學號：B04902044 系級： 資工二 姓名：朱柏澄

1. 請問softmax適不適合作為本次作業的output layer? 寫出你最後選擇的output layer並說明理由。

不適合。softmax 適合用在單一 label 的預測；但這題是 multiple label，比較適合用 sigmoid 作為 output layer。

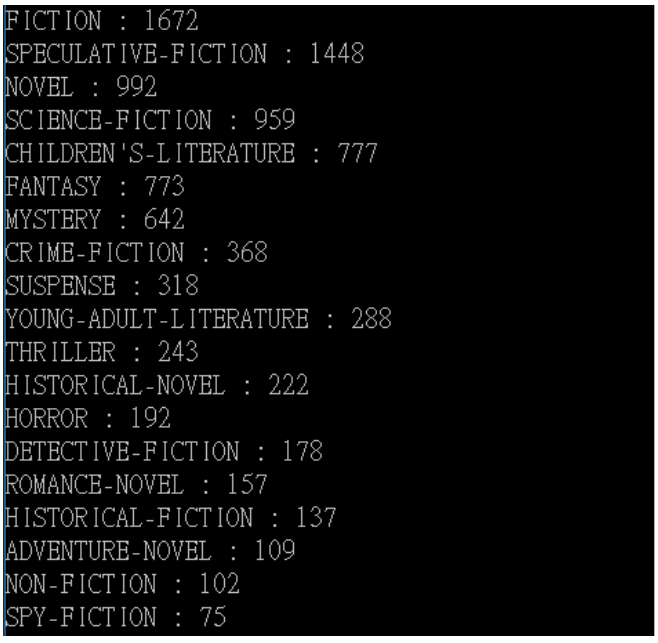
這題用 softmax 的話，train 的時候就會有問題，例如有一筆資料的答案有兩個 label，那麼 training 的目標應該是使那兩個 label 的 output 越接近 1 越好，但 softmax 卻只能使那兩個 label 都是 0.5。另一個問題是在預測答案的時候，很難有一個判斷的標準。如果 output 中其中三個是 0.5 0.3 0.2，此時不好判斷出答案究竟是一個或是兩個、甚至三個 label。以上兩點是用softmax 最大的問題，如果使用 sigmoid ，將每個 label 視為獨立的答案，根據輸出的答案是否超過門檻值，作為判斷的依據，就可以避免這些問題，所以我最後選擇 sigmoid 作為 output layer。

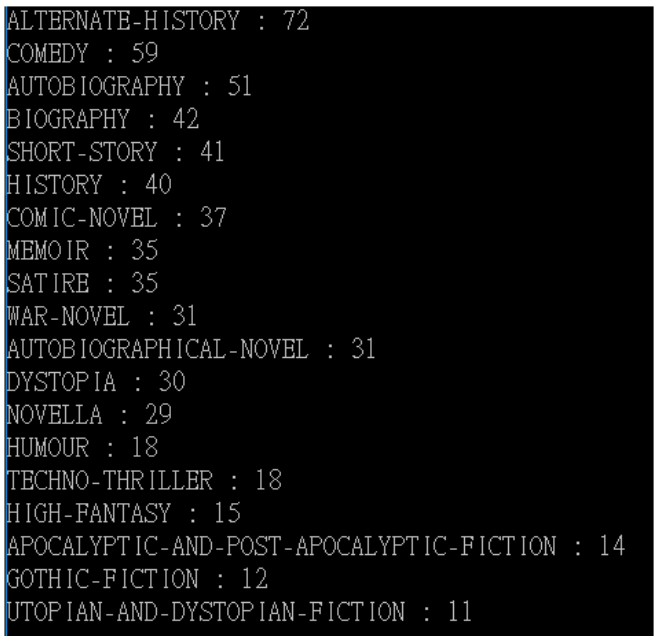
2. 請設計實驗驗證上述推論。

我用同一個架構，分別使用 sigmoid 和 softmax 作為 output layer，sigmoid 的 model 設 0.4 為 threshold，softmax 的取前 k 大且總和超過全部的 0.7 的 k 個做為答案。
實驗結果如下：
sigmoid : 0.5103 (on validation data)
softmax : 0.3341 (on validation data)

3. 請試著分析tags的分布情況(數量)。

training data:





fiction 相關的類別佔了最大部分。

4. 本次作業中使用何種方式得到word embedding?請簡單描述做法。

我用的是 trian 好的 GloVe
GloVe 的想法是來自 co-ocurrence matrix，作者認為 ratios of co-occurrence probabilities 比一般的條件機率更能表示詞和詞之間的關係。
由 co-ocurrence matrix 推導出的 cost function

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

做 gradient descent，就可以得到 GloVe。

5. 試比較bag of word和RNN何者在本次作業中效果較好。

我覺得差不多，kaggle public 上最好的成績分別是：
bag of word: 0.50710
RNN: 0.50116