CQA Duplicate Detection

Team Name: Ultron

Pulle Sri Satya Sravya
Sushmitha Rajeswari Muppa
Moneesh Shashank

Our understanding of the project:

The goal of the project is to detect duplicate questions or questions with same intent from the dataset provided. The main aim is to detect whether two sentences deliver the exact same meaning or not. For example, consider the below questions:

Q1:

1. What would a Trump presidency mean for current international master’s students on an F1 visa?
2. How will a Trump presidency affect the students presently in US or planning to study in US?

These two questions have the same answer, just the way of asking is different. So, they are classified as duplicate questions.

Now, consider the following questions:

Q2:

1. What are the laws to change your status from a student visa to a green card in the US, how do they compare to the immigration laws in Canada?
2. What are the laws to change your status from a student visa to a green card in the US? How do they compare to the immigration laws in Japan?

These two questions are different because one deals with comparison between US and Canada, while the other deals with comparison between US and Japan.

Approach 1:

1. Identify named entities in both the sentences and do POS tagging to both of them
2. Compare named entities, nouns, adjectives, adverbs etc. of the two sentences.
3. Give a similarity score based on something like how many nouns/adjectives are similar etc.

Baseline we choose to implement:

We will be implementing NGram matching and neural network as baselines.

NGram matching:

Ngram match between the two questions using a similarity score like Jaccard's. If the similarity is above a threshold value, consider it a duplicate.
This works well only if:

1. Same words are used but ordering of words in the sentences changed
2. Same words are present in both the sentences but there are some grammatical errors.

   This algorithm sometimes mis-classifies non-duplicates as duplicates if the sentences are similar except for one/two important differences. This happens in the Q2 described in the first section.

Neural Network Approach:

- A basic neural network whose inputs are two vectors of questions and output is the score of similarity. Each vector is the average of word embeddings in the question.

Neural networks generally perform better than many other approaches if we give them a large proper training data. We will be implementing a basic neural network model for duplicate detection.

Bayesian networks can also be used to solve this problem since we have large training data. But, we have already implemented Bayes a lot of times before, so we didn't pick Bayes this time.

Dataset we will be working on:

The dataset we'll be using is in
https://www.kaggle.com/quora/question-pairs-dataset