DL Final Competition Report

學生：0513404 姜林寬

如果助教開啟這個pdf檔排版有跑掉，麻煩移駕到以下網址，謝謝~

https://hackmd.io/1XbEwR00TNC8EYhorr5E1g (https://hackmd.io/1XbEwR00TNC8EYhorr5E1g)

Public Score: 0.92321

Rank:62/144(截至比賽結束前兩小時)

執行方式：直接執行

實作過程

這個競賽是新聞分類的題目，屬於NLP的一環，一開始我就直接用jieba的斷詞把train data的title分詞，然後跟keyword混在一起做embedding，再塞進LSTM，經過一些微調最後得到了0.84416的score，一開始覺得還行，但後來名次就一直往後掉使得我必須找到一個更好的方法，於是上網找NLP做這類題目相關的文章來看。最後在一篇中國人寫的文章中看到他使用不同方法做新聞分類這個題目，令他驚訝的是他用單純的貝氏分類器這種較不複雜的作法竟然得到了比他第一版的LSTM還要好的結果，於是我仿照了貝氏分類的核心觀念跟一些前面做LSTM得到的一些對於資料處理影響結果的小結論自己做了一個類似的分類器，馬上得到了0.91263的score，在經過一些研究又改了幾次最終以0.92321的結果收工，雖然名次只能在差不多中段，但看了許多文章後我覺得如果還有更多時間以及增加對於其他類似統計分類方法的了解，也許可以再往前爬一些。

實作原理

1. 首先我用jieba把train data的title做斷詞，再加上每個title的keyword，可以得到某個label所對應的一個word pool，我去統計每個單詞在各個分類中出現過幾次建成一個字典。

2. 接著餵進test data，test data的title斷詞後加上keyword一樣可以得到一些詞，用這些詞去算各label字典中出現的次數然後除以label字典值的加總，可以得到(n,10)的矩陣，n是title斷詞出的詞數加上keyword詞數的和。

3. 設定參數與normalization，為了不讓某個單詞影響過大，對該單詞所代表的那行(1,10)vector做normalization，我分別做了softmax跟同除該行總合兩種做法，實驗證明後者效果較佳。

4. 為了不讓title或keyword影響過大，分別給他們加上一個參考權重w1和w2，然後將(n,10)矩陣依照column加總，某row是title取出的字則乘以w1，是keyword取出的字則乘以w2，可以得到一個(1,10)的矩陣，最後看哪一類分數較高就選哪行。

結論

下圖是隨機取出train data 5000筆data答錯的confused matrix，label=row predict=column，可以看出6與7最容易搞混，也就是軍事與國際，我有把答錯的新聞拉出來看，蠻多用人腦也判斷不太出來的，軍事與國際混淆的大都是國家之間的戰爭，例如美國與中東各國，畢竟國與國之間發生的戰爭本來就算是國際事務，也算是軍事新聞，在train data沒有明確的判斷標準的情況下此分類器也很難判

斷，但只要是確定只符合一個分類的新聞，此分類器都可以分出來。因此根據資料label的正確率，也許可以得到更佳的score也說不定。

```
[[ 0.  3.  0.  1.  2.  3.  0.  2.  2.  0.]
 [ 9.  0.  1.  1.  0.  1.  1.  0.  0.  1.]
 [ 0.  0.  0.  0.  2.  1.  0.  0.  1.  0.]
 [ 2.  0.  3.  0.  0.  7.  1.  0.  1.  0.]
 [ 0.  1.  0.  0.  0. 13.  1.  4.  0.  0.]
 [ 5.  0.  0. 13.  2.  0.  2.  0.  1.  5.]
 [ 4.  0.  0.  3.  1.  2.  0. 31.  1.  1.]
 [ 9.  0.  0.  2.  0.  5. 10.  0.  1.  1.]
 [ 1.  0.  1.  1.  3.  3.  1.  1.  0.  0.]
 [ 6.  3.  0.  1.  3.  6.  1.  1.  0.  0.]]
```

我有寫一個news_analysis的function，可以看看到底錯在哪，如下圖這篇新聞，正確答案是科技，但實際上分到汽車也沒有什麼不對，畢竟是在討論行車記錄器。另外可以特別注意的是若是兩類都可以的話，從最後給出的分數來看，此分類器確實能找出最佳的兩類，正確答案的科技為第二高分。這類的例子是錯誤中最多的。

```
news_analysis(data_f.iloc[6])
```

```
title: 行车记录仪内存满了怎么办？
['内存', '记录仪', '行车', '怎么办']
[0.0, 0.0, 0.0, 0.009, 0.0, 0.861, 0.0, 0.0, 0.0, 0.13]
[0.0, 0.0, 0.0, 0.884, 0.0, 0.116, 0.0, 0.0, 0.0, 0.0]
[0.0, 0.0, 0.0, 0.935, 0.0, 0.065, 0.0, 0.0, 0.0, 0.0]
[0.023, 0.024, 0.129, 0.117, 0.282, 0.175, 0.088, 0.035, 0.072, 0.056]
keyword: nan
label: 5 科技
predict: 3 汽車
[0.013, 0.017, 0.206, 1.8980000000000001, 0.286, 1.1360000000000001, 0.095, 0.036, 0.095, 0.217]
```

同類型例子二：最常見的就是這種軍事新聞，就算給人分有時也會分到國際，所以分類起給這兩類也給了22、23差不多的分數

```
title: 战火烧到边境，枪炮声不断，紧急预案启动，真的没有岁月静好
['静好', '枪炮声', '火烧', '岁月', '预案']
[0.091, 0.0, 0.455, 0.0, 0.0, 0.0, 0.273, 0.091, 0.091, 0.0]
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.167, 0.833, 0.0, 0.0]
[0.375, 0.0, 0.0, 0.125, 0.0, 0.0, 0.125, 0.125, 0.25, 0.0]
[0.283, 0.1, 0.0, 0.05, 0.05, 0.0, 0.067, 0.35, 0.067, 0.033]
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.5, 0.5, 0.0, 0.0]
keyword: 马六甲海峡,孟加拉湾,常任理事国,大国崛起,缅甸
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.611, 0.389, 0.0, 0.0]
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.5, 0.417, 0.083, 0.0]
[0.0, 0.0, 0.0, 0.022, 0.0, 0.0, 0.378, 0.6, 0.0, 0.0]
[0.0, 0.0, 0.286, 0.0, 0.143, 0.0, 0.571, 0.0, 0.0, 0.0]
[0.01, 0.02, 0.025, 0.01, 0.005, 0.0, 0.142, 0.741, 0.025, 0.02]
label: 6 軍事
predict: 7 國際
[0.513, 0.22200000000000003, 4.694, 0.332, 1.288, 0.0, 22.745, 23.079, 1.893, 0.215]
```

還有一種新聞也無法判斷，像是下面這種，不可思議的經歷跟答案的科技也毫無關係，看上去也想不到是分哪類，keyword也沒有提供，自然是分不出來。

```
title: 你有过哪些不可思议的经历？
['不可思议', '哪些', '经历']
[0.163, 0.143, 0.0, 0.041, 0.0, 0.122, 0.163, 0.184, 0.0, 0.184]
[0.089, 0.061, 0.063, 0.095, 0.138, 0.196, 0.08, 0.079, 0.045, 0.152]
[0.132, 0.075, 0.019, 0.028, 0.132, 0.321, 0.028, 0.066, 0.028, 0.17]
keyword: nan
label: 5 科技
predict: 9 電競
[0.256, 0.23299999999999998, 0.148, 0.125, 0.316, 0.522, 0.34199999999999997, 0.402, 0.111, 0.544]
```

然後訓練資料中也可以找到一些本來就標錯的data，像是下面這篇毫無疑問就是在討論王者榮耀，關鍵字中也可以看到不知火舞、英雄聯盟、玩家、遊戲等字，但訓練資料卻給了科技這類，這類的例子甚至蠻多個的。

title: 王者荣耀最多还能火多长时间？
['王者', '多长时间', '荣耀', '最多']
[0.004, 0.012, 0.0, 0.009, 0.002, 0.023, 0.002, 0.001, 0.0, 0.948]
[0.018, 0.089, 0.0, 0.214, 0.107, 0.232, 0.214, 0.036, 0.071, 0.018]
[0.003, 0.012, 0.0, 0.002, 0.0, 0.086, 0.005, 0.001, 0.0, 0.892]
[0.026, 0.026, 0.158, 0.132, 0.053, 0.079, 0.184, 0.184, 0.0, 0.158]
keyword: 游戏,王者荣耀,不知火舞,玩家,网络游戏,妈妈是超人,奔跑吧,英雄联盟,鲁班
[0.006, 0.002, 0.0, 0.001, 0.004, 0.037, 0.003, 0.001, 0.001, 0.946]
[0.005, 0.004, 0.0, 0.001, 0.003, 0.089, 0.004, 0.0, 0.0, 0.895]
[0.032, 0.0, 0.0, 0.0, 0.0, 0.016, 0.0, 0.0, 0.0, 0.952]
[0.013, 0.001, 0.001, 0.003, 0.001, 0.012, 0.001, 0.001, 0.0, 0.969]
[0.01, 0.0, 0.0, 0.0, 0.123, 0.056, 0.005, 0.0, 0.0, 0.805]
[0.959, 0.0, 0.0, 0.005, 0.032, 0.005, 0.0, 0.0, 0.0, 0.0]
[0.898, 0.083, 0.0, 0.005, 0.005, 0.003, 0.0, 0.0, 0.0, 0.005]
[0.007, 0.009, 0.0, 0.001, 0.005, 0.028, 0.0, 0.0, 0.0, 0.949]
[0.0, 0.0, 0.0, 0.0, 0.024, 0.006, 0.0, 0.0, 0.0, 0.97]
label: 5 科技
predict: 9 電競
[18.541, 1.231, 0.268, 0.41900000000000004, 2.599, 2.2830000000000004, 0.621, 0.254, 0.118, 67.655]

同類型例子二

title: 王者荣耀职业选手梦泪现在在职业选手里还是不是个人能力最强的？
['选手', '职业', '梦泪', '王者', '荣耀']
[0.114, 0.332, 0.0, 0.005, 0.041, 0.009, 0.009, 0.0, 0.0, 0.491]
[0.005, 0.071, 0.003, 0.011, 0.464, 0.044, 0.019, 0.008, 0.041, 0.332]
[0.0, 0.056, 0.0, 0.0, 0.009, 0.0, 0.0, 0.0, 0.0, 0.935]
[0.004, 0.012, 0.0, 0.009, 0.002, 0.023, 0.002, 0.001, 0.0, 0.948]
[0.003, 0.012, 0.0, 0.002, 0.0, 0.086, 0.005, 0.001, 0.0, 0.892]
keyword: nan
label: 1 體育
predict: 9 電競
[0.08600000000000001, 0.40599999999999997, 0.006, 0.02, 0.555, 0.123, 0.041, 0.011, 0.057, 3.6950000000000003]

同類型例子三

title: 绝地求生全军出击载具大全 绝地求生全军出击有什么载具
['载具', '求生', '绝地', '出击', '全军']
[0.0, 0.0, 0.0, 0.071, 0.0, 0.0, 0.071, 0.0, 0.0, 0.857]
[0.008, 0.009, 0.002, 0.005, 0.002, 0.018, 0.031, 0.011, 0.001, 0.913]
[0.009, 0.008, 0.003, 0.004, 0.001, 0.014, 0.035, 0.017, 0.001, 0.909]
[0.034, 0.085, 0.0, 0.068, 0.017, 0.051, 0.237, 0.119, 0.017, 0.373]
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.603, 0.0, 0.0, 0.397]
keyword: UAZ,吉普车,Dacia,绝地求生,全军
[0.0, 0.0, 0.0, 0.667, 0.0, 0.0, 0.333, 0.0, 0.0, 0.0]
[0.027, 0.0, 0.0, 0.351, 0.081, 0.0, 0.162, 0.0, 0.0, 0.378]
[0.0, 0.0, 0.0, 0.75, 0.0, 0.0, 0.25, 0.0, 0.0, 0.0]
[0.009, 0.005, 0.001, 0.002, 0.0, 0.026, 0.02, 0.011, 0.001, 0.925]
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.603, 0.0, 0.0, 0.397]
label: 6 軍事
predict: 9 電競
[0.261, 0.118, 0.028, 14.888, 0.982, 0.253, 17.762, 0.28500000000000003, 0.046, 20.377]