

# 金融科技與文字探勘

K O L、網路情緒指標與投資人行為之早期預測

組員：

台大財金所碩一 謝侑軒  
台大農經系大六 莊浩平  
師大數學所碩一 蔡承男  
政大企管系大三 曾琮傑  
政大科智所碩一 劉珈卉

指導老師：石百達、張智星 老師

指導業師：Tony / 黃俊哲

野村投信 題目一  
組別二

# Agenda

01

研究目的

02

研究方法

03

PTT研究結果

04

強基金研究結果

05

結論與應用程式



# 研究目的



# 研究目的

## ● 現行狀況

- 理想：長期的穩健投資及勝率為最高的投資原則
- 現實：投資人易受短期市場影響，追高殺低

## ● 問題

- 市場事件個人解讀不同
- 資產管理業者只能被動出擊，無法事前協助投資人
  - (如：發布研究報告、個別溝通來改變投資人偏誤)

## ● 解決方法

- 建立贖回率變動指標，讓業者可事前/即時與投資人溝通



# 研究方法



# 研究對象

將此兩種資料來源合併分析，得到更完整的輿情結果。

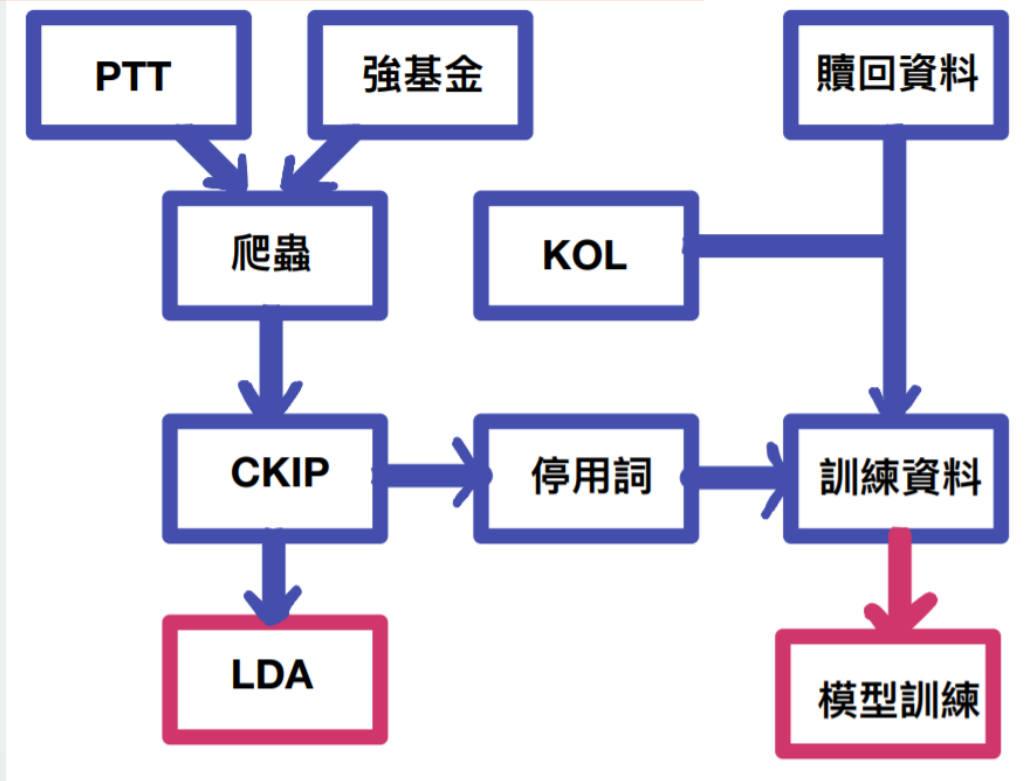
	PTT_股票版	強基金論壇
樣本來源	內文	內文 + 回文
使用者習慣	閒聊	詢問申購事宜
文章種類標籤	有	無
發文、回文頻率	高(每日50篇↑)	低
內容主要組成	閒聊、標的分享、新聞	理財詢問、投資建議
使用者	年輕群眾	多為青壯年
選擇原因	Stock版為熱門看板，討論度比Fund版熱烈，產生較多意見領袖	專注討論基金相關議題，選取資料涵蓋各個討論版



# 研究方法

透過KOL的文本來預測贖回率是否上升

- 指標建立
  - KOL定義
  - 贖回率指標
- 研究方法
  - 斷詞
  - 找尋主題
  - 找尋關鍵字
- 模型結果
  - 字典法+傳統機器學習
  - 自然語言處理-BERT



# 研究方法 PTT KOL指標分析

KOL

	指標一	指標二	指標三	指標四
挑選標準	文章數前十名	推文數超過30的文章前十名	扣除 "新聞"與"其他" 後發文數最多	PageRank "心得"與"標的"
文章組成	新聞	公告、閒聊	心得、標的	心得、標的
優缺點	新聞 非個人見解	公告 較無參考意義	文章數大幅減少，比起NLP方法較適合使用字典法	
選取與否				



# KOL指標

指標三：

扣除「新聞」及「其他」後，發文數最多

author	
vendan5566	264
hito21	175
justforsing	137
hrma	128
tim0259	114
epson5566	102
j2708180	100
zesonpso	90
Ejaculation	76
zakjudelo	61

	time	author	type	title	score	filename
9641	2018-10-01 05:18:14	epson5566	標的	2330 台積電 空	39	pttdata\M.1538342296.A.937.json
9796	2018-10-04 00:59:41	epson5566	標的	9955 佳能多	7	pttdata\M.1538585983.A.AD6.json
9859	2018-10-05 01:52:46	epson5566	標的	2330 中性偏空	7	pttdata\M.1538675568.A.EEC.json
10000	2018-10-08 10:03:57	epson5566	標的	2330台積電 多	54	pttdata\M.1538964240.A.424.json
10800	2018-10-23 14:40:30	epson5566	謠言	中國的美元儲備是不是有點少啊?	10	pttdata\M.1540276832.A.AA1.json
...	...	...	...	...	...	...
32746	2020-04-02 19:15:18	epson5566	心得	整理一些昨日美股收盤後的期權倉位	11	pttdata\M.1585826120.A.4F4.json
32831	2020-04-05 00:34:56	epson5566	心得	週五美股REITs暴跌	63	pttdata\M.1586018098.A.22B.json
32838	2020-04-05 12:46:54	epson5566	心得	美國PPP計畫的混亂可能衝擊小型企業	13	pttdata\M.1586062018.A.3C1.json
33512	2020-04-17 11:16:25	epson5566	心得	盤勢的一些看法	16	pttdata\M.1587093387.A.CB2.json
33930	2020-04-24 00:39:31	epson5566	心得	整理一些體質良好的個股資料	120	pttdata\M.1587659973.A.D33.json

Epson5566的文章種類分配性質理想

type	
心得	85
標的	10
謠言	7

[心得] 最後的逃命波

看板 [Stock \(股票\)](#)  
作者 [epson5566 \(ep\)](#)  
時間 3月前 (2020/03/24 22:39)  
推噓 133 (166推 33噓 100→)  
留言 299則, 224人參與, 3月前最新  
討論串 [1/4 \(看更多\)](#) < > > >

真的語重心長

千萬不要在這個時機點入場做多了

個人研判 這波反彈的持續時間會非常非常的短

主要理由如下

首先 在今日股市那麼樂觀的情緒下

一些金融指標 幾乎都沒有緩解的情況

推 avrw  
美股24 台股22

推 a989876  
好 我歐印了

推 kurapical106  
epson5566的文章大家可以查一下 有很多可以學習的

→ kurapical106  
地方

→ SkySoldier55  
美股今日暴漲，至少明天還穩定

推 m791226  
27直噴破萬

推 xhs  
感謝政府讓你逃命，之後就是溜滑梯了

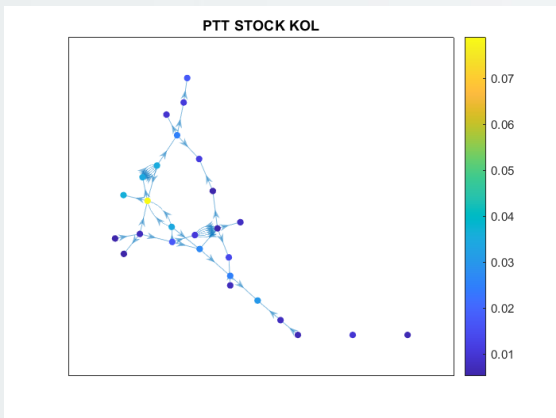
# 版上KOL

指標四：

利用PageRank概念

找出分數最高的人做KOL

(以「心得」與「標的」文發文者為主)



Zesonpso的  
文章類型半數以上為心得文

type	
其他	3
心得	54
新聞	22
標的	20

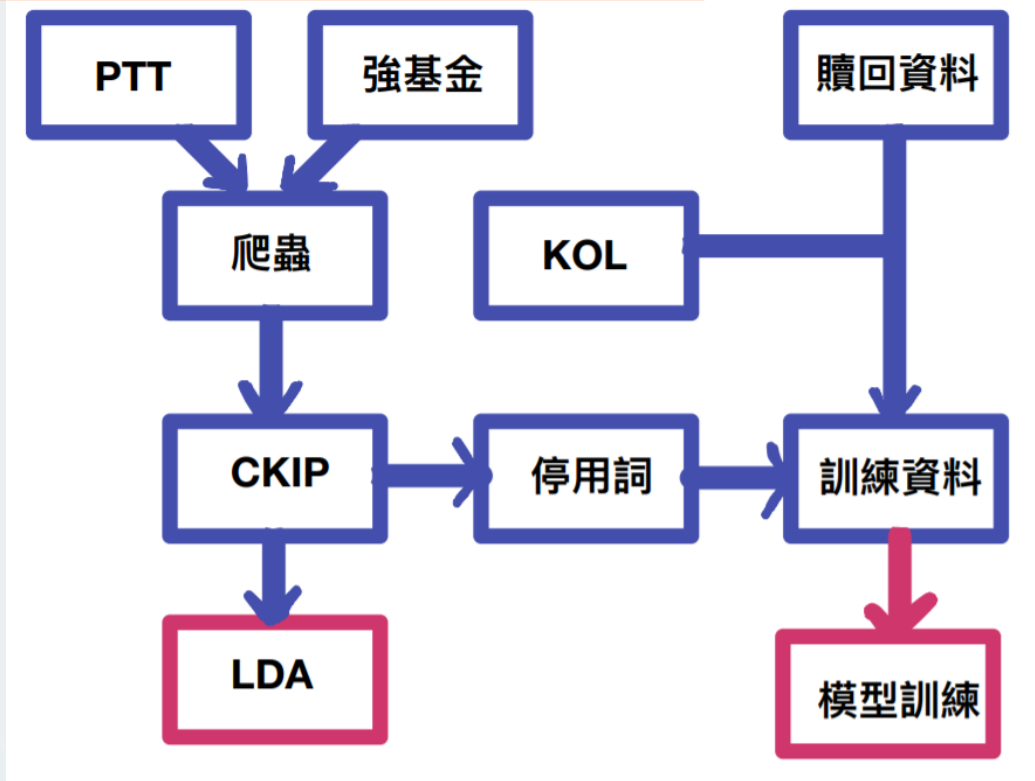
1 Name	2 PageRank	3 InDegree	4 OutDegree
'zesonpso'	0.0790	813	2
'nenin2003'	0.0362	60	0
'LittlePong'	0.0349	125	2
'zzz499'	0.0346	22	6
'civelant'	0.0344	501	3
'drakd4d'	0.0304	76	0
'finalwave'	0.0266	260	1
'jazz8010...	0.0249	42	1
'optimize'	0.0249	218	3
't200830...	0.0175	147	0

推 kiversonx17 : 頭推  
推 nano5421 : 沙發推  
推 clvmomo : 3  
推 uuu5566 : 鬆哥第一 鬆哥第一 鬆哥第一  
推 justin037666 : 推  
推 sismiku : 推  
推 uoho : 推  
推 coolschumi : 終於等到鬆哥文!  
推 oldEn15 : 大多頭開始 股神又紛紛登場了

# 研究方法

## 透過KOL的文本來預測贖回率是否上升

- 指標建立
  - KOL定義
  - 贖回率指標
- 研究方法
  - 斷詞
  - 找尋主題
  - 找尋關鍵字
- 模型結果
  - 字典法+傳統機器學習
  - 自然語言處理-BERT



# 找尋贖回率指標


根據贖回率相關指標，才能接下來找尋有影響力的關鍵字，或是進行模型預測

## 贖回率指標建立步驟

1. 收集國內基金贖回率資料
2. 鎖定「野村」的「證券型」基金
3. 根據每個月結算資料做算術平均
4. 計算每月差異
5. 將大於上月的月份標出

```
eqt_fund_name = [  
    "野村環球基金",  
    "野村優質基金",  
    "野村台灣高股息基金",  
    "野村全球品牌基金",  
    "野村中國機會基金",  
    "野村歐洲中小成長基金",  
    "野村日本領先基金",  
    "野村e科技基金",  
    "野村高科技基金",  
    "野村中小基金",  
    "野村鴻運基金",  
    "野村台灣通籌基金",  
    "野村成長基金",  
    "野村積極成長基金",  
    "野村大俄羅斯基金",  
    "野村巴西基金",  
    "野村印尼潛力基金",  
    "野村印度潛力基金",  
    "野村泰國基金",  
    "野村新馬基金",  
    "野村全球高股息基金",  
    "野村亞太高股息基金"  
]
```

1



	ym	redemption	delta	signal
60	2019-11-30	0.040687	0.009933	True
61	2019-12-31	0.045125	0.004438	True
62	2020-01-31	0.063234	0.018109	True
63	2020-02-29	0.063234	0.000000	False
64	2020-03-31	0.114438	0.051203	True

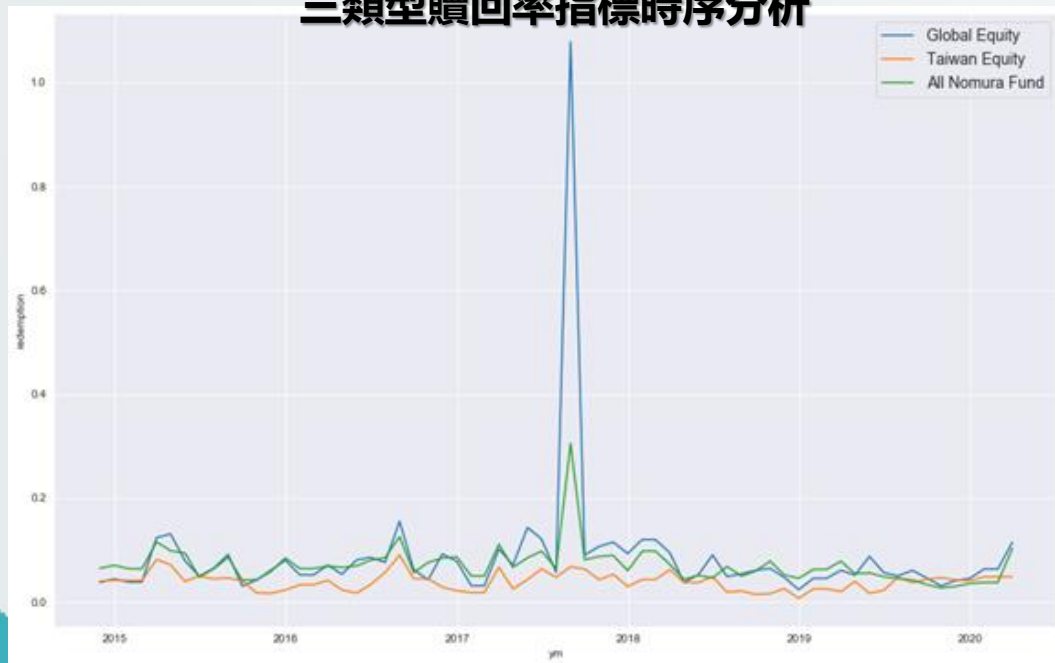
# 找尋贖回率指標

除了野村全部股票型基金，我們也使用同樣方法製作以下兩種資料的贖回率指標

1. 野村投信的「台灣股票型基金」
2. 野村投信的「全部基金」

## 三類型贖回率指標時序分析

```
eqt_fund_name = [  
    # "野村環球基金",  
    # "野村優質基金",  
    "野村台灣高股息基金",  
    # "野村全球品牌基金",  
    # "野村中國機會基金",  
    # "野村歐洲中小成長基金",  
    # "野村日本領先基金",  
    "野村e科技基金",  
    "野村高科技基金",  
    "野村中小基金",  
    "野村鴻運基金",  
    "野村台灣運籌基金",  
    "野村成長基金",  
    "野村積極成長基金",  
    # "野村大俄羅斯基金",  
    # "野村巴西基金",  
    # "野村印尼潛力基金",  
    # "野村印度潛力基金",  
]
```



可以看到大致上贖回率的走勢都滿接近的，除了在**2017下半年**，全球股市型基金贖回率指標有出現一個很大的跳點

在接下來的分析，我們先是使用指標一的**全球股票型基金**來分析，之後會探討改成台灣股票型基金的狀況。

# 研究方法

透過KOL的文本來預測贖回率是否上升

- 指標建立

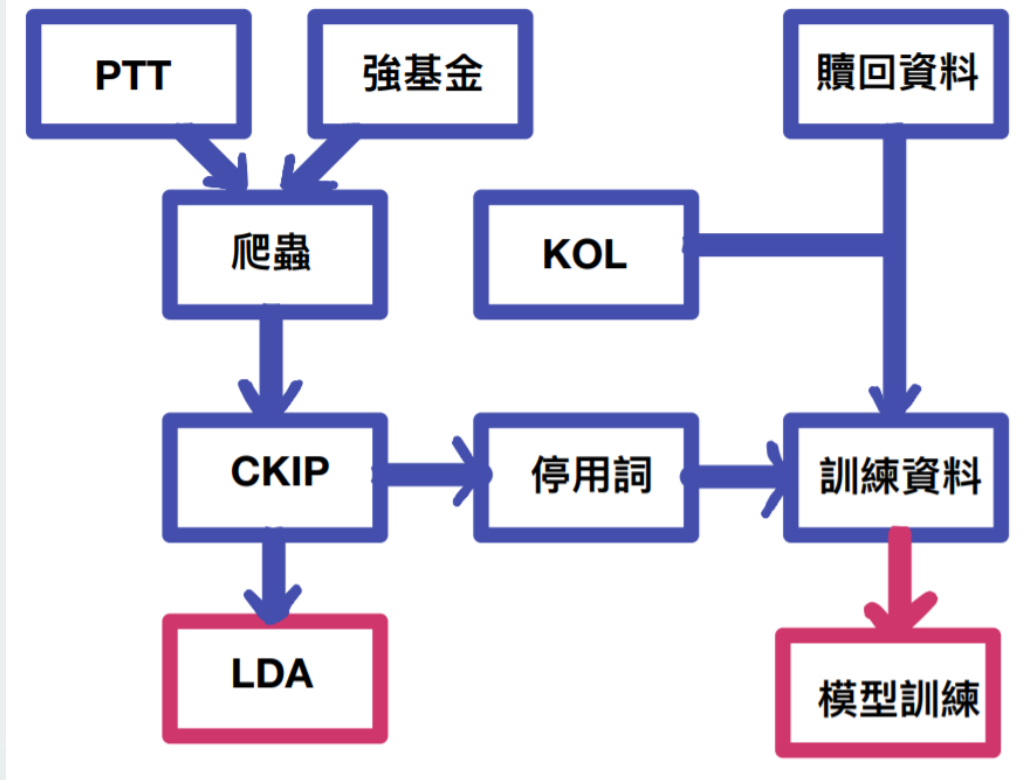
- KOL定義
- 贖回率指標

- 研究方法


- 斷詞
- 找尋主題
- 找尋關鍵字

- 模型結果

- 字典法+傳統機器學習
- 自然語言處理-BERT



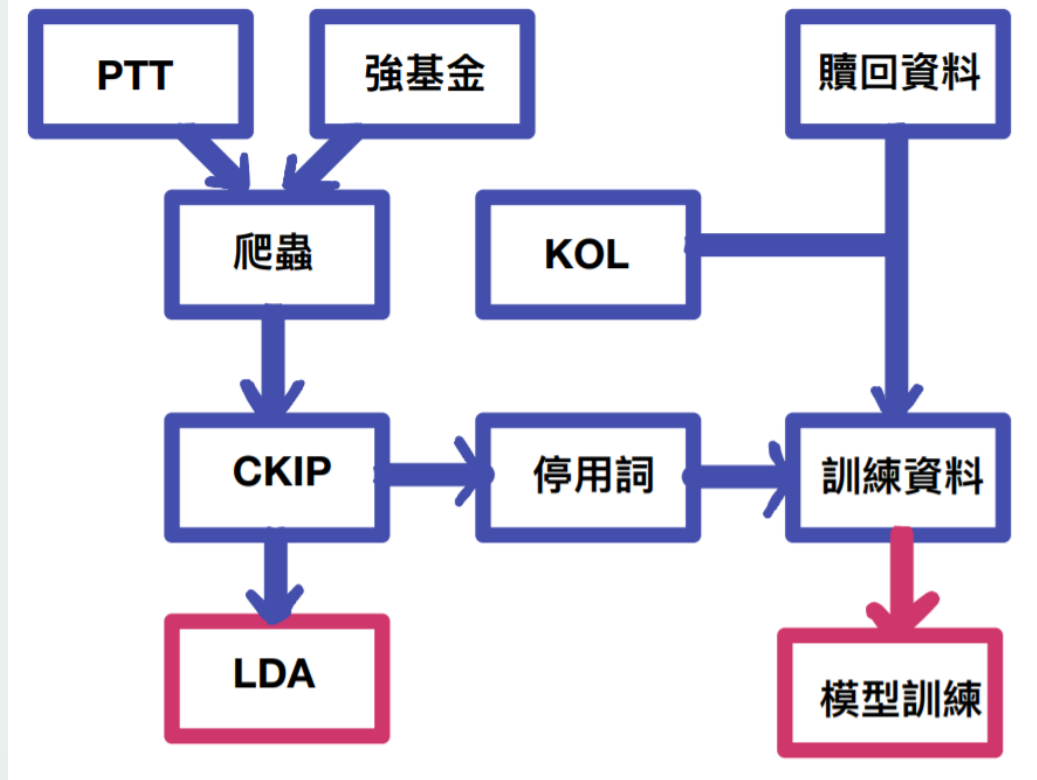
# 找尋關鍵字 -以贖回率指標回頭找有顯著差異的用字(以epson5566為例)

方法	說明	結果(K-fold AUC)
一	<ol style="list-style-type: none"> <li>選出限定的資料子集</li> <li>篩選贖回率上升/未上升時，文章的前80個用字。</li> <li>鎖定各自前20個字，比對後將與前80字重複者刪除。</li> </ol>	Mean: 0.8792 Std: 0.1242
二	<ol style="list-style-type: none"> <li>Scikit-Learn建立文章的TF-IDF矩陣，分數越高表示某個字在該篇文章中的重要性與代表性較高。</li> <li>各個字在各文章的分數加總，取出TF-IDF最高的前50字作標準。</li> </ol>	Mean: 0.7000 Std: 0.2005 直接用TF-IDF，未做類似方法一的分類，預測效果較差
三	<ol style="list-style-type: none"> <li>依照label分類，再使用Scikit-Learn分別建立文章的TF-IDF矩陣，分數越高表示某個字在該篇文章中的重要性以及代表性較高。</li> <li>各個字在各label下的分數加總，手動去除無意義和重複出現字詞後，各取出TF-IDF最高的前25個字當作標準。</li> </ol>	Mean: 0.7708 Std: 0.2695 預測效果較穩定，用計量學理常見Logit效果良好，第二名是XGBoost。
 <b>四</b> <b>T-test</b> <b>顯著程度</b>	<ol style="list-style-type: none"> <li>選出限定的資料子集選出</li> <li>依照每個字為單位，統計在各篇文章出現的次數，並標註是正向文章還是負面文章，之後就可以將此兩個序列做2樣本T檢定。(用Y=1次數減去Y=0次數)</li> <li>將T-ratio依照絕對值之後的值排序，找出影響力強烈的字詞。</li> </ol>	Mean: 0.7972 Std: 0.2399 效果與方法一差不多好，但標準差較高，不過比較能有一致性的判斷標準

# 研究方法

## 透過KOL的文本來預測贖回率是否上升

- 指標建立
  - KOL定義
  - 贖回率指標
- 研究方法
  - 斷詞
  - 找尋主題
  - 找尋關鍵字
- 模型結果
  - 字典法+傳統機器學習
  - 自然語言處理-BERT





## ● PyCaret



- 使用PyCaret可以輕鬆的比較各個分類模型的預測力，挑選出適當的模型

## ● Google BERT

- 遷移學習，站在巨人的肩膀上省錢省力





# PTT研究結果



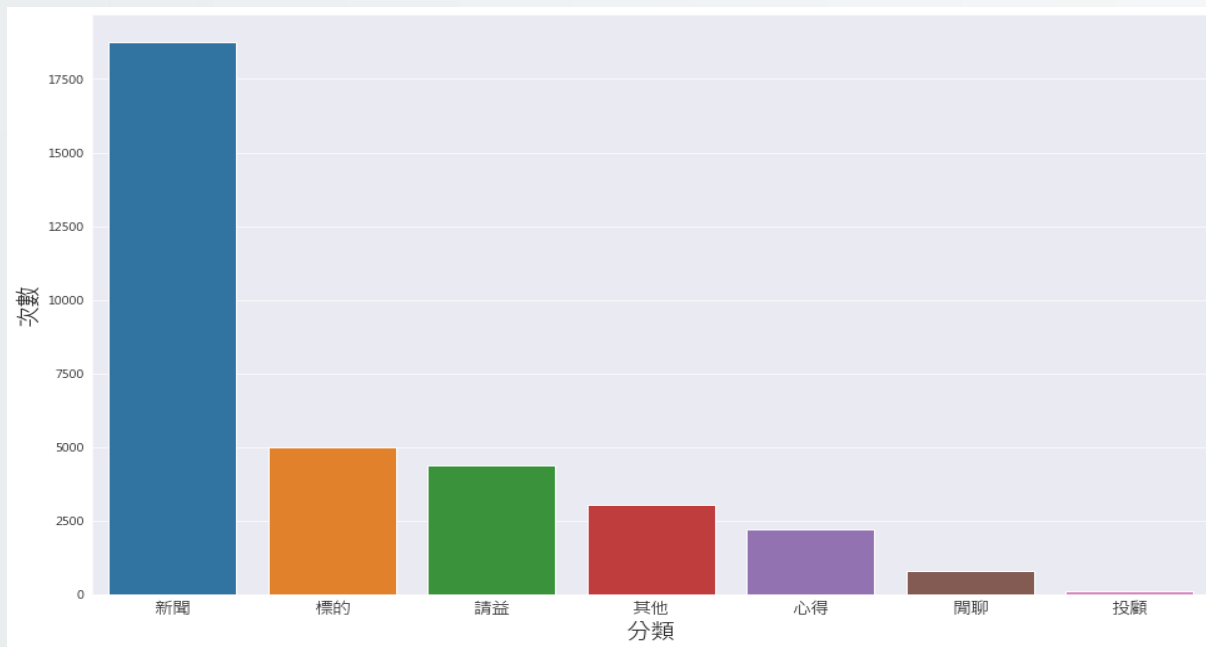
# PTT資料樣態

資料區間：2018/02/01 – 2020/05/01

資料筆數：34354

資料來源：PTT Stock版

- 資料特色：各個文章有類別標籤(如右圖)
- 採用文章正文



# 資料探勘

## 使用贖回率指標做資料探勘

底色為贖回率增加的時間

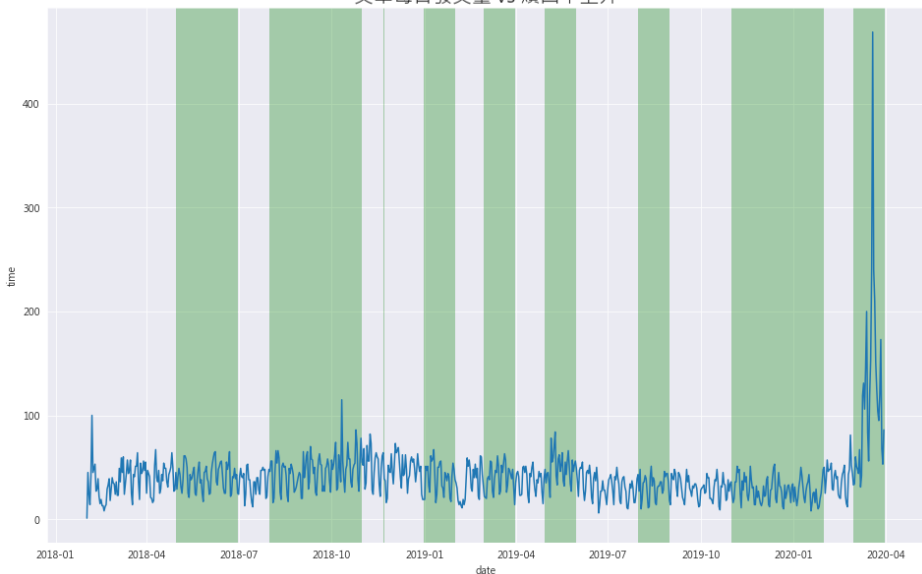
```
signal  
False 14358  
True  18115
```



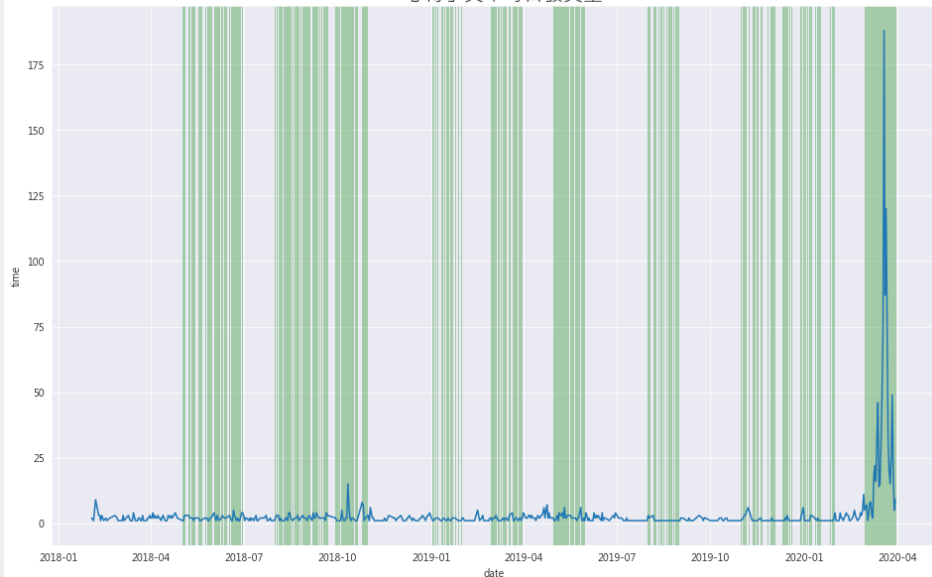
type	signal	
其他	False	1261
	True	1621
心得	False	526
	True	1472
投顧	False	36
	True	50

註：為了獲得更多資訊，這邊設定如果完全沒有文章的日期，就不會畫綠色線，因此心得文某些區間綠色底色呈現稀疏樣態

文章每日發文量 vs 贖回率上升



「心得」文章每日發文量



# 模型結果

## 使用T-test顯著性篩選關鍵字

## 以epson5566為例

由指標三選出，在97篇文章中  
40篇Y=1， 57篇Y=0

	t-ratio	p-value	abs_tratio
關稅	-5.239830	1.644165e-07	5.239830
白宮	-4.408803	1.051665e-05	4.408803
盤	-4.312124	1.634414e-05	4.312124
買入	4.200072	2.694381e-05	4.200072
應該	3.944542	8.056790e-05	3.944542
美股盤	-3.912335	9.209019e-05	3.912335
協議	-3.811656	1.389724e-04	3.811656
美股	-3.804072	1.432920e-04	3.804072
指數	-3.786857	1.535711e-04	3.786857
評論	-3.774198	1.615685e-04	3.774198
新高	-3.657921	2.557412e-04	3.657921
發文	3.638000	2.763223e-04	3.638000
道瓊	-3.629329	2.857563e-04	3.629329
底部	3.615507	3.014217e-04	3.615507
ADR	-3.596216	3.246330e-04	3.596216
進口	-3.574429	3.528566e-04	3.574429
美	-3.528269	4.204013e-04	3.528269
指	-3.436644	5.916685e-04	3.436644
言論	-3.435087	5.950743e-04	3.435087
週	-3.254811	1.138931e-03	3.254811



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
0	Naive Bayes	0.8190	0.7917	0.900	0.8300	0.8579	0.6008
1	Ada Boost Classifier	0.7857	0.7833	0.875	0.8150	0.8334	0.5017
2	Gradient Boosting Classifier	0.7833	0.8521	0.900	0.7983	0.8406	0.4857
3	Logistic Regression	0.7690	0.7792	0.875	0.7867	0.8213	0.4909
4	Extra Trees Classifier	0.7690	0.8250	0.900	0.7717	0.8272	0.4777
5	Ridge Classifier	0.7381	0.0000	0.825	0.7650	0.7860	0.4403
6	SVM - Linear Kernel	0.7000	0.0000	0.725	0.6917	0.6933	0.4004
7	Decision Tree Classifier	0.6952	0.6542	0.825	0.7283	0.7645	0.3096
8	Random Forest Classifier	0.6905	0.7062	0.775	0.7667	0.7546	0.3003
9	K Neighbors Classifier	0.6690	0.6812	0.975	0.6543	0.7810	0.1864
10	Extreme Gradient Boosting	0.6500	0.7417	0.750	0.7050	0.7152	0.2523
11	Quadratic Discriminant Analysis	0.6381	0.6000	0.800	0.6971	0.7340	0.1957
12	Linear Discriminant Analysis	0.5786	0.6146	0.600	0.6983	0.6226	0.1373



```
lr = create_model("lr")
```

	Accuracy	AUC	Recall	Prec.	F1	Kappa
0	0.5714	1.0000	0.2500	1.0000	0.4000	0.2222
1	0.7143	0.8333	0.7500	0.7500	0.7500	0.4167
2	0.8571	0.9167	0.7500	1.0000	0.8571	0.7200
3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0.7143	1.0000	0.5000	1.0000	0.6667	0.4615
5	0.8571	1.0000	1.0000	0.8000	0.8889	0.6957
6	0.5714	0.7500	0.7500	0.6000	0.6667	0.0870
7	0.7143	0.6667	1.0000	0.6667	0.8000	0.3636
8	0.5714	0.5833	0.7500	0.6000	0.6667	0.0870
9	0.5000	0.2222	0.6667	0.5000	0.5714	0.0000
Mean	0.7071	0.7972	0.7417	0.7917	0.7267	0.4054
SD	0.1514	0.2399	0.2250	0.1870	0.1630	0.3070

AUC在0到1之間，越接近1代表模型表現越好

# 穩健性測試

野村證券型基金，不同KOL比較

	Epson5566	Zesonpso	整體心得文
T-test Logit	Mean:0.7972 Std:0.2399	Mean:0.6889 Std:0.1759	Mean:0.6767 Std:0.1924

※Zesonpso這位版友文章數較少  
較難當作預測用的主力KOL



# 穩健性測試

Epson5566，不同贖回率指標比較

	野村全球 股票型基金	野村台灣 股票型基金
T-test Logit	Mean:0.7972 Std:0.2399	Mean:0.7742 Std:0.2213



# 字典法機器學習小結

- T-test尋找關鍵字的方法可以更有效率地泛用在不同資料集
- 更換成另外一個Page Rank選出的KOL，也可以得到不錯的預測結果
- 更改贖回率指標為台灣股票型基金，不會影響原先預測的模型準確度
- 字典法針對特定的KOL預測時效果最好

使用關鍵字進行PTT的**特定KOL**預測的最佳實務是：  
使用**T-test方法**尋找關鍵字並進行機器學習**分類模型**訓練





# BERT



資料子集	「新聞」	「閒聊」	「心得」
Cross-validation Accuracy	0.6153	0.5032	0.8925

- 心得文雖然有資料數目不均等的狀態，但是我們的預測效果仍然勝過壓單邊的狀況。
- 造成此結果的可能是因為原先BERT訓練時使用大量語料，因此如果綜合各個心得文行家的說法，會比單一KOL的可能會有特殊用語習慣時，更有預測的能力。

以「心得」做BERT模型的訓練資料集有最佳效果

# BERT小結

- 使用BERT來進行KOL講話的預測效果不太好，原因可能是因為資料筆數少，或是選到的KOL不夠有解釋力，以及個人的習慣用字可能也會影響到訓練的結果。
- 反觀如果使用以「心得」集結各家說法，得出來的預測力較高，並且可以看到預測效果遠高於「閒聊」。
- 相比於每天高頻的日常文章，心得文的發文頻率較低，和每月公布一次的贖回率的對應相性可能比較好

**使用BERT進行PTT預測的最佳實務是：**  
**針對同類型文章跨作者來進行訓練及預測**





# 強基金研究結果



# 強基金論壇資料樣態

資料區間：2011/01/01 – 2020/05/09

資料筆數：54985

資料來源：強基金論壇

強基金論壇的特點是，留言回復也很有資訊含量，且專注於基金的討論，年齡層相較PTT更成熟一些

- 論壇資料：主要還是以強基金論壇作為文字資料以及模型 fitting、training 的來源
- 該論壇運行時間約從 2011 年開始
- 文章數量約為 8000 篇
- 3000 位不重複使用者的發言紀錄
- 不計文章正文，約有 55000 篇回文

Re: 摩根士丹利印度基金-27 %，該續扣養大？還是該痛下決心贖回？

jett



天使人

刪除

註冊時間:

2016-11-18, 07:43

文章: 400

發表於：2020-06-16, 08:08

定期定額很久了吧？我猜低檔應該沒加扣，報酬率才會如此，不是基金的問題，是你操作習慣的問題

一堆資深強友正在討論加碼印度基金，你選擇在此時退出，反而可能賣在低點

股災買了東協、泰國、印尼、印度、馬來西亞、巴西、南韓基金，績效不錯

<https://forum.fundhot.com/viewtopic.php?t=8230>

當然如果你困擾到睡不著，也沒太多錢可加碼，就贖回吧

重來一次，請看FBI逢低加扣，效果才會好

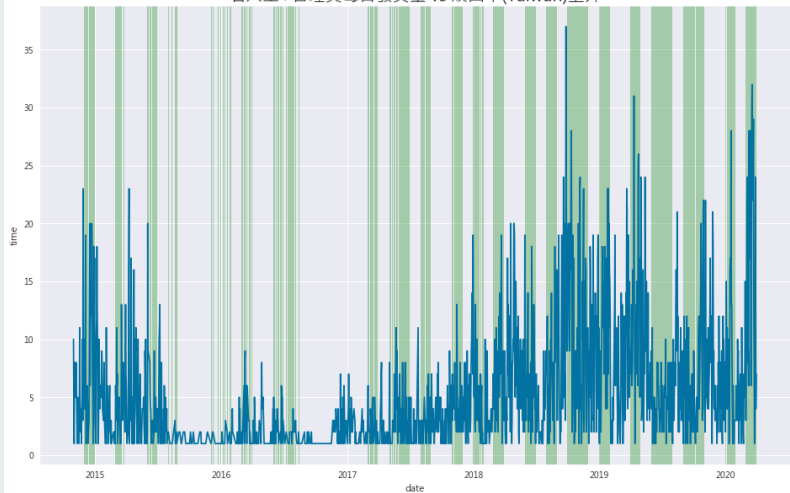
「修正是被圍困的城堡，城外的人想衝進去，城裡的人想逃出來」

# 使用資料

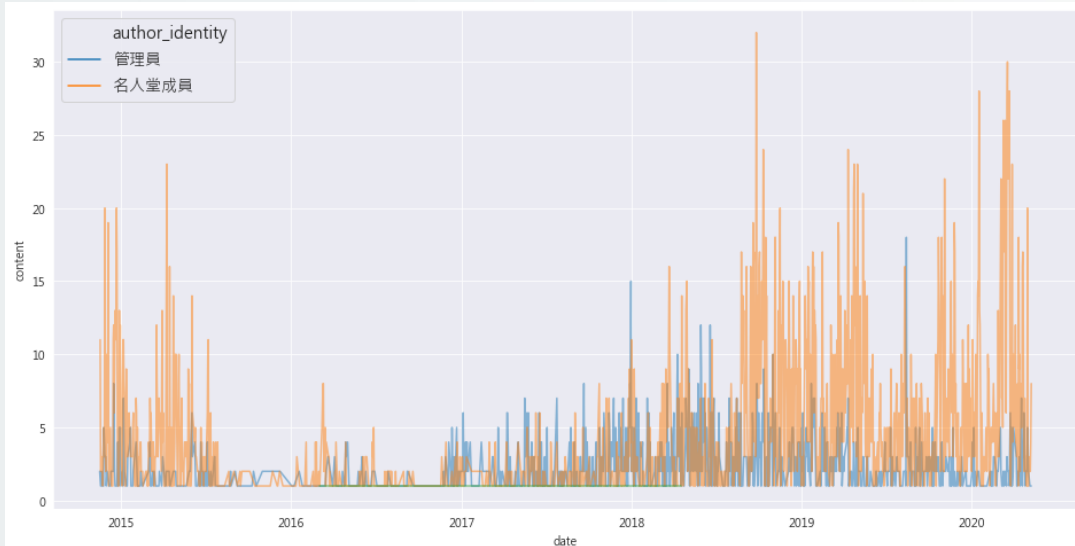
我們使用「名人堂成員和管理員」的文字資料，預測「野村的台灣股票型基金」贖回率上升現象

```
signal  
False 4270  
True 4609
```

名人堂+管理員每日發文量 vs 贖回率(Taiwan)上升



管理員與名人堂成員發文數比較



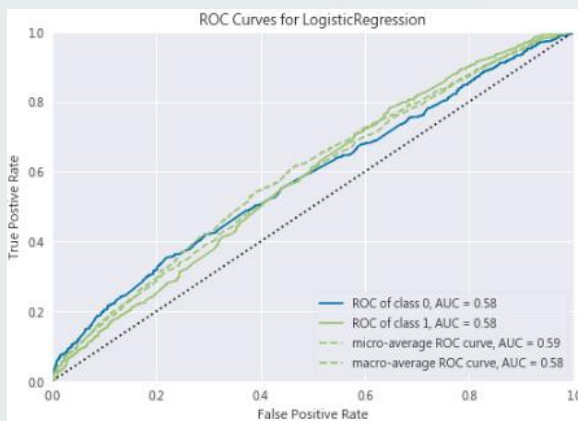
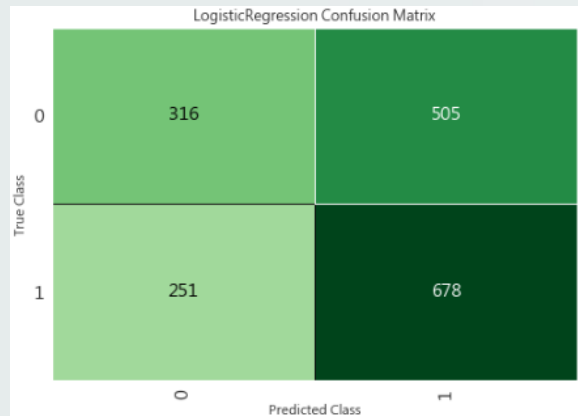
可以看到在贖回率上升與否下，發文文章數沒有特殊差異

# 訓練模型

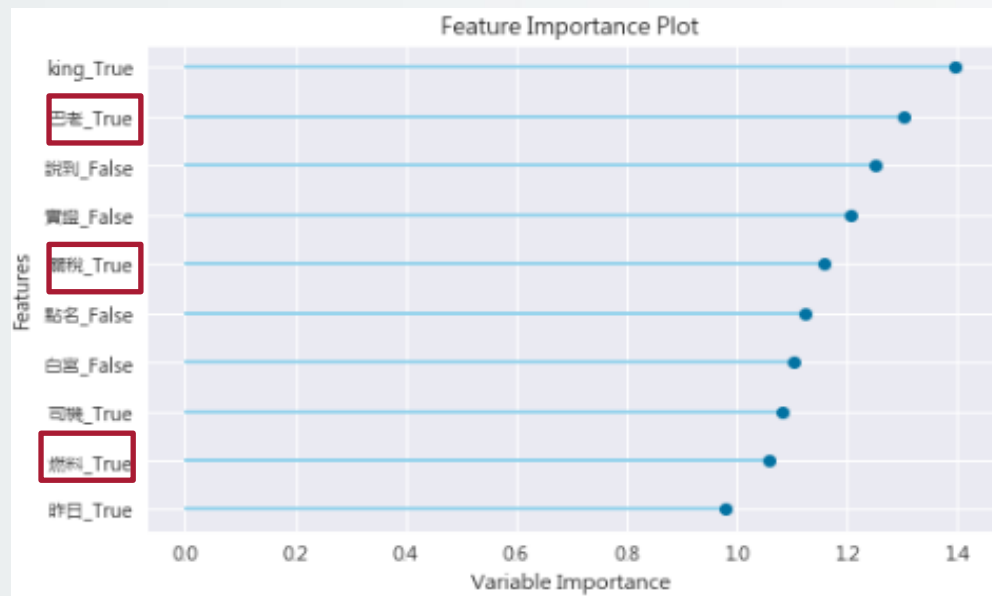
- 可以看到AUC大部分表現都僅落在0.5~0.6之間。
- 依照Accuracy排序下，最佳的模型來自脊回歸以及Logit等傳統計量方法模型。

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
0	Ridge Classifier	0.5758	0.0000	0.7388	0.5786	0.6488	0.1326
1	Logistic Regression	0.5753	0.5993	0.7439	0.5776	0.6501	0.1309
2	Linear Discriminant Analysis	0.5751	0.6004	0.7388	0.5780	0.6484	0.1310
3	Ada Boost Classifier	0.5707	0.5998	0.7554	0.5724	0.6512	0.1196
4	Gradient Boosting Classifier	0.5702	0.5915	0.7997	0.5674	0.6637	0.1133
5	Extreme Gradient Boosting	0.5643	0.5856	0.8011	0.5629	0.6611	0.1003
6	Extra Trees Classifier	0.5530	0.5687	0.5880	0.5769	0.5819	0.1017
7	Random Forest Classifier	0.5449	0.5645	0.6184	0.5654	0.5905	0.0806
8	Decision Tree Classifier	0.5371	0.5451	0.5469	0.5655	0.5557	0.0729
9	SVM - Linear Kernel	0.5351	0.0000	0.6418	0.5987	0.5047	0.0588
10	K Neighbors Classifier	0.5285	0.5409	0.6331	0.5490	0.5875	0.0437
11	Naive Bayes	0.5036	0.5832	0.0886	0.7852	0.1589	0.0581
12	Quadratic Discriminant Analysis	0.4898	0.5875	0.0747	0.8016	0.1223	0.0321

# 機器學習預測



## 對Logit模型做10-fold



- 雖然模型本身預測效果不好，但是feature importance plot可以給我們一些有趣的insight，例如當論壇討論「巴老」(巴菲特)、「關稅」、「燃料」等**關鍵字**時，很有可能會影響**贖回率**上升

# BERT

- 每經過一個Epoch，就相當於把資料集完整丟入模型訓練一輪
- 可以看到經過5個Epoch之後，交叉驗證的Accuracy結果仍然維持在0.56上下，並且很難再提升
- BERT的預測效果和字典法的結果差不多

Epoch	Cross-vaildation Accuracy
1	0.4657
2	0.4690
3	0.5676
4	0.5619
5	0.5518



# 機器學習分析小結

- 即使是「名人堂成員」或是站長等級的用戶，有時也會是純粹閒聊的內容。而強基金論壇不若PTT有明確的文章**類型標籤**可供識別，因此造成機器學習模型的預測效果較差。
- 受限資料的關係，不管使用機器學習或是BERT，得到的結果都不算非常理想
- 在Feature Importance Plot中，可以給我們一些有趣的洞見，例如當論壇討論特定**關鍵字**時，較有可能會對**贖回率上升**產生影響

透過其他研究方法來獲得更多改進方向



# 計量方法研究



# 計量方法研究目的

- 強基金論壇是我們更重視的資料來源，因此我們除了沿用機器學習方法之外，也加入了另一種方法，希望能獲得更多見解
- 目的
  1. 連結詞彙與市場買賣超的關係，並量化出每個詞彙對於買賣超情形的影響，未來能透過量化結果了解文章屬於買、賣超的情緒傾向
  2. 猜測各論壇討論的調性略有差異（風格可能與年紀有關？），瞭解強基金在輿論資訊中所扮演的腳色，也可以將此做為基礎區分在預測未來市場時各論壇所隱含的不同資訊
- 資料
  - 以「0050 每日買賣超」作為研究的主要變數。直覺上買賣超情形是能夠連結到市場氛圍，也希望藉由比較精細的日資料可以獲得比較精準並細緻的結果



# 計量方法研究流程

以 0050 的「一般戶買賣超情形」做為基金贖回率的替代指標，並定義每日為買超日或賣超日

初步找出「訓練資料集」中，高頻在買超日、賣超日出現的詞彙  
( 並以 proportion t-test 、 OLS test 分別進行統計顯著性的驗證 )

根據各詞彙影響買賣超的程度，綜合考量各詞彙 TF-IDF 之權重，為每個詞彙定義出【買超分數】以及【賣超分數】以評估各個詞彙傾向買超行為以及賣超行為的情形

將「訓練集」計算出各詞彙的買超、賣超分數套用在「驗證集」上，並以之計算出各文章的【買超分數】以及【賣超分數】來代表該文章的情形分數，並進行後續的因果關聯驗證 ( OLS 、 Quantile Regression )

- 訓練集：2019/12/31之前
- 測試集：2020/01/01之後

※避免模型學習到「事件」，而不是長期出現的「情緒詞」做為買賣超情形的判斷

如：

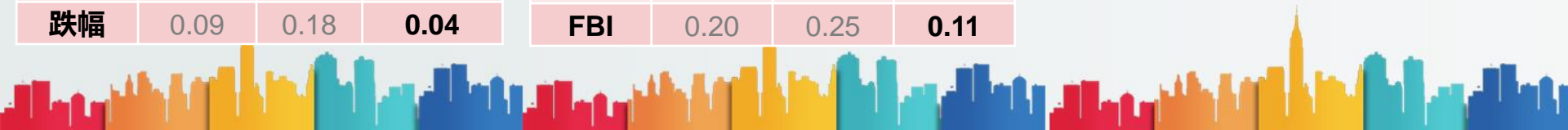
隨機選取到「肺炎」連結到市場賣超，  
未來「肺炎」不一定再次主導市場，  
但會讓我們目前模型的結果過度樂觀

買超日常出現詞彙	買超分數	TF-IDF	結果
現金	0.73	0.15	<b>0.28</b>
撲滿	0.24	0.32	<b>0.13</b>
沒錢	0.28	0.21	<b>0.13</b>
天然資源	0.23	0.28	<b>0.12</b>
組合	0.23	0.18	<b>0.10</b>
心情	0.18	0.28	<b>0.10</b>
跌破	0.12	0.59	<b>0.09</b>
加油	0.11	0.53	<b>0.08</b>
小弟	0.18	0.16	<b>0.07</b>
大跌	0.17	0.14	<b>0.06</b>
連動	0.14	0.18	<b>0.06</b>
明年	0.12	0.19	<b>0.05</b>
台股	0.09	0.22	<b>0.04</b>
事件	0.10	0.19	<b>0.04</b>
跌幅	0.09	0.18	<b>0.04</b>

賣超日常出現詞彙	賣出分數	TF-IDF	結果
紅利	0.62	0.94	<b>0.74</b>
恭喜	0.60	0.93	<b>0.72</b>
NN	0.40	1.00	<b>0.63</b>
生技	0.38	0.89	<b>0.55</b>
贖回	0.17	0.66	<b>0.28</b>
大漲	0.25	0.52	<b>0.26</b>
謝謝	0.22	0.46	<b>0.22</b>
新年快樂	0.64	0.27	<b>0.22</b>
利安	0.72	0.24	<b>0.21</b>
表單	1.00	0.21	<b>0.21</b>
去年	0.13	0.57	<b>0.20</b>
強網	0.41	0.30	<b>0.19</b>
一支	0.26	0.28	<b>0.14</b>
天后	0.87	0.15	<b>0.14</b>
FBI	0.20	0.25	<b>0.11</b>

在此先列出強基金論壇在買超日以及賣超日常出現的詞彙

- 在買超日常常出現詞彙有
  1. 大跌、跌破 ( 逢低點進場 )
  2. 現金、沒錢 ( 討論各自手頭上資產的流動性 ) .....等
- 而賣超日常常出現詞彙有
  1. 恭喜、謝謝、贖回、大漲、去年、一支 ( 分享基金成效、描述購買基金的脈絡 )
  2. 另外也有天后、FBI 等該網站的專門術語 ( FBI 是該論壇專門用以評估相對低點的指數 )



Y = 買賣超情形	Est.	S.E.	t-value	Pr(> t )
(Intercept)	2.35	0.04	57.20	<0.00
賣超分數	14.94	8.31	1.80	② 0.07
買超分數	-12.70	3.03	-4.20	① 0.0002

使用 0050 的買賣超數量，針對強基金常分別在買超日、賣超日出現的詞彙進行複回歸分析，結果發現

- **【買超分數】**有較低的 p-value 代表著較好的顯著情形，也說明「買超分數」其實較能夠較好地描述市場上「買賣超情形」( ① )
- **【賣超分數】**的表現則較差，統計顯著性僅在常使用的 0.05 邊緣 ( ② )，顯示這個變數對於「買賣超情形」的解釋能力其實差強人意
- 推測
  - 進行質性的觀察後，我們認為**強基金上面的文章大多以詢問投資標的為主**，而抒發情緒類文章相對較少。相較於購入基金大家會傾向借重群眾的智慧找到最適合的標的，基金贖回則是比較屬於個人停損點的評估。因此我們認為比較難就單一論壇上的討論結果進行未來市場情形的預測。



賣超情形時	Est.	SE	Pr(> t )	AIC
(截距)	-3.57	0.041	<0.00	<b>3226.75</b>
賣超分數	32.74	32.91	0.32	
買超分數	0.27	0.35	0.44	

買超情形時	Est.	SE	Pr(> t )	AIC
(截距)	4.38	0.008	<0.00	<b>2087.93</b>
賣超分數	6.30	2.77	0.02	
買超分數	-3.31	0.55	<0.00	

- 為了驗證買賣超情形不同時，會影響到我們模型預測的能力，在這邊進行分量回歸進行分析，他可以幫助我們「買超情形」以及「賣超情形」兩種不同的情形模型分別的解釋效果
  - 在此分別取前 10% 買超情形以及前 10% 賣超情形進行觀察
- 結果顯示這個模型在買超情形底下這個模型的解釋能力會較好，在市場氣氛傾向買超的時候，每個係數都變得更顯著了 ( ③ )，AIC 也下降了 ( AIC 下降代表模型預測能力提升 )
  - 如果最近市場氛圍使散戶投資人的交易情形為買超時，這個模型在強基金論壇上找出買賣超詞彙，可以較為準確地掌握市場脈動，並利未來進行相對的策略舉措



# 強基金論壇計量分析小結

- 可能是基於用戶年齡層的關係，強基金論壇和PTT相比，文章較少情緒字眼
- 「買超分數」其實較能夠較好地描述市場情形，而「賣超分數」的表現則較差
- 如果最近市場氛圍使散戶投資人的交易情形為買超，可以較為準確地掌握市場脈動
- 每天的文章數量沒有大到一定要靠機器學習才能獲得每天的洞見，只要市場情形好的時候，質性地研究強基金論壇的想法，然後對他們自己內部的投資人去做行銷(ex.現在版上哪種基金熱度高他們就也在內部推對應的基金之類的) 應該就會有不錯的效果了

強基金論壇相比於預測贖回，更適合在市場樂觀情形下預測**申購率**







# 結論與應用程式



# 結論 - PTT

- 良好的訓練資料子集

1. 指標三(排除新聞與公告)選出的Epson5566
2. 指標四(PageRank)選出的Zesonpso
3. 利用BERT來訓練整體的「心得」文章

- 關鍵字提取方法

- ✓ 方法四的T-test方法較不需要外生給定參數，為最佳實務方法

- 訓練模型

- ✓ 運用字典法時**Logit**、**Naïve Bayes**、**Gradient Boosting**幾乎都能提供良好預測效果
- ✓ 使用NLP方法時，資料筆數必須要夠多才能產生有意義的結果



# 結論 - 強基金

## ● 機器學習方法部分

- 受限於論壇沒有文章類型標籤，使得找尋KOL的指標不盡理想
- 在篩選出的資料上，應用字典法和NLP的預測表現都不如在PTT時理想
- 透過Feature Importance做關鍵字的探索容易得到有洞見的詞彙

## ● 計量方法部分

- 相比於PTT，較少情緒用字，更專注投資標的上的討論
- 「買超分數」比起「賣超分數」能夠較好地描述市場情形
- 在市場樂觀時，預測申購率和贖回率效果都更好
- 由於版上討論風氣，強基金論壇相比於預測贖回，更適合預測申購率



# 總結

- PTT部分

- 如果要針對特定KOL分析，使用T-Test尋找關鍵字效果較好
- 在我們採用的方法中，使用BERT對「心得」文章訓練效果最好

- 強基金部分

- 計量方法部分

- 「買超分數」比「賣超分數」更能夠描述市場情形
    - 市場樂觀時，預測買賣超的效果都更良好

- 機器學習方法部分

- 用字典法和NLP表現都不如在PTT時理想，需再對資料深入處理

# 可改進之處

## ● 整體部分

- 完善資料管道，實現日頻的結果更新，並且集成各個模型，產生贖回率的輿情指標
- 透過野村內部提供的週頻贖回率資料，應能產生更高頻更細緻的模型預測結果

## ● PTT部分

- 建立資料庫，長期累積PTT的語料，避免太久遠的文章被系統回收刪除
- 建立回覆型文本的分類標準，擴大可用資料範圍

## ● 強基金部分

- 站內文章沒有特定標籤，因此可能參雜閒聊類型的文章降低預測力，如果能先透過篩選閒聊模型過濾，再進行我們的贖回率預測，應能幫助我們更加精準的提取該網站的洞見
- 透過人工判斷挑選KOL，彌補論壇上「沒有文章標籤輔助我們篩選KOL」的限制



# 應用程式

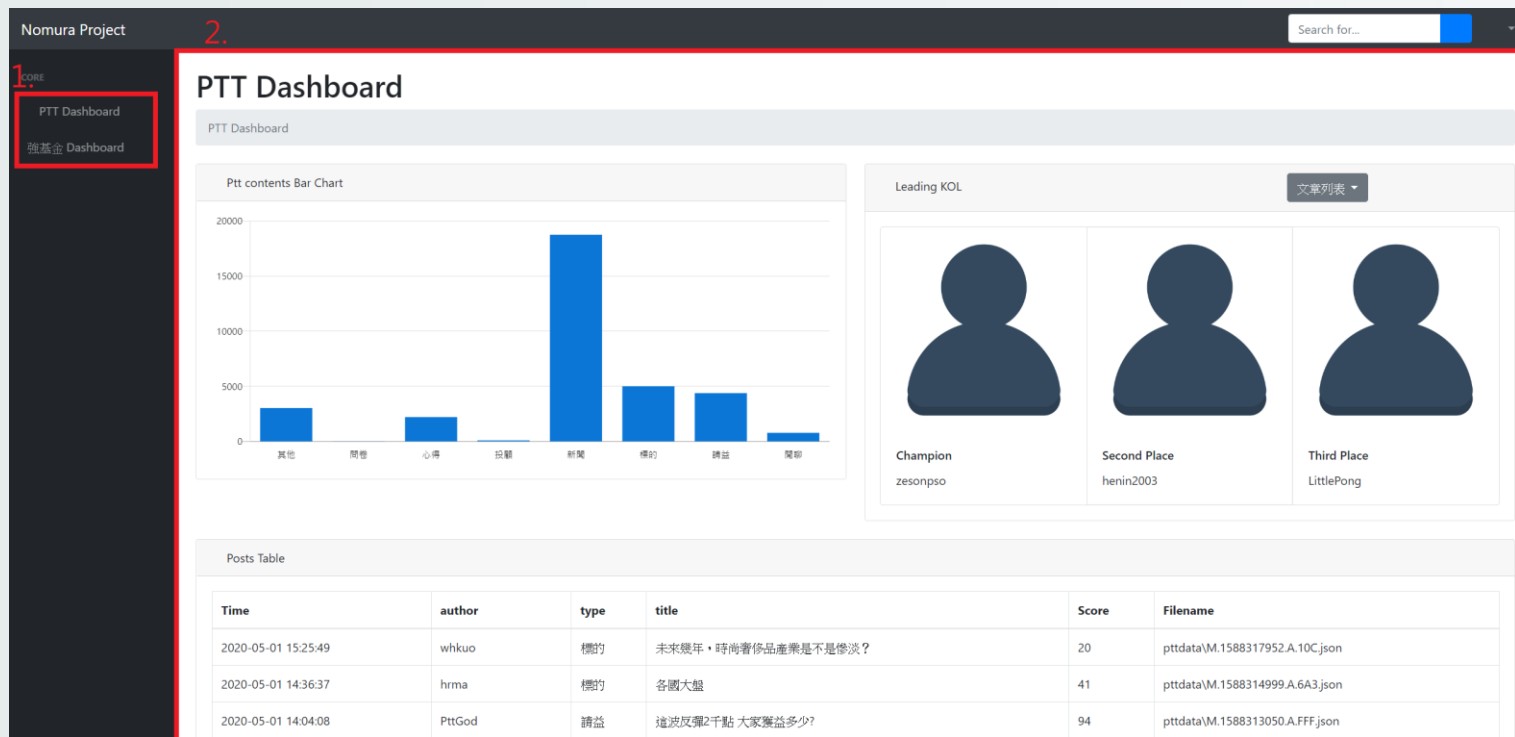
<http://3.134.79.174>

- Rwd的資料統計Dashboard
  - 前後端分離
  - Rwd
  - Python and javascript

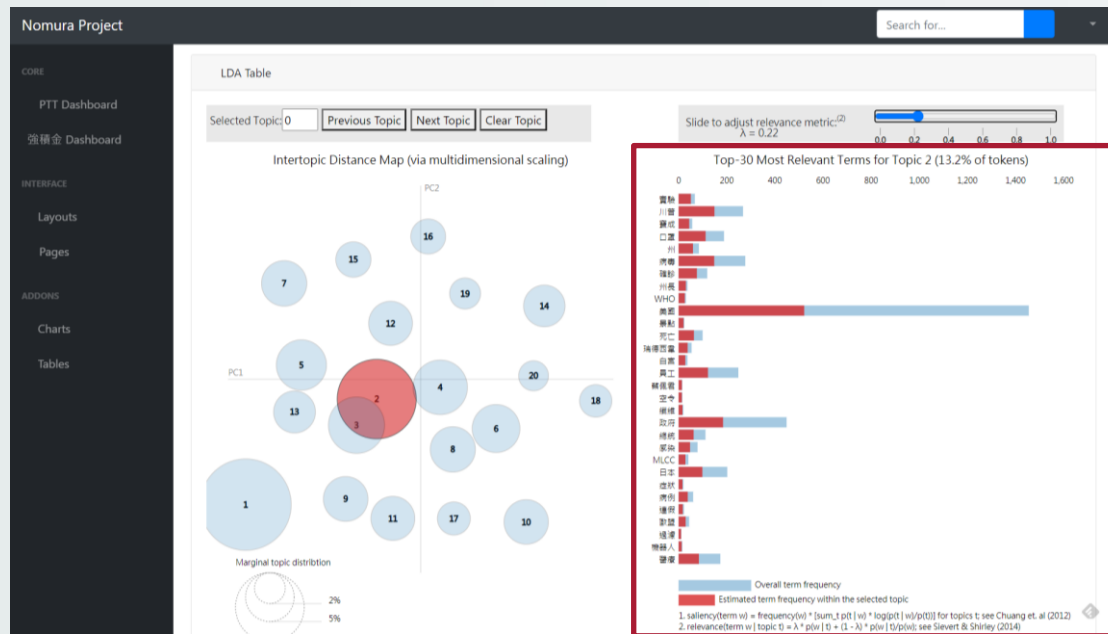


# 介面呈現

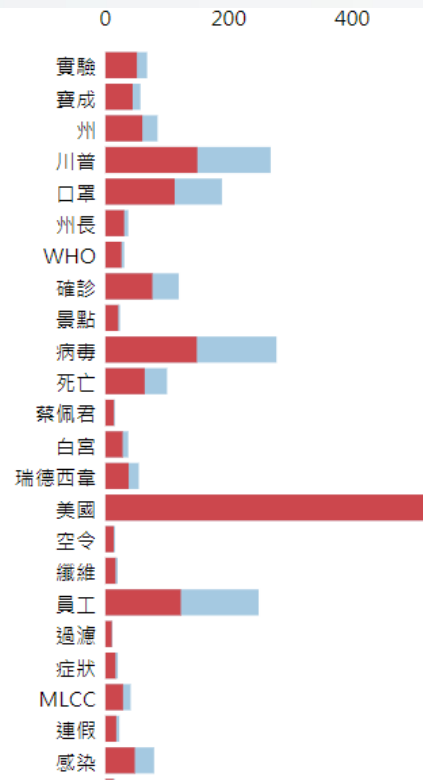
- 1.選擇資料來源
- 2.資料視覺化Dashboard



# LDA



可以看到PTT的主題二應該是在討論Covid-19疫情相關的議題





# 感謝您的聆聽



**GitHub**

qoo121314/Fintech\_Nomura\_Group2

Github Repo及完整版PPT

