Quora question Pair similarity

**Introduction**

In this project we will be dealing with the task of pairing up the duplicate questions from Quora. More formally, the followings are our problem statements.

* 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.

**Note-** we are talking about the semantic similarity of the questions.

**2. Problem Statement:**

The data is in a csv file named “Train.csv.”

* Train.csv contains 5 columns: qid1, qid2, question1, question2, is\_duplicate
* Size of Train.csv — 60MB
* Number of rows in Train.csv = 404,290
* ‘qid1’ and ‘qid2’ are the ids of the respective questions, ‘question1’ and ‘queston2’ are the question bodies themselves and ‘is\_duplicate’ is the target label which is 0 for non-similar questions and 1 for similar questions.
* This can be also thought as if ‘qid1, qid2, question1, question2,’ are the x labels and ‘is\_duplicate’ is the y labels.

3. Dataset:

The project utilizes a labelled dataset consisting of question pairs from Quora. Each question pair is labelled with a similarity score indicating whether the questions are similar or Not similar. The dataset is split into a training set, a validation set, and a test set.

4. Approach:

The system employs a machine learning model trained on the Quora question pairs dataset. The model uses natural language processing (NLP) techniques to extract meaningful features from the questions and learn patterns that determine similarity.

5. Feature Extraction:

- Pre-processing: The questions are pre-processed by removing stop words, punctuation, and converting text to lowercase.

- Word Embeddings: Word embeddings like Word2Vec or GloVe are used to represent words as dense vectors capturing semantic relationships.

6. Model Training:

- Feature Combination: The extracted features are combined to form a feature vector representing each question pair.

- Model Selection: Several machine learning models, such as logistic regression, support vector machines, or neural networks, can be trained on the feature vectors.

- Hyperparameter Tuning: The model's hyperparameters are optimized using techniques like grid search or random search to improve performance.

7. Model Evaluation:

The trained model is evaluated using appropriate evaluation metrics such as accuracy, precision, recall, and F1 score. The model's performance is measured on the validation set to select the best model for deployment

8. Model Deployment:

Once the best model is selected, it can be deployed to a production environment where it can be integrated with the Quora platform. The system can provide real-time predictions on the similarity of question pairs, helping users and moderators identify potential duplicates.

9. Performance Monitoring:

It is crucial to monitor the performance of the deployed system over time. Regular evaluations can be performed using a holdout dataset or by collecting user feedback to ensure the system's accuracy and reliability.

10. Conclusion:

The Quora Question Pair Similarity project provides an effective solution for determining the similarity between question pairs on the Quora platform. By accurately identifying similar questions, the system enhances user experience and improves the quality of information available to Quora users.

Please note that this is a general outline of a project documentation for the Quora Question Pair Similarity task. The specific implementation details and algorithms may vary based on the preferences and choices made during the project development.