**Course: Introduction to Machine Learning and Data Mining**

**Project Proposal: “Sentence to Sentence semantic similarity”**

Name: Mohammed Bilal Ansari (Individual Project)

**Objective:**

To identify the Quora question pairs that have the same intent but have been phrase in a different ways. In this project, we will try to predict which of the provided pairs of questions contain two question with similar meaning.

**Data-set:**

We will use “Quora Question Pairs” data-set on Kaggle for this project. Set of labels in this dataset have been supplied by human experts. Human labeling is also a 'noisy' process because true meaning of sentences can never be known with 100% certainty. It may include few incorrect labeling in the dataset.

Data fields

* id - the id of a training set question pair
* qid1, qid2 - unique ids of each question (only available in train.csv)
* question1, question2 - the full text of each question
* is\_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

[Dataset Link](https://www.kaggle.com/c/quora-question-pairs/data)

**Proposed Approach:**

We will try to solve this problem using the Neural Network. The idea is to train the neural network using the work embedding on given question pairs. Dense embedded vectors will be learned for questions using either CBOW/Skip-Gram model. After getting the dense vector representation of question pairs, we will feed them into our neural network for classification whether the pair of question is duplicate or not and then will try minimize the loss function using Back-propagation.