學號：R06921077 系級： 電機碩一 姓名：黃詩凱

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：使用skip-gram+兩層LSTM(512)，並接到Fully Connected Network去train

Batch size=64, epoch=10, word vector維度100，並padding到長度140

沒有用semi-supervised

句子沒有特別處理，標點符號都保留，單詞出現次數小於3次就捨棄

最後分數：0.80169

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：

使用內建Tokenizer直接建字典，並直接把input sequences轉BOW形式，接到一層LSTM+Fully Connected Network去train

一樣保留標點符號，沒有特別過濾和處理

Batch size=64, epoch=10, word vector長度140, 沒有semi-supervised

分數：0.78751

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：

RNN:

[[0.9875202]

[0.3768344]]

從結果來看，第一個在我的RNN model被當作是正面的 (>0.5)

第二句則是負面的(<0.5)

BOW:

[[0.7248613]

[0.7248613]]

從結果來看，我的BOW model把這兩句當作是正面的 (>0.5)

但BOW的結果相對較弱一些，只有0.7

兩者的差異主要是BOW不會去記憶前面出現過的字詞，所以理論上預測起來會容易比RNN model預測出來的意思還要偏 (BOW不考慮順序及文法)，但也有可能是我的Code寫爛就是了...

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：

   以原本的RNN model來比較
   有包含標點符號下去train，出來的準確率是：0.80169
   但是如果去掉標點符號(包含一些像亂碼的字元)，出來的準確率：0.76320
   直覺來看，後者準確率應該有可能比較高，因為data很多亂碼
   但從結果來看卻相反，猜測可能是因為標點符號也會影響句子的文意理解，單就逗號來說，標在不同位置可能就會語意造成極大的差異。
   我想這個差異比亂碼等符號的影響更大，所以才會有這個結果。

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：

   如果predict結果>0.7，就標記該筆data的label = 1

   如果小於0.3就標0

標完之後繼續train，最後出來的結果準確率有上升一點點

從accuracy 0.78多上升到0.80多。