]: import pickle

          pickle.dump(rf,open('model.pkl','wb'))
          pickle.dump(cv,open('cv.pkl','wb'))
```

ii. Pycharm code for stream-lit module

```
import streamlit as st
import helper
import pickle

model = pickle.load(open('model.pkl','rb'))

st.header('Duplicate Question Pairs')

q1 = st.text_input('Enter question 1')
q2 = st.text_input('Enter question 2')

if st.button('Find'):
   query = helper.query_point_creator(q1,q2)
   result = model.predict(query)[0]

   if result:
       st.header('Duplicate')
   else:
       st.header('Not Duplicate')
```

iii. Demo for Duplicate question-pairs



Duplicate Question Pairs

Enter question 1

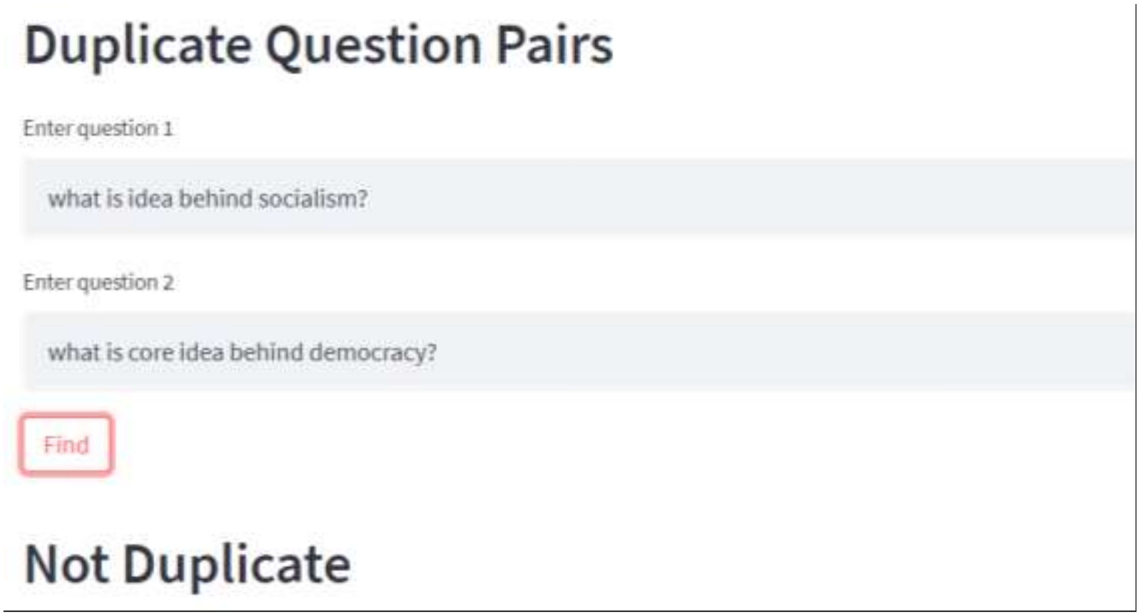
what is idea behind democracy?

Enter question 2

what is core idea behind democracy?

Duplicate

iv. Demo for not duplicate question-pairs



Duplicate Question Pairs

Enter question 1

what is idea behind socialism?

Enter question 2

what is core idea behind democracy?

Not Duplicate

Chapter 5

SUMMARY

Part-1) CONCLUSION

The project introduces a model to predict whether the entered question pairs are duplicate or not. We have used algorithms like Random forest classifier that gives an accuracy of 78% whereas XGB classifier that gives an accuracy of 79%. Though it looks like XGB classifier has a higher accuracy still it’s not efficient. This can be proved by finding the confusion matrix. It depicts that even if question pairs are not duplicate it shows to be duplicate. This error is much more avoided while using random forest algorithm and hence it’s more reliable to use. Basic exploratory data analysis id performed to know more about the dataset and act accordingly. The machine learning implementations from jupyter notebook are imported into pycharm. This is done so to create and host a website on stream-lit.

Research reported within the current paper allows for further understanding of the methods of obtaining automatic recurring questions and their use. The most conclusion is that the foremost complex sort of reference method, i.e. in-depth reading, apparently works far better than other methods, as long because it isn't trained in a very database large enough. Interestingly, in small training databases, shows a transparent advantage over other more complex methods. In future work, it'll be interesting to review the apparent asymptotic progression of the educational curve of a technique supported deep neural convoluted networks in order that you'll understand, once these behaviors are confirmed, what's the order size of the training data set to make sure optimal performance of this method. This may help to style applications where duplicate query detection is embedded, especially so as to direct a good collection of appropriate training databases of sufficient size.

Part-2) FUTURE PROSPECTS

1) The mechanism is quite an effective use case that helps websites like Quora and StackOverflow to effectively eliminate duplicate questions from their site, and provide a clean User Interface.

2) NLP's Duplicate question detection method is widely getting used in today's world even for sentiment analysis.

3) A large dataset can be used in future and algorithms like random forest classifier and XGB classifier to increase the accuracy of database.

4) By splitting the dataset into train and test unbiased and correct predictions are made upto 80%. This can further increase by performing more processing steps over the dataset.

5) Instead of just stating the prediction value by duplicate and not duplicate, we can give a probability range value between 0 and 1. This will be more precise in prediction purpose.

Chapter 6

REFERENCES

1] On Application of Natural Language Processing in Machine Translation of IEEE written by Zhaorong Zong, Changchun Hong on 6th Sept 2018.

2] Translation of natural language queries to structured data sources of IEEE written by Ruslan Posevkin, Igor Bessmertny on30th Nov 2020.

3] Information Processing and Retrieval from CSV File by Natural Language of IEEE written by Chalermpol Tapsai on 30th Dec 2018.

4] Research on Data Preprocessing and 3D Matrix Model written by Liu Hongling, Wan Di on 25th Oct 2020.

5] Pointer-Generator Abstractive Text Summarization Model with Part of Speech Features written by Shuxia Ren, Zheming Zhang on 20th Oct 2019.

6] ATC: An Automatic Text Comparison Tool Based on Diff Algorithm Search of IEEE written by Li Lixun, Wang Gaoshan, Dou Zengjie, Feng Yan on 20th March 2020.

7] Improve Quora Question Pair Dataset for Question Similarity Task of IEEE written by Huong T. Le, Dung T. Cao on 1st Sept 2021.

8] An Effective Crop Prediction Using Random Forest Algorithm of IEEE written by Dr. V.Geetha A. Punitha M. Abarna M. Akshaya Mishra on 4th July 2020.

9] Learning Profiles in Duplicate Question Detection of IEEE written by Chakaveh Saedi, Jo˜ao Rodrigues, Jo˜ao Silva, Ant´onio Branco on 6th Aug 2017.

10] Bengali Words Classification by Its Prefix Using Machine Learning Classifiers of IEEE written by K.M. Shahriar Islam, Sharun Akter Khushbu on 8th July 2021.

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my guide Mrs.Pranali Wagh, co-guide Mr. Santosh Rathod, our principal Mr. Bhavesh Patel and the faculty of Shah And Anchor Kutchhi Engineering College for this wonderful project on the topic 'Duplicate text-pair detection mechanism' which also helped me in doing a lot of Research. It had been a great journey in analysing, understanding and realizing the implementation of a real time project in frontend areas of stream-lit module of python and backend areas of machine learning algorithms in python. It helped me grow a lot not only academically but also in a broader and knowledgeable aspect.

I would like to express my very great appreciation to my group members for their valuable and constructive execution and suggestions during the planning and development of this research work and in making of the project. The team building, managing and co-operative skills were also been imparted in us. Hence making us more professional and patient in work