INNOMATICS RESEARCH LABS

Quora Question Pair Similarity

Project Title	**Quora Question Pair Similarity**
Technologies	**ML, NLP, MLOPs**
Domain	**Social Networking**
Project Difficulties Level	**Advance**

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. Quora values canonical questions because they provide a better experience to active seekers and writers, and offer more value to both of these groups in the long term. The main aim of the project is 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.

Domain: Social Network

Prerequisites: Python programing language, Machine Learning, NLP and MLOPs

Project Difficulty Level: Advance

Real World/Business Objectives and Constraints:

- The cost of a mis-classification can be very high.
- You would want a probability of a pair of questions to be duplicates so that you can choose any threshold of choice.
- No strict latency concerns.
- Interpretability is partially important.

Tasks to perform:

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

- Load the dataset
- Text Preprocessing:
    - Removing html tags
    - Removing Punctuations
    - Performing stemming
    - Removing Stop words
    - Expanding contractions etc.
- Apply Tokenization
- Apply Stemming
- Apply POS Tagging
- Apply Lemmatization
- Apply label encoding
- Feature Extraction
- Text to Numerical vector conversion
    - Apply BOW
    - Apply TFIDF vectorizer
    - Apply Word2Vector vectorizer
    - Apply Glove
- Data preprocessing
- Model Building
- Evaluate the model
    - Confusion Matrix
    - Classification report

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

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

Datalink:

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

Type of Machine Learning Problem

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