2018
下學期機器學習
Final Project

組員 : 0616018 林哲宇 0616084 林柏均
0516017 李柏毅 0313340 邱韡
0413233 李彥儒

一、題目與動機

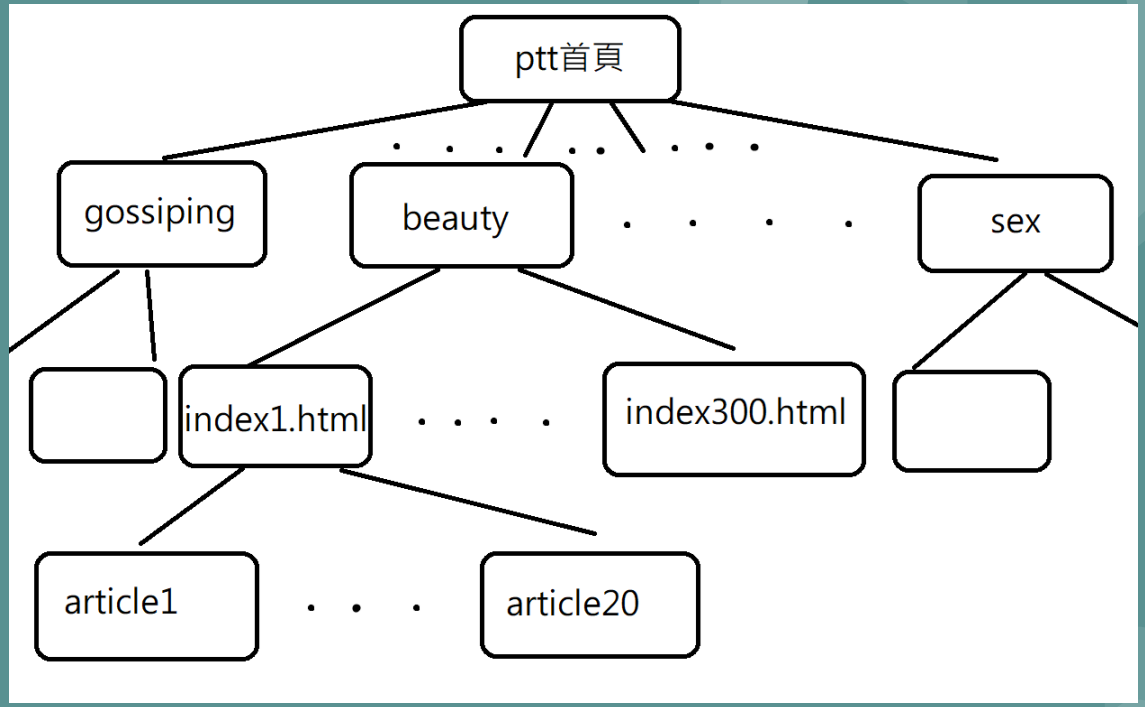
題目:文章檢測器

動機:

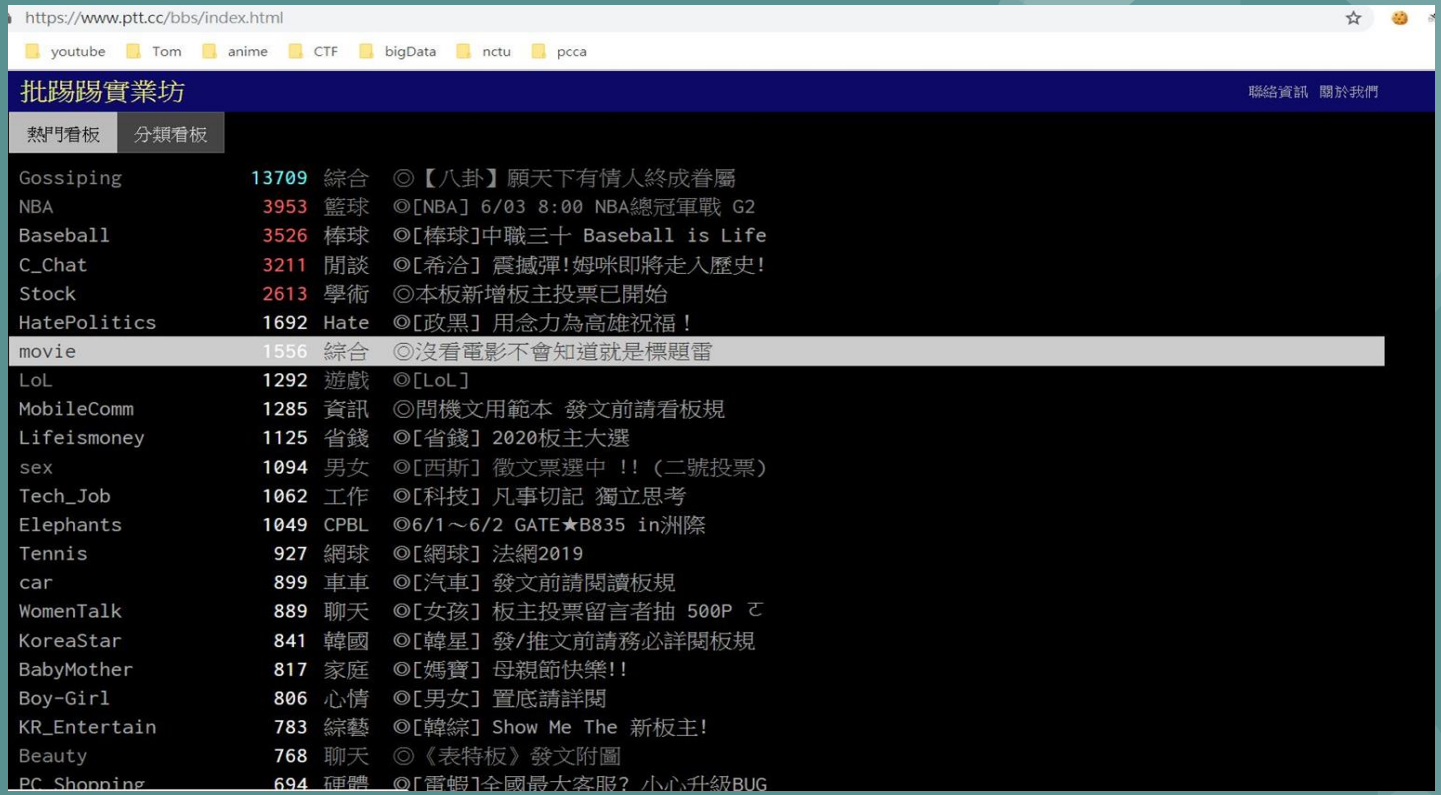
1. ptt 版主目前的工作

2. 本身的 ptt 使用經驗

二、取得資料


ptt首頁
gossiping
beauty
sex
index1.html
index300.html
article1
article20

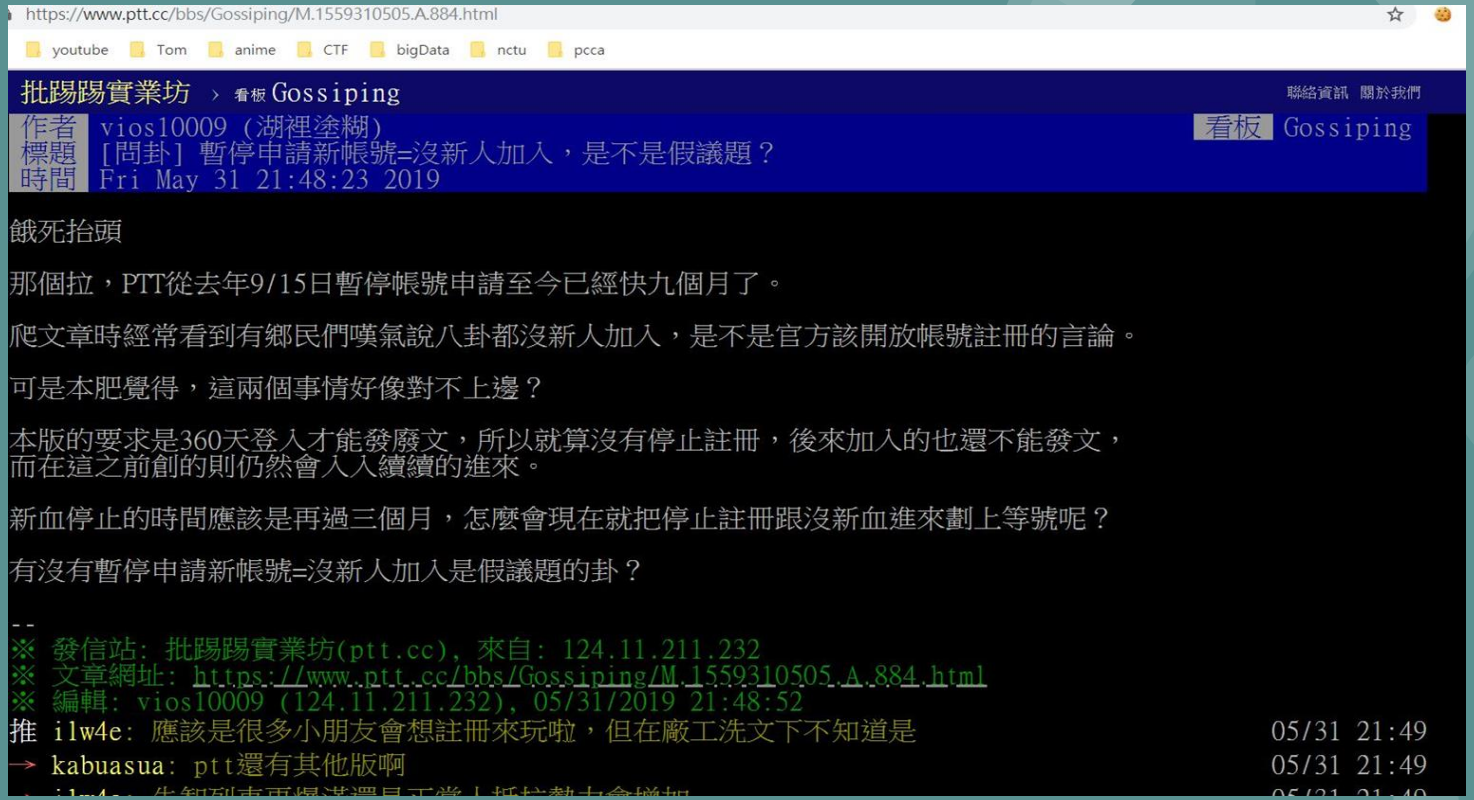
ppt首頁


https://www.ptt.cc/bbs/index.html
youtube
Tom
anime
CTF
bigData
nctu
pcca
批踢踢實業坊
聯絡資訊 關於我們
熱門看板
分類看板
Gossiping 13709 綜合 ◎【八卦】願天下有情人終成眷屬
NBA 3953 籃球 ◎[NBA] 6/03 8:00 NBA總冠軍戰 G2
Baseball 3526 棒球 ◎[棒球]中職三十 Baseball is Life
C_Chat 3211 閒談 ◎[希洽] 震撼彈!姆咪即將走入歷史!
Stock 2613 學術 ◎本板新增板主投票已開始
HatePolitics 1692 Hate ◎[政黑] 用念力為高雄祝福！
movie 1556 綜合 ◎沒看電影不會知道就是標題雷
LoL 1292 遊戲 ◎[LoL]
MobileComm 1285 資訊 ◎問機文用範本 發文前請看板規
Lifeismoney 1125 省錢 ◎[省錢] 2020板主大選
sex 1094 男女 ◎[西斯] 徵文票選中 !! (二號投票)
Tech_Job 1062 工作 ◎[科技] 凡事切記 獨立思考
Elephants 1049 CPBL ◎6/1～6/2 GATE★B835 in洲際
Tennis 927 網球 ◎[網球] 法網2019
car 899 車車 ◎[汽車] 發文前請閱讀板規
WomenTalk 889 聊天 ◎[女孩] 板主投票留言者抽 500P ㄛ
KoreaStar 841 韓國 ◎[韓星] 發/推文前請務必詳閱板規
BabyMother 817 家庭 ◎[媽寶] 母親節快樂!!
Boy-Girl 806 心情 ◎[男女] 置底請詳閱
KR_Entertain 783 綜藝 ◎[韓綜] Show Me The 新板主!
Beauty 768 聊天 ◎《表特板》發文附圖
PC_Shopping 694 硬體 ◎[電蝦]全國最大客服? 小心升級BUG

ptt 分類 index


https://www.ptt.cc/bbs/Gossiping/index.html
youtube
Tom
anime
CTF
bigData
nctu
pcca
批踢踢實業坊 › 看板 Gossiping
聯絡資訊 關於我們
看板
精華區
最舊
‹ 上頁
下頁 ›
最新
搜尋文章…
4 [問卦] 暫停申請新帳號=沒新人加入，是不去假議
vios10009 5/31
7 [問卦] 白石 茉莉奈這天是我最後能為妳做的事....
alan4ni 5/31
1 [問卦] 老闆把我的雞排切了老板切了我的雞排
hongou 5/31
[新聞] 蘇丹示威抗議不斷 軍方關閉半島電視台
JyumonjiKaho 5/31
[問卦] 西門町沒落時是什麼樣子
w854105 5/31
3 [新聞] 一條毛巾在太空飄了10年 終於被太空人取回
maruEX 5/31
[問卦] 手機吃到飽是台灣特有的產物嗎?
safelyfuck 5/31
[問卦] 爺孫戀還在嗎
OGC100times 5/31
[問卦] 這樣算抓到女同學是鄉民嗎?

文章內頁

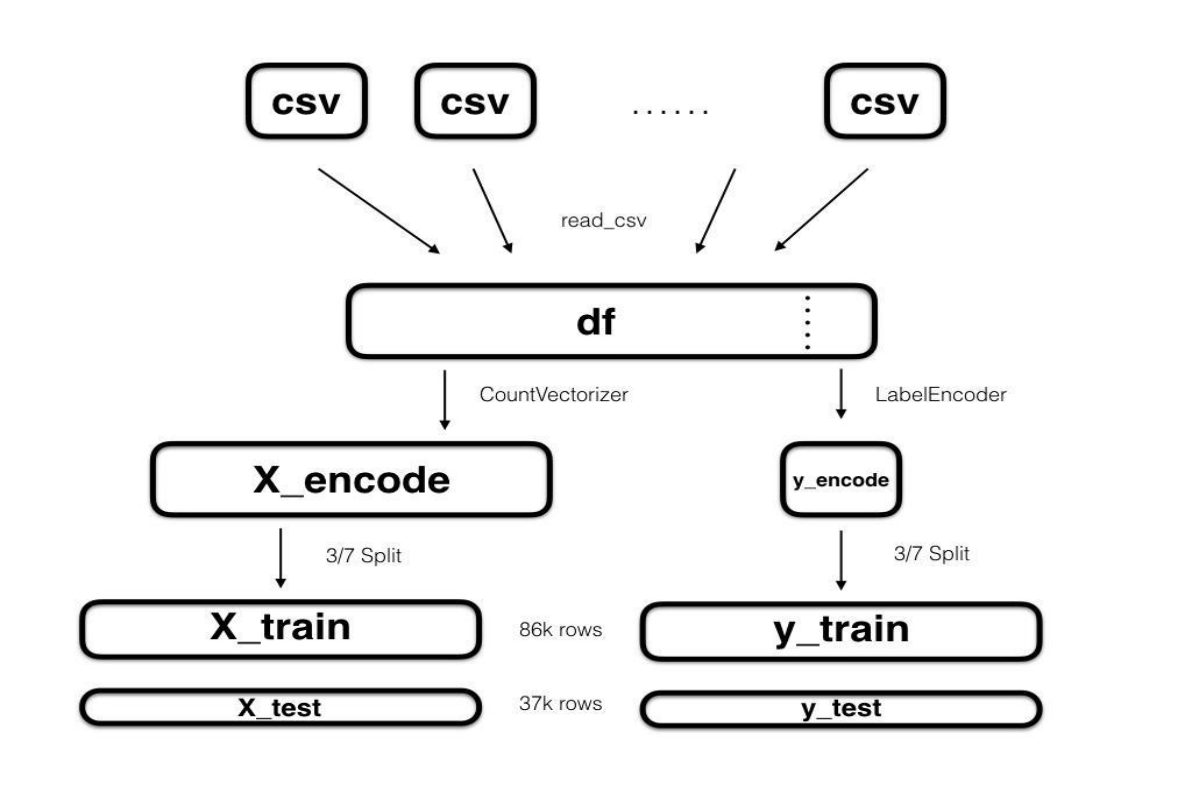

https://www.ptt.cc/bbs/Gossiping/M.1559310505.A.884.html
youtube
Tom
anime
CTF
bigData
nctu
pcca
批踢踢實業坊 › 看板 Gossiping
聯絡資訊 關於我們
作者 vios10009 (湖裡塗糊)
標題 [問卦] 暫停申請新帳號=沒新人加入，是不是假議題？
時間 Fri May 31 21:48:23 2019
看板 Gossiping
餓死抬頭
那個拉，PTT從去年9/15日暫停帳號申請至今已經快九個月了。
爬文章時經常看到有鄉民們嘆氣說八卦都沒新人加入，是不是官方該開放帳號註冊的言論。
可是本肥覺得，這兩個事情好像對不上邊？
本版的要求是360天登入才能發廢文，所以就算沒有停止註冊，後來加入的也還不能發文，
而在這之前創的則仍然會人入續續的進來。
新血停止的時間應該是再過三個月，怎麼會現在就把停止註冊跟沒新血進來劃上等號呢？
有沒有暫停申請新帳號=沒新人加入是假議題的卦？
--
※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 124.11.211.232
※ 文章網址: https://www.ptt.cc/bbs/Gossiping/M.1559310505.A.884.html
※ 編輯: vios10009 (124.11.211.232), 05/31/2019 21:48:52
推 ilw4e: 應該是很多小朋友會想註冊來玩啦，但在廠工洗文下不知道是 05/31 21:49
→ kabuasua: ptt還有其他版啊 05/31 21:49

輸出格式

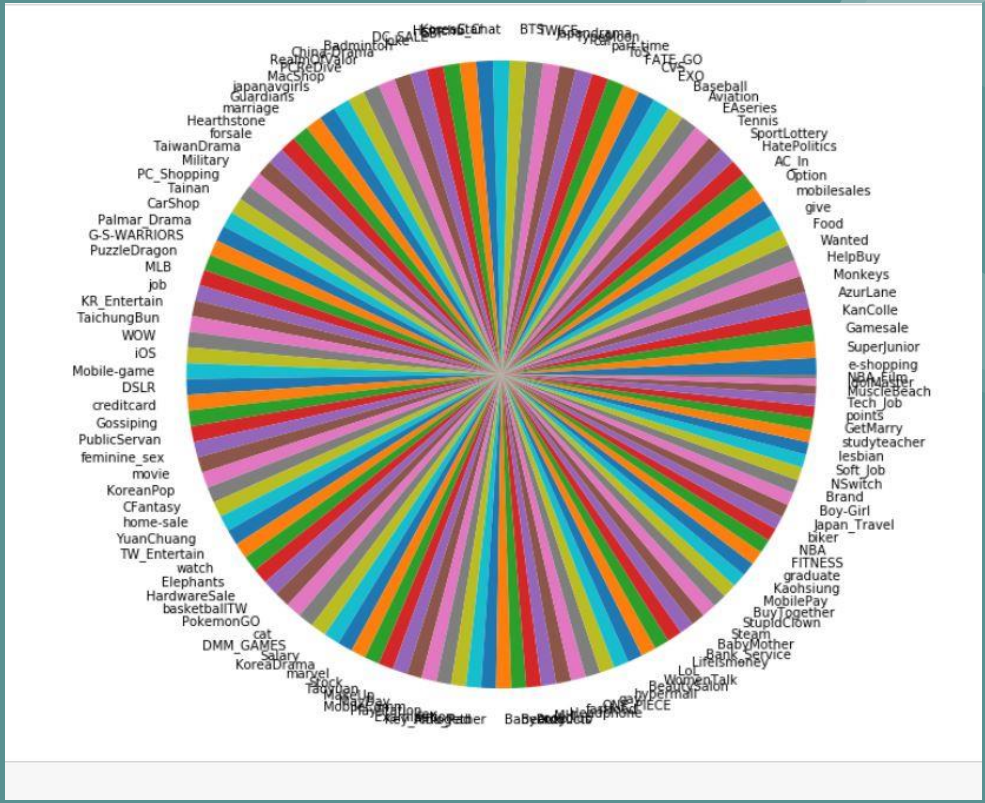
id	Category	words
1	Gossiping	"明星","外遇","柏均"
2	NBA	"投出","五個","三分球"
3	tech_job	"監督","公司","出貨"

id	Category	words
1	Gossiping	"明","星","外","遇","柏","均"
2	NBA	"投","出","五","個","三","分","球"
3	tech_job	"監","督","公","司","出","貨"

三、資料前處理


csv
csv
.....
csv
read_csv
df
CountVectorizer
LabelEncoder
X_encode
y_encode
3/7 Split
3/7 Split
X_train
86k rows
y_train
X_test
37k rows
y_test

資料視覺化



128 categories

same ratio articles

四、訓練模型

i.Bayes

ii.Desicion Tree

iii.Random Forest

iv.NN

v.Linear Regression

預測結果

Accuracy

Bayes gaussian:0.44

Desicion Tree : 0.63

Random Forest : 0.55

NN : 0.799

Linear Regression :0.43

五、遇到難題

1. 抓取完整的資料遇到的各種狀況

solved

2. 透過 jieba 所得到字詞分類會讓整個 array 變得太大

solved……but too late

3. 希望能處理各國字詞與圖片

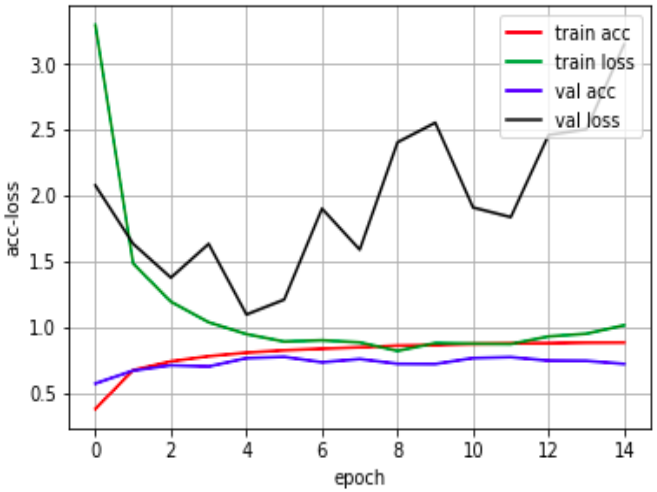
cannot be done before deadline

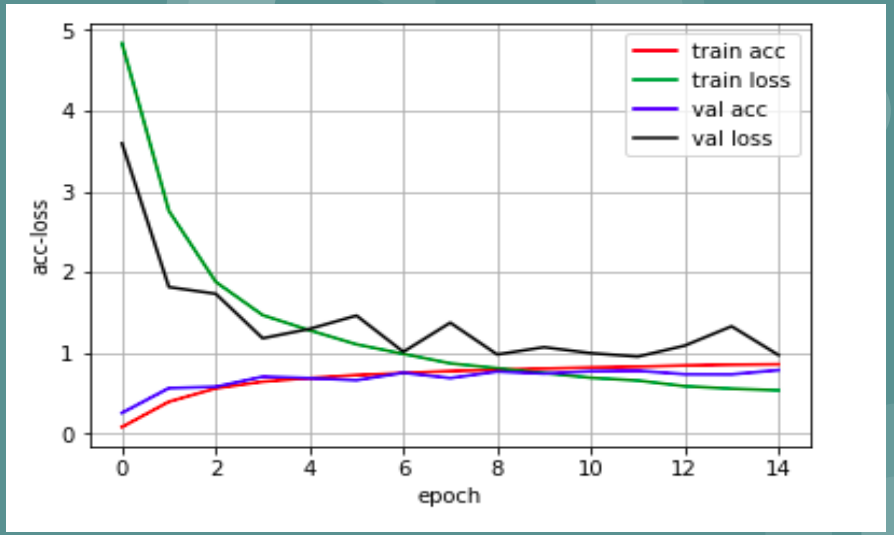
解決方法

Gossiping	NBA	tech_job
"明星","外遇","柏均"	"投出","五個","三分球"	"監督","公司","出貨"

"明星","在","公司","投出","三分球"				
id	category	Gossiping	NBA	tech_job
1	gossiping	1	2	1

nn overfitting






```
37091/37091 [==============================] – 15s 410us/step
acc: 0.6069936103113857
```

```
37091/37091 [==============================] – 13s 351us/step
acc: 0.7996549028082177
```

謝謝大家