學號：B04505021 系級： 工海三 姓名：黃廉弼

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：

   Summary:

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 40)	0
embedding_1 (Embedding)	(None, 40, 128)	2560000
lstm_1 (LSTM)	(None, 512)	1312768
dense_1 (Dense)	(None, 256)	131328
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 1)	257

Total params: 4,004,353
Trainable params: 4,004,353
Non-trainable params: 0

   Parameters:

   Epoch: 20
   Optimizer : adam
   Loss Fuction : binary_crossentropy
   Dropout rate : 0.5
   Learning rate : 0.001

   ACC: 0.80843/0.80764(Public/ Private)

   說明: 這次的 RNN 主要是靠 keras 的 Tokenizer 先建立辭典，接著依照辭典的內容，經過 embedding layer 後送入 LSTM，後面接著很基本的 DNN 架構。

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：

   Summary:

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 20000)	0
dense_1 (Dense)	(None, 256)	5120256
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 1)	257

Total params: 5,120,513
Trainable params: 5,120,513
Non-trainable params: 0

Parameters:

Epoch: 20

Optimizer : adam

Loss Fuction : binary_crossentropy

Dropout rate : 0.5

Learning rate : 0.001

ACC: 0.72171/0.72441(Public/ Private)

說明: 這次的 BOW 主要是靠 keras 中 Tokenizer 的套件先統計各詞彙出現的情形，然後接的是很基本的 DNN 架構。

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"
BOW	0.483444	0.483444
RNN	0.734555	0.859974

BOW 因為是以統計的方式處理，故當兩個句子組成相同僅排列不同時，會 Mapping 到相同的分數；相對的，RNN 的架構會考慮到語句前後關聯，故當順序改變會有所差異。

這樣的結果，明顯看得出來，有考慮前後關聯的 RNN 能做出較貼近自然語言的判斷。

不過值得討論的是，在結果中 RNN 在"today is a good day, but it is hot"中判斷的情緒分數為 0.734555 為正面情緒，與直覺看下來其實是不太一致的，從這邊其實可以看出，這次實作出來的 RNN 其實還有改善空間。

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：

	無標點符號	有標點符號
ACC	0.80241/0.80150	0.80832/0.80755

   處理方式主要透過 Keras 中 Tokenizer 的 constructor 裡的參數 filter 進行調整。
   由原先的 ACC 0.80241/0.80150 進步到 0.80832/0.80755，可看出標點符號對語氣判斷確實是有幫助的。
   尤其是 training data 以及 testing data 都是十分生活化的文句。不論是”!!!!!”或是”…….”都有帶有語氣強烈的意味，因此也進一步加深的標點符號的重要性，故在這項測試中，兩者所表現出來的準確率會有如此差異之處。

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：本次 semi-supervised 的部分主要是以 self-training 的方式去時實作，把 un-labeled 的資料餵到之前用 labeled-data train 出來的 model 中

   而經過 semi-supervised 後，ACC 也僅從原先的 0.80241/0.80150 提升至 0.80767/0.80669，其提升的幅度相對於所給的資料量其實是十分有限的。