Quora Question Pairs

METHODIST

COLLEGE OF
ENGINEERING & TECHNOLOGY
O.U.

Group Number: **13**

Project Members:-

- **Sukanya Behera [160714733008]**
- **Chandrabhatta Sriram [160713733042]**
- **Abul Faiz Mohammed Abdul Hadi Mouzzam [160714733108]**

Project Guide: **E. Shailaja**
Asst. Professor (Dept. of CSE)

Abstract

Quora is a question-and-answer site where questions are asked, answered, edited and organized by its community of users. 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(s), and make writers feel they need to answer multiple versions of the same question.

Quora has publicly released the data set to mitigate the inefficiencies of having duplicate question pages at scale. Which gives us our problem statement: An automated way of detecting if pairs of question text actually correspond to semantically equivalent queries. Previously, Quora uses a Random Forest model to identify duplicate questions.

The goal is to predict which of the provided pairs of questions contain two questions with the same meaning. The ground truth is the set of labels that have been supplied by human experts. The ground truth labels are inherently subjective, as the true meaning of sentences can never be known with certainty. Human labelling is also a 'noisy' process, and reasonable people will disagree. As a result, the ground truth labels on this dataset should be taken to be 'informed' but not 100% accurate, and may include incorrect labelling. The labels, on the whole, represent a reasonable consensus, but this may often not be true on a case by case basis for individual items in the dataset.

Steps Involved:-

1. **Exploratory Data Analysis [EDA]**: EDA is an approach to analysing data sets to summarize their main characteristics, often with visual methods.
2. **Dimensionality Reduction & Visualization**: In Machine Learning and Statistics, Dimensionality Reduction is the process of reducing the no. of Random Variables under consideration, via obtaining a set of Principal Variables.
3. **Classifying Data Using Classification & Regression Models**: k-Nearest Neighbours, Naïve Bayes, Logistic Regression, Linear Regression & Stochastic Gradient Descent.

Data Type: .CSV [Comma Separated Values] files. **Data Size**: 130MB.

- Train data: train.csv (id, qid1, qid2, question1, question2, is_duplicate).
- Test data: test.csv (id, qid1, qid2, question1, question2).
- Total number of records in train data: 404351.

Approx. Project Development Inception Date: 1st or 2nd week of February, 2018.

Approx. Project Deployment Date: 4th week of March, 2018 or, 1st week of April, 2018.

References:-

Dataset Source: https://www.kaggle.com/c/quora-question-pairs

https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning