

社群媒體分析

第三次讀書會報告_第十二組

一、分析主題

透過分析不同版面的政治、社會、國際版面新聞分析新聞主題分布。

二、組員

N114320002_李湘怡
N114320008_陳家卉
N114320009_林育心
N114320017_黃敏喻
N114320020_曾子瑋
N114320027_吳政翰
N114320030_吳郁文
N114320031_洪宜綾

三、分析工具

- 中山大學工作流程平台：Tarflow
- 工作流程名稱：第三次讀書會
- 主題模型

四、動機目的

小組從東森新聞政治、社會、國際版面蒐集新聞內容，並運用自訂的詞彙字典以及 Tarflow 平台提供的 LDA 主題模型及 GuidedLDA 模型，來進行相關主題的內容分布分析。

針對去年重要新聞事件期間，以巴衝突、總統大選時期(2023/10/01-2023/12/31)，最終目的為識別在關鍵時期內新聞報導的主題偏重，以獲得有價值的資訊和洞察，進而更深入了解台灣在政治、社會和國際事務上的動態。

五、資料來源

- 資料來源：東森新聞資料庫之政治、社會、國際看板。
- 關鍵字：無
- 搜尋日期：2023/10/01~2023/12/31
- 資料筆數：6292 筆

東森新聞爬蟲 (112)

參數設定
任務結果

選擇看板 *

- living(生活)
- politics(政治)
- society(社會)
- sport(體育)
- story(新聞)
- travel(旅遊)
- world(國際)

搜尋關鍵字 ⓘ

以換行區隔, e.g.
國立中山大學
西子灣
...

搜尋起始日期

搜尋結束日期

任務結果

東森新聞爬蟲 (112)

參數設定
任務結果

統計資訊

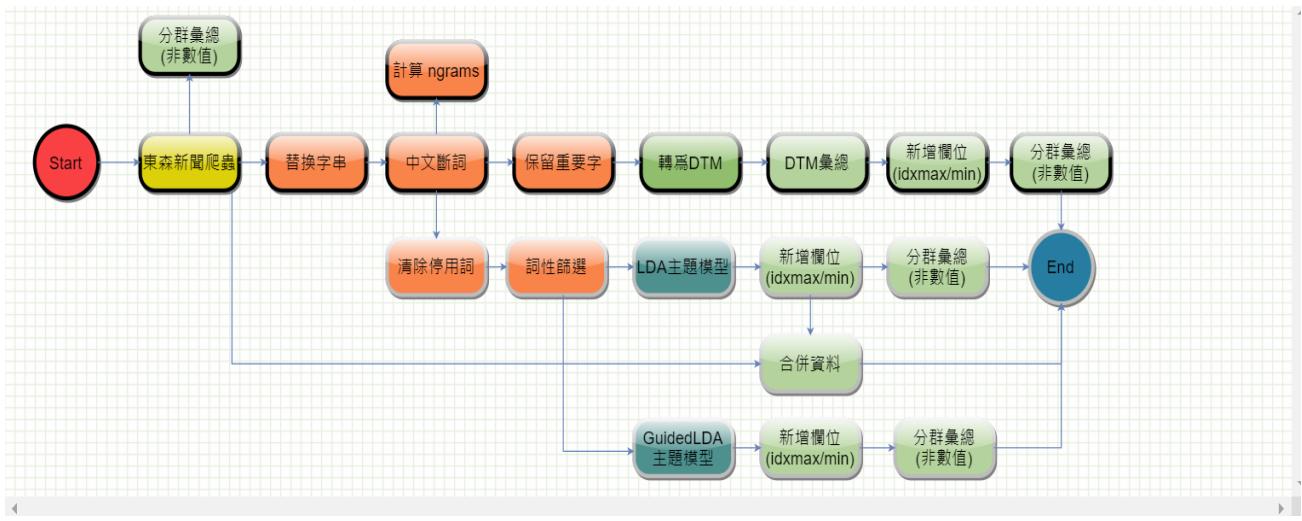
6 欄位數		6292 資料筆數	
-----------------	--	---------------------	--

任務結果

Show 10 entries	Search:					
system_id	artTitle	artUrl	artDate	artCategory	artContent	dataSource
	總統大選最新民調曝！吳子嘉：侯		2023-10-01		美麗島電子報30日公布最新總統大選民調，三腳督情況下民進黨領滑得39.4%、國	
	全螢幕瀏覽	點我下載完整CSV資料	點我下載完整Rdata	點我下載完整json資料		

六、系統流程圖

- 6-1. 資料前處理：以東森新聞爬蟲資料庫，進行資料觀察(各版文章數量)和前處理(斷詞、停用詞)。
- 6-2. 分群彙總：了解東森新聞原資料集之新聞其各類別分別為多少數量。
- 6-3. 將資料進行預處理(替換字串、中文斷詞、保留重要字、清除停用詞、詞性篩選)
- 6-4. 透過三種方式訓練與判斷該文章為何主題。(自建字典轉 DTM、LDA 主題模型、GuildedLDA 主題模型)



任務結果

分群彙總 (非數值) (121)

參數設定	Input - 112	任務結果
統計資訊		
3 群組數量		
任務結果		
Show 10 entries	Search: <input type="text"/>	
artCategory	system_id@count	
國際	1904	
政治	1901	
社會	2487	
Showing 1 to 3 of 3 entries		Previous 1 Next

七、資料預處理-替換字串、中文斷詞、詞性篩選、保留重要字

7-1 將東森新聞透過正規表達式把 URL 都替換掉

參數設定	Input - 112	任務結果
選擇處理欄位 *	artContent	替換字串設定 ⓘ <pre>((http ftp https)://)[((a-zA-Z0-9\-_]+\.[a-zA-Z]{2,6}) ([0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}) ([0-9]{1,3})) ([0-9]{1,4})* ([a-zA-Z0-9\&%_\/\-\~\^*\?]>></pre>
選擇替換規則檔案 ⓘ	-----請選擇-----	
<input type="button" value="儲存更改"/>		

7-2 賦予專有名詞權重提升中文斷詞的精準度

三腳督 500
藍白合 500
總統大選 500
市長 500
立法委員 500
潛艦國造 500
Linbay 好油 500
國安單位 500
市政單位 500
哈瑪斯 500
烏俄戰爭 500
美麗島電子報 500

中文斷詞 (101)

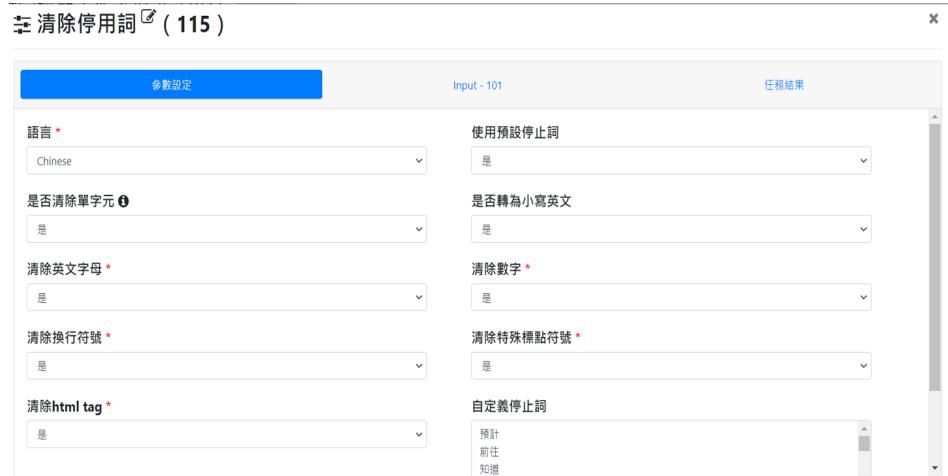


7-3 清除停用字

將以下文字設為停用字：

預計、前往、知道、完成、日期、決定、朋友、情人、確定、需要、特別、今天、看到、八卦、早上、晚上、下午、之後、現在、提供、目前、現在

清除停用詞 (115)



任務結果

停用字 1087421 個

參數設定	Input - 101	任務結果
清除停用詞 (115)		

統計資訊

1087421



任務結果

Show 10 entries

Search:

7-4 保留重要字

透過自建字典(檔案: data0506-1.csv)保留重要字詞，並且自定義其分類類別(戰爭、治安、交通事故、選舉)。

自建字典預覽圖如下，且預先定義其字詞 class(戰爭、治安、交通事故、選舉)及同義字詞 alias

Show entries

	name	class	alias
0	以色列	戰爭	以色列
1	巴勒斯坦	戰爭	巴勒斯坦
2	哈瑪斯	戰爭	哈瑪斯
3	以巴衝突	戰爭	以巴戰爭
4	戰爭	戰爭	戰爭
5	加薩	戰爭	加薩
6	軍事	戰爭	軍事
7	攻擊	戰爭	攻擊
8	武器	戰爭	武器
9	和平	戰爭	和平

Showing 1 to 10 of 130 entries

Previous 1 2 3 4 5 ... 13 Next

三 保留重要字 (139)

參數設定

Input - 101

任務結果

設定保留詞彙 ⓘ

以換行符號區隔, e.g.

西子灣
壽山
駁二
...

選取字典

data0506-1.csv

字典欄位

name

儲存更改

任務結果

保留住自建字典之字詞。

三 保留重要字 (139)

參數設定 Input - 101 任務結果

任務結果

Show 10 entries Search:

system_id	result
1	[民謽, 民進黨, 賴清德, 國民黨, 侯友宜, 民眾黨, 柯文哲, 賴清德, 侯友宜, 柯文哲, 侯友宜, 民謽, 賴清德, 民進黨, 賴清德, 侯友宜, 民謽, 柯文哲, 柯文哲, 柯文哲, 柯文哲, 賴清德, 民謽, 民謽, 柯文哲, 柯文哲, 柯文哲, 柯文哲, 柯文哲, 藍白台]
2	[法院, 法官, 立法委員, 檢察官, 法官]
3	[民進黨, 賴清德, 民進黨, 民進黨, 國民黨, 民進黨, 國防, 國民黨, 民進黨, 國民黨, 立法委員, 民進黨, 國民黨, 國防, 國防, 外交, 民進黨, 國民黨, 攻擊, 民進黨, 民進黨, 國防, 候選人, 民謽, 候選人]
4	[民進黨, 總統, 賴清德, 民進黨, 總統, 賴清德, 賴清德, 侯友宜, 民進黨, 總統, 侯友宜, 賴清德, 國民黨, 總統, 侯友宜, 賴清德, 賴清德, 侯友宜, 賴清德, 侯友宜]
5	[立法委員, 國民黨, 立法委員, 國民黨, 民進黨, 民進黨, 國民黨, 國民黨]

全螢幕瀏覽 點我下載完整CSV資料 點我下載完整Rdata 點我下載完整json資料

7-5 詞性篩選:為了將過多無意義的出現頻率太高動詞去除，篩選名詞與專有名詞作為判斷字詞。

三 詞性篩選 (135)

參數設定 Input - 115 任務結果

語言 * Chinese 選擇保留詞性 *

- Noun
- Proper Noun
- Verb
- Adjective
- Adverb

儲存更改

八、DTM

8-1 保留重要字詞後轉為 DTM 作詞頻計算

三 轉為DTM (141) 參數有做更動，建議重新執行

參數設定 Input - 139 任務結果

保留詞彙 ① 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 ② 200

儲存更改

任務結果

轉為DTM (141) 參數有做更動，建議重新執行

參數設定		Input - 139																				任務結果			
Show	10 entries																					Search:			
system_id	ohca	中共	乘客	事故	以色列	侯友宜	俄羅斯	候選人	偷拍	傷害	兩岸	刑事	刑警	判刑	刺殺	前線	副總統	加薩	勞工	危險	台日	台獨	台美	吳欣盈	吸毒
1	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	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	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	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	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

全螢幕瀏覽 紮我下載完整CSV資料 紮我下載完整Rdata 紮我下載完整json資料

8-2 DTM彙總

透過彙總自建辭典的重要關鍵字出現次數，得到每篇文章分類與關鍵字字頻計算
DTM彙總 (143)

參數設定 Input - 141 任務結果

參數設定		Input - 141																				任務結果	
選擇匯總字典 *																							
data0506-1.csv																						儲存更改	

任務結果

DTM彙總 (143)

參數設定		Input - 141																				任務結果			
任務結果																									
Show	10 entries																					Search:			
system_id		戰爭	治安	交通事故	選舉																				
1		0.0	0.0	0.0	33.0																				
2		0.0	4.0	0.0	1.0																				
3		2.0	0.0	0.0	24.0																				
4		0.0	0.0	0.0	24.0																				
5		0.0	0.0	0.0	8.0																				
6		0.0	0.0	0.0	10.0																				
7		1.0	0.0	0.0	52.0																				

全螢幕瀏覽 紮我下載完整CSV資料 紮我下載完整Rdata 紮我下載完整json資料

8-3 新增欄位

將出現頻率最高(max)字詞以其所屬之 class 判定為該文章之 class。

新增欄位：判定該文章之主題 topic

■ 新增欄位 (idxmax/min) (146)

參數設定

Input - 143

任務結果

匯總函數 * ⓘ

max

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

system_id
戰爭
治安
交通事故
選舉

新增的欄位名稱 *

topic

儲存更改

任務結果

參數設定

Input - 143

任務結果

任務結果

Show 10 entries

Search:

system_id	戰爭	治安	交通事故	選舉	topic
1	0.0	0.0	0.0	33.0	選舉
2	0.0	4.0	0.0	1.0	治安
3	2.0	0.0	0.0	24.0	選舉
4	0.0	0.0	0.0	24.0	選舉
5	0.0	0.0	0.0	8.0	選舉
6	0.0	0.0	0.0	10.0	選舉
7	1.0	0.0	0.0	52.0	選舉

全螢幕瀏覽

點我下載完整CSV資料

點我下載完整Rdata

點我下載完整json資料

8-4 分群彙總

計算(count)歸類主題(topic)文章數如下：

■ 分群彙總(非數值) (148)

參數設定

Input - 146

任務結果

使用...欄位進行分群(按住ctrl(Windows)或command(MAC)可以複選) *

system_id
戰爭
治安
交通事故
選舉
topic

匯總函數 * ⓘ

count
unique
min
max
first
last
sum

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

system_id
戰爭
治安
交通事故
選舉
topic

儲存更改

任務結果

根據自定義的分類詞典，可以看到新分類的四個主題：選舉、治安、戰爭、交通事故佔比中，以選舉最為多，其次為交通事故。

比較原始資料東森新聞各版面（國際、政治、社會）數量，國際及政治數量近乎相同。然而，新分類出的主題，選舉卻遠高於戰爭主題數，研判可能是因為在台灣總統大選期間，東森新聞更聚焦於選情報導，且國際新聞的報導範疇則更加廣泛。此外，在社會版報導內容，可以發現交通事故文章數遠高於治安事件，這部分與實際情況相符。

分群彙總 (非數值) (148)

任務結果

topic	system_id@count
交通事故	1629
戰爭	1054
治安	1172
選舉	2176

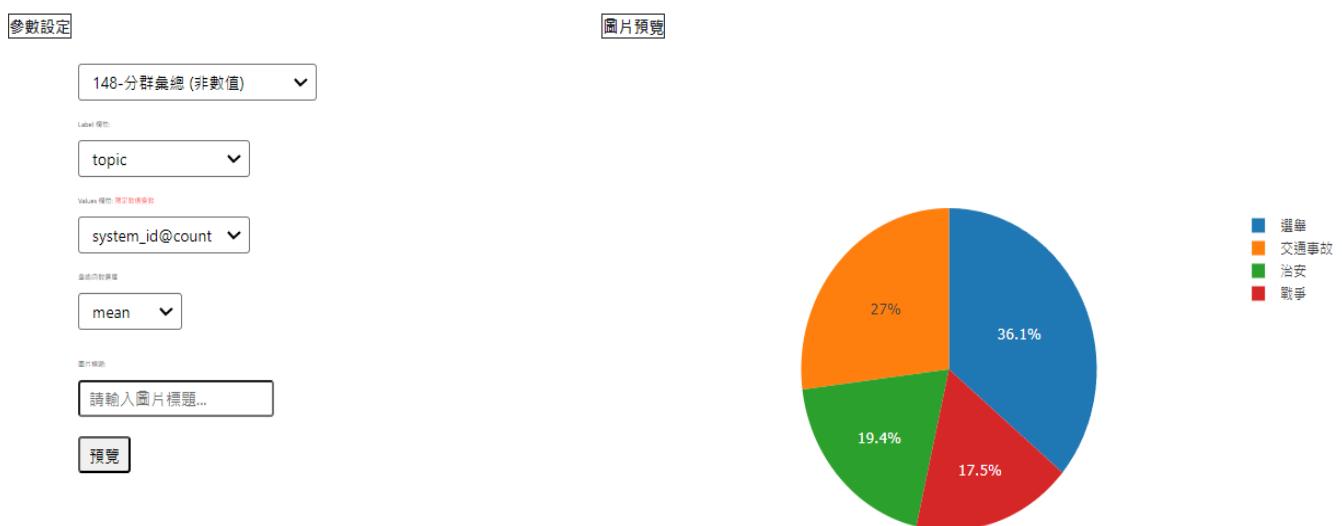
Show 10 entries Search:

Showing 1 to 4 of 4 entries Previous 1 Next

全螢幕瀏覽 點我下載完整CSV資料 點我下載完整Rdata 點我下載完整json資料

8-5 視覺化儀表板呈現結果

資料期間為 2023/第四季，正值台灣總統大選，因此可以發現在選舉相關新聞的篇幅最高(36.1%)，其次為交通事故(27%)與治安(19.4)，整體來說台灣內部新聞佔比高達 82.5%，最後為與國際新聞相關的戰爭議題為 17.5%，這與台灣新聞媒體為人所詬病缺乏國際觀的新聞報導印象相符。



九、LDA 主題模型

9-1 建立 LDA 參數

設定四個主題

LDA主題模型 (117)

參數設定

Input - 135

任務結果

目標欄位 *

result

迭代次數

100

主題數 *

4

主題保留關鍵字數量

20

詞彙頻率下限 ⓘ

40

詞彙頻率上限 ⓘ

0.7

alpha

預設為主題數/50

Beta

預設為0.1

chuckszie ⓘ

預設為2000

update_every ⓘ

1

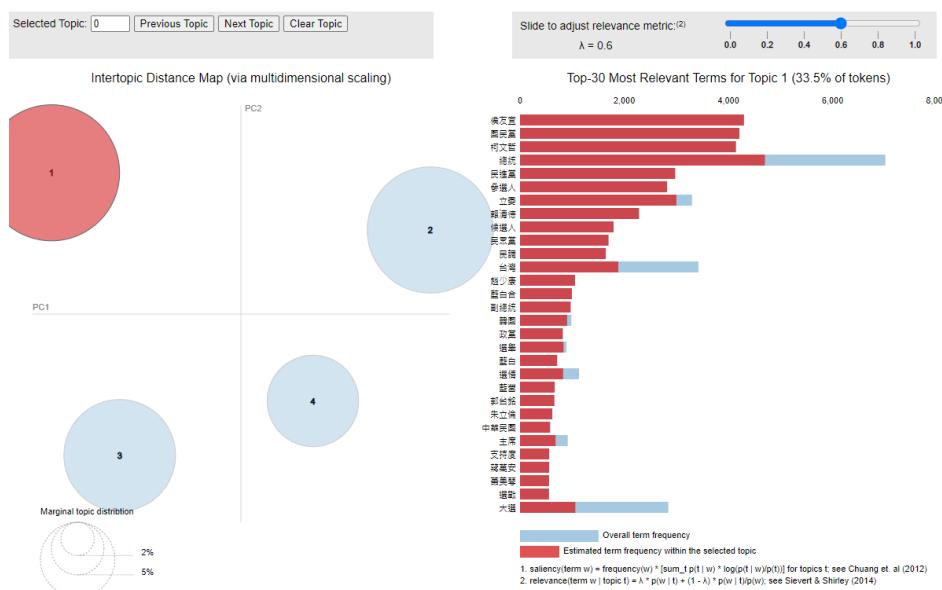
是否輸出字典

任務結果

主題 1

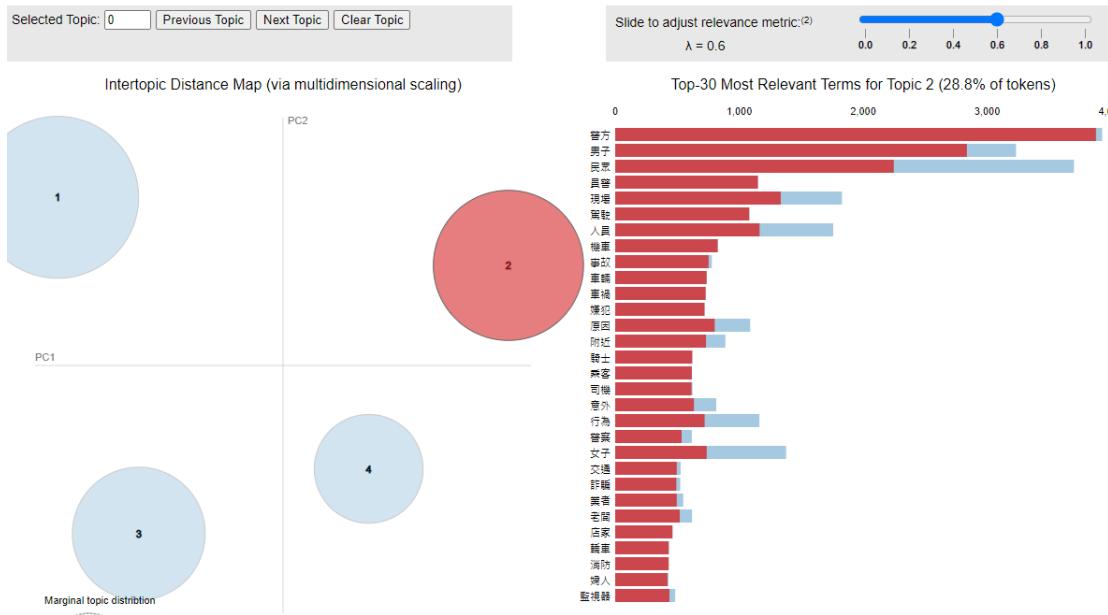
從關鍵字清單中，可以發現主題 1 判別結果與我們自定義字典的選舉主題相似。從分析圖中可以看出，包括了一些政治人物和政黨的名稱，顯示這部分的內容主要涉及政治選舉。

LDA Vis



主題 2

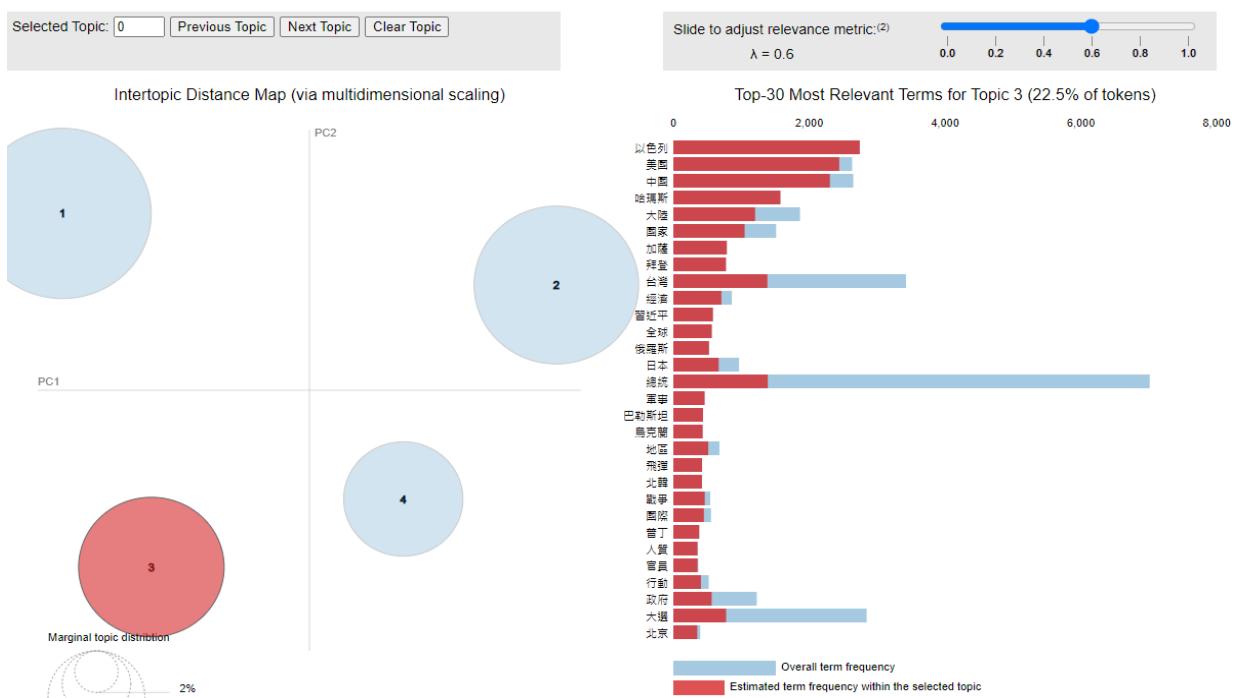
主題 2 判別結果接近我們自定義字典的交通事故類別，其詞彙文檔也相對廣泛，判斷是交通事故此類別內容本身較無特定的專有名詞和詞彙，內容會因不同人、事、時、地、物而有所不同，但嫌犯、現場、騎士、警方、司機、駕駛，這些詞彙與交通事故最為相關，也較不會出現在政治、國際類別文章，因此藍色條狀佔比幾乎是 0。



主題 3

由於數據來源於 2023 年 10 月到 12 月期間，該時段恰逢以巴戰爭，主題 3 因此偏向戰爭主題。分析圖顯示以色列、美國、巴基斯坦的提及頻率最高。雖然在戰爭主題中有提及台灣、總統、政府、中國等詞彙，但這些詞彙在其他主題中也有出現，尤其是在政治相關的討論中。

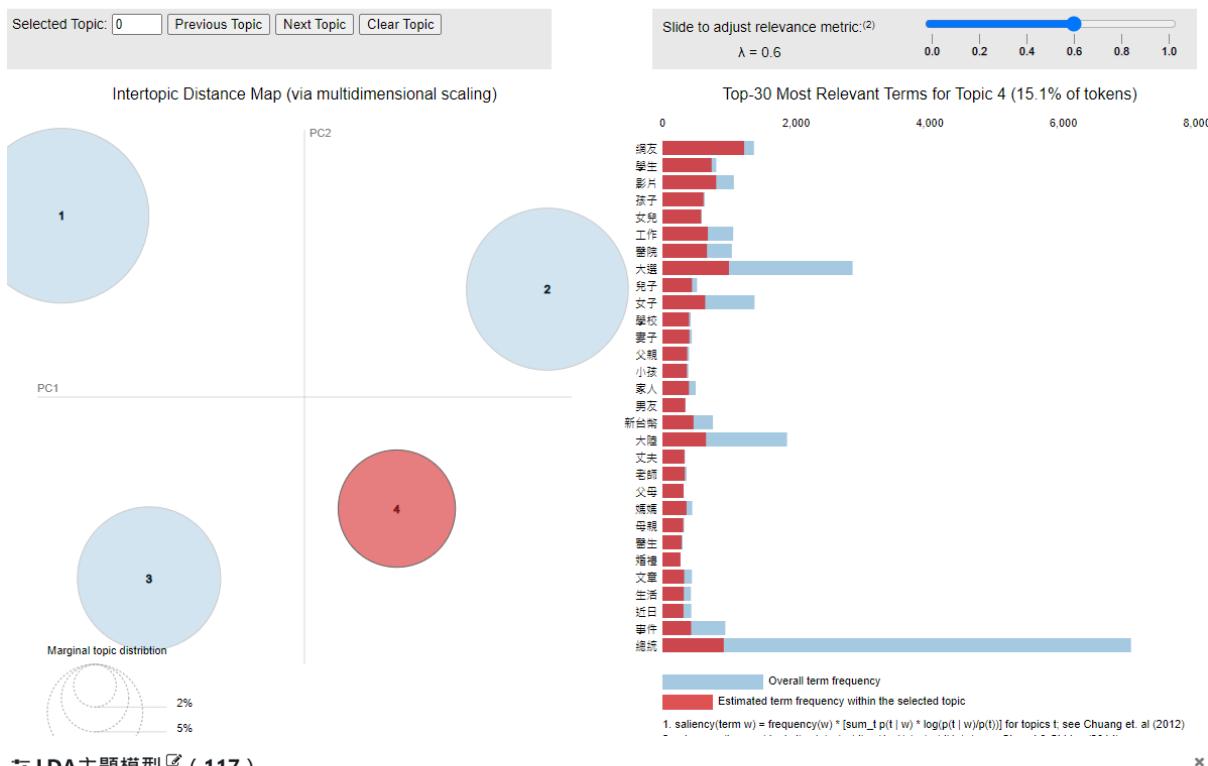
LDA Vis



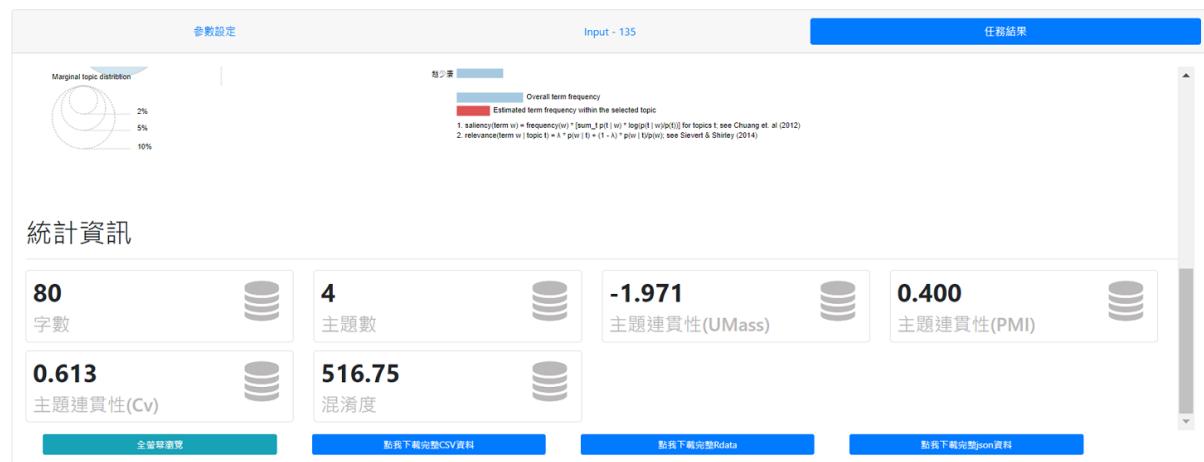
主題 4

從主題 4 顯示的關鍵詞清單，較難以看出實質分類意義，研判是社會新聞涵蓋範圍較廣的緣故。舉例而言，「網友」、「大選」、「總統」、「報導」、「網路」與「文章」，其主題可能與當前事件或社會新聞報導相關；「學生」、「學校」、「老師」這類詞彙為教育主題；還有一系列與家庭關係相關的詞，如「男子」、「媽媽」、「兒子」、「女兒」、「家人」，這些詞彙則偏向家庭及個人生活主題。

LDA Vis



LDA 主題模型 (117)



整體來看，這四個主題的分佈較為明顯，各主題中的文章關鍵字分佈有較少的重疊。透過分析關鍵字出現的頻率，可以較清楚地劃分出各主題的範圍。在選舉相關版面中，由於政治人物和政黨的觀點及對立、相關訊息在總統大選期間較為多元化，主題分散程度也相對較高。

9-2 新增欄位 topic

計算其所屬分類主題矩陣，取最大值(max)判斷為該所屬 topic

■ 新增欄位 (idxmax/min) (128)

參數設定

Input - 117

任務結果

匯總函數 * ⓘ

max

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

system_id
0
1
2
3

新增的欄位名稱 *

topic

儲存更改

任務結果

■ 新增欄位 (idxmax/min) (128)

參數設定

Input - 117

任務結果

任務結果

Show 10 entries

system_id	0	1	2	3	topic
1	0.000000	0.000000	0.000000	0.996635	3
2	0.000000	0.000000	0.837831	0.157993	2
3	0.000000	0.000000	0.000000	0.996974	3
4	0.122303	0.094453	0.000000	0.782195	3
5	0.209879	0.000000	0.000000	0.786589	3
6	0.000000	0.000000	0.110701	0.886731	3
7	0.000000	0.000000	0.000000	0.997154	3
8	0.000000	0.082113	0.178922	0.737917	3
9	0.000000	0.000000	0.000000	0.996770	3

全螢幕檢覽

點我下載完整CSV資料

點我下載完整Rdata

點我下載完整json資料

9-3 分群彙總

計算(count)四個主題(0123)之文章數量

■ 分群彙總 (非數值) (126)

參數設定

Input - 128

任務結果

使用...欄位進行分群(按住ctrl(Windows)或command(MAC)可以複選) *

system_id
0
1
2
3
topic

匯總函數 * ⓘ

count
nunique
min
max
first
last
sum

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

system_id
0
1
2
3
topic

儲存更改

任務結果

三 分群彙總 (非數值) (126)

參數設定 Input - 128 任務結果

4 群組數量

任務結果

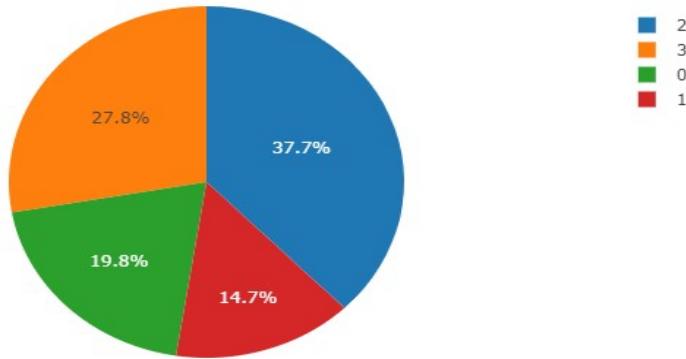
Show 10 entries Search:

topic	system_id@count
0	1243
1	928
2	2373
3	1748

Showing 1 to 4 of 4 entries Previous 1 Next

全螢幕觀賞 點我下載完整CSV資料 點我下載完整Rdata 點我下載完整json資料

9-4 視覺化儀表板呈現結果



9-5 資料合併

為幫助了解 rawdata 與 LDA 主題模型分類是否符合預期，以 system_id 為主，將原始資料內容和 LDA 分類結果做合併以利審視。

三 合併資料 (129)

參數設定 Input - 112 Input - 128 任務結果

JOIN規則

新增規則 刪除規則

任務一欄位 任務二欄位

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

儲存更改

任務結果

三 合併資料 (129)

參數設定		Input - 112		Input - 128		任務結果							
任務結果													
Show 10 entries													
system_id	artTitle	artUrl	artDate	artCategory	artContent	dataSource	0 1 2 3 topic						
1	總統大選最新民調爆！吳子嘉：侯友宜支持度升高 東森新聞	https://news.ebc.net.tw/news/article/384821	2023-10-01 13:14:00	政治	美麗島電子報30日公布最新傳統大選民調，三腳督情況下民進黨領悟帶39.4%、國民黨侯友宜23.5%、民眾黨柯文哲19.0%；若是四腳督對戰，賴清德36.6%，侯友宜21.6%，柯文哲16.2%，馮海鵬...	EBC	0.000000 0.000000 0.000000 0.996635 3						
2	高虹安涉詐領助理費案 北院2日傳喚王郁文 東森新聞	https://news.ebc.net.tw/news/article/384825	2023-10-01 13:16:00	政治	台北地方法院審理高虹安涉嫌偽領助理費案，9月25日傳喚高虹安前助理黃惠玲等3名被告，3人均認罪，並請法官減刑且宣告緩刑。北院明天將傳喚萬虹安第1名被控助理王郁文出庭，新竹市長高虹安被控立法院委員任內收民國1...	EBC	0.000000 0.000000 0.837831 0.157993 2						
					民進黨立委高嘉瑜vs.選民：「賴清德、高嘉瑜、東勢、東勢、東勢、」								
全萤幕瀏覽		點我下載完整CSV資料		點我下載完整Rdata		點我下載完整json資料							

三 合併資料 (129)

參數設定		Input - 112		Input - 128		任務結果	
				訪,進行座談,國民黨總統參選人侯友宜:... 教育部廣告片:「媽,明天我的公仔會寄過來!你倒幫我收一下。(蛤,你還在玩那些『娃娃』哩。)後來錢呢,(那確實實更有用的東西嘛。)」這公仔的頭卻被媽媽拿去賣育兒用品,教育部上個月推出的「0到6歲居家養護...」			
93	0-6歲居家一起養育廣告炎 上! 教育部推廣「玻璃」 東森新聞	https://news.ebc.nettw/news/article/386036	2023-10-09 18:40:00	政治	EBC	0.000000 0.591497 0.145773 0.260465	1
94	獨家 / 揭密「私底下的侯友宜」專訪直呼忙得不習慣 東森新聞	https://news.ebc.nettw/news/article/386042	2023-10-09 20:40:00	政治	EBC	0.000000 0.000000 0.047769 0.949533	3
95	藍白會談二對二? 侯陳營將 添金湧跪 黃健庭 東森新	https://news.ebc.nettw/news/article/386046	2023-10-09 21:50:00	政治	EBC	0.000000 0.000000 0.000000 0.996855	3

小結

在未給予種子字的情況下，LDA 主題分群與東森新聞版分類比對如下，主題分群效果不錯，如 0 代表國際新聞、2 代表社會新聞、3 代表政治新聞。

東森新聞		LDA 主題模型分類				
版分類		0	1	2	3	總計
社會	19	319	2142	7		2487
政治	72	45	51	1733		1901
國際	1152	564	180	8		1904
總計	1243	928	2373	1748		6292

十、GuildedLDA 模型

10-1 主題、主題種子、迭代次數、詞彙頻率下限/上限

因預設主題數為 4，故設定 4 個主題種子字，來影響主題分類結果。
下限設定為 40，低於 40 篇文章次數的排除；同時排出過多頻率的詞（是一些不重要的詞彙）。

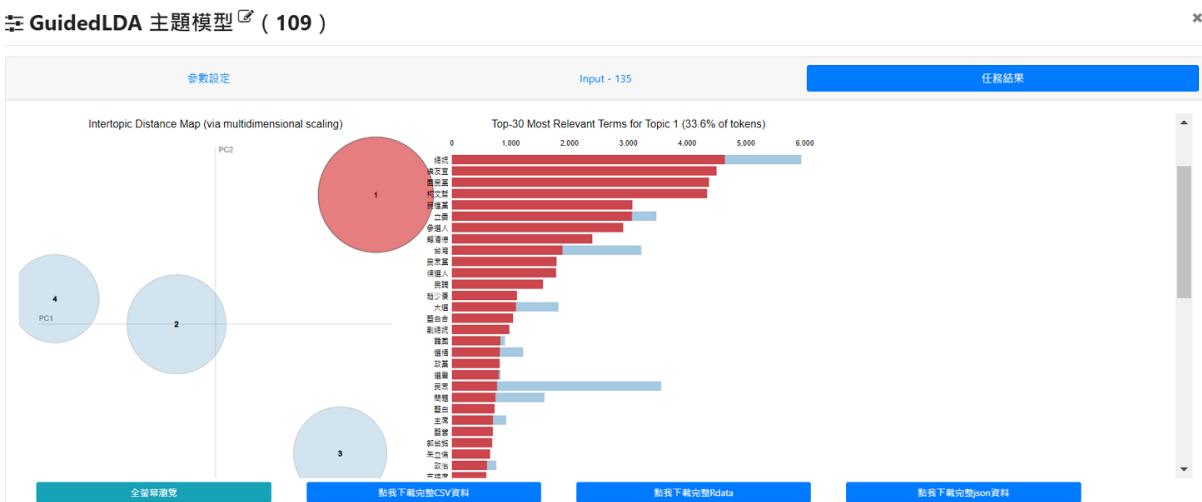
總統,大選,立法委員,立委,政黨
以色列,巴勒斯坦,哈瑪斯,衝突
車禍,酒駕,毒駕,交通,事故
命案,刑警,犯罪,槍擊,暴力,毒品



任務結果

主題 1

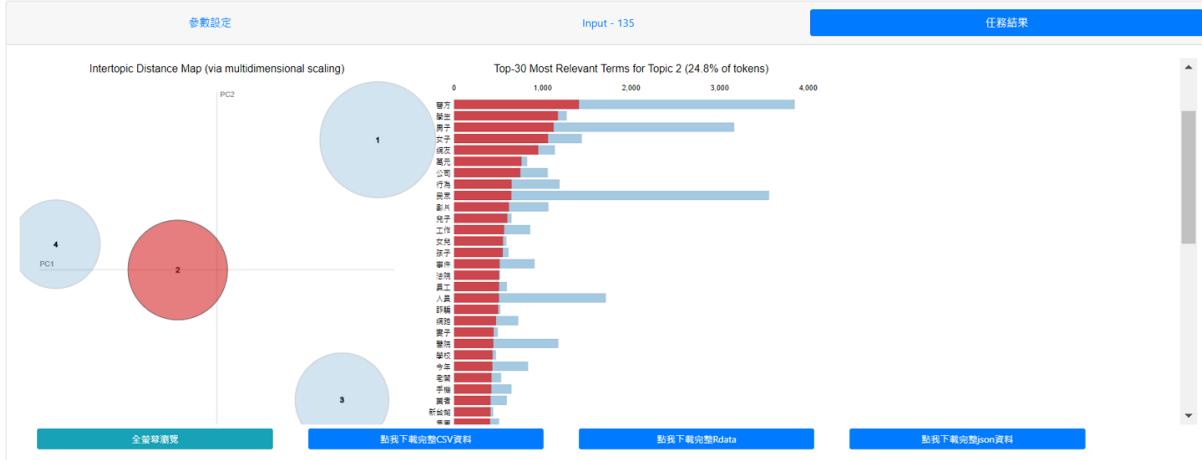
與未設主題種子字的 LDA 結果相似，主題 1 的關鍵字內容與選舉相關。從圖中可以發現設定主題種子的主題 1 圓圈相較其他主題大，表示主題內的文本量較多。此外，選舉相關文章在內容上與其他類別的文章有明顯距離，推測有可能是政治相關的專有名詞和候選人名字多，與其他類別的內容重疊較少。理論上新聞詞彙分佈廣泛，但從主題 1 的藍色和紅色詞彙的比例可見，這些詞彙很少出現在其他文章類別中。



主題 2

設定種子字的LDA主題2與未設定種子字的LDA主題4相似，難以看出實質分類意義。但是，其圓圈相較於未設定種子字的圓圈更大，顯示該分類文本量反而變得更多

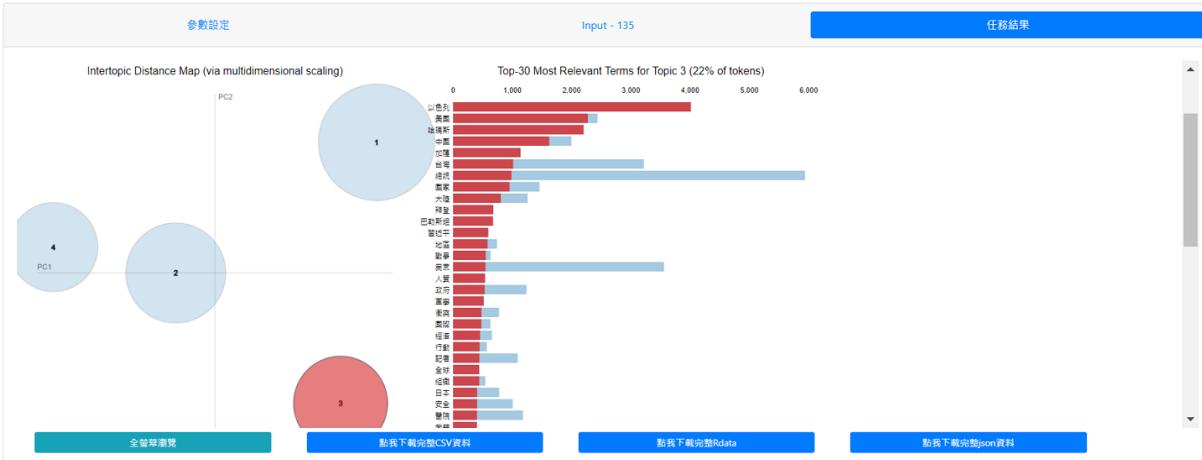
■ GuidedLDA 主題模型 (109)



主題 3

主題 3 的判斷結果與我們自定義的戰爭類別字典相近。不過，關鍵詞如「總統」、「國家」、「大陸」及「民眾」，由於也與選舉主題相關，因此出現在兩個不同的主題中。推測可能是資料時間為以巴衝突及台灣總統大選，在國際政治上，美國、大陸、國家、總統皆有相關性，故藍色條狀分佈出現比率相對高。

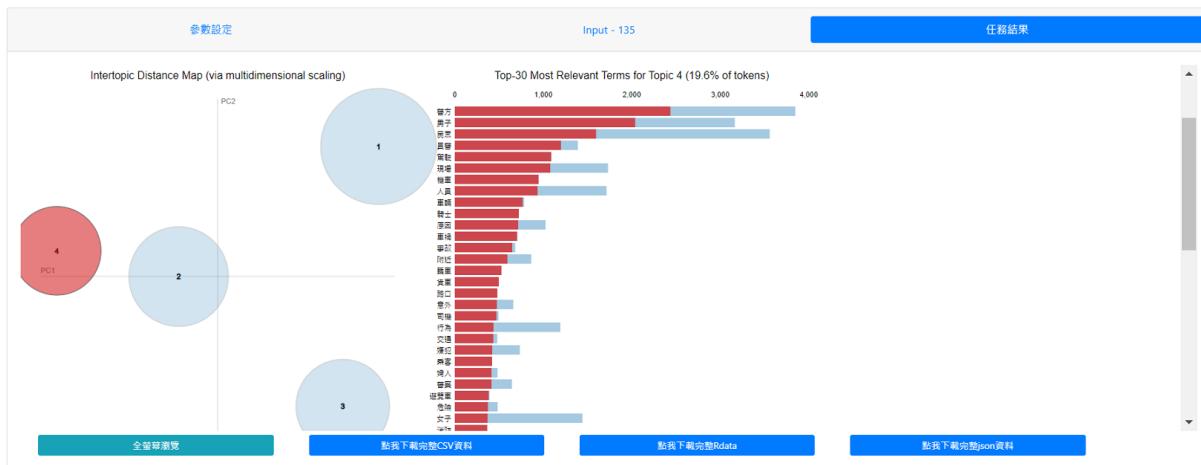
三 GuidedLDA 主題模型 (109)



主題 4

在設定種子字的 LDA 中，主題 4 與未設定種子字的主題 2 相似，但顯示的關鍵字更集中於交通事故。相比之下，未設定種子字的主題中仍含有如「詐騙」、「萬元」等不相關字詞。設定種子字後，主題聚焦於交通事故，且該圓圈大小縮小，表示涵蓋的文本量變少。

二 GuidedLDA 主題模型 (109)



10-2 新增欄位 topic

計算其所屬分類主題矩陣，取最大值(max)判斷為該所屬 topic

三 新增欄位 (idxmax/min) (151)

參數設定 Input - 109 任務結果

匯總函數 *

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

system_id
0
1
2
3

新增的欄位名稱 *

topic

儲存更改

三 新增欄位 (idxmax/min) (151)

參數設定 Input - 109 任務結果

任務結果

Show 10 entries

system_id	0	1	2	3	topic
1	0.982614	0.015143	0.001122	0.001122	0
2	0.002088	0.002088	0.002088	0.993737	3
3	0.996974	0.001009	0.001009	0.001009	0
4	0.761006	0.197589	0.001048	0.040356	0
5	0.994704	0.001765	0.001765	0.001765	0
6	0.691271	0.001284	0.001284	0.306162	0
7	0.997154	0.000949	0.000949	0.000949	0
8	0.603774	0.001048	0.001048	0.394130	0
9	0.996771	0.001076	0.001076	0.001076	0

全螢幕瀏覽 點我下載完整CSV資料 點我下載完整Rdata 點我下載完整json資料

10-3 分群彙總

計算(count)四類主題(0123)之文章數量

分群彙總 (非數值) (153)

參數設定

使用...欄位進行分群(按住ctrl(Windows)或command(MAC)可以複選) *

system_id
0
1
2
3
topic

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

system_id
0
1
2
3
topic

Input - 151

匯總函數 * ⓘ

count
nunique
min
max
first
last
sum

任務結果

儲存更改

分群彙總 (非數值) (153)

參數設定

統計資訊

4
群組數量

Input - 151

任務結果

Show **10** entries

topic	system_id@count
0	1719
1	1197
2	1673
3	1703

Search:

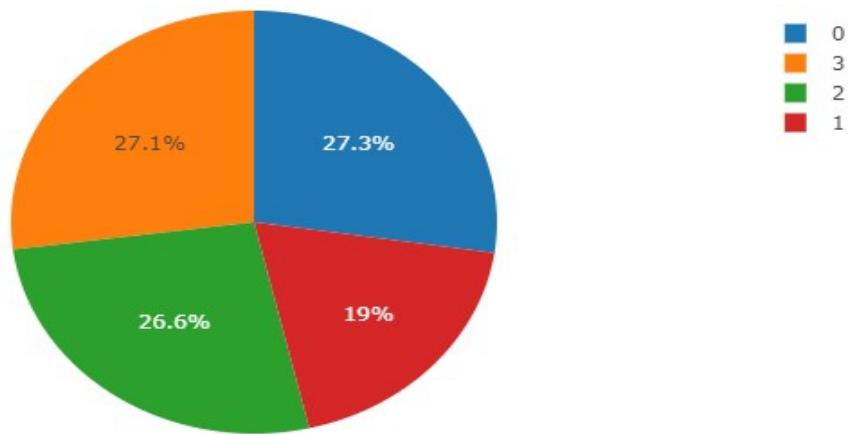
[全螢幕檢覽](#)

[點我下載完整CSV資料](#)

[點我下載完整RData](#)

[點我下載完整json資料](#)

10-4 視覺化儀表板呈現結果



小結

在給予種子字後，GuidedLDA 主題分群與東森新聞版分類比對如下，大致還是可以分出主題，如 0 代表政治新聞、1 代表國際新聞、2 代表社會新聞，分群效果與未給種子字的 LDA 模型有差異，**分群效果不如 LDA 犀利**。

東森新聞 版分類	GuidedLDA 主題模型分類				
	0	1	2	3	總計
社會	5	19	1568	895	2487
政治	1703	59	11	128	1901
國際	11	1119	94	680	1904
總計	1719	1197	1673	1703	6292