學號:R06945003 系級： 生醫電資碩一 姓名：林鈺盛

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: NO)

   答：

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 39, 256)           16793856
_________________________________________________________________
bidirectional_1 (Bidirection (None, 128)               394240
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 129
_________________________________________________________________
activation_1 (Activation)    (None, 1)                 0
=================================================================
Total params: 17,188,225
Trainable params: 394,369
Non-trainable params: 16,793,856
_________________________________________________________________
```

以上為我的 model 架構，我的 w2v 是用了全部的 data(train_label ,train_nolebel, test)來 train

epoch:12

batch size:64

optimizer:Adam

loss function: binary_crossentropy

validation:用 0.1 的 training data(最後 20000 筆)

最後在 kaggle 上得到的準確率為 0.82312, val_acc 也大約為 0.82

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: NO)

   答：

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_7 (Dense)              (None, 1000)              20001000
_________________________________________________________________
dense_8 (Dense)              (None, 1)                 1001
_________________________________________________________________
activation_3 (Activation)    (None, 1)                 0
=================================================================
Total params: 20,002,001
Trainable params: 20,002,001
Non-trainable params: 0
_________________________________________________________________
```

epoch:5

batch size:64

optimizer:Adam

loss function: binary_crossentropy

validation:用 0.1 的 training data(最後 20000 筆)

最後在 kaggle 上得到的準確率為 0.80083, val_acc 為 0.79550

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: NO)
   答：

BOW:兩句話的分數皆為 0.69583458，因為 BOW 只考慮詞出現的次數，而不考慮出現的順序
RNN:
"today is a good day, but it is hot"為 0.53232586
"today is hot, but it is a good day"為 0.99920446
這兩句話的差別我認為在於有 but 這個詞，第二句：雖然今天很熱，但是今天是個美好的一天，確實讓人感到更多喜悅、正面的成分，因此得到較高的情緒分數也是很合理的結果。

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: NO)
   答：

沒有包含標點符號得到的準確率為 0.82312(kaggle)，而包含標點符號的情況下雖然同樣可以得到 0.82 左右的 val_acc，但是 kaggle 上得到的準確率都只有 0.79 左右。

我想原因在於其實標點符號對於一句話代表的情緒不會有太大影響，相反地像是開心或生氣都有可能會用到驚嘆號，反而會造成判別上的困難，不考慮標點符號的情況下，train 出來的 w2v 我想更能代表詞與詞之間的相關程度。

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:NO)
答：

由於訓練時間的考量，我只取了 400000 筆 no label 的 data 進行 semi-supervised,而我標記 label 的方法滿單純的，對預測的情緒分數四捨五入，最後加上原本 18 萬筆 training data 總共 58 萬筆進行 training，最後得到的 kaggle 分數為 0.82240，跟原本的結果相比其實是沒有進步的，不過結果也差不多。在 training 的過程中發現加入更多 data 卻更容易產生 overfit，我想原因可能是我 label 的方法不夠嚴格。