CS 410 Project Documentation

Yiao Ding (Captain, Team Leader)

yiaod2@illinois.edu

1. **How to use the software**
   You should see three docs in the repo: projectModel.py, test.txt as well as train.txt. The software is just the projectModel.py, and test.txt / train.txt are just some input data which will be processed within projectModel.py. To successfully run the software (which is projectModel.py), you need to first install all required dependencies, including numpy, torch, sklearn and gensim. After that, like you run every other Python file, you can run the projectModel.py and you will see the result printed in the console.

2. **How the software is implemented**
   First the goal of this project is to create a machine learning model so that it can predict the similarity between sentences, note that the sentence to be judged here will be in the form of subject, predicate, and object. We use convolutional neural network for our model. The model requires training data, so what we did here is to use Word2Vec technique to transform each word from subject, predicate and object to vectors so that each sentence can be represented as a matrix, specifically, we used “twitter 50” from Gensim to turn each word into a vector of length 50. After obtaining those input data, we pass it to our model. There are total 2 convolution layers utilized in our model, and each of them has a kernel with size of 3 by 3. After convolution we add a rectified linear unit activation function. Also, there is a 3-max pooling layer after these 2 convolution layer. This is the general architecture of the model. Beside this, to implement this software some parameters also required to be set. We use 64 as the training data batch size, and within the pooling layer and convolution layer we set padding equal to one. We also set learning rate to 0.001 and set the number of epochs as one. In conclusion, with the general architecture and detailed parameters above, we are able to implement this software to make a prediction of the similarity between two sentences.