作業三Report B05702095 會計四 黃禹翔

1. 執行環境

Windows PowerShell  
  
2. 程式語言

Python 3.7  
  
3. 執行方式

$ python hw3-b05702095.py



我使用的非原生套件有：

1. nltk：協助文字的前處理

2. numpy：協助數值、向量、矩陣的運算

3. num2words：協助把文字的數字加到停用字裡

以上三個套件需要 pip install。

我把檔案路徑在程式裡面寫死了，如果需要重新執行，需要進到資料夾裡執行。

4. 作業處理邏輯說明

第一步先把訓練資料吃進來。

接著做資料前處理，這部分與我的作業一樣，基本上就是先Tokenize、去

除停用字、再取詞幹。

然後先訓練最基本的 Multi-Nomial Naïve Bayes Classifier與Bernoulli Naïve

Bayes Classifier，利用 add-one smoothing 解決可能出現機率為零的情況。

兩者的訓練都是先算出先驗各類別的機率，以及在各類別的出現的條件

下，各個term出現的機率。預測的時候在裡用這些機率取log相加，把出

現最大機率的類別挑出來。如果遇到測試資料裡有沒出現過的term

(OOV)，就直接忽略不計。

最後，再利用chi-score特徵選擇，挑出前五百重要的特徵，重新預測一

次，預測方式與前面雷同，只是再預測時把不是前五百重要的term直接去

掉，讓模型專注在chi-score高的特徵上。

最後Kaggle Public Score結果如下：

|  |  |  |
| --- | --- | --- |
|  | Multi-Nomial NB | Bernoulli NB |
| 無feature selection | 0.92222 | 0.94444 |
| Chi-score feature selection | 0.96444 | 0.96888 |

可以看到，利用Chi-score做特徵選擇確實可以讓模型的表現有所提升。而

在這次的任務上，Bernoulli model的表現要比 multinomial model來得好。