

Report

The most important step in this assignment is the preprocess. Our team figure out a few ways to maintain a good preprocess.

<1>Remove the stop words: Since stop words are nonemotional, which in this assignment will be treated as noise, so it's important to remove them. The way we determine the stop words are those words given in the python file, as well as those words, have a low correlation coefficient with emotional words.

<2> Remove the punctuation: Punctuation will also affect the accuracy of this assignment. Since that punctuation will mislead the machine, we have to remove them to get a better message for the machine.

After that, we think we should start to focus on the how to build a neural network to train this data well. First, we start to use one layer LSTM, after we check all the available resources from the Internet, we have a draft out and results percentage is about 80, then we start to modify this one layer LSTM to two layers because we want to improve the results.

However, the training data results are very nice, but it tends to show a serious overfitting in the test data, which make us very confused and we try to use L1 regularization but overfitting doesn't be solved. Therefore we have to shift to the initial version 1-layer LSTM and try to modify the variables like max-words or batch size to get a better result.