學號：B04901067 系級： 電機三 姓名：陳博彥

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：（註）兩個gru都是bidirectional

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
gru_1 (GRU)                  (None, 40, 128)           87936
_________________________________________________________________
gru_2 (GRU)                  (None, 128)               98688
_________________________________________________________________
dense_1 (Dense)              (None, 10)                1290
_________________________________________________________________
dropout_1 (Dropout)          (None, 10)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 11
=================================================================
Total params: 187,925
Trainable params: 187,925
Non-trainable params: 0
_________________________________________________________________
```

epochs=72, batchsize=512, validation=0.1, optimizer=Adam, loss_function=binary_cross_entropy

public score	private score
0.83046	0.82878

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：（註）我使用keras的texts_to_matrix

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 512)               512512
_________________________________________________________________
dense_2 (Dense)              (None, 128)               65664
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 129
=================================================================
Total params: 578,305
Trainable params: 578,305
Non-trainable params: 0
_________________________________________________________________
```

epochs=4, batchsize=512, validation=0.1, optimizer=Adam, loss_function=binary_cross_entropy

public score	private score
0.76898	0.76876

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：（註）顯示的分數是sigmoid的output，越接近一表示越正面。

	BOW分數	RNN分數
today is a good day, but it is hot	0.66298014	0.15379485
today is hot, but it is a good day	0.66298014	0.96670616

Bag of word只會記錄句子中出現的單字以及次數，對於BOW來說，這兩個句子是完全一樣的，因此理所當然有一樣的輸出。但也許是因為句子中出現了正面字眼good，所以BOW的model還是把它辨別成偏向正面的結果。

RNN不但會考慮出現的單字，更會把出現的順序納入考量，這兩個句子對他來說是不一樣的，又因為model被train得還不差，兩者的分數差異明顯。

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：

	public score	private score
有標點(!?.)	0.83046	0.82878
無標點	0.82230	0.82124

由於我實作「有標點」的方法，是把較能表現語意的標點留下（留下! ? . 刪掉 / ‘ ’ “ ”~等等沒用的標點），這些訊息對於語意的判讀是有意義的，能讓程式做出更好的預測。

少了標點，準確率稍微差一些，但也不算很爛。

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：

方法一：用unlabeled data來train RNN(self learning)
對unlabeled data作predict，把output大於0.8和小於0.2的unlabeled data當做ground truth，與labeled data合在一起再train一次。

方法二：用unlabeled data來train word2vec(no self learning)
用labeled跟unlabeled data一起train gensim的word2vec，但在train RNN的時候，不使用unlabeled data。

	public score	private score
不使用unlabeled data	0.78000	0.77664
方法一	0.74377	0.74065
方法二	0.81379	0.81072

由結果可見，方法一爛掉，方法二有顯著提升。

方法一爛掉的原因：被當作ground truth的unlabeled data有六十多萬筆，數量比真正的labeded data還多，卻不見得每筆都是對的，方法一爛掉的原因應該是因為training data中有不少錯誤的label。

方法二變好的原因：由於gensim在訓練的過程中，本來就不需要label，在使用大量unlabeled data之後，可以生產出更好的word2vec模型，幫助之後的RNN做出更好的預測。

REFERENCE:

1.https://machinelearningmastery.com/sequence-classification-lstm-recurrent-neural-networks-python-keras/

2.https://machinelearningmastery.com/predict-sentiment-movie-reviews-using-deep-learning/

3.https://vgpena.github.io/classifying-tweets-with-keras-and-tensorflow/

4.https://cloud.google.com/blog/big-data/2017/10/intro-to-text-classification-with-

keras-automatically-tagging-stack-overflow-posts

5.https://www.kaggle.com/ngyptr/lstm-sentiment-analysis-keras

6.手把手小老師