社群媒體分析
第二次讀書會報告

教 授 ：黃三益
第11組：
N124320004_ 莊筑雅 / N124320006_ 陳勁廷
N124320009_ 吳亭緯 / N124320027_ 楊毅婷
N124320030_ 傅才容 / N124320023_ 蕭瑜垚
N124320016_ 李明容 / N124320003_ 馬永恩

文本資料介紹

1. 林襄資料集

- 資料來源：
  八卦版的林襄相關文章
- 日期區間：
  2022-01-01 ~ 2025-03-31

2. 新聞資料集

東森新聞網
聯合新聞網

●模型訓練資料內容：
資料來源：東森新聞網(EBC)
版別：體育、國際、財經
區間：2024/09/01-2025/02/28

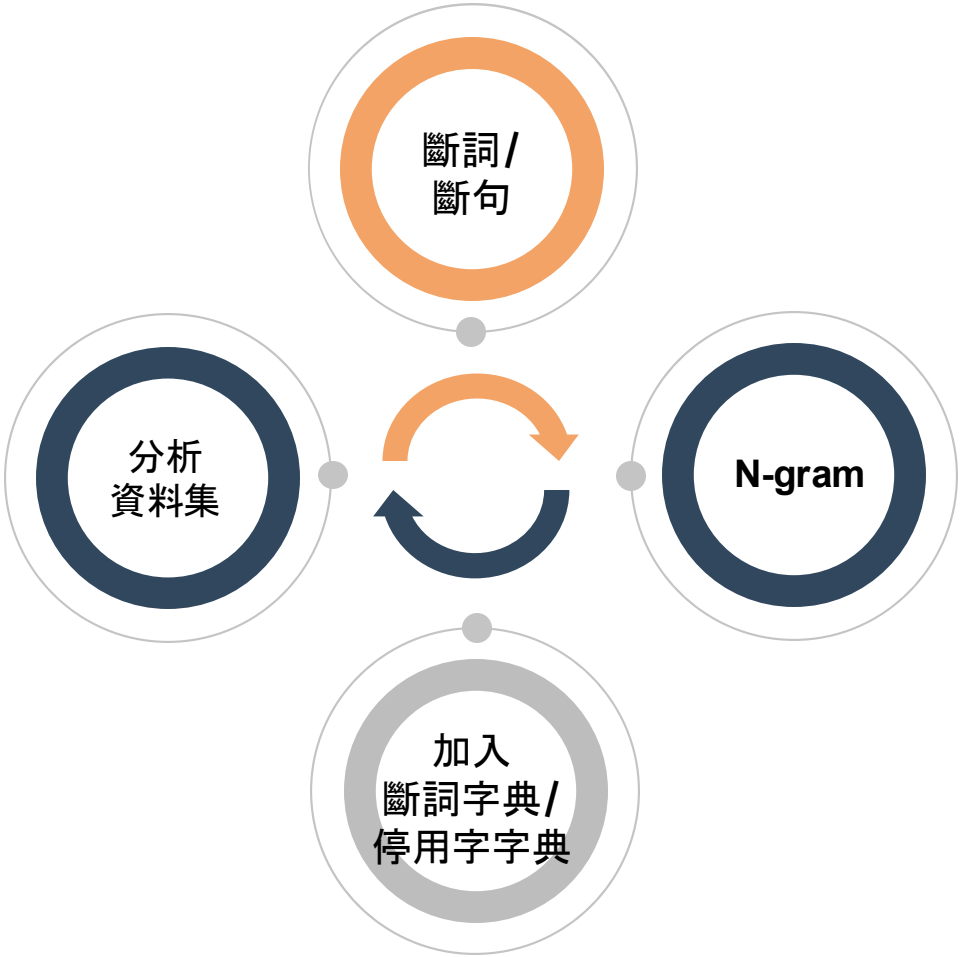
●測試資料內容：
資料來源：聯合新聞網(UDN)
版別：運動、產經、股市、全球
區間：2025/03/15-2025/03/31

分析PTT八卦版 林襄 相關文章之詞彙關係

詞彙關係

文件分類

主題模型



資料清洗

欲分析的資料集先進行斷詞/斷句，再利用 bigram 和 trigram 方式反覆測試結果，將各字組以參數之形式手動加入字典中，幫助斷詞能更準確，直到最終沒有可以合併的字詞。

N-gram最終結果

▼Bigram

	word	count
0	小妹 台大	71
1	肥宅 小妹	56
2	台灣 美女	48
3	女神 林襄	47
4	啦啦隊女神 林襄	41
5	看到 林襄	41
6	企鵝 姐姐	41
7	林襄 李多慧	38
8	隔壁 房間	38
9	房間 睡覺	35

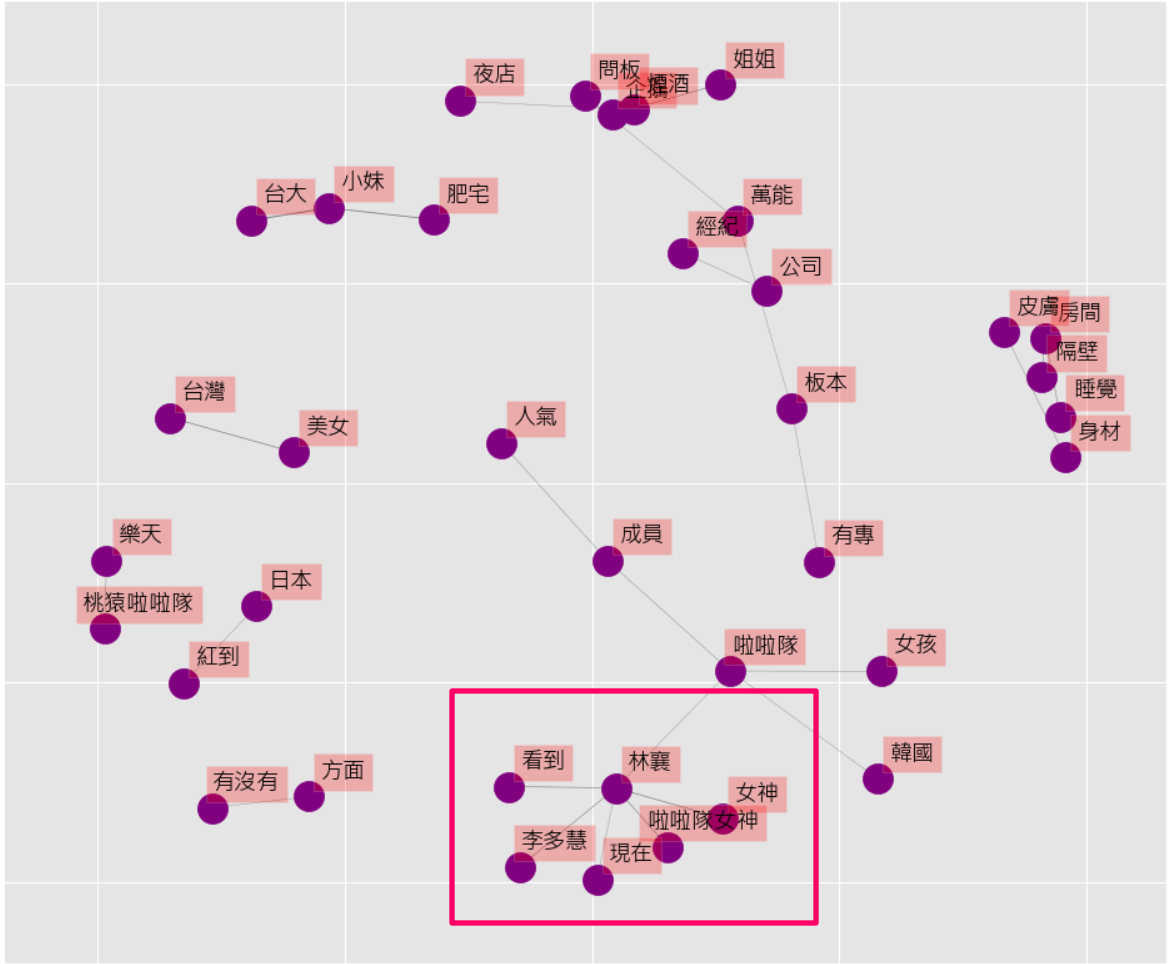
▼Trigram

	word	count
0	肥宅 小妹 台大	56
1	隔壁 房間 睡覺	35
2	有專 板本 萬能	23
3	板本 萬能 問板	23
4	政治 問卦 嚴重	22
5	問卦 嚴重 板論	22
6	不煙 不酒 夜店	19
7	萬能 問板 三則	16
10	問板 三則 政治	15
9	三則 政治 問卦	15

Bigram視覺化

取得 bigram 斷詞間的出現頻率

可看到("**林襄**"、"**李多慧**"、"**啦啦隊女神**")等字詞常一起出現討論。



詞彙相關性(一)

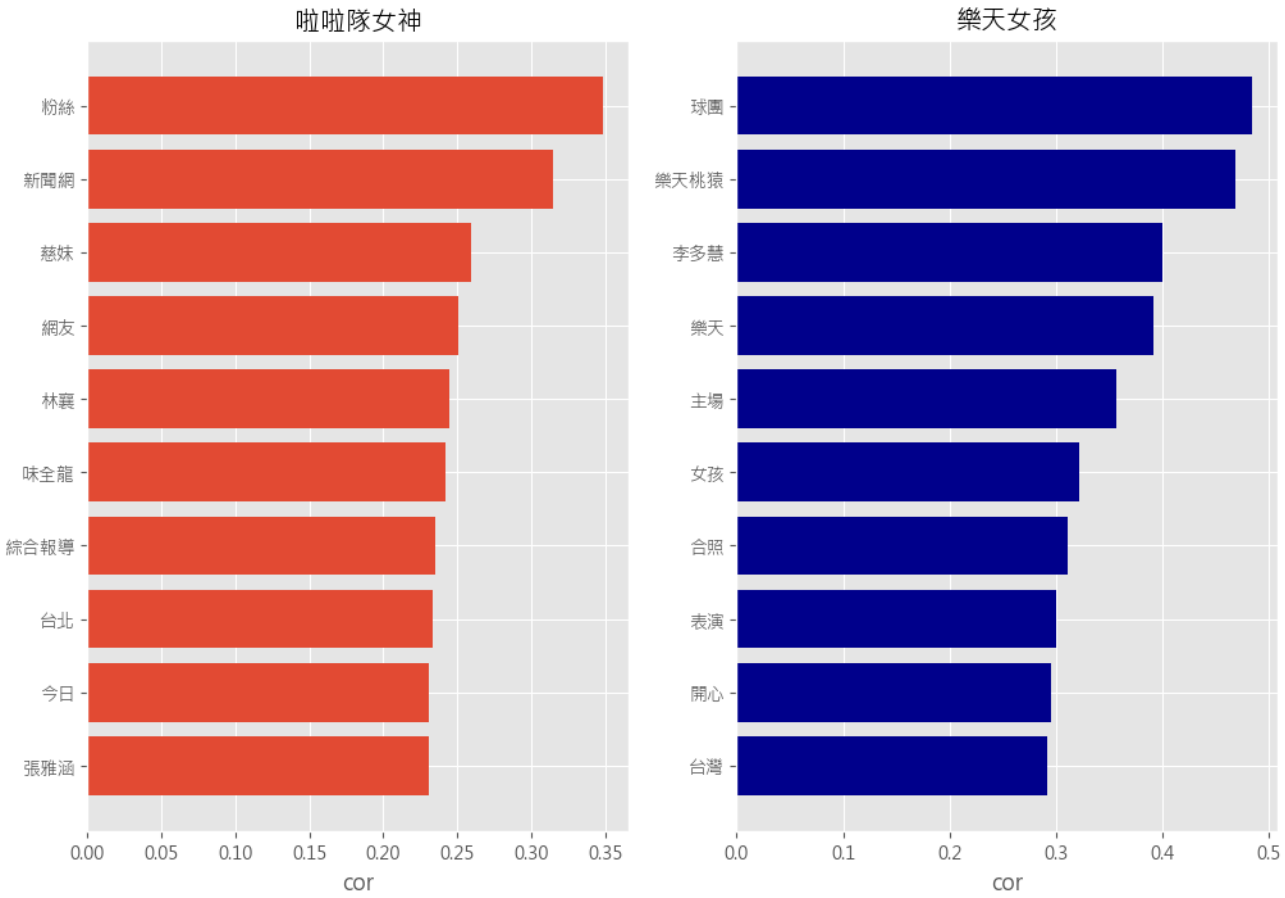
利用Pearson correlation計算兩個詞彙間的相關性

條件：
篩選至少出現在5篇文章以上且詞頻前300的詞彙

圖左可看到 **啦啦隊女神** 和**("粉絲"、"慈妹"、"林襄")**這些詞的相關性比較高。

圖右可看到 **樂天女孩** 和**("球團"、"李多慧")**這些詞的相關性比較高。

找出和「啦啦隊女神」「樂天女孩」相關性最高的10個詞彙



詞彙相關性(二)

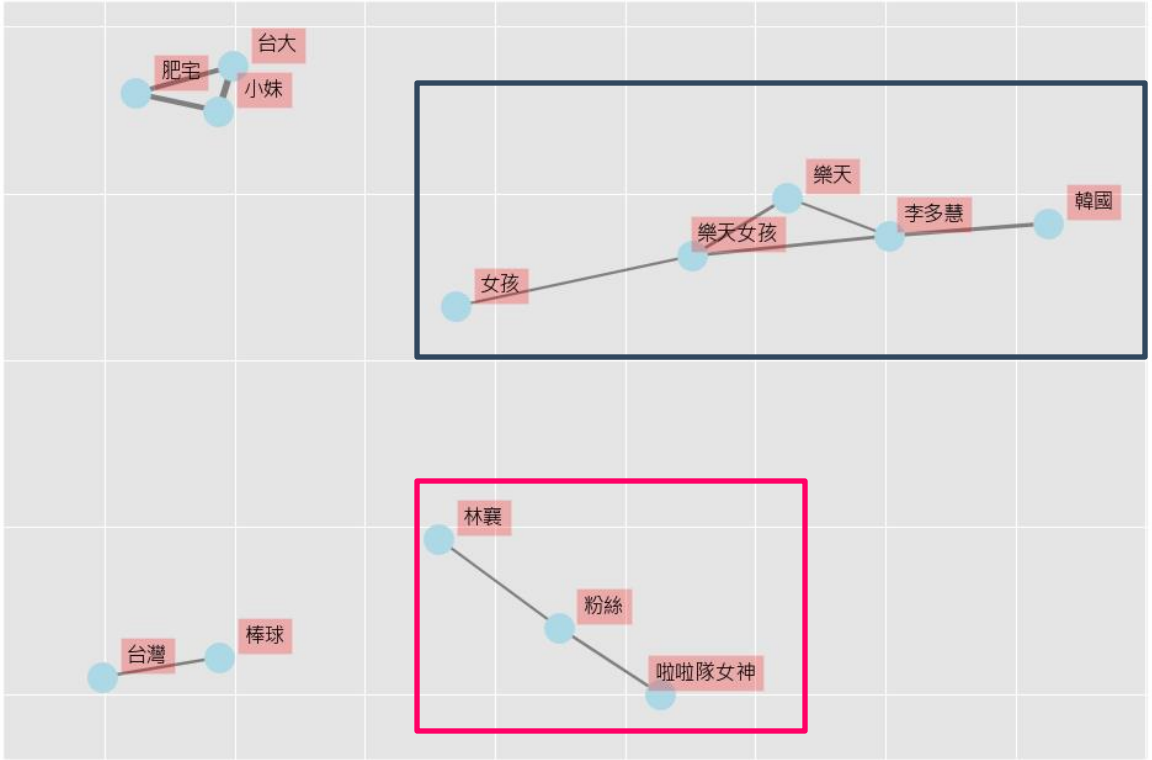
透過DTM找出詞頻高的詞彙

	word1	word2	cor
0	小妹	台大	0.777559
5	台大	小妹	0.777559
20	小妹	肥宅	0.627336
6	肥宅	小妹	0.627336
19	台大	肥宅	0.576018
1	肥宅	台大	0.576018
9	韓國	李多慧	0.466964
21	李多慧	韓國	0.466964
8	樂天女孩	李多慧	0.400447
15	李多慧	樂天女孩	0.400447
13	樂天女孩	樂天	0.391354
16	樂天	樂天女孩	0.391354
2	棒球	台灣	0.363027
11	台灣	棒球	0.363027
3	粉絲	啦啦隊女神	0.348416
17	啦啦隊女神	粉絲	0.348416
4	樂天女孩	女孩	0.321881
14	女孩	樂天女孩	0.321881
10	粉絲	林襄	0.313476
18	林襄	粉絲	0.313476
12	李多慧	樂天	0.302909
7	樂天	李多慧	0.302909

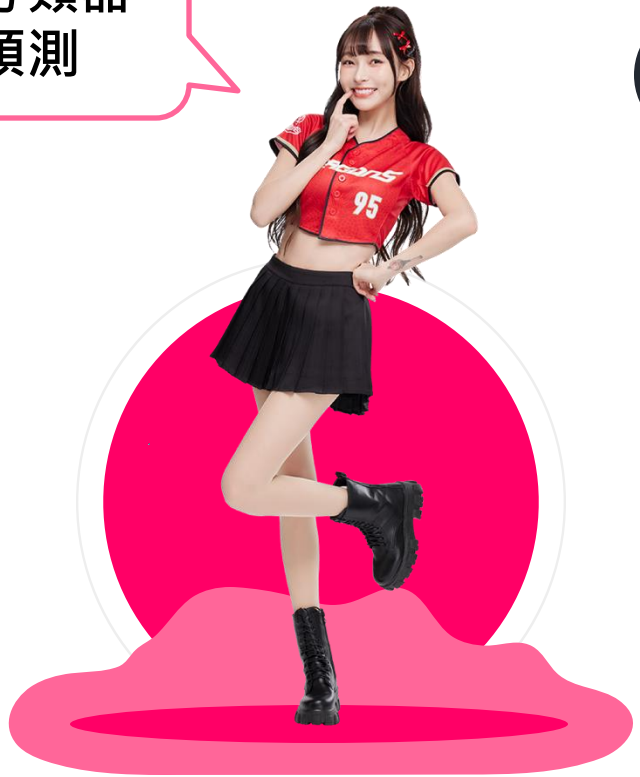
條件：

以詞頻前60為節點且相關性高於0.3的組合

可看到(“林襄”、“粉絲”、“啦啦隊女神")、("樂天女孩"、"李多慧"、"韓國")這些字眼常一起出現。



新聞網分類器
建模及預測



詞彙關係

文件分類

主題模型

```
number of posts: 6309
date range: (Timestamp('2024-09-01 06:00:00'), Timestamp('2025-02-28 22:57:00'))
category:
artCatagory
國際    2644
體育    1621
財經     980
Name: count, dtype: int64
```

```
▷Initialize Reactive Jupyter | Sync all Stale code
# 設定每類別抽取的最大篇數（例如 1000 篇）
target_samples = 1000
# 分類別抽取樣本，並保證每個類別最多抽取 target_samples 篇
ebc = ebc_raw.groupby('artCatagory').apply
(lambda x: x.sample(n=min(target_samples, len(x)), random_state=42)).reset_index(drop=True)
# 顯示抽樣後的結果
print(f"Balanced category distribution: \n{ebc['artCatagory'].value_counts()}")
# 檢視抽樣後的資料
ebc.head(3)
```

```
Balanced category distribution:
artCatagory
國際    1000
體育    1000
財經     980
```

東森新聞文本資料抓同樣區間：2024/09/01-2025/02/28，發現各文本類別筆數差異過大。

為避免分類器模型，容易學會只預測「多數類別」就能得到高準確率類別，財經類別最少為980篇，故將「國際」及「體育」版別資料，隨機抽樣1,000篇以平衡類別。

清洗與斷詞：

-統一日期格式
-過濾NAN資料
-移除網址格式
-只留中文字
-斷詞
-篩掉停用字與字元數大於1的詞彙

清洗/斷詞 → 東森新聞-分類模型訓練 → 最佳模型 → 聯合新聞- 分類預測 → 洞見觀察

```
raw data percentage :
artCatagory
國際    33.557047
體育    33.557047
財經    32.885906
Name: proportion, dtype: float64

train percentage :
artCatagory
體育    33.796740
國際    33.557047
財經    32.646213
Name: proportion, dtype: float64

test percentage :
artCatagory
國際    33.557047
財經    33.445190
體育    32.997763
Name: proportion, dtype: float64
```

原始文本

訓練集

測試集

- train_cv() ：進行CV、指標評估、混淆矩陣視覺化
- 比較四者模型的結果

now training: clf_logistic

	precision	recall	f1-score	support
國際	0.95	0.96	0.95	700
財經	0.96	0.95	0.95	681
體育	0.99	0.99	0.99	705
accuracy			0.97	2086
macro avg	0.97	0.97	0.97	2086
weighted avg	0.97	0.97	0.97	2086

now training: clf_svm

	precision	recall	f1-score	support
國際	0.95	0.96	0.95	700
財經	0.96	0.95	0.95	681
體育	0.99	0.99	0.99	705
accuracy			0.97	2086
macro avg	0.97	0.97	0.97	2086
weighted avg	0.97	0.97	0.97	2086

now training: clf_dtree

	precision	recall	f1-score	support
國際	0.85	0.89	0.87	700
財經	0.90	0.88	0.89	681
體育	0.96	0.94	0.95	705
accuracy			0.90	2086
macro avg	0.90	0.90	0.90	2086
weighted avg	0.90	0.90	0.90	2086

now training: clf_rf

	precision	recall	f1-score	support
國際	0.94	0.96	0.95	700
財經	0.97	0.94	0.95	681
體育	0.98	0.98	0.98	705
accuracy			0.96	2086
macro avg	0.96	0.96	0.96	2086
weighted avg	0.96	0.96	0.96	2086

清洗/斷詞 → **東森新聞-分類模型訓練** → 最佳模型 → 聯合新聞- 分類預測 → 洞見觀察

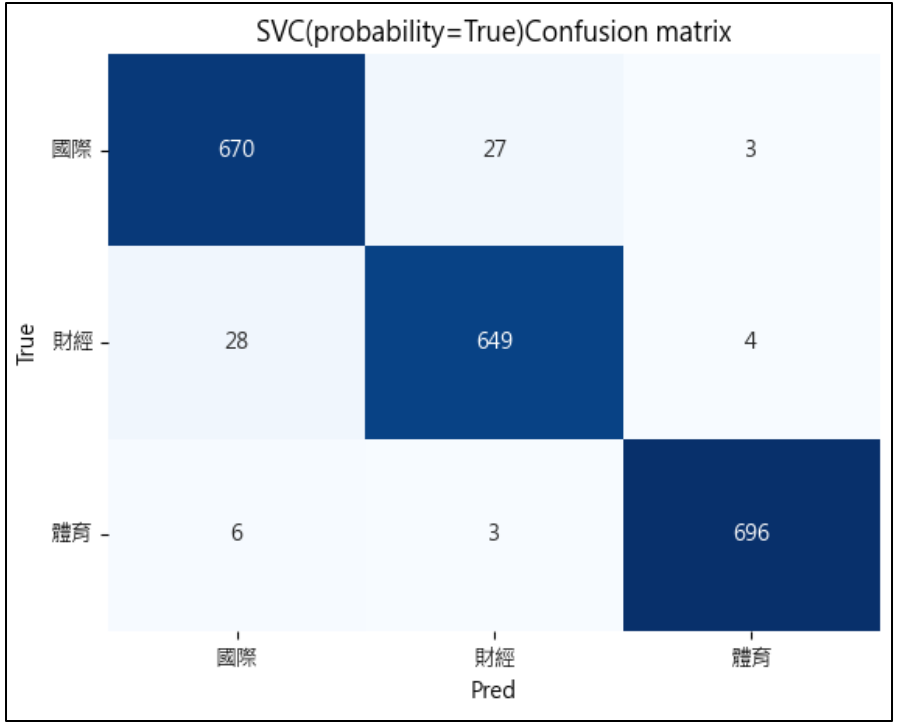
★ 找出f1-score表現最好的模型作為分類器

測試	Vectorizer	Best Model	Macro avg F1-Score
1	CountVectorizer()	RandomForestClassifier()	0.9654
2	TfidfVectorizer()	svm.SVC(probability=True)	0.9658

測試兩者差異不大，測試2效能較佳，故選用

→`TfidfVectorizer()`
→ `svm.SVC(probability=True)`

```
best model: clf_svm
```



清洗/斷詞 → 東森新聞-分類模型訓練 → 最佳模型 → 聯合新聞- 分類預測 → 洞見觀察

```
date range: ('2025-03-16 01:11:00', '2025-03-31 23:56:00')
category:
artCatagory
產經    2460
全球    2345
股市    1435
運動     874
```

- 將聯合新聞的類別給重新mapping到東森新聞分類器的類別中，接著開始進行分類任務
  - EBC東森新聞類別：體育、財經、國際
  - UDN聯合新聞類別：運動(→體育)、產經(→財經)、股市(→財經)、全球(→國際)

	precision	recall	f1-score	support
國際	0.92	0.83	0.87	2345
財經	0.90	0.96	0.93	3895
體育	0.98	0.97	0.98	874
accuracy			0.92	7114
macro avg	0.94	0.92	0.93	7114
weighted avg	0.92	0.92	0.92	7114

●測試資料：聯合新聞網(UDN)
版別：運動、產經、股市、全球
區間：2025/03/15-2025/03/31

●訓練分類器由東森新聞資料集EBC訓練，在聯合新聞UDN預測分類表現還不錯

●「國際」類別表現較差，F1-Score僅0.87

●進一步研究分類的結果，找出問題的原因

清洗/斷詞 → 東森新聞-分類模型訓練 → 最佳模型 → 聯合新聞- 分類預測 → 洞見觀察

```
# 將錯誤分類的結果篩選出來
false_pred = ct.query("artCatagory != pred").loc[:,['words', 'artCatagory',"pred"]]
false_pred
```

	words	artCatagory	pred
22	金融時報 投資人 看好 川普 放寬 制裁 尋求 獲利 投資人 看好 美國 總統 川普對 俄羅...	國際	財經
189	地緣 通訊 風險 台灣 聯外 通訊 命懸 海底 電纜 地緣 政治 風險 上升 中國 大陸 漁...	國際	財經
195	有能 避開 川普 關稅 貿易戰 逃生 交易 基金 經理 這裡 總部 設在 倫敦 環球 資產 ...	國際	財經
197	OECD 下修 全球 經濟 展望 普調 關稅 減弱 美墨加 經濟 成長 路透 報導 經濟 合...	國際	財經
199	首付款 蒸發 川普 關稅 拖累 奄奄一息 美國 房市 周刊 Newsweek 17 分析 川...	國際	財經
...	...	...	...
6881	世界 迎去 中化 轉單 地緣 政治 轉單 題材 延燒 世界 先進 5347 24 股價 早盤...	財經	國際
6909	精剛法 受惠 關稅 樂觀 前景 特殊鋼 材料廠 精剛 1584 26 日法 表明 美國 關稅...	財經	國際
6910	精剛 轉單 營運 特殊鋼 材料廠 精剛 1584 26 日法 表明 美國 關稅 中國 大陸 ...	財經	國際
6924	台灣 穩鴻 31 登錄 創櫃板 再添 生力軍 櫃買 中心 Taipei Exchange 創...	財經	國際
6949	野獸 國財報 2024 每股 1.65 擬配息 0.14 野獸 6995 28 發布 財報 ...	財經	國際

580 rows × 3 columns

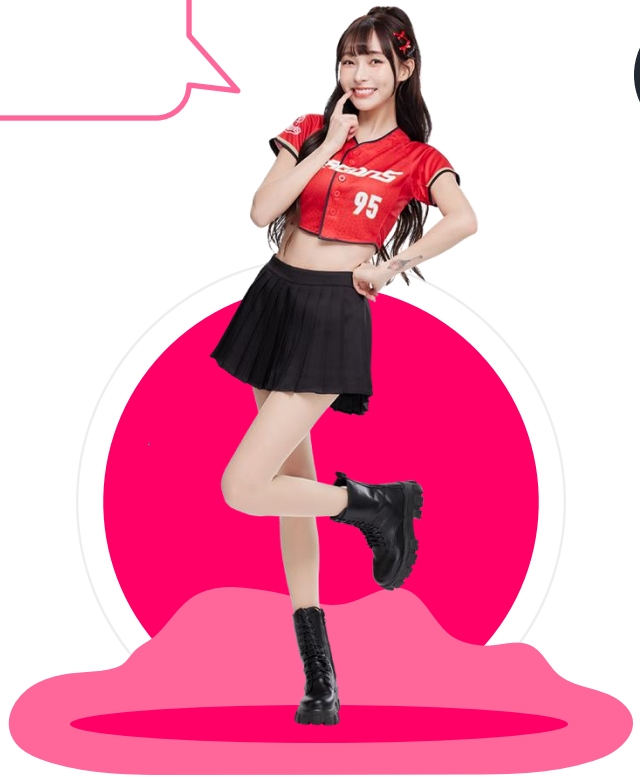
●聯合新聞網(UDN)版別：
運動(→體育)、產經(→財經)
股市(→財經)、全球(→國際)

●從預測結果來看，可以發現聯合新聞UDN的「全球」(→國際)類別，也常常會報導“國外財經”的資訊，常常出現財經用詞基金、市場、股票等字，使模型預測成「財經」；「財經」類別也多出現多國國家名，使模型預測為「國際」。

●整體分類器預測聯合新聞網UDN類別，F1-Score結果「國際」-0.87 /「財經」-0.93 /「體育」-0.98 表現結果都算相當不錯！

清洗/斷詞 → 東森新聞-分類模型訓練 → 最佳模型 → 聯合新聞- 分類預測 → 洞見觀察

請補充



詞彙關係

文件分類

主題模型

訓練 topic model(LDA)

```python
#建立模型
ldamodel = LdaModel(
    corpus=corpus,
    id2word=dictionary, # 字典
    num_topics=9, # 生成幾個主題數
    random_state=2024, # 亂數
)
```

✓ 1.4s Python

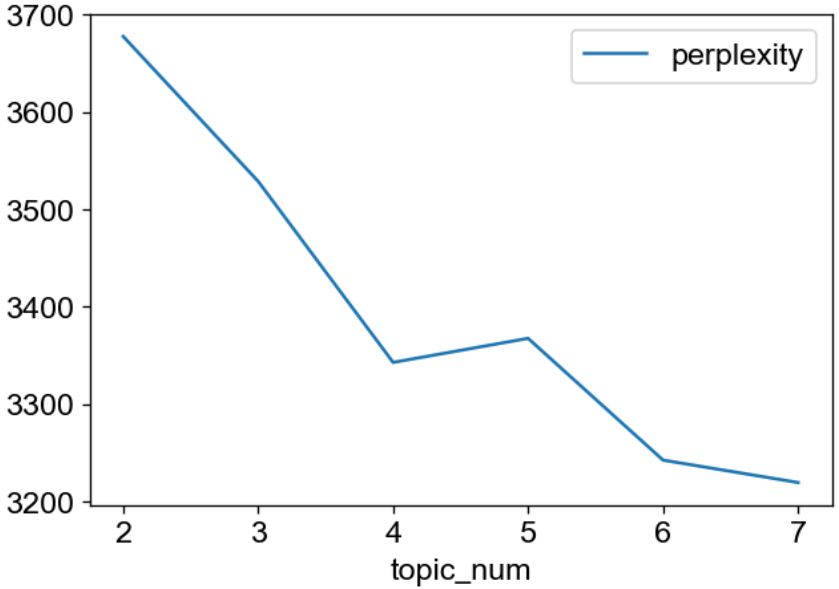
●從2980篇文檔中自動識別9個主題

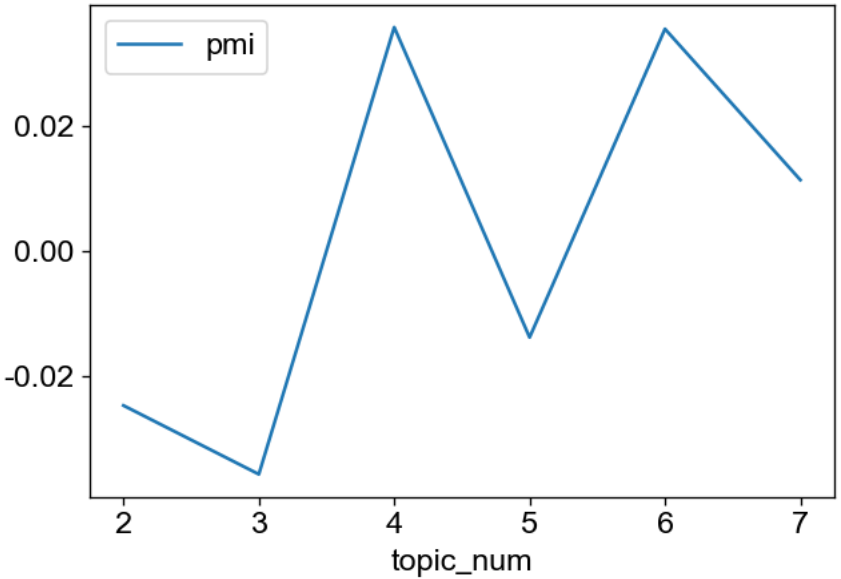
	# topic_num	# perplexity	# pmi
0	2	3677.4630647671506	-0.024664118542167105
1	3	3528.586055236737	-0.035639079011401543
2	4	3342.589110813794	0.03572894812202629
3	5	3367.3054289198853	-0.013789040750584496
4	6	3242.1888745592446	0.03545781794230064
5	7	3219.2034280537646	0.011344763561994809

透過GENSIM進行LDA訓練找出最佳主題數

PERPLEXITY：隨著主題數量增加，困惑度整體呈現下降趨勢
1.在主題數為7時達到最低值(約3200)
2.主題數從2增加到4時有最明顯的改善(從3700降至3350)，從4到7的改善相對緩慢

PMI：PMI值在主題數為4和6時達到峰值
主題數為4時達到最高點





故以下BEST_MODEL建立將選擇4作為NUM_TOPICS挑選：

- 主題數為4時PMI達到最高
- 雖然PERPLEXITY在主題數為7時最低，但從4到7的改善相對有限
- 主題數4提供了困惑度和一致性之間的最佳平衡點
- 考慮到模型簡潔性和可解釋性，選擇較少的主題數

Selected Topic: 3 Previous Topic Next Topic Clear Topic

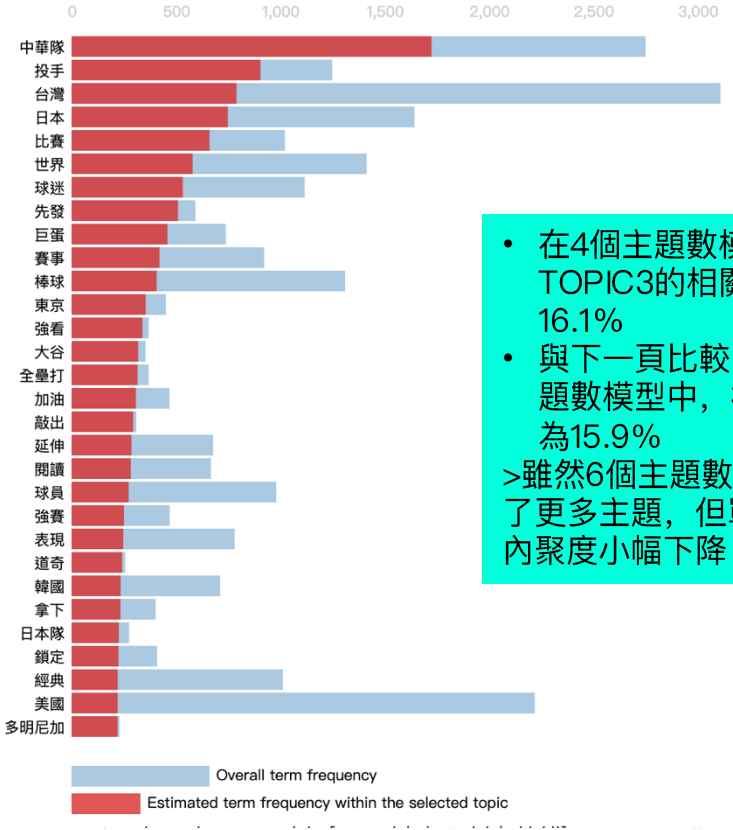
Slide to adjust relevance metric:(2)

λ = 1

0.0 0.2 0.4 0.6 0.8

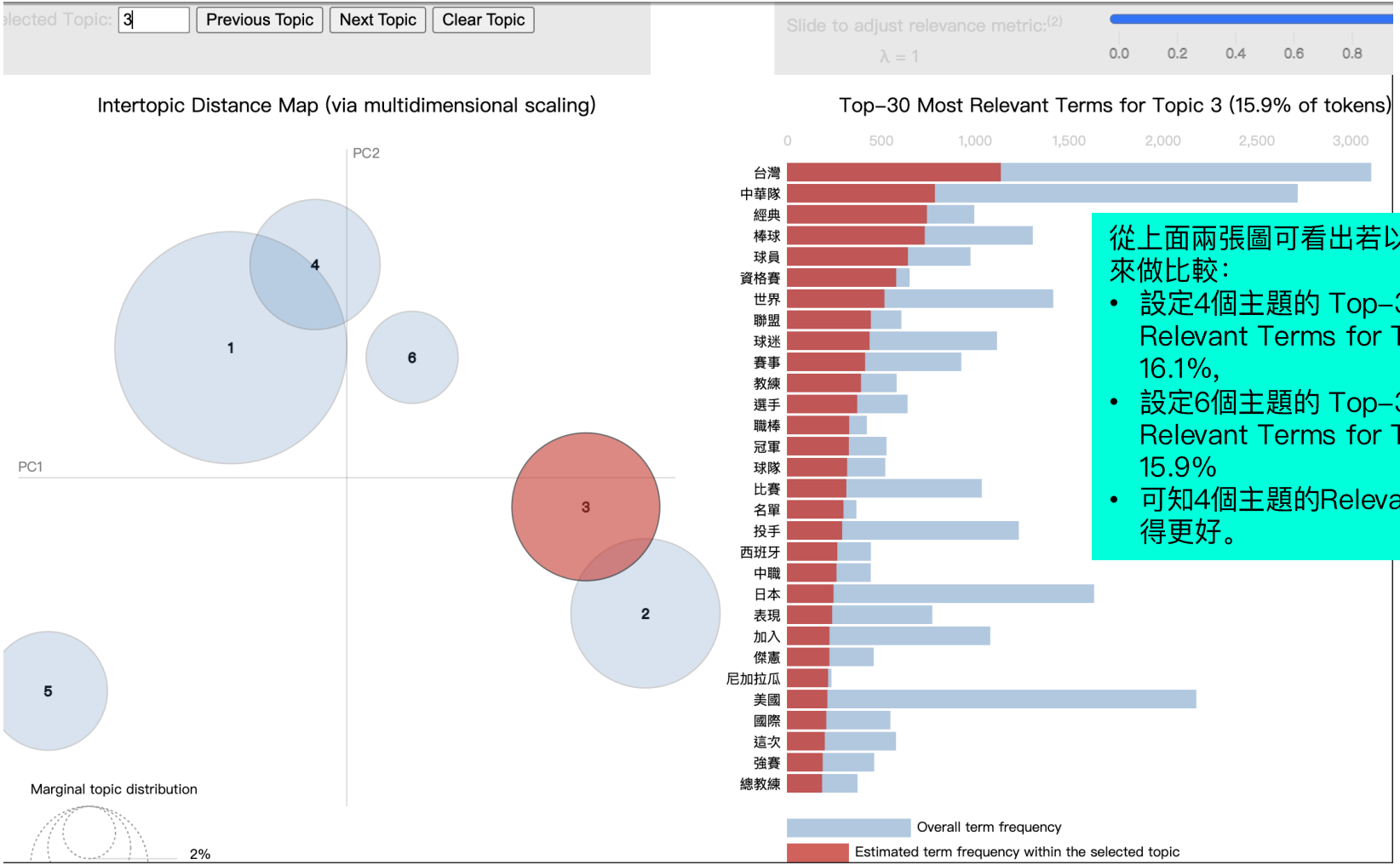
Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 3 (16.1% of tokens)



- 在4個主題數模型中TOPIC3的相關詞佔比是16.1%
- 與下一頁比較，在6個主題數模型中，相關詞佔比為15.9%

>雖然6個主題數的模型提供了更多主題，但單一主題的內聚度小幅下降



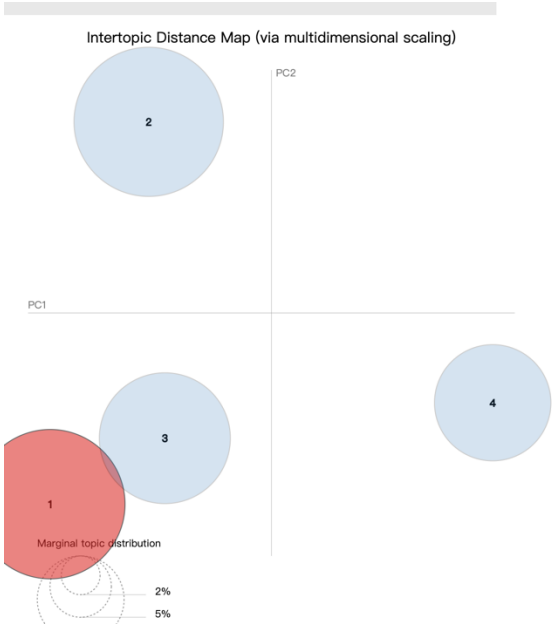
從上面兩張圖可看出若以Topic 3為例來做比較：

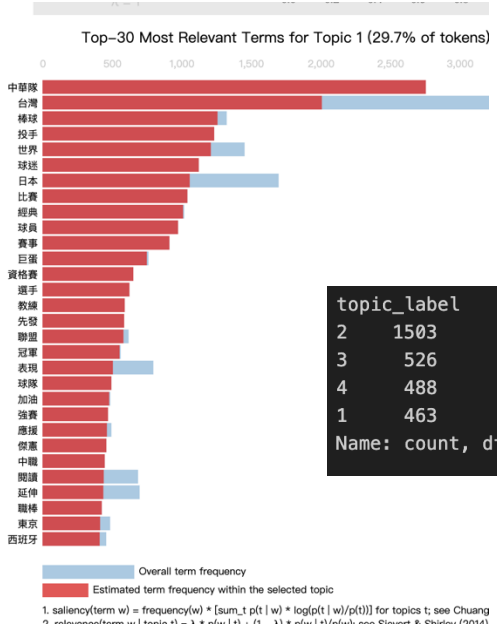
- 設定4個主題的 Top-30 Most Relevant Terms for Topic 3 為 16.1%,
- 設定6個主題的 Top-30 Most Relevant Terms for Topic 3 為 15.9%
- 可知4個主題的Relevant Terms來得更好。

GuidedLDA

自行設定種子詞列表 ＞達到話題半監督的目的

```python
# 自行設定種子詞列表
seed_topic_list = [
    # 主題1：體育新聞（棒球相關）
    ["投手", "中華隊", "棒球", "道奇", "球員", "先發", "球迷", "敲出", "世界", "經典"],
    # 主題2：國際政治新聞
    ["美國", "中國", "總統", "政府", "川普", "戰爭", "國際", "俄羅斯", "關係", "外交"],
    # 主題3：財經新聞
    ["台股", "漲幅", "股價", "市場", "台積", "台積電", "投資", "基金", "金融", "經濟"],
    # 主題4：台灣民生新聞
    ["台灣", "民眾", "生活", "事件", "社會", "調查", "問題", "活動", "服務", "教育"]
]
```

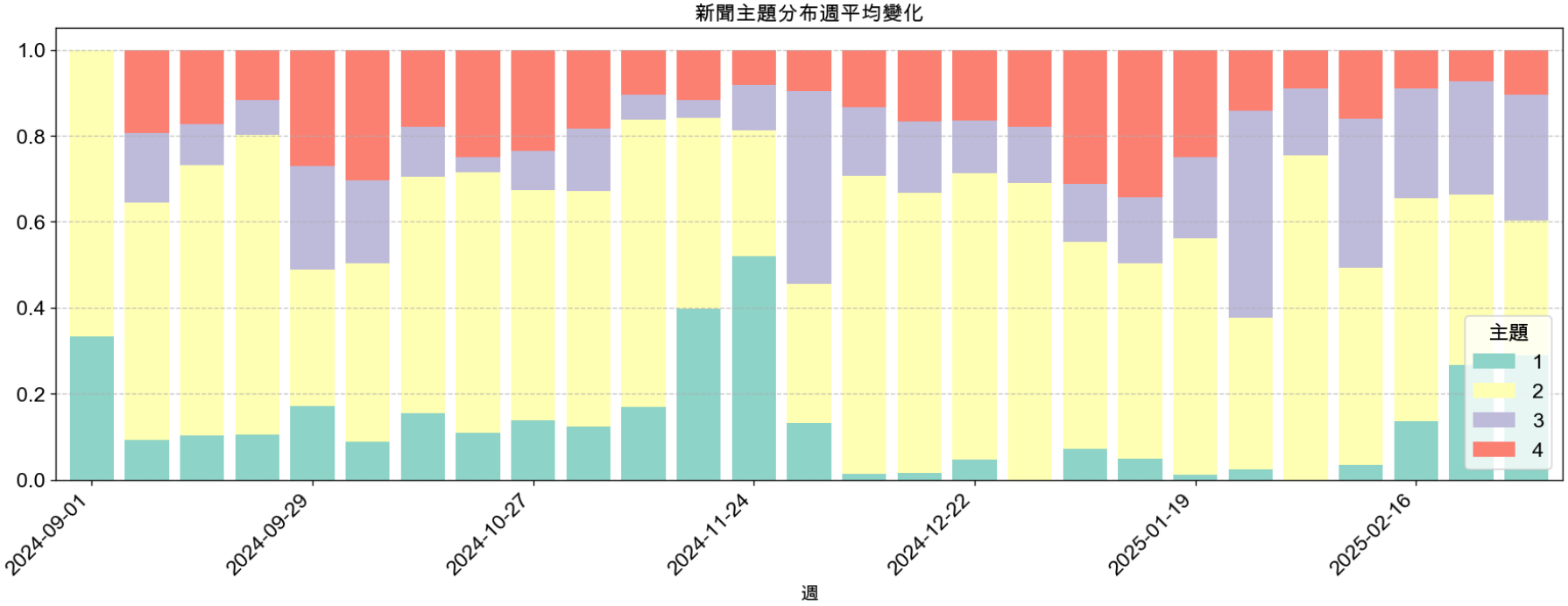




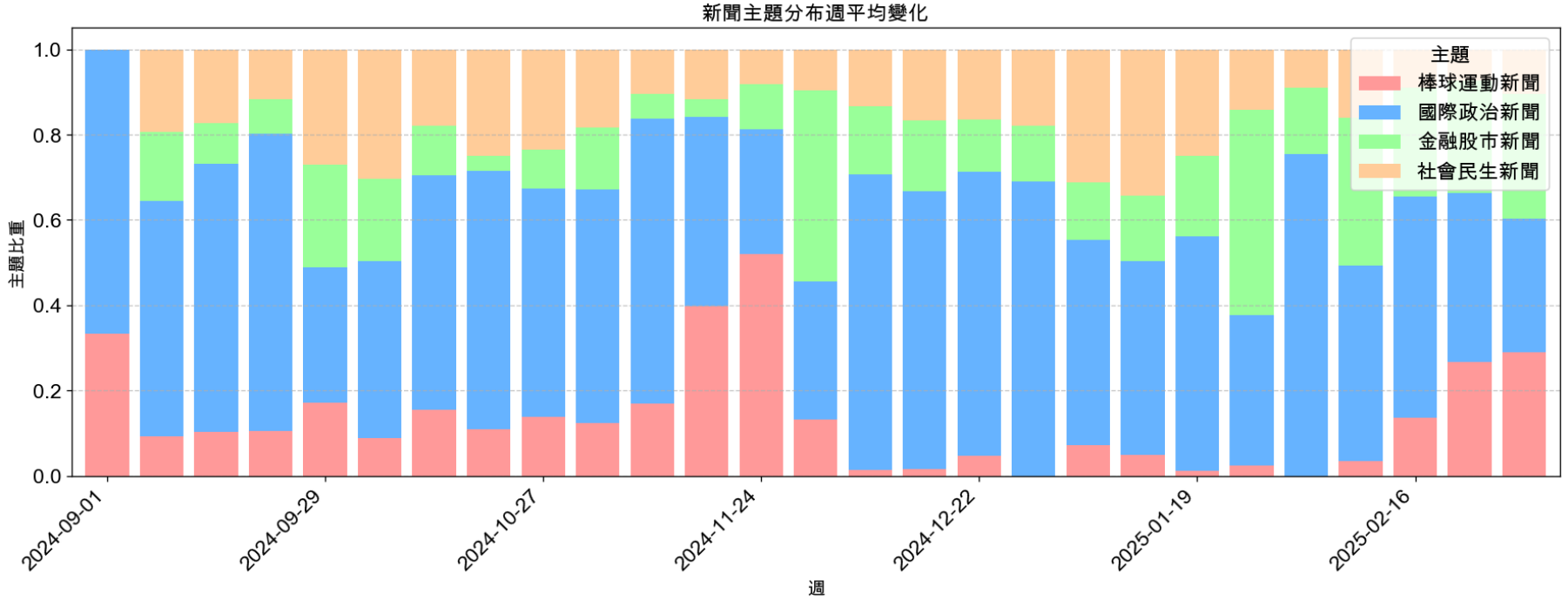

```
topic_label
2    1503
3     526
4     488
1     463
Name: count, dtype: int64
```

主題標籤的解釋

- 主題2：國際政治新聞
- 主題3：金融股市
- 主題4：社會新聞
- 主題1：體育賽事（棒球）



- 主題1（綠色）在特定時間點有明顯的比例增加，特別是11月中旬的顯著上升，很大可能是因為中華隊參與世界棒球錦標賽時，吸引媒體關注度大幅提升。



- 2024年11月中下旬有明顯的棒球新聞增加
- 國際政治新聞（藍色）在大部分時間維持較高比例，尤其在2024年9月初及12月
- 金融股市新聞（綠色）在2025年1月底至2月增加
- 社會民生新聞（橘色）作為基礎類別始終存在