**Quora Duplicate Questions**

**Data Available :** Pair of questions with flag of duplication or not

**Feature Engineering** : Both the questions were processed with Punctuation removal, conversion to lower, removal of stopwords, stemming and stripping the Space, and the corpus was created.

1. Set of basic features were created for both the given questions :
   1. Total word count in the original set of questions
   2. Total word count in the reformed set of questions, after above processing
   3. Total unique word count in the original set of questions
   4. Total unique word count in the reformed set of questions
   5. Total character count in the original set of questions
   6. Total character count in the reformed set of questions
   7. Total count of the unique common words in both the questions
   8. Percent of the unique common words in both the questions

The above set of features were put in the XGBoost model and predictions were made on the test data, after computing the same set of features on test data.

1. Set of Fuzzy words features :
   1. Partial Ratio
   2. Partial Token Set Ratio
   3. Partial Token Sort Ratio
   4. QRatio
   5. WRatio
   6. Token Set Ratio
   7. Token Sort Ratio

The above set of features were put in the XGBoost model along with the above basic set of features and predictions were made on the test data, after computing the same set of features on test data.

**Model Selection** : XGBoost model has been selected for the current purpose. The problem has been classified as a classification problem, and further features are created for it. XGBoost works well for such problem cases.

**Hyper-parameter Tuning** : A default set of parameters has been taken into account in the first iteration, and the hyper parameters have been further tuned by following a Grid Search based approach. The parameters tuning have been done on the basis of the priority and significance of each of the parameters in the model. The grid values chosen by the model have been done with the aim of maximizing the roc-auc. (Further change of the metric to log-loss could be done!)

**Additional Steps** **for better accuracy** :  
1. More features creation : TF-IDF set of features and word2vec set of features (in-progress, will be updated on github profile)  
2. Different models : As the data is enough to train a deep learning model, Deep Siamese LSTM networks could be used for generating a text similarity