學號：B05901189 系級： 電機一 姓名：吳祥叡

1. (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。

softmax 不適合作為 output layer.我選用 sigmoid.

原因如這篇文章提到：

Sigmoid, unlike softmax don't give probability distribution around $nclasses$ as output, but independent probabilities.

https://stats.stackexchange.com/questions/207794/what-loss-function-for-multi-class-multi-label-classification-tasks-in-neural-n

也就是說因為 softmax 會讓全部 label 加總是 1,label 之間會互相影響,不適合用在 multilabel 的狀況下,.

2. (1%)請設計實驗驗證上述推論。

我將我做出來最好的 model 的 output layer 改成 softmax activation。然後觀察他的 performance.

在 training 階段如預期的 f1_score 都是 0，因為沒有一個 tag 的預測值會超過 threshold.

接著觀察訓練完的 model 在 training set 上的預測. 第一筆是這樣：

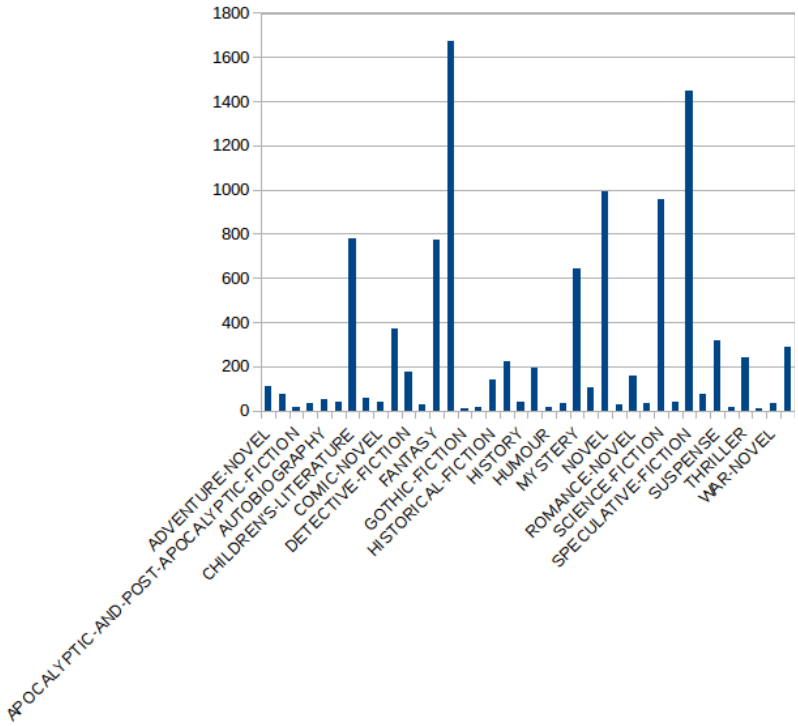
[0.12999448 0.1653344 0.11967589 0.08315536 .....]

正確的 predict 會是前兩個數字要遠大於其他的，從這裡可以看出他有學出這件事，但是因為加總為 1 的這個性質讓他的 prediction 被分散，都過不了 threshold.

最後提出如果一定要用 softmax 的改良方法：我認為必須要有替代 threshold 的方法，比如說找每個 tag 的預測值可能會差數倍，把相對的關係當作 threshold 可能也是一種作法。

3. (1%)請試著分析 tags 的分布情況(數量)。

各個 tag 數量如下表：



4. (1%)本次作業中使用何種方式得到 word embedding?請簡單描述做法。

使用 GloVe.他使用 count based 的方法，在他們的論文（https://nlp.stanford.edu/pubs/glove.pdf 中提出利用 co-occurrence matrix 和 word vector 所希望要有的性質導出關係:

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik}) . \qquad (7)$$

但因為 log 在 0 發散因此定義 Loss function 包含定義快速收斂的 weight:

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

(8

利用這個 Loss function 去做訓練。

5. (1%)試比較 bag of word 和 RNN 何者在本次作業中效果較好。

原先我是拿助教的 sample code 改成 bag of word，做出來 kaggle 分數和 validation 可以到.48 ，和 RNN 做出來的差不多，不過 training 時間遠少於 RNN。

但是後來助教修正 f1 score 以後用同樣的架構的 bag of word 就 train 不起來了，但 RNN 和修正前差不多，原因我還不清楚。

但是我的電腦記憶體不夠跑 bag of words（連減少到 1000 個字都不行），之前申請的工作站又到期了所以就沒有繼續做下去了。