

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 11
Text Embeddings



Week 12
BERTopic



Week 11
Text Embeddings

WEEK11

基本介紹

資料集

使用爬蟲抓取。

- ptt速食版
- 關鍵字：麥當勞
- 時間：2022-06-21 ~ 2025-03-21
- 資料筆數：1000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



WEEK11

WORD2VEC模型建立

利用麥當勞資料集建立word2vec模型

```
w2v_model.wv.most_similar('套餐',topn=10)
```

```
[('雞塊', 0.7287017703056335),  
 ('麥香魚', 0.6616314053535461),  
 ('飲料', 0.6578062176704407),  
 ('甜心卡', 0.6360278725624084),  
 ('中薯', 0.6310502290725708),  
 ('加價', 0.5952916741371155),  
 ('中杯', 0.580108642578125),  
 ('單點', 0.5408732295036316),  
 ('大薯', 0.5388852953910828),  
 ('大麥克', 0.5101746320724487)]
```

```
w2v_model.wv.similarity("大薯","冰炫風")
```

0.84286463

```
w2v_model.wv.similarity("優惠","漲價")
```

0.35325685

WEEK11

WORD2VEC模型建立

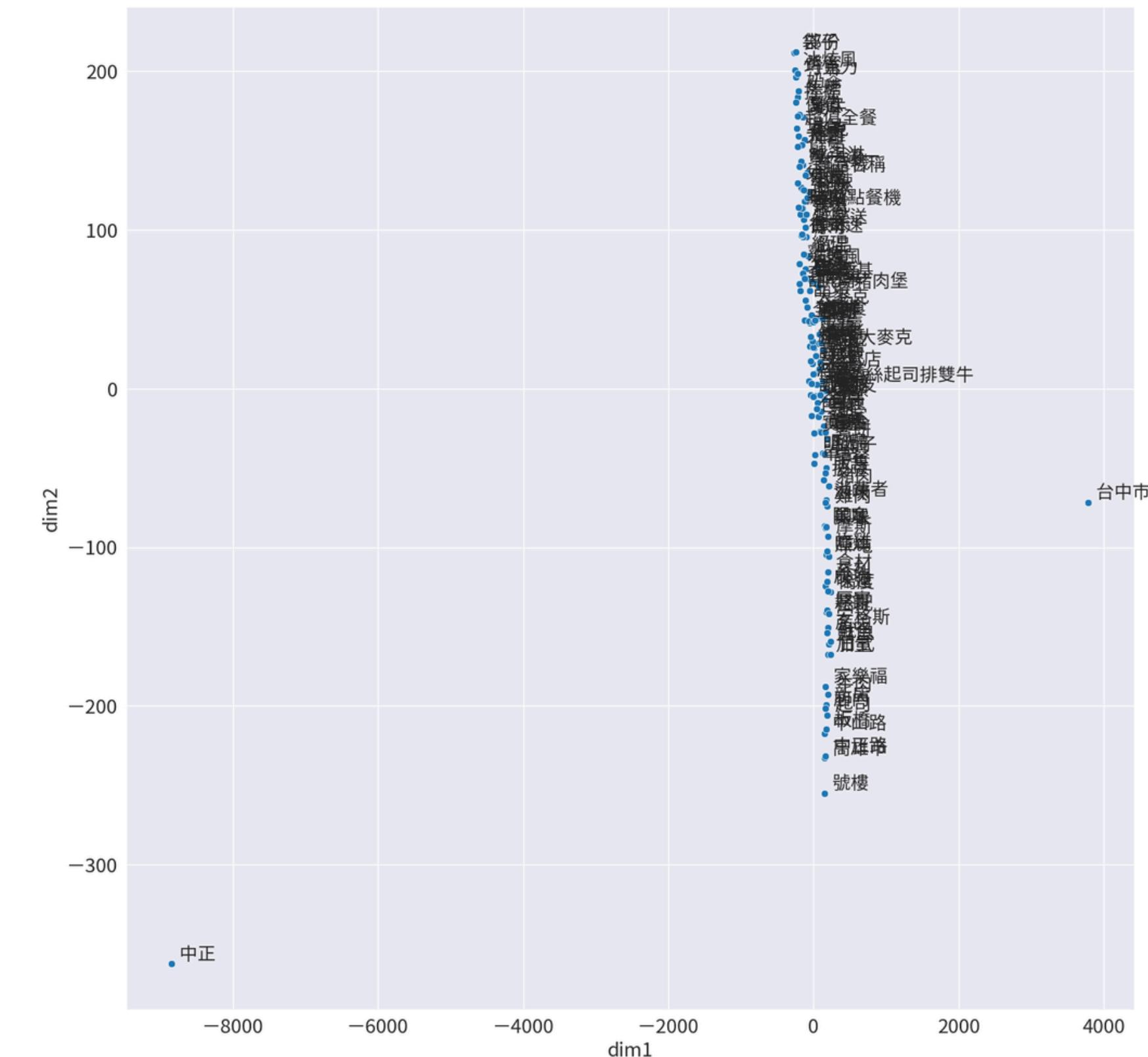
```
# 跟兩個字最不相關  
w2v_model.wv.most_similar(negative=['大麥克', '套餐'], topn=10)  
  
[('新北市', 0.43770653009414673),  
 ('調整', 0.43545466661453247),  
 ('台北市', 0.42688611149787903),  
 ('號樓', 0.4250276982784271),  
 ('中山', 0.41516992449760437),  
 ('高雄市', 0.4106994569301605),  
 ('台中市', 0.4083278477191925),  
 ('桃園', 0.4015451669692993),  
 ('中正', 0.3921082615852356),  
 ('家樂福', 0.37845054268836975)]
```

```
expandPosWord(w2v_model, ['大薯', '冰炫風'], top_n = 10)  
  
['巧克力',  
 '搖搖',  
 '超值全餐',  
 '儲值',  
 '草莓',  
 '環保',  
 '袋子',  
 '辣味',  
 '焦糖',  
 '小杯',  
 '冰炫風',  
 '喜歡',  
 '部份']
```

WEEK11

二維圖

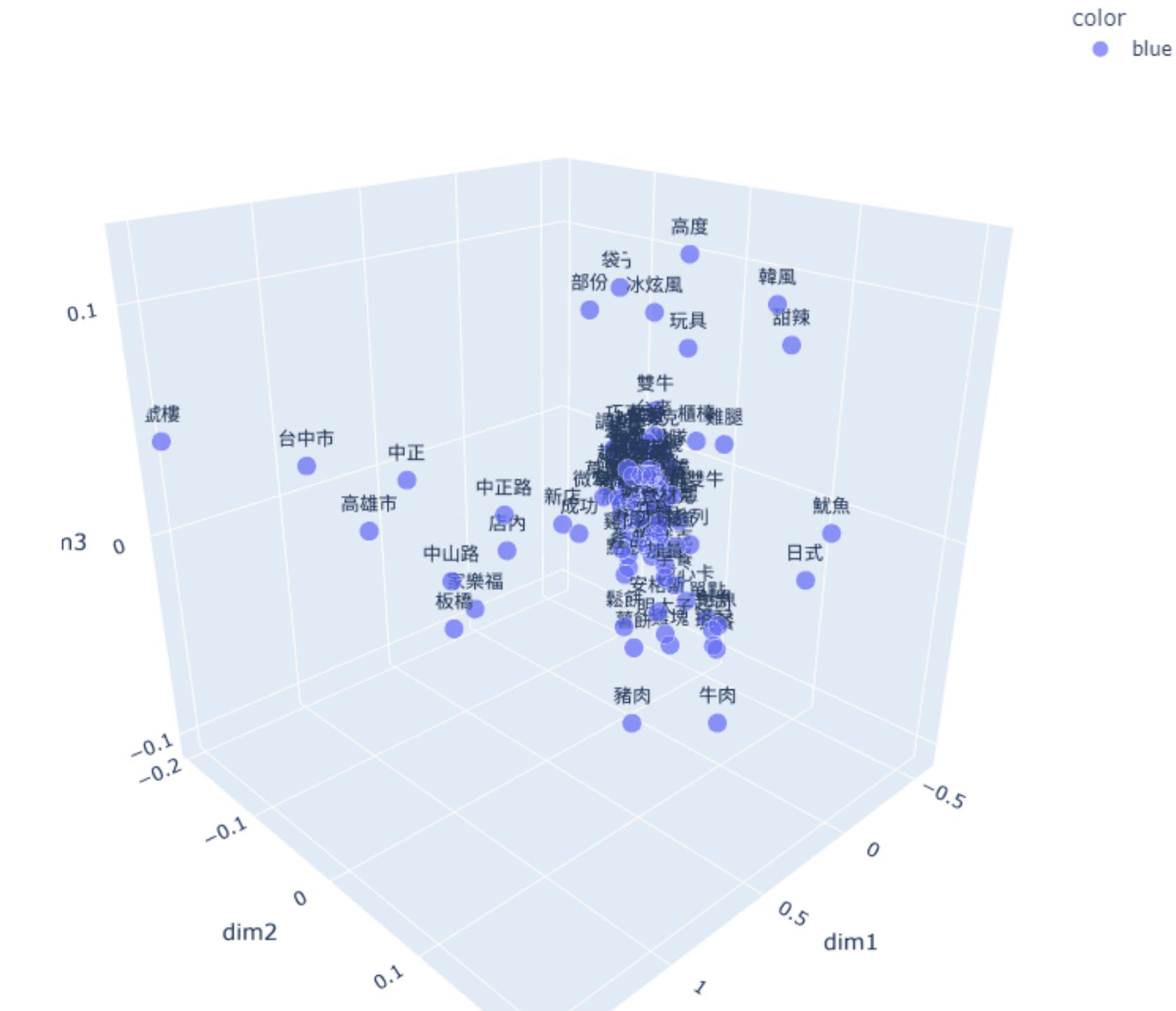
- 在二維空間相當難以辦別出有用資訊
- 改為利用三維圖觀察



WEEK11

三維圖

- 在三維空間能夠更清楚看出向量分布
 - 還是無法辨識有效資訊



WEEK11

三維圖 - KMEANS

- 分為四類
 - 每類前10名詞彙

食材與餐點組合

加量, 產品, 鮭魚, 厚實, 安格斯, 搭配, 鬆餅, 日式, 牛肉, 高麗

 促銷行銷語與商品推薦

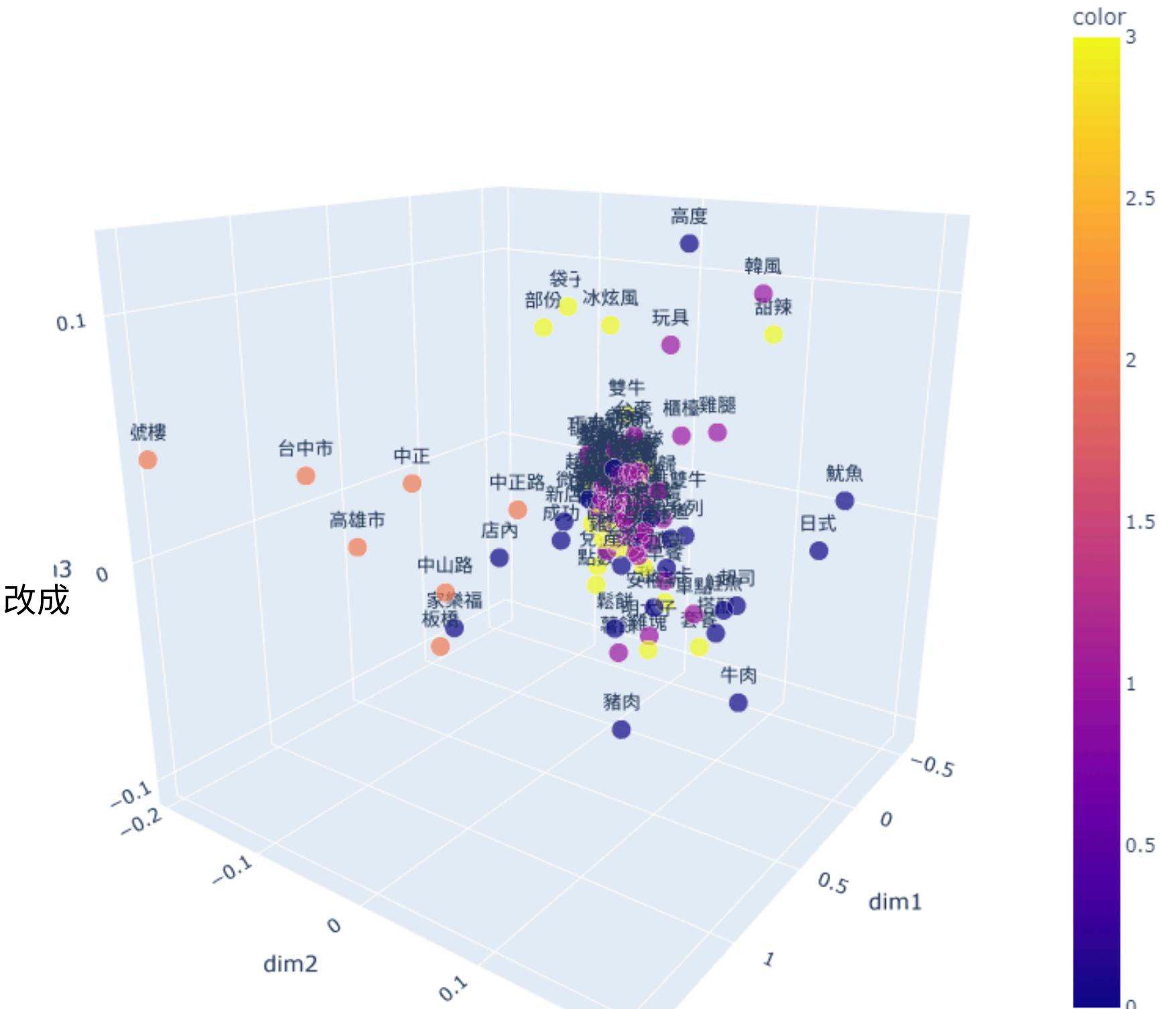
會員，醬汁，這款，微牽絲起司排雙牛，香港，正式，人氣，辦法，持續，改

店鋪地點與地址

中正, 高雄市, 中正路, 台中市, 中山路, 板橋, 號格

◆ 價格優惠與活動資訊

抽到 顧客 便宜 星級 漲價 提醒 買一送一 時段 冰淇淋 台幣



WEEK11

相似文件

- Embedding模型: bert-base-chinese
- 指定文章索引，找出相似文章
- 企業可以根據一篇負面文章，找出其他大眾的抱怨點

Query: [情報] 麥當勞2025甜心卡點餐價格

資料集中前五相似的文章：

- [情報] 麥當勞2025甜心卡點餐價格 (Score: 1.0000)
- [討論] 麥當勞點數回饋率計算 (Score: 0.9657)
- [新聞] 麥當勞全新「1+1」只要89元！ (Score: 0.9595)
- Re: [討論] 麥當勞100元內最佳搭配 (非早餐時段) (Score: 0.9583)
- Re: [新聞] 麥當勞全新「1+1」只要89元！ (Score: 0.9576)

Query: [討論] 麥當勞員工手機掉油鍋

資料集中前五相似的文章：

- [討論] 麥當勞員工手機掉油鍋 (Score: 1.0000)
- [抱怨] 15:09更新後續在推文請問麥當勞滿福堡這是...蟑螂嗎？ (Score: 0.9557)
- [問題] 麥當勞監視器影像保留時間 (Score: 0.9497)
- [討論] 麥當勞疑似把1+1 50藏起來？ (Score: 0.9470)
- Re: [抱怨] 麥當勞點餐點到超挫折 (Score: 0.9465)

WEEK11

文件分類

- 資料集: ptt-韓劇版、歐美影集版、玄幻版、台劇&陸劇
(2024-10-05~2025-04-05)
- Embedding模型: bert-base-chinese
- 7:3切分訓練集
- LogisticRegression()

	Train (4720)
KoreaDrama	1346 (28.5%)
CFantasy	1217 (25.8%)
EAseries	1180 (25%)
TWCN_Drama	977 (20.7%)

WEEK11

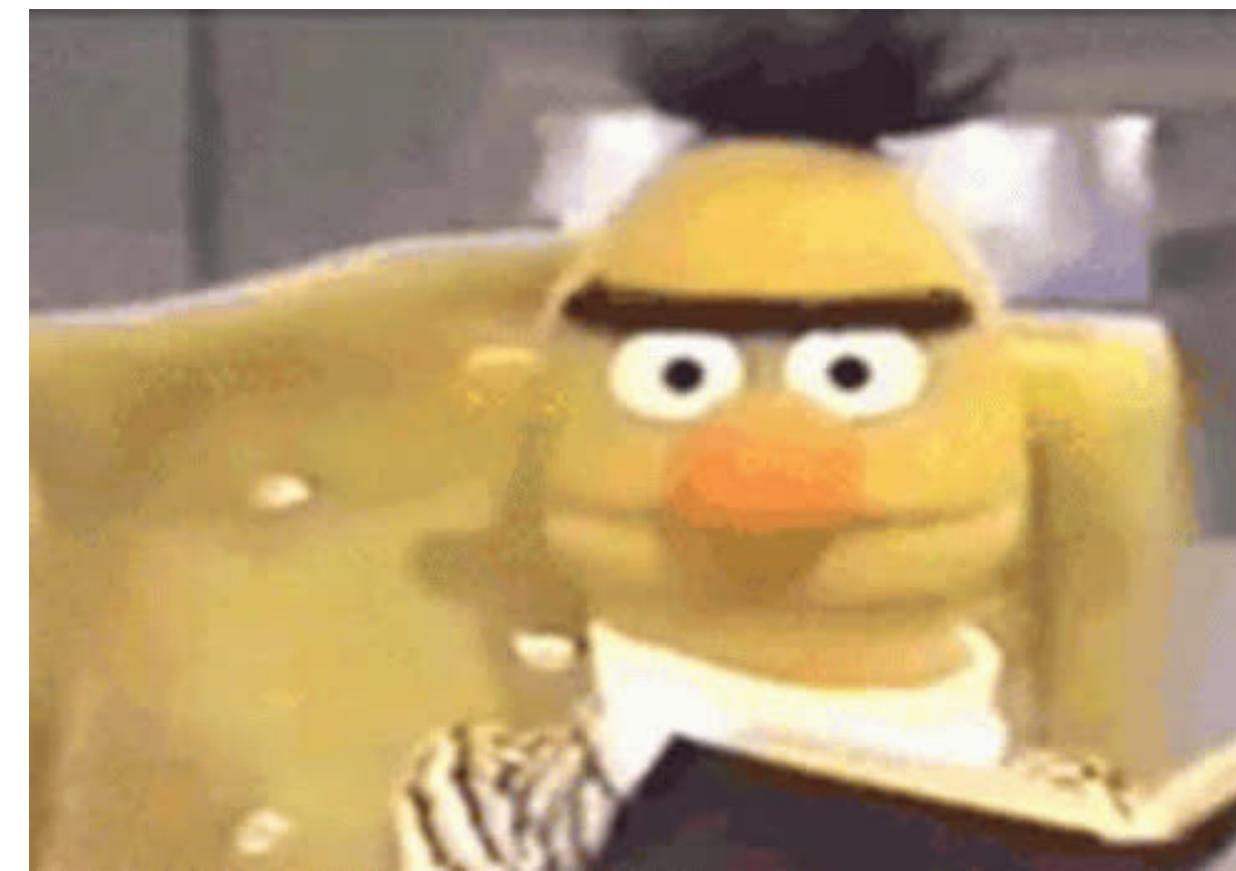
文件分類

- 第七週結果
 - TF-IDF
 - SVM
 - 當時最佳結果

	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

- 本週結果
 - Embeddings
 - LogisticRegression()
 - 整體結果大幅提升

	precision	recall	f1-score	support
CFantasy	0.96	0.96	0.96	367
EASeries	0.97	0.94	0.96	354
KoreaDrama	0.97	0.97	0.97	384
TWCN_Drama	0.90	0.93	0.92	311
accuracy			0.95	1416
macro avg	0.95	0.95	0.95	1416
weighted avg	0.95	0.95	0.95	1416



Week 12

BERTopic

WEEK12

基本介紹

資料集與前處理：皆與 Week 11 相同



資料集

使用爬蟲抓取。

- ptt速食版
- 關鍵字：麥當勞
- 時間：2022-06-21 ~ 2025-03-21
- 資料筆數：1000筆

基本處理 與 結果觀察

資料清理、斷詞

自訂：辭典與停用字詞典

WEEK12

抓取NER

	sentence	packed_sentence	entities
0	日本麥當勞起兒童餐玩具哆啦夢	日本(Nc) 麥當勞(Nc) 起(VC) 兒童餐(Na) 玩具(Na) 哆啦夢(Nb)	[NerToken(word='日本', ner='GPE', idx=(0, 2)), N...
1	介紹影片	介紹(VE) 影片(Na)	[]
2	資料來源	資料(Na) 來源(Na)	[]
3	廠商分店別麥當勞苗栗店	廠商(Na) 分店別(Nc) 麥當勞(Nc) 苗栗店(Nc)	[NerToken(word='麥當勞', ner='ORG', idx=(5, 8)), ...]
4	商品名稱草莓優格雙餡派	商品(Na) 名稱(Na) 草莓(Na) 優格(Na) 雙餡派(A)	[]
5	評分	評分(VA)	[]
6	想到之前有些特殊口味的派總是很快售完	想到(VE) 之前(Ng) 有(V_2) 些(Nf) 特殊(VH) 口味(Na) 的(DE)...	[]
7	錯過就沒機會	錯過(VJ) 就(D) 沒(VJ) 機會(Na)	[]
8	所以這次決定先買來嚐鮮	所以(Cbb) 這(Nep) 次(Nf) 決定(VE) 先(D) 買來(VC) 嚐(VC) ...	[]
9	結論並不理想	結論(Na) 並(D) 不(D) 理想(VH)	[]

被抓出實體有國家、機構

Relation Extraction

使用model：xlm-roberta-base

		sentence	label	score
0		日本麥當勞起兒童餐玩具哆啦夢	LABEL_1	0.560699
1		秒	LABEL_1	0.561102
3		介紹影片	LABEL_1	0.570859
5		第彈金木	LABEL_1	0.561496
6		第彈金木	LABEL_1	0.561496
...	
525		吃完頗有飽足感	LABEL_1	0.565534
527		更新修正南京以及竹山大明店名	LABEL_1	0.563867
528		加上雙和太平洋	LABEL_1	0.563212
529		先說結論目前還差以及號店沒有找到	LABEL_1	0.560061
530	商工公示資料上有註冊但還沒找到店號的有忠孝三台北吳興台北漢中台北南京四台北復興台北西寧台北天...	LABEL_1	0.562133	

結果：皆被認定為是『LABEL_1』的關係

因沒找到針對『繁中』的RE模型，因此結果並不是很好

應先事先label，結果才會是正確的

WEEK12

主題建模

使用 BERTopic 進行主題模型建立

使用模型：

- embedding model : uer/sbert-base-chinese-nli
 - 把句子轉成向量 (embedding)
- 聚類模型 : HDBSCAN (預設)
 - 將上一步的 embedding 向量做 主題分群
- 詞頻模型 : CountVectorizer (預設)
 - 幫助萃取每個主題的關鍵詞
- 主題模型 : BERTopic
 - 將以上模型組合建構出主題架構與結果

WEEK12

主題建模：嘗試其他Embedding model

因 google-bert/bert-base-chinese 輸出的向量為token-level，
需要手動 mean-pooling，對分群結果可能模糊

因此：使用已設計好 pooling、sentence-level 的
sentence embedding models

模型名稱：

shibing624/text2vec-base-chinese

基本介紹：

常用於中文主題建模、主題分類、關鍵詞易解釋

uer/sbert-base-chinese-nli

對短句分類、相似句子分類表現穩定

paraphrase-multilingual-MiniLM-L12-v2

適用於多語言混合資料

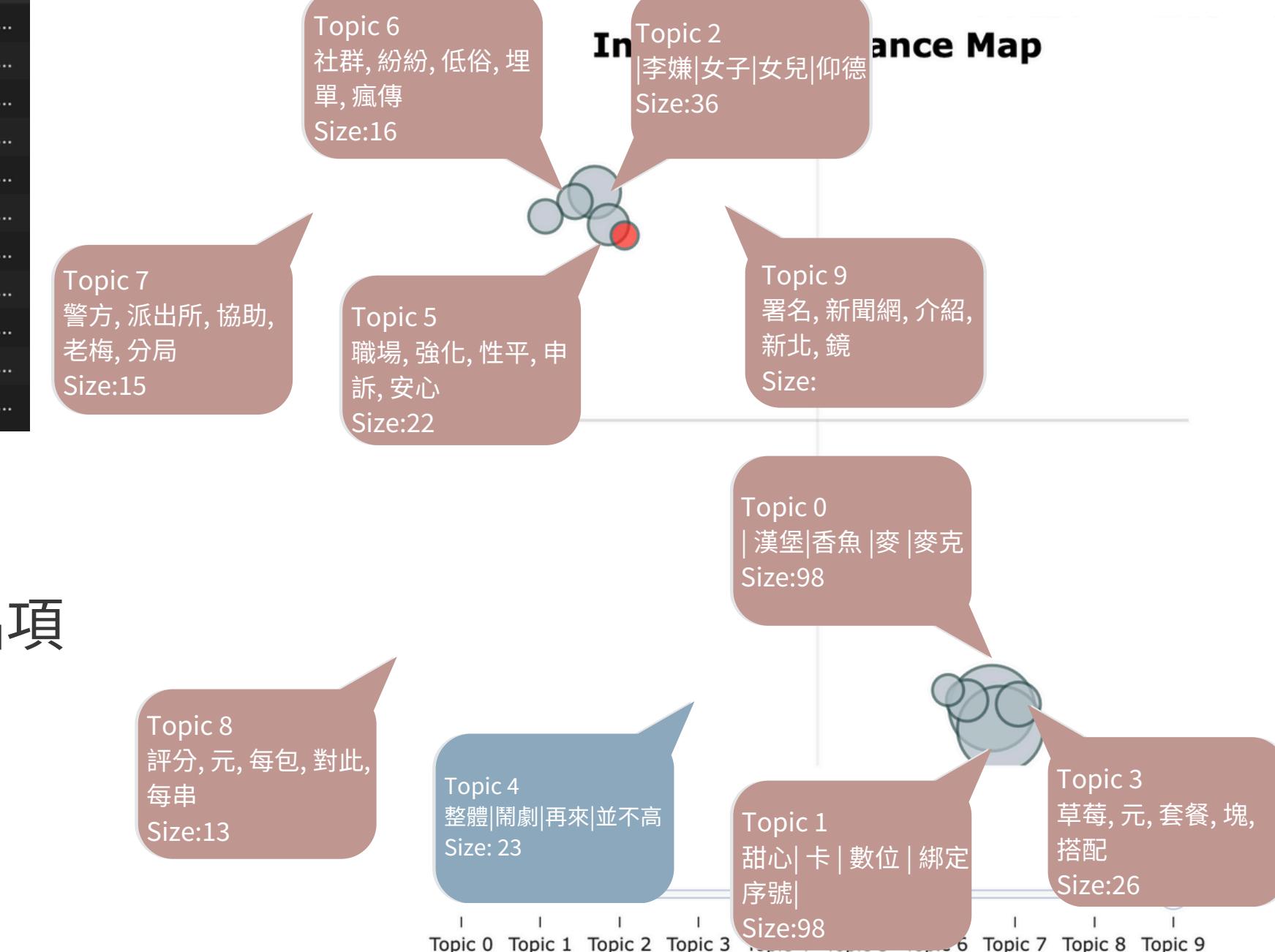
WEEK12 試用不同Embedding model: google-bert/bert-base-chinese (助教提供)

Topic	Count	Name	Representation
-1	143	-1_中_店_家樂福	[, 中, 店, 家樂福, 廠商, 板, 行走, 大雨, 挪威克朗, 徒步, 標題新聞, ...]
0	98	0_漢堡_香魚_麥	[, 漢堡, 香魚, 麥, 麥克, 套餐, 司片, 打開, 無敵, 額外, 吃, 雙人, ...]
1	98	1_甜心_卡_數位_綁定	[甜心, 卡, 數位, 綁定, 序號, 優惠, 優惠券, , 兌換, 實體, 領取, 送,...]
2	36	2_李嫌_女子_女兒_	[李嫌, 女子, 女兒, , 仰德, 紅髮, 小茹, 離家, 一名, 二代, 涉案, 該名...]
3	26	3_草莓_元_套餐_塊	[草莓, 元, 套餐, 塊, 搭配, 優格, 紅茶, 雞, 果醬, 雙餡, 派, 冰炫風, ...]
4	23	4_整體_鬧劇_再來	[整體, , 鬧劇, 再來, 並不高, 值得, 值, 排序, 不斐, 不及格, 超級, 台...]
5	22	5_職場_強化_性平_申訴	[職場, 強化, 性平, 申訴, 安心, 協助, 培訓, 走失, 管道, 團體, 團隊, 身...]
6	16	6_社群_紛紛_低俗_埋單	[社群, 紛紛, 低俗, 埋單, 瘋傳, 不雅片, , 網怒, 唯一, 引, 發起, 痛轟...]
7	15	7_警方_派出所_協助_老梅	[警方, 派出所, 協助, 老梅, 分局, 警, 家屬, 金山, 溝通, 求助, 檢方, 耐...]
8	13	8_評分_元_每包_對此	[評分, 元, 每包, 對此, 每串, 食用, 不高, 卡面, 一張, , 款, , , ...]
9	10	9_署名_新聞網_介紹_新北	[署名, 新聞網, 介紹, 新北, 鏡, 郭世賢, 週刊, 蕭涵, 葵, 莊雅婷, 綜合, ...]

共分出10類主題

集中於兩大主題：社會事件、麥當勞品項

主題四並不是很符合品項相關

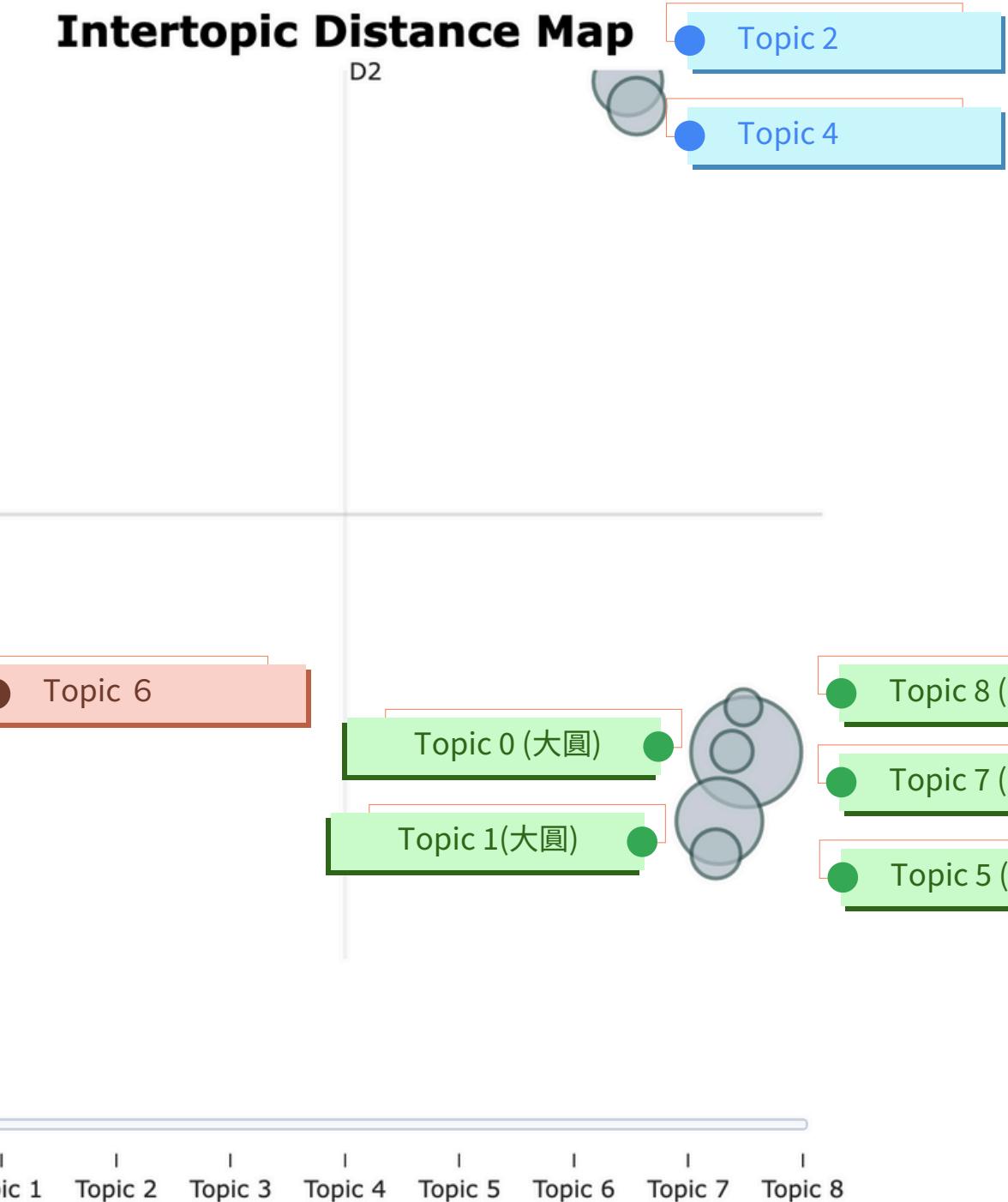


WEEK12 試用不同Embedding model: shibing624/text2vec-base-chinese

Topic	Count	Name	Representation
-1	106	-1_評分_女子_麥克	[, 評分, 女子, 麥克, 卡, 遭性, 侵, 女兒, 挪威克朗, 甜心, 女模, 小茹...
0	120	0_協助_警方_職場	[, 幫助, 警方, 職場, 金山, 中, 大雨, 派出所, 強化, 性平, 返家, 社群...
1	73	1_數位_甜心_綁定_序號	[數位, 甜心, 綁定, 序號, 卡, 優惠券, , 優惠, 領取, 實體, 兌換, 即可...
2	47	2_草莓_元_紅茶_套餐	[草莓, 元, 紅茶, 套餐, 冰炫風, 塊, , 優格, 搭配, 雞, 口感, 果醬, ...]
3	45	3_二代_李嫌_廠商_家樂福	[二代, 李嫌, 廠商, 家樂福, 店, , 富, 背景, 仰德, 頭份, 今日, 公里, ...]
4	32	4_香魚_漢堡_麥_薯餅	[香魚, 漢堡, 麥, 薯餅, 司片, 打開, , 供應, 薯條, 起司片, 半片, 大份...
5	24	5_雙人_早餐_點餐機_套餐	[雙人, 早餐, 點餐機, 套餐, , 自動, 蛇, 開春, 完, 鮑餐一頓, 維護, 太...
6	23	6_甜心_卡_優惠_自助	[甜心, 卡, 優惠, 自助, 點餐機, 買, 兌換, 卡今, , 速, 辦法, 開賣, ...]
7	17	7_自看_板_新聞網_署名	[自看, 板, 新聞網, 署名, , 莊雅婷, 薀, 晟, 新北, 區於日, 技術, 學院...
8	13	8_就業網_找到_開賣_查	[就業網, 找到, 開賣, 查, 查不到, 意義, 家裡, 先說, 就點, 忘, 太慢, 徒...

共分出9類主題

- 分類好：只有右上 【麥當勞品項】
- 分類差：
 - 左下：主題三與六的關係並無高度相關，分類效果不好
 - 右下：有社會事件、支付方式等不同，上下兩圓理應不該有重疊



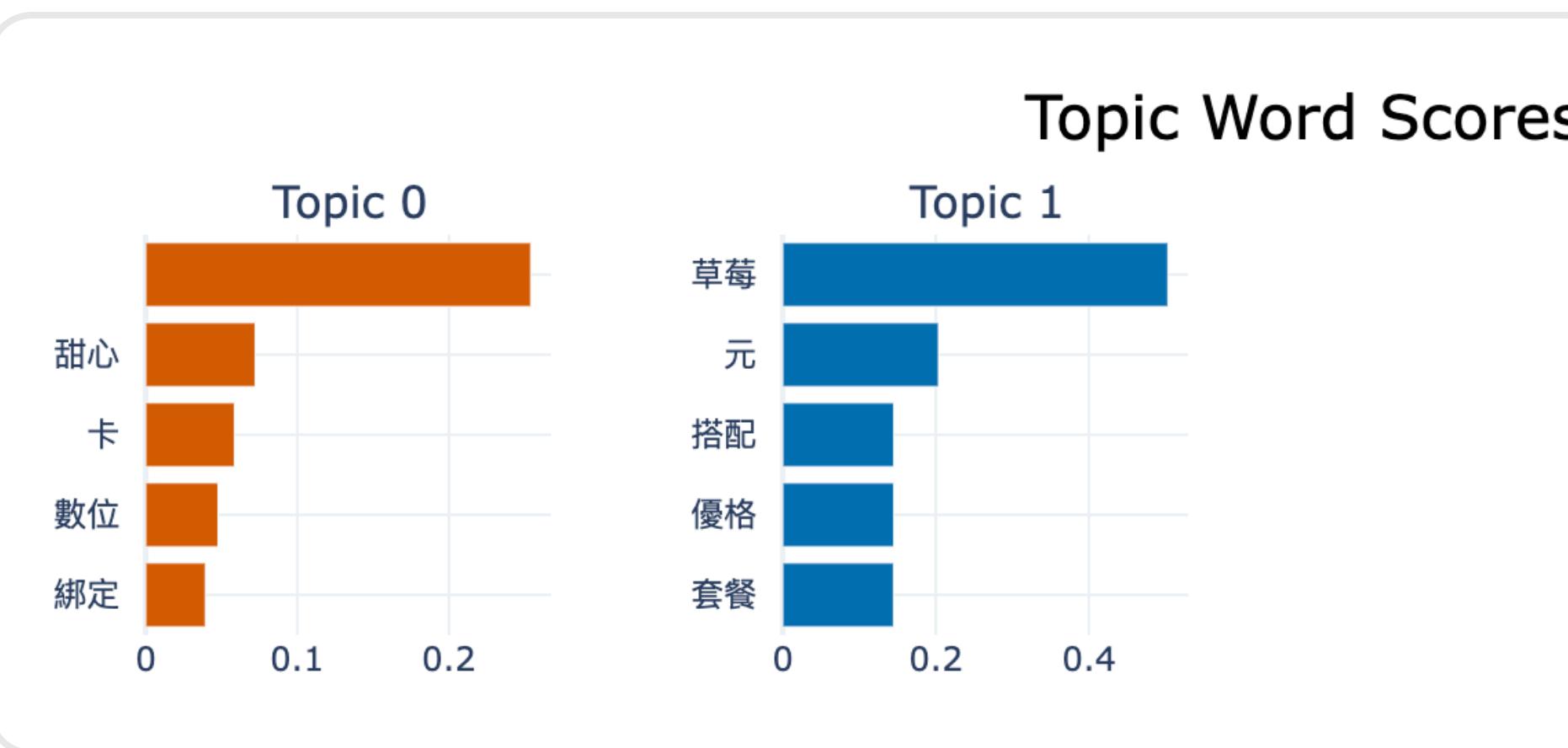
WEEK12 試用不同Embedding model: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Topic	Count	Name	Representation
0	482	0_甜心_卡_數位	[, 甜心, 卡, 數位, 綁定, 優惠, 序號, 漢堡, 優惠券, 香魚, 麥克, 麥,...
1	18	1_草莓_元_搭配_優格	[草莓, 元, 搭配, 優格, 套餐, , 獨家, 雙牛堡, 小薯, 雙餡, 官網, 果醬...

共分出2類主題

- 因目前的主題數量太少，且主題只有一筆資料，導致降維失敗。

用關鍵字做表示：



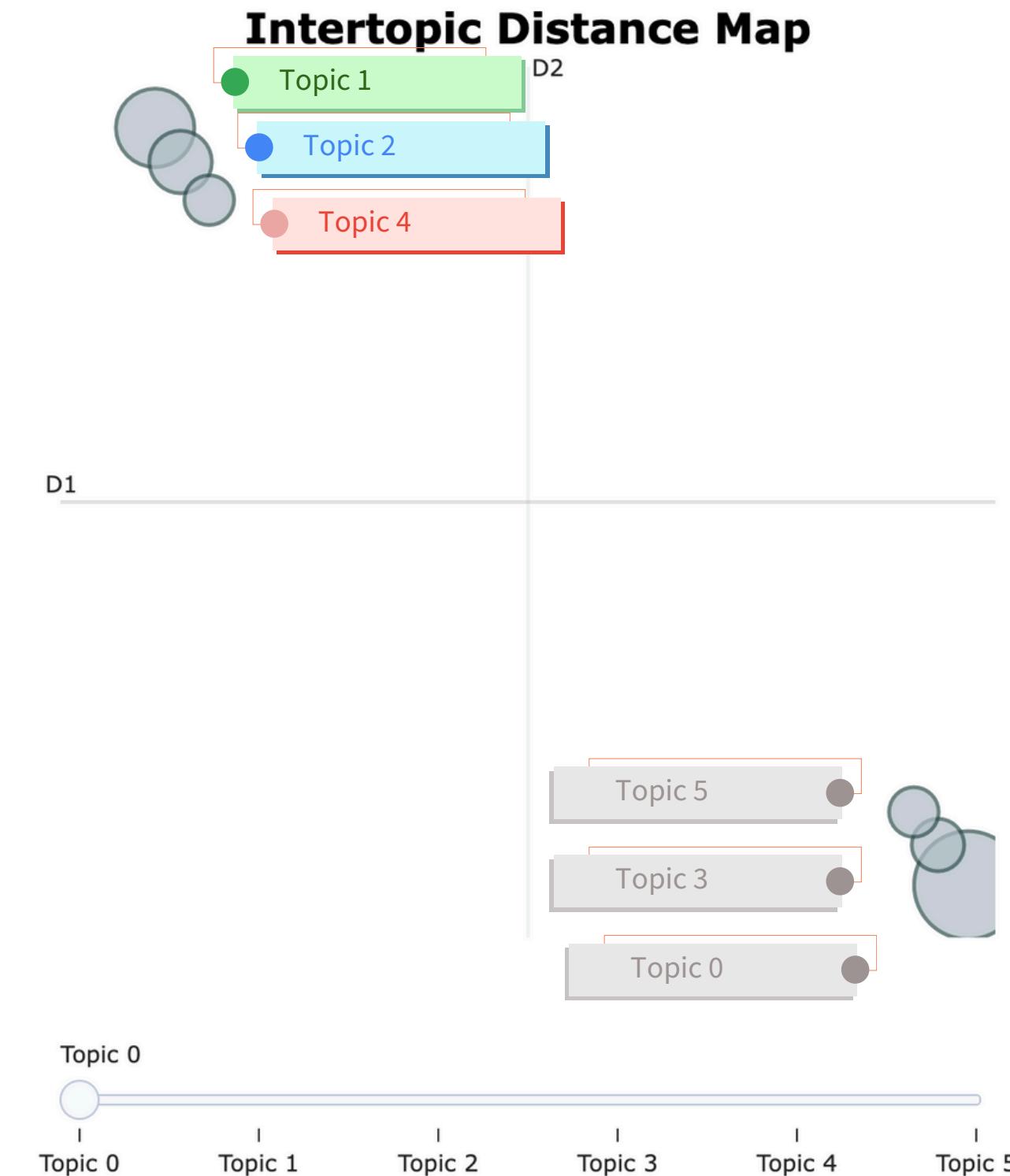
分類結果可見：
大部分都屬於主題一，忽略了資料
中的『社會新聞』相關的主題
分類結果差

WEEK12 試用不同Embedding model: uer/sbert-base-chinese-nli

Topic	Count	Name	Representation
-1	94	-1_序號_數位_綁定	[, 序號, 數位, 綁定, 家樂福, 店, 廠商, 麥克, 套餐, 期限, 每組, 頭份...]
0	161	0_協助_警方_評分	[, 協助, 警方, 評分, 綁定, 性平, 安心, 強化, 成功, 派出所, 職場, 挪...
1	85	1_甜心_卡_優惠	[甜心, 卡, , 優惠, 優惠券, 數位, 序號, 兌換, 實體, 得來, 買, 領取, ...]
2	54	2_漢堡_草莓_香魚_麥	[漢堡, 草莓, 香魚, 麥, 元, , 套餐, 麥克, 紅茶, 薯餅, 薯條, 塊, 優...
3	38	3_越_一場_結論	[, 越, 一場, 結論, 吃, 超級, 說, 社會風氣, 起床, 痛定思痛, 帶來, 退...
4	34	4_大雨_女子_紅髮	[, 大雨, 女子, 紅髮, 行走, 社群, 一名, 中, 李嫌, 二代, 新北女, 點名...
5	34	5_卡_不適_用於	[, 卡, 不適, 用於, 未, 甜心, 雨具, 餐巾紙, 攜帶, 就業網, 數位, 辦法...

共分出6類主題

- 主題四【社會事件】相關因內文有『餐點內容』，因此與主題二有些微重疊
- 右下皆是新聞相關資訊、且新聞主角皆有女性，因此有些微重疊



WEEK12

針對最好的向量模型分類結果，各主題前10關鍵字

因分類皆無非常完善，粗略認定結果最好順序為：

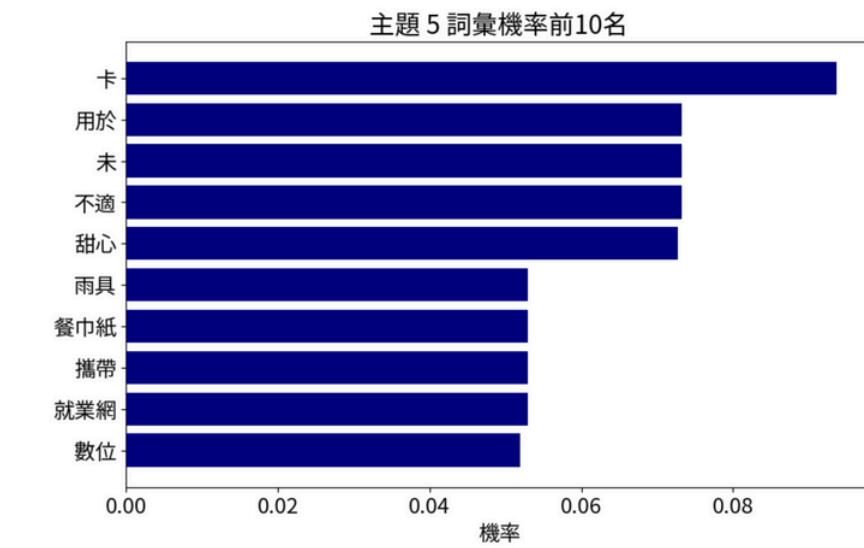
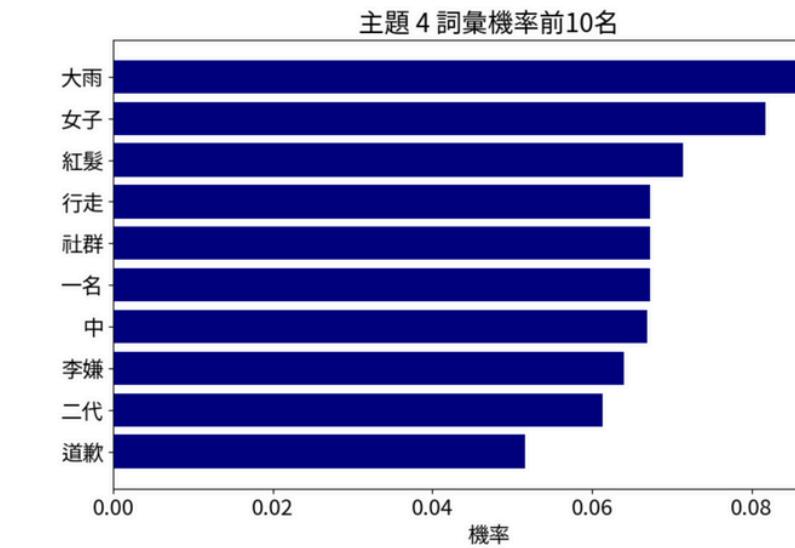
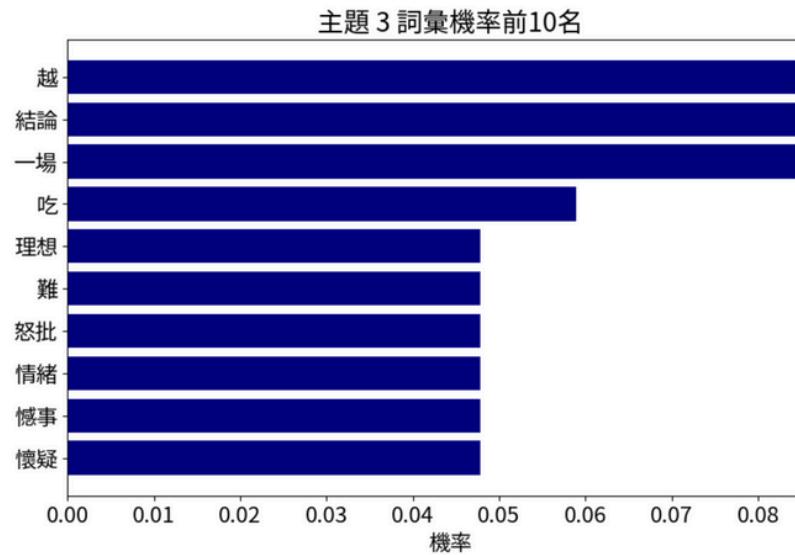
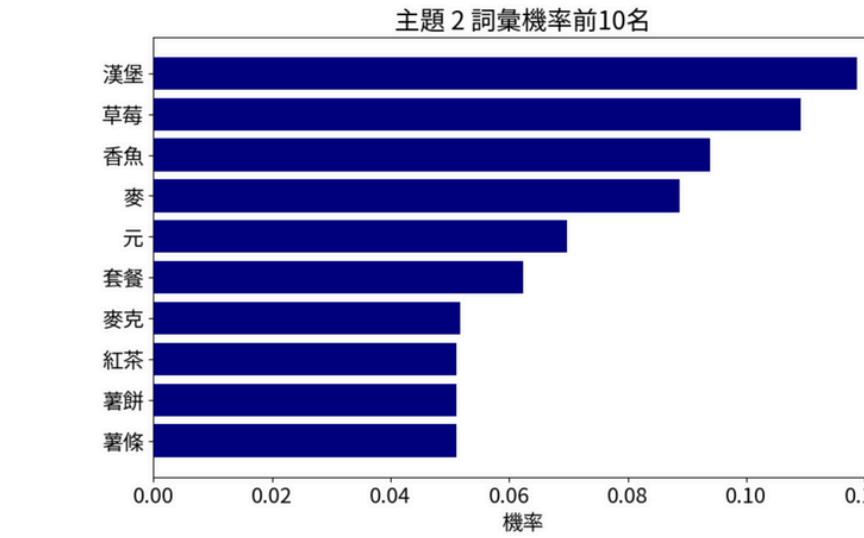
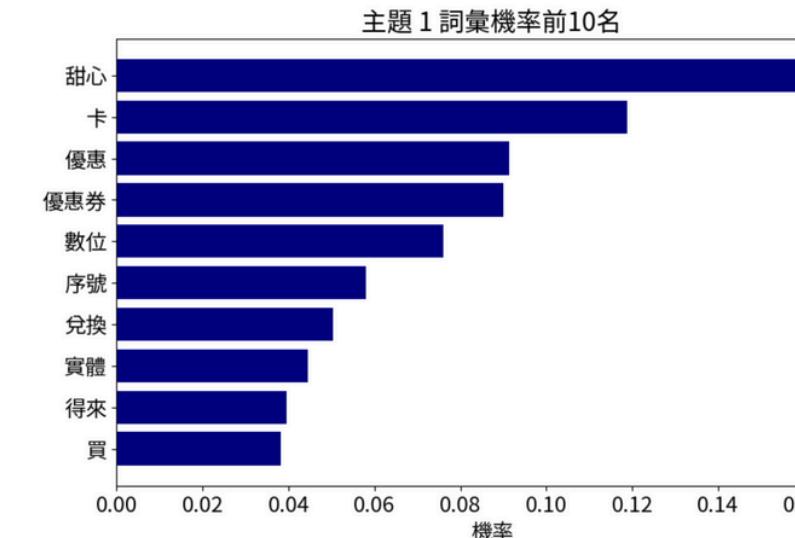
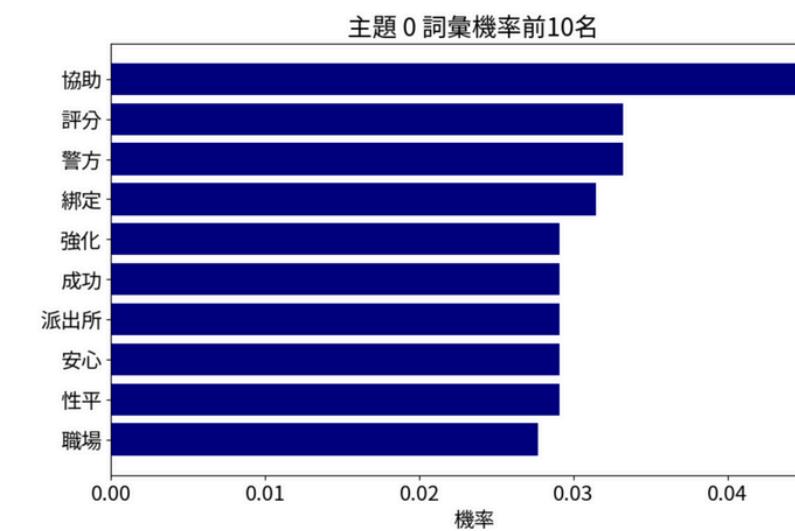
uer/sbert-base-chinese-nli > google-bert/bert-base-chinese > shibing624/text2vec-base-chinese > paraphrase-multilingual-MiniLM-L12-v2

有明確分布，關鍵字具代表性。
雖主題重疊但屬於可接受範圍。

明顯存在主題過於相近 (0,1)、
且部分主題樣本數過少

主題間分群過於密集，
有多個重疊，關鍵字差異度低

主題只有 2 類，其中一類只有 1 篇，
降維直接失敗



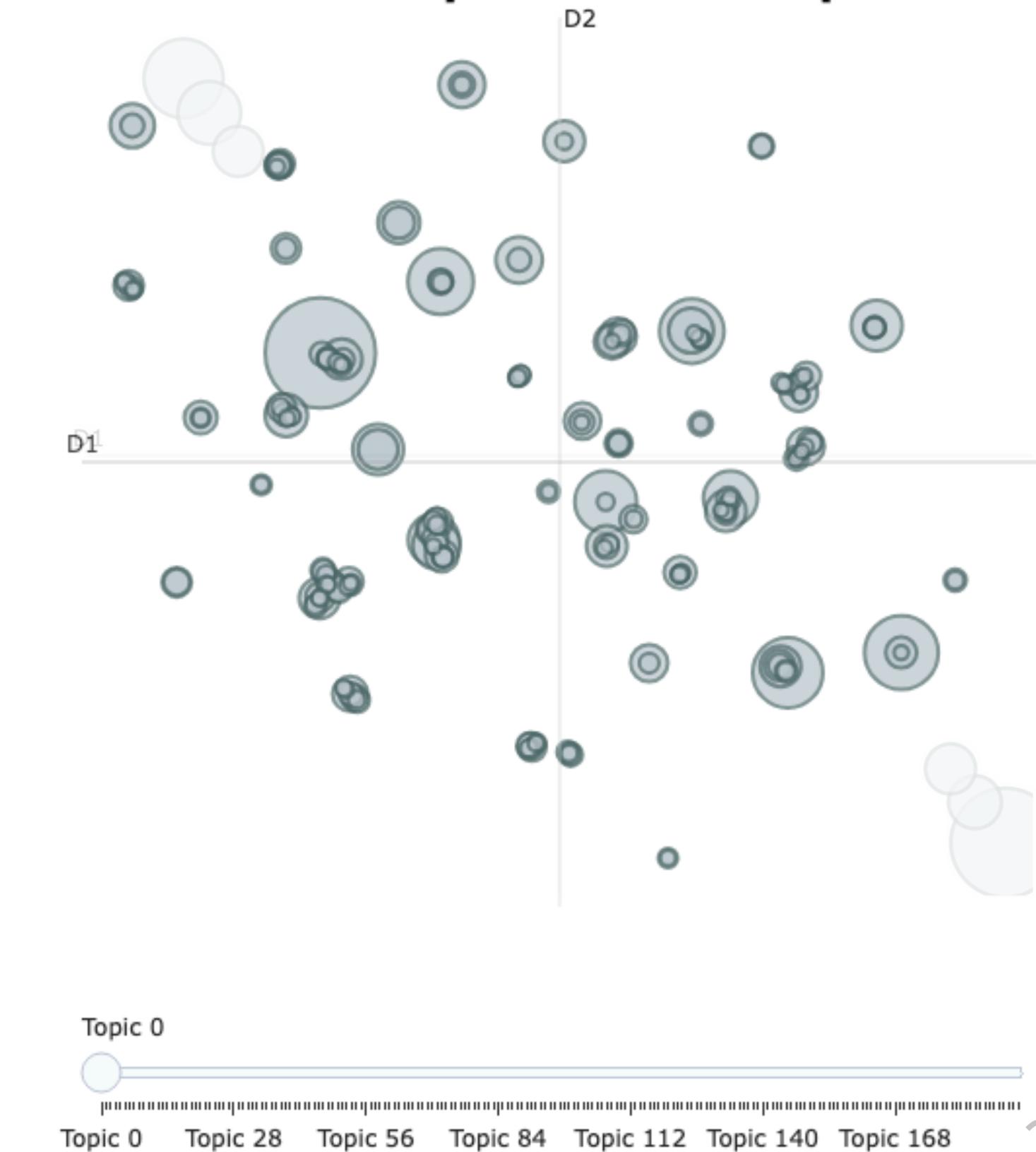
WEEK12

未經前處理的分類結果

如果沒有經過前處理的資料，
BertTopic最後呈現的結果共會有
191個主題

底下為透明圓圈為：經前處理的分類結果

Intertopic Distance Map





Thank You!

