

社群媒體分析第三次讀書會作業

指導老師：黃三益 教授

組別：第六組

組員：

N104020001 李采容

N104020002 廖英捷

N104020007 郭育雯

N104020008 游淑媛

N104020009 蔡雨臻

N104020010 盧貴聰

B096060020 黃湘安

M106020015 林猷盛

目錄

一、分析議題說明	3
二、工作流程設計	3
三、爬蟲、詞頻計算與資料清理	5
四、文件分類	8
五、主題模型	16
六、視覺化	26
七、結論	33

一、分析議題說明

- 主題：聯合新聞網中包含「ChatGPT」文章的文章分類與主題模型建立
- 議題發想：

近年來，人工智慧技術已日益成熟，其中自然語言處理技術更是被廣泛運用。OpenAI 開發的ChatGPT是其中一個非常成功的自然語言處理模型，能夠讓人們與機器之間進行更加自然和流暢的對話。為了進一步研究ChatGPT在網路上引起的效應，本組將運用社群媒體分析課程所學，進行以下三點討論：

1. 建立與ChatGPT議題相關文章之文件分類器
 2. 網路上對於ChatGPT的相關字詞網絡圖有哪些？
 3. 網路上對於ChatGPT的相關主題分類有哪些？
- 使用平台：文字探勘工作流程設計平台

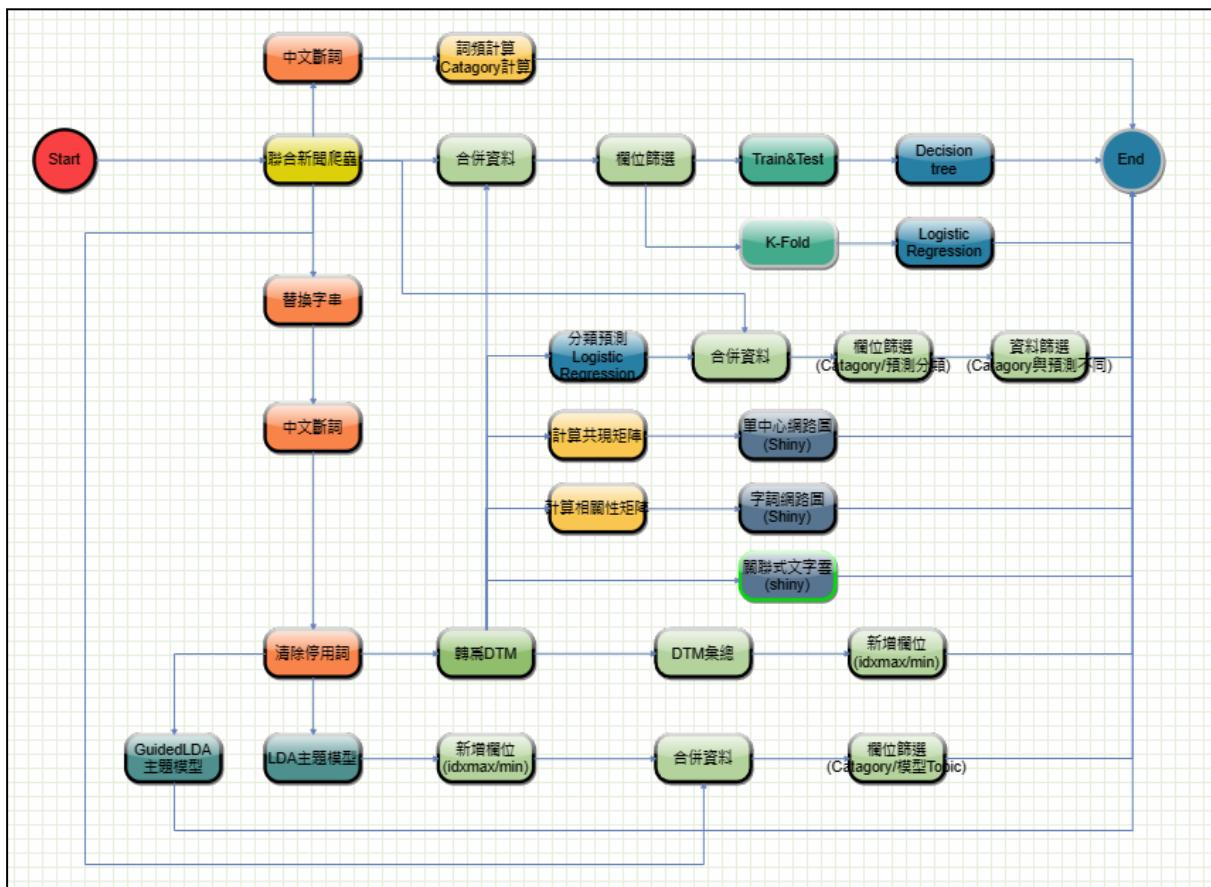
二、工作流程設計

- 工作流程：flow3_N010
- 字典：topicClass_N104020010.csv
- 資料來源：聯合新聞網（文教、股市、生活、產經、數位、其他，共六個版）
- 分析期間：112/01/01 ~ 112/05/06
- 流程概述：
 1. 指定chatGPT相關的關鍵字（GPT、gpt），爬取聯合新聞網今年度至5/6發布之新聞內容。
 2. 進行資料清理，以「替換字串」將語意、主題相近的詞彙替換成同一個字詞，再以「中文斷詞」將新聞內容分解成字詞單位，並使用「清除停用詞」將不

必要的符號、單字元去除，最後設定停用字以過濾出現頻率高但無意義的字詞。

3. 以「中文斷詞」結果中的「catagory」投入「詞頻計算」中，以得出爬蟲結果中六個文章類別的數量。
4. 將清理好的資料轉為DTM，進行文件分類的流程，並將完成的預測模型進行分類預測。
5. 將清理好的資料轉為DTM，計算相關性矩陣與共現矩陣，產生字詞網路圖、單中心網路圖跟關聯式文字雲分析與GPT相關的其他字詞。
6. 將清理好的資料進行LDA主題模型和GuidedLDA主題模型，針對各主題出現之字詞進行分析。

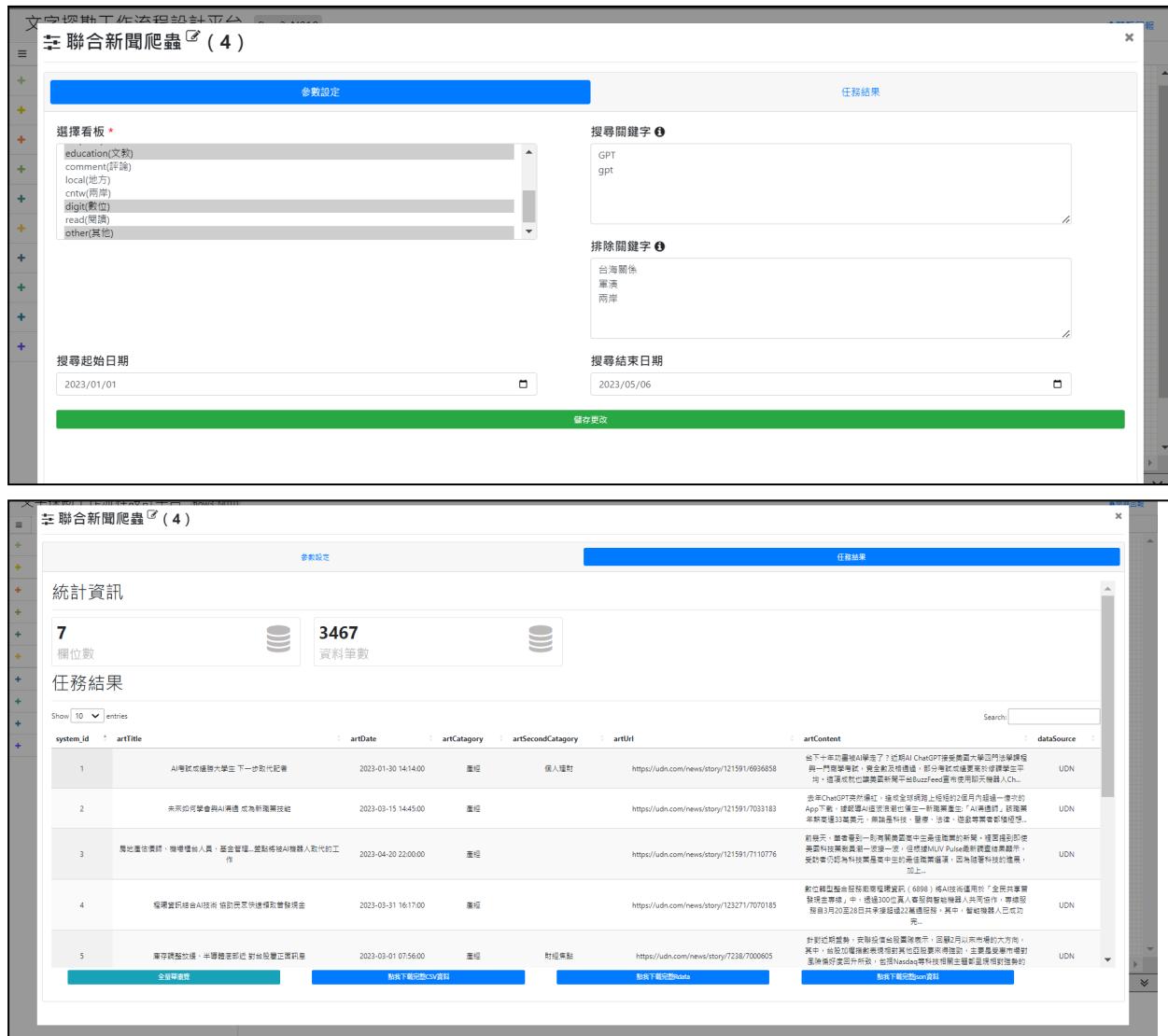
- 總工作流程設計圖一覽：



三、爬蟲、詞頻計算與資料清理

1.聯合新聞網爬蟲

爬取112/01/01 ~ 112/05/06期間，文教、股市、生活、產經、數位、其他，共六個版與GPT相關之聯合新聞網資料，共3467筆資料。



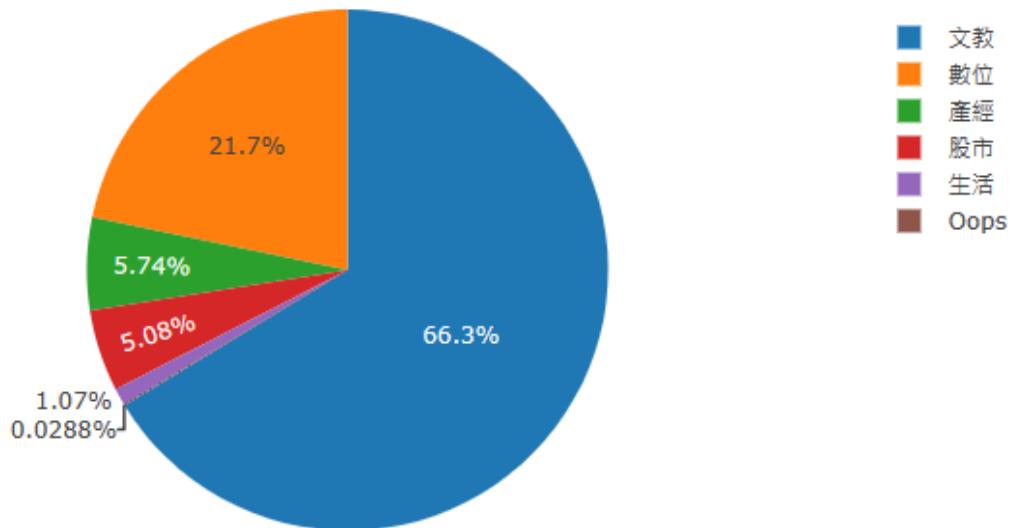
2. 詞頻計算

將資料中的「catagory」欄位進行中文斷詞，並進行詞頻計算，得出各個版面的文章數量，結果如下圖，此可用於後續文件分類結果的檢視。由生成的視覺化圓餅圖可發現文教版文章佔多數(66.3%)。

任務結果

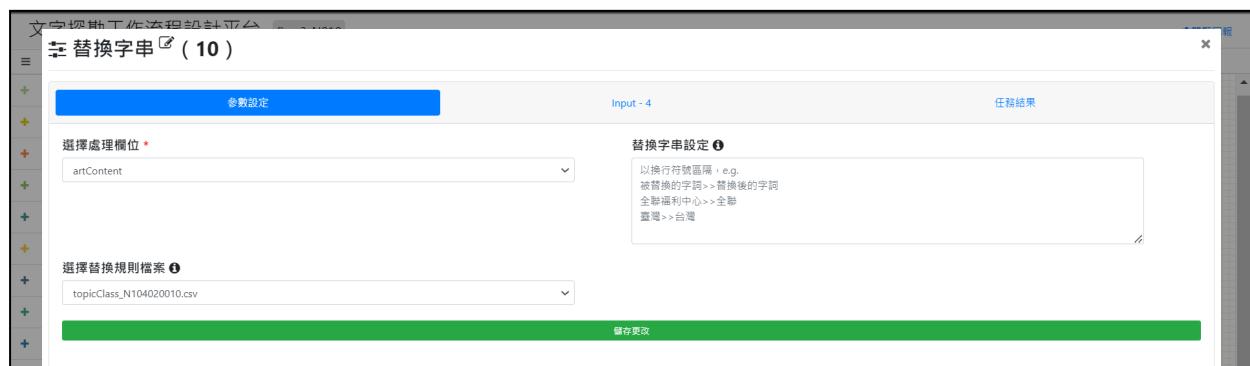
Show	10	entries	Search:
n			Term
	2300		文教
	754		數位
	199		產經
	176		股市
	37		生活
	1		Oops

各版比例圖

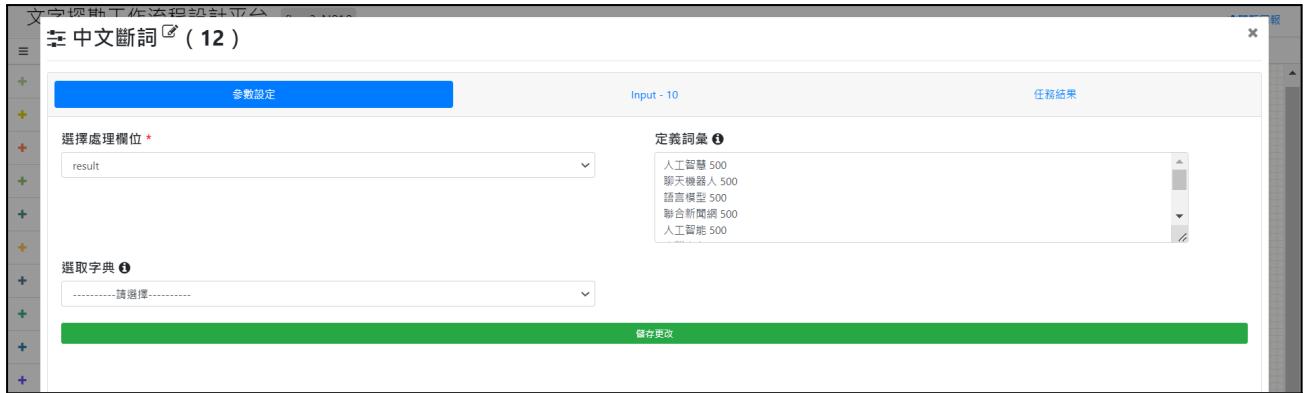


3.資料清理

替換字串 : 使用topicClass_N104020010.csv字典，將相近詞換替換成同一詞彙。

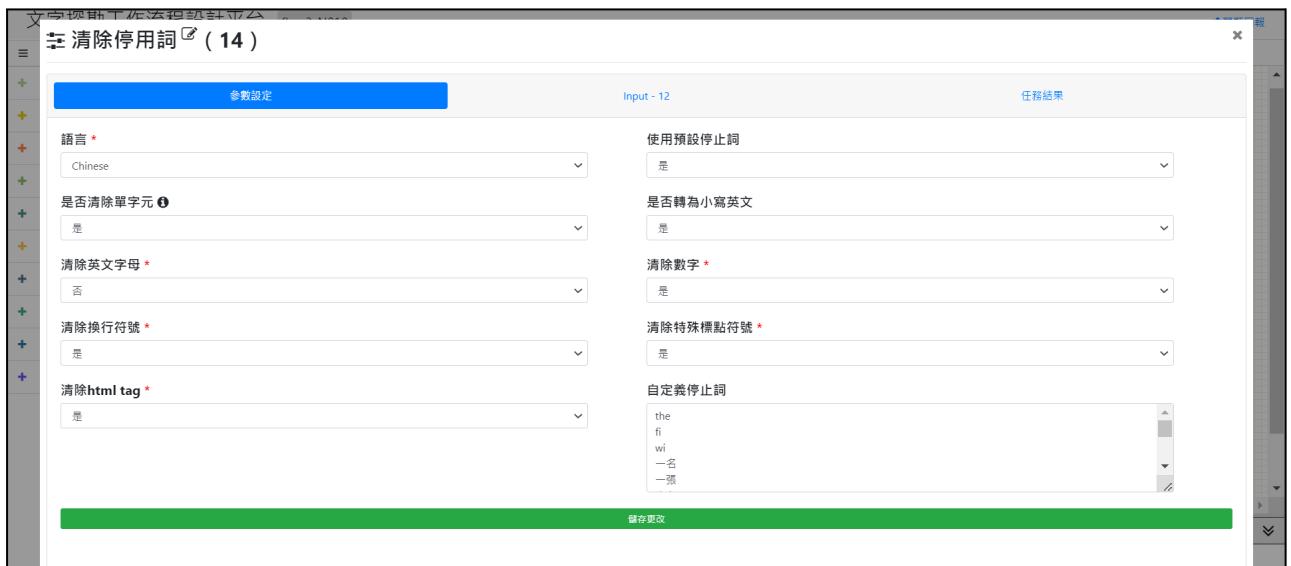


中文斷詞 : 將不應被斷開的字詞設定權重(例如:人工智慧、語言模型、聊天機器人等等)

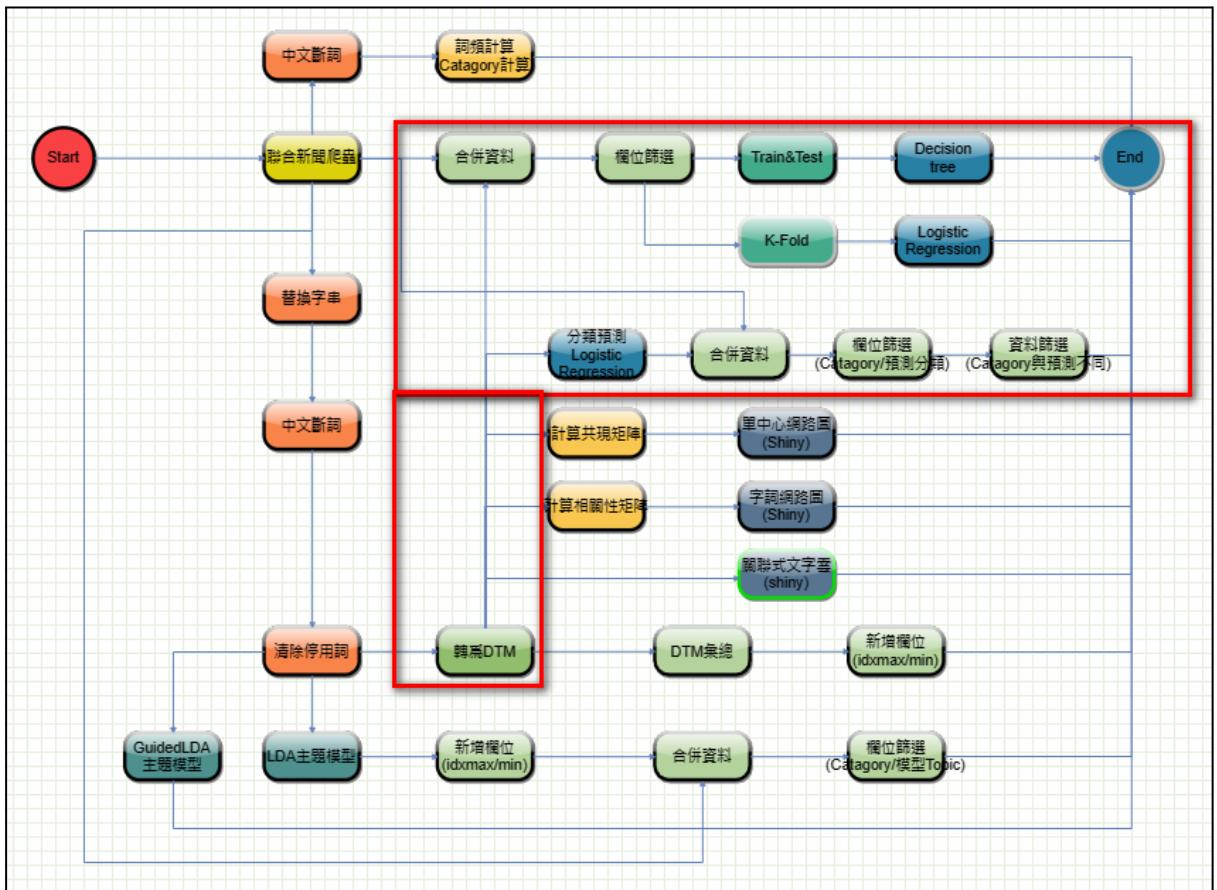


清除停用字: 清除不必要的符號與單字元, 並將頻率高但無意義的字詞設定自定義停止

詞(例如:一張、XD等), 此外因為本次主題為GPT, 清除英文字母的部分設定為否。



四、文件分類



- 流程概述：

1. 將清理後之資料轉為DTM
2. 將DTM與原資料做合併，並保留原資料Catagory欄位與DTM欄位
3. 利用Train&Test與K-Fold模塊切割資料成訓練集與測試集
4. 分別使用Decision Tree 與 Logistic Regression演算法來訓練預測模型
5. 最後再拿原資料來驗證Logistic Regression訓練出來的模型成果
6. 透過資料篩選後可以得知，模型預測率非常高，高達3465/3467

- 流程實作說明：

1. 轉為DTM：最多篩選詞彙數量設定為500，統計每個詞彙出現次數

轉爲DTM (16)

參數設定

Input - 14

任務結果

保留詞彙 ?

以換行符號區隔, e.g.
國立中山大學
西子灣
壽山...

最多篩選詞彙數量 ?

500

儲存更改

2. 合併資料：將DTM之資料與原本的資料集合併

合併資料 (35)

參數設定	Input - 4	Input - 16	任務結果
JOIN規則			
<input style="background-color: #007bff; color: white; border: none; padding: 5px 10px; margin-right: 10px; border-radius: 5px; font-weight: bold; width: fit-content; height: fit-content;" type="button" value="新增規則"/> <input style="background-color: #dc3545; color: white; border: none; padding: 5px 10px; border-radius: 5px; font-weight: bold; width: fit-content; height: fit-content;" type="button" value="刪除規則"/>			
left_key	right_key		
system_id	system_id		
-----請選擇-----	-----請選擇-----		
<input style="background-color: #007bff; color: white; border: none; padding: 10px 20px; border-radius: 5px; font-weight: bold; width: fit-content; height: fit-content;" type="button" value="儲存更改"/>			

參數設定		Input - 4				Input - 16				任務結果			
system_id	artTitle	artDate	artCategory	artSecondCatagory	artUrl	artContent	dataSource	ai	ar	udn	7.0		
1	AI考試成績勝大學生下一步取代記者	2023-01-30 14:14:00	產經	個人理財	https://udn.com/news/story/121591/6936858	台下十年功盡被AI學走了？近期AI ChatGPT接受美國大學四門法學課程與一門商學考試，竟全數及格通過，部分考試成績更高於修課學生平均。這項成就也讓泰國新聞平台	UDN	7.0					

3. 欄位篩選: 篩選DTM之所有詞彙與文章分類之欄位

欄位篩選 (47)

參數設定

Input - 35

任務結果

選擇要保留的欄位(按住ctrl(Windows)或command(MAC)可以複選)*

- artDate
- artCategory
- artSecondCategory
- artUrl
- artContent
- dataSource
- ai

儲存更改

參數設定																Input - 35		
system_id	artCatagory	ai	and	app	apple	bic	bing	camera	chatgpt	esim	google	gpt	iphone	line	of	op		
1	產經	7.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
1	產經	7.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
2	產經	19.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0		
2	產經	19.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0		
3	產經	37.0	1.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0		
3	產經	37.0	1.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0		
4	產經	3.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		

4. Tran&Test: 將篩選後之資料切割並進行訓練

Train&Test (49)

參數設定

Input - 47

任務結果

目標欄位 *

測試資料切割比率 *

artCategory

0.2

是否隨機排序資料

亂數種子

是

487

儲存更改

Train&Test (49)		參數設定	Input - 47	任務結果
資料切割資訊				
Show <input type="button" value="10"/> entries Search: <input type="text"/>				
_id	system_id	train_idx	test_idx	target_column result
64573ebc72bc01d565c2ad1f	1	[5369, 4374, 2722, 6683, 6576, 154, 1150, 107, 1961, 3762, 3781, 6145, 590, 3434, 5837, 6157, 3615, ...]	[5901, 144, 5136, 2591, 2114, 2927, 889, 1403, 5050, 2041, 2803, 2794, 1681, 661, 6586, 1959, 658, 1...]	artCatagory []

5. Decision tree: 訓練後之模型準確度、精確度與召回率都高於80%

Decision tree (55)

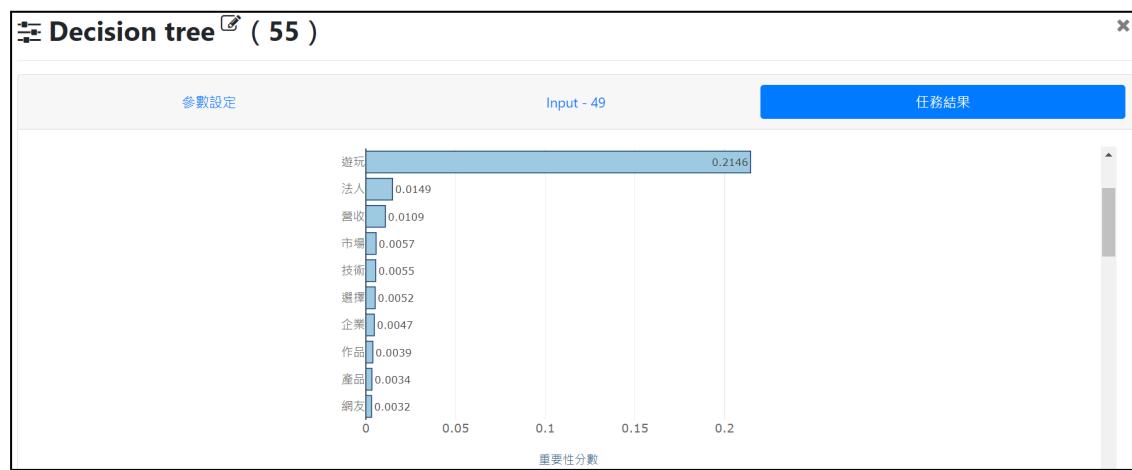
參數設定

Input - 49

任務結果

計算方式 *	gini	最大深度
最小分割條件樣本數 *	2	葉節點最小樣本數 *
最大特徵選用數		隨機種子
最大頁節點數		

儲存更改





6. K-Fold: 將篩選後之資料切割並進行第二種訓練

K-Fold (51) 參數有做更動，建議重新執行

參數設定	Input - 47	任務結果
目標欄位 *	artCatagory	分割數 5
是否隨機排序資料	是	亂數種子 487
儲存更改		

K-Fold (51) 參數有做更動，建議重新執行

參數設定	Input - 47	任務結果																										
<h3>資料切割資訊</h3> <table border="1"> <thead> <tr> <th>Show 10 entries</th> <th>Search:</th> </tr> </thead> <tbody> <tr> <th>_id</th> <th>system_id</th> <th>train_idx</th> <th>test_idx</th> <th>target_column</th> <th>result</th> </tr> <tr> <td>64573ebc72bc01d565c2ad1e</td> <td>1</td> <td>[0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 14, 15, 18, 19, 21, 25, 26, 27, 28, 30, 31, 32, 33, 34, 36, 37, ...]</td> <td>[1, 11, 12, 16, 17, 20, 22, 23, 24, 29, 35, 40, 42, 47, 49, 52, 53, 54, 63, 75, 79, 80, 82, 86, 88, ...]</td> <td>artCatagory</td> <td>[]</td> </tr> <tr> <td>64573ebc72bc01d565c2ad20</td> <td>2</td> <td>[1, 5, 6, 7, 9, 10, 11, 12, 14, 16, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 34, 35, 36, ...]</td> <td>[0, 2, 3, 4, 8, 13, 15, 18, 31, 32, 33, 37, 38, 43, 46, 48, 51, 56, 57, 73, 74, 87, 89, 93, 99, 103, ...]</td> <td>artCatagory</td> <td>[]</td> </tr> <tr> <td>64573ebc72bc01d565c2ad22</td> <td>3</td> <td>[0, 1, 2, 3, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 29, 30, 31, 32, 33, ...]</td> <td>[5, 9, 19, 25, 26, 27, 28, 34, 41, 59, 60, 67, 70, 77, 81, 85, 90, 94, 96, 98, 100, 114, 116, 119, 1...]</td> <td>artCatagory</td> <td>[]</td> </tr> </tbody> </table>			Show 10 entries	Search:	_id	system_id	train_idx	test_idx	target_column	result	64573ebc72bc01d565c2ad1e	1	[0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 14, 15, 18, 19, 21, 25, 26, 27, 28, 30, 31, 32, 33, 34, 36, 37, ...]	[1, 11, 12, 16, 17, 20, 22, 23, 24, 29, 35, 40, 42, 47, 49, 52, 53, 54, 63, 75, 79, 80, 82, 86, 88, ...]	artCatagory	[]	64573ebc72bc01d565c2ad20	2	[1, 5, 6, 7, 9, 10, 11, 12, 14, 16, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 34, 35, 36, ...]	[0, 2, 3, 4, 8, 13, 15, 18, 31, 32, 33, 37, 38, 43, 46, 48, 51, 56, 57, 73, 74, 87, 89, 93, 99, 103, ...]	artCatagory	[]	64573ebc72bc01d565c2ad22	3	[0, 1, 2, 3, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 29, 30, 31, 32, 33, ...]	[5, 9, 19, 25, 26, 27, 28, 34, 41, 59, 60, 67, 70, 77, 81, 85, 90, 94, 96, 98, 100, 114, 116, 119, 1...]	artCatagory	[]
Show 10 entries	Search:																											
_id	system_id	train_idx	test_idx	target_column	result																							
64573ebc72bc01d565c2ad1e	1	[0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 14, 15, 18, 19, 21, 25, 26, 27, 28, 30, 31, 32, 33, 34, 36, 37, ...]	[1, 11, 12, 16, 17, 20, 22, 23, 24, 29, 35, 40, 42, 47, 49, 52, 53, 54, 63, 75, 79, 80, 82, 86, 88, ...]	artCatagory	[]																							
64573ebc72bc01d565c2ad20	2	[1, 5, 6, 7, 9, 10, 11, 12, 14, 16, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 34, 35, 36, ...]	[0, 2, 3, 4, 8, 13, 15, 18, 31, 32, 33, 37, 38, 43, 46, 48, 51, 56, 57, 73, 74, 87, 89, 93, 99, 103, ...]	artCatagory	[]																							
64573ebc72bc01d565c2ad22	3	[0, 1, 2, 3, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 29, 30, 31, 32, 33, ...]	[5, 9, 19, 25, 26, 27, 28, 34, 41, 59, 60, 67, 70, 77, 81, 85, 90, 94, 96, 98, 100, 114, 116, 119, 1...]	artCatagory	[]																							

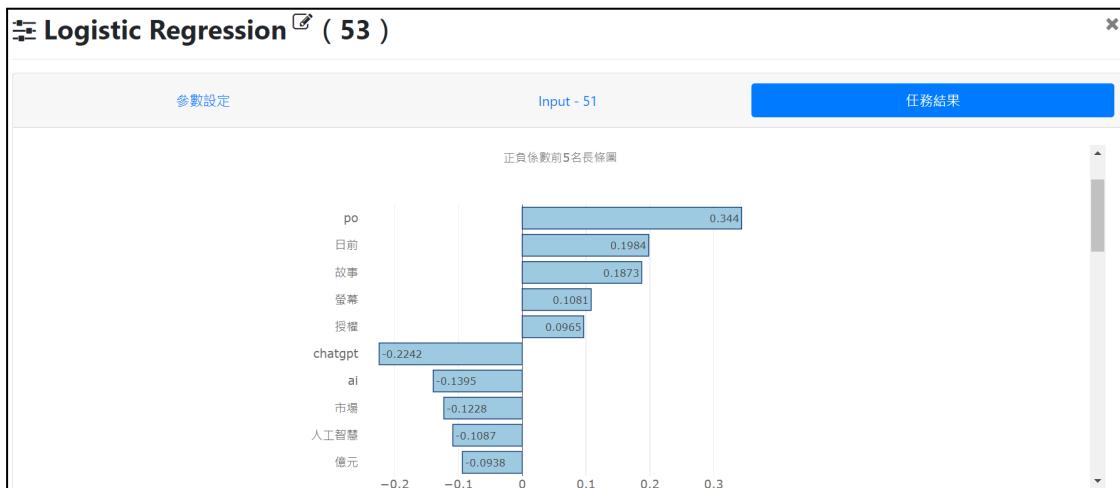
K-Fold (51) 參數有做更動，建議重新執行

參數設定		Input - 47															任務結果	
system_id	artCatagory	ai	and	app	apple	bic	bing	camera	chatgpt	esim	google	gpt	iphone	line	of	oi		
1	產經	7.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
1	產經	7.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
2	產經	19.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0		
2	產經	19.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0		
3	產經	37.0	1.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0		
3	產經	37.0	1.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0		
4	產經	3.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		

7. Logistic Regression: 訓練後之模型準確度、精確度與召回率都高於80%

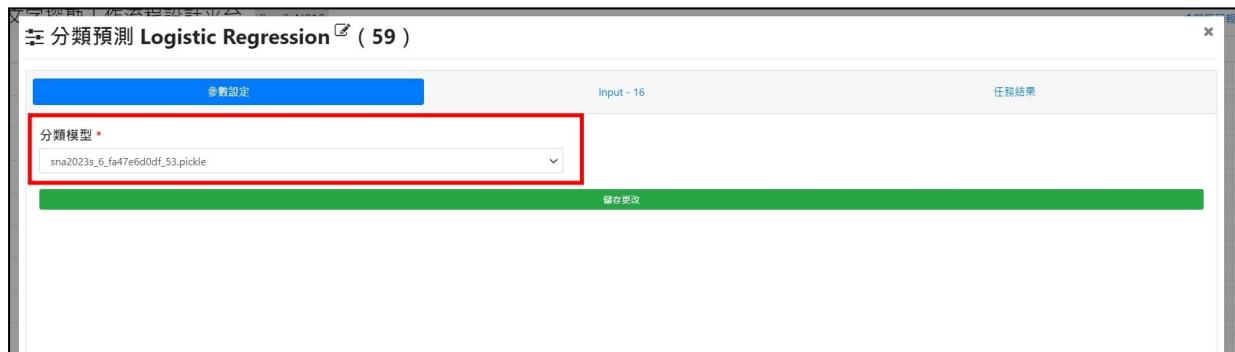
Logistic Regression (53)

參數設定		Input - 51															任務結果	
正規化懲罰	I2	懲罰程度(0<C<=1, 越小懲罰越多)															1	
隨機種子	487																儲存更改	

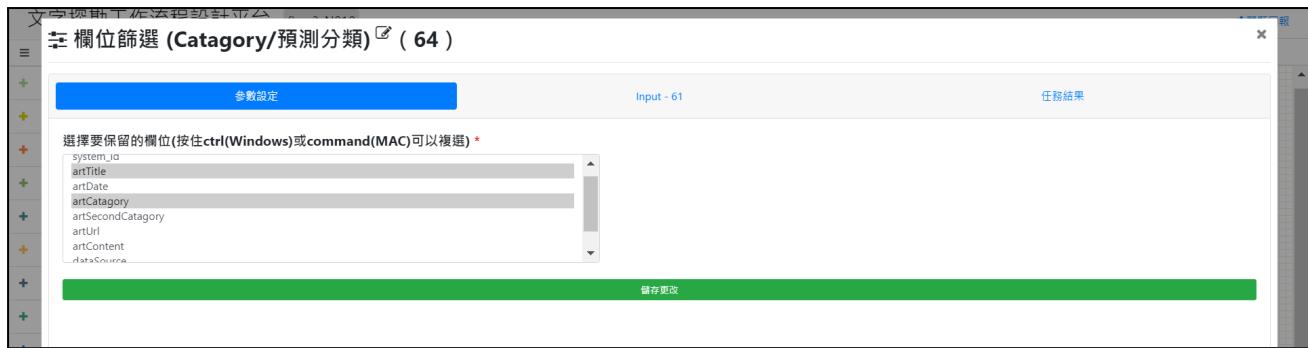
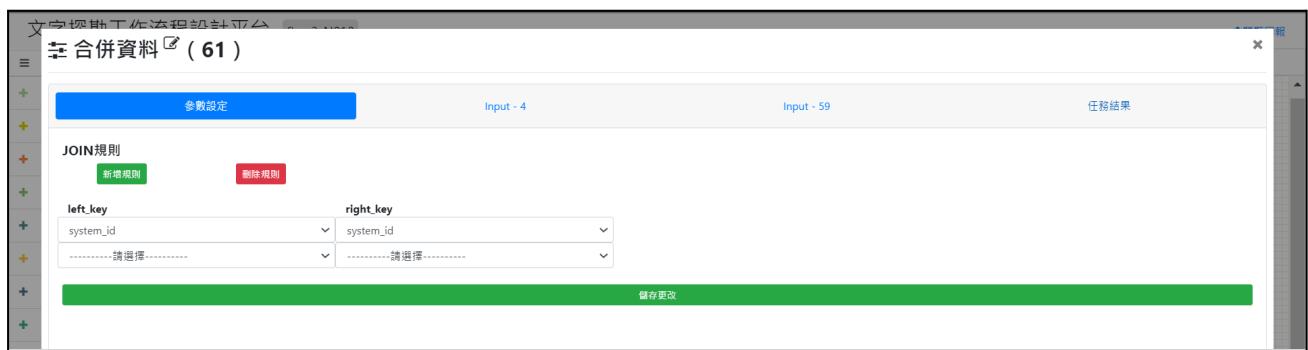




完成文件分類模型之後，將模型進行分類預測：



將預測之結果與原資料合併，並進行欄位篩選(選取文章標題、實際的文章分類與預測的文章分類三個欄位)：



進行資料篩選，篩選出實際與預測之分類不同之資料，得出結果為兩筆，由此可知預測之模型準確率非常高，高達3465/3467。

The screenshot shows a data filtering interface with the following details:

Top Panel: A blue header bar labeled "參數設定" (Parameter Settings) with a dropdown menu "Input - 64". On the right, there is a "任務結果" (Task Result) section.

Condition Input: A text input field containing the condition: "\$artCategory != \$result_ClassifierPrediction".

Bottom Panel: A green progress bar labeled "儲存更改" (Save Changes).

Second Screenshot: A detailed view of the task results.

Section Headers: "統計資訊" (Statistics) and "任務結果" (Task Results).

Statistics: Shows "3465 移除數量" (Deleted Count) and "2 保留數量" (Kept Count).

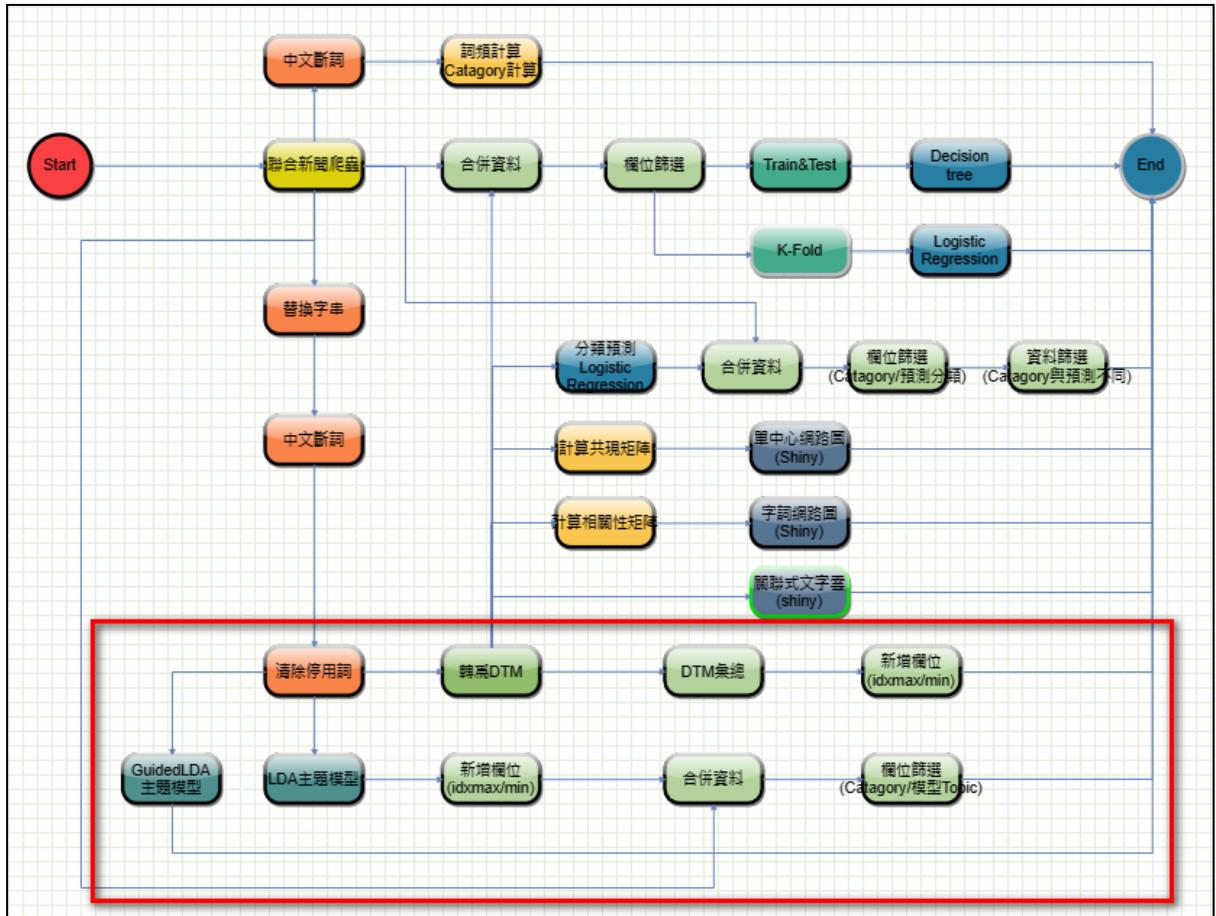
Task Results Table:

system_id	artTitle	artCategory	result_ClassifierPrediction	result
153	股王信託：市況沒想像中樂觀 趨下半年業績回溫	產經	股市	_selected
162	國巨報喜 全年喊賺四股本	產經	股市	_selected

Table Footer: Shows "Showing 1 to 2 of 2 entries", "Previous", "1", and "Next".

Bottom Buttons: "全螢幕瀏覽" (Full Screen View), "點我下載完整CSV資料" (Download Complete CSV Data), "點我下載完整Rdata" (Download Complete Rdata), and "點我下載完整json資料" (Download Complete json Data).

五、主題模型



- 流程概述：

1. 清除停用詞：

- a. 「清除英文字母」設定為「否」
- b. 使用預設停止詞，置換英文字母為小寫
- c. 清除單字元、英文字母、換行符號、html tag、數字、特殊標點符號
- d. 本組並自定義停止詞，將一些文章常用字詞清除

清除停用詞 (14)

參數設定	Input - 12	任務結果
語言 *	Chinese	使用預設停止詞
是否清除單字元 ⓘ	是	是否轉為小寫英文
清除英文字母 *	否	清除數字 *
清除換行符號 *	是	清除特殊標點符號 *
清除html tag *	是	自定義停止詞
		the fi wi —名 —張

2. 轉為DTM

a. 清除停用字後，將結果轉為DTM矩陣，以便做後續分析應用，本組

保留500個篩選詞彙

轉為DTM (16)

參數設定	Input - 14	任務結果
保留詞彙 ⓘ 以換行符號區隔，e.g. 國立中山大學 西子灣 壽山...	最多篩選詞彙數量 ⓘ 500	

b. 取得各種詞彙在文本出現次數，結果如下：

system_id	ai	and	app	apple	bic	bing	camera	chatgpt	esim	google	gpt	iphone	line	of	openai	po	pro	vr
1	7.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2	19.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0
3	37.0	1.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0
4	3.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
6	1.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
7	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
8	1.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

3. DTM彙總

- a. 將事先上傳的類別字典檔「topicClass_N104020010.csv」作為匯總字典

DTM彙總 (18)

參數設定	Input - 16	任務結果
選擇匯總字典 *	topicClass_N104020010.csv	

- b. 運用字典檔，可以有效將DTM矩陣，其相關詞彙對應為某一主題(產業、教育、答題、全球、生活、技術)

system_id	產業	教育	答題	全球	生活	技術
1	8.0	14.0	8.0	6.0	7.0	6.0
2	43.0	2.0	2.0	5.0	19.0	5.0
3	100.0	4.0	6.0	6.0	38.0	30.0
4	21.0	0.0	3.0	1.0	23.0	15.0
5	6.0	0.0	0.0	2.0	1.0	0.0
6	28.0	6.0	2.0	21.0	8.0	12.0
7	6.0	0.0	0.0	2.0	1.0	2.0

4. 新增欄位(idxmax/min)(89)

- a. 為有效歸類每篇文本對應某個主題，本組利用函數「max」擷取自定義的六個主題(產業、教育、答題、全球、生活、技術)中最大數值，作為該文本的主題

匯總函數 * ⓘ

max

計算欄位(按住ctrl(Windows)或command(MAC)可以複選) *

system_id

產業
教育
答題
全球
生活
技術

新增的欄位名稱 *

max_topic

b. 利用新增欄位「max_topic」來儲存主題

system_id	產業	教育	答題	全球	生活	技術	max_topic
1	8.0	14.0	8.0	6.0	7.0	6.0	教育
2	43.0	2.0	2.0	5.0	19.0	5.0	產業
3	100.0	4.0	6.0	6.0	38.0	30.0	產業
4	21.0	0.0	3.0	1.0	23.0	15.0	生活
5	6.0	0.0	0.0	2.0	1.0	0.0	產業
6	28.0	6.0	2.0	21.0	8.0	12.0	產業
7	6.0	0.0	0.0	2.0	1.0	2.0	產業
8	27.0	3.0	3.0	13.0	8.0	18.0	產業
9	4.0	0.0	1.0	11.0	2.0	1.0	全球
10	14.0	14.0	3.0	3.0	17.0	8.0	生活

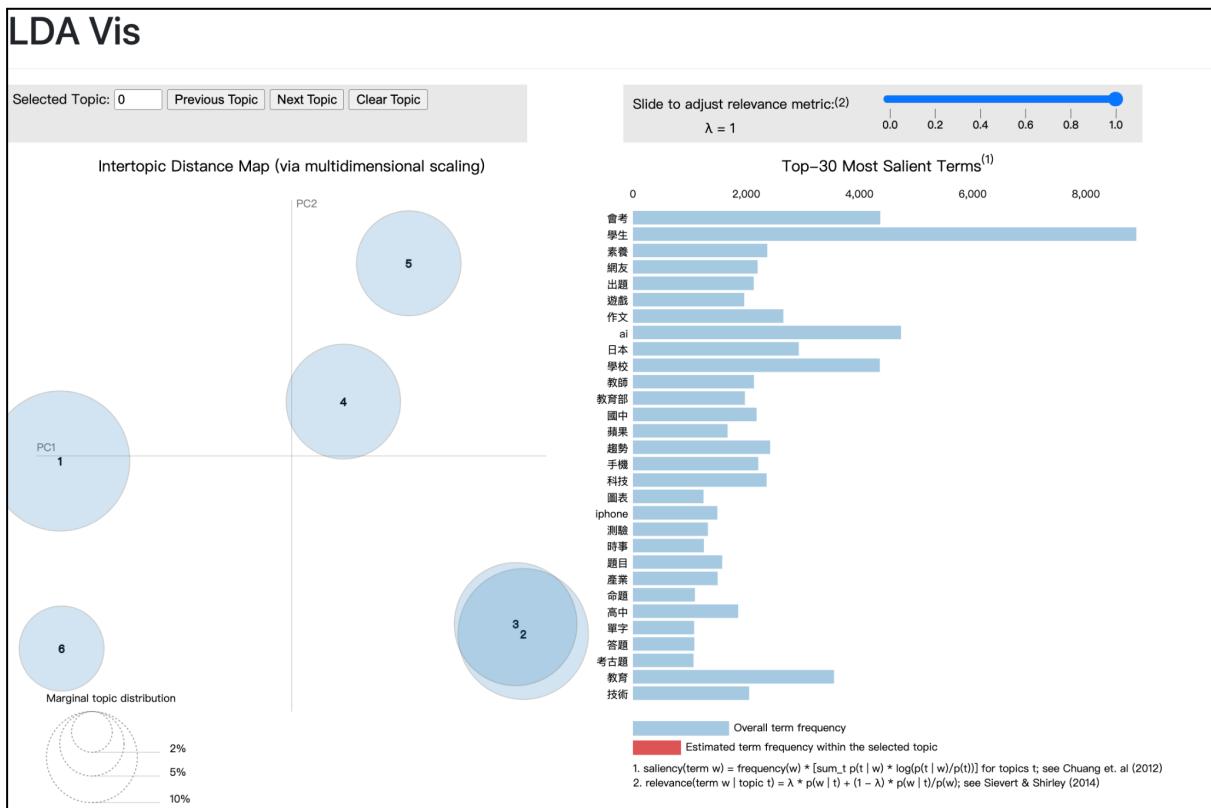
5. LDA模型

- a. 將清理後之詞彙利用LDA主題模型模組做主題隸屬機率劃分。
- b. 取得各則文章分屬最高機率之主題後與原始資料作合併並使用欄位篩選保留catagory與機率最高之主題，確認catagory與模型預測之主題是否相似。
- c. 設定主題數為6個；迭代次數為150次；保留關鍵字20個。
- d. 設定詞彙頻率下限為20，該詞彙最少要出現在20篇文章中。
- e. 設定詞彙頻率上限為0.6，在所有文章中高於0.6者將排除計算。

LDA主題模型 (20)

參數設定	Input - 14	任務結果	
目標欄位 *	result	迭代次數	150
主題數 *	6	主題保留關鍵字數量	20
詞彙頻率下限 ⓘ	20	詞彙頻率上限 ⓘ	0.6
alpha	預設為主題數/50	Beta	預設為0.1
chuckszie ⓘ	預設為2000	update_every ⓘ	1
是否輸出字典	是		

f. LDA主題連貫性結果數值如下, 6個主題的相符程度排行: Topic 1 (23.5%) > Topic 2 (20.5%) > Topic 3 (18.2%) > Topic 4 (15.8%) > Topic 5 (13.2%) > Topic 6 (8.7%)

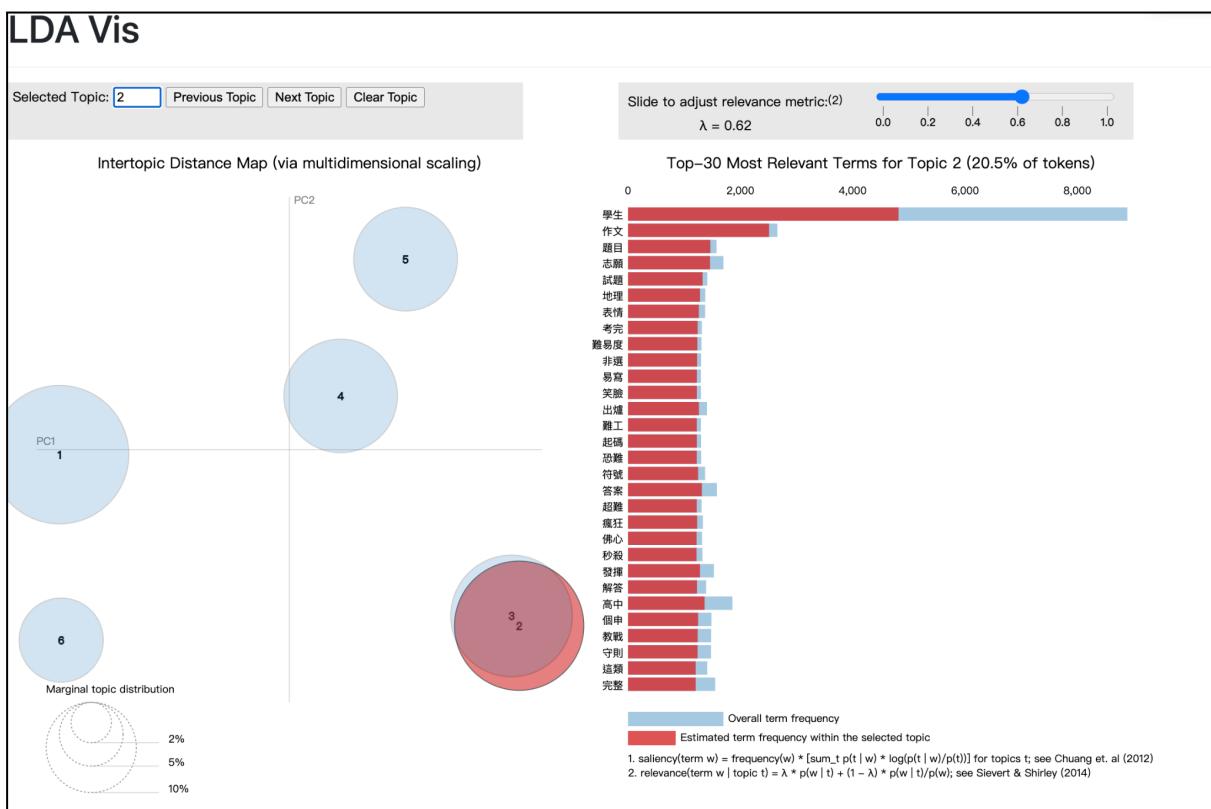
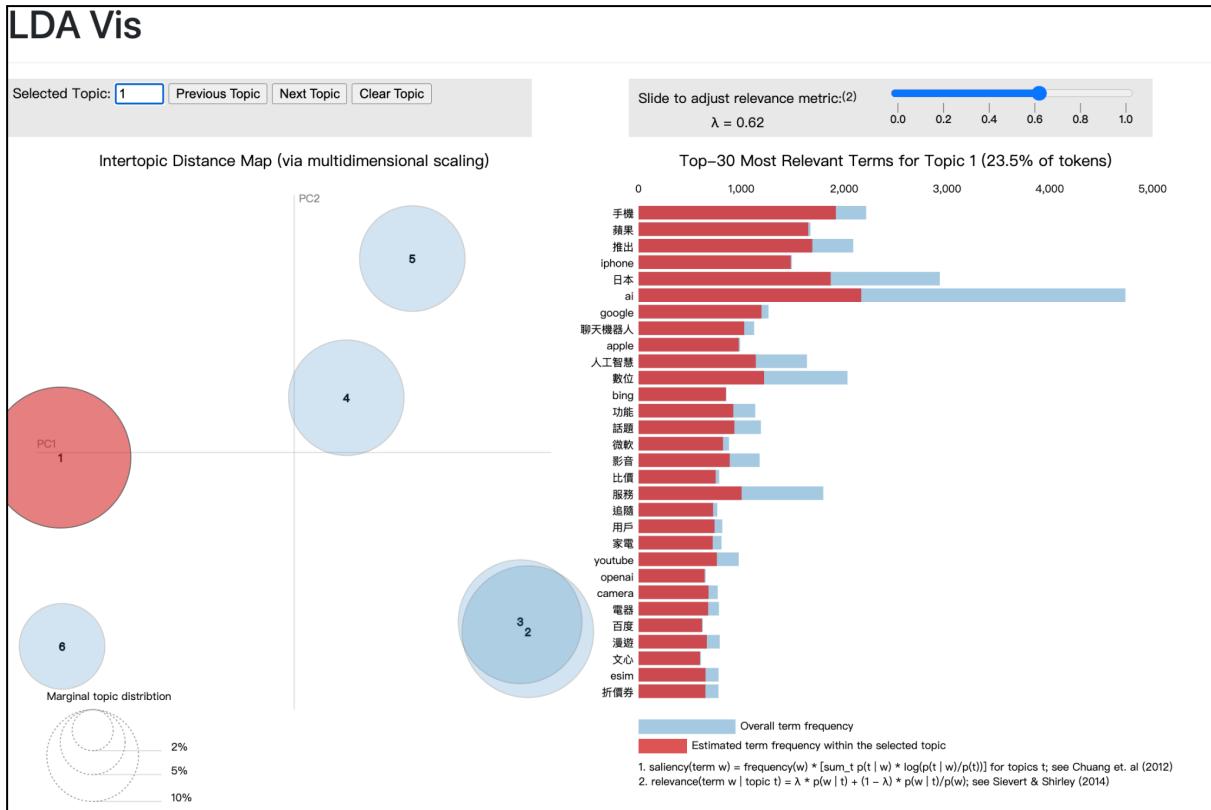


g. 主題模型結果, 以主題連貫性的頻率來看, 各主題前三大詞彙:

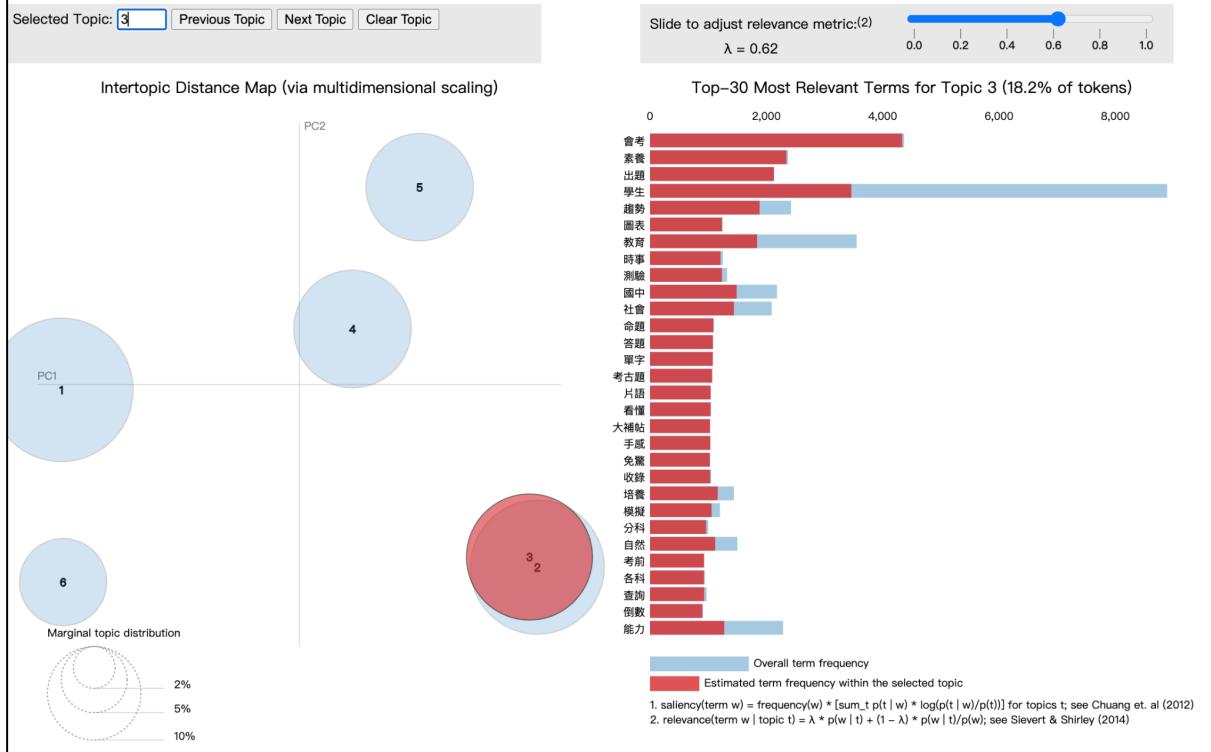
- 主題1: ai > 手機 > 日本
- 主題2: 學生 > 作文 > 題目
- 主題3: 會考 > 學生 > 素養
- 主題4: ai > 台灣 > 科技

- 主題5:教師 > 學校 > 教育部

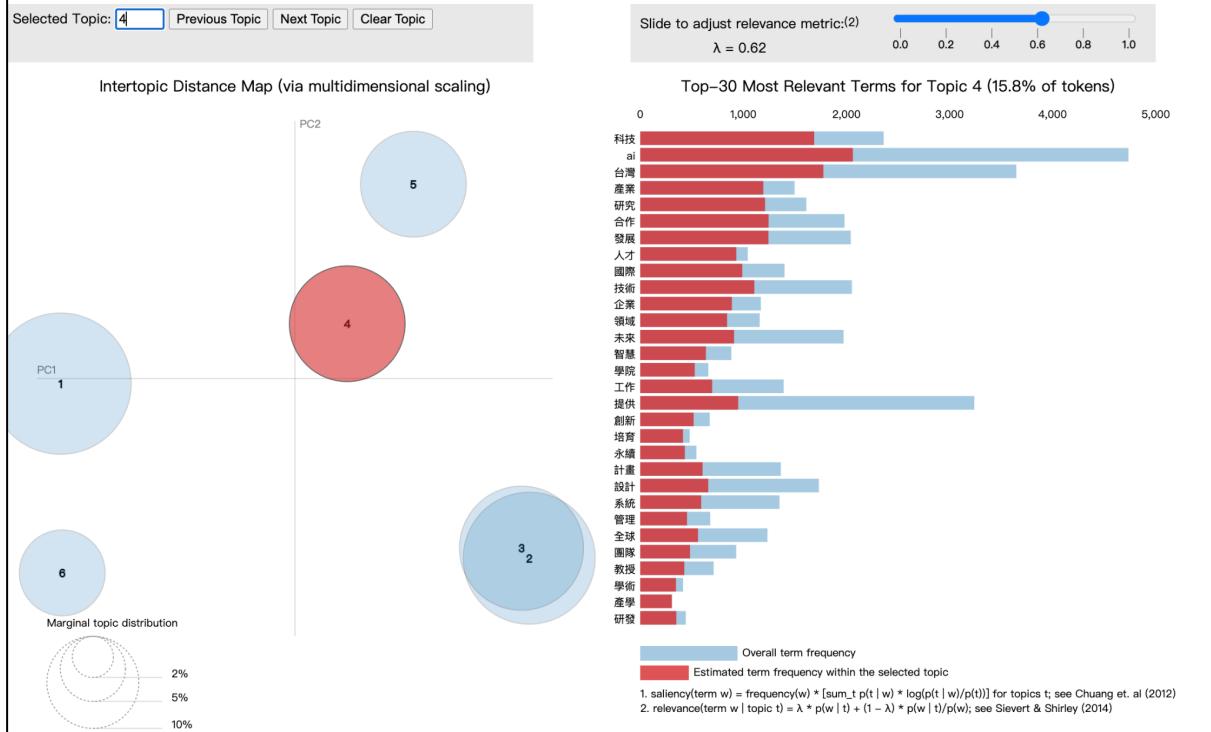
- 主題6:遊戲 > 網友 > 日本



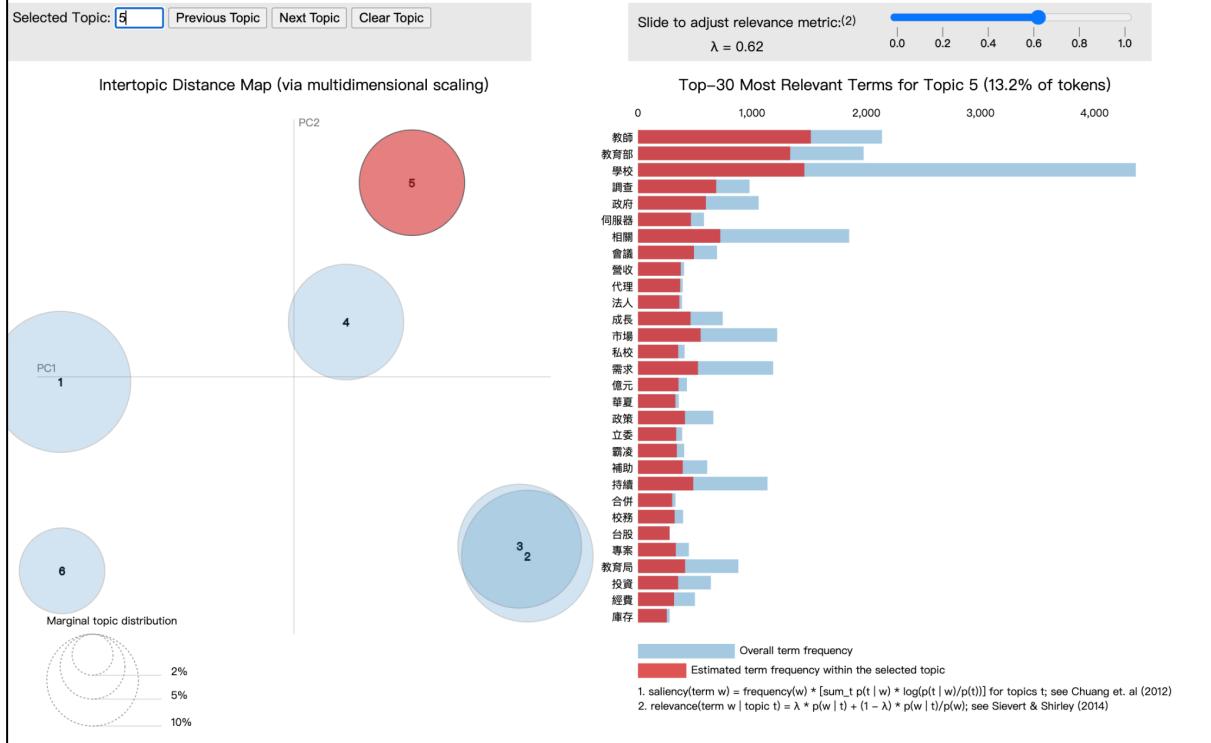
LDA Vis



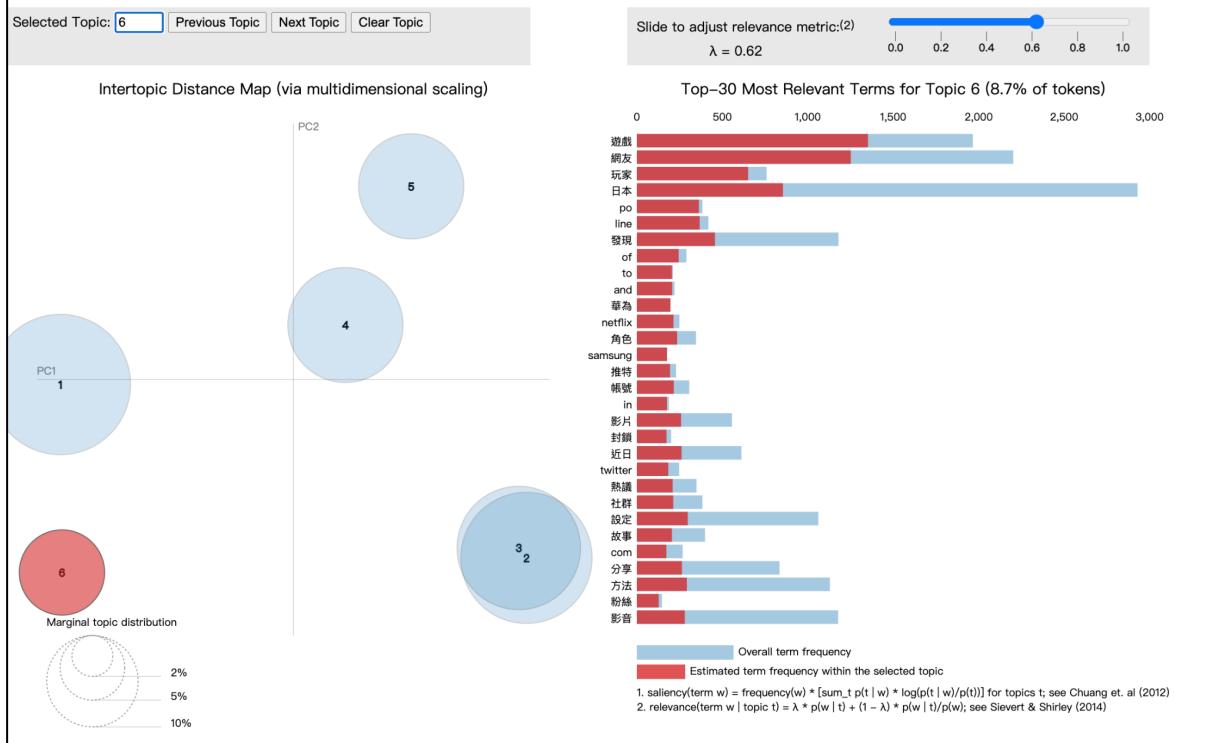
LDA Vis



LDA Vis



LDA Vis



6. 新增欄位(idxmax/min)(22)

- a. 本組利用函數「max」，將LDA分數計算每篇文本的最大值，定義為新欄位「top_topic」

參數設定

Input - 20

任務結果

匯總函數 * 1
max

計算欄位(按住ctrl(Windows)或command(MAC)可以複選) *

system_id
0
1
2
3
4
5

新增的欄位名稱 *

top_topic

- b. 利用新增欄位「top_topic」來儲存主題編號

system_id	0	1	2	3	4	5	top_topic
1	0.150931	0.000000	0.108563	0.369274	0.253541	0.116948	3
2	0.066402	0.000000	0.456730	0.379343	0.000000	0.096895	2
3	0.060342	0.000000	0.094109	0.691096	0.000000	0.149759	3
4	0.212345	0.000000	0.143940	0.641578	0.000000	0.000000	3
5	0.000000	0.000000	0.905195	0.091888	0.000000	0.000000	2
6	0.000000	0.000000	0.264719	0.704543	0.000000	0.030077	3
7	0.000000	0.000000	0.440099	0.553718	0.000000	0.000000	3
8	0.000000	0.000000	0.395736	0.576448	0.000000	0.027127	3
9	0.000000	0.000000	0.886461	0.111084	0.000000	0.000000	2
10	0.029980	0.165981	0.000000	0.754562	0.000000	0.048534	3

7. 合併資料

- a. 將原始聯合新聞爬蟲結果與LDA所計算結果，利用system_id關聯，進行資料合併

合併資料 (77)

參數設定

Input - 4

Input - 22

任務結果

JOIN 規則

新增規則

刪除規則

left_key right_key

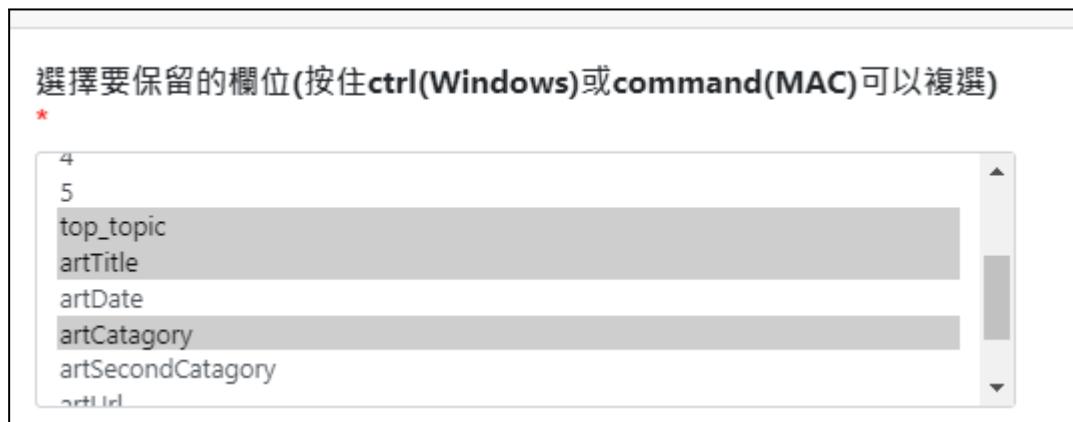
system_id	system_id
-----請選擇-----	-----請選擇-----

b. 產生結果，可作為後續進一步優化模型使用

參數設定	Input - 4	Input - 22	任務結果
1 0.150931 0.000000 0.108563 0.369274 0.253541 0.116948 3	AI考試成績勝大學 生下一步取代記者	2023-01-30 14:14:00	產經 個人理財 https://udn.com/

8. 欄位篩選(Category/模型Topic)

a. 保留「top_topic」、「artTitle」、「artCategory」欄位

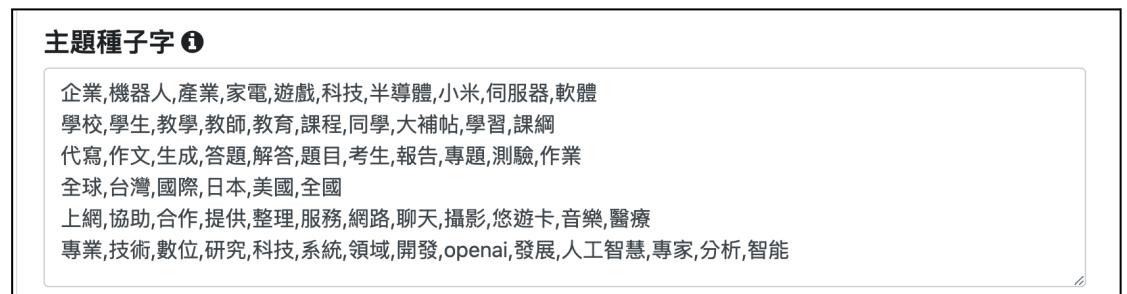


b. 檢視其文章標題(artTitle)、文章類別(artCategory)是否可對應到正確的主題(top_topic)。例如第一篇文章屬於「產經」類別, top_topic為3, 對應之前主題3的前三詞彙(會考 > 學生 > 素養), 由於該篇文章是提到「研究者讓ChatGPT接受美國大學四門法學課程與一門商學考試, 而全數及格通過, 高於修課學生平均值, 因此可能造成未來記者工作因此縮編」。模型或許因為考試等關鍵字將其歸類為主題3, 當然我們可以進一步考量是否要進行優化調整

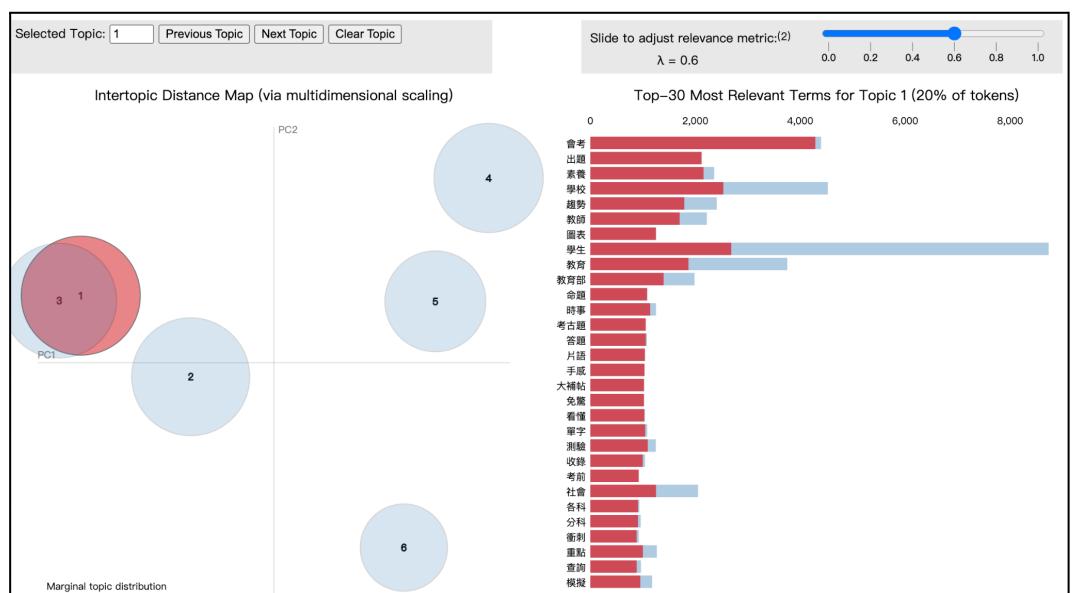
system_id	top_topic	artTitle	artCategory
1	3	AI考試成績勝大學生 下一步取代記者	產經
2	2	未來如何學會與AI溝通 成為新職業技能	產經
3	3	房地產估價師、機場櫃台人員、基金管理...盤點將被AI機器人取代的工作	產經
4	3	程蹟資訊結合AI技術 幫助民眾快速領取曾發現金	產經
5	2	库存調整放緩 半導體底部近 對台灣正面訊息	產經
6	3	中台灣服務業近七成出現缺工危機 旅宿業最嚴峻	產經
7	3	缺服務業六缺工 商總：加薪找不到人、年輕人不愛輪班	產經
8	3	中部服務業有近七成出現缺工危機 六成二計畫開缺徵才	產經
9	2	麻克林證券投顧展望第2季：聚富組合降波動、願收益	產經
10	3	清大校園登場250攤位搶才 欣興電子碩士新鮮人起薪5萬	產經

9. GuidedLDA：

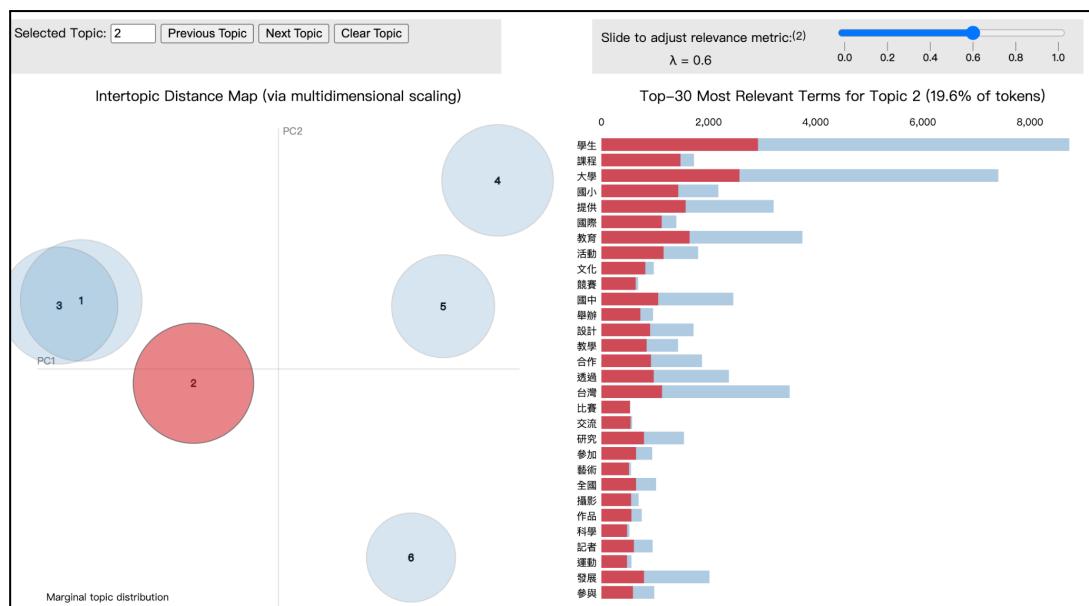
- a. 從原本的字典中篩選整理出來的種子字做主題分屬機率計算。



- b. 設定主題數為6個；迭代次數為150次；保留關鍵字20個。
- c. 設定詞彙頻率下限為20，該詞彙最少要出現在20篇文章中。
- d. 設定詞彙頻率上限為0.7，在所有文章中高於0.7者將排除計算。
- e. 第一個主題佔整體20%，和考試答題有關。



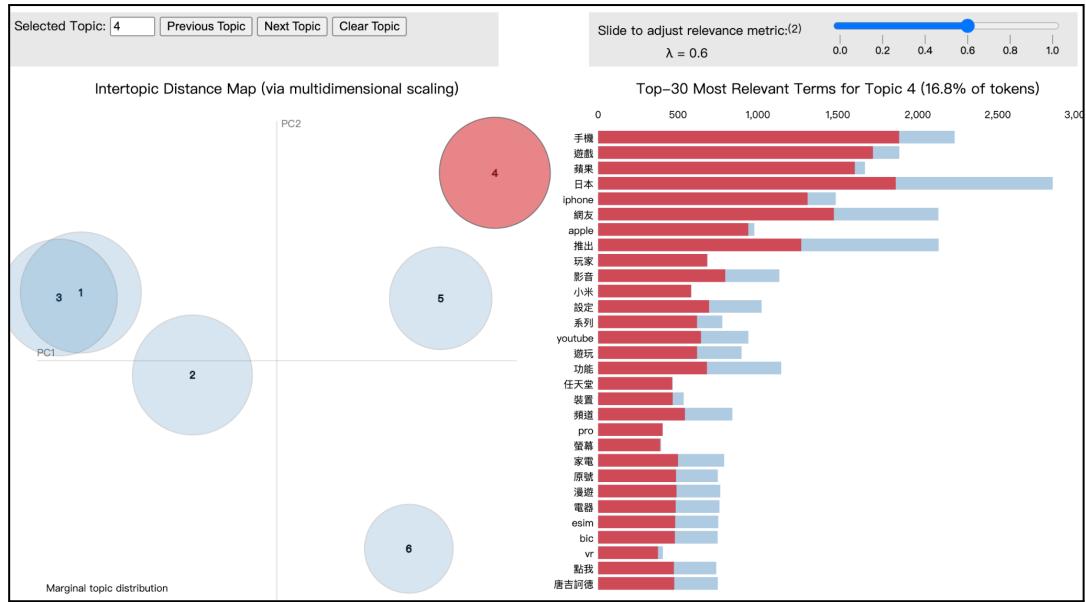
f. 第二個主題佔整體19.6%，和教育有關。



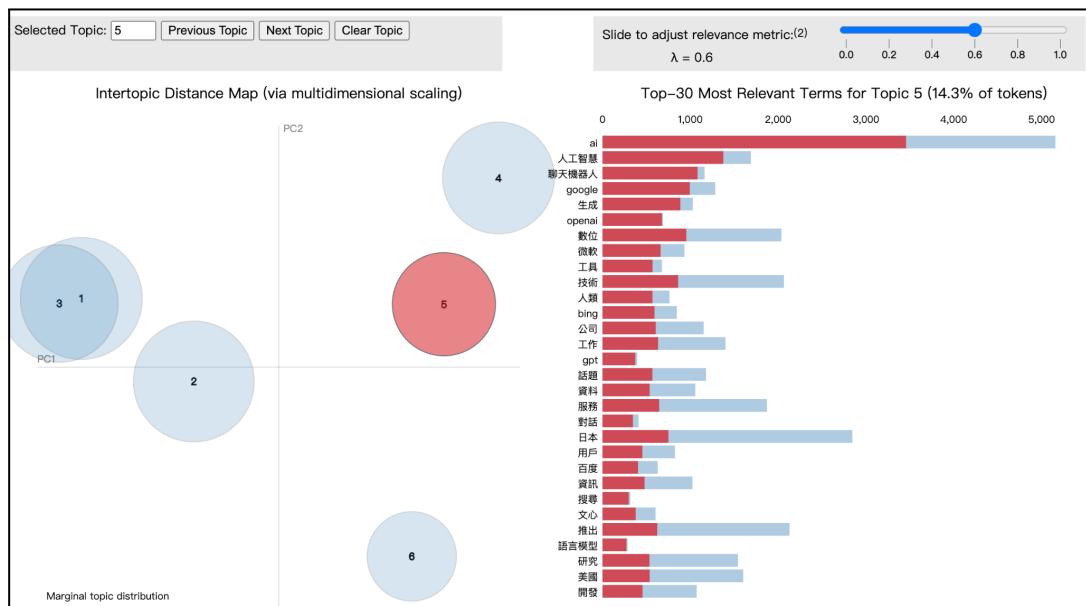
g. 第三個主體佔整體18.6%，和第一主體有多數重複。



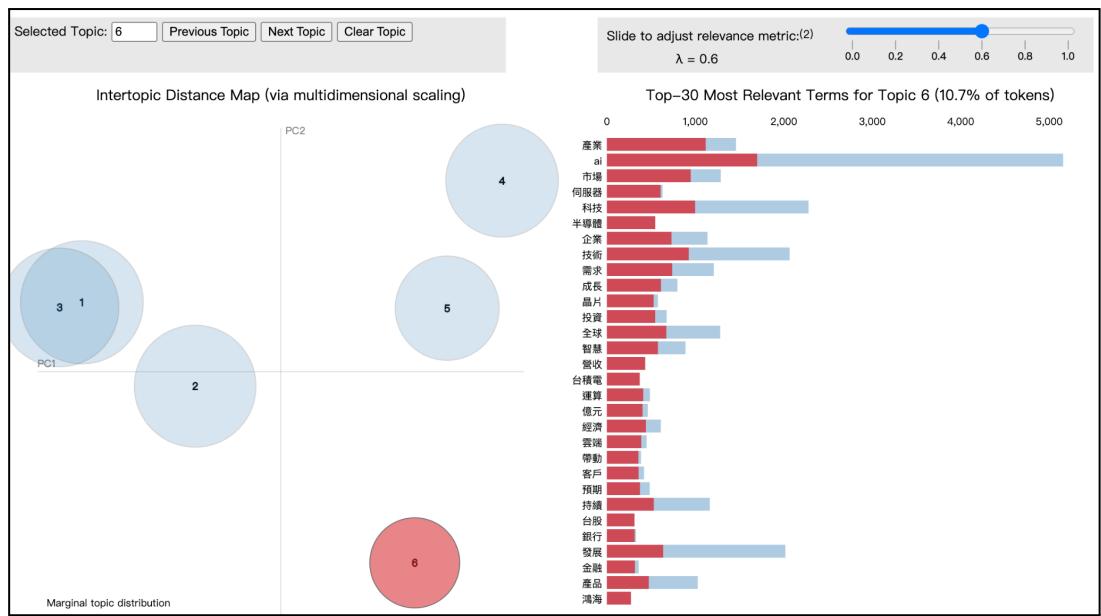
h. 第四個主題佔整體16.8%，和生活有關。



i. 第五個主題佔整體14.3%，和技術有關。



j. 第六個主題佔整體10.7%，和產業有關。



k. 模型的PMI為-1.175，混淆度為1624.74。

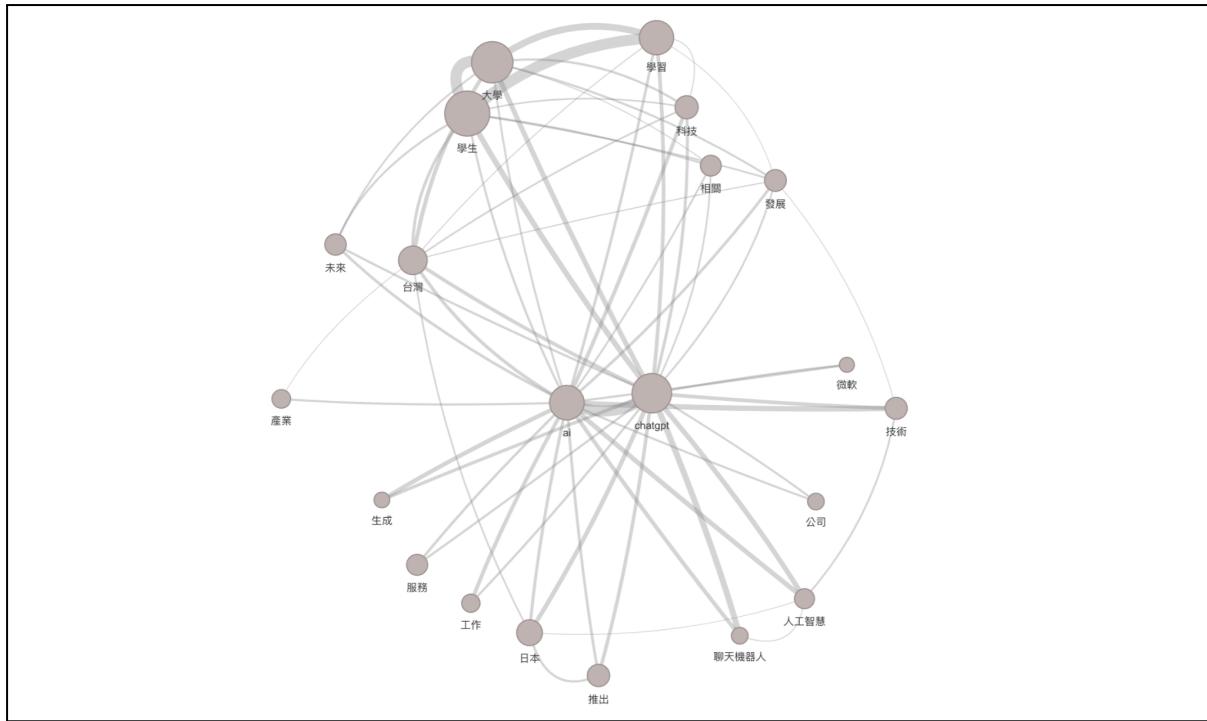


六、視覺化

1. 字詞網路圖：以DTM結果計算相關性矩陣，得出字詞網路圖。詳細設定之節點數量：20，關聯強度：0.7。

由圖可發現「ai」、「chatgpt」兩字詞的關聯程度高，因為chatGPT可說是今年度被普羅大眾廣泛認識、使用以及討論度最高的AI技術的應用。

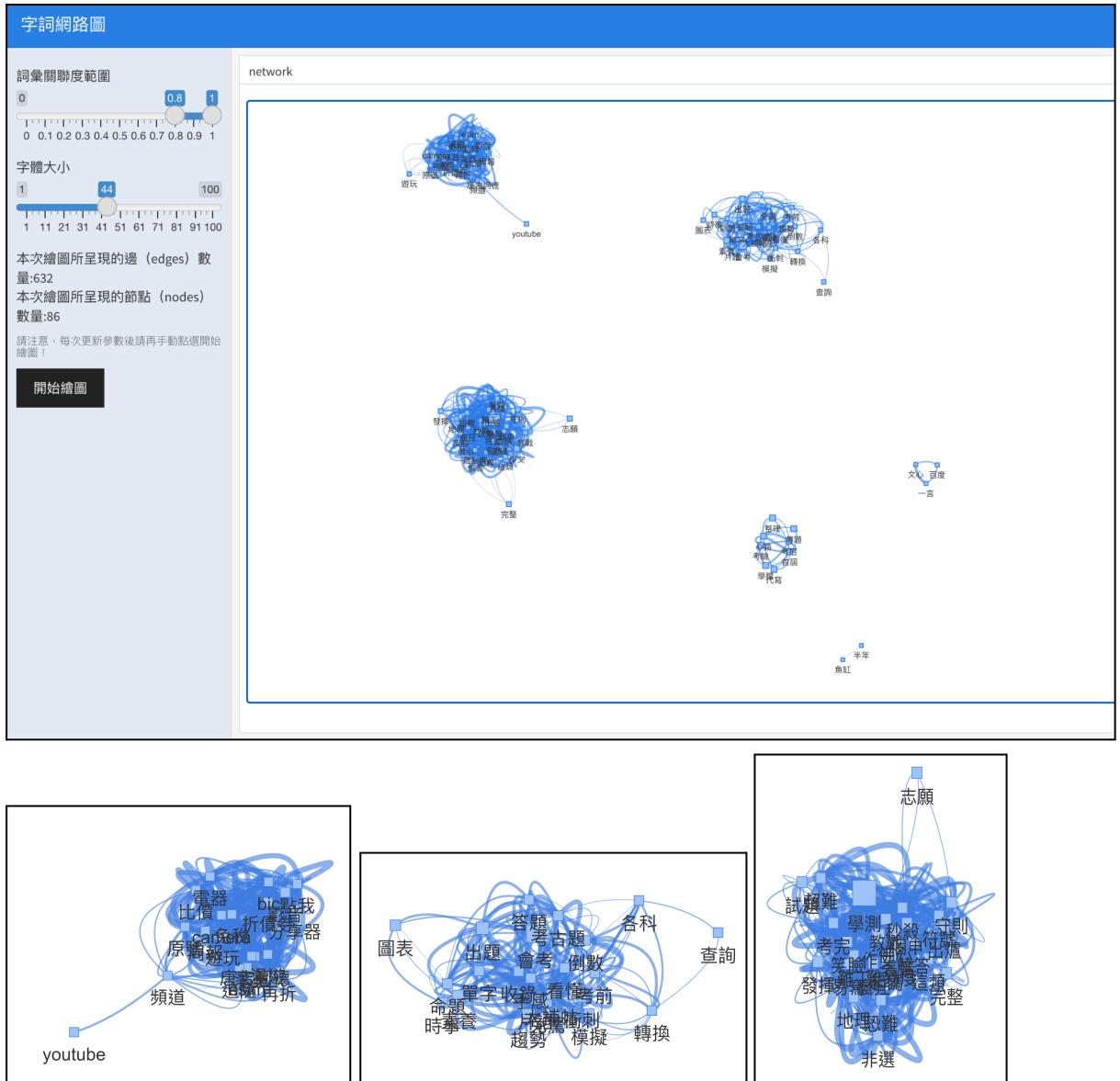
另外，「大學」、「學生」、「學習」三字詞之間的相關性也極為強烈，推測原因可能是chatGDP對於傳統心得、報告等文案撰寫有極高的替代性，因此在新聞內文中較常被討論。



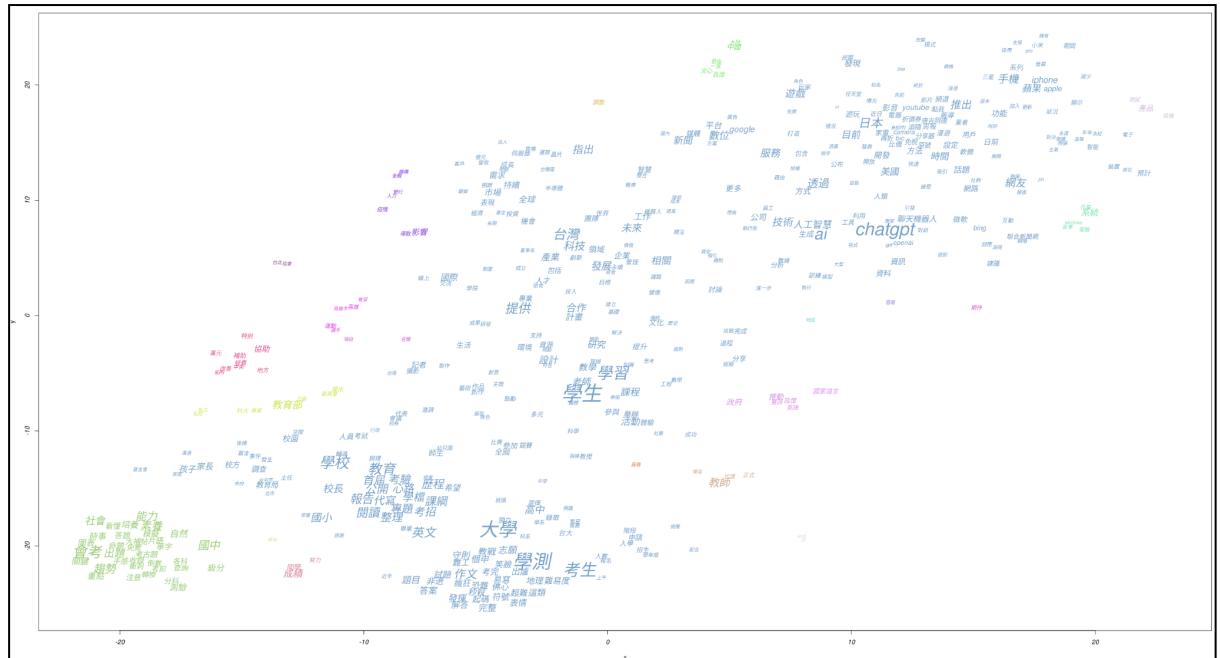
2. 單中心網路圖:利用DTM結果進行共現矩陣計算, 後繪製出字詞網路圖。

因為邊(edges)與節點(nodes)數量及多, 因此將詞彙關聯度範圍設定為0.8至1, 共獲得632的邊與86個節點。

由圖可發現字詞與chatGPT議題關聯較小, 推測是因為爬蟲時會同時抓取到新聞本文以外的推薦新聞欄位, 且本分析時間跨度包含學測以及會考, 因此推薦的熱門新聞多為此主題, 加上熱門新聞會不斷被重複推薦與計算, 導致分析結果與預期不符。



3. 關聯式文字雲: 利用DTM結果繪製結果如下。



將與chatGPT相關的部分擷取放大，可觀察到文字雲確實有捕捉到討論此議題時常用的詞彙，包含「ai」、「聊天機器人」等等。



另外，文字相關圖之中還有許多與chatGPT無直接關連的字詞，例如左下角多半是與升學相關的內容。推測原因同上：本分析納入文教版，且爬蟲時會讀入新聞內文以外的推薦熱門新聞，加上搜尋時間跨度包含學測以及會考，因此推薦新聞的標題大多與此相關，且其會不斷被重複計算於各篇爬蟲結果，使得升學相關的新聞也會因為推薦而被爬蟲捕捉到。

七、結論

本次以ChatGPT為發想，透過課程所學來分析網路上對於ChatGPT討論之主題分類為何。此次資料以聯合新聞網的幾個版為來源，在分析的時候有發現聯合新聞網收集的資料會超出新聞本身，連下方的相關新聞連結也會被收入，因此在判讀上會有不小的誤差。

LDA主題模型產生的主題一為企業和應用相關，可以看到裡面出現的企業都為與AI、ChatGPT有關的企業或應用，雖然在此主題前三大詞彙為ai、手機、日本，但可以拼湊的出此類為與日本產品有關或是ai應用的各企業應用，也就是在觀察的期間內，各界搭上ChatGPT風潮開始發展自己的展品及應用。

主題二和主題三受相關新聞影響，開始出現一些誤差，但從前三大詞彙中仍可看出主題二的範圍為討論ChatGPT如何應用在學生課業上，而主題三就很明顯已經無法辨識出與ChatGPT的關聯。

主題四可以歸類為是台灣在ChatGPT應用上的教育和技術發展，可以看到學院、人才、產業等，可以得出此主題為談論國內大專教育及產業發展目標。

主題五較為複雜，混合了教育到股市的詞彙，取其前三大詞彙也無法看出與在此主題其他詞彙的關聯性，猜測也是因為受到國中會考和大學學測等議題影響，在本來談及有關股市的主題納入了教育相關。

GuidedLDA主題模型產生的主題一為考試相關，主題一即受到上述問題影響的主題，其內容多為近期熱門議題（或主題），即為國中會考登場和解答，內容多半受此事件（國中會考）影響而有所偏差，即便本身即為考試相關主題，但受到觀察外的資料汙染因此主題一需要做改善。

主題二為教育相關，其因教育議題，因此也受國中會考問題汙染，但在主題二看起來問題較不嚴重，可以看到會考以外的項目，如設計、攝影、藝術等。

主題三與主題一高度重複且更偏向了國中會考和大學學測等事件，因此主題三也是需要再做調整的。

主題四則以手機和家電及相關應用有關，此類主題應是在討論家電的相關應用。

主題五就比較明顯回歸到ChatGPT的技術面，可以看出此主題為討論相關技術、相關企業、相關應用等。

主題六則為討論近期受ChatGPT影響之公司和產業在其營收和股價上的波動。

綜合來說，本次選用聯合新聞網並指定議題探索，在聯合新聞網的資料收集部分就出現了不小的問題，此問題也影響到了後續主題分類的判讀，即便本組仍能從中挖掘出原先的樣子並加以觀察和分析，但其影響也不小。而在ChatGPT本身的觀察可以發現企業相關、股票相關、學生應用、技術應用等四個大項目在去拆分合併成五至六個主題，但其中都脫離不出彼此的關聯，也就是說ChatGPT影響範圍不是在單一或少數主題及內容，其影響力之大以至於每一個主題之間都是可以有所關聯的，而目前也正處於ChatGPT的發展階段，未來資料量更大之後將更容易看出每一個主題彼此互相串聯。