學號：R06922116 系級： 資工所 姓名：賴柏恩

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:)

模型架構:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 37, 100)           7939700
_________________________________________________________________
bidirectional_1 (Bidirection (None, 37, 1024)          2510848
_________________________________________________________________
dropout_1 (Dropout)          (None, 37, 1024)          0
_________________________________________________________________
bidirectional_2 (Bidirection (None, 512)               2623488
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               65664
_________________________________________________________________
dropout_3 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_4 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 65
=================================================================
Total params: 13,148,021
Trainable params: 13,148,021
Non-trainable params: 0
_________________________________________________________________
```

訓練細節:

epoch:3

optimizer:adam

loss_function:binary_crossentropy

以 validation data 準確率最高的 model 當作最佳 model

準確率：0.80123

得到準確最高的 model 是第一個 epoch 產生的, 因此可以知道或許後面就是 overfit 了。在做這個 rnn model 時遇到很大的困難, 因為沒有ＧＰＵ可以使用所以直接用 mac 訓練，一個 epoch 就要 3 個小時。但是因為這次 model 很容易就 fit data 不需要 train 太多次

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 #
=================================================================
dense_1 (Dense)              (None, 256)               2560256
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               32896
_________________________________________________________________
dropout_2 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_3 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 65
=================================================================
Total params: 2,601,473
Trainable params: 2,601,473
Non-trainable params: 0
_________________________________________________________________
```

epoch:3　　optimizer:adam　loss function: binary crossentropy

在 kaggle 上分數為 0.79311。 在做ＢＯＷ時, 一開始因為字數太多電腦 memory 會不足, 但後來試著調整取字的方式，只考慮較常出現的字。但ＢＯＷ試起來的準確率都很差，調整了很多次還是沒辦法通過 baseline

(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：在這兩句中ＢＯW的分數皆為 0.68237281，這是因為ＢＯW本來就不會去考率字的順序，只會考慮字出現的數目，因此分數會依樣。
但在ＲＮＮ中分數分別為，0.32488743,0.83447891，這是因為ＲＮＮ本來就會考慮到字出現的順序，因此準確率會較高。

3. (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：對包含標點符號我的處理方式是，將連續的重複標點符號縮為單一一個，如...縮減成.，這樣子去減少不必要的過多重複標點符號，而對不包含標點符號的處理方式是，將所有標點符號去除。包含標點符號的準確率為 0.80123 而不包含的為 0.79196。皆為使用 rnn 去進行運算。應該是因為標點符號也是會影響語意的東西，因此有標點符號的準確率會較高。但標點符號的處理方式或許可以更好，因為或許有些標點符號有特殊意義，可以去做處理。

4. (1%) 請描述在你的 semi-supervised 方法是如何標記 label，並比較有無 semi-supervised training 對準確率的影響。
(Collaborators:)
答：我先用 rnn model 用 training data traing 一次，接著用 nolabel 的 data 預測，並且取 threshold 為 0.2，也就是取 0.2 以下,0.8 以上的資料，加入 training set 中，以上過程重複十遍。取 validation data 準確率最高的當作最佳 model。
但在做 semi-supervised training 我的準確率反而比 rnn model 降低了，變成 0.7992，可能是因為 threshold 太高，或是說