1. Jyothi Vishnu Vardhan Kolla
2. Sanjay Prabhakar

## Abstract:

Every public discussion forum has many questions that are asked very frequently. Most of these questions are similar or redundant and effort must be put to answer these similar questions multiple number of times, there should be a way to map a new question to an existing question and retrieve the answer already provided to the existing question.

## Introduction:

In this project, we are planning to build a machine learning model that can Identify whether a new question that is being asked is same as an already existing question, the problem is formulated as a machine learning problem as follows where the dataset that is being used to train the model will have examples of questions and a binary variable describing whether the questions are identical or not, and once the model finishes its training, it will have the ability to identify whether a new incoming question is same as the question that is already in the database.

The solution to this problem will be very useful for applications like chatbots where the chatbot can match the question from a database of questions and retrieve the existing answers for the question. It can also be used in applications such as piazza where there will be so many repetitive questions asked, and the instructors must put an extra effort to answer similar questions multiple times which can be tackled by training the model using the data from piazza.

## Proposed Project Specifics

### Problem Formulation:

We have mapped out the problem we want to solve as a classification machine learning problem, the output of the problem will be a binary value either ‘0’ or ‘1’, where ‘0’ indicates the questions are not similar and ‘1’ indicates that the questions are similar.

### Dataset used and working of the model:

The dataset we are using to build the model for our problem statement is Quora Question Pairs dataset from Kaggle which contains a total of 404,290 questions pairs which is basically 404,290 rows and it has five attributes which are as follows:

1. **ID:** A unique identifier to identify each row.
2. **Question-1:** The first question in the pair.

3. **Question-2:** The second question in the pair.
4. **Qid1:** A unique identifier for the first question in the pair.
5. **Qid2:** A unique identifier for the second question in the pair
6. **Is_duplicate:** A binary label indicating whether the two questions in the pair are semantically equivalent or not. A value of 1 indicates that the questions are equivalent, and a value of 0 indicates that they are not.

**Feature Extraction:**

The features are obtained using some of the standard nlp techniques such as Bag of words with n-grams and TF-Idf Vectorizer to convert the text in the questions into numerical vectors after removing unwanted characters in the text such as punctuations, html tags during data preprocessing phase.

**Conclusion:**

Finally, the workflow of the project is as follows, firstly the model is trained using the Quora question pairs dataset by experimenting with several machine learning algorithms and picking the one with the best performance using metrics such as Accuracy, Precision, Recall, F1-score, Log-loss, and then during the testing phase the given test data point is compared with all the questions in the database choose the match with probability score.