QUORA QUESTION PAIR SIMILARITY

Using ML, NLP , MLOPS

Guided by Innomatics Research Labs:
Sandhya
Abilash Patel Routhu

Presented by:

1. Avinash Durugkar
2. Kevin Dausi
3. Cinthiya M
4. Mani Teja
5. Chekuri Nikhil

INTRODUCTION

On Quora, users can ask questions and receive responses from a variety of users. Though the questions may frequently be formulated differently, the goal remains the same.

The ideal scenario is that after a question is posted, Quora would apply some sort of "technique" to identify a part of its current question from the database, and that part would comprise questions that are "similar" to or on the same topic as the newly posted question. Once this selection has been found, Quora will use machine learning to check if any questions in this selected subset are duplicates.

Goal : To predict whether a pair of questions are similar or not

PROBLEM STATEMENT

Identify which questions asked on Quora are duplicates of questions that have already been asked.

This could be useful to instantly provide answers to questions that have already been answered. We are tasked with predicting whether a pair of questions are duplicates or not.

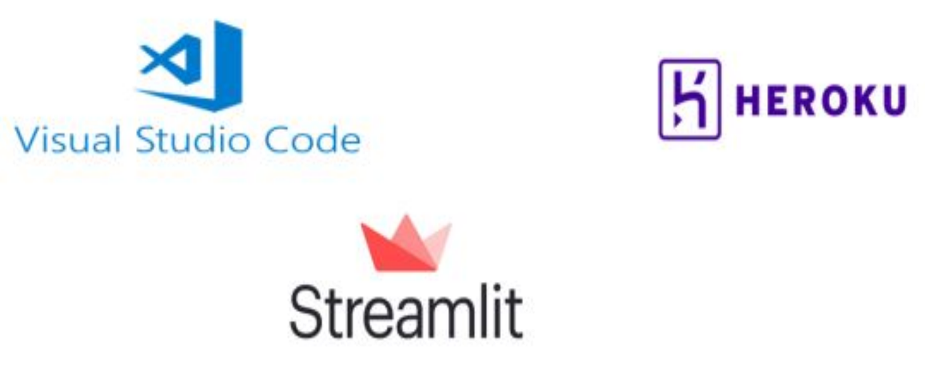
PROJECT REQUIREMENTS

HARDWARE REQUIREMENTS

- Processor i3 or more
- RAM 8GB or more

SOFTWARE REQUIREMENTS

- OS - Windows7 or more
- Language - Python
- Jupyter Notebook
- IDE - Vscode



DATA SET

The dataset is a collection of data. It can be a single table, or it can be many tables that are related to each other. Datasets are often used in machine learning and data mining applications.

Team working on Dataset provide by Innomatix Research Labs

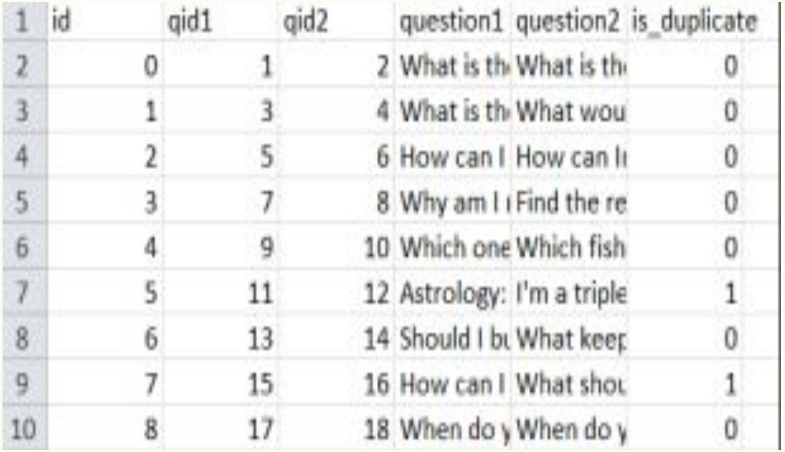
Resource: https://github.com/Koorimikiran369/Quora-Question-Pairing/blob/main/train.csv.zip

Data Overview

- Data will be in a file Train.csv
- Train.csv contains 5 columns

  qid1, qid2, question1, question2, is_duplicate

- Number of rows in Train.csv = 404,290



1	id	qid1	qid2	question1	question2	is_duplicate
2	0	1	2	What is th	What is th	0
3	1	3	4	What is th	What wou	0
4	2	5	6	How can I	How can I	0
5	3	7	8	Why am I (	Find the re	0
6	4	9	10	Which one	Which fish	0
7	5	11	12	Astrology:	I'm a triple	1
8	6	13	14	Should I bu	What keep	0
9	7	15	16	How can I	What shou	1
10	8	17	18	When do y	When do y	0

STEPS INVOLVED

Dataset and library

- Import the General libraries, NLP module, and Machine learning modules
- Load the dataset

Natural Language Processing (NLP)

- Text Preprocessing:
- Apply Tokenization
- Apply Stemming
- Apply POS Tagging
- Apply Lemmatization
- Apply label encoding
- Feature Extraction

STEPS INVOLVED PART - 2

Exploratory Data Analysis

- Data preprocessing
- Text to Numerical vector conversion

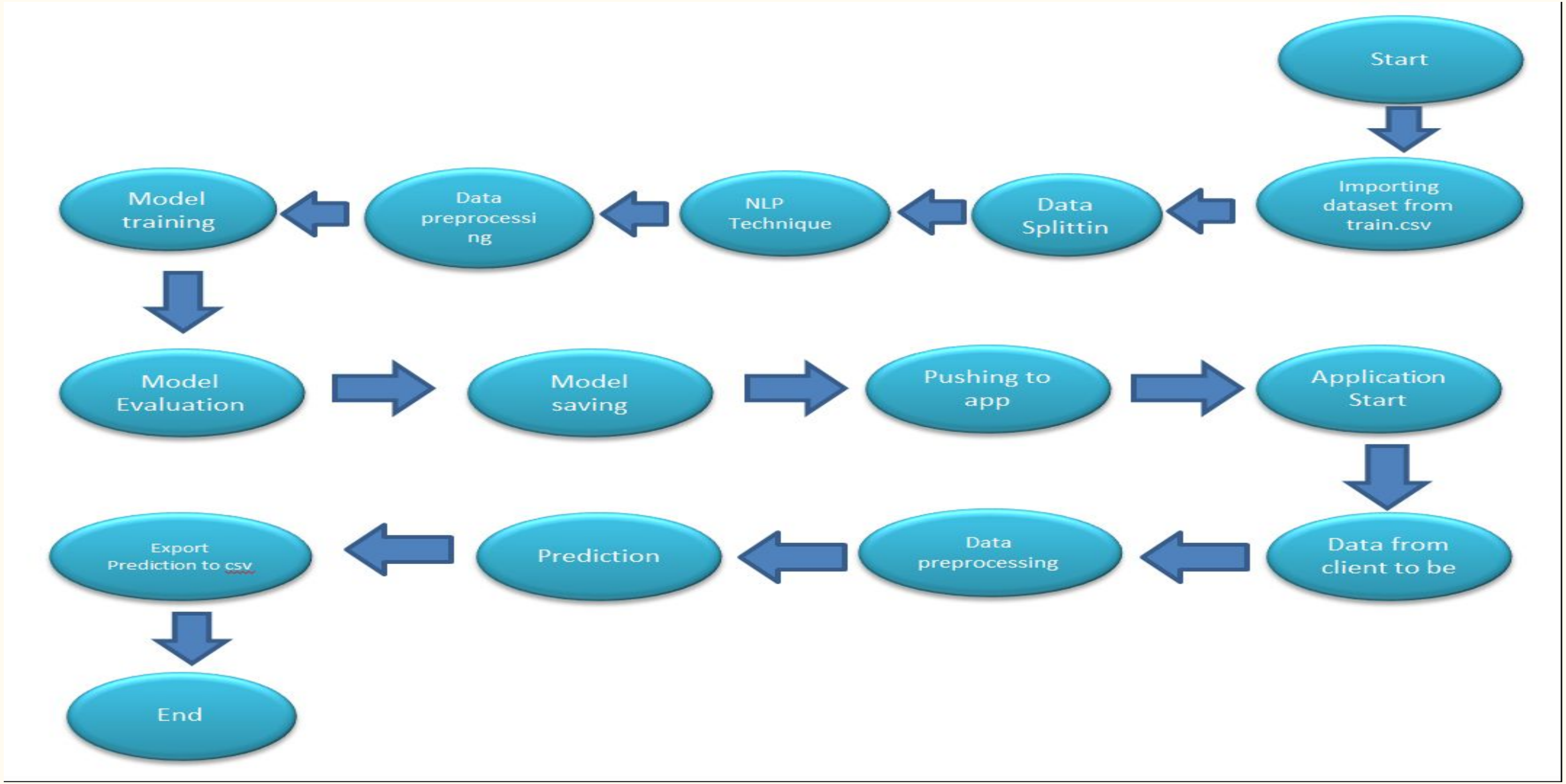
Machine Learning

- Model Building
- Evaluate the model

MLFlow

- Track your experiments with the help of MLFlow
- Break your code into production ready script
- Automation using Workflow Orchestration - Prefect (OPTIONAL)

ARCHITECTURE


Start
Importing dataset from train.csv
Data Splittin
NLP Technique
Data preprocessi ng
Model training
Model Evaluation
Model saving
Pushing to app
Application Start
Data from client to be
Data preprocessing
Prediction
Export Prediction to csv
End