IR HW3 資管碩二 R05725034 張鑑霖

1. 執行環境 & 作業系統

Jupyter & win7

1. 程式語言

Python3

1. 作業處理邏輯說明

建立前處理後的每篇文章，存在doc\_tf/

建立training data (class 1~13)的dictionary

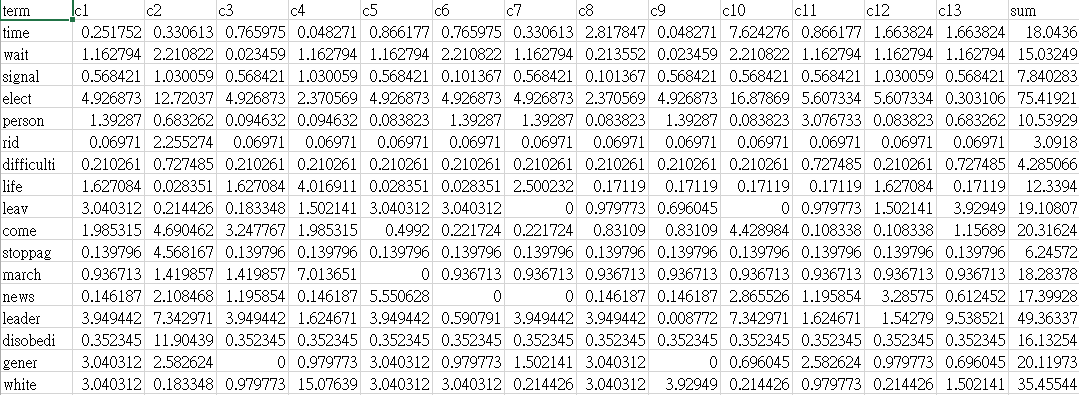
**前處理：**

對此dictionary的每個term算它在每個class的likelihood ratio

Likelihood ratio: n11 = term在class中出現次數，n10 = 15-n11

n01 = term沒在class中出現次數，n00 = 180 – n01，其結果大概如下圖，

最後我們家總後排序取前500個當term



**Training:**

開啟文章，根據每個class的training文章裡面的term算出現的次數，例如opposite在 各個class的出現次數，有了這個就可以根據此數據算出機率每個term對應到每個class的機率。



**Testing:**

最後每篇文章的class分數都從0開始，並根據加上上面train出來的機率，最大的就是哪個class