學號：B05902042 系級：資工二 姓名：林瑋毅

1. (1%) 請說明你實作的 RNN model，其模型架構、訓練過程和準確率為何？
   (Collaborators:)

   答：

我實做的 model 架構如下：

```
LSTMClassifier(
    (embedding): Embedding(46596, 128)
    (lstm): LSTM(128, 128, num_layers=2, bidirectional=True)
    (fc): Sequential(
        (0): Linear(in_features=256, out_features=2)
    )
)
```

首先將各個字 embedding 到一個 128維的空間，再進行 bidirectional 的 LSTM（ hidden dimesion 為 128）選取兩個方向最後出來的向量並 concatenate 起來，形成一個 256 維的向量，再用一層全連接層到兩個 output，分別代表 positive 及 negative，最後用 softmax 得出分別的機率，並選取較大者最為預測答案。

訓練過程使用 Adam (Pytorch 預設參數)，batch size 為 128，訓練 4 個 epoch。

另外，preprocess 的部分包含濾除除了問號、驚歎號外的標點符號、全部換成小寫字母，word embedding 使用 gensim 的 word2vec，參數為 iter=10 ，其餘預設參數。

準確率為：

private / public = 0.82370 / 0.82542

2. (1%) 請說明你實作的 BOW model，其模型架構、訓練過程和準確率為何？
   (Collaborators:)

   答：

我實做的 model 架構如下：

```
BOW(
    (fc): Sequential(
        (0): Linear(in_features=46596, out_features=128, bias=True)
        (1): ReLU()
        (2): Linear(in_features=128, out_features=32, bias=True)
        (3): ReLU()
        (4): Linear(in_features=32, out_features=2, bias=True)
    )
)
```

首先將句子依照各個字出現的次數，製作成一個 bag of word 的向量，再丟到如上所示

的 deep neural network，最後有兩個 output，分別代表 positive 及 negative，並且用 softmax 得出分別的機率，並選取較大者最為預測答案。

訓練過程使用 Adam (Pytorch 預設參數)，batch size 為 128，訓練 4 個 epoch。

另外，preprocess 的部分包含濾除除了問號、驚歎號外的標點符號、全部換成小寫字母。

準確率為：

private / public = 0.78357 / 0.78612

3. (1%) 請比較bag of word與RNN兩種不同model對於"today is a good day, but it is hot"與"today is hot, but it is a good day"這兩句的情緒分數，並討論造成差異的原因。
   (Collaborators:)

   答：

RNN（使用第一題所述的模型）：0.5016 / 0.9307
BOW（使用第二題所述的模型）：0.9128 / 0.9128

可見 RNN 能抓到句子的前後文關係，所以知道第一句與第二句相比沒有那麼 positive，而 BOW 因為只有考慮句子出現什麼字，因此第一句與第二句的分數會相同，無法透過前後文關係發現兩者的不同。

4. (1%) 請比較"有無"包含標點符號兩種不同tokenize的方式，並討論兩者對準確率的影響。
   (Collaborators:)

   答：

有標點：private / public = 0.82370 / 0.82542
無標點：private / public = 0.82103 / 0.82255

兩者皆使用與第一題同的模型、方法、訓練參數，可以發現有標點符號的準確率稍微高一點，但幅度不大，推測原因為標點符號可以提供更多的資訊（例如：驚歎號可能是更強的情緒、問號是不確定的語氣等）因此可以稍微增加準確率。

5. (1%) 請描述在你的semi-supervised方法是如何標記label，並比較有無semi-surpervised training對準確率的影響。

(Collaborators:)

答：

此題使用與第一題相同的模型、方法、訓練參數，不同的是，用 label data 訓練完之後，先幫所有 unlabeled data 用當前的 model 標記，並且將 threshold 設為 0.9 及 0.1，再用這組資料多 train 1 個 epoch（optimizer 用 SGD(lr=1e-4, momentum=0.9，epoch 數量由 validation 的結果選擇）。此模型的準確率如下：

private / public = 0.82172 / 0.82405

不做 semi-supervised 的準確率為：

private / public = 0.82370 / 0.82542

發現結果反而略微衰退，不做 semi-supervised 反而比較好，此外，如果用 unlabeled data 繼續 train 下去，validation 的結果會更糟糕，因此推測用這個方法沒辦法會使模型更好，反而會加大 overfit 的程度。