學號：R06921082　系級：電機碩一　姓名：陳與賢

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：

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

loss function 使用 binary crossentropy、optimizer 為 Adam、validation ratio 為 0.1、dropout rate 為 0.3、使用 LSTM、semi-supervised 的 threshold 為 0.1、supervised 的 epoch 設定 20、semi-supervised 的 epoch 設定 2 且 iteration 設定 15，另外我有將 testing_data 一起拿來 train semi-supervised。

Semi-supervised validation 的準確率從一開始的 0.8012 到最後最佳的是 0.8052，predict 後上傳至 kaggle 分數 0.80524。

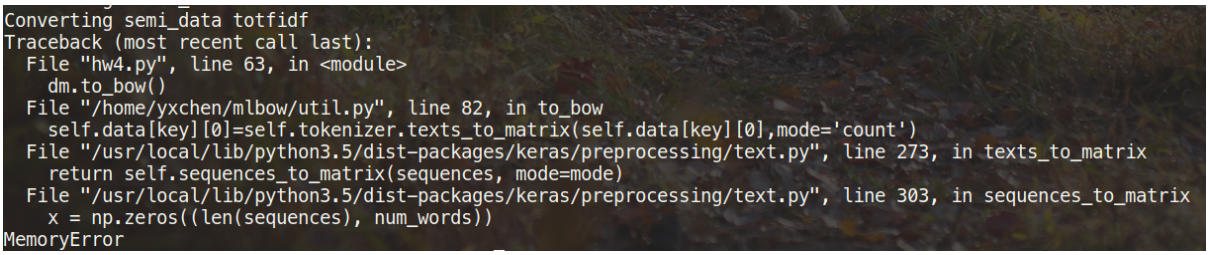
2. (1%) 請說明你實作的 BOW model，其模型架構、訓練過程和準確率為何？
答：

架構跟 RNN 差不多，只是直接將 RNN 的 LSTM 跟 embedding 拿掉而已。

Layer (type)	Output Shape	Param #
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

loss function 使用 binary crossentropy、optimizer 為 Adam、validation ratio 為 0.1、dropout rate 為 0.3、semi-supervised 的 threshold 為 0.1、supervised 的 epoch 設定 20、semi-supervised 的 epoch 設定 2 且 iteration 設定 10。



因為 server 的電腦不給力(如上圖)…nolabel 資料量太大 memory 不夠用，所以我有將 nolabel 的 data 刪到只剩 20 萬筆資料。

Semi-supervised validation 的準確率從一開始的 0.7220 到最後最佳的是 0.7368，predict 後上傳至 kaggle 分數 0.73389。

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：

RNN 對第一句 predict 出來為 0.420319、對第二句 predict 出來為 0.993098
BOW 對第一句 predict 出來為 0.657915、對第二句 predict 出來為 0.657915

可以看出來 BOW 因為沒有考慮先後順序，而單純考慮每個單字的出現次數之下，兩句的預測是一樣的（因為 input 一樣）。

而 RNN 因為有考慮出現的前後順序所以預測出來的情況較準確。

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：

上傳 kaggle 後有包含標點符號的分數為 0.80524，沒有包含標點符號的分數為 0.80472。

兩者是差不多的，而有包含標點符號高一點點，我想可能是因為加入標點符號可以讓 model 針對語句的文法及語氣的轉換掌握的更好。

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：

threshold 我是直接用助教設定的值，也就是 0.1，代表說只會取 predict 出來 > 0.9 及< 0.1 的 nolabel data 來加入 labeled data 繼續 train。

上傳至 kaggle 後，有 semi-surpervised training 的分數為 0.80524、沒有 semi-surpervised training 的分數為 0.80414，有 semi-surpervised 高約 0.001，雖然看起來很少，但少這 0.001 我的排名會直接下降十幾名，所以 semi-surpervised training 還是蠻有用的。