**MACHINE LEARNING ENGINEER NANODEGREE** RITU YADAV

**CAPESTONE PROPOSAL** DATE: 17-06-2018

# QUORA QUESTION PAIRS: IDENTIFYING QUESTIONS WITH SAME INTENT

## Domain Background:

Quora is a place to gain and share knowledge—about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world.

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.

## Problem Statement:

The goal 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

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)

## Solution Statement:

Given the dataset, I will train several classification models and ensemble methods such as K-Nearest Neighbors, Logistic Regression, SVMs, Random Forest, XGBoost and evaluate each of them to find the best suitable model for solving the above stated problem. And I will further try to optimize the best model selected by tuning parameters to predict the pair of questions with same intent.

## Benchmark Model:

Benchmarking here means, a standard solution which already performs well. Thus,

1. First, I will keep a very naïve model as my benchmarking model: x% change of getting a duplicate question pair, which is , nothing but the percentage of total duplicate pairs in the dataset i.e., (Number of duplicate pairs in dataset)/ (Total number of records in dataset) * 100

2. Second, keeping in mind that currently, Quora uses a Random Forest model to identify duplicate questions. So, first I will implement Random Forest Model to identify the duplicate questions and then I will use that model as a benchmark model and by applying several different models listed above, I will try to either cross or at-least match the benchmark model's accuracy.

## Evaluation Metrics:

Our goal is to predict the probability that the questions are duplicates (a number between 0 and 1 for each ID in the test set. The results obtained will be evaluated on the log loss between the predicted values and the ground truth.

For handling the stated problem, I will try to minimize the log loss obtained for each model implemented. The model having the least log loss value will be evaluated as the best model in order to tackle this problem.

Also, one of the best ways to evaluate the performance of the benchmark model and the solution model is to measure the accuracy using cross validation. For this, I will construct a confusion matrix to show the prediction results.

## Project Design:

The whole project flow is outlined below:

1. Environment Used:

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

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

2. Data Exploration:

   The statistics of both the training and testing data will be explored. The given dataset will be observed carefully and also I will be identifying the most common word occuring in the training dataset.

3. Data Preprocessing and Feature Visualization:

   After exploring the dataset,

   I will construct some the features of the training set such as,

   - length of the question (number of characters in the question)
   - number of words in the question
   - Ratio of word share between the questions

   As part of data preprocessing part, I will be calculating the TF-IDF word share between the question pairs.

   Afterwards, I will understand the characteristics of the dataset by visualizing each of the features mentioned above using matplotlib library.

4. Model Building and Performance Evaluation

   Under this section, I will implement different classification and ensemble methods such as K-Nearest Neighbors, Logistic Regression, SVMs, Random Forest, XGBoost and then will select the best model by evaluating each of the performance matrix. Further I will try to optimize the best model selected by tuning certain parameters.

---

References:


1. http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
2. https://www.kaggle.com/quora/question-pairs-dataset
3. https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
4. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html