NLP Project Proposal:

Mohit Nihalani (mn2643)
Rohit Saraf (rs6785)
Sushanth Samala (ss12852)

Problem Statement: Quora Question Pairs

On Quora, 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.

In this project, we want to tackle this natural language processing problem by applying advanced techniques to classify whether question pairs are duplicates or not. To summarize the problem we are trying to Identify which questions asked on Quora are duplicates of questions that have already been asked.

In this project, we are planning to experiment with various text preprocessing, feature engineering techniques and comparing the performance of various machine learning algorithms

Text Processing and Feature Engineering:

1. Removing stopwords, HTML tags
2. TFIDF weighted unigram and bigram feature vectors
3. Word2Vec embeddings using Glove
4. Stemming
5. Replacing Nouns with Named Entity Recognition
6. Distance Features such as word movers distance, cosine similarity, Euclidean distance using word vectors of pretrained glove embeddings and TF-IDF weighting
7. Fuzzy ratios, common word ratio

Machine Learning Models for Training and Testing:

1. Logistic Regression
2. SVM
3. XgBoost
4. Random Forest
5. LSTM
6. GRU

Performance Metrix: Log Loss

Dataset: Quora Dataset