學號：b03902101 系級： 資工四 姓名：楊力權

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: No)

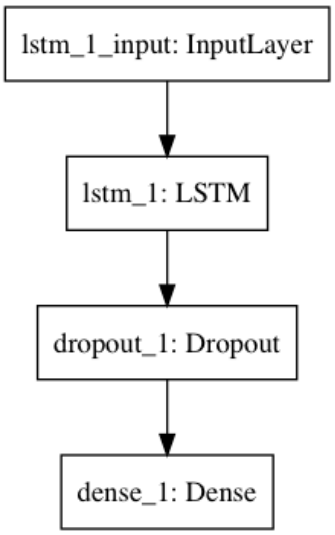
答：

(a.)模型架構

lstm_1：LSTM units=200

dropout_1：Dropout rate=0.5

dense_1：Dense units=1 activation=sigmoid



(b.)訓練過程

pretrain：Gensim word2vector 100維

loss function：binary_crossentropy

optimizer：adam

epoch：10

batch size：64

過程：把句子的每一個字轉成100維的word embedding，然後再把每個句子pad sequence成長度(字數)40，每個字維度100，字數不足的句子後面全補0，得到shape(200000, 40, 100)的陣列，丟入rnn訓練。

(c.)準確率

kaggle score　public：0.81861　　private：0.81629

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: No)

答：

(a.)模型架構

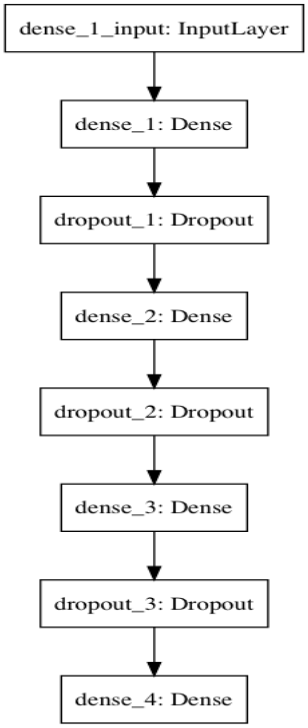
dense_1：units=512 activation=relu

dense_2：units=256 activation=relu

dense_3：units=128 activation=relu

dense_4：units=1 activation=sigmoid

dropout rate全部都是0.5



(b.)訓練過程

loss function：binary_crossentropy

optimizer：adam

epoch：10

batch size：64

過程：把每個句子變成一個一維的bag of word，每個index代表對應到的字出現次數，丟進DNN訓練。

(c.)準確率

kaggle score　public：0.79588　　private：0.79607

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: No)

	today is a good day, but it is hot	today is hot, but it is a good day
BOW	0.63041413	0.63041413
RNN	0.66038901	0.95214069

雖然兩種方式的兩句預測結果都大於0.5，也就是label皆是1，但是很明顯可以發現BOW兩種句字的分數一樣，而RNN有很大的差異。
BOW只判斷句子中某些字出現的次數，因此單字一樣但是排列不同的兩個句子在model前面其實是同一個input因此得到的prediction當然相同。
RNN有著記憶前後文的性質，因此不同出現順序的字會得到不同的結果，因而得到截然不同的分數，更可以很大膽的預測第二句是個正面句。

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: No)

	public	private
有標點符號	0.81861	0.81629
無標點符號	0.80795	0.80417

討論：有加入標點符號的結果叫好，標點符號對句子的轉折以及表達有一定的影響，因此在去掉標點符號後，使model難以辨認轉折點而不易判斷正反面。

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: No)

先用有label的data訓練出一個模型，用這個模型對unlabeled的data做預測，若分數高於自設的threshold(0.9)就給label 1，低於0.1就給label 0，因此得到更多有label的training data，再用這些training data重新train一個model。

	public	private
supervised	0.81861	0.81629
semi-supervised	0.82498	0.82296

討論：semi-supervised的結果較好，因為在training data二十萬筆與no label data一百三十萬比的差距之下，用semi-supervise增加十分確定label的data，能夠增加不少trainging data，對訓練模型更有幫助。