# Project Proposal

**Topic**: Identify question pairs that are duplicate i.e have same intent.

**Overview:** Goal is to predict which of the provided pairs of questions have the same meaning. Such problem is prevalent in online Q&A forums like Quora and Stack Overflow. Combining answers for duplicate questions will provide improved experience for users.

**Dataset:** Dataset is obtained from Kaggle.com. The dataset contains csv files already divided into test and training sets, and each files contains the following columns:

- **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.

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

**Techniques:** We plan on using deep neural network based LSTM models for solving this problem. LSTM are known to work very well with temporal data. This is applicable to the given problem as the given sequence of the words in the questions are useful factor in determining if the questions are duplicate or not.

**Team:**
Nikhil Lahoti (012448514)
Vedashree Bhandare(012416924)