學號：B04705043 系級： 資管三 姓名：張凱庭

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：
(1)文字處理:
使用gensim的套件，gensim.parsing.porter.PorterStemmer()，對文字進行處理，去掉文尾，刪去一些雜亂的資訊，但是仍保留語意，且保留所有標點符號
(2)模型架構:
如下圖，embedding layer 使用gemsim 的 word2vec pretrained的詞向量(dimension 200維，取出現次數大於10的詞進行訓練)，接著兩層bidrectional LSTM，output分別為256和128，中間加入batchnormalization layer，最後兩層Dense layer，output分別為32和2，loss使用categorical_crossentropy

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 39, 200)           13278000
_________________________________________________________________
spatial_dropout1d_2 (Spatial (None, 39, 200)           0
_________________________________________________________________
bidirectional_3 (Bidirection (None, 39, 256)           336896
_________________________________________________________________
bidirectional_4 (Bidirection (None, 128)               164352
_________________________________________________________________
batch_normalization_2 (Batch (None, 128)               512
_________________________________________________________________
dense_3 (Dense)              (None, 32)                4128
_________________________________________________________________
dense_4 (Dense)              (None, 2)                 66
=================================================================
Total params: 13,783,954
Trainable params: 505,698
Non-trainable params: 13,278,256
```

(3)訓練過程:
validation set取4000筆， batch_size 設定為64，在第11個epoch時候收斂，最後valid_loss=0.39304，valid_acc = 0.82925，訓練時間約一個半小時
(4)結果:

public score	0.82595
private score	0.82507

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：

(1)文字處理:
使用gensim的套件，gensim.parsing.porter.PorterStemmer()，對文字進行處理，去掉文尾，刪去一些雜亂的資訊，但是仍保留語意，且保留所有標點符號
(2)模型架構:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 64)                4196480
_________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 32)                2080
_________________________________________________________________
dropout_2 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 66
=================================================================
Total params: 4,198,626
Trainable params: 4,198,626
Non-trainable params: 0
_________________________________________________________________
```

(3)訓練過程:
validation set取4000筆， batch_size 設定為64，在第2個epoch時候收斂，最後 valid_loss=0.4616，valid_acc = 0.7970，訓練時間約10分鐘
(4)結果:

public score	0.78795
private score	0.78845

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：(1)"today is a good day, but it is hot":

	負面	正面
RNN	0.8505494	0.14945062
BOW	0.42041725	0.57958275

(2)"today is hot, but it is a good day":

	負面	正面
RNN	0.05856774	0.9414323
BOW	0.40273115	0.5972688

BOW將兩句話都歸類為正面情緒，而RNN將第一句歸類為負面，第二句為正面情緒。原因可能在於BOW，比較不能判斷出轉折的語氣，兩句話組成一樣，導致兩種情緒的分數很接近，而RNN有考慮到前後語意，所以在轉折語意的判斷上就有所不同了

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：

   (1)包含標點符號:

public score	0.82595
private score	0.82507

   (2)不包含標點符號:

public score	0.81422
private score	0.81290

   討論:有標點符號準確率大約提高1%，直觀來看，標點符號在句子中也扮演重要的腳色，例如!通常伴隨驚訝的情緒，大多都是正向的，因此包含標點符號的準確率較高
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：我把先前用來預測的model拿來對no label的data標記，將threshold設為0.97，這樣多生成了373389筆資料，把這些資料加進原本label的在重新訓練一次

public score	0.80376
private score	0.80517

   從結果來看反而退步了，可能要把valid set的大小做調整，w2v也要重新訓練，但因為資料數量大幅增加，訓練時間也大大增長，在第二次訓練時的模型可能必須做出調整，才能反映在準確率上。