HW5 Report

學號：b04901009 系級： 電機三 姓名：林孟瑾

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？
   答： 將字串先用 gensim 的 Word2Vec 轉成 vector 後，擷取出此 Word2Vec 的 model 中的 embedding 層，再經過第一層 LSTM（units 數為 512）和 Dropout(0.25)層，然後第二層 LSTM(units 數為 256) 和 Dropout(0.25)層，接著進入三層 DNN，第一層 Dense 的 units 數為 256，第二層 Dense 的 units 數為 128，第三層 Dense 的 units 數為 1 且 activation 為 sigmoid。訓練 model 時，用的 loss 是 binary_crossentropy，optimizer 是 adam，epoch 為 5。最後出來的準確率 public 為 0.82693，private 則是 0.82523。

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？
   答：

將字串用 keras 的 Tokenizer()中的 text_to_matrix（mode 為 tfidf）轉成 bag of words 的形式之後，進入一層 Dense(units 數為 256)，再進入 units 數為 1 且 activation 為 sigmoid 的 Dense。訓練 model 時，用的 loss 是 binary_crossentropy，optimizer 是 adam，epoch 為 5。最後出來的準確率 public 為 0.79530，private 則是 0.79562。

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：

	第一句	第二句
Bag of word	0.67	0.68
RNN	0.31	0.89

Bag of word 處理字串時不會加入 word 之間順序的因素，所以兩句話的情緒分數相差不大。而 RNN 會考慮 word 之間的順序，因此處理此兩句話的情緒分數相差很大。

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。 答：

從結果中可以看出，包含標點符號的準確率較高，因為標點符號能增加訓練 RNN 的 data，讓 model 學習辨認情緒的效果變好。

	Private score	Public score
有標點符號 ， 用 gensim word2vec	0.82523	0.82693
沒標點符號 ， 用 gensim word2vec	0.81680	0.81770

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。

答：threshold 為 0.1，預測值大於 0.9 的 data 才標為 1，然後小於 0.1 的 data 才標為 0。用 semi supervised 的方法訓練出的 model 比原本的 model 準確率稍低，原因為有些 data 可能在標記 label 時預測錯誤，導致準確率變低。

	Private score	Public score
No semi ， 用 keras Tokenizer()	0.80200	0.80271
Semi ， 用 keras Tokenizer()	0.79854	0.79828