野村投信題目一 第一組

運用熱門主題及字組提升行銷轉換成率之探索性研究

指導教授：蔡芸琤 教授

野村投信Mentor：張孝璿(Kian), 鄒志斌(Eddie)

組員：台大財金碩一 陳昱嘉, 台大國發所碩二 黃緯易, 台大財金四 方婕薰, 東吳巨資二 呂承翰

NOMURA

專案動機

傳統行銷流程：TOP-DOWN

先從主推什麼產品，再去分析TA，最後才選擇行銷管道及最後規劃促銷內容。

逆向行銷流程：BUTTOM-UP

先從挖掘熱門投資話題及關鍵字詞，再反向來發展行銷策略。

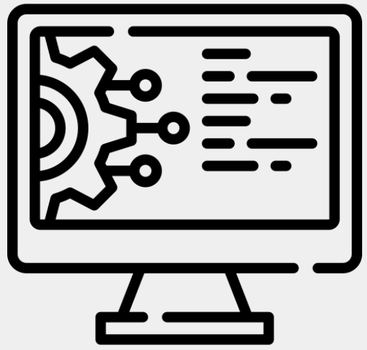
Outline

1. 程式功能簡介
2. 專案所應用之技術
3. 專案開發流程
4. 網頁Demo
5. 未來可能發展建議

程式功能簡介

Input

- 論壇文章與留言
- 新聞
- 投信官網主頁資訊



Output

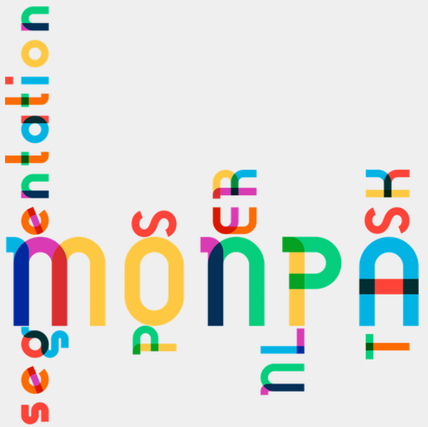
- 熱門關鍵字
- 關鍵字組合分析

→ 解決投信業者利用傳統人工方式瀏覽資料找尋熱門議題，耗時費力之痛點

專案所應用之技術

技術	實現套件
資料爬蟲	Beautiful Soup
斷詞	monpa (在CKIP基礎上優化後的套件)
資料儲存	pandas
繪圖	Plotly
前端網頁設計	flask

*CKIP: 中研院中文斷詞系統



專案流程

資料爬蟲
與設定資料區間

- PTT
- Dcard
- 各大新聞網站
- 投信業者官網

資料預處理

- 斷詞
- 資料清洗

找出熱門關鍵字

- 詞頻計算
- TF-IDF出現文件數

關鍵字組合分析

- 所有相鄰字組
- TF-IDF關鍵字
  與TF-IDF關鍵字之詞組
- TF-IDF關鍵字
  與非TF-IDF關鍵字之詞組

建立資料庫
蒐集每日分析結果

網頁架設
與資料更新

資料爬蟲與設定資料區間

資料來源	抓取內容	抓取最新資料 (定義資料抓取日為t)	資料分析區間
ptt 基金版	內文、留言(一則留言視為一文件)	t-1	t-1~t-14(兩周)
Dcard 理財版	內文、留言(一則留言視為一文件)	t-1	t-1~t-14(兩周)
新聞：工商時報、中國時報、鉅亨網基金、UDN基金	內文	t-1	t-1(一天)
投信業者官網	精選基金、Banner	t-1	t-1(一天)

→ **根據資料來源分為四個專案個別分析: ptt, Dcard, 新聞, 投信**

- 根據資料提供群體不同，分開可以做出區隔
- 避免資料量不平衡導致資料被稀釋的問題

以PTT資料為例
(抓取日期：6/22、分析資料區間：6/21~6/07)

資料預處理

若未自行定義，可能被斷成: (景氣, 循環, 股)

```
景氣循環股 100 NER
多重資產 100 NER
非投資級 100 NER
墜落天使 100 NER
公司債 100 NER
階梯到期 100 NER
目標到期 100 NER
```

斷詞：

除了使用monpa套件內建字典外，

另外使用自訂義字典找出**基金相關名詞**

資料清洗：

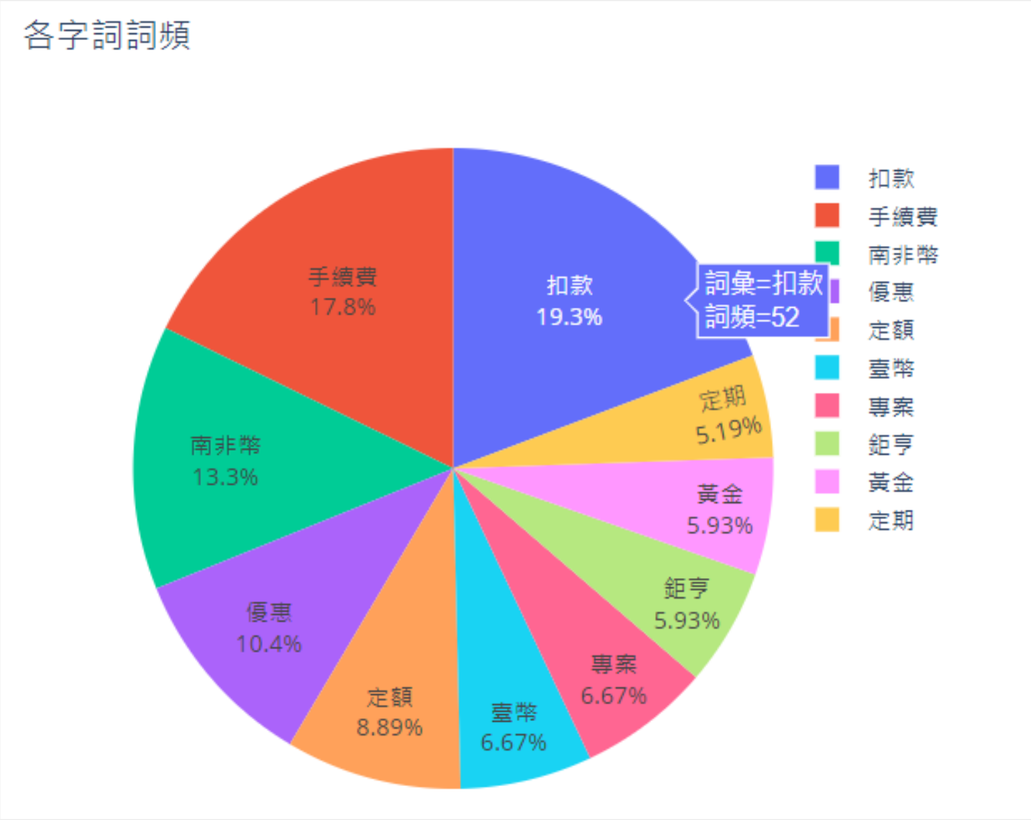
使用**停用詞字典**將無法給出明確insight的詞

(包含無意義符號及不重要的詞性)去除

```
類型
等級
多重
方式
問題
很多
```

找出熱門字詞(1) 詞頻計算

- 方法：計算每個字詞出現的數量
- 意義：從所有詞當中，找到出現次數最高的詞



找出熱門字詞(2)　TF-IDF關鍵字出現文件數

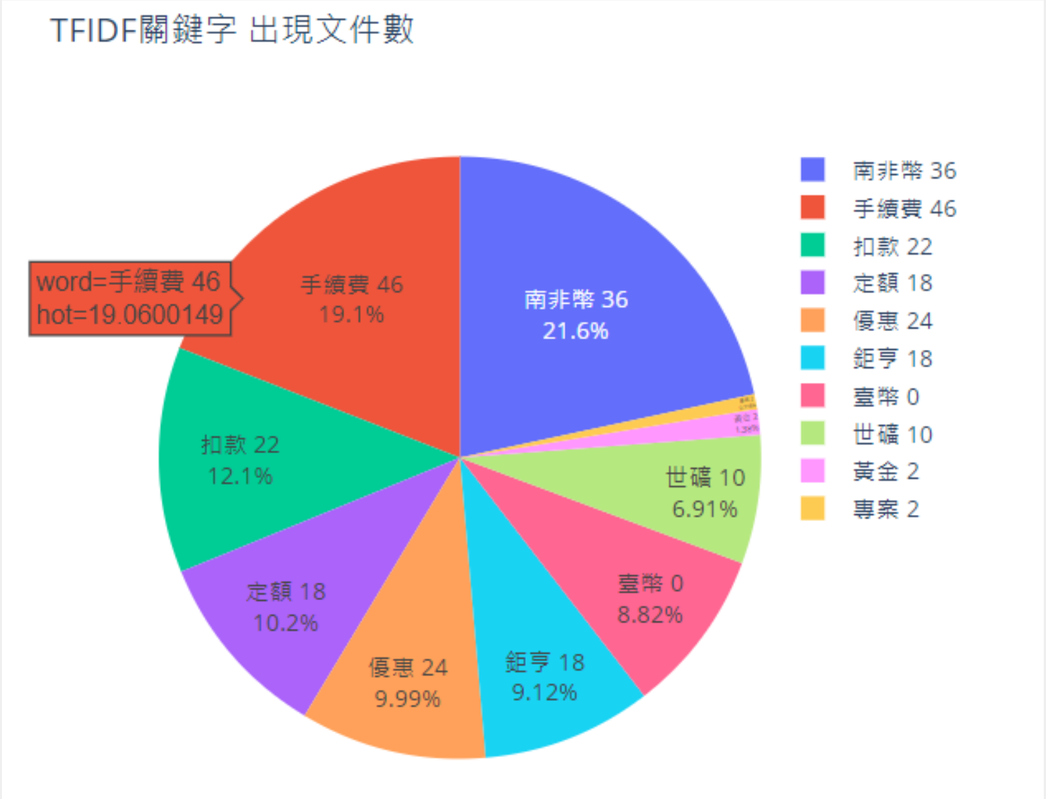
- **方法：**
  Step 1. 透過TF-IDF找出每份文件關鍵字

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

- TF ：第 i 個字在第 j 個文件出現頻率
- IDF：有出現第 i 個字的文件數佔所有文件的比率 (逆向：佔比愈高，IDF值愈低)

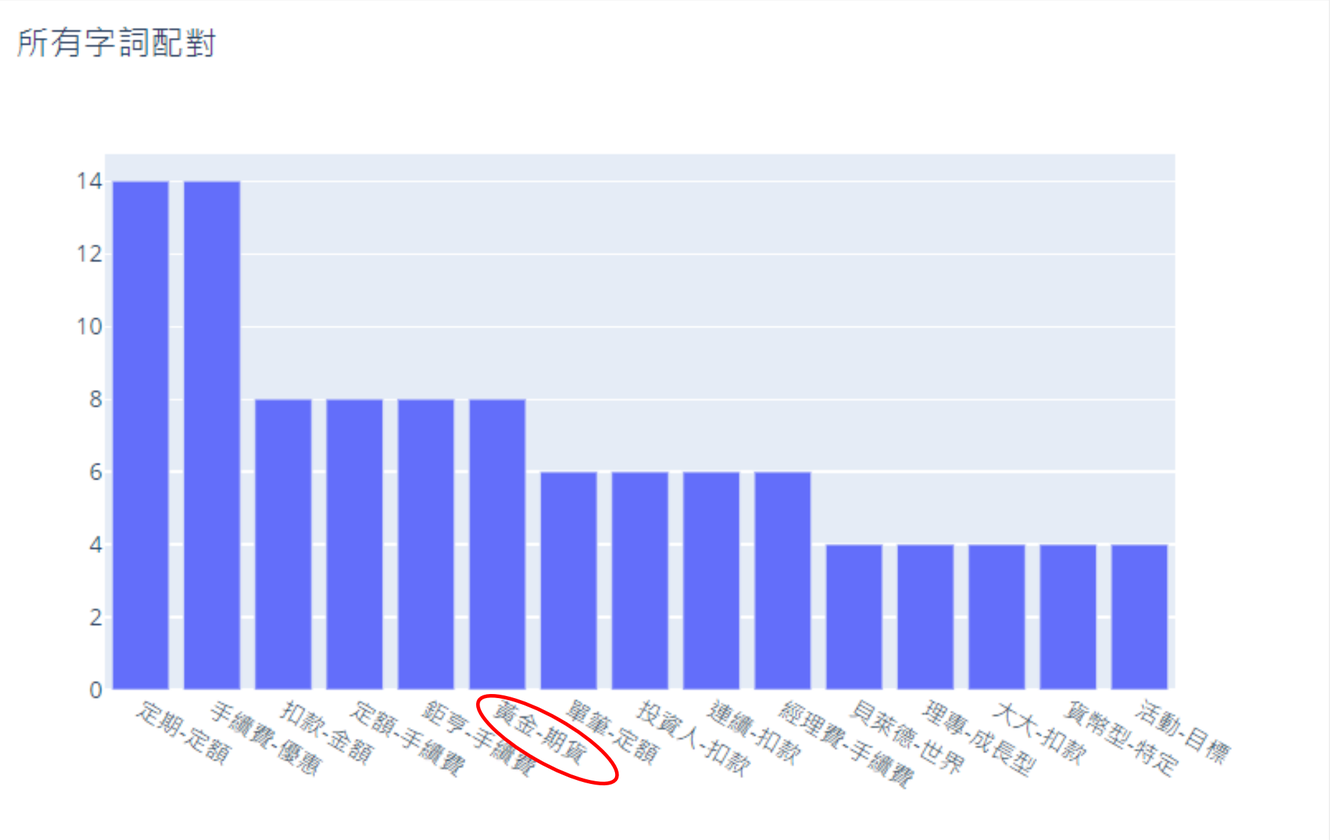
Step 2. 計算所有TF-IDF關鍵字出現的文件數

- **意義：**
  關鍵字在越多份文件被討論，熱門程度越高



關鍵字組合分析(1) 所有相鄰字組

- 方法：計算所有相鄰字組出現的次數
- 意義：可以看到哪些相鄰詞組在所有文件當中出現最多次



關鍵字組合分析(2) TF-IDF熱門關鍵字相互配對

- **方法：TF-IDF關鍵字排名前10熱門的關鍵字(參考P. 11右圖結果)，兩兩一組，計算每個詞組中，這兩個詞在每篇文件的平均距離，以及這兩個詞在所有文章中所出現了幾對(必須要同時出現在同一文件才能成對)**

平均距離計算方法：假設第j種詞組有n_j次配對

$$n_j^{-1}\sum_{i=1}^{n_j} dist_{ij}$$

$dist_{ij}$表示第j種詞組在第i次配對之距離
(距離：配對中字詞間隔字數)

- 平均距離可以代表兩個詞之間的相關性，距離愈近，表示愈有可能被同時提到；
- 出現的對數則代表這個詞組的熱門程度。

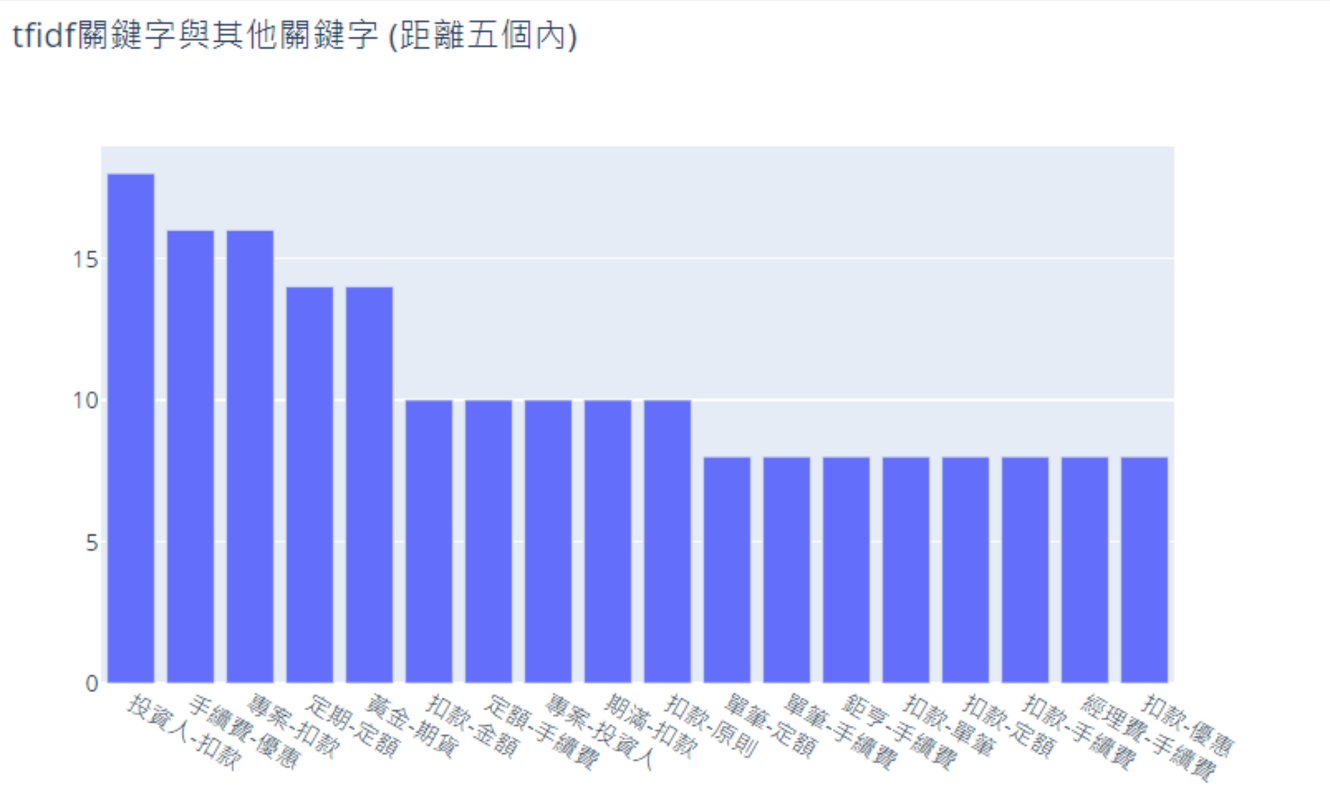
TFIDF配對TFIDF

關鍵字1	關鍵字2	平均距離	對數
鉅亨	手續費	1	8
臺幣	扣款	1	4
扣款	臺幣	1	2
手續費	鉅亨	1	2
定額	鉅亨	1	2
手續	鉅亨	1	2
鉅亨	優惠	2	2
臺幣	定額	2	2
定額	臺幣	5	2
南非幣	臺幣	6	8
手續費	專案	31	36

- **意義：可能可以找到較具意義的熱門字組**

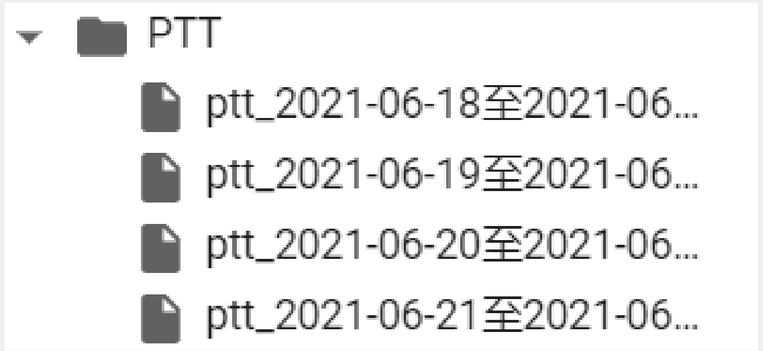
關鍵字組合分析(3) TF-IDF熱門關鍵字與其它字詞配對

- 方法：定義"距離五個詞彙"以內都算是是一組配對，並計算TF-IDF詞彙與其他詞彙配對數
- 意義：可以看出哪些TF-IDF沒有萃取出的詞，可能會經常跟TF-IDF字詞組合。避免有些重要詞組因為沒有被視為TF-IDF關鍵字而被忽略



網頁架設與資料更新

以網頁方式呈現每日分析結果，存取每日匯出之網頁檔，並依照資料來源不同以網頁檔形式存取在不同的資料夾，之後可以從歷史網頁檔觀察每日熱門議題趨勢



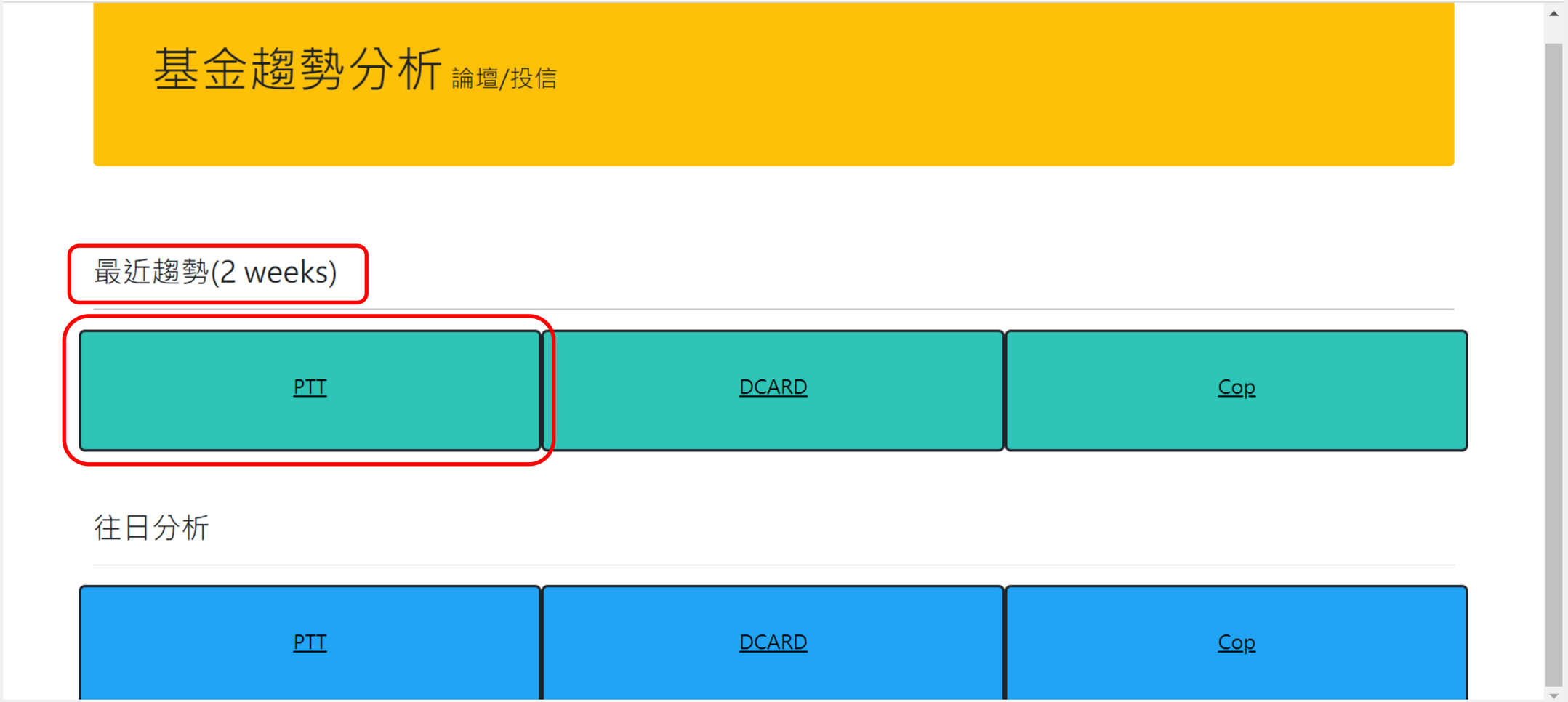
網頁架設與資料更新

將所有資料來源之分析結果的網頁檔整合，依設計之網頁架構，將資料更新至主網頁

```
 * Serving Flask app "__main__" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off
06/21/2021 16:11:16 - INFO - werkzeug -  * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
 * Running on http://f62705492a70.ngrok.io
 * Traffic stats available on http://127.0.0.1:4040
```

程式跑完後點選此連結即可進入

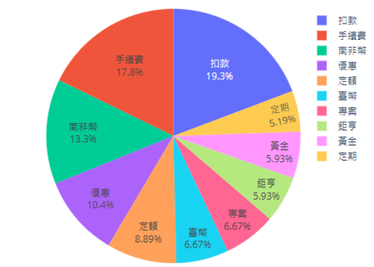
網頁Demo



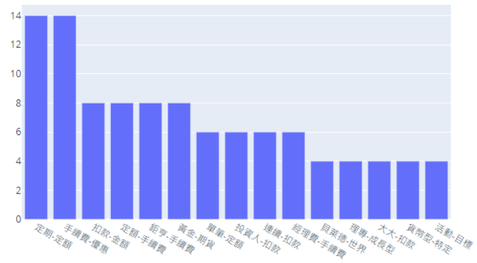
PTT 趨勢分析 2021-06-21至2021-06-07

文章數量	留言數量	不重複的字詞/總共字詞
13	273	500/1742

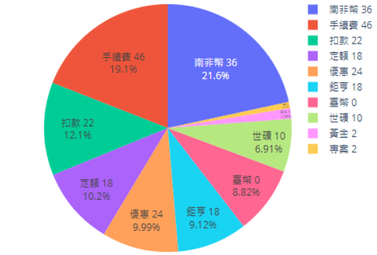
各字詞詞頻



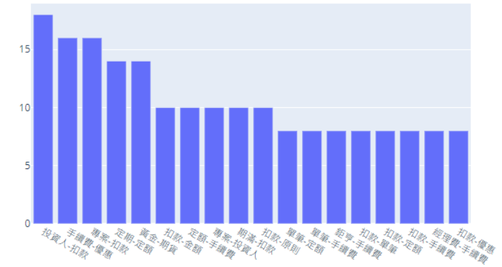
所有字詞配對



TFIDF關鍵字 出現文件數



tfidf關鍵字與其他關鍵字 (距離五個內)

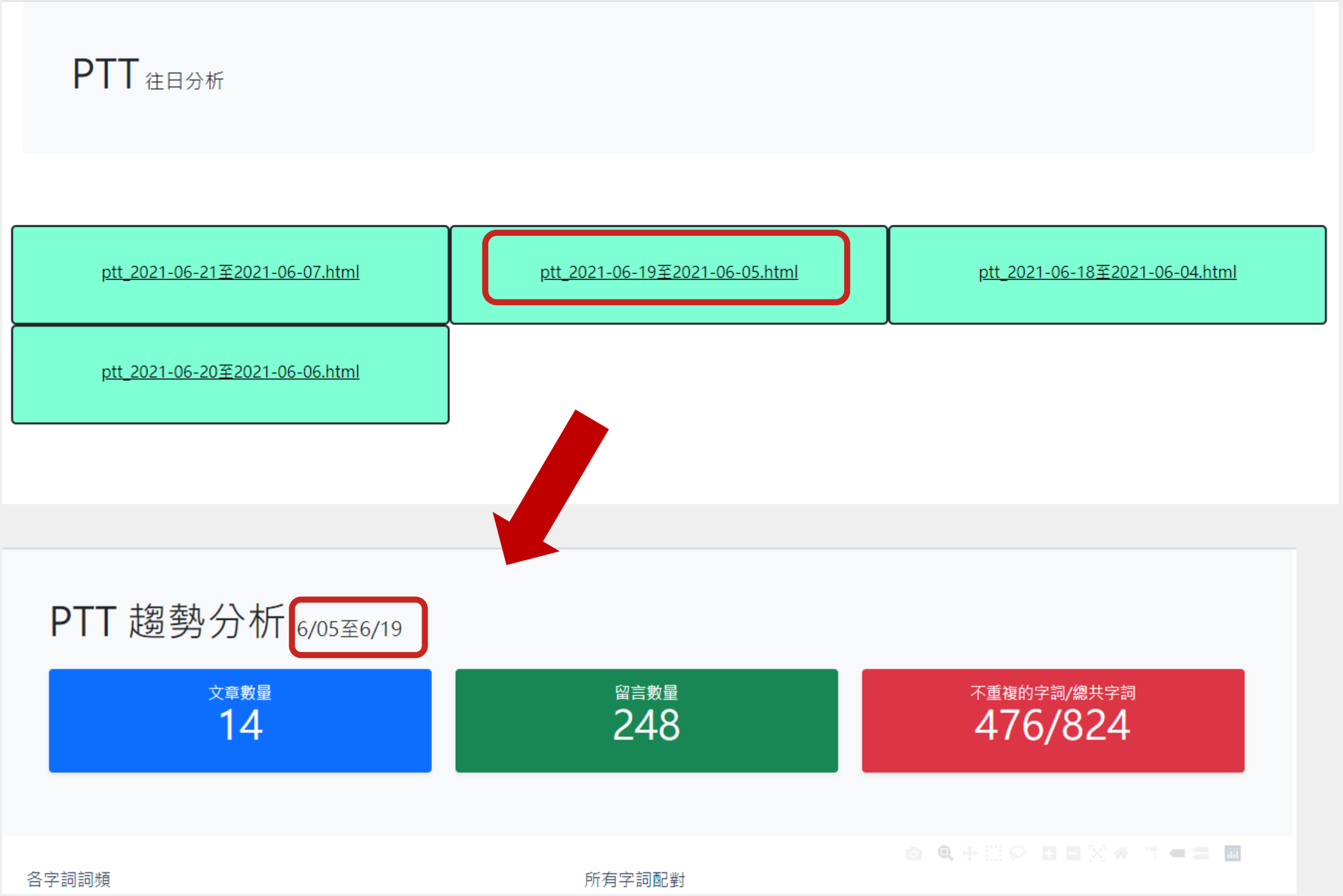


TFIDF配對TFIDF

關鍵字1	關鍵字2	平均距離	對數
矩亨	手續費	1	8
臺幣	扣款	1	4
扣款	臺幣	1	2
手續費	矩亨	1	2
定額	矩亨	1	2
手續	矩亨	1	2
矩亨	優惠	2	2
臺幣	定額	2	2

熱門文章(討論度>=20)

標題/時間/討論度
[新聞]「好享退」將屆期滿！持續扣滿 48 個月， 6/16 25
[問題]被理專推的配息基金會賴回嗎? 6/12 25



未來可能發展建議

1. GUI設計
2. 能自行定義想分析的資料區間
3. 透過多標籤分類方式將文件分類，並計算每一種標籤之數量，藉以掌握熱門議題

台灣

投資標的區域

科技

投資標的產業

一直 有 想 買 野村 基金 之前 有 存 過 006208 一 開始 看 重 野村 優質 基金 最近 發現 野村 優質 持股 成分 沒 含 金融股 科技股 比重 較 高 當走 空頭 時 持股 希望 能 較 分散 點 後來 看到 野村 成長 野村 積極 成長 這 2 支 很 相同 與 富邦 006208 相近 想 知道 富邦 006208 VS 野村 積極 成長 基金 2 支 差異

股票型

基金類型

分工表

台大財金碩一 陳昱嘉	專案發想、構想熱門字組分析方法、PPT製作
台大國發所碩二 黃緯易	資料爬蟲、程式實現字組分析方法、網頁視覺化、自動分類資料庫
台大財金四 方婕薰	專案發想、構想熱門字組分析方法、PPT製作
東吳巨資二 呂承翰	資料爬蟲、影片製作

Thank You