**CS434 Proposal**

**Team Composition**

Michael Lee, leemi

Alex Nguyen, nguyalex

**Competition:**

Quora Question Pairs - https://www.kaggle.com/c/quora-question-pairs

**Method**:

Python and Keras as a platform

**Data Input and Prep:**

Pandas to import all of our data. Prep our sentences by first elimination all of the unneeded or “useless” english words like “the”, “and”, etc. We want to obtain the minimum sentences that will still convey the original meaning of the questions.

**Data modeling:**

There are a lot of pre-trained models that seem to cover natural language processing [1]. The approach would be to build up upon one of these pre-trained models using our specific criteria and improving the accuracy we get from the model. These are some of the more promising possibilities that we’ve been able to find: Google’s word2vec [2], GloVe word vectors[3], and Facebook’s fastText[4]. Essentially these models are pre-trained word vectors that are used to distinguish words of similar meanings, and we’ll be using these models to determine similar meanings between our two questions. In all, our approach is to use one of the models to convert each Quora question to word vectors. Using the word vectors we will use a Convolutional Neural Network to check if the two input word vectors are similar. If their similarity passes a threshold, it will predict that it is a duplicate. Our plan is similar to a paper we found from the discussions on kaggle [5].

**Questions:**

* Any advice on how to tackle the problem?
* How should we determine the best way to utilize our data?
* If the competition ends early, should we be submitting our answers before the competition deadline?
* What if our methodology fails, how will we be graded on the

**Reference Links:**

1. <https://www.kaggle.com/c/quora-question-pairs/discussion/30286>
2. <https://code.google.com/archive/p/word2vec/>
3. <https://nlp.stanford.edu/projects/glove/>
4. <https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md>
5. <https://aclweb.org/anthology/K15-1013>

**Helpful Kaggle Discussion links**

1. <https://www.kaggle.com/fernandocanteruccio/quoras-question-pairs-modeling-notebook/comments/notebook>
2. <https://www.kaggle.com/c/quora-question-pairs/discussion/30411>
3. <https://www.kaggle.com/c/quora-question-pairs/discussion/30340>