Description

For data preprocessing I used padding of the sentences for the maximum length of 300 as per suggested by BERT. Then I converted the given training and validation files into lower cases as I used lowercase pre-trained BERT versions. After that I did the fine tuning of the model on the GPU. It took approximately 15 minutes for the model to get fine tuned on the GPU. As per BERT Suggestion I used 32 batches and 10 epochs.
I manually took all the tags from the data set and created a dictionary for that. After that I gave sentence numbers to each sentence that will help later in combining the words into a sentence. Also, gives names to the columns like 'Word', 'Tags' and 'Sentence #'. Then I wrote two methods 'SentenceGetter' and 'tokenize_and_preserve_labels'. **SentenceGetter** makes two arrays for words and tags each and **tokenize_and_preserve_labels** gives labels to unlabelled entities like **(**, **)** or others. After that I tried to removes **##** that gets generated by the BERT model itself during training and fine tuning. Then I group by the sentences, words using spaces and labels using comma. Lastly I test the models and obtained accuracy and F1 score.