學號：R05942056 系級： 電信碩一 姓名：時丕澔

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。

若對這兩個function做一些研究，可以發現softmax會傾向於只讓其中一個特別突出；也就是說，若原本有兩個輸出值突出，則softmax傾向於將其中一個壓低，這種情形在sigmoid的情況不會發生。

故考慮到這次作業的目的multilabeling會需要讓多個值都夠大，可以預測sigmoid函數會比softmax更適合當作output的activation function。

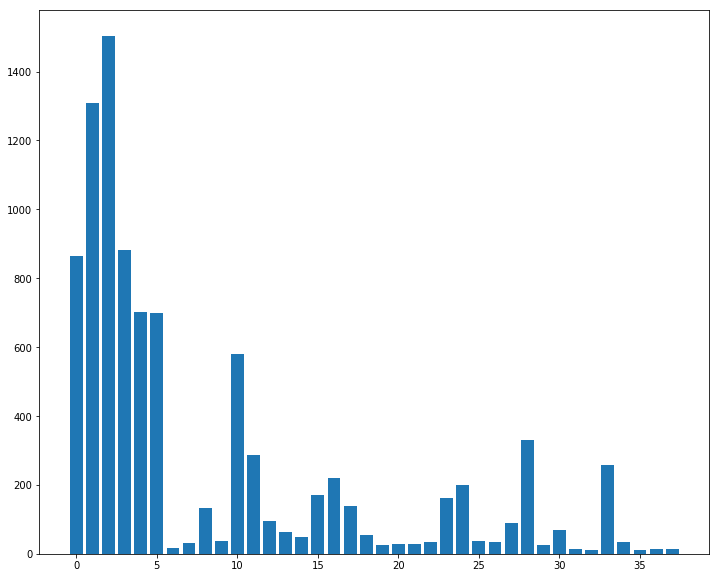
1. (1%)請設計實驗驗證上述推論。

我在此作業中試著分別使用了softmax和sigmoid當作output層的activation function，架構皆為embedding層→兩層GRU(128,tanh)後→三個Dense層(256,128,64,RELU)→output層(38,softmax/sigmoid)，使用f1\_score (threshold=0.4)當作validation metric。

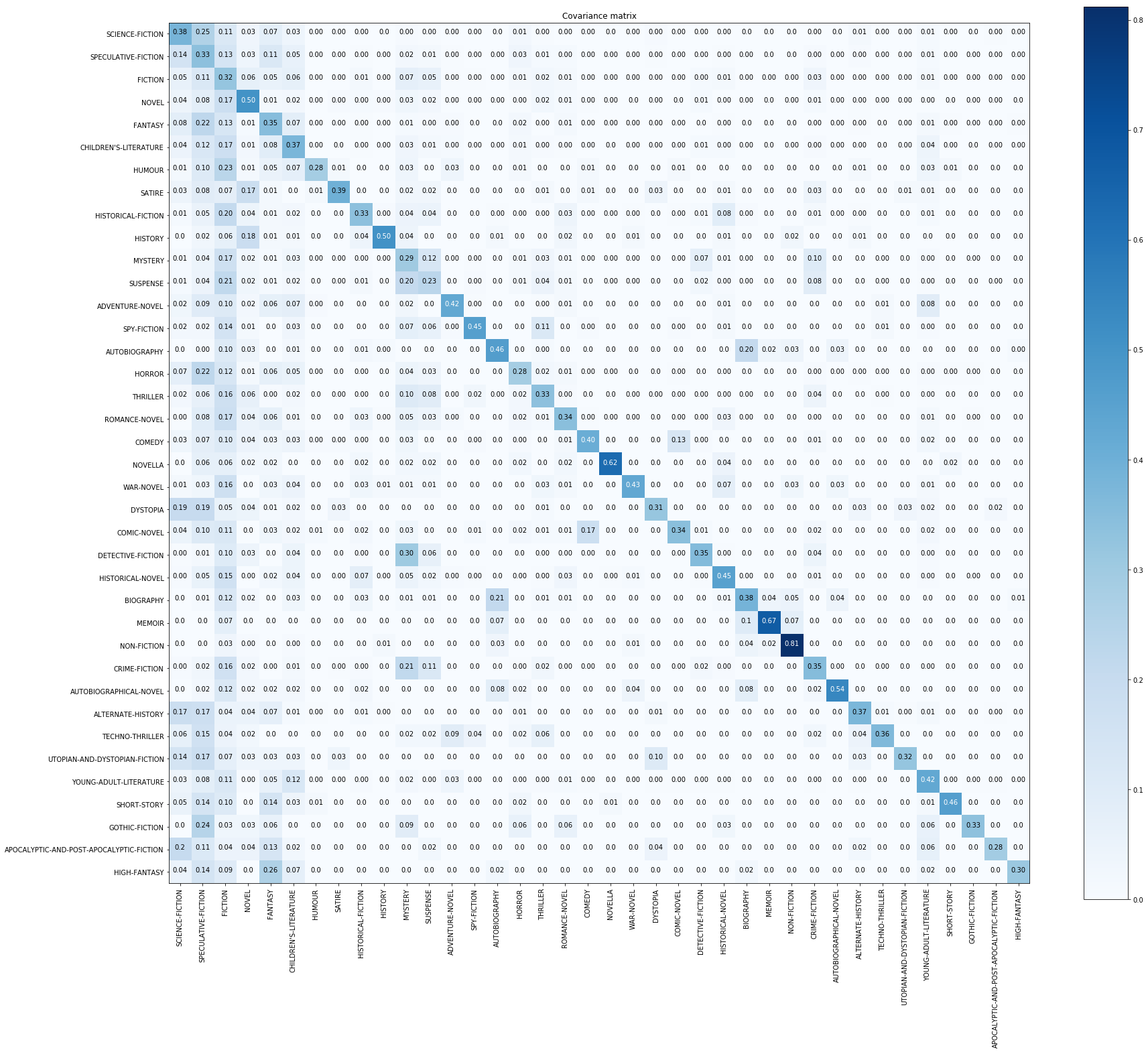
最後得出的結果為softmax 漸增到f1\_score=0.25, sigmoid漸增到f1\_score=0.53。可以看到softmax在比sigmoid早很多的地方就停下來了，證實第一題推論的合理性。

有趣的是，如果在使用softmax 當作activation function 的情況下學出來的 model，最後在prediction的時候將threshold降低至0.13~0.2左右，f1\_score就有機會和sigmoid做出來的模型差不多，從這情況可以看出softmax的情況也有train出對的model，但會因為原本function的限制導致f1\_score上不去。這情況也和預期softmax的特性相符。

1. (1%)請試著分析tags的分布情況(數量)。



　　上圖是training set中所有tag的總出現次數。可以看出前六個tag是比較常使用的，分別為SCIENC-FICTION, SPECULATIVE-FICTION, FICTION, NOVEL,FANTASY還有CHILDREN-LITERATURE，而第九個HISTORICAL-FICTION也是比較多的種類。



　　上圖為對training set的tag matrix M做MTM的row normalize，可以從圖中看出每兩種tag之間的關係，例如FICTION較常伴隨其他tag一起出現、而NON\_FICTION比較常單獨出現，DETECTIVE\_FICTION和MYSTERY很常一起出現……等等。

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

　　我使用了GloVe網站上提供的事先train好的word vector來做embedding，使用的data是維基百科和Gigaword的開放資料。

　　GloVe使用的是log-bilinear的model，透過看整個corpus句子的co-occarence來決定兩個字之間的條件機率關係，進而決定他們之間的距離，而距離則可以用兩個向量之間的內積來換算，這樣就可以連接成一套最佳化問題，最後GloVe選擇了weighted mean square 來當整個最佳化問題的objective function。

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

　　為了跟和2.中設計的model有差不多的參數量(約20w)，直接使用Bag of word的tag丟進DNN(256,256,128,64,38,RELU/sigmoid)學習，結果的f1\_score只有大約0.21左右，就算和前述model一樣先使用word embedding 再丟進DNN，f1\_score也只有約0.39左右，還是沒有使用RNN layer的結果來的突出。

　　不過值得注意的是這些模型中最耗時間的元素就是RNN layer。要不要使用其實是可以再討論的事情。