HW5 REPORT

學號：B04705026 系級： 資管三 姓名：林彥廷

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？

- **架構**

# layer	Layer	Units	Actication	
6	Dense	2	softmax	
5	Dense	32	relu	L2 regularized
4	Bidirectional LSTM	64	LSTM cell	Dropout=.2
3	Bidirectional LSTM	64	LSTM cell	Dropout=.2
2	Embedding	-		Trained on Gensim W2V
1	Input	Dim=37		37為設定最長句子辭彙數

- **訓練細節與參數**
  - **Feature Extract.**: W2V trained on Gensim (包含標點符號)
  - **Loss Func.** : categorical crossentropy
  - **Optimizer**: Nesterov Adam (參數使用原論文建議)
  - **Epochs**: 第7個epochs時收斂到val_loss=0.4168
  - 隨機取10% Validation Set判斷模型表現
- **模型表現**

	RNN model
public score	0.82546
private score	0.82433

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？

- **架構**

# layer	Layer	Units	Actication	
4	Dense	2	softmax	
3	Dense	64	relu	Dropout=.5
2	Dense	128	relu	Dropout=.5
1	Input	Dim=82117		

- **訓練細節與參數**
  - **Feature Extraction**: Count (計算字詞出現次數)

- **Loss Func.** : categorical crossentropy
- **Optimizer**: Nesterov Adam (參數使用原論文建議)
- **Epochs**: 第2個epochs時收斂到val_loss=0.4746
- 隨機取10% Validation Set判斷模型表現

- **模型表現**

	BOW model
public score	0.77975
private score	0.77956

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。

	Today is a good day, but it is hot		Today is hot, but it is a good day	
情緒分數	RNN	BOW	RNN	BOW
P(Negative)	0.7591841	0.11468845	0.10389365	0.11468845
P(Positive)	0.24081592	0.8853115	0.8961063	0.8853115

- BOW模型並沒有考慮字詞出現順序，對BOW來說兩句話是相同的，並且模型判斷為正面很有可能是因為出現"Good"這個詞。
- RNN會考慮字詞出現順序以及彼此之影響(Eg.介系詞之影響)，能準確判斷出"But"後面的字句才是主要情緒。

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。

- **標點符號**：'!"#$%&()*+,-./:;<=>?@[\]^_`{|}~'
- **討論**
  - 同樣使用第1小題的架構與相同訓練細節，包含標碘符號的模型表現比較好。
  - 標點符號雖然不是文字，但有時可以些微偷漏出語意(Eg.'?'可能包含比較多負面情緒)，模型可以有額外的信息來判斷正反情緒，但有些標點符號並沒有意義({%})可能增加Noise。
- **模型表現**

RNN Model	包含標點符號	不包含標點符號
public score	0.82546	0.81451
private score	0.82433	0.81446

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。

- **Semi-supervised方法 - Self - Training**
  - 設定機率門檻值 = 0.9
    1. 使用Training data訓練Base learner
    2. 預測no label data正反情緒之機率
    3. 幫機率高於門檻值之no label data標上預測標籤
    4. Traning data增加新標記資料，從no label data中移除新標記資料
    5. 重新訓練模型，紀錄val accu.最好的模型
    6. 重複步驟2（2次/no label data被標記完）
- **討論**
  - 使用self-training表現並沒有顯著提升，可能原因如下
    - Self-training容易放大前期標注錯誤的Noise：如果標注no label data時就已經標錯，後面的訓練只是overfit雜訊而已。
      藉由觀察訓練過程，val loss相較於supervised的模型高出不少，所以可能是前期label問題。
      Ref: http://pages.cs.wisc.edu/~jerryzhu/pub/sslicml07.pdf
- **模型表現**

	Supervised	Semi-supervised
public score	0.82546	0.82445
private score	0.82433	0.82224