Machine Learning Homework5 Report

學號：B02901124 系級： 電機四 姓名：黃柏翔

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 在本次作業中不適合作為 output layer。因為 softmax 是要讓所有的 output 的值加起來等於 1，通常是用在只要預測一種類別的 classification，使該 label 的值越趨近 1 而剩下的 label 的值越趨近 0。而在這次的作業中，由於所要預測的為 multi-label，若是選擇 softmax 去作為輸出層，會使得最後分數的值分散掉，很難去設定一個 threshold 判定哪些 label 是對的。
   而我所使用的 output layer 是 sigmoid，會個別的將每個 label 的值變到[0,1]之間，不會與其他 label 有關，因此較為合適。

2. **(1%)請設計實驗驗證上述推論。**

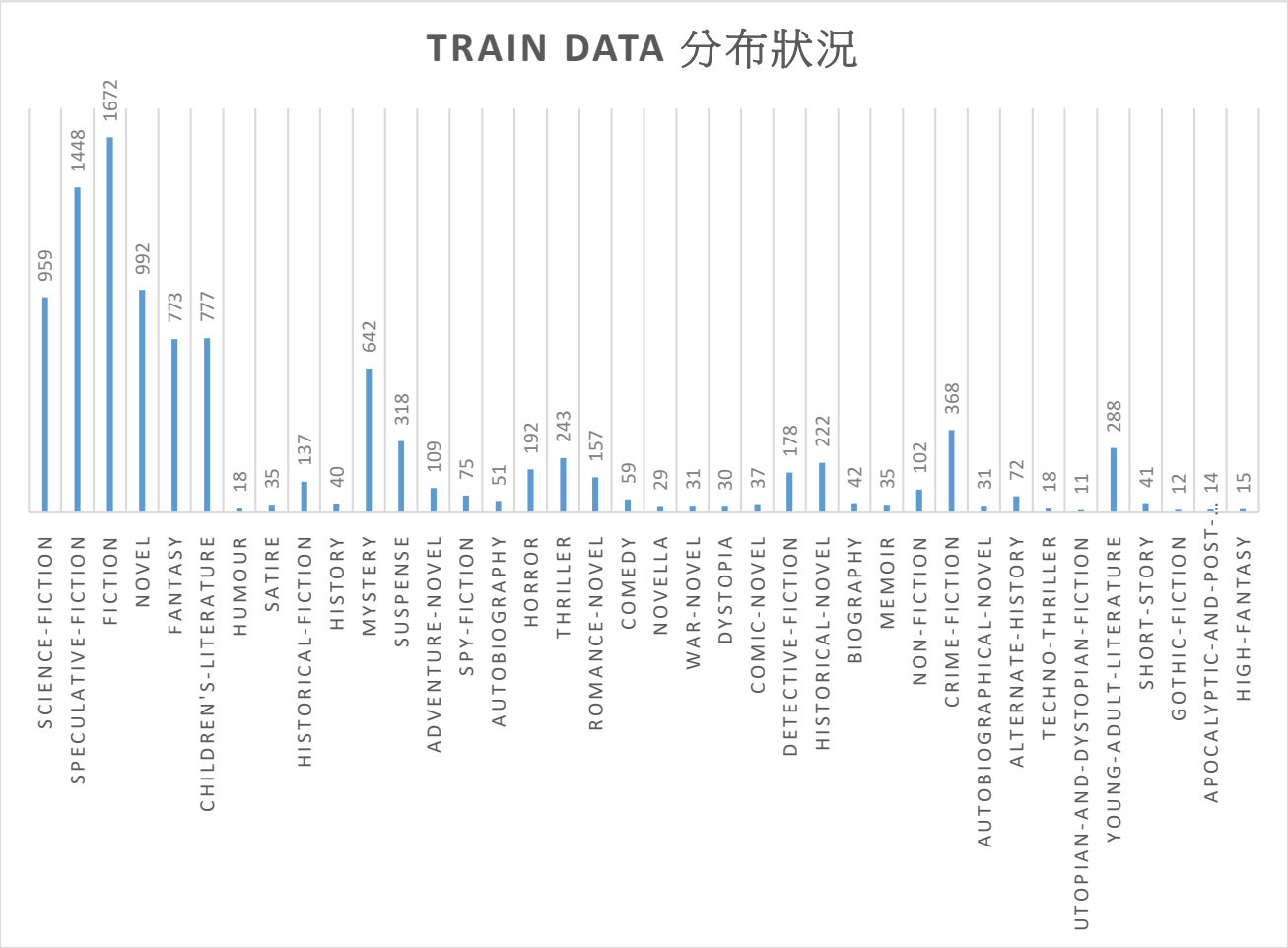
   我利用同樣架構的 model ，只代換最後的 output layer，分別為 softmax 級 sigmoid 去作 training，測試出來的結果如下表。

Activation	Threshold = 0.2	Threshold = 0.3	Threshold = 0.35	Threshold = 0.4
Softmax	0.3768	0.3957	0.4222	0.4453
Sigmoid	0.4394	0.4688	0.4802	0.4809

   由實驗數據可以明顯看出 sigmoid 的效果較 softmax 突出許多。

   而丟到 kaggle 上的結果差異更大但在此沒有多作分析。

3. (1%)請試著分析 tags 的分布情況(數量)。



SCIENCE-FICTION	959
SPECULATIVE-FICTION	1448
FICTION	1672
NOVEL	992
FANTASY	773
CHILDREN'S-LITERATURE	777
HUMOUR	18
SATIRE	35
HISTORICAL-FICTION	137
HISTORY	40
MYSTERY	642
SUSPENSE	318
ADVENTURE-NOVEL	109
SPY-FICTION	75
AUTOBIOGRAPHY	51
HORROR	192
THRILLER	243
ROMANCE-NOVEL	157

COMEDY	59
NOVELLA	29
WAR-NOVEL	31
DYSTOPIA	30
COMIC-NOVEL	37
DETECTIVE-FICTION	178
HISTORICAL-NOVEL	222
BIOGRAPHY	42
MEMOIR	35
NON-FICTION	102
CRIME-FICTION	368
AUTOBIOGRAPHICAL-NOVEL	31
ALTERNATE-HISTORY	72
TECHNO-THRILLER	18
UTOPIAN-AND-DYSTOPIAN-FICTION	11
YOUNG-ADULT-LITERATURE	288
SHORT-STORY	41
GOTHIC-FICTION	12
APOCALYPTIC-AND-POST-APOCALYPTIC-FICTION	14
HIGH-FANTASY	15

總共 38 個 tags，其分布狀況是 fiction 最多，Utopian-And-Post-Apocalyptic-Fiction 是最少的

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

讀取 test 與 train data 後利用 keras 的 tokenizer 統計出現過的字並給予 index，再利用 GloVe pre-train 好，100 維的 word vector 去表示這些單字。換言之就是講每個單字用一個 100×1 的陣列表示。此過程即為 word embedding。

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

我在本次作業中利用 bag of word 所得到的效果比較好。主要的原因可能是在於這次作業的目的只在餘分類這些字句的段落，可能只需要看到那些單字的出現跟 label 的相關性，並不用去考慮整個文章的意義。又或者太簡單架構的 RNN 並沒有辦法輕易利用字句的意義去判斷文章的分類，但單就比較簡單的架構而言，bag of word 所得到的效果是比 RNN 來的好的。