學號：B03701221 系級： 工管四 姓名：王逸庭

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：

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 30)	0
embedding_1 (Embedding)	(None, 30, 300)	6000000
lstm_1 (LSTM)	(None, 300)	721200
dense_1 (Dense)	(None, 150)	45150
dropout_1 (Dropout)	(None, 150)	0
dense_2 (Dense)	(None, 1)	151

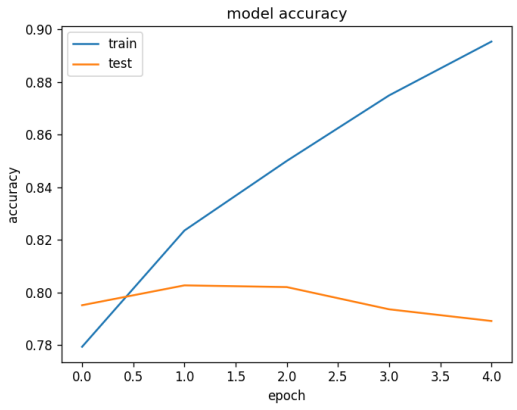
Total params: 6,766,501
Trainable params: 6,766,501
Non-trainable params: 0

None

以文章中最常出現的 20000 字作為辭典，對原始資料進行處理。

模型主要先疊 embedding layer(dim =300)，再疊一個 LSTM(hidden size =300)，最後再接兩個 Dense，層數分別是 LSTM hidden size 的一半 150，和 1。Batch 使用

原本預設的 epoch 為 100，但在訓練過程中使用 early-stopping，停止於第五個 epoch。



右圖為 train 和 validation 在每一個 epoch 的變化

Test acc：(0.80566+0.80601)/2 = 0.805835

LSTM_4.csv a day ago by kelly wang add submission details	0.80566	0.80601

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:)

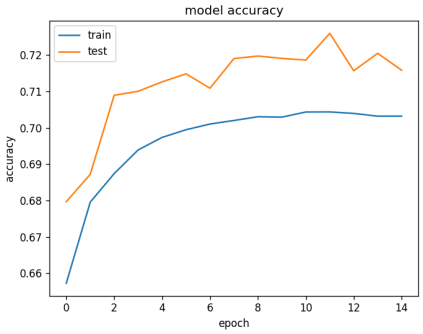
答：

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 5000)              0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               2560512
_________________________________________________________________
dropout_1 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 513
=================================================================
Total params: 2,561,025
Trainable params: 2,561,025
Non-trainable params: 0
_________________________________________________________________
None
```

取前 5000 個常見的字作為 tokenizer，計算每句對應對應的 Bag of words。

模型主要疊兩層 Dense，層數分別是 512,1。

右圖為 train 和 validation 在每一個 epoch 的變化，訓練過程如圖，訓練停止於第 14 個 epoch。



Test acc：(0.72233+0.72035)/2 = 0.72134

bow_1.csv a minute ago by kelly wang add submission details	0.72233	0.72035

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：

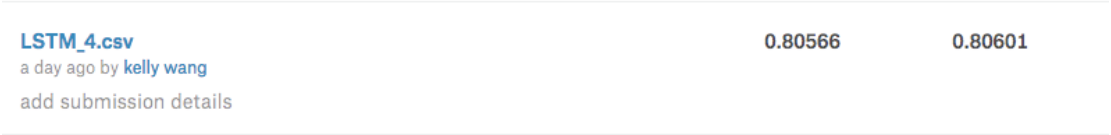
	"today is a good day, but it is hot"	"today is hot, but it is a good day"
RNN	0.20322265	0.95238316
BOW	0.80584508	0.80584508

上述兩句話的差別只有在於 hot, a good day 在語句的位置。而這個差異只會影響 RNN 的結果，因為模型是有考慮字詞的順序（位置），並不會影響 Bag of words，因為 bag of word 只考慮字詞出現的次數，不考慮其所在的位置。

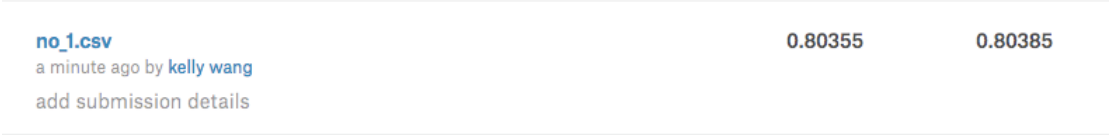
4. (1%) 請比較"有無"包含標點符號兩種不同 tokenize 的方式，並討論兩者對準確率的影響。
(Collaborators:)

答：

有標點符號：`(0.80566+0.80601)/2 = 0.805835`



無標點符號：`(0.80355+0.80385)/2 = 0.8037`



兩者差異大約在 0.02，代表有無標點符號對於結果影響不大。

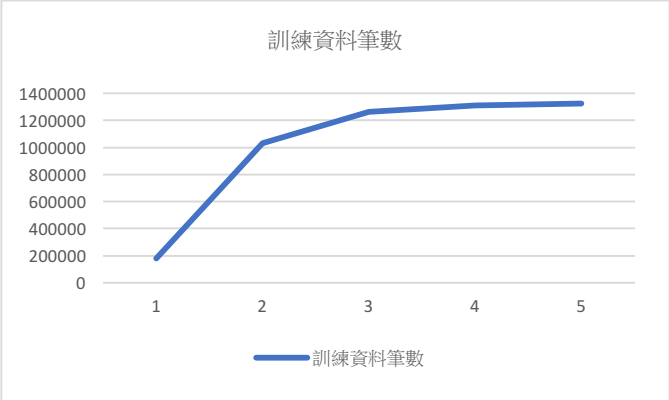
5. (1%) 請描述在你的 semi-supervised 方法是如何標記 label，並比較有無 semi-surpervised training 對準確率的影響。
(Collaborators:)

答：

每一個 epoch 使用上一個 epoch 訓練出來的 model 來 label，threshold 設定 0.3，只有超過 threshold 的資料才拿來加入訓練資料。

訓練資料筆數變化

179991,1032515,1264435,1312176,1326839



train acc = 0.9673

val acc = 0.8141907095

test acc = (0.80619+0.80642)/2 = 0.806305

semi_4.csv a few seconds ago by kelly wang add submission details	0.80619	0.80642

在 test data 上，與一般的 RNN 比，semi-supervised 的準確率大概多了 0.008。而在 validation set 上準確率多了 0.1 左右。代表增加 semi-supervised 準確率有變好，但效果不明顯。