P T T 文章分析與統計

使用工具

PYTHON

1. Scrapy 爬取資料
2. MongoDB 儲存資料
3. pandas / seaborn / matplotlib / jieba / numpy / Jupyter Notebook

已做進度

資料蒐集

來源

1. 政黑板：https://www.ptt.cc/bbs/HatePolitics/index.html
2. 女版：https://www.ptt.cc/bbs/WomenTalk/index.html

實作方法

1. 爬取一週的資料，共 19261 篇文章，53646 位使用者資料

問題

1. 八卦版以往以 cookies 選取 18 歲以上選項的方法失效，導致無法正確取得文章，目前只有約三天的資料。


```
_id: ObjectId("5cd3db0f9fb0bb5f9d2c1255")
board: "Gossiping"
title: "[問卦] 有沒有數學課的八卦?"
category: "問卦"
author: "wellYM"
date: 2019-05-09T15:44:31.000+00:00
      "

      在高雄某國小數學課堂上

text: 數學老師問: 3 X 4 = ?

      韓同學回答說: 3+4=7 我不是不會數學。謝謝！（坐下）

          ..."
› img_link: Array
          "125.227.143.128
ip_author: "
url: "https://www.ptt.cc/bbs/Gossiping/M.1557387873.A.DFC.html"
› comment: Array
score: -1
› text_cut: Array
› text_speech: Array
```


- 資料庫內容

利用 IP 篩選正常使用者

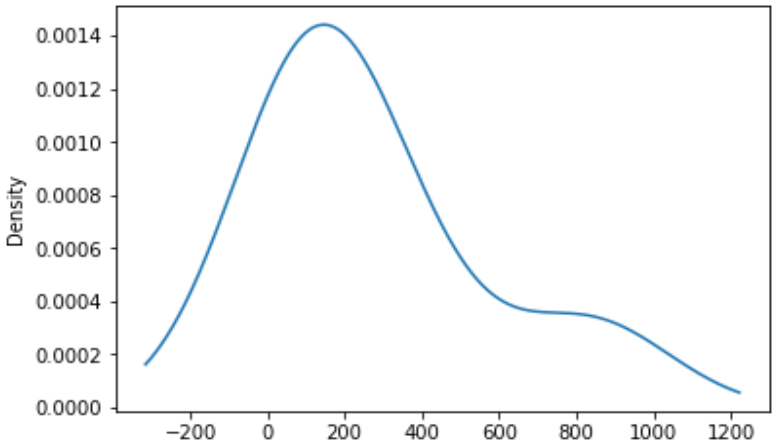
1. 有發過文&推文的使用者：53351 人
2. 有發過文(有被記錄 IP)：5297 人
3. 若該使用者 IP 數量為離群值，列為非正常使用者

經離群值淘汰前

IP 個數	使用者人數
1~5	4988
6~10	226
11~15	51
16~20	15
21~30	13
30~40	3
55~60	1

經離群值淘汰後

IP 個數	使用者人數
1	3498
2	838
3	351
4	182
5	119
總和	4988



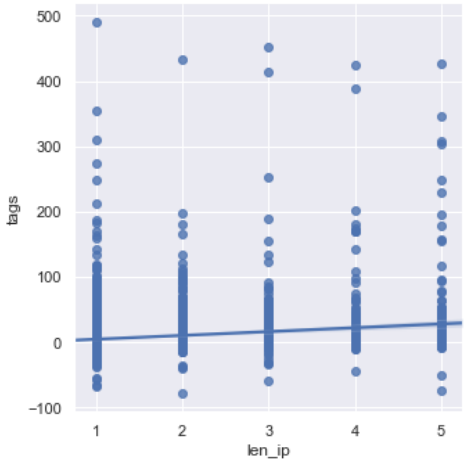
- KDE 核密度圖

CODE

https://github.com/aqweteddy/CourseStatistic/blob/master/ipAnalsis.ipynb

活動數量和 IP 數量關係

1. 以 ip 數量做為 x 軸、發文＋推噓文數量作為 y 軸，計算相關係數
2. 去除 ip 數量所得的離群值
3. r = 0.223373，很明顯，IP 數量和活動數量沒有關係



- 點圖與回歸直線

CODE

https://github.com/aqweteddy/CourseStatistic/blob/master/ipAnalsis.ipynb

政黑版不正常 IP 個數的平均數會大於母體平均數

1. 以 Z 檢定驗證
   - ✧ 我們假設政黑版 IP 個數的平均數會大於母體平均數，所以假設 H_0: μ = 平均數， H_1: μ > 平均數(主張)

population_mean	2.73
sample_mean	1.52
stdev	1.51
z	16.52

CODE

https://github.com/aqweteddy/CourseStatistic/blob/master/ipAnalsis.ipynb

未來想做的議題

蒐集哪一天有甚麼風波，之後針對它做分析

1. 以假設檢定驗證
2. 困難點:需要斷詞，但需要大量時間處理，目前已在學校工作站上斷詞，但不知道何時可以處理完
    - ✧ CODE: https://github.com/aqweteddy/CourseStatistic/blob/master/cut_db_content.py

文章有什麼字的時候推文特別多

1. 困難點:需要斷詞，但需要大量時間處理，目前已在學校工作站上斷詞，但不知道何時可以處理完
2. 已有小樣本的測試