

SMA TEAM 05

M134020003 陳亭聿、M134020019 黃亮慈
M134020027 陳昱璋、M134020044 李晉豪
B104020024 蔡尚宸、B104020037 董昱辰、M133040018 鄭惠心

INDEX



Week 6
文辭和文件分析



Week 7
文件分類



Week 9
主題分析



Week 6

文辭和文件分析

WEEK06

基本介紹

資料集

使用 Tarflow 平台抓取。

- 全部看板
- 關鍵字：麥當勞
- 時間：2024/08/31 - 2025/03/13
- 資料筆數：共 2036 筆

基本處理 與 結果觀察

資料清理、斷詞、自訂辭典、停用字詞典

Bigram, Trigram, 視覺化, 與特定詞相關性最高前10等 (後續詳細說明)



WEEK06

BIGRAM 視覺化

透過 N-gram 幫助建立斷詞字典：

斷詞字典建立原則：

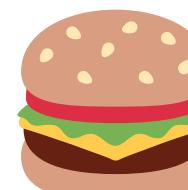
- 基本上皆是麥當勞的餐點名稱
 - 多為對同個商品的略縮語、代稱

例子：



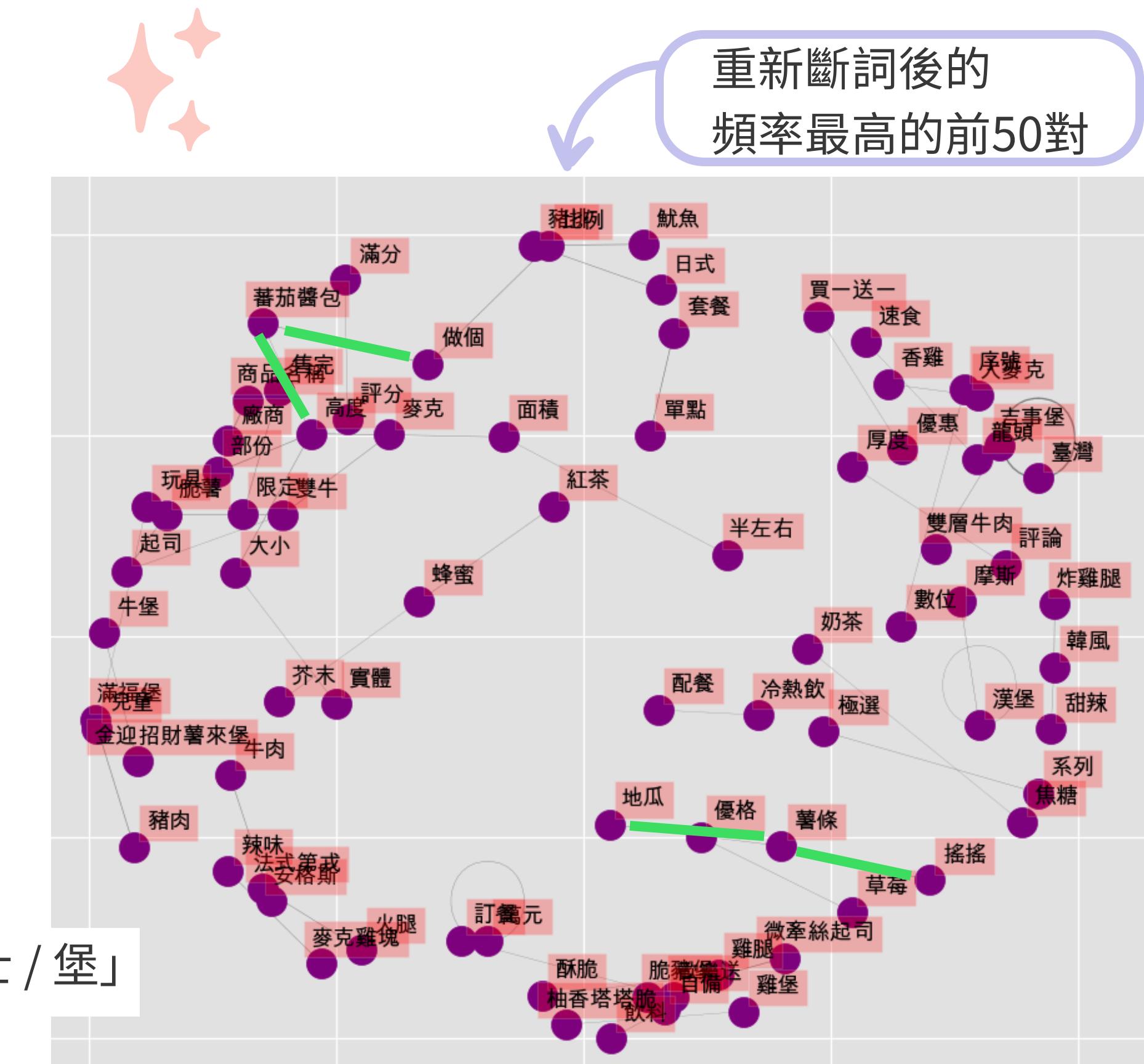
「微牽絲起司排雙牛脆豬堡」

- 會被亂切一堆奇怪的詞



「雙層牛肉吉士堡」

- 會被錯切為「雙層 / 牛肉 / 吉士 / 堡」



WEEK06

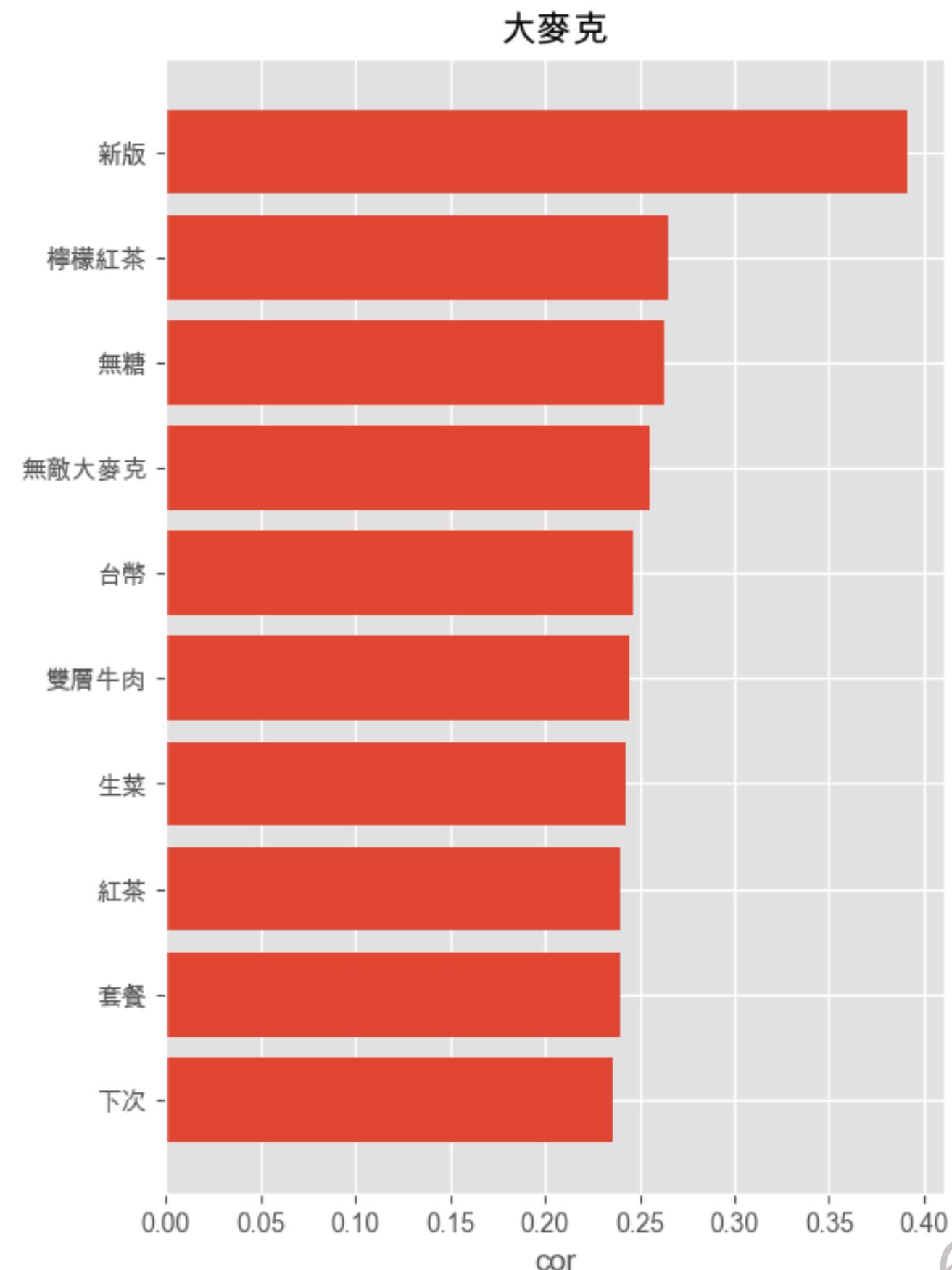
PAIRWISE CORRELATION

接著計算兩個詞彙之間的相關性

挑選關鍵詞『大麥克』，與各關鍵詞之間的相關程度

大麥克：

- 「新版」 - 2024年中後大麥克有進行改版，許多網友在討論新舊版大麥克之間的差異
- 飲品搭配關聯高，可能常將飲品與主餐一起討論
- 「雙層牛肉」、「無敵大麥克」可能是與相似的品項作討論



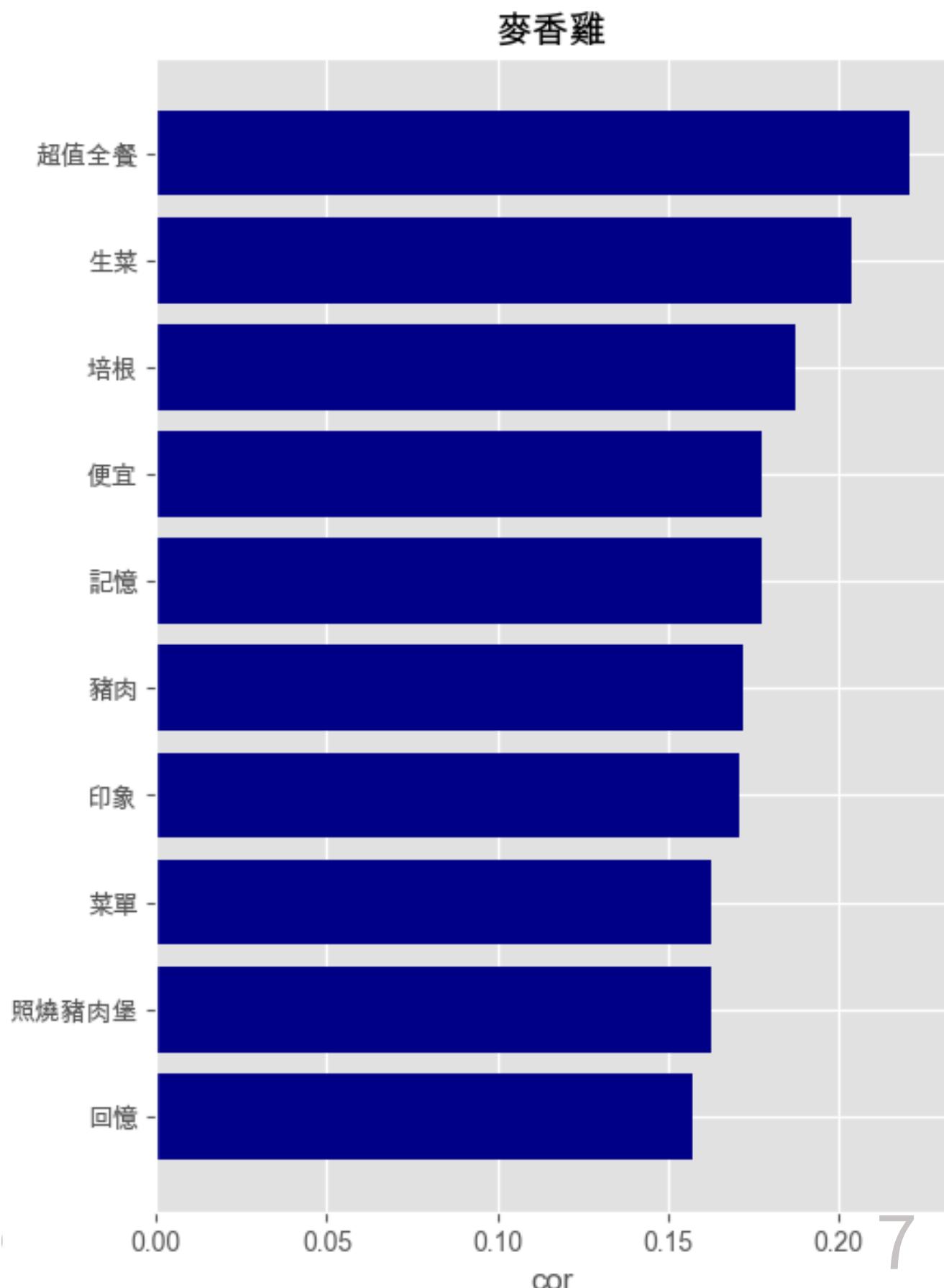
WEEK06

PAIRWISE CORRELATION

挑選關鍵詞『麥香雞』，與各關鍵詞之間的相關程度

麥香雞：

- 「超值全餐」出現頻率最高，可能顯示出消費者習慣
- 「便宜」、「記憶」、「印象」、「回憶」，這種傾向於情感，
麥香雞可能具有「童年記憶」或「經典口味」的地位

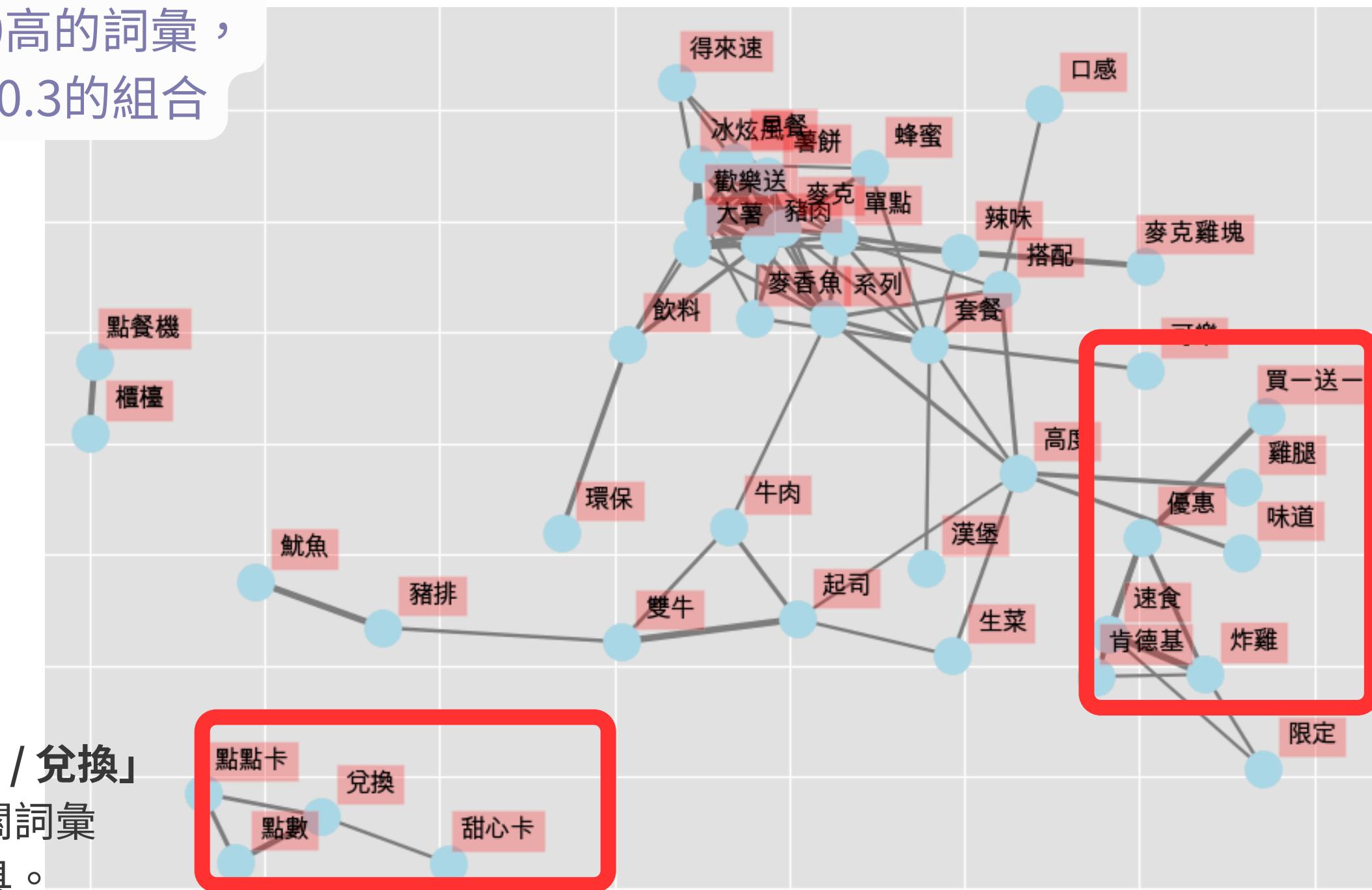


WEEK06

PAIRWISE CORRELATION

透過DTM找出 詞頻 前60高的詞彙，
並篩選出 相關係數 大於0.3的組合

右方「炸雞 / 肯德基 / 速食 / 限定 / 買一送一」形成一群
可能對應網友討論促銷與炸物商品的主題



下方「點點卡 / 甜心卡 / 點數 / 兌換」
也是一群明確的優惠行為相關詞彙
顯示出使用者在乎的優惠工具。



Week 7

文件分類

WEEK07

介紹

訓練集: ptt-韓劇版、歐美影集版、玄幻版、台劇&陸劇
(2024-10-05~2025-04-05)

測試集: ptt-韓劇版、歐美影集版、玄幻版、台劇&陸劇
(2024-04-05~2024-10-05)

Drama Dictionary: 同時爬取每篇文章專有名詞
(ex: 劇名、作品名)，以確保模型運作更順利

	Train (4720)	Test (5241)
KoreaDrama	1346 (28.5%)	1377 (26.3%)
CFantasy	1217 (25.8%)	1321 (25.2%)
EAseries	1180 (25%)	1405 (26.8%)
TWCN_Drama	977 (20.7%)	1138 (21.7%)



WEEK07

訓練模型

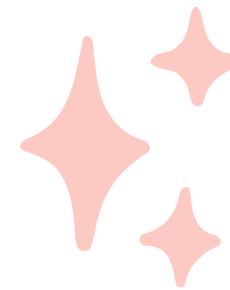


根據7:3的比例切分資料集，並使用LogisticRegression()進行訓練

		Confusion matrix			
		CFantasy	EAseries	KoreaDrama	TWCN_Drama
True	CFantasy	351	7	1	8
	EAseries	10	315	14	15
KoreaDrama	5	13	341	25	
TWCN_Drama	10	18	25	258	
		CFantasy	EAseries	KoreaDrama	TWCN_Drama
	Pred				

	precision	recall	f1-score	support
CFantasy	0.93	0.96	0.94	367
EAseries	0.89	0.89	0.89	354
KoreaDrama	0.90	0.89	0.89	384
TWCN_Drama	0.84	0.83	0.84	311
accuracy			0.89	1416
macro avg	0.89	0.89	0.89	1416
weighted avg	0.89	0.89	0.89	1416

訓練模型：使用TF-IDF



	precision	recall	f1-score	support
CFantasy	0.94	0.96	0.95	367
EAseries	0.91	0.92	0.91	354
KoreaDrama	0.88	0.91	0.90	384
TWCN_Drama	0.90	0.83	0.86	311
accuracy			0.91	1416
macro avg	0.91	0.90	0.90	1416
weighted avg	0.91	0.91	0.91	1416

Accuracy稍微上升，可能是因為各個版的用字遣詞在TF-IDF訓練下，更能顯示出差異。

WEEK07

訓練模型

使用LogisticRegression()進行訓練

LogisticRegression()Confusion matrix

		LogisticRegression()Confusion matrix			
		CFantasy	EASeries	KoreaDrama	TWCN_Drama
True	CFantasy	820	13	7	10
	EASeries	30	735	30	31
	KoreaDrama	12	50	852	48
	TWCN_Drama	28	38	47	553

	precision	recall	f1-score	support
CFantasy	0.92	0.96	0.94	850
EASeries	0.88	0.89	0.88	826
KoreaDrama	0.91	0.89	0.90	962
TWCN_Drama	0.86	0.83	0.85	666
accuracy			0.90	3304
macro avg	0.89	0.89	0.89	3304
weighted avg	0.90	0.90	0.90	3304

WEEK07

訓練模型

DecisionTree()進行訓練

DecisionTreeClassifier()Confusion matrix

		Pred			
		CFantasy	EASeries	KoreaDrama	TWCN_Drama
True	CFantasy	762	38	12	38
	EASeries	42	633	86	65
	KoreaDrama	10	76	777	99
	TWCN_Drama	40	77	86	463

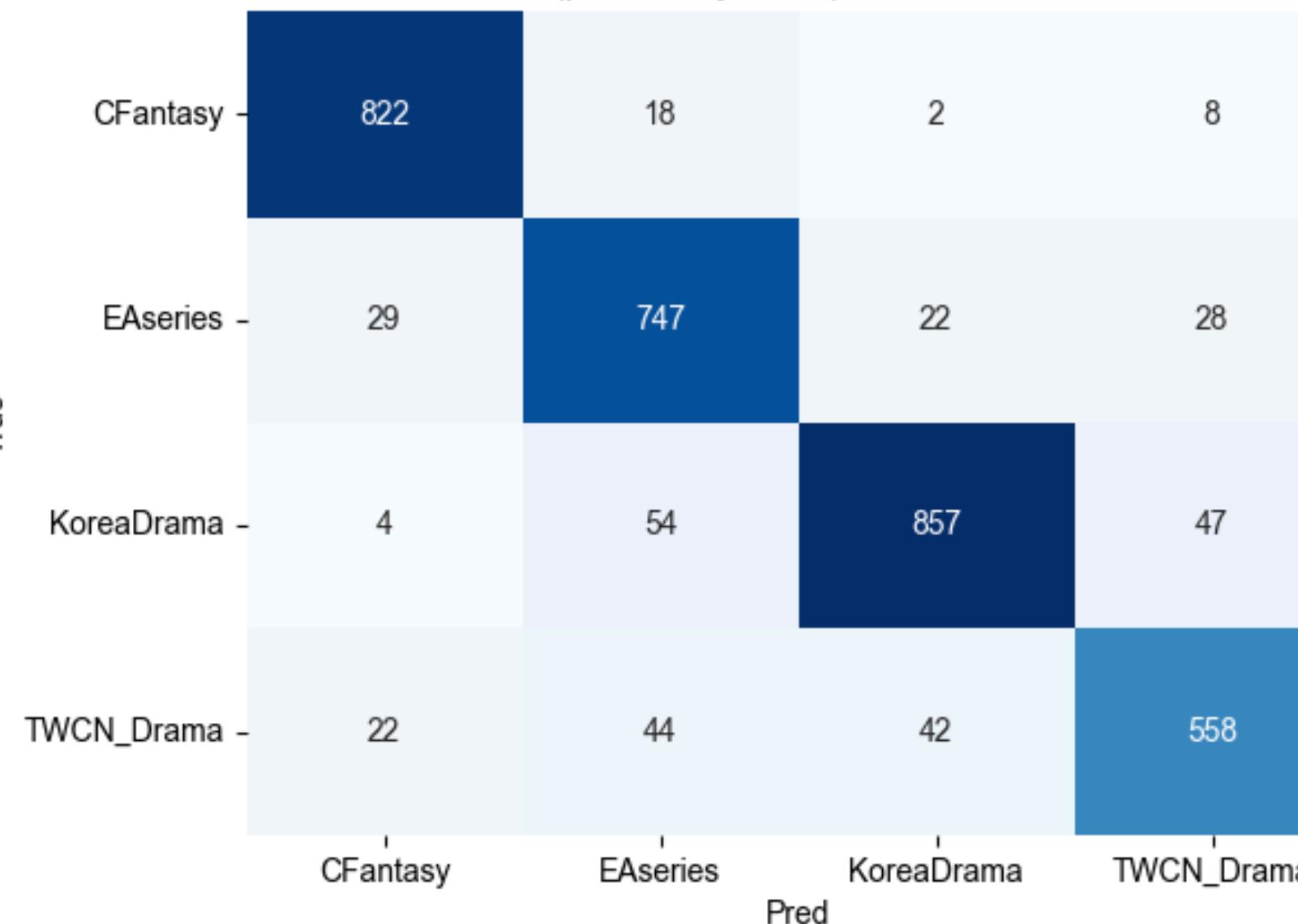
	precision	recall	f1-score	support
CFantasy	0.89	0.90	0.89	850
EASeries	0.77	0.77	0.77	826
KoreaDrama	0.81	0.81	0.81	962
TWCN_Drama	0.70	0.70	0.70	666
accuracy			0.80	3304
macro avg	0.79	0.79	0.79	3304
weighted avg	0.80	0.80	0.80	3304

WEEK07

訓練模型

使用SVM()進行訓練

SVC(probability=True)Confusion matrix



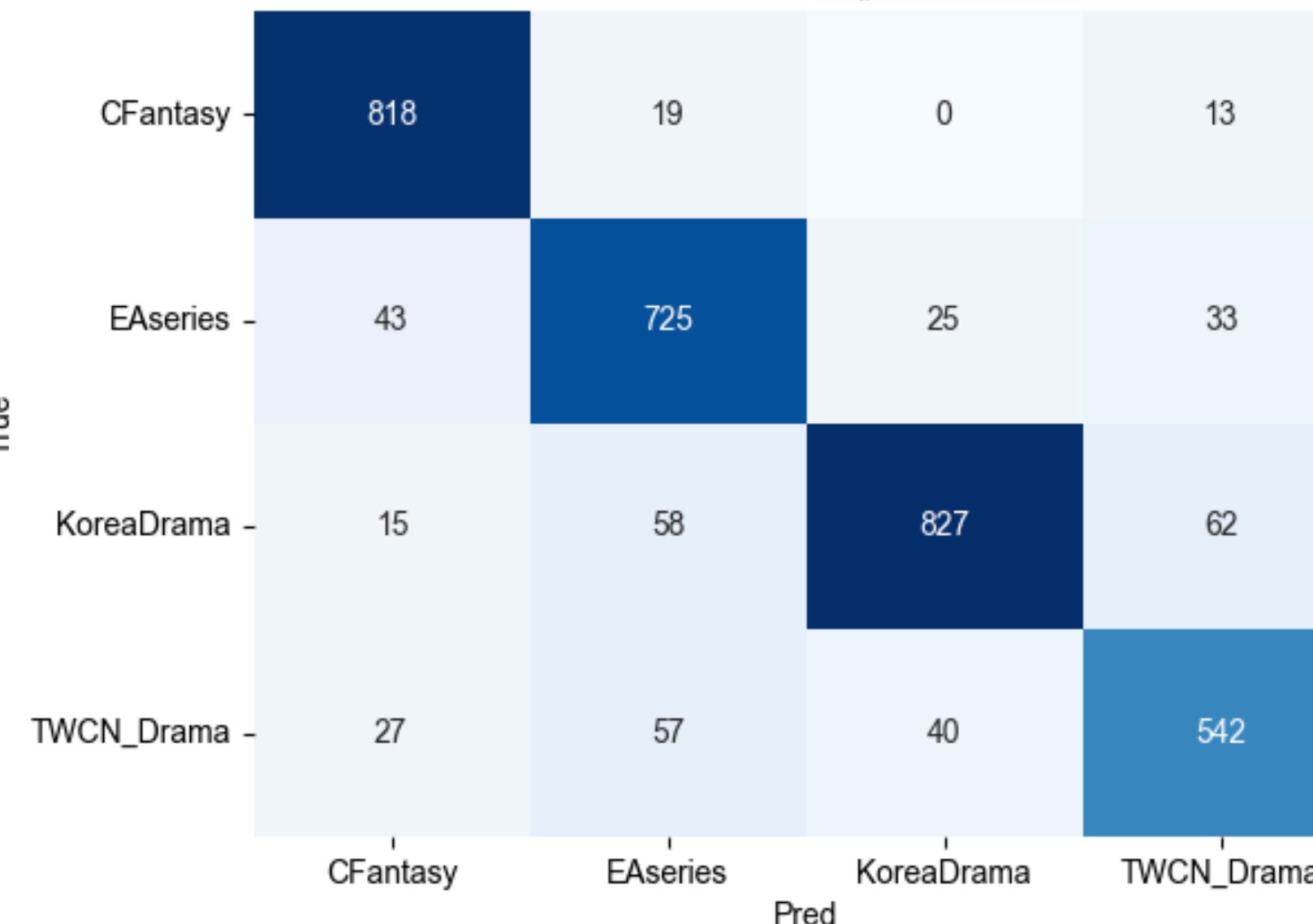
	precision	recall	f1-score	support
CFantasy	0.94	0.97	0.95	850
EAseries	0.87	0.90	0.88	826
KoreaDrama	0.93	0.89	0.91	962
TWCN_Drama	0.87	0.84	0.85	666
accuracy			0.90	3304
macro avg	0.90	0.90	0.90	3304
weighted avg	0.90	0.90	0.90	3304

WEEK07

訓練模型

使用Random Forest訓練

RandomForestClassifier()Confusion matrix



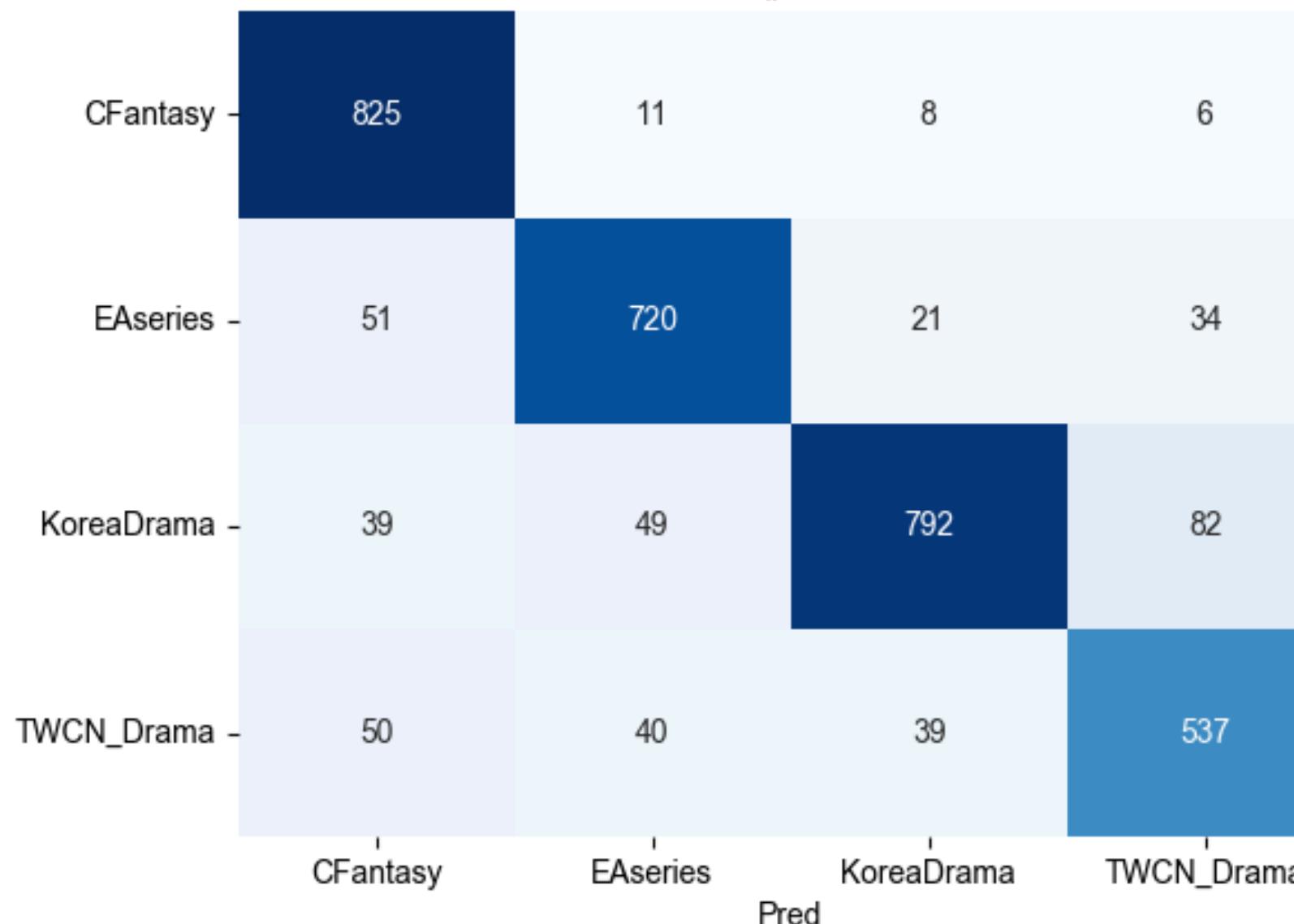
	precision	recall	f1-score	support
CFantasy	0.91	0.96	0.93	850
EAseries	0.84	0.88	0.86	826
KoreaDrama	0.93	0.86	0.89	962
TWCN_Drama	0.83	0.81	0.82	666
accuracy			0.88	3304
macro avg	0.88	0.88	0.88	3304
weighted avg	0.88	0.88	0.88	3304

WEEK07

訓練模型

使用MultinomialNB()進行訓練

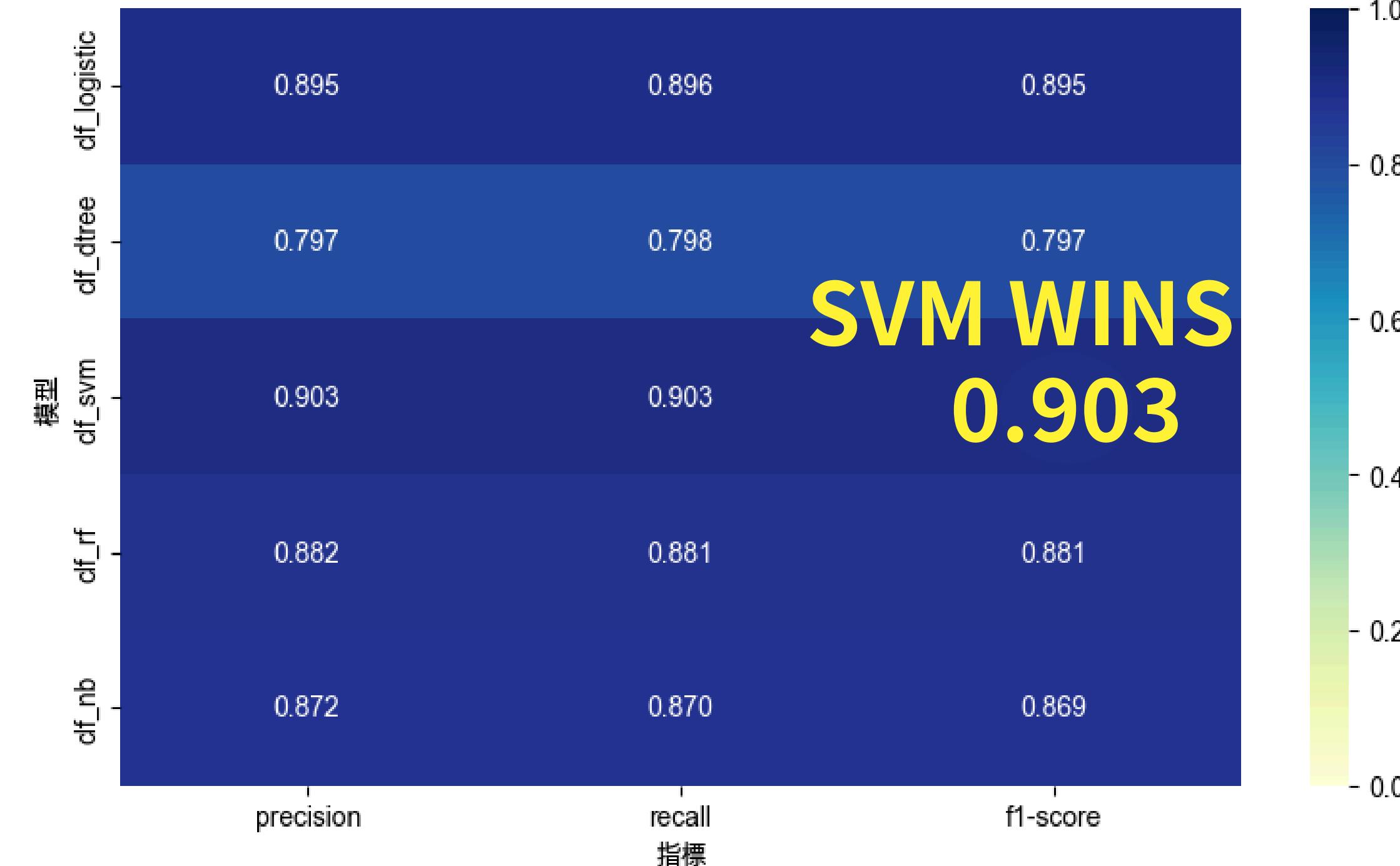
MultinomialNB()Confusion matrix



	precision	recall	f1-score	support
CFantasy	0.85	0.97	0.91	850
EAseries	0.88	0.87	0.87	826
KoreaDrama	0.92	0.82	0.87	962
TWCN_Drama	0.81	0.81	0.81	666
accuracy			0.87	3304
macro avg	0.87	0.87	0.87	3304
weighted avg	0.87	0.87	0.87	3304

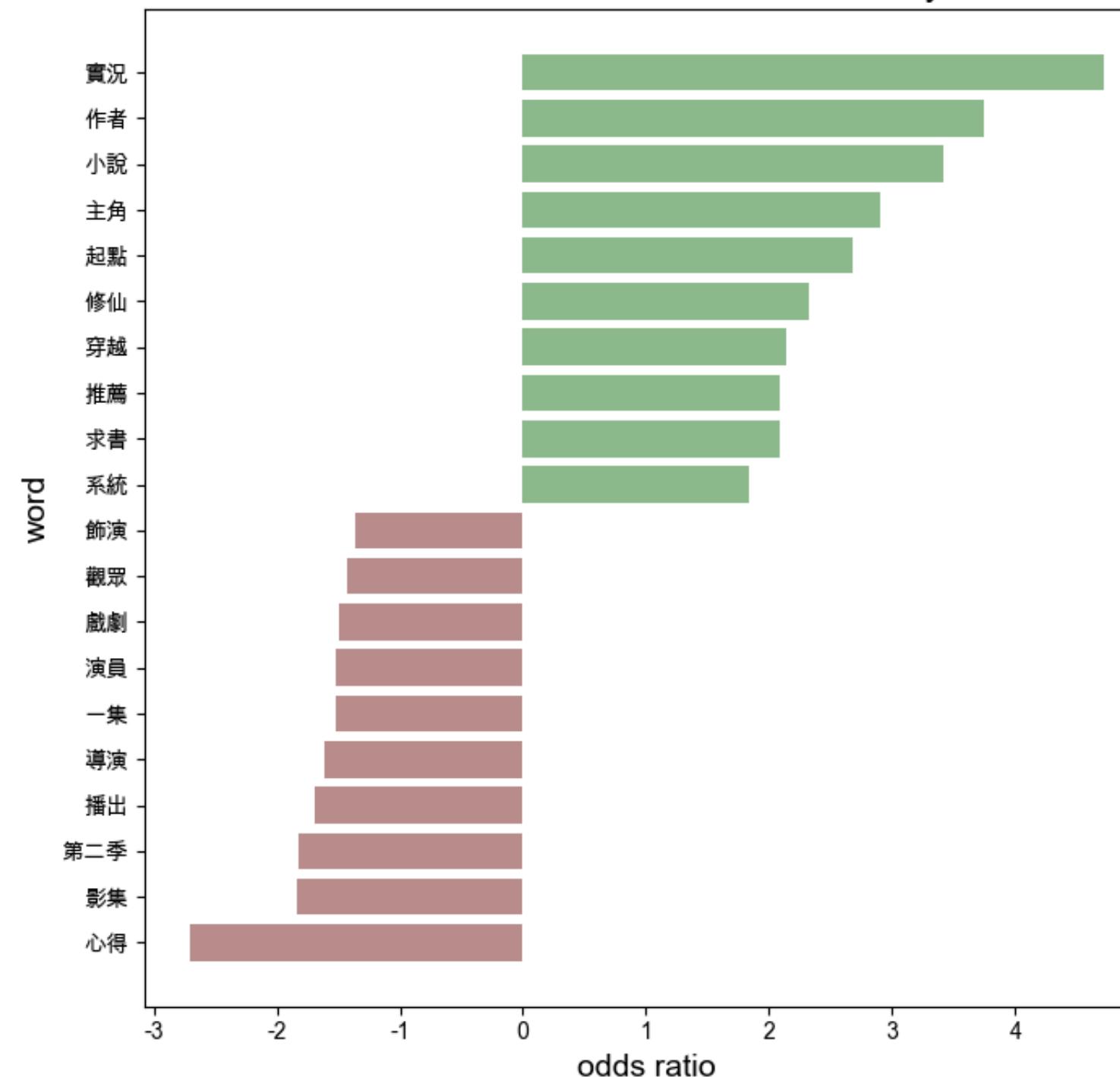
模型比較

模型 weighted avg 分類指標熱力圖

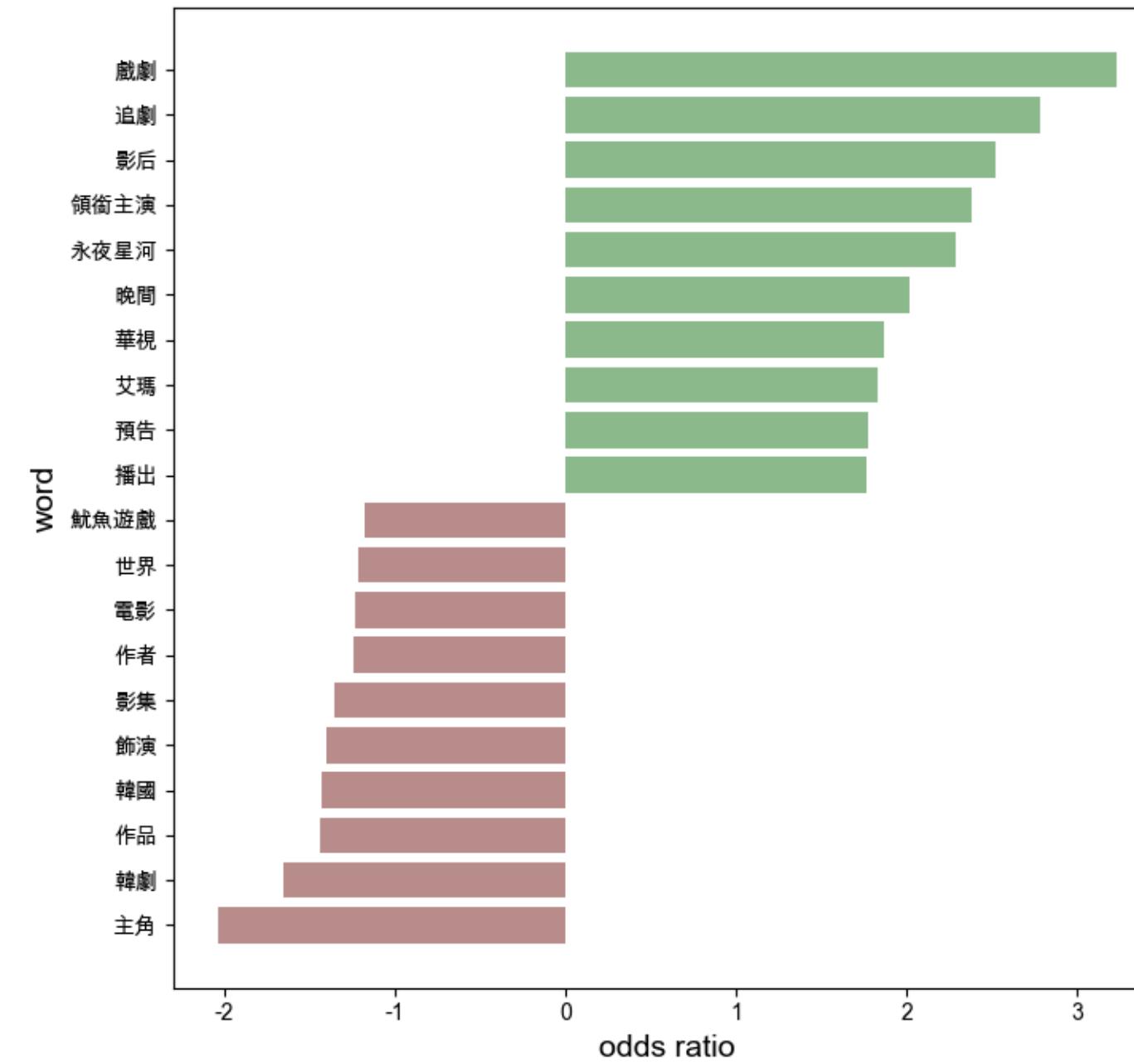


各版字詞

Coeff increase/decrease odds ratio of 「CFantasy」 label the most



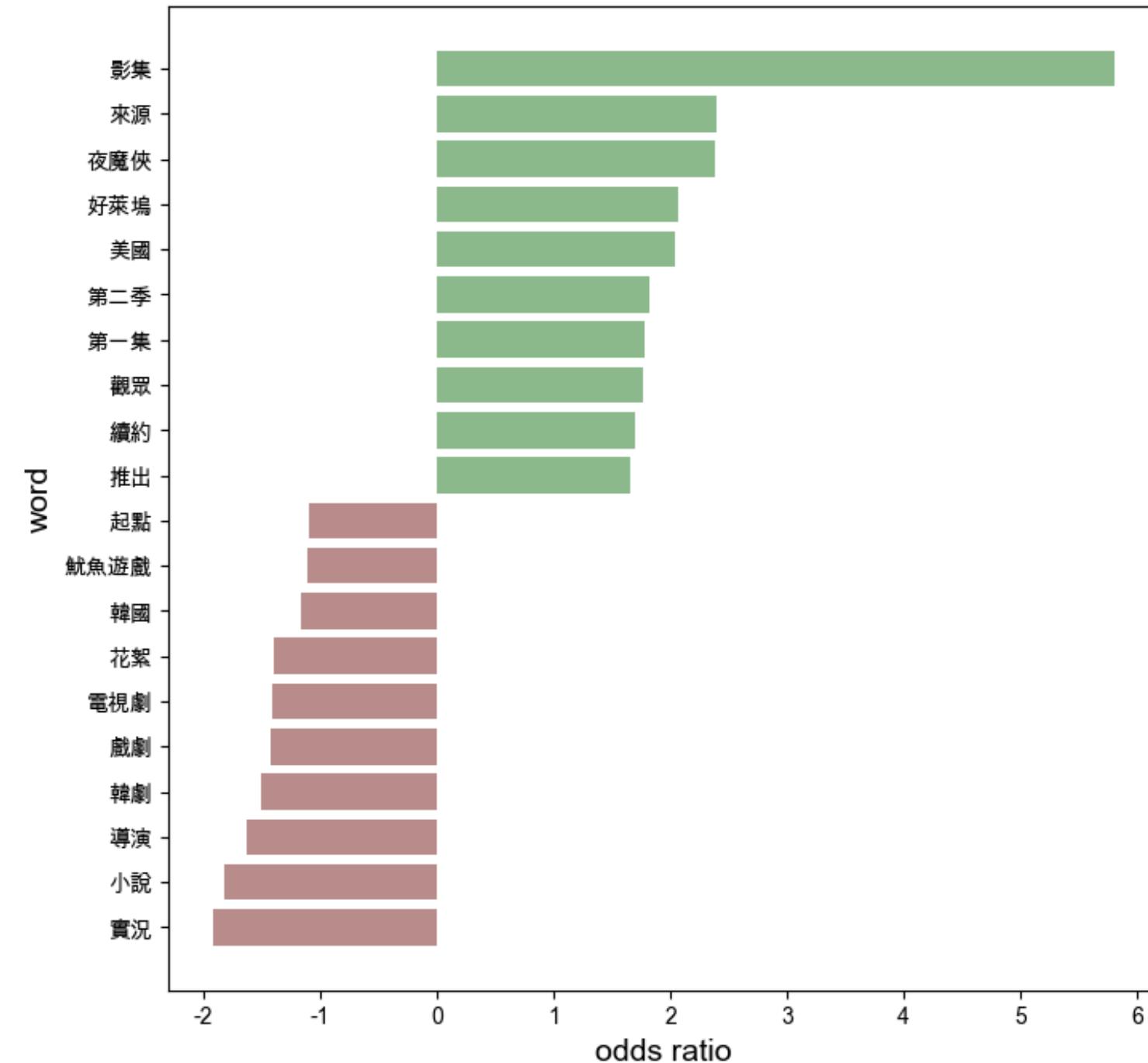
Coeff increase/decrease odds ratio of 「TWCN_Drama」 label the most



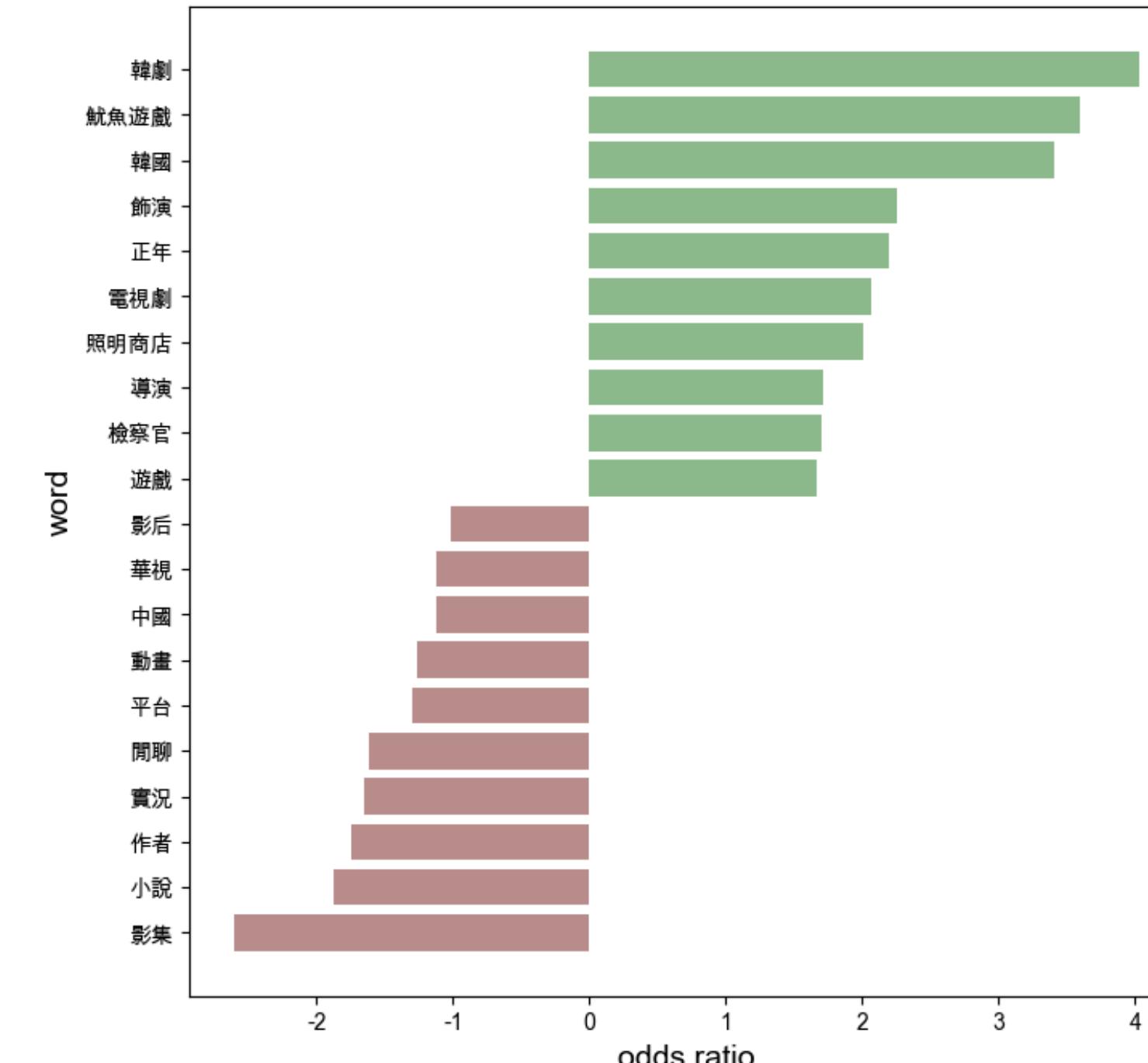
WEEK07

各版字詞

Coeff increase/decrease odds ratio of 「EAseries」 label the most



Coeff increase/decrease odds ratio of 「KoreaDrama」 label the most



模型測試

	precision	recall	f1-score	support
CFantasy	0.89	0.95	0.92	1321
EAseries	0.81	0.87	0.84	1405
KoreaDrama	0.80	0.80	0.80	1377
TWCN_Drama	0.74	0.60	0.67	1138
accuracy			0.81	5241
macro avg	0.81	0.81	0.81	5241
weighted avg	0.81	0.81	0.81	5241

使用表現最佳的 SVM 模型對測試集進行預測，結果顯示「玄幻」分類的準確度最高，推測因其文章中常出現具辨識度的專用詞彙；相對地，「台陸劇」的預測較弱，可能因為其用語較為籠統、類別邊界模糊所致。

WEEK07



```
false_pred.loc[false_pred['artCatagory']=='KoreaDrama', :].head(20)
```

		words	artCatagory	pred
2088	心得 淚之女王 一股 煩躁 如題 老婆 淚之女王 推說 好看 感人 第三集 之有 滿滿的 尴...		KoreaDrama	TWCN_Drama
2094	篡位 一集 配樂 主題曲 配樂 三首 主題曲 一首 搜尋 一首 開頭 謝謝		KoreaDrama	EASeries
2098	閒聊 七人的復活 順玉繼 上流 戰爭 成功 去年 推出 七人 逃脫 原本 狗血 保證 沒想到...		KoreaDrama	EASeries
2099	閒聊 殺人者的購物中心 全劇 李棟旭 頹喪 叔叔 模樣 圈粉 男主死 穿插 回憶 女主還 武...		KoreaDrama	TWCN_Drama
2106	心得 寄生獸 普雷日 韓兩版 防雷 一二三四五六七八九十 開門見山 喜歡 日版 原因 循序漸...		KoreaDrama	CFantasy
2107	心得 網飛 寄生獸 灰色 部隊 第一季 先說 結論 看過 寄生獸 相關 作品 影集 精彩 網...		KoreaDrama	EASeries
2112	寄生獸 灰色 部隊 第集 有雷 初次 發文 格式 怪怪 請見 原作 書粉 原作 設定 記沒 ...		KoreaDrama	TWCN_Drama
2113	心得 寄生獸 觀後 心得 大部份 韓版 缺點 分享 心得 結尾 只能 震撼 幾次 結尾...		KoreaDrama	TWCN_Drama
2118	淚之女王 結尾 音樂 無雷 請問 預告 這段 好好		KoreaDrama	EASeries
2119	閒聊 小吐槽 與惡魔有約 有雷 真的 宋江 顏值 攤起 劇看 第一集 先猜 結局 宋江 復活...		KoreaDrama	TWCN_Drama
2125	心得 淚之女王 喜歡 片段 淚之女王 喜劇 中帶 沈重 生病 妻子 狀況 慢慢 拾回 愛情 ...		KoreaDrama	TWCN_Drama
2133	情報 第屆 百想 藝術 大賞 入圍 名單 電視 百想 入圍 名單 激烈 名單 想必 意外 原...		KoreaDrama	TWCN_Drama
2141	心得 淚之女王 觀感 十集 感觸 有雷先 幾個 可惜 點畫 切換 細節 幾集 至少 有次 原...		KoreaDrama	TWCN_Drama
2142	心得 淚之女王 集二刷 幾個 漫長 一週 等待 重刷 幾個 一刷時 地方 劇本 用心 沒什麼...		KoreaDrama	TWCN_Drama
2144	情報 上線 日期 預告 上線 日期 預告 八個 參加 一檔 誘人 危險 遊戲 節目 被困 一...		KoreaDrama	TWCN_Drama
2146	心得 兩個 編劇 不可能的婚禮 完食 心得 終於 時間 這篇 心得 文了 每週 週三 乖乖 ...		KoreaDrama	TWCN_Drama
2147	情報 淚之女王 更新 劇照 編配 文讓 看情 聖賢 相簿 海仁 花海 仁親 海仁笑 全世界 ...		KoreaDrama	TWCN_Drama
2150	閒聊 支配物種 人工 培植 取代 動物 肉雷 韓劇 題材 真的 突破 故事 劇本 未來 十年...		KoreaDrama	EASeries
2156	心得 七人的復活 狗血 有沒有 沒差 集數 種花 集會 拆成 兩段 快速 複習 第一季 七人...		KoreaDrama	EASeries
2168	閒聊 淚之女王 更新 重看 忘記 本集 心動 橋段 要來 補充 我要 呐喊 追檔 折磨 熬夜...		KoreaDrama	TWCN_Drama

Case Study

```
pprint(false_pred['words'][2106])
```

```
('心得 寄生獸 普雷日 韓兩版 防雷 一二三四五六七八九十 開門見山 喜歡 日版 原因 循序漸進 過程 安排 完整 主角 剛被 寄生 慢慢 寄生獸 寄生獸  
'融入 社會 只會 慢慢 融入 社會 寄生獸 寄生獸 不吃 導入 主角 戰鬥能力 寄生獸 變強 設定 級別 寄生獸 省略 太多 省略 設定 感情 戰術 '  
'主角 寄生獸 暫時 分離 日版 寄生獸 壓迫感 較強 韓版 爬彈槍 能幹 寄生獸 韓版 單調 寄生獸 組織 莫名 韓版 寄生獸 實力 像是 惡靈古堡 '  
'裡的 舐爺 濕臨絕種 舐爺 人類 部隊 完虐 寄生獸 劇情 寄生獸 生存 還在 內部 大屠殺 團結 生存 更久 背叛 稍微 砍死 還算 隊友 殺光 '  
'吃錯藥 隊友 死光 位置 自我 繁殖')
```

可以看到在第2106筆資料中，有較多CFantasy版中常出現的字詞
(如：砍死、殺光、戰鬥能力等)，因此模型判斷錯誤。



Week 9

主題分析

基本介紹：

資料集: ptt-速食版、美妝版、裝潢版、道路版、育嬰版、陸劇版 (2024-07-01~2025-04-15)

Catagory	data (4441)
fastfood	571 (12.9%)
MakeUp	442 (10%)
Interior	970 (21.8%)
Road	519 (11.7%)
BabyMother	951(21.4%)
China-Drame	988(22.2%)

模擬生活中食衣住行育樂
對應不同版進行主題分析

LEXICON-BASED / 人工給定主題的主題模型

食 (fastfood)

- 品牌 | 麥當勞、肯德基...
- 主餐類型 | 漢堡、雞塊...
- 配餐 | 薯條、薯餅...
- 口味 | 起司、明太子...
- 飲品類 | 飲料、冰淇淋...
- 消費 | 單點、套餐...
- 評價 | 好吃、嚐鮮...

衣 (MakeUp)

- 彩妝品類 | 粉底液、粉底...
- 眼妝系列 | 眼影、眼線液...
- 唇妝系列 | 唇膏、唇彩...
- 工具 | 化妝刷、粉撲...
- 護膚 | 精華、乳液...
- 品牌 | 香奈兒、資生堂...
- 妆感特性 | 妆效、妝感...

住 (Interior)

- 空間類型 | 客廳、玄關...
- 主要材料 | 地板、磁磚...
- 施工工法 | 水電、油漆...
- 家具 | 家具、櫥櫃...
- 設計流程 | 裝潢、裝修...
- 風格取向 | 北歐、簡約...

行 (Road)

- 道路類型 | 高架、國道...
- 交通管控 | 交通、車道...
- 路網結構 | 匝道、隧道...
- 施工維護 | 施工、封路...
- 管理單位 | 交通局、監理...

育 (BabyMother)

- 嬰兒關鍵字 | 寶寶、嬰兒...
- 用品 | 奶瓶、奶嘴...
- 護理流程 | 產檢、產後...
- 日常照護 | 保母、月子中心
- 醫療 | 疫苗、護理...
- 活動 | 親子、幼兒園...

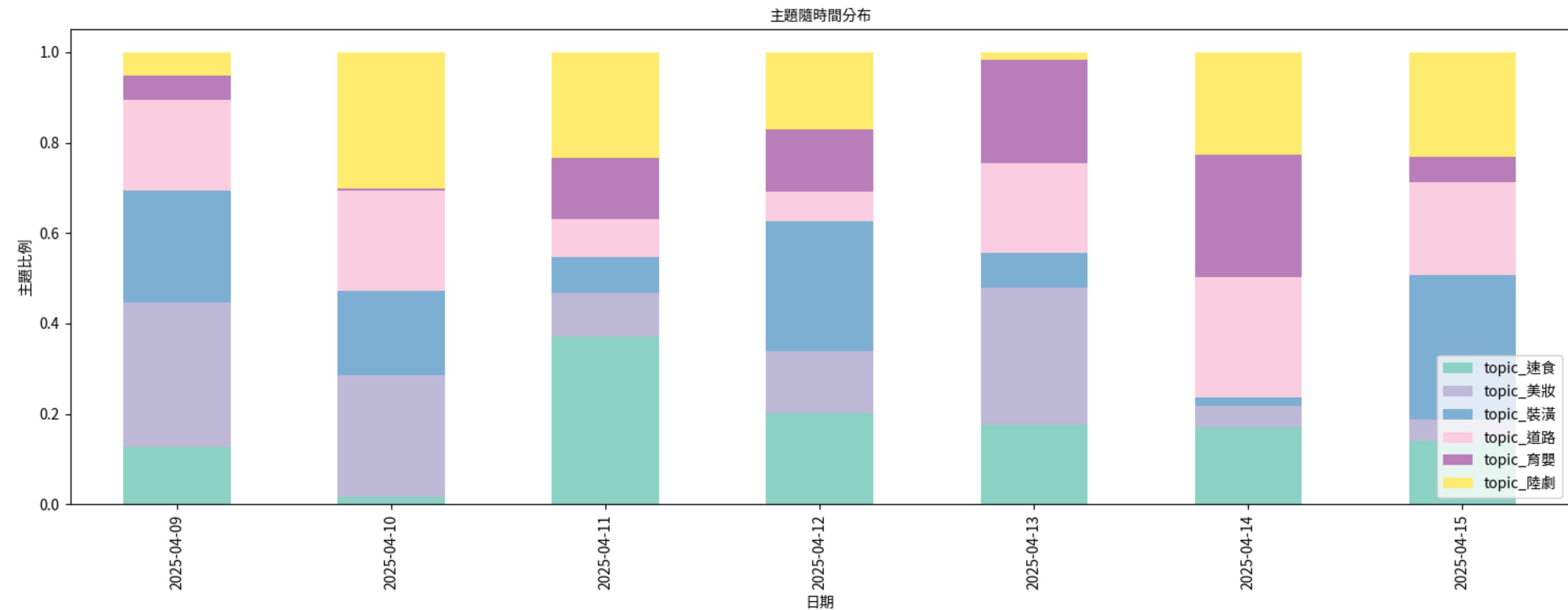
樂 (China-Drama)

- 劇情 | 劇情、角色...
- 製作環節 | 劇集、劇本...
- 演員團隊 | 演員、主演...
- 播放平台 | 騰訊、愛奇藝...
- 互動形式 | 預告、花絮...

WEEK09

人工給定主題分佈

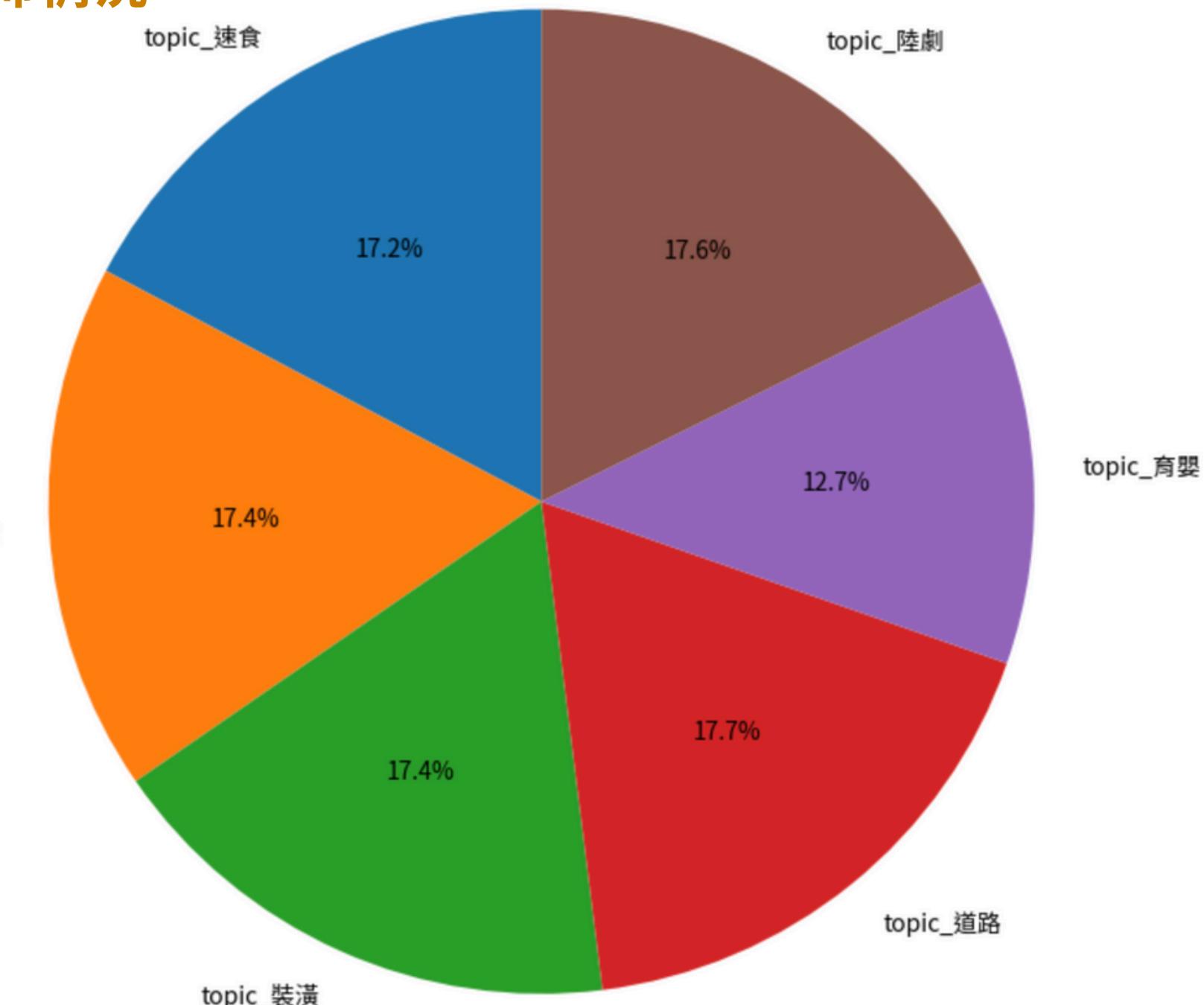
2025/4/9-2025/4/15 主題分佈圖



人工給定主題分佈

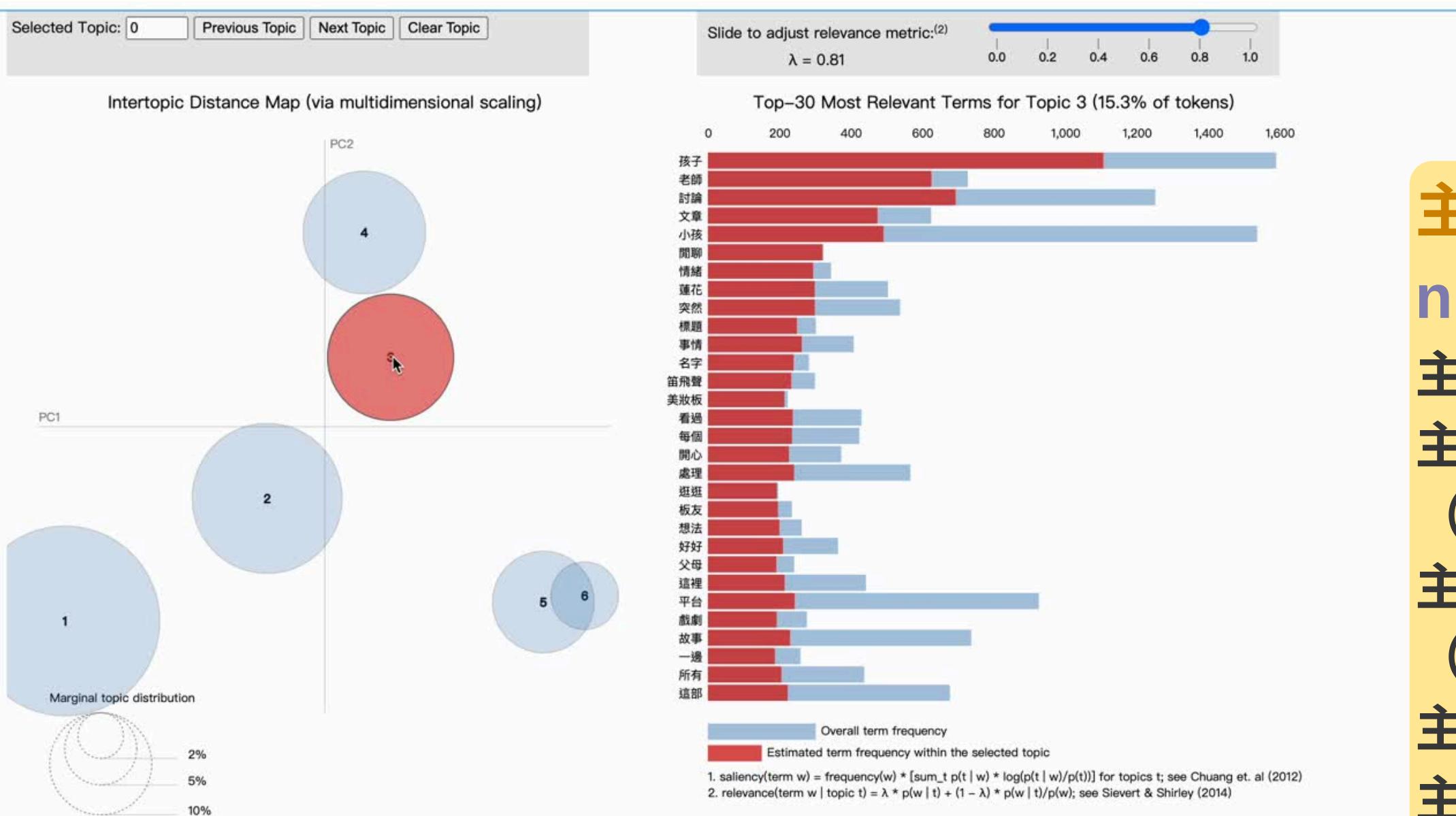
2025/4/9-2025/4/15 總體圓餅圖主題分佈情況

分佈趨於平均
可以看出育嬰相關的話題討論度最少



WEEK09

LDA 主題模型



主題解說：

`num_topics = 6`

主題1:道路施工

主題2:媽媽懷孕期間就醫情況
(醫院附近有麥當勞)

主題3:麥當勞推出新活動
(附近交通情況)

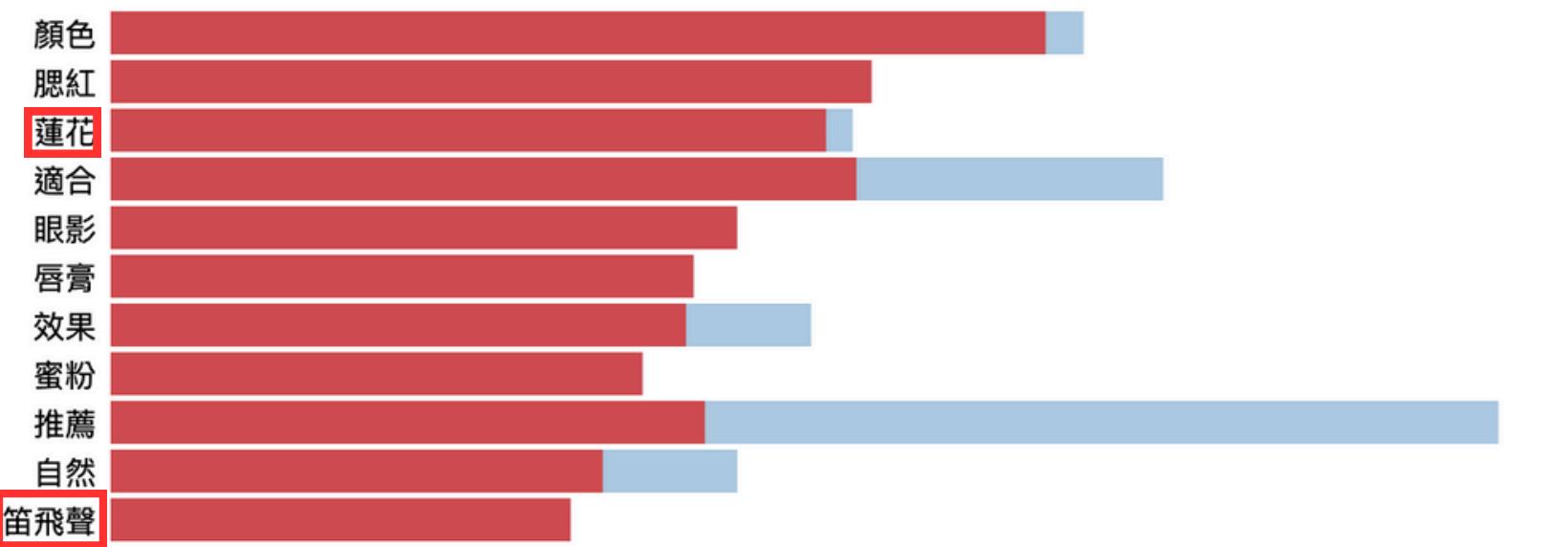
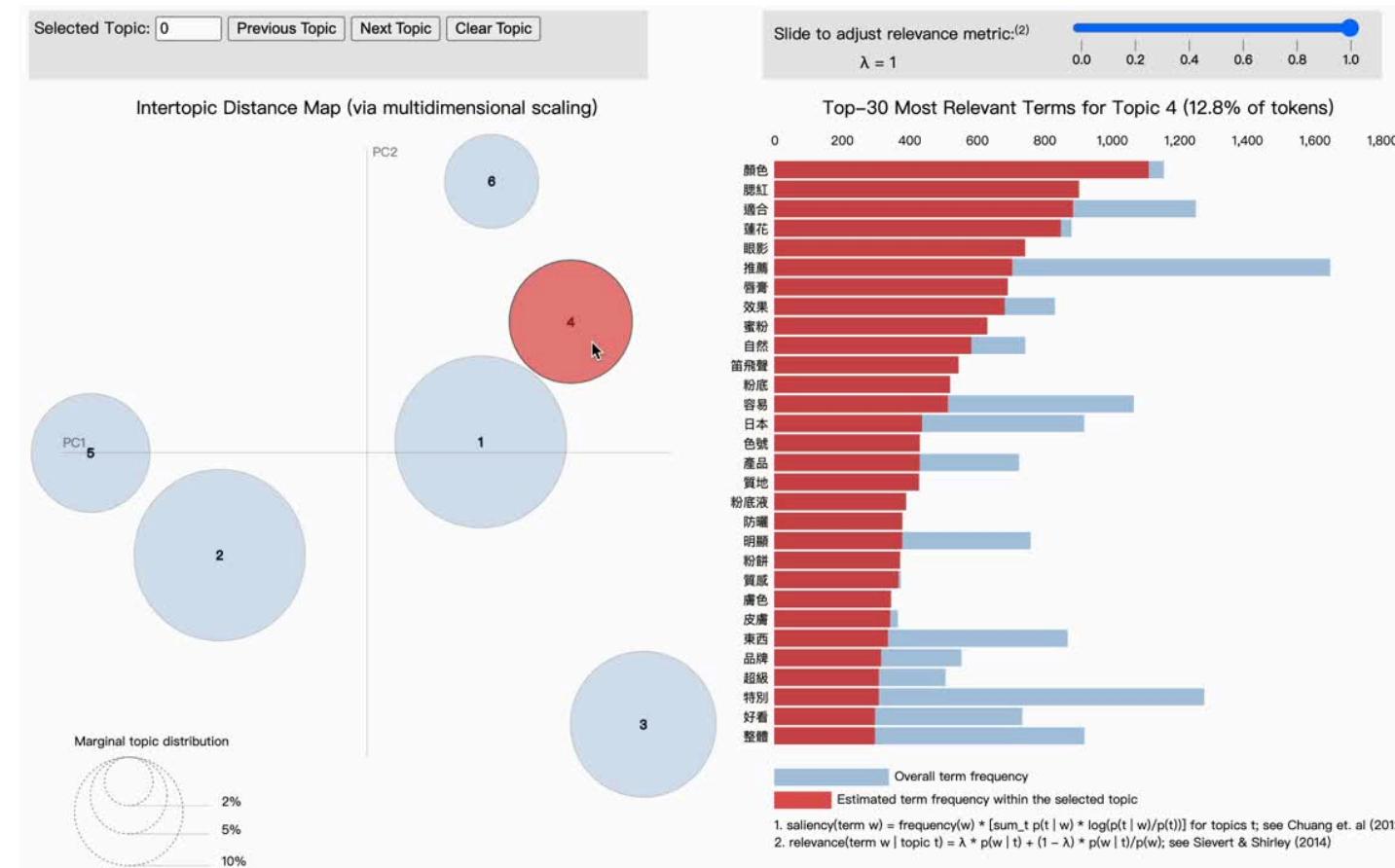
主題4:討論劇情

主題5:美妝試用效果

主題6:美妝推薦

WEEK09

GUIDED LDA



設定關鍵字-食衣住行育樂

seed_topic_list = [

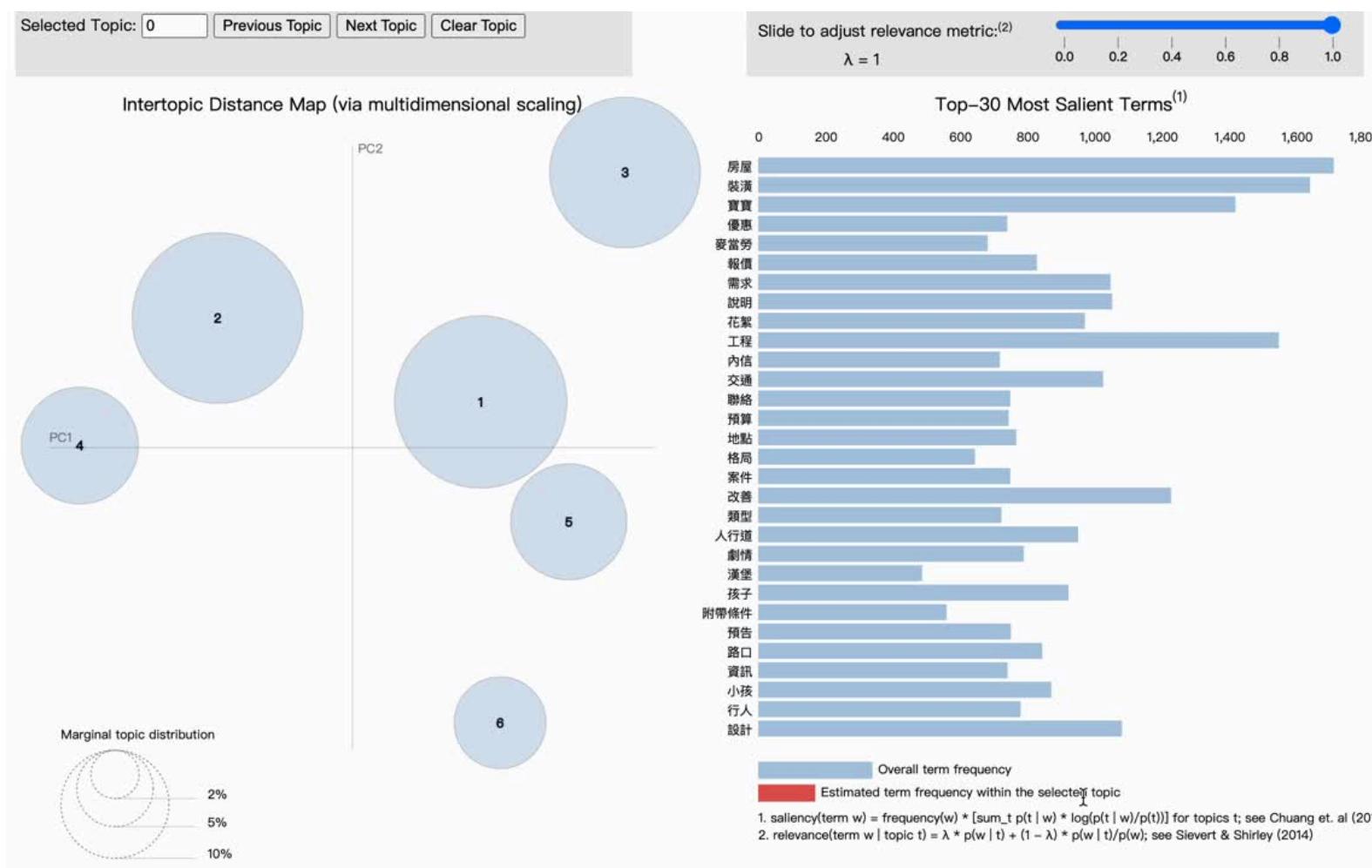
["麥當勞", "炸雞", "薯條", "漢堡", "套餐"],
["眼影", "粉底", "唇膏", "定妝", "妝容"],
["房屋", "裝潢", "廚房", "坪數", "浴室"],
["交通", "車道", "工程", "施工", "人行道"],
["寶寶", "媽媽", "孩子", "照顧", "家庭"],
["劇情", "演員", "角色", "小說", "花絮"]

]

主題4 包含了美妝和陸劇（蓮花樓）

WEEK09

GUIDED LDA



調整關鍵字：

seed_topic_list = [

["麥當勞", "炸雞", "薯條", "漢堡", "套餐"],
["眼影", "粉底", "唇膏", "定妝", "妝容"],
["房屋", "裝潢", "廚房", "坪數", "浴室"],
["交通", "車道", "工程", "施工", "人行道"],
["寶寶", "媽媽", "孩子", "照顧", "家庭"],
["劇情", "演員", "角色", "小說", "花絮",
"蓮花", "笛飛聲"],

]

調整後有很好的被分類為食衣住育樂

主題1：有關家庭育嬰相關(育)、主題2：道路施工(行)、主題3：陸劇蓮花樓 (樂)

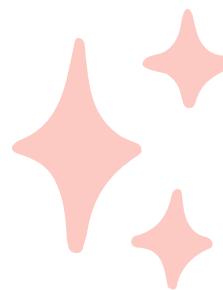
主題4：房屋裝潢花費 (住)、主題5：美妝保養產品推薦 (衣)、主題6：速食產品分享 (食)

WEEK09

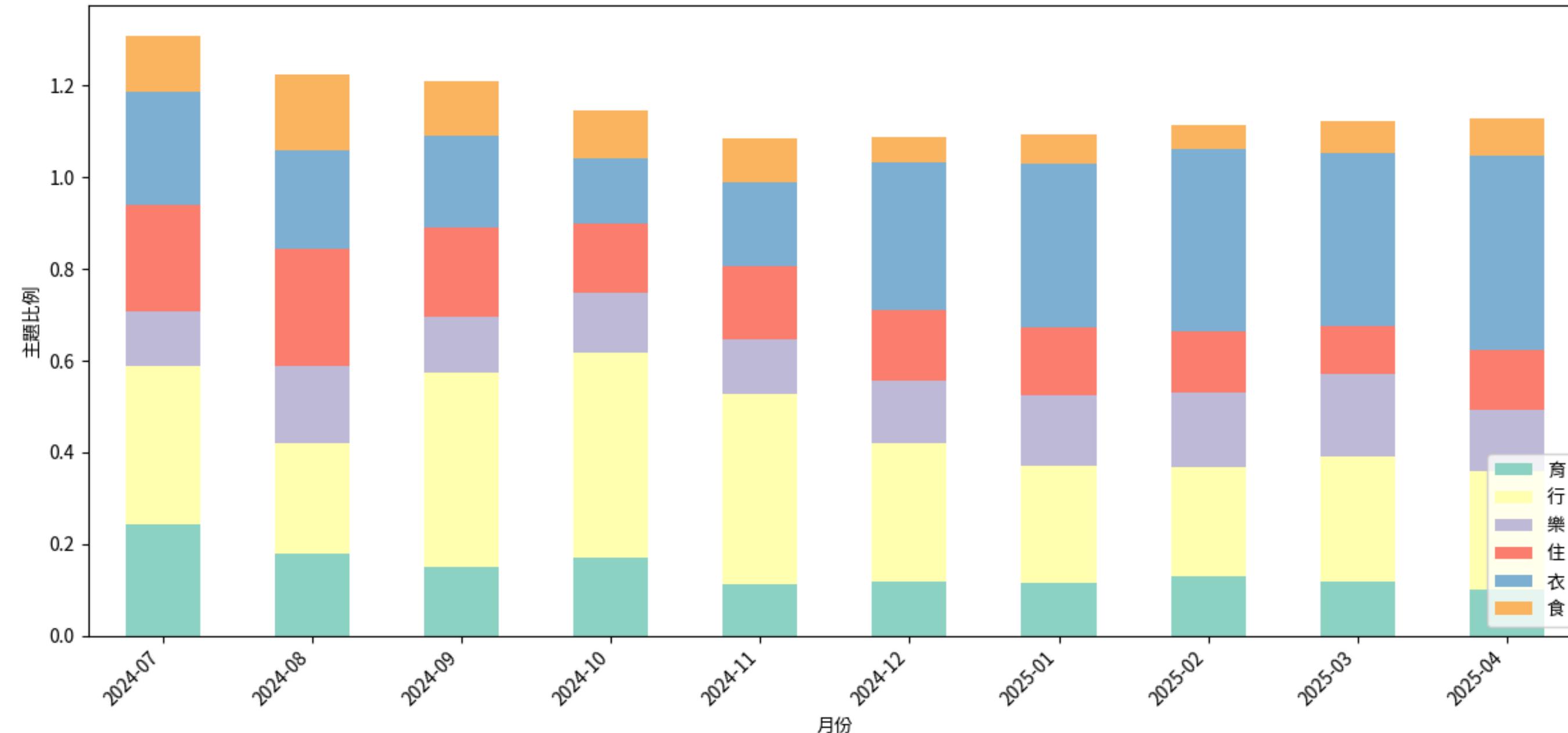
每月主題文章佔比變化

透過模型觀察每月文章佔比

2024/7-2025/4 主題分佈圖



每月主題分布變化



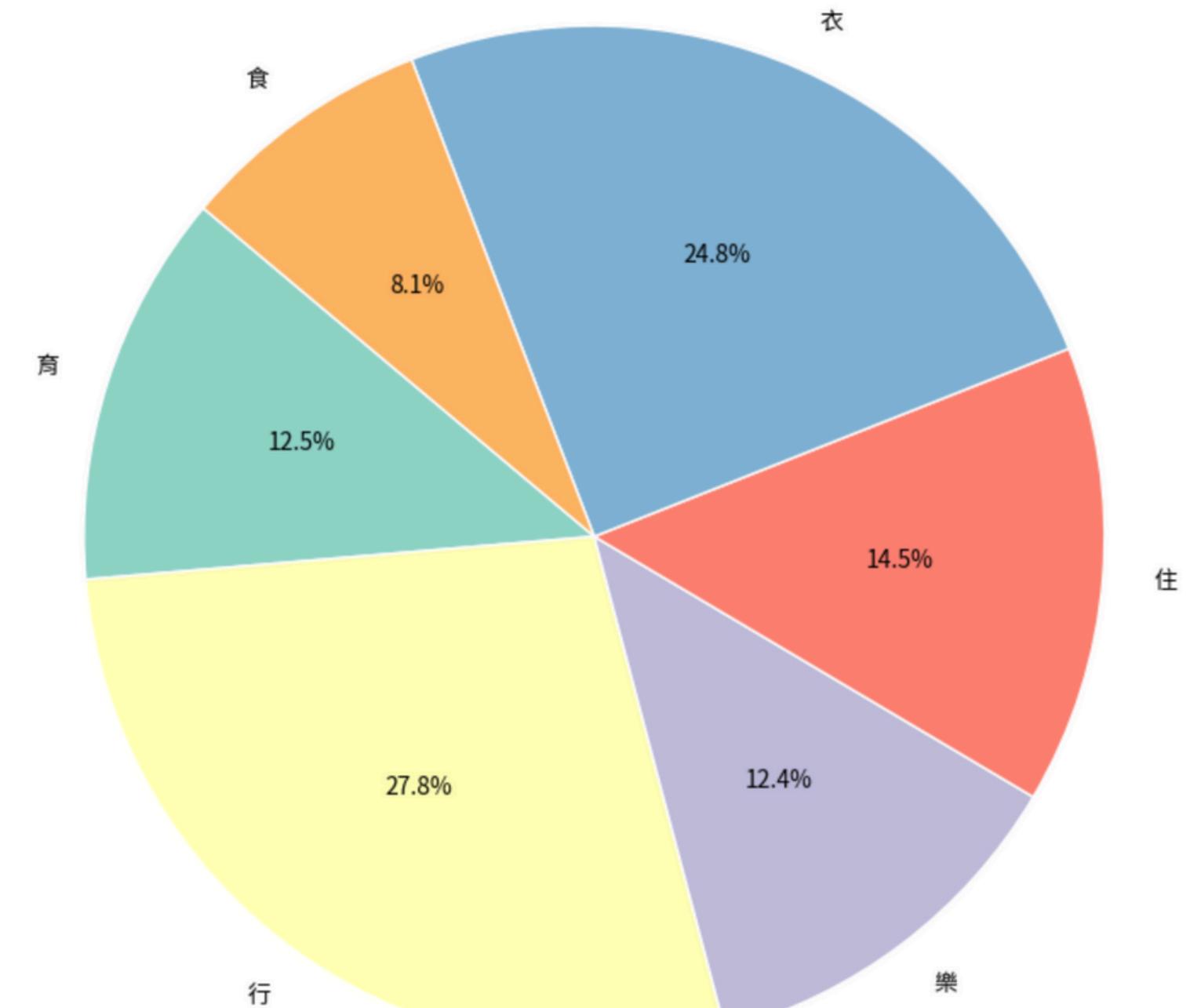
WEEK09

每月主題文章佔比變化

2024/7/1-2025/4/15 整題圓餅圖主題分佈情況

可以看出在2024/7/1-2025/4/15
行(道路施工)和衣(美妝)討論是最多的

整體主題分布圓餅圖





Thank You!

