1132_社群媒體分析

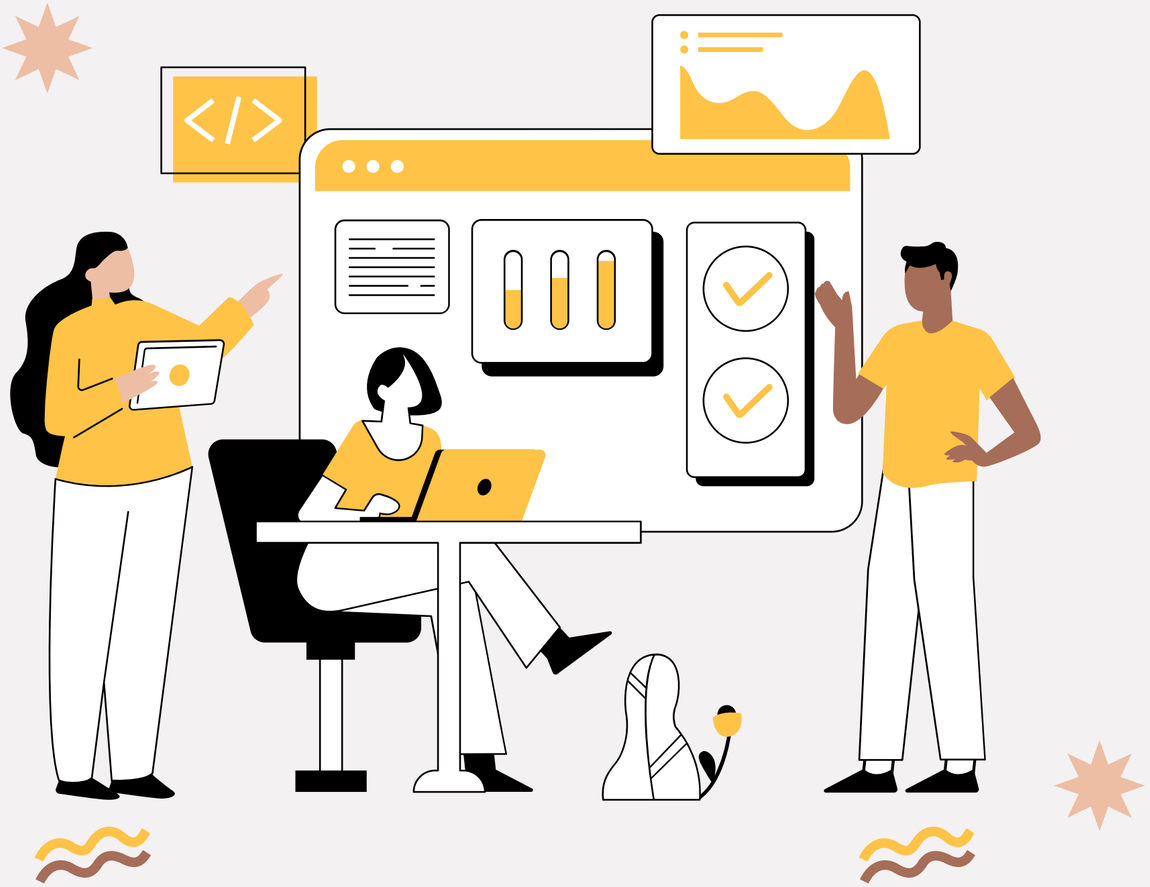
第二次讀書會專案

DCARD
美食、研究所、股票版
主題分析

第四組

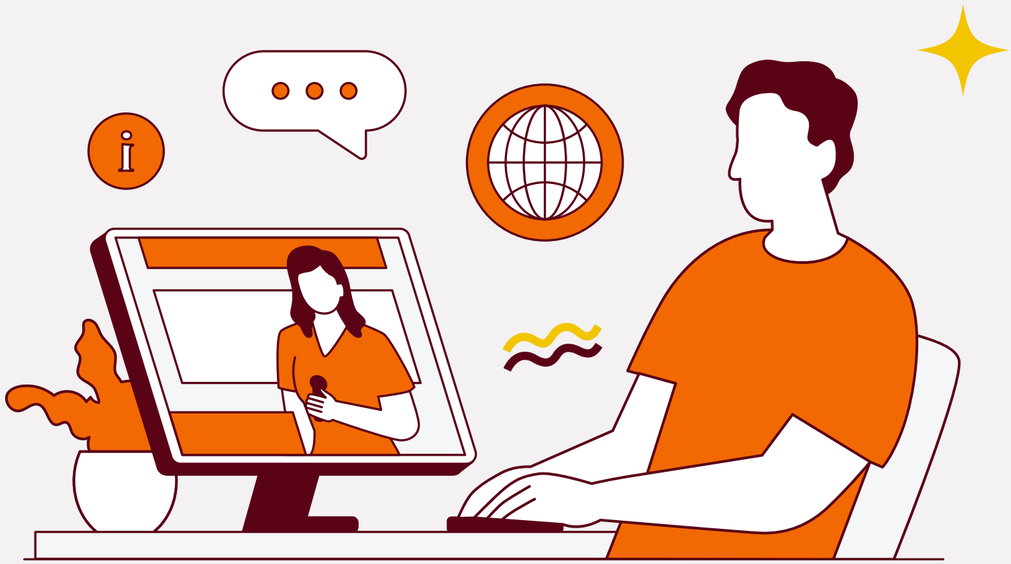
M134020043 張恩綮
M134020002 陳亭瑄
M134020018 吳信輝
B104020027 蕭宇廷
B104020028 林晏甄
M13B020007 陳學蒨

目錄



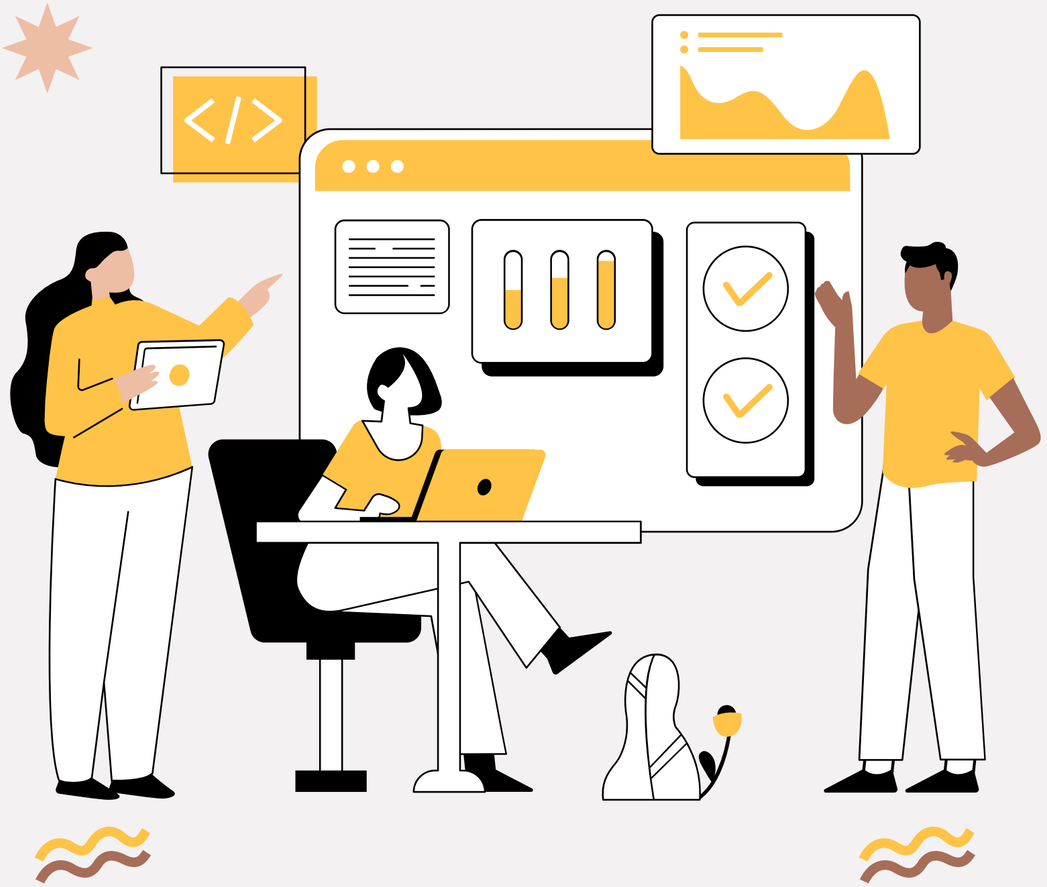
01
專案動機

專案動機



- 專案目標：探討跨平台文章分類模型的適應能力與效果
- 訓練資料：以 Dcard 文章作為模型訓練資料
- 測試資料：應用於
  - 性質相似的社群平台：PTT
  - 性質差異較大的新聞媒體：聯合新聞網

- 重點：
  - 模型在不同平台與語境下的分類表現
  - 文本風格與內容類型差異對模型效果的影響
  - 分析分類準確度與表現落差背後的原因

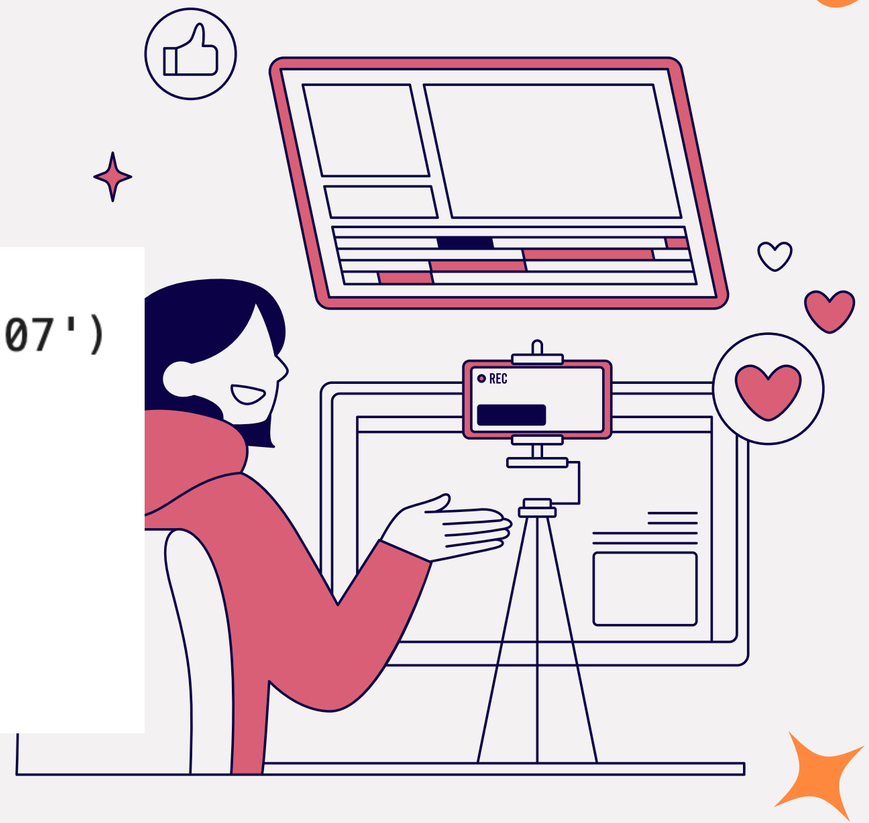


02
資料前處理

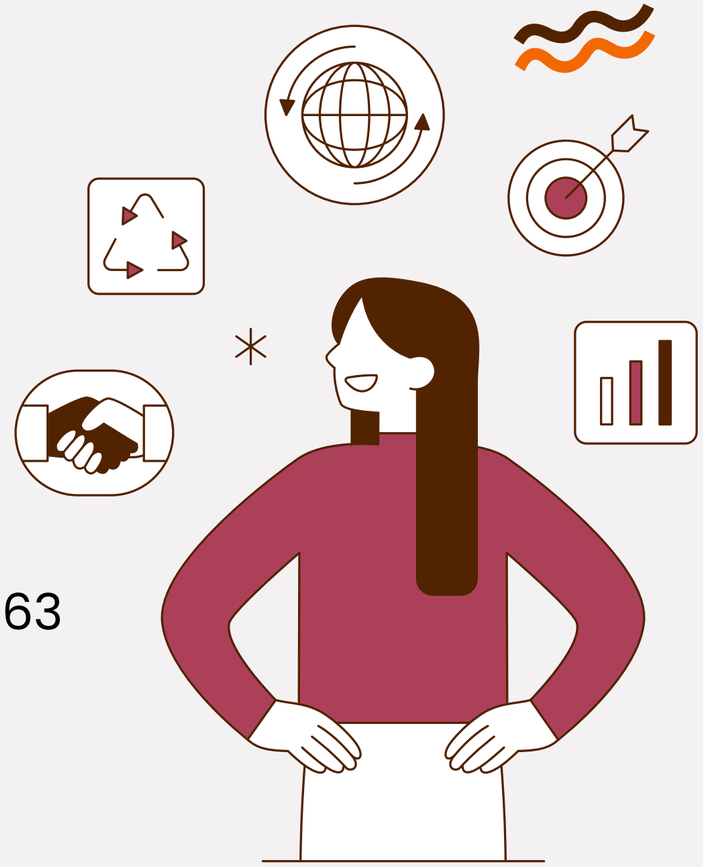
資料抓取

```
number of posts: 6262
date range: ('2025-01-01 01:00:04', '2025-02-22 17:36:07')
category:
artCatagory
stock                2165
graduate_school      2093
food                 2004
Name: count, dtype: int64
```

Donec quis erat et quam iaculis faucibus at sit amet nibh.



文字前處理



01

清理

將欄位中有空值的資料刪除，並對剩下的文本資料移除網址和英文字。

02

斷詞

將'artTitle'和'artContent'欄位合併成'content'後進行斷詞，並將結果存在'words'欄位。

03

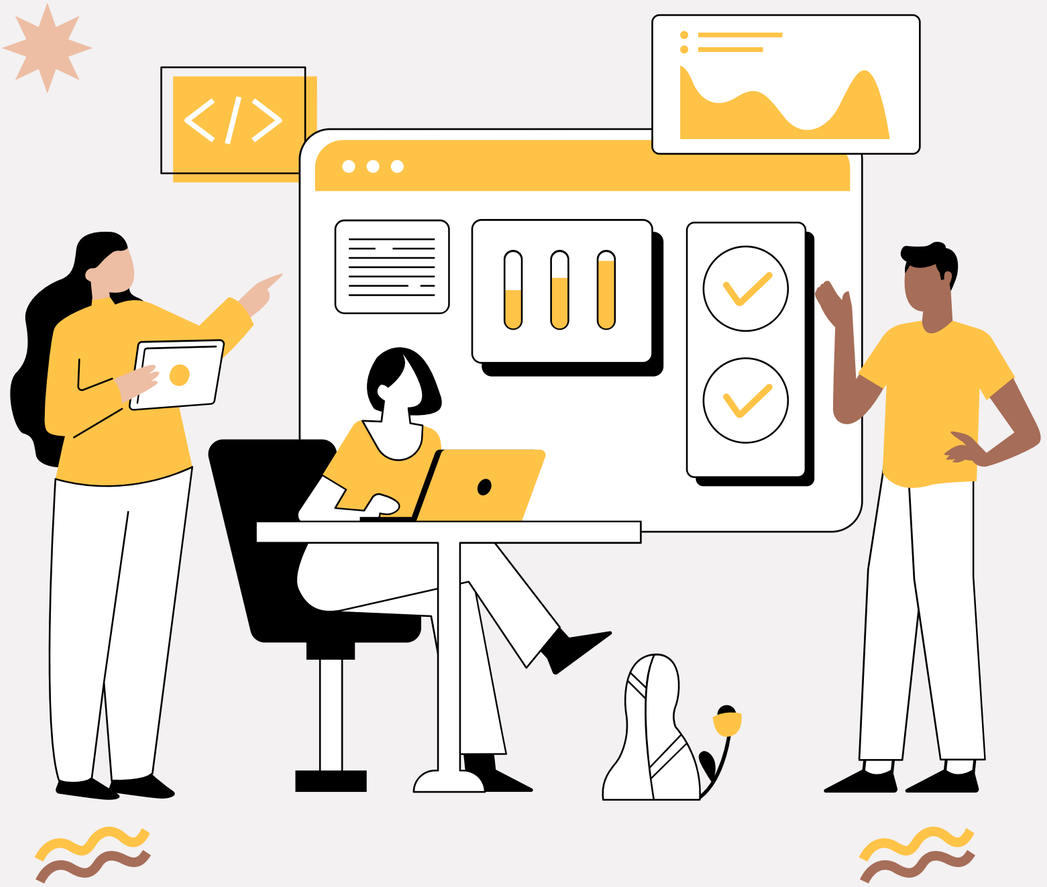
檢視結果

資料筆數由6262 → 5963

研究所： 2074

股票：2039

美食：1850



03
文件分類

資料分割

將資料切分為7:3

```
raw data percentage :
artCatagory
graduate_school    34.781150
stock              34.194198
food               31.024652
Name: proportion, dtype: float64

train percentage :
artCatagory
graduate_school    34.595113
stock              34.163872
food               31.241016
Name: proportion, dtype: float64

test percentage :
artCatagory
graduate_school    35.215204
stock              34.264952
food               30.519843
Name: proportion, dtype: float64
```

將文章轉為DTM

將文章轉為DTM並用LogisticRegression進行模型訓練

依據詞頻

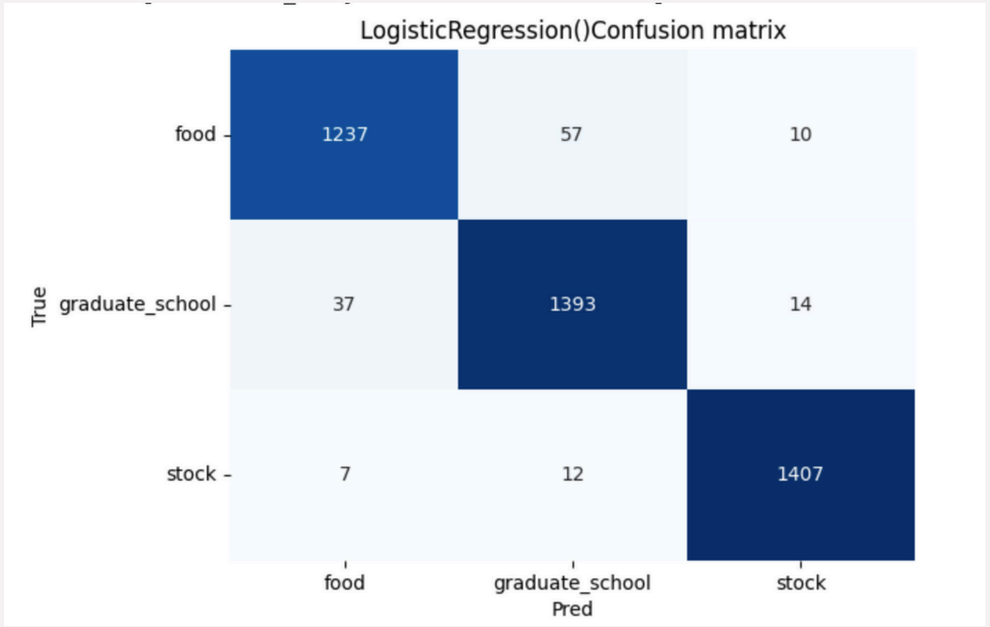
	precision	recall	f1-score	support
food	0.95	0.93	0.94	546
graduate_school	0.93	0.95	0.94	630
stock	0.98	0.97	0.97	613
accuracy			0.95	1789
macro avg	0.95	0.95	0.95	1789
weighted avg	0.95	0.95	0.95	1789

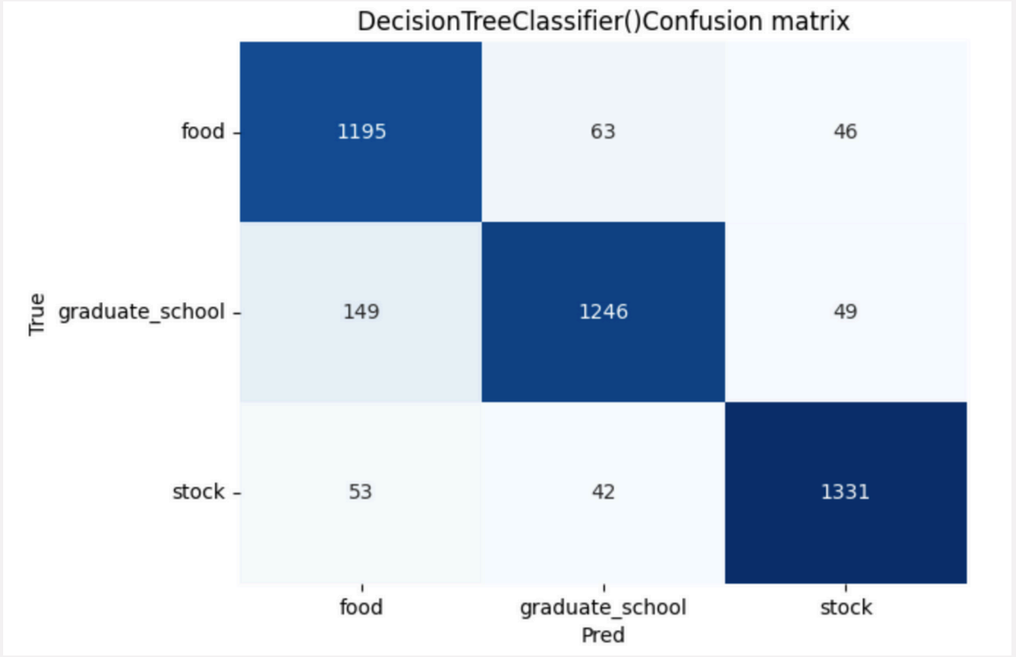
依據TF-IDF

	precision	recall	f1-score	support
food	0.97	0.94	0.96	546
graduate_school	0.94	0.97	0.96	630
stock	0.99	0.98	0.98	613
accuracy			0.97	1789
macro avg	0.97	0.96	0.97	1789
weighted avg	0.97	0.97	0.97	1789

模型比較

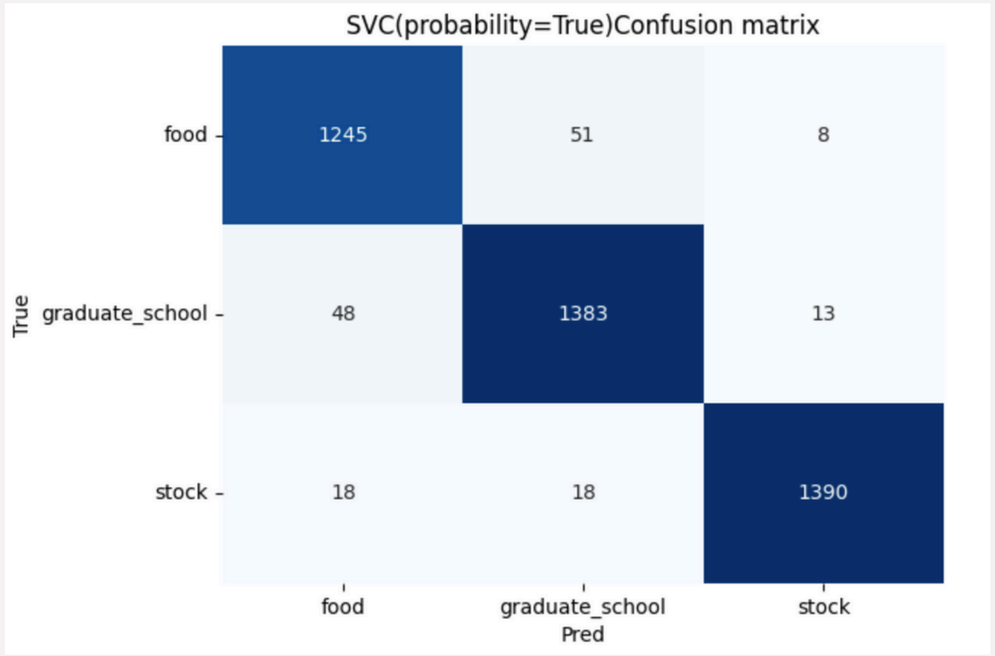
TF-IDF結果好一點點，所以我們用TF-IDF來訓練其他模型進行比較。

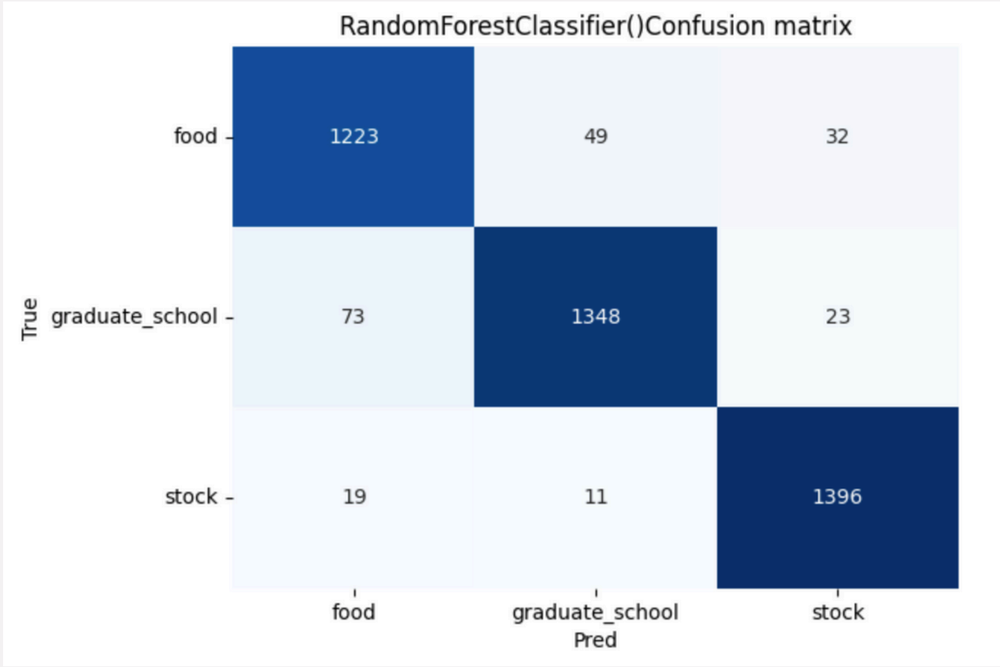




模型比較

TF-IDF結果好一點點，所以我們用TF-IDF來訓練其他模型進行比較。





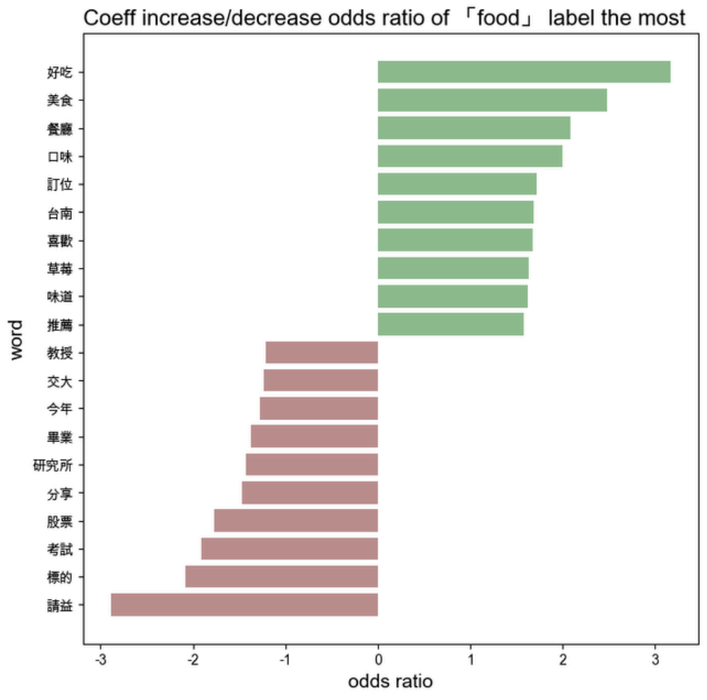
比較結果

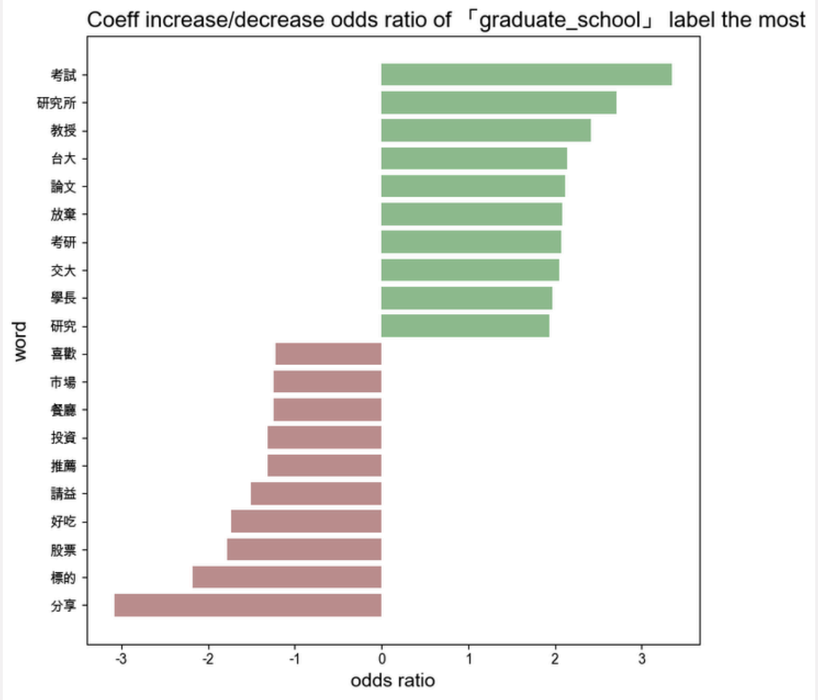
我們以f1-score為指標
最終LogisticRegression為最佳模型

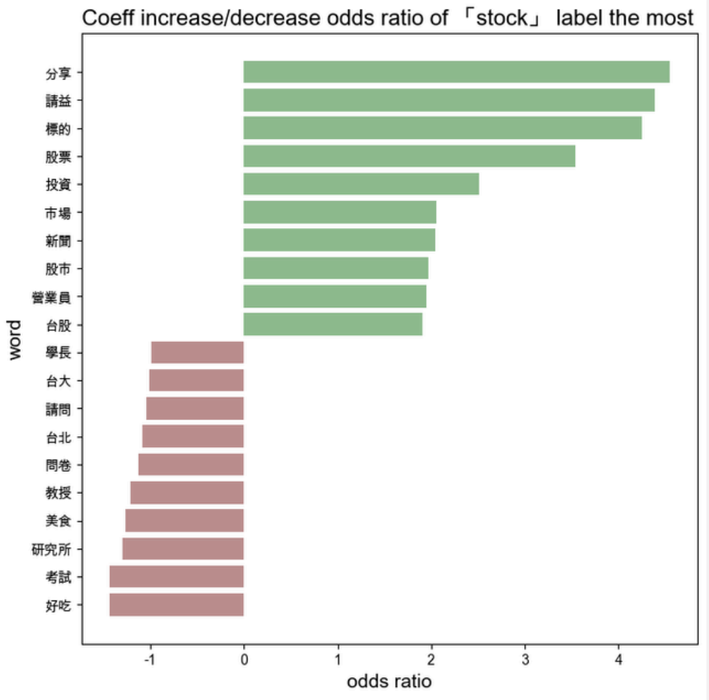
```
y_pred = model_set['clf_logistic'].predict(vectorizer.transform(X_test).toarray())
print(classification_report(y_test, y_pred))
```

```
                precision    recall  f1-score   support

          food       0.97      0.94      0.96       546
graduate_school       0.94      0.97      0.96       630
         stock       0.99      0.98      0.98       613

      accuracy                           0.97      1789
     macro avg       0.97      0.96      0.97      1789
  weighted avg       0.97      0.97      0.97      1789
```

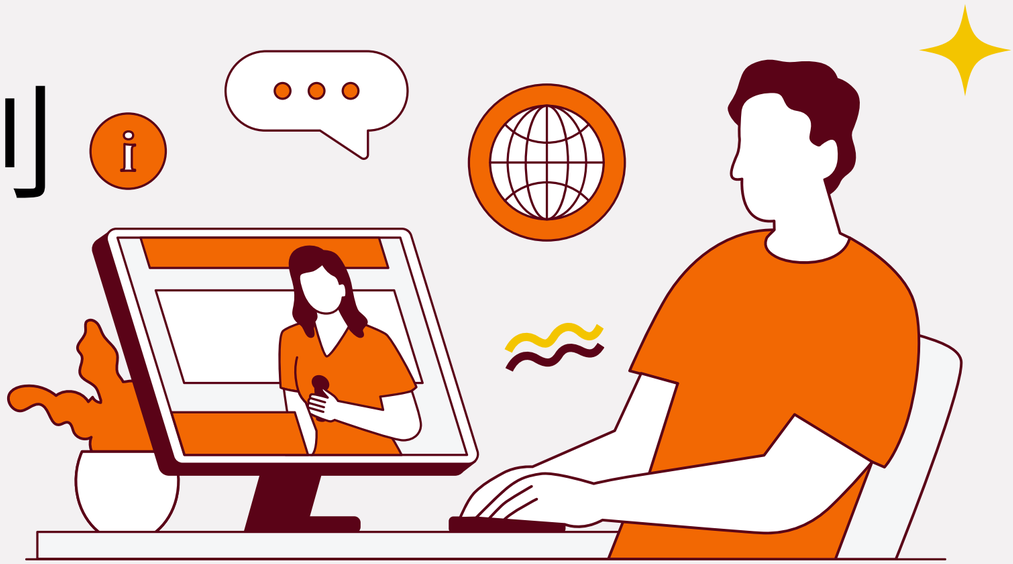
分析可解釋模型的結果


Coeff increase/decrease odds ratio of 「food」 label the most
word
好吃
美食
餐廳
口味
訂位
台南
喜歡
草莓
味道
推薦
教授
交大
今年
畢業
研究所
分享
股票
考試
標的
請益
-3
-2
-1
0
1
2
3
odds ratio


Coeff increase/decrease odds ratio of 「graduate_school」 label the most
word
考試
研究所
教授
台大
論文
放棄
考研
交大
學長
研究
喜歡
市場
餐廳
投資
推薦
請益
好吃
股票
標的
分享
-3
-2
-1
0
1
2
3
odds ratio


Coeff increase/decrease odds ratio of 「stock」 label the most
word
分享
請益
標的
股票
投資
市場
新聞
股市
營業員
台股
學長
台大
請問
台北
問卷
教授
美食
研究所
考試
好吃
-1
0
1
2
3
4
odds ratio

預測其他媒體的資料類別



美食、研究所：PTT

股票：聯合新聞網

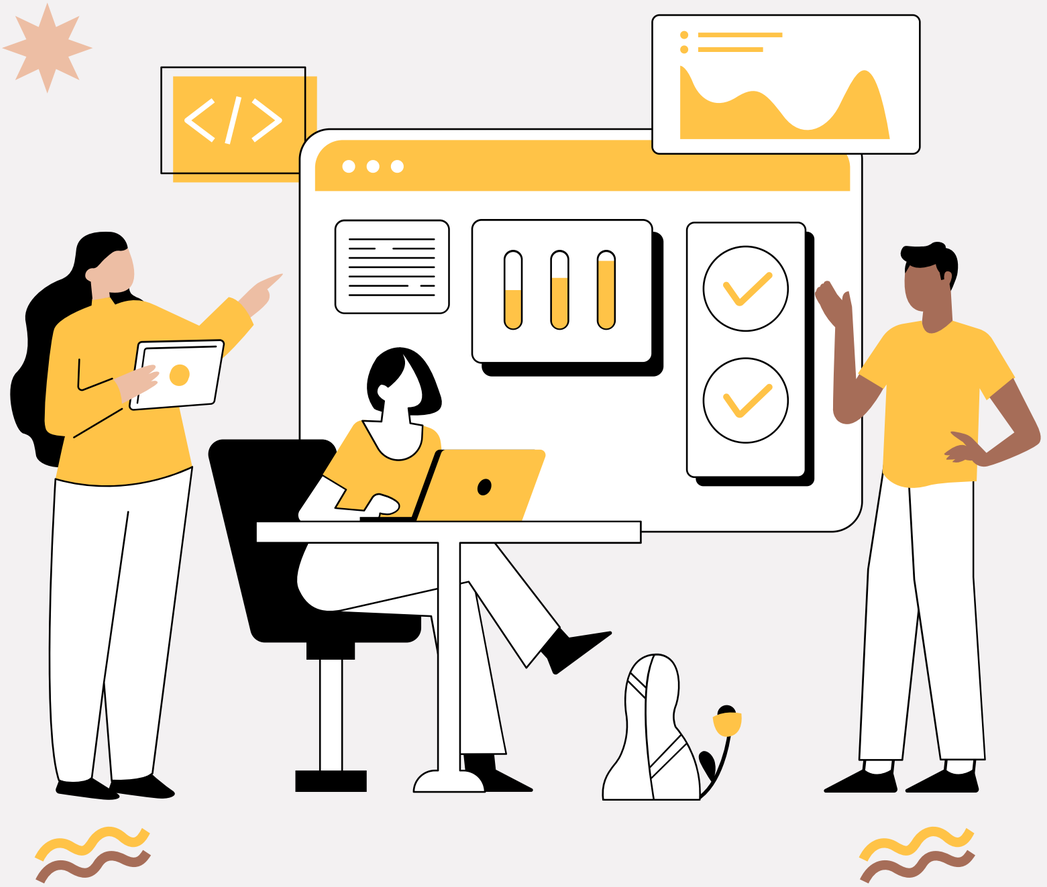
	precision	recall	f1-score	support
food	1.00	1.00	1.00	2373
graduate_school	0.97	0.99	0.98	596
stock	1.00	0.99	1.00	2100
accuracy			0.99	5069
macro avg	0.99	0.99	0.99	5069
weighted avg	0.99	0.99	0.99	5069

預測其他媒體的資料類別

```
false_pred.loc[false_pred['artCatagory']=='stock', :]
```

	words	artCatagory	pred
974	高鐵 便當 連漲 利多 預言 下週 股價 網酸 該漲 台灣 高鐵 公司 表示 高鐵 便當 因...	stock	food
1040	承業醫 下午 點重 記者會 說明 董事會 重大 決議 事宜 承業醫 4164 於今 15 0...	stock	food
1055	證交所 發布 IFRS 永續 揭露 準則 S2 氣候 相關 揭露 參考 範例 協助 企業 順...	stock	graduate_school
1085	潤泰 創新 國際 違反 重大 訊息 內部 控制 制度 規定 證交所 違約金 萬元 潤泰 創新...	stock	graduate_school
1099	證券 承銷商 業務 宣導 13 登場 臺灣 證券 交易所 預計 13 舉行 證券 承銷商 業...	stock	graduate_school
1266	翔耀 違反 重訊及 內控 規定 證交所 違約金 20 萬元 翔耀 實業 2438 2024 ...	stock	graduate_school
1290	證交所 舉辦 證券商 導入 金融 信任 架構 作業 問卷 說明會 臺灣 證券 交易所 12 ...	stock	graduate_school
1321	證券商 導入 金融 信任 架構 作業 問卷 證交所 教戰 臺灣 證券 交易所 日前 舉辦 推...	stock	graduate_school
1639	證交所 新代 科技 派司 上市 申請 及熙 特爾 新能源 申請 創新 上市 臺灣 證券 交易...	stock	graduate_school

'高鐵便當連漲2年是利多？他預言下週**股價跟漲** 網酸：該漲的是它台灣高鐵公司表示，「**高鐵便當**」因應**食材原物料**等成本持續增加，自今年4月起調整售價為120元。高鐵公司表示，便當菜色仍將堅持採用在地化特色食材，提供旅客搭車時更美味、方便的餐食選擇。網友不解表示，該漲的應該是票價，而不是便當，也有一票人直呼「台鐵便當完勝！」「高鐵便當」從2015年4月1日起至2024年3月31日，長達9年售價維持在100元，然因**食材、原物料等成本持續增加**，去年4月1日首次調漲至110元，今年4月起再調整售價為120元。此外， TGO會員95折便當優惠至今年3月底截止。一名網友在PTT轉題新聞並表示，高鐵便當已經比去年初上漲20%，過去是凍漲9年，如今卻連續兩年調漲，他笑稱「這樣下週對高鐵應該是利多吧？股價是不是準備噴噴了！」其他人留言「板橋，台北，南港這三站的旅客可以改買台鐵便當再進去搭高鐵」、「坐高鐵吃台鐵便當」、「台鐵便當屌打」、「我會寧願去其他地方買的說，這CP值真的不高」、「還好我都買台鐵便當」。不少人認為該漲的是車票，留言「最好票價也漲20%」、「漲便當衝三小？自由座給我先漲20%」、「票價也應該漲」、「票價不漲，便當來補」、「電費漲，開漲票價了吧」、「怎不敢漲票價」、「不懂為啥票價不漲」。'



04
主題分類模型 LDA

將主題數設為10

當分為10個類別可以發現指標普遍<0

->重分！！！

```
ldamodel = LdaModel(
    corpus=corpus,
    id2word=dictionary, # 字典
    num_topics=10, # 生成幾個主題數
    random_state=2024, # 亂數
)
```

✓ 5.1s

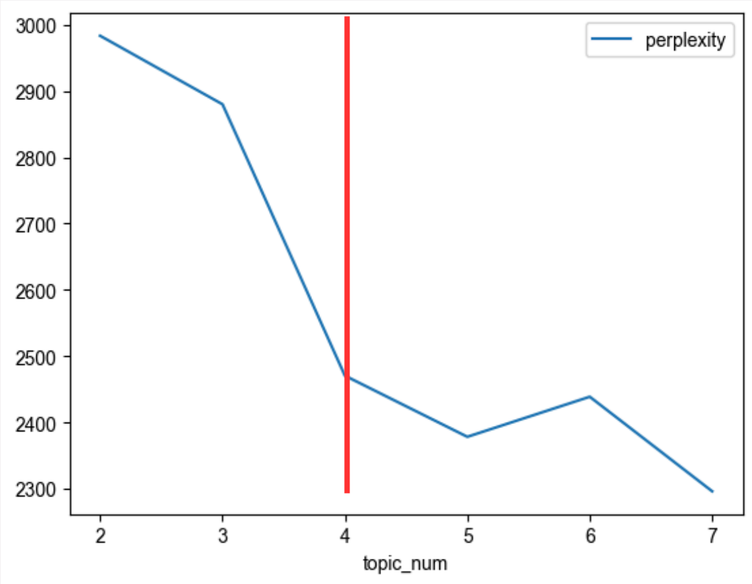
```
NPMI_model_lda.get_coherence_per_topic()
```

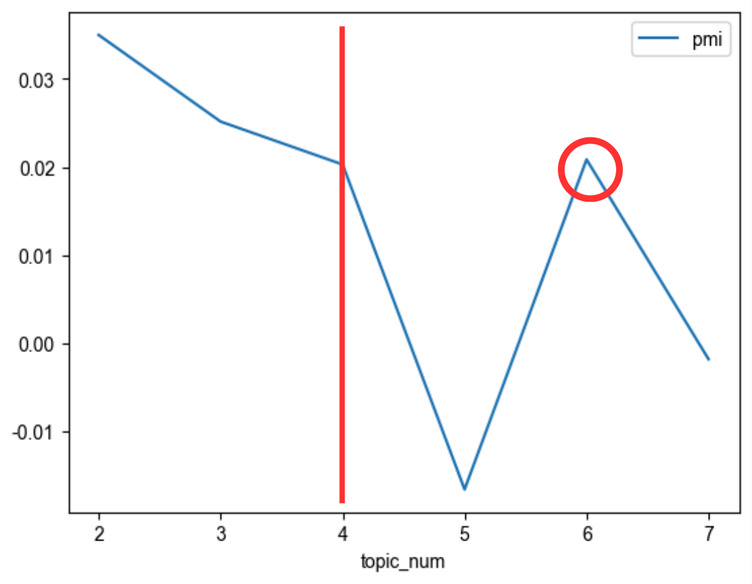
✓ 0.3s

```
[-0.08712904260751009,
 -0.21220646093527773,
 0.0038690881519432103,
 -0.17878548452798343,
 -0.1717886075471941,
 -0.10389117115855821,
 0.027925053073445706,
 -0.0938136737312601,
 -0.1081585042526105,
 -0.08893315513674334]
```

找出最佳指標數

將主題設為2~8

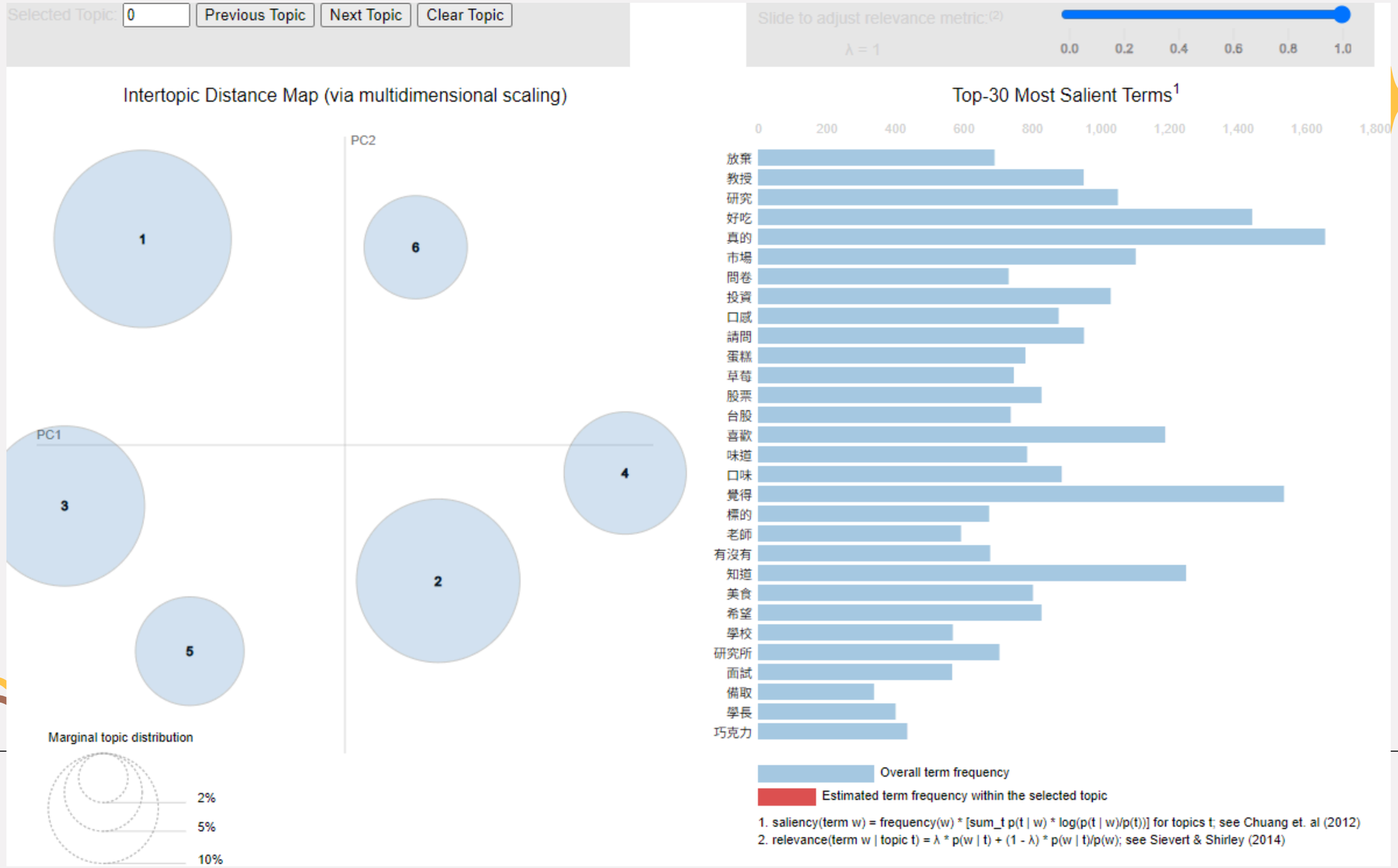




INTERTOPIC DISTANCE MAP（主題距離圖）

KEITHSTON AND PARTNERS

WWW.REALLYGREATSITE.COM



GUIDEDLDA

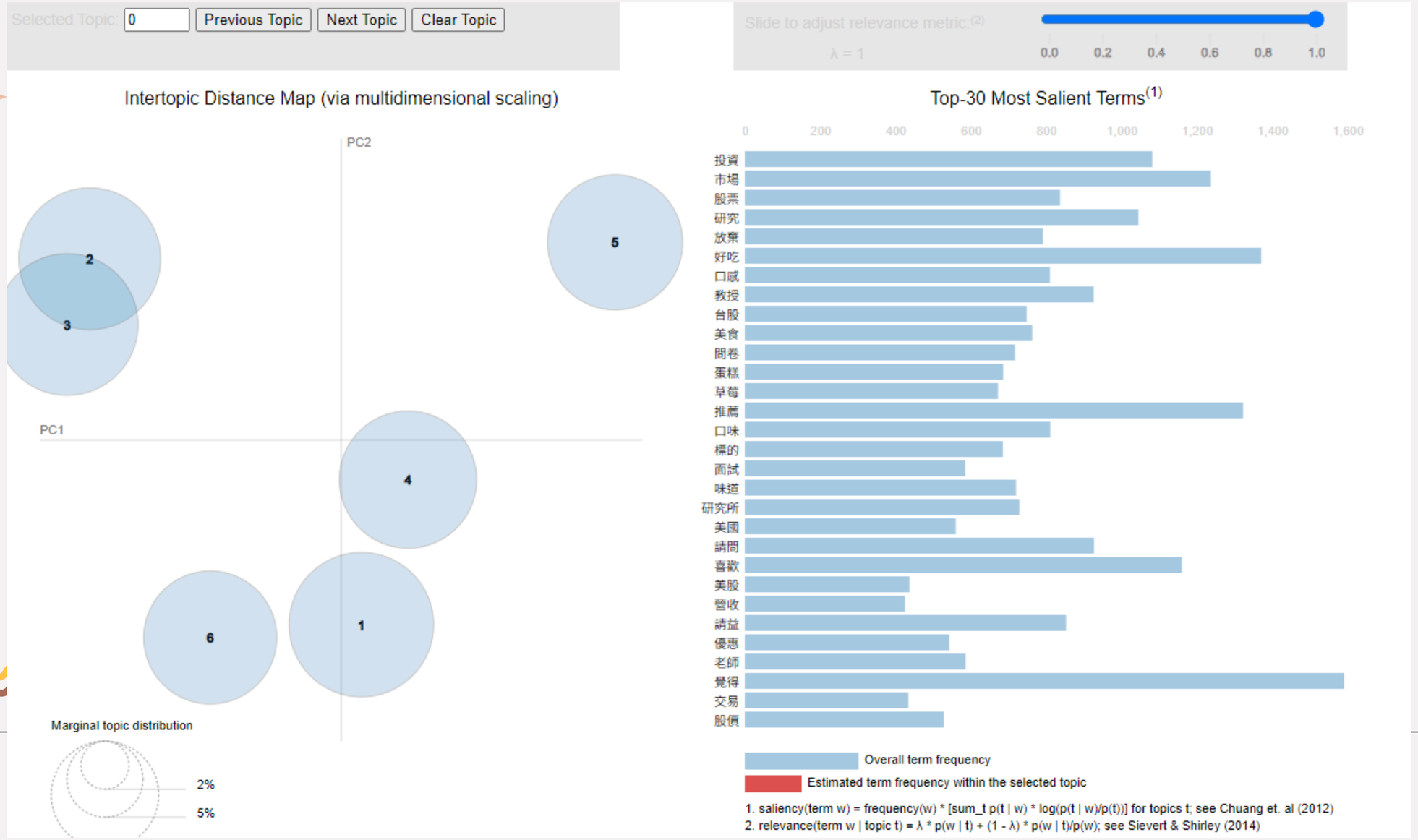
```
seed_topic_list = [
    ["投資", "股票", "台股", "美股", "分析"],  #股市與投資討論
    ["蛋糕", "巧克力", "口味", "料理", "火鍋"],#飲食經驗分享（料理篇）
    ["研究", "教授", "考試", "研究所", "面試"],#升學與學術壓力
    ["放棄", "希望", "覺得", "有沒有", "謝謝"],#情緒與求助分享
    ["好吃", "推薦", "美食", "餐廳", "口味"],  #美食推薦與評價
    ["口感", "味道", "草莓", "蛋糕", "巧克力"] #甜點風味與感官描述
]
```

✓ 0.0s

INTERTOPIC DISTANCE MAP（主題距離圖）

KEITHSTON AND PARTNERS

WWW.REALLYGREATSITE.COM



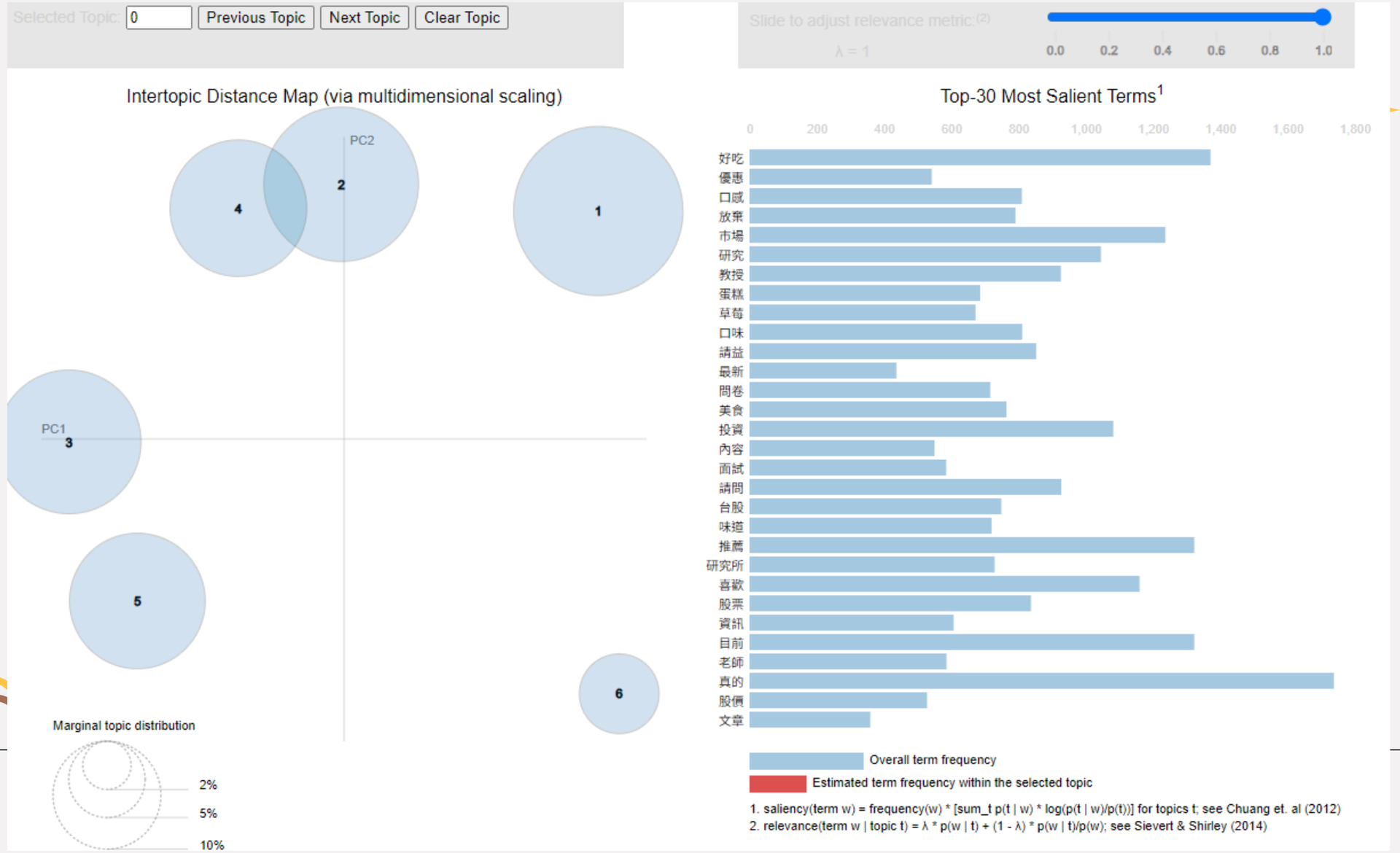
GUIDEDLDA

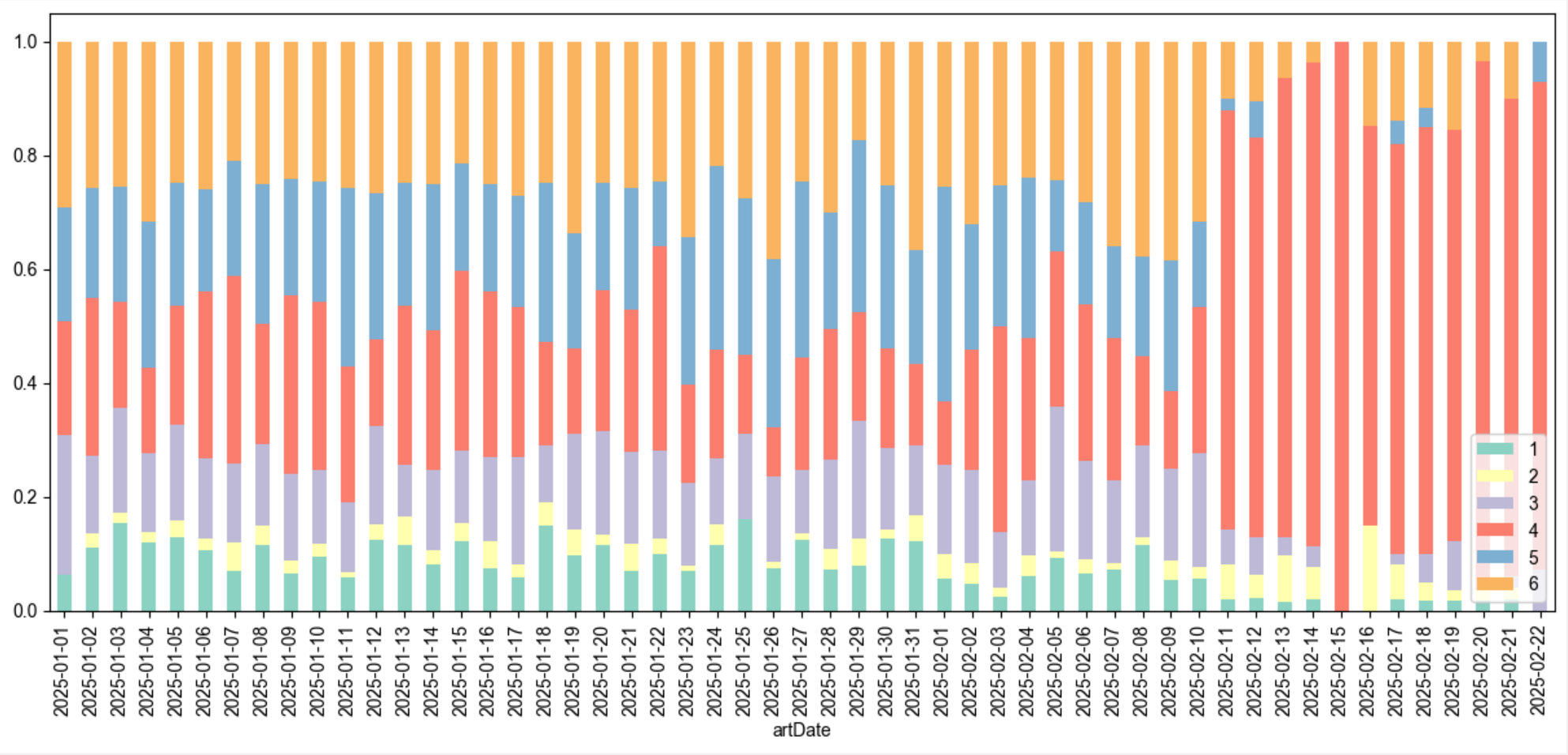
```
seed_topic_list = [
    # 股市與投資討論
    ["投資", "股票", "台股", "美股", "分析", "報酬", "財報", "市場", "基金", "利率", "股市", "股價", "經濟", "操盤"],

    # 飲食經驗分享（料理篇）
    ["火鍋", "料理", "麻辣", "炸物", "日式", "韓式", "中餐", "晚餐", "便當", "燒烤", "套餐", "點餐", "食材"],

    # 升學與學術壓力
    ["研究", "教授", "考試", "研究所", "面試", "筆試", "報考", "成績", "推甄", "論文", "實驗室", "科系", "口試", "準備", "學歷"],

    # 情緒與求助分享
    ["放棄", "希望", "覺得", "謝謝", "壓力", "崩潰", "心情", "求救", "有人", "煩惱", "憂鬱", "焦慮", "很累", "痛苦"],

    # 美食推薦與評價
    ["好吃", "推薦", "美食", "餐廳", "名店", "排隊", "小吃",  "用餐", "服務", "訂位", "評價", "食記"],

    # 甜點風味與感官描述
    ["蛋糕", "巧克力", "口感", "味道", "草莓", "甜點", "奶油", "脆皮", "綿密", "苦甜", "布丁", "慕斯", "甜食"]
]
```

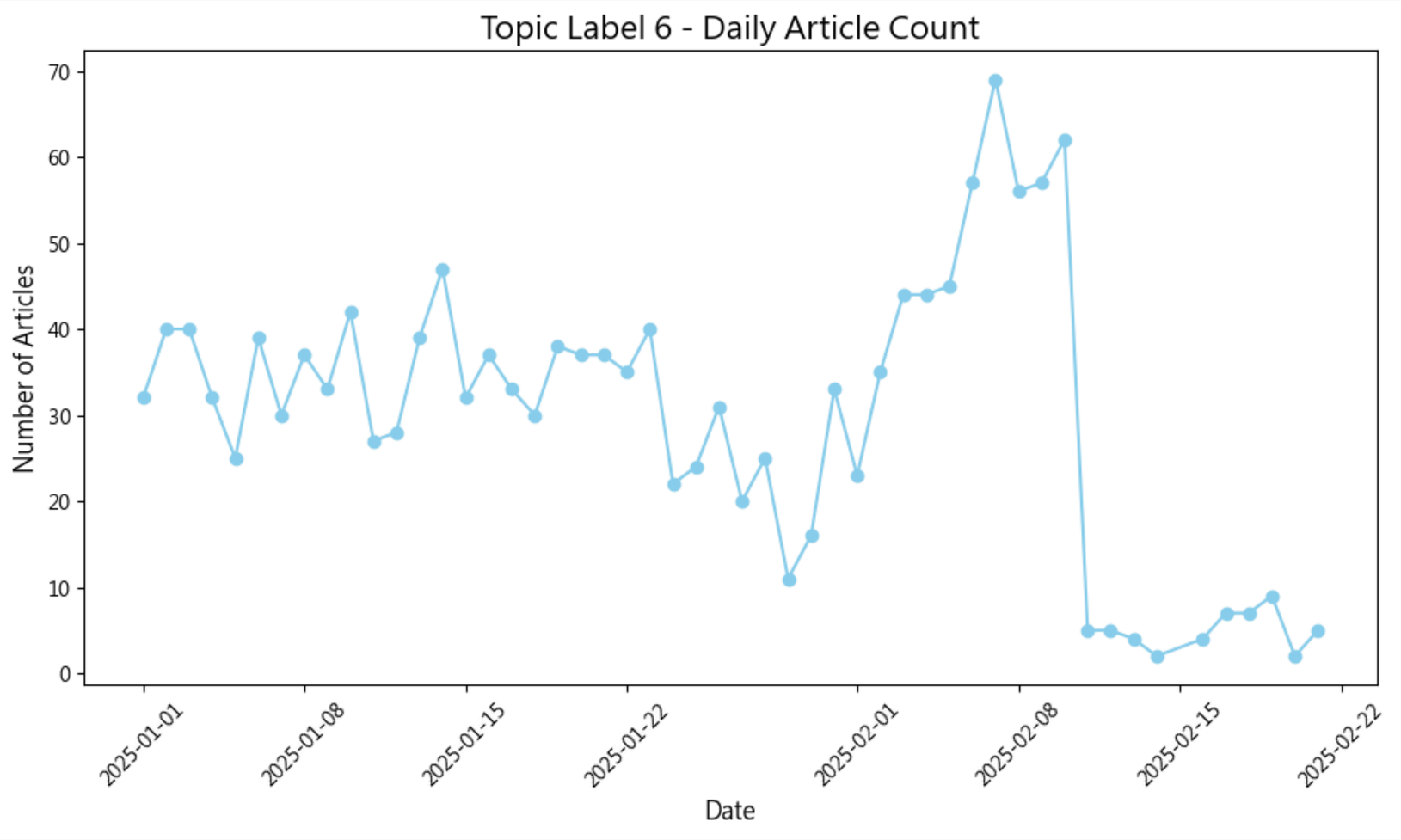
INTERTOPIC DISTANCE MAP（主題距離圖）

KEITHSTON AND PARTNERS

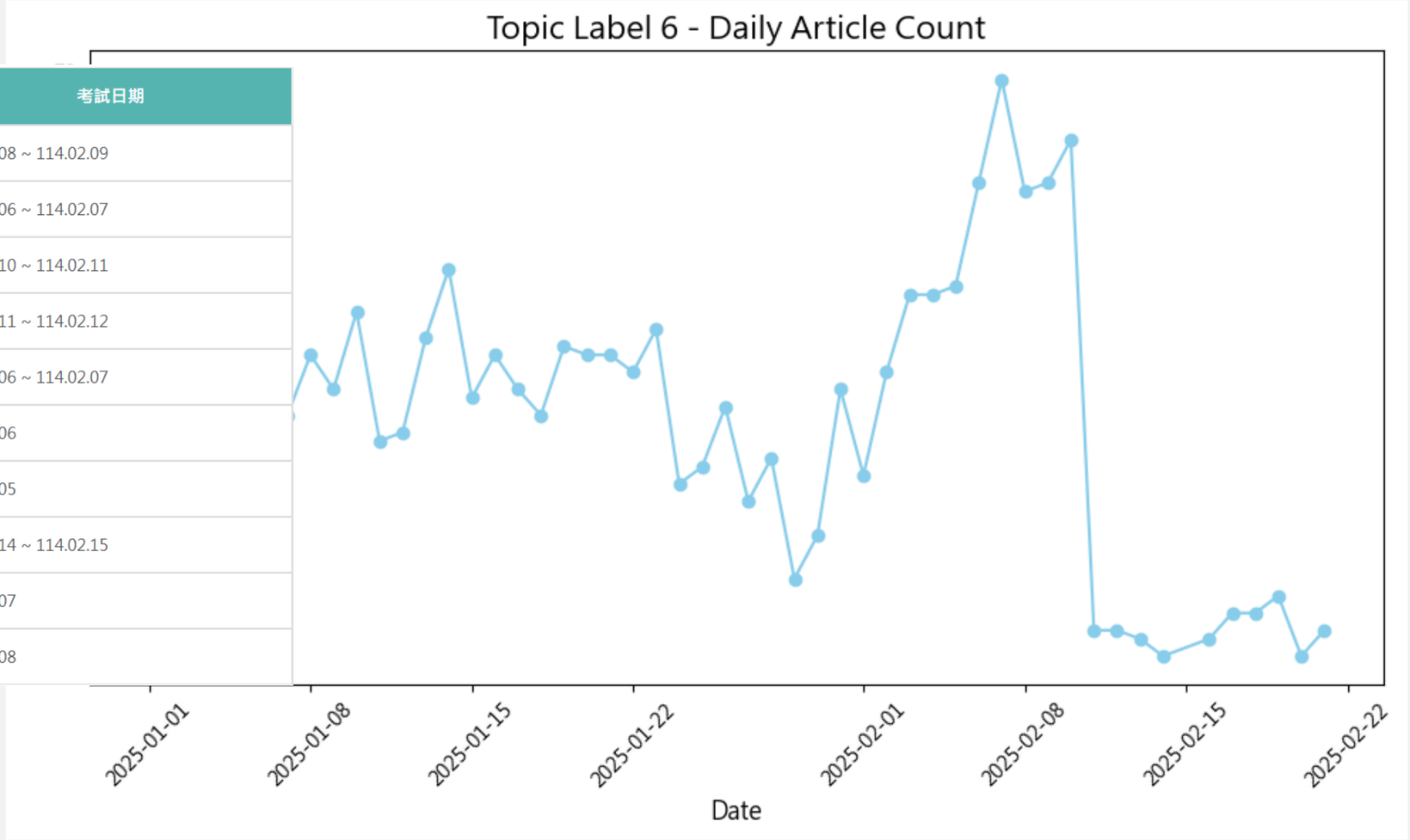
WWW.REALLYGREATSITE.COM

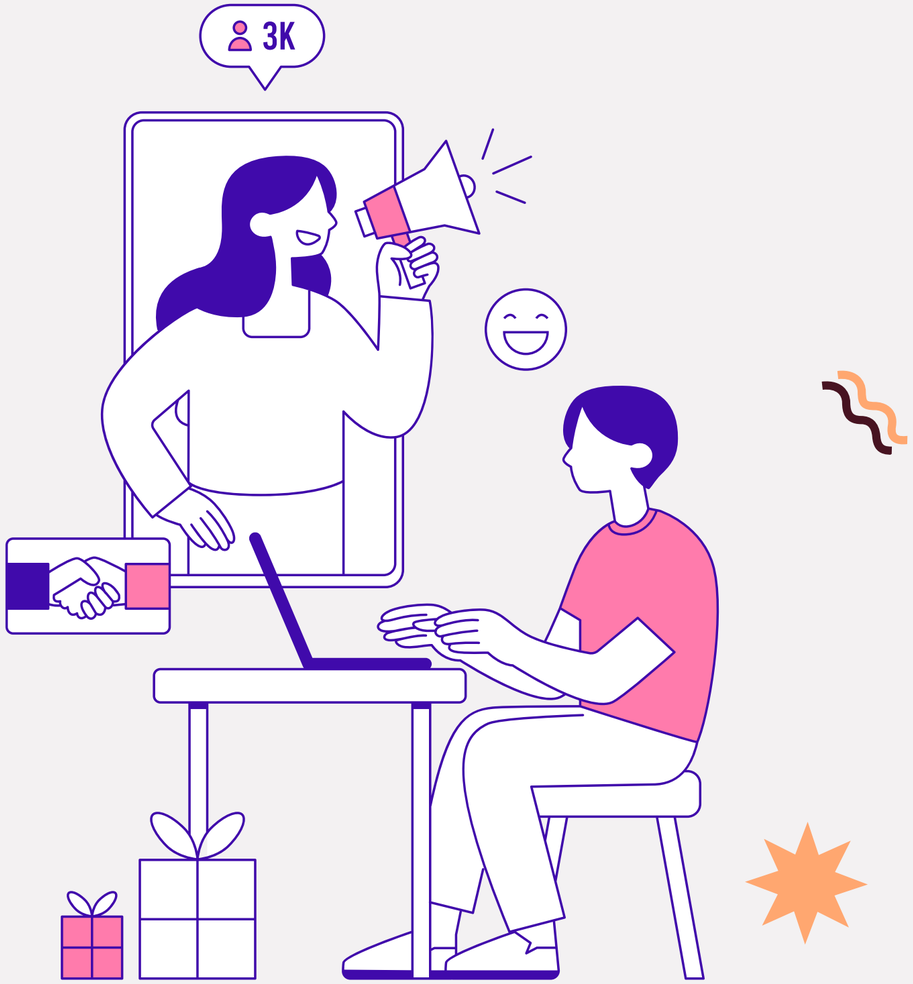



1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
2025-01-01
2025-01-02
2025-01-03
2025-01-04
2025-01-05
2025-01-06
2025-01-07
2025-01-08
2025-01-09
2025-01-10
2025-01-11
2025-01-12
2025-01-13
2025-01-14
2025-01-15
2025-01-16
2025-01-17
2025-01-18
2025-01-19
2025-01-20
2025-01-21
2025-01-22
2025-01-23
2025-01-24
2025-01-25
2025-01-26
2025-01-27
2025-01-28
2025-01-29
2025-01-30
2025-01-31
2025-02-01
2025-02-02
2025-02-03
2025-02-04
2025-02-05
2025-02-06
2025-02-07
2025-02-08
2025-02-09
2025-02-10
2025-02-11
2025-02-12
2025-02-13
2025-02-14
2025-02-15
2025-02-16
2025-02-17
2025-02-18
2025-02-19
2025-02-20
2025-02-21
2025-02-22
artDate


Topic Label 6 - Daily Article Count
Number of Articles
0
10
20
30
40
50
60
70
2025-01-01
2025-01-08
2025-01-15
2025-01-22
2025-02-01
2025-02-08
2025-02-15
2025-02-22
Date

學校	考試日期
國立台灣大學	114.02.08 ~ 114.02.09
國立清華大學	114.02.06 ~ 114.02.07
國立成功大學	114.02.10 ~ 114.02.11
國立政治大學	114.02.11 ~ 114.02.12
國立陽明交通大學	114.02.06 ~ 114.02.07
國立中正大學	114.02.06
國立中山大學	114.02.05
國立中央大學	114.02.14 ~ 114.02.15
國立中興大學	114.02.07
國立臺灣科技大學	114.02.08

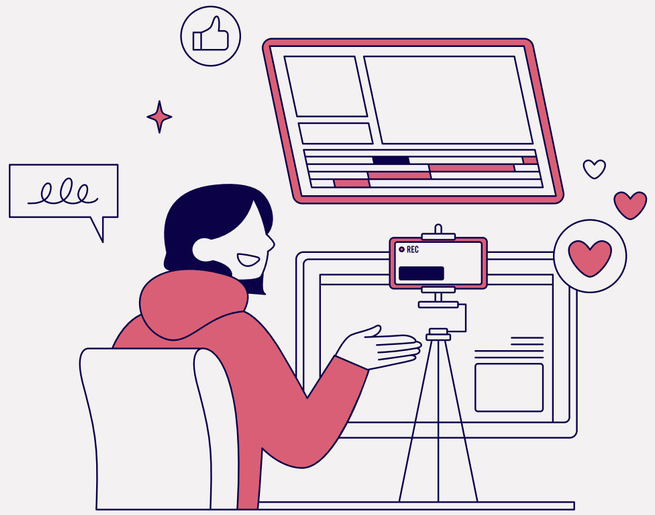




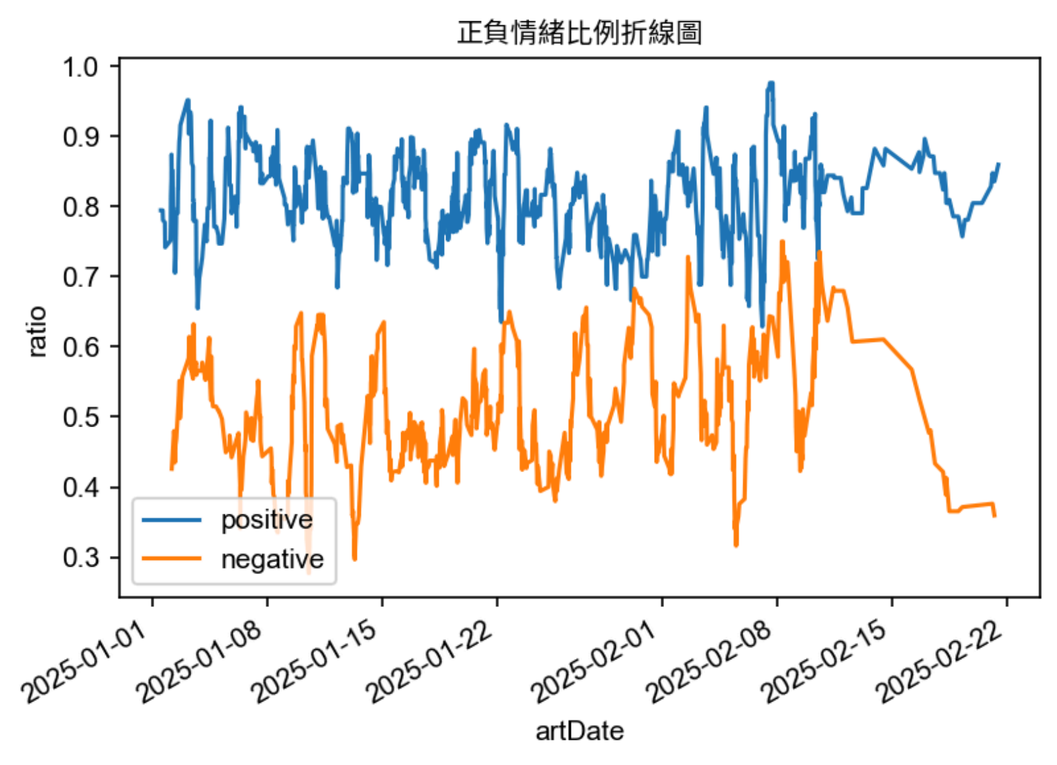
05
結合情緒分析

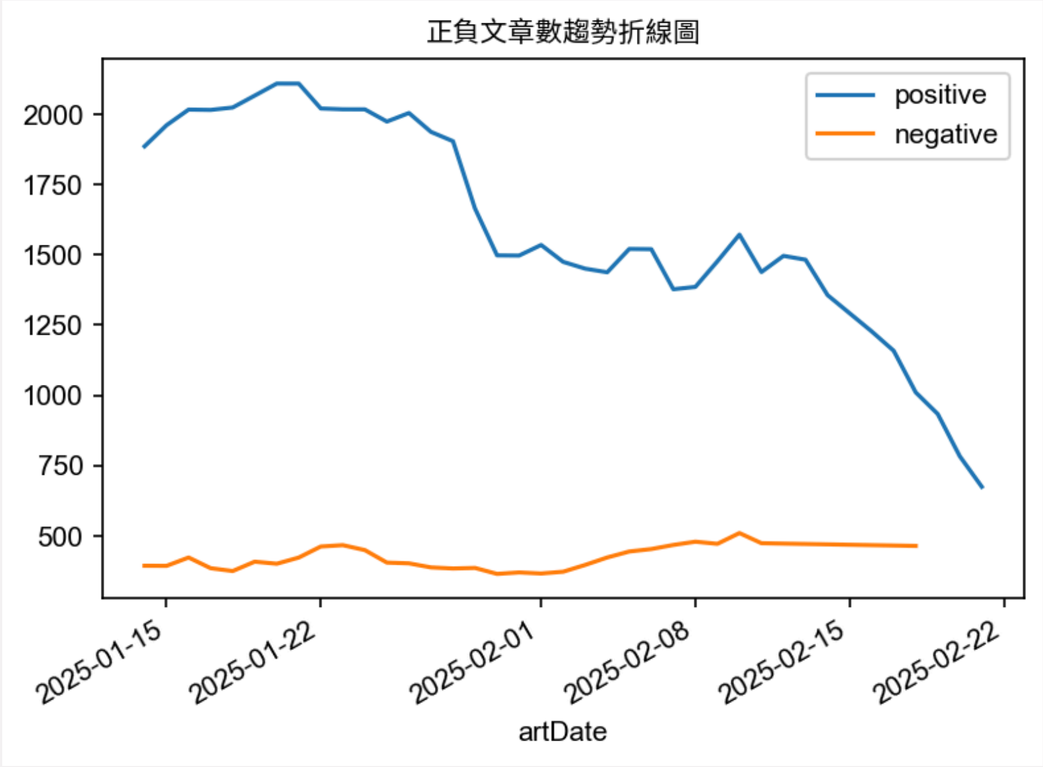
查看各類別的文章數

```
dcard_lable_4["artCatagory"].value_counts()
✓ 0.0s
artCatagory
graduate_school    1397
stock               105
food                 33
Name: count, dtype: int64
```



正負向情緒詞彙比例折線圖 & 文章為單位的情緒分析





查看2025-02-08 ~2025-02-20 這段時間出現了什麼關鍵字

負面

正面

Thank you.