Name: Rutuja Bhandigani

Task 1: Simple Bidirectional LSTM model

1. Task 1 involves creating a Bidirectional LSTM model for Named Entity Recognition (NER).
2. The data files (train, dev/validation, and test) are read and converted into a list of lists, where each list represents a sentence, and each word is a list element.
3. BiLSTM network architecture:
Embedding Layer: vocab size * embedding_dim = 30292 * 100
LSTM Layer: embedding_dim = 100, hidden_dim = 256, num_lstm_layers = 1, bidirectional = True
Dropout: 0.33
Linear Layer: input_dim= 512 (hidden_dim * 2(bidirectional)), linear_out_dim = 128
ELU Layer: alpha = 0.01
Classifier Layer: input_dim = 128

4. Data loader and collator functions are used to manage variable-length sentences and pad them with the maximum length sentence in the batch. A dictionary is created to store all the words and NER tags in the dataset (train, dev and test), with each word and tag assigned a unique index.
5. Class weights are used to address imbalanced class issues during training ('O' NER-tag was overpowering the model, initially predicting every word as 'O'). Class weights are calculated as the total number of labels divided by their respective class frequency, multiplied by a hyperparameter, and taking the log of scores less than 1.0.
6. The model is trained for 200 epochs using Cross Entropy loss and SGD optimizer, with class weights passed to the Loss Function and a learning rate of 0.1 and momentum of 0.9 in SGD.
7. We use the best-performing models for validation and testing on separate datasets.

HyperParameters:
Embedding Dimension = 100
Hidden Dimension = 256
Linear Output Dimension = 128
Bidirectional = True
Dropout = 0.33
Number of LSTM layers = 1
Batch Size = 4
Learning Rate = 0.1
Momentum = 0.9
Epochs = 200

Using the evaluation script on pred1.out:

Accuracy: 95.45%
Precision: 79.40%
Recall: 75.36%
F1: 77.33

```
!perl conll03eval.txt < pred1.out

processed 51578 tokens with 5942 phrases; found: 5640 phrases; correct: 4478.
accuracy:  95.45%; precision:  79.40%; recall:  75.36%; FB1:  77.33
              LOC: precision:  86.72%; recall:  82.47%; FB1:  84.54  1747
             MISC: precision:  79.96%; recall:  78.74%; FB1:  79.34  908
              ORG: precision:  73.72%; recall:  69.87%; FB1:  71.75  1271
              PER: precision:  75.85%; recall:  70.58%; FB1:  73.12  1714
```

Task 2: Using GloVe word embeddings

1. Download the GloVe word embeddings with 100 dimensions.
2. Create an embedding matrix using the Vocab Dictionary that we created in Task 1.
3. Since the GloVe word embeddings only contain lowercase words, we need to handle uppercase words by adding a small value (5e-3) to the values for their lowercase counterparts. This is because uppercase words are often similar to their lowercase counterparts.
4. Load the embedding matrix into an embedding layer in a new class called BiLSTM_glove. This class has the same architecture as the BiLSTM class from Task 1, except that it takes an additional parameter embedding_matrix which is used to load the embedding matrix into the embedding layer.
5. Train the BiLSTM_glove model using the same hyperparameters as in Task 1, except for the batch size which is set to 8. We train the model for 50 epochs.
6. Test the trained BiLSTM_glove model on the validation and testing datasets.

Hyper Parameters:
Embedding Dimension = 100
Hidden Dimension = 256
Linear Output Dimension = 128
Bidirectional = True
Dropout = 0.33
Number of LSTM layers = 1
Batch Size = 8
Learning Rate = 0.1
Momentum = 0.9
Epochs = 50

Using the evaluation script on pred2.out:
Accuracy: 98.06%
Precision: 89.50%
Recall: 89.97%
F1: 89.74

```
!perl conll03eval.txt < pred2.out

processed 51578 tokens with 5942 phrases; found: 5973 phrases; correct: 5346.
accuracy:  98.06%; precision:  89.50%; recall:  89.97%; FB1:  89.74
              LOC: precision:  93.61%; recall:  94.18%; FB1:  93.89  1848
             MISC: precision:  82.19%; recall:  83.08%; FB1:  82.63  932
              ORG: precision:  84.42%; recall:  84.04%; FB1:  84.23  1335
              PER: precision:  92.73%; recall:  93.54%; FB1:  93.14  1858
```