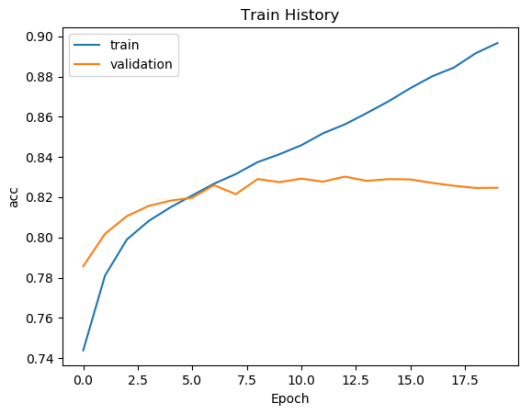
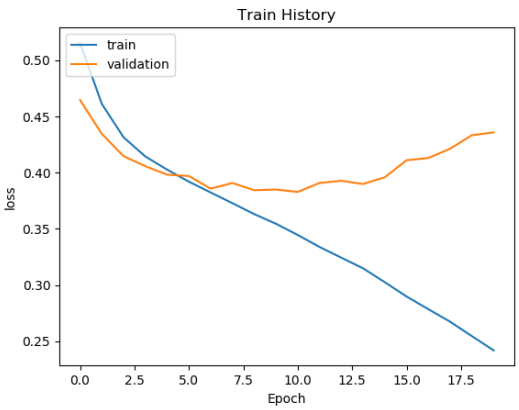
學號：R05546014 系級： 工工二 姓名：洪紹綺

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: Keras Sample Code)

答：使用 gensim 套件訓練詞向量，資料包含 training_nolabel、training_label 以及 testing_data，設定 sg=1，(對出現頻率低詞彙也具敏感度)，min_count=1(沒重複出現的字也訓練)，詞向量設為 300 維度(但因為這樣 model 超大)。此問題為二分類問題，RNN Model Loss Function 取對數損失函數(binary_crossentropy)；Optimizer 取 Adam；訓練 20 個 Epoch 後取第 14 個 Epoch，Validation 準確率為 0.8296 的 Model，Kaggle 準確率為 0.83056(Public)。

```
word_vectors.shape (256834, 300)
Training...
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, None, 300)         77050200
_________________________________________________________________
lstm_1 (LSTM)                (None, 500)               1602000
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 501
=================================================================
Total params: 78,652,701
Trainable params: 1,602,501
Non-trainable params: 77,050,200
```

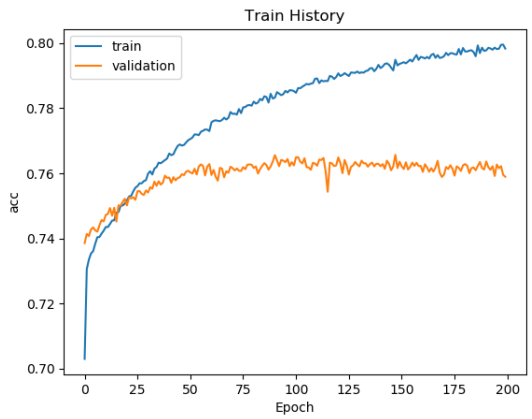


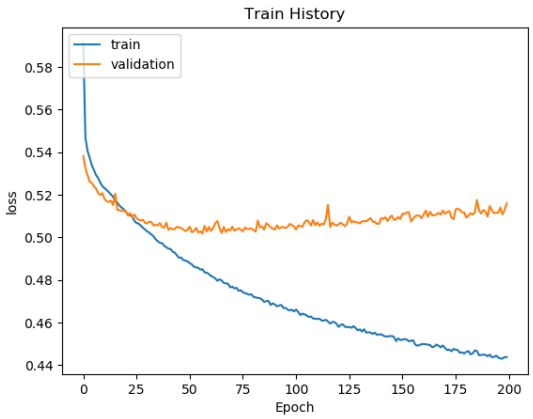


2. (1%) 請說明你實作的 BOW model，其模型架構、訓練過程和準確率為何？
(Collaborators: 0)

答：此問題為二分類問題，Loss Function 取對數損失函數(binary_crossentropy)；Optimizer 取 Adam；訓練 200 個 Epoch 後取第 148 個 Epoch，Validation 準確率為 0.7657 的 Model，Kaggle 準確率為 0.72449 (Public)。

```
Training Bow Model...
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 150)               45150
_________________________________________________________________
dropout_1 (Dropout)          (None, 150)               0
_________________________________________________________________
dense_2 (Dense)              (None, 256)               38656
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 257
=================================================================
Total params: 84,063
Trainable params: 84,063
Non-trainable params: 0
```





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: 0)

答：BOW Model 不考慮單字出現順序，因此用字完全相同的句子不僅被歸為同一個類別，情緒分數也完全相同；RNN Model 則是考慮完整的句意，因此 but 子句暗示的情緒差異，直接表現在 RNN 情緒分數上，且被歸在不同類別。

TEXT	情緒分數	
	BOW Model	RNN Model
today is a good day, but it is hot	0.8660423	0.240648
today is hot, but it is a good day	0.8660423	0.961589

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: 0)

答：RNN Model 包含標點符號準確率高 2%左右， 推測 RNN 有學到像是！！、…… 也可以表達情緒。

Tokenizer	RNN Model	
Validation 準確率	包含標點符號	不含標點符號
	0.83056	0.81165

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: 0)

答：

以前述最好的 RNN model 預測完 Unlabel 資料後擴充我的 training data， pos_threshold、neg_threshold 分別設為 0.85 及 0.2，Unlabel 預測機率大於 0.85 或小於 0.2 才加入 training data。Semi-supervised 在 Kaggle Public 準確率大約較原 RNN Model 高 0.3%。