# QUORA QUESTION PAIRS: IDENTIFYING QUESTIONS WITH SAME INTENT

## Project Overview

The goal of the project is to build a binary classification model using a simulated dataset containing a pair of questions and a binary class label stating whether a pair is duplicate or not. In this project, I will be handling this problem by applying advanced techniques (Random Forest, K-Means, SVM, XGBoost etc.) to classify whether question pairs are duplicates or not. After applying several models, I'll be comparing the accuracy obtained with each model.

Doing so will make it easier to find high quality answers to questions resulting in an improved experience for Quora writers, seekers, and readers.

**Datasets and Inputs:**

I'll be using the Quora dataset provided on the Kaggle Competition ( ~ 4,00,000 records) : https://www.kaggle.com/quora/question-pairs-dataset

**Features:**

- id - the id of a training set question pair – (Numeric)
- qid1, qid2 - unique ids of each question (only available in train.csv) – (Numeric)
- question1, question2 - the full text of each question – (String)

**Target Variable:**

- is_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise. – (Numeric)

**Environment Used:**

The project will be written in Python 3.5 and the following libraries will be used:

- Pandas
- Numpy
- Scikit-Learn
- MatplotLib
- XGBoost
- Seaborn

**Link to Capestone Proposal Review:** https://review.udacity.com/#!/reviews/1286079