學號：R06944031 系級： 網媒碩一 姓名：林蔚廷

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:

架構以及訓練過程:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, None, 256)         21233920
_________________________________________________________________
bidirectional_1 (Bidirection (None, None, 512)         1050624
_________________________________________________________________
bidirectional_2 (Bidirection (None, 256)               656384
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 257
=================================================================
Total params: 22,941,185
Trainable params: 22,941,185
Non-trainable params: 0
_________________________________________________________________
Epoch 00001: val_acc improved from -inf to 0.77140, saving model to model.h5
Epoch 00002: val_acc improved from 0.77140 to 0.79050, saving model to model.h5
Epoch 00003: val_acc improved from 0.79050 to 0.80245, saving model to model.h5
Epoch 00004: val_acc improved from 0.80245 to 0.80780, saving model to model.h5
Epoch 00005: val_acc improved from 0.80780 to 0.80995, saving model to model.h5
Epoch 00006: val_acc improved from 0.80995 to 0.81055, saving model to model.h5
Epoch 00007: val_acc improved from 0.81055 to 0.81435, saving model to model.h5
Epoch 00008: val_acc did not improve
Epoch 00008: early stopping
```

在第 8 個 epoch early stop，並且使用第 7 個 epoch 的 model 來預測 testing data

準確度:

在 validation 中 accuracy 可達 0.81435，上傳 kaggle 在 public 得到 0.81765

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:

架構以及訓練過程:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 2)                 30002
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 3
=================================================================
Total params: 30,005
Trainable params: 30,005
Non-trainable params: 0
_________________________________________________________________
Epoch 00001: val_acc improved from -inf to 0.77845, saving model to bow_model.h5
Epoch 00002: val_acc improved from 0.77845 to 0.78760, saving model to bow_model.h5
Epoch 00003: val_acc did not improve
Epoch 00003: early stopping
```

在第 3 個 epoch early stop，在第二個 epoch 的 validation accuracy 只有 0.78760

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。

答:

BOW:

```
prediction:
today today is a good day, but it is hot : [ 0.58346522]
today is hot, but it is a good day : [ 0.63457161]
```

RNN:

```
prediction:
today today is a good day, but it is hot : [ 0.50408459]
today is hot, but it is a good day : [ 0.48788291]
```

可以看到因為 BOW 沒有考慮字句的先後順序，所以預估出來兩個結果都是正面，而 RNN 會記前幾個字，所以預估出來有一個是正面一個是負面。

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:

在我最好的 model 中是有包含標點符號的，在 validation 中可以得到 0.81435 的準確度。

實驗無標點符號的方式，發現表現比起有標點符號的方式差一點

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, None, 256)         21043968
_________________________________________________________________
bidirectional_1 (Bidirection (None, None, 512)         1050624
_________________________________________________________________
bidirectional_2 (Bidirection (None, 256)               656384
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 257
=================================================================
Total params: 22,751,233
Trainable params: 22,751,233
Non-trainable params: 0
_________________________________________________________________
Epoch 00001: val_acc improved from -inf to 0.76770, saving model to model.h5
Epoch 00002: val_acc improved from 0.76770 to 0.78955, saving model to model.h5
Epoch 00003: val_acc improved from 0.78955 to 0.79610, saving model to model.h5
Epoch 00004: val_acc improved from 0.79610 to 0.80080, saving model to model.h5
Epoch 00005: val_acc improved from 0.80080 to 0.80505, saving model to model.h5
Epoch 00006: val_acc did not improve
Epoch 00006: early stopping
```

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:

使用訓練好的 model 來預測 non-label 的資料，並且將預估結果大於 0.9 以及小於 0.1 的資料，加入原本的訓練資料中。

在我最好的 model 中，validation 可以得到 0.81435 的準確度。

只做一次取樣，約加入了 8 萬筆 data，但是在我的 model 表現沒有進步，validation 準確度達 0.81015

```
Epoch 00005: val_acc improved from 0.80875 to 0.81015, saving model to model.h5
Epoch 00006: val_acc did not improve
Epoch 00006: early stopping
```