第8組

第3次讀書會報告



組長:N104020023林義行

組員:N104020018謝明和

N104020020李宜臻 N104020021馮慧嬌 N104020022凌嫈琪 N104020024林品均 N104020025陳姿樺

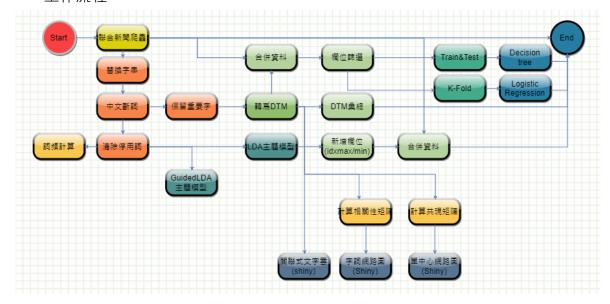
指導老師:黃三益教授

2023.05

目錄

_ `	工作流程
Ξ,	資料前處理9
(–)	替换字串9
(=)	中文斷詞9
(≡)	清除停用字10
(四)	詞頻分析10
(五)	分類器11
三、	文章主題分析:14
(–)	LDA 主題模型14
(=)	Guided LDA 主題模型19
四、	視覺化呈現25
(–)	關聯式文字雲25
(=)	字詞網路圖26
(≡)	單中心網路圖
五、	結論29

一、工作流程



(流程檔案名稱:0507_8-3)

資料來源:選擇工作平台資料集「聯合新聞爬蟲」,期間設定為 2023/03/01~2023/03/31,因為資料量較大,故資料以運動、全球及產經看板為設定參數,主題為近期討論度較高的三個議題分別為經典賽、烏俄衝突、普發 6000。

自定義字典:

name	class	alias
烏俄戰爭	戰爭	烏俄戰爭 俄烏戰爭
歐洲糧倉	戰爭	歐洲糧倉
烏克蘭	戰爭	烏克蘭 烏國
俄羅斯	戰爭	俄羅斯
頓內茭克	戰爭	頓內茨克
美國	戰爭	美國
入侵	戰爭	入侵
白俄羅斯	戰爭	白俄羅斯
核能武器	戰爭	核能武器
人道救援	戰爭	人道救援
軍事援助	戰爭	軍事援助 軍事 援助
經濟制裁	戦争	經濟制裁
天然資源	戰爭	天然資源

傭兵	戰爭	傭兵
馬立波	戰爭	馬立波
北約	戰爭	北約
莫斯科	戰爭	莫斯科
克理姆林宮	戰爭	克理姆林宮
普丁	戰爭	普亭 普丁
普亭	戰爭	普亭 普丁
澤倫斯基	戰爭	澤倫斯基
俄羅斯屠夫	戰爭	俄羅斯屠夫
馬立波屠夫	戰爭	馬立波屠夫
瓦格納集團	戰爭	瓦格納集團
普里格津	戰爭	普里格津
敘利亞屠夫	戰爭	敘利亞屠夫
德沃爾尼科夫	戰爭	德沃爾尼科夫
特別軍事行動	戰爭	特別軍事行動
美國	戰爭	美國
俄軍	戰爭	俄軍
戰爭	戰爭	戰爭
總統	戰爭	總統
制裁	戰爭	制裁
入侵	戰爭	入侵
基輔	戰爭	基輔
衝突	戰爭	衝突
歐洲	戰爭	歐洲
戰火	戰爭	戰火
英國	戰爭	英國
獨立	戰爭	獨立
卅年	戰爭	卅年
拜登	戦争	拜登
攻擊	戰爭	攻擊 攻打

歐盟	戰爭	歐盟
部隊	戰爭	部隊
國防部	戰爭	國防部
能源	戰爭	能源
世界	戰爭	世界
佐佐木朗希	運動	佐佐木朗希
美津濃	運動	美津濃
王柏融	運動	王柏融
多明尼加	運動	多明尼加
經典賽	運動	經典賽
WBC棒球經典賽	運動	WBC 棒球經典賽 WBC
世界棒球經典賽	運動	世界棒球經典賽
國際棒球賽	運動	國際棒球賽
美國職棒大聯盟	運動	美國職棒大聯盟
世界棒球經典賽會內資格賽	運動	世界棒球經典賽會內資格賽
世界棒球經典賽預賽	運動	世界棒球經典賽預賽
世界棒球經典賽分組賽	運動	世界棒球經典賽分組賽
2023 經典賽	運動	2023 經典賽
大谷翔平	運動	大谷翔平
投手	運動	投手
Mariano Rivera	運動	Mariano Rivera 李維拉
張育成	運動	張育成 部長 國防部長
布萊勒文	運動	布萊勒文 Bert Blyeven
皮耶薩	運動	皮耶薩 Mike Piazza
中華隊	運動	中華隊
台中洲際棒球場	運動	台中洲際棒球場
中華隊教練團	運動	中華隊教練團
棒球	運動	棒球
比賽	運動	比賽
賽事	運動	賽事

球員 運動 球音 預奪 運動 預奪 或翰 運動 戰餘 球隊 運動 球隊 資數 球 要報 球逃 運動 球逃 選事 大聯盟 MLB 大聯盟 運動 大發 運動 先發 運動 比分 運動 上分 安打 運動 安打 聯盟 運動 海軍 記事 巴拿馬 上場 運動 上場 運動 運動 運動 連動 教練 運動 教練 日本隊 運動 本區 出資 日本隊 運動 本區 運動 棒球場 職株 運動 構株球場 運動 古巴隊 總教練 運動 古巴隊 總教練 運動 也數教練 本季 運動 本季			
戰績 運動 戰隊 賽程 運動 賽程 球速 運動 球速 選手 運動 送手 大聯盟 運動 大聯盟MLB 分組 運動 先發 比分 運動 先發 比分 運動 比分 安打 運動 安打 聯盟 運動 房軍 巴拿馬 運動 巴拿馬 主場 運動 主場 運動 運動 養級 對身賽 運動 景級 香級 運動 青級 教練 運動 本局 日本隊 運動 本局 出賽 運動 本局 銀數 棒球場 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 運動 60 總教練 運動 60	球員	運動	球員
球隊 運動 球隊 賽程 運動 球迷 球並 運動 球迷 選手 運動 大聯盟 MLB 分組 運動 大聯盟 MLB 分組 運動 先發 比分 運動 安打 聯盟 運動 港軍 巴拿馬 運動 巴拿馬 主場 運動 上場 運動 運動 主場 運動 主場 運動 教身賽 運動 無身賽 音級 運動 日本隊 本屆 運動 本屆 出賽 運動 本屆 出賽 運動 棒球場 職棒 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 總教練 運動 為教練	預賽	運動	預賽
賽程 運動 賽程 球迷 運動 球迷 選手 運動 大聯盟 MLB 分組 運動 大聯盟 MLB 分組 運動 先發 比分 運動 比分 安打 運動 安打 聯盟 運動 冠動 起軍 巴拿馬 運動 主場 運動 主場 運動 運動 養動 熱身賽 運動 合級 教練 運動 台級 教練 運動 日本隊 本屆 運動 本屆 出賽 運動 本屆 出賽 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 運動 右巴隊 總教練 運動 總數	戰績	運動	戰績
球迷 運動 球迷 選新 大聯盟 大聯盟 MLB 分組 運動 今組 光發 運動 先發 比分 運動 比分 安打 運動 安打 聯盟 運動 房軍 巴拿馬 運動 巴拿馬 主場 運動 主場 運動 運動 連動 教練 運動 計算 日本隊 運動 本屆 出賽 運動 本屆 出賽 運動 本屆 出賽 運動 本屆 大樓球場 運動 棒球場 職棒 運動 本品 大門隊 運動 本品 大門 東面 大門 大門 大門 東面 大門 大門 大門 東面 大門 大門 大門 東面 大門 大門 大門 東面 大門 大門 </td <td>球隊</td> <td>運動</td> <td>球隊</td>	球隊	運動	球隊
選手 大聯盟 天聯盟 MLB	賽程	運動	賽程
大聯盟 運動 大聯盟 MLB 分組 運動 分組 先發 運動 先發 比分 運動 比分 安打 運動 安打 聯盟 運動 房 電車 運動 一个 电拿馬 運動 上場 運動 運動 連動 大學 運動 一个 大學 運動 大學 日本隊 運動 本屆 出審 運動 本屆 出審 運動 「日本隊 「中央 運動 「日本隊 大田市 運動 「日本隊 日本隊 運動 「日本隊 日本隊 「運動 「日本隊 日本隊 「運動 「日本隊 日本球場 「運動 「日本球場 日本球場 「日本球場 「日本球場 「日本球場 日本球場 「日本球場 「日本球場 「日本球場 日本球場 「日本球場 「日本球場 「日本球場 日本球 「日本球場 「日本球	球迷	運動	球迷
分組 運動 分組 先發 運動 先發 比分 運動 比分 安打 運動 安打 聯盟 運動 房 建動 房 房 建動 建動 建動 連動 登 日本 本島 運動 本島 出審 運動 本島 出審 運動 本島 山審 運動 本島 大島 運動 大島 大島 運動 大島 大島 大島 大島 大島	選手	運動	選手
先發 運動 先發 比分 運動 比分 安打 運動 安打 聯盟 運動 聯盟 冠軍 運動 冠軍 巴拿馬 運動 上場 運動 運動 運動 熱身賽 運動 台級 教練 運動 日本隊 本屆 運動 日本隊 本屆 運動 出賽 領先 運動 標本球場 職棒 運動 荷蘭隊 古巴隊 運動 總教練	大聯盟	運動	大聯盟 MLB
比分 運動 比分 安打 運動 安打 聯盟 運動 聯盟 冠軍 運動 冠事 巴拿馬 運動 主場 運動 運動 運動 熱身賽 運動 音級 教練 運動 音級 技術 運動 日本隊 本屆 運動 本屆 出賽 運動 出賽 領先 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 運動 右巴隊 總教練 運動 總教練	分組	運動	分組
安打 運動 安打 聯盟 運動 聯盟 冠軍 運動 冠馬 建動 巴拿馬 主場 運動 主場 運動 運動 熱身賽 音級 運動 音級 教練 運動 日本隊 本屆 運動 日本隊 本屆 運動 出賽 領先 運動 標本球場 職棒 運動 構体 荷蘭隊 運動 古巴隊 總教練 運動 總教練	先發	運動	先發
聯盟 運動 聯盟 冠軍 運動 冠軍 巴拿馬 運動 巴拿馬 主場 運動 主場 運動 連動 熱身賽 晉級 運動 晉級 教練 運動 日本隊 本屆 運動 本屆 出賽 運動 本屆 領先 運動 棒球場 職棒 運動 棒球場 荷蘭隊 運動 古巴隊 總教練 運動 總教練	比分	運動	比分
冠軍 運動 冠軍 巴拿馬 運動 巴拿馬 主場 運動 運動 熱身賽 運動 熱身賽 晉級 運動 晉級 教練 運動 教練 日本隊 運動 本屆 出賽 運動 本屆 出賽 運動 領先 棒球場 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 運動 總教練	安打	運動	安打
巴拿馬 運動 已拿馬 主場 運動 主場 運動 運動 熱身賽 晉級 運動 晉級 教練 運動 教練 日本隊 運動 日本隊 本屆 運動 本屆 出賽 運動 場先 標本球場 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 運動 總教練	聯盟	運動	聯盟
主場 運動 主場 運動 運動 運動 熱身賽 運動 熱身賽 晉級 運動 晉級 教練 運動 大本區 山賽 運動 本區 出賽 運動 出賽 領先 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 運動 總教練	冠軍	運動	冠軍
運動 運動 熱身賽 運動 晉級 運動 音級 運動 老線 運動 日本隊 運動 本屆 運動 出賽 運動 領先 運動 棒球場 運動 職棒 運動 荷蘭隊 運動 古巴隊 運動 總教練 運動	巴拿馬	運動	巴拿馬
熱身賽 運動 熱身賽 晉級 運動 晉級 教練 運動 教練 日本隊 運動 日本隊 本屆 運動 本屆 出賽 運動 出賽 領先 運動 棒球場 職棒 運動 職棒 荷蘭隊 運動 古巴隊 總教練 運動 總教練	主場	運動	主場
晉級 運動 晉級 教練 運動 教練 日本隊 運動 日本隊 本屆 運動 本屆 出賽 運動 出賽 領先 運動 標財 棒球場 運動 職棒 荷蘭隊 運動 古巴隊 總教練 運動 總教練	運動	運動	運動
教練 運動 教練 日本隊 運動 日本隊 本屆 運動 本屆 出賽 運動 出賽 領先 運動 棒球場 職棒 運動 職棒 荷蘭隊 運動 荷蘭隊 古巴隊 運動 想教練	熱身賽	運動	熱身賽
日本隊 運動 日本隊 本屆 運動 本屆 出賽 運動 出賽 領先 運動 棒球場 職棒 運動 職棒 荷蘭隊 運動 古巴隊 總教練 運動 總教練	晉級	運動	晉級
本屆 運動 本屆 出賽 運動 出賽 領先 運動 領先 棒球場 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 運動 古巴隊 總教練 運動 總教練	教練	運動	教練
出賽 運動 出賽 領先 運動 領先 棒球場 運動 職棒 荷蘭隊 運動 荷蘭隊 古巴隊 運動 古巴隊 總教練 運動 總教練	日本隊	運動	日本隊
領先 運動 領先 棒球場 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 運動 古巴隊 總教練 運動 總教練	本屆	運動	本屆
棒球場 運動 棒球場 職棒 運動 荷蘭隊 古巴隊 運動 古巴隊 總教練 運動 總教練	出賽	運動	出賽
職棒 運動 職棒 荷蘭隊 運動 荷蘭隊 古巴隊 運動 古巴隊 總教練 運動 總教練	領先	運動	領先
荷蘭隊 運動 荷蘭隊 古巴隊 運動 古巴隊 總教練 運動 總教練	棒球場	運動	棒球場
古巴隊 運動 古巴隊 總教練 運動 總教練	職棒	運動	職棒
總教練 運動 總教練	荷蘭隊	運動	荷蘭隊
	古巴隊	運動	古巴隊
本季 運動 本季	總教練	運動	總教練
	本季	運動	本季

144 '全	雷利	11枚 宝
聯賽	運動	聯賽
荷蘭	運動	荷蘭
義大利	運動	義大利
義大利隊	運動	義大利隊
登場	運動	登場
球星	運動	球星
洲際	運動	洲際
打擊	運動	打擊
球團	運動	球團
塞佩達斯	運動	塞佩達斯
孟卡達	運動	孟卡達
先發	運動	先發 先發投手
雙殺	運動	雙殺
三振	運動	三振
全壘打	運動	全壘打 轟
外野	運動	中外野 左外野 右外野
內野	運動	一壘 二壘 三壘 游擊
富邦悍將	運動	富邦悍將
普發 6000	產經	普發 6000
普發現金	產經	普發現金
6000	產經	6000
普發	產經	普發
6 🕇	產經	6 🕇
全民共享	產經	全民共享
六千	產經	六千
郵局領現	產經	郵局領現
ATM 領現	產經	ATM 領現
全民普發	產經	全民普發
普發六千	產經	普發六千
超徵稅收	產經	超徵稅收

詐騙集團	產經	詐騙集團
時代力量	產經	時代力量
政府	產經	政府
台灣	產經	台灣
經濟	產經	經濟
全民	產經	全民
民眾	產經	民眾
預算	產經	預算
民進黨	產經	民進黨
立委	產經	立委
總統	產經	總統
立法院	產經	立法院
現金	產經	現金
發放	產經	發放
條例	產經	條例
蔡英文	產經	蔡英文
蘇貞昌	產經	蘇貞昌
代領	產經	代領
登記	產經	登記
管道	產經	管道
領取	產經	領取
線上登記	產經	線上登記
身份證	產經	身份證
金融機構	產經	金融機構
還稅於民	產經	還稅於民
刺激經濟	產經	刺激經濟
刺激消費	產經	刺激消費
審計	產經	審計
舉債	產經	舉債
發現金	產經	發現金

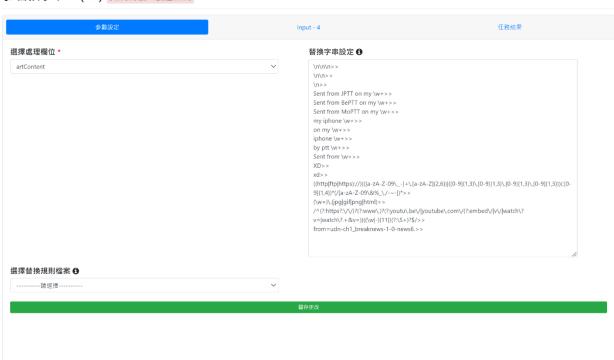
加碼	產經	加碼
中央	產經	中央
直接入帳	產經	直接入帳
造冊發放	產經	造冊發放
數位發展部	產經	數位發展部
1988	產經	1988
線上登記	產經	線上登記
分流	產經	分流
政策	產經	政策
紅包	產經	紅包
怎麼花	產經	怎麼花
通膨	產經	通膨
升息	產經	升息

二、資料前處理

(一) 替換字串

將空格與一些常出現較無意義的符號及網址做替換,讓搜尋出來的資料比較容易 閱讀。

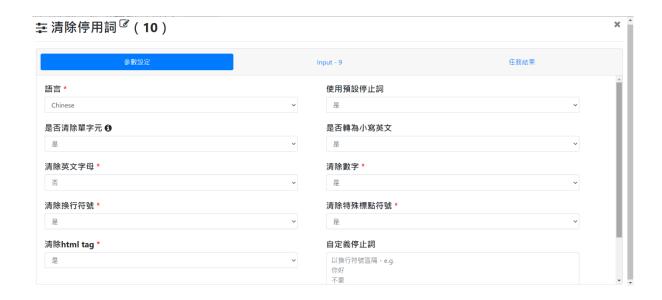
章 替換字串 ^② (7) 參數有做更動·建議重新執行



(二) 中文斷詞

至中文斷詞♂(9)		ж
多數設定	Input - 7	任務結果
選擇處理欄位 *	定義詞彙 ❸	
result	以換行符號區隔, e.g. 詞量 權重 國立中山大學 1000 西子灣 500	
選取字典		
Dri ACL I平	儲存更改	

(三)清除停用字



(四) 詞頻分析

討論度較高的字詞,即為相對重要的參數。同時也對應並改善我們定義的字典詞 彙。

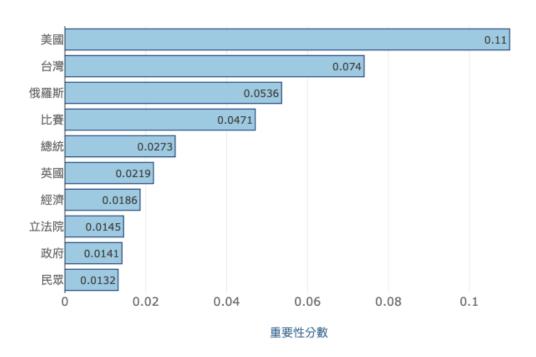


(五) 分類器

模型的準確度等參數、參數的重要性分數,以及重要的前幾個參數

1. Decision tree

參數重要性長條圖



統計資訊



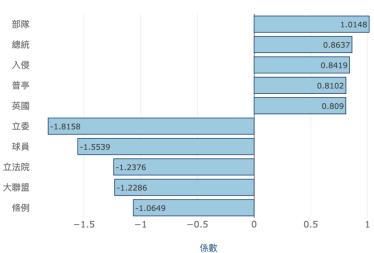
2. Logistic regression

根據本組不同主題分類,看前幾個重要的參數中分別接近0或1的分佈情形。

視覺化

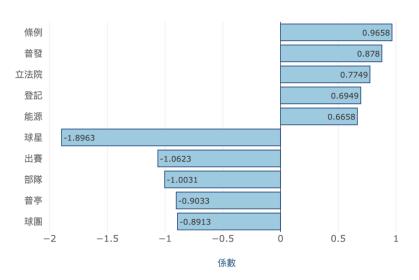
分類: 全球

正負係數前5名長條圖



分類: 產經

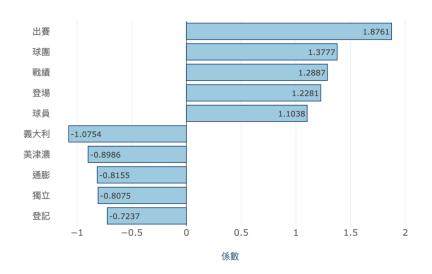
正負係數前5名長條圖



12

分類: 運動 ~

正負係數前5名長條圖



統計資訊



三、文章主題分析:

(一) LDA 主題模型

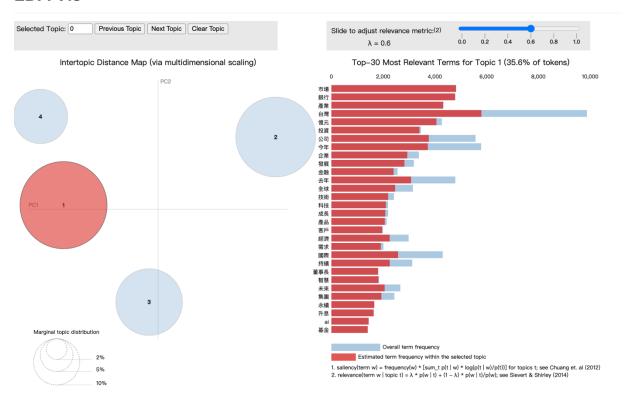
根據事先定義好的主題,用 LDA 模型觀察新聞內容的分佈,並同時觀察是否有其他主題出現。

幸 LDA主題模型 [☑] (17)



1. 第1主題:

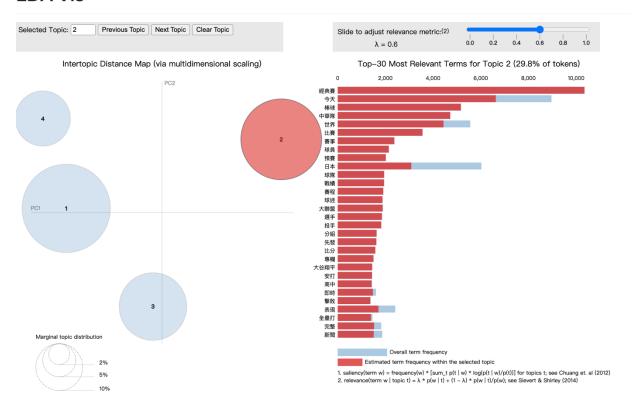
LDA Vis



與第1主題相關的文章中,出現較多的關鍵字如「市場」、「銀行」、「產業」,有些關鍵字如「台灣」出現在較多主題分類中的文章中,因此鑑別度較低。由這些關鍵字看來,第1群主題是跟「產業經濟」相關主題的內容。

2. 第 2 主題:

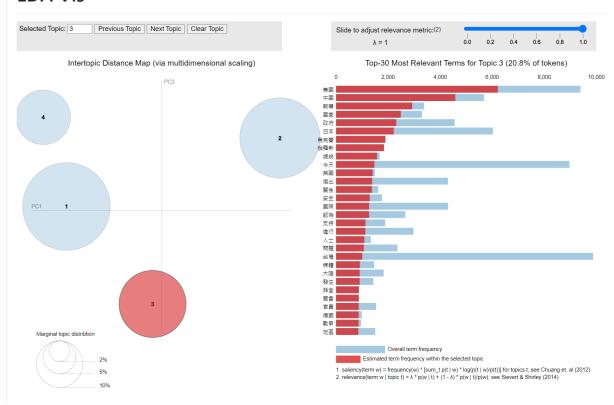
LDA Vis



與第2主題相關的文章中,出現較多的關鍵字如「經典賽」、「棒球」、「中華隊」,有些關鍵字如「今天」、「世界」、「日本」出現在較多主題分類中的文章中,因此鑑別度較低,然而此主題分類中,大部分關鍵字都跟我們所想要看的主題很相關,所以從這些關鍵字可以很明確找到我們所想要的主題文章。由這些關鍵字來看,第2群主題是跟「經典賽」相關主題的內容。

3. 第3主題:

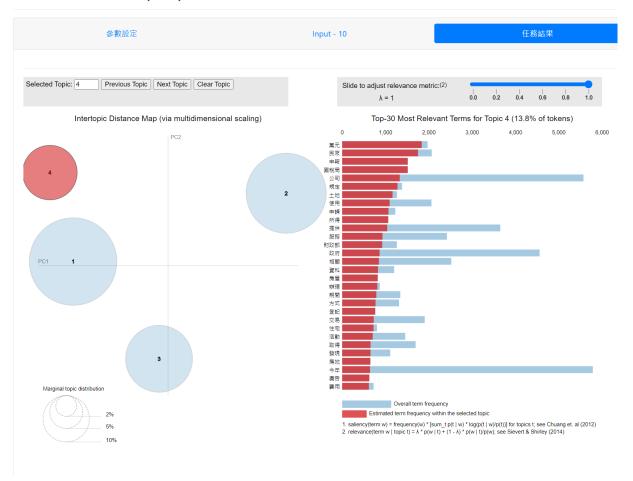
LDA Vis



與第3主題相關的文章中,出現較多的關鍵字如「美國」、「中國」、「報導」等,有些關鍵字如「日本」、「政治」、「美國」等出現在較多主題分類中的文章中,因此鑑別度較低。有些出現頻率較低的字詞,在其他主題出現的機率也較低,如「烏克蘭」、「俄羅斯」、「軍事」等,反而比詞頻高的字詞有鑑別度。由這些關鍵字來看,可以從中找出與我們所希望看到的主題的共同點,由此可知第3群主題是跟「俄烏戰爭」相關主題的內容。

4. 第 4 主題:

註 LDA主題模型 ^② (17)



與第4主題相關的文章中,出現較多的關鍵字如「國稅局」、「申報」等,有些關鍵字如「公司」、「使用」等也會出現在其他主題分類中的文章中,因此鑑別度較低,若要以有出現這些字詞來判斷文章是否屬於此主題則相對不易。第4主題及第1主題相關的文章較有相關性,從這些字詞來細看,第1群主題是跟「產業經濟」相關主題的內容,第4群文章數量較少,且較限縮在經濟範圍中討論「財政」、「稅務」方面的文章。

(二) Guided LDA 主題模型

我們對於看板內其他新聞內容,對新聞做主題分類。設定主題種子字,作為導引其他新聞主題的依據。一開始設定的看板有全球、運動、與生活三大類,但我們自訂的分類主題為戰爭、運動、與生活。因此在主題種子字的設定中,秉持與三大類主題相關的重點關鍵字做為種子主題字依序為:

- A. 經典賽,日本,大聯盟,大谷翔平,先發,中華隊
- B. 烏克蘭,俄羅斯,美國,政府,戰爭
- C. 普發,升息,通膨,登記,回饋

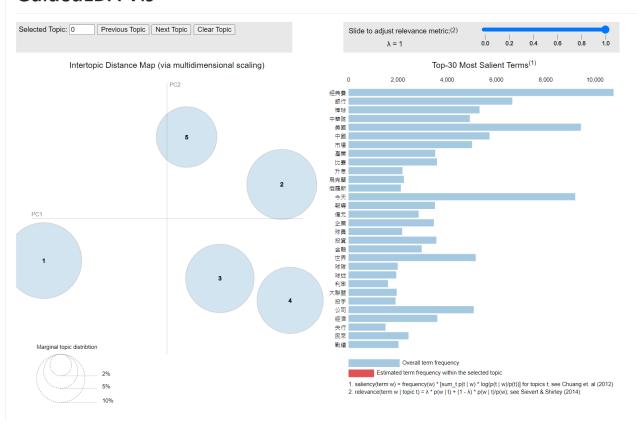
並設定主題數為5個

幸 GuidedLDA 主題模型[☑](15)

參數設定	Input - 10	任務結果
目標欄位 *	迭代次數	
result	50	
主題數 *	主題保留關鍵字數量	
5	20	
詞彙頻率下限 ❶	詞彙頻率上限 €	
30	0.6	
alpha	Beta	
預設為主題數/50	預設為0.1	
主題種子字❶	是否輸出字典	
經典賽,日本,大聯盟,大谷翔平,先發,中華隊 烏克蘭,俄羅斯,美國,政府戰爭	是	>
普發,升息,通膨,登記,回饋		
	le la	

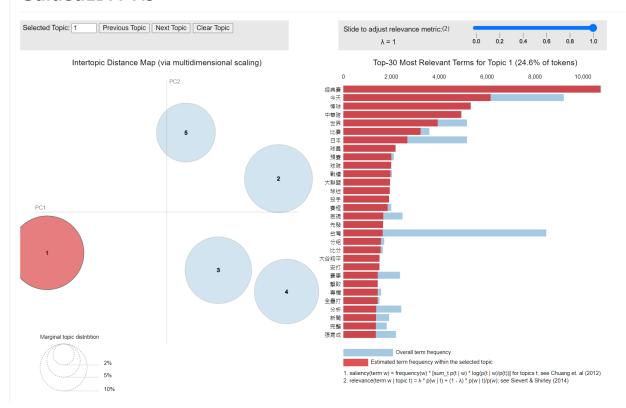
任務結果如下圖所示:

GuidedLDA Vis



1. Guided LDA 第 1 主題:

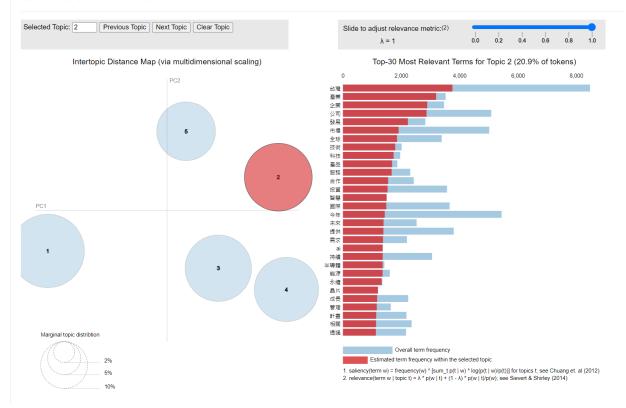
GuidedLDA Vis



第1主題以「經典賽」、「中華隊」、「棒球」等相關詞彙,主題可以明確歸納世界經典棒球賽,資料期間正是 WBC 世界棒球經典賽比賽期間,所有新聞幾乎都圍繞在WBC 世界棒球經典賽在討論,由於棒球是台灣國球,最容易激起並團結台灣全民的情緒,全民一心一起為中華台北棒球隊加油。因此期間運動看板中報導最多的就是經典賽棒球的新聞。

2. 第2主題:

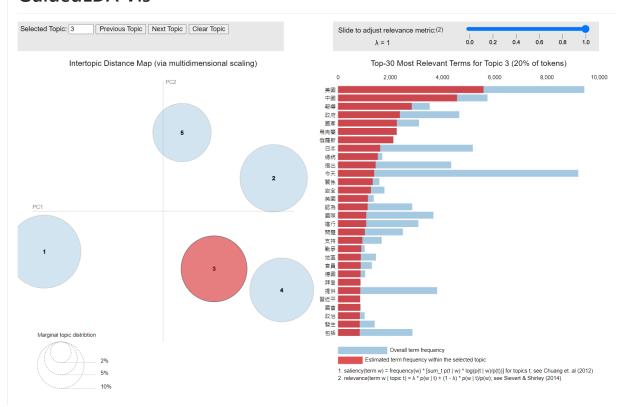
GuidedLDA Vis



第2主題從「企業」、「產業」、「公司」等相關詞彙,不難看出本主題主要敘述產業企業的發展,台灣的護國神山半導體、台積電等也有出現在裡面,而充滿紅色區塊的亦即未出現在其他主題的,包含 ai 智慧永續創新,主要談論企業未來發展,智慧應該是指人工智慧,斷詞可再加重權重可能會更精準。

3. 第3主題:

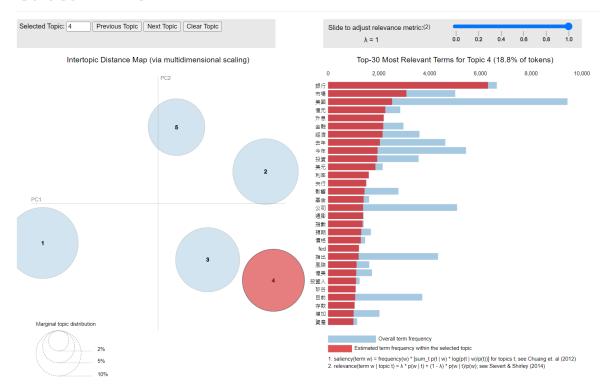
GuidedLDA Vis



第 3 主題中可以看到紅色佔滿全部的包含烏克蘭、俄羅斯、軍事及援助等,亦即 這幾個詞彙未出現在其他文章中,可看出主要是在議論烏俄戰爭,比較特別的是拜登 與習近平兩個詞彙也有相同的情況,推測是新聞中兩國總統發言主要跟烏俄戰爭比較 有關係。

4. 第4主題:

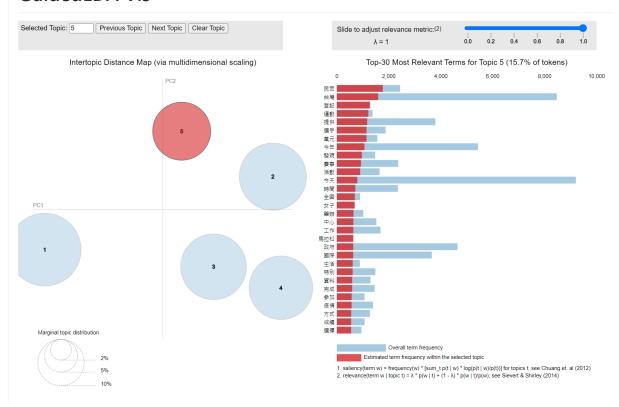
GuidedLDA Vis



第4主題主要是講述全球經濟的部分,其中重要詞彙包含「聯準會」、「升息」、「FED」、「通膨」等等,而「美國」雖然紅色佔總體藍色比例不高,但主因是它對很多主題均有影響,因此在本主題中亦很重要,而非不重要,而對比的是「今年」這個詞彙也有這個狀況,不過「今年」這個詞彙就比較通用,相較之下在本主題來說沒這麼重要。

5. 第5主題:

GuidedLDA Vis



第 5 主題比較看不出來彼此之間的關聯性,內容較為繁雜,其中包含運動項目如:馬拉松、亞錦賽;民生方面如:普發、民眾、房屋等等,因此比較像是前幾個主題以外的其他類主題。

四、視覺化呈現

(一) 關聯式文字雲

進行完主題區分之後,對照關聯式文字雲可以看出明顯的分群,即每一個主題都 有一定的辨別度。

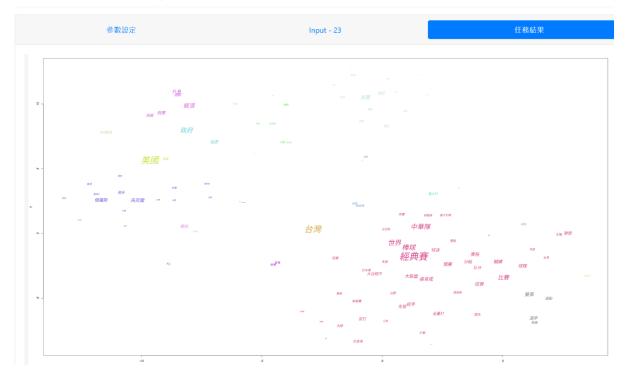
參數設定:

幸 關聯式文字雲 (shiny) [☑] (58)

多數設定	Input - 23	任務結果	
分群數 *	文字雲顯示字數 * 1000		
迭代次數(最少 250 次)	距離計算公式 * euclidean	V	
聚合演算法。 single	v		
留有更改			

呈現關聯式字雲圖:

幸 關聯式文字雲 (shiny) ^፪ (58)



(二) 字詞網路圖

在字詞網路圖中,除了主題區分明確,圖中每一個主題也可以讓我們更好地觀察 詞彙彼此間的關係。

熱圖:

幸 字詞網路圖 (Shiny) [☑] (65)



呈現單字詞網路圖

幸 字詞網路圖 (Shiny) ^፪ (65)



(三)單中心網路圖

使用單中心網路圖的互動性功能,可在次層詞類選擇字詞,再進一步觀察與它關聯的字詞網路,並可進一步抓出提到其字詞的文章。

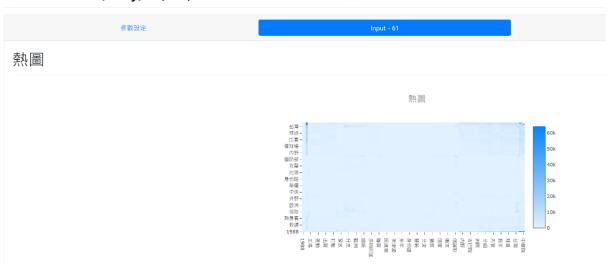
參數設定:

幸 單中心網路圖 (Shiny) ^② (64)

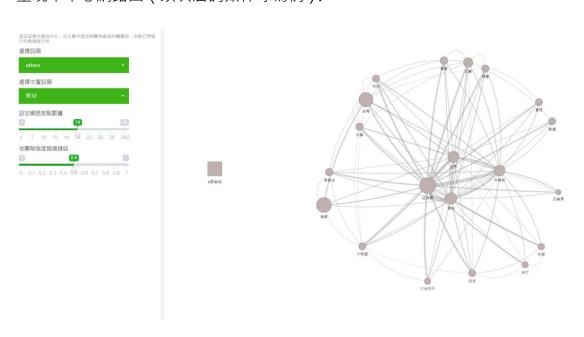


計算共現矩陣:

幸 單中心網路圖 (Shiny) [☑] (64)



呈現單中心網路圖 (以次層詞類棒球為例):



万、結論

- 1. 在一開始選擇資料來源時,文章主題要選差異較大的主題,否則比較不好區分文章分類。
- 2. 自定義字典的字詞量越多,對模型訓練越有幫助。
- 3. 產經和國際兩個主題中,有些字詞頻繁出現如「美國」、「中國」,雖然無法以此參數來判斷文章的主題,但可以想見,產經和國際等議題多數都與美國、中國兩大經濟體相關,兩者有強大的影響力。
- 4. LDA 及 GuidedLDA,雖都有區分主題的功能,但在增加主題種子字後, GuidedLDA 有較好的表現,1 到 4 的主題區分都很明確,且對與想要找出分析主題,或者找出關聯性當相當容易,不過在前處理及係數的設定上,可能還是要多嘗試才能更精準。
- 5. 我們在分析時,應該要對主題有一些深入的認識,才能再針對分析結果做細部人工調整,例如「美津濃」、「義大利」這兩個參數應該是跟經典賽高度相關主題,但做出來的重要性分數卻不高。