

中研院數位人文研究平台 使用紀錄與心得

NTNU 文本分析與程式設計 期末作業
41247013S 王修佑

資料說明

本次所使用的資料集分為「分析文本」與「權威檔文本」。

分析文本

- 公視新聞網 - 副刊報導
 - 年份：2023
 - 來源：<https://news.pts.org.tw/supplement>
- 公視新聞網 - 觀點新聞
 - 年份：2023
 - 來源：<https://news.pts.org.tw/opinion>
- 公視新聞網 - 以哈戰爭報導
 - 年份：2023
 - 來源：<https://news.pts.org.tw/hotTopic/280>

運用 Python requests 與 bs4 爬取成 pandas DataFrame 後，輸出為 xlsx 檔案，上傳人文研究平台匯入群組。

爬蟲程式檔案：https://github.com/whyhugo/Textual-Data-Analysis/blob/main/final/pts_report_crawler.ipynb

資料集：<https://github.com/whyhugo/Textual-Data-Analysis/tree/main/final/data>

權威檔文本

- 廣大之陸上自然地理區域-佛典規範地名資料庫
 - 來源：平台開放權威檔
- 廣大之陸上人文地理區域-佛典規範地名資料庫
 - 來源：平台開放權威檔
- 地點-佛典規範地名資料庫
 - 來源：平台開放權威檔
- 台灣地名
 - 來源：平台開放權威檔

實際操作紀錄

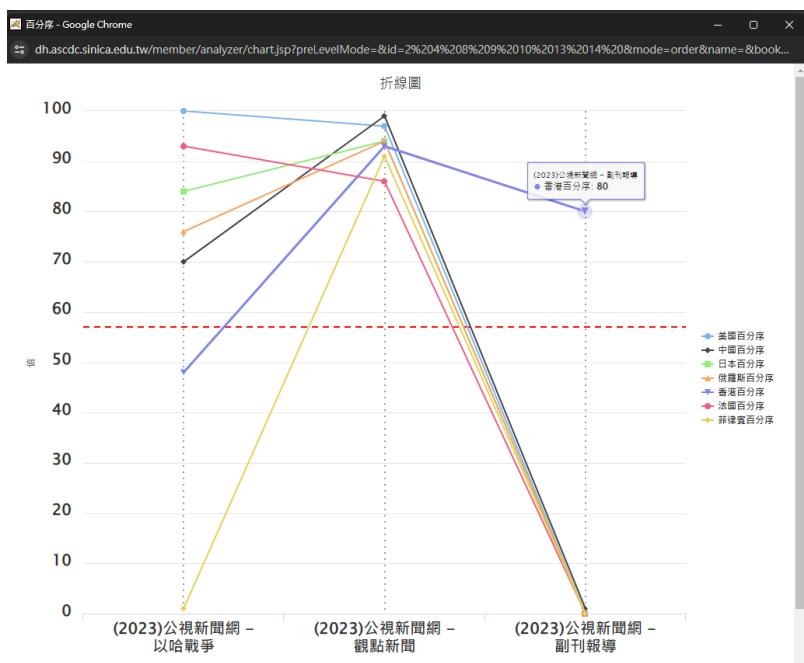
權威詞頻率分析與視覺化

1. 加入文本與權威詞

勾選比對文本與權威檔，即可出現此畫面，分析結果分為「百分序」與「百分序」兩種顯示方式。

2. 權威詞的百分序

例如，在以哈戰爭報導中「美國」的百分序為 100、「中國」為 70；在觀點報導中「美國」的百分序為 97、「中國」為 99。



以折線圖方式呈現勾選的單詞在各文本中的百分序變化，上圖以「美國」、「中國」、「日本」、「俄羅斯」、「香港」、「法國」與「菲律賓」為例。

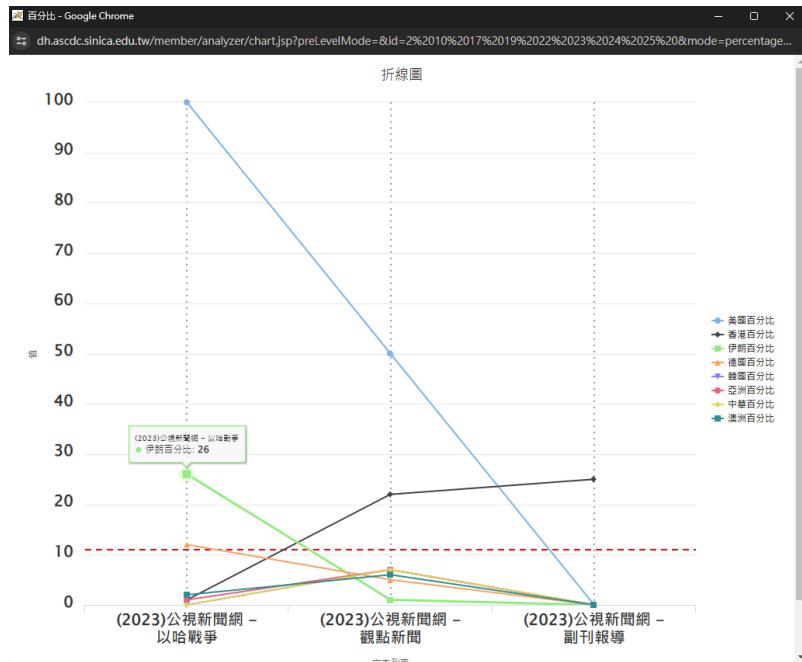
3. 權威詞的百分比

將該權威詞頻率，除以最高權威詞的頻率。

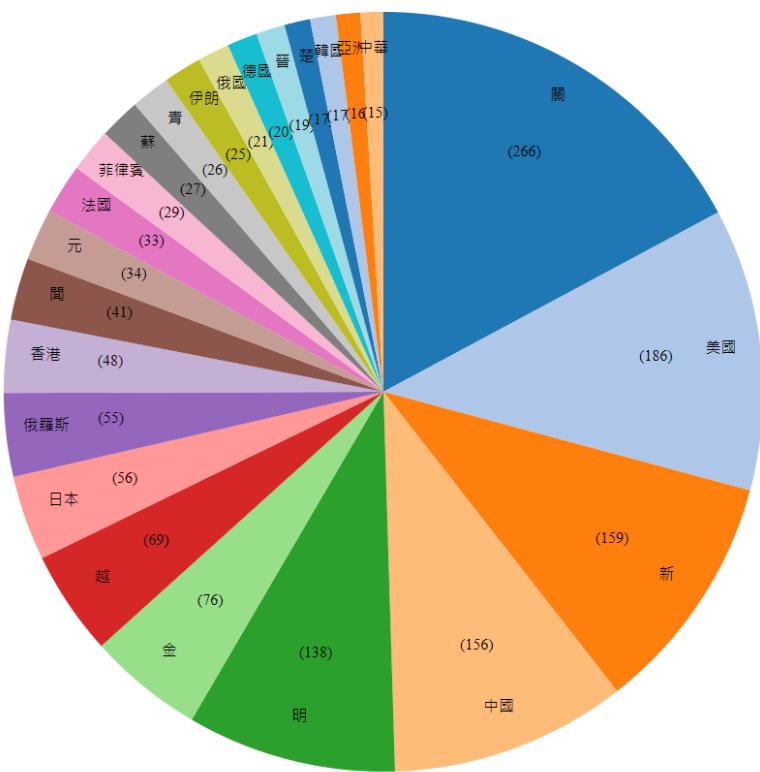
The screenshot shows a table titled '文本辭匯' (Text Vocabulary) with columns for '勾選' (Selected), '關鍵字' (Keyword), '平均詞頻' (Average Frequency), '平均百分比' (Average Percentage), '變異' (Variation), and several columns for different news sources from 2023. The table includes rows for '華語' (Chinese), '美國' (USA), '新' (New), '中華' (China), '俄' (Russia), '日' (Japan), '法' (France), '俄羅斯' (Russia), '香港' (Hong Kong), '元' (Yuan), '法國' (France), '菲律賓' (Philippines), and '日' (Japan). The '變異' column shows values ranging from 1 to 17. The '平均百分比' column shows values ranging from 0 to 100.

勾選	關鍵字	平均詞頻	平均百分比	變異	(AD 2023) 公視新聞網 - 以哈戰爭	百分比	(AD 2023) 公視新聞網 - 觀點新聞	百分比	(AD 2023) 公視新聞網 - 副刊報導	百分比
	華語	266	70	59	176	70	617	100	5	41
1	美	186	50	100	249	100	309	50	0	0
2	新	159	65	69	92	36	374	60	12	100
3	中	156	28	91	8	3	459	74	1	8
4	俄	138	35	37	90	38	321	52	2	16
5	日	76	19	28	21	8	206	33	2	16
6	法	69	23	35	14	5	190	30	4	33
7	日本	56	10	25	22	8	147	23	0	0
8	俄羅斯	55	10	29	14	5	150	24	0	0
9	香港	48	16	30	3	1	138	22	3	25
10	華	41	11	10	23	9	100	16	1	8
11	元	34	11	13	14	5	85	13	2	16
12	法國	33	9	22	52	20	47	7	0	0
13	菲律賓	29	5	19	1	0	87	14	0	0
14	日	27	8	17	26	10	55	8	1	0

點擊各個 column bar 上的箭頭即可切換升冪或降冪排序。也可輸入條件進行數值篩選。



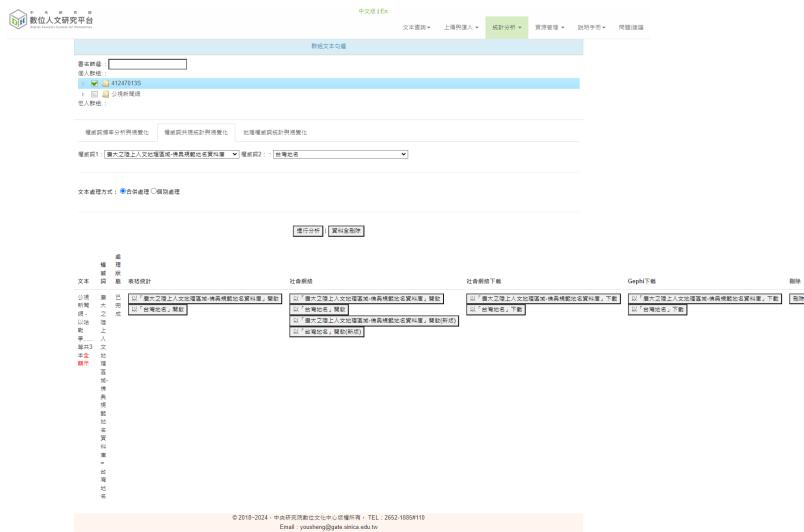
以折線圖方式呈現勾選的單詞在各文本中的百分比變化，上圖以「美國」、「伊朗」、「德國」、「韓國」、「香港」、「亞洲」、「澳洲」與「中華」為例。



亦可直接產生圓餅圖（也可特別選擇 Top 25）與文字雲。

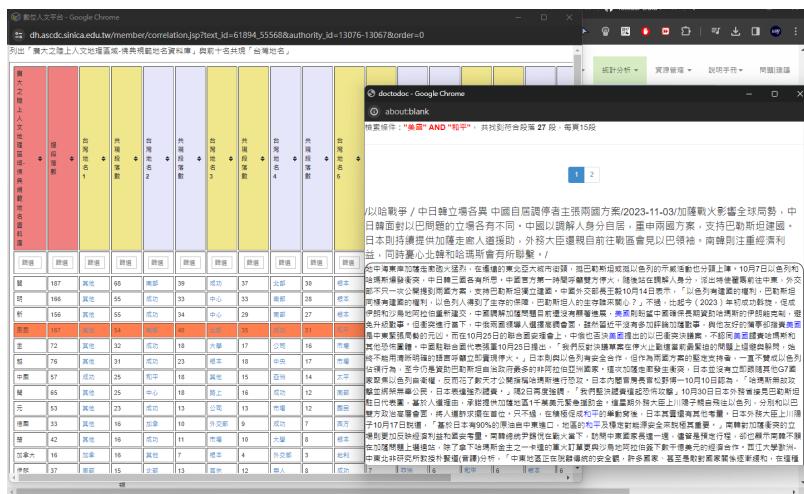
權威詞共現與視覺化

1. 加入文本與權威詞



我選擇「廣大之陸上人文地理區域-佛典規範地名資料庫」與「台灣地名」這兩個權威詞檔進行操作。

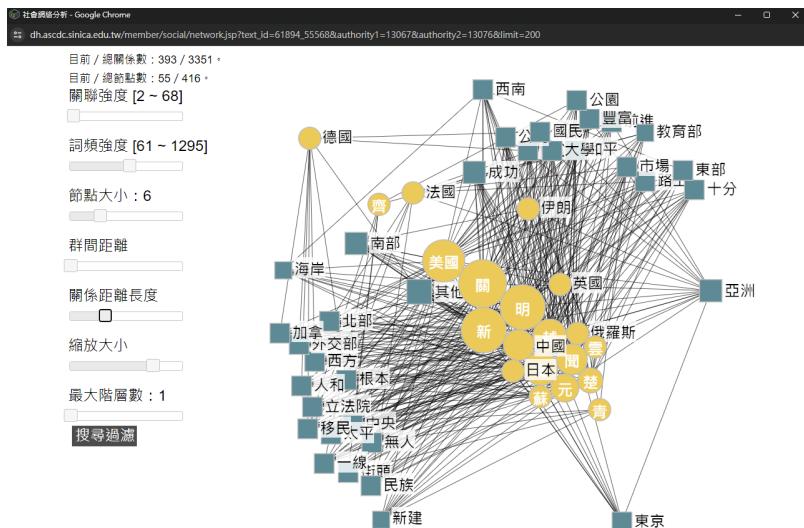
2. 以表格呈現分析結果

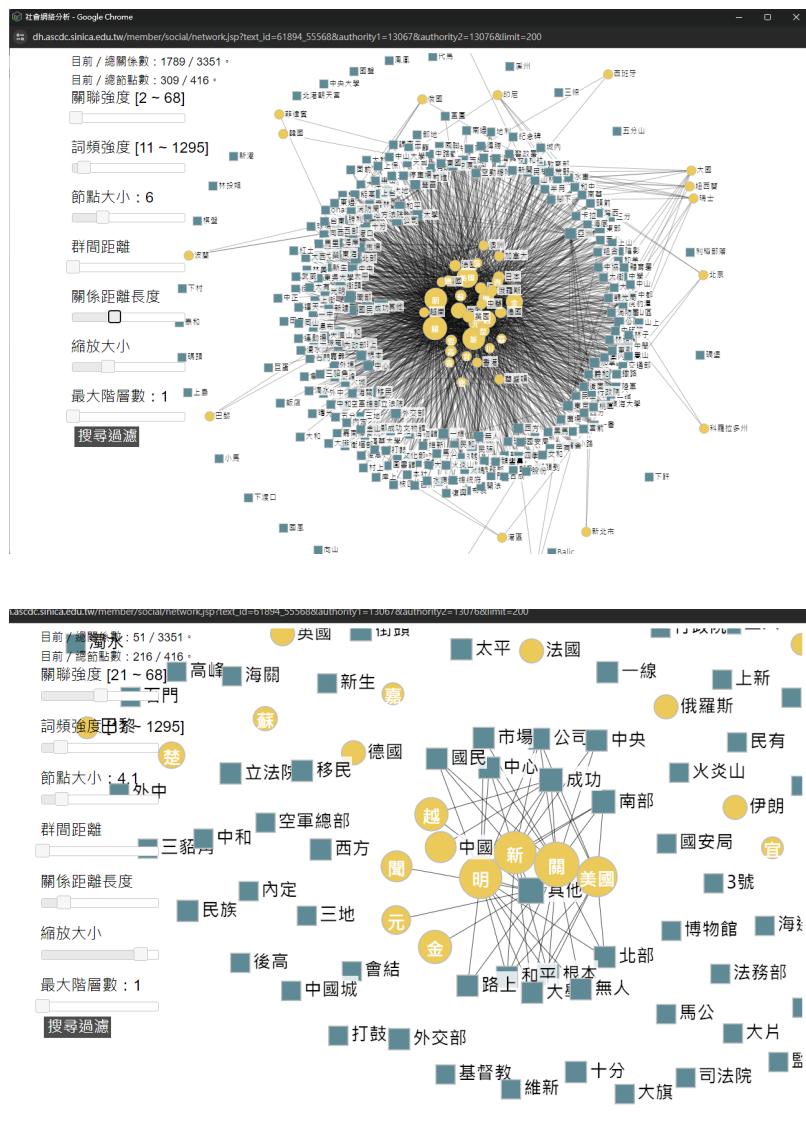


以表格方式列出「廣大之陸上人文地理區域-佛典規範地名資料庫」與前十名共現「台灣地名」，此處選擇檢索條件：「美國」AND「和平」，共找到符合段落 27 段，同時可直接檢視原始文本。

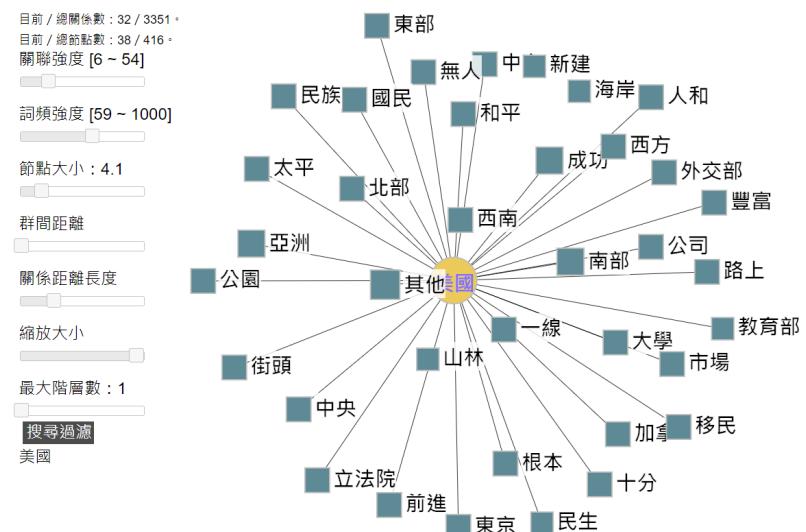
3. 社會網絡視覺化

可以透過調整關聯強度、詞頻大小等方式來調整社會網絡的呈現樣貌；





上圖中可以明顯發現，詞頻強度調高後只呈現出現次數較多的權威詞，因此畫面中的單詞數降低、關聯強度調高後節點連線數量明顯減少。



除此之外也可以輸入關鍵字計進行過濾。上圖中以篩選「美國」為例，並且調整關聯強度、詞頻強度、關係距離、縮放大小來呈現社網絡圖。

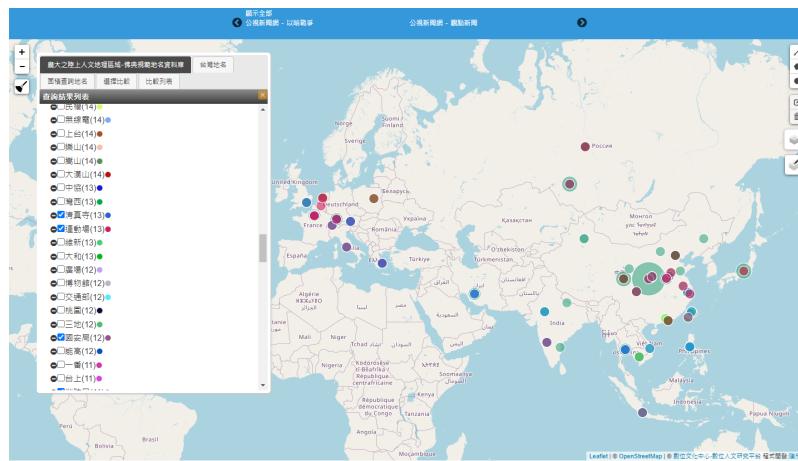
地理權威詞統計與視覺化

1. 加入文本與地理權威詞

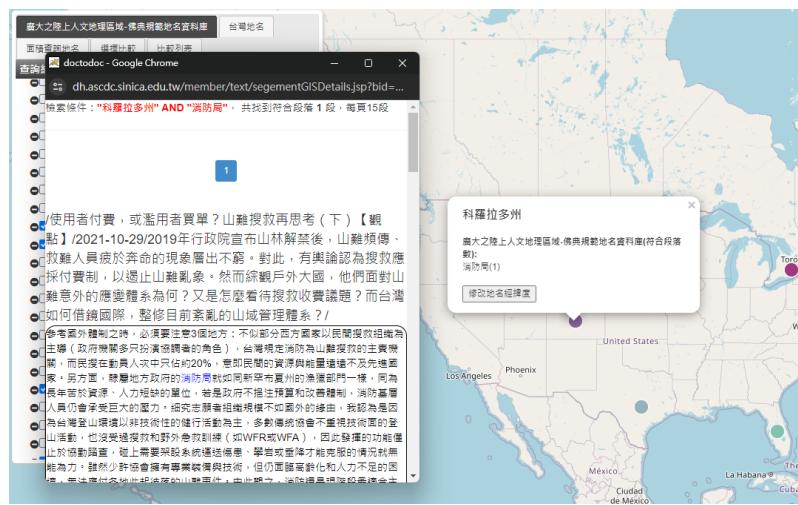
The screenshot shows the DH platform's text analysis interface. At the top, there are tabs for '中文版' (Chinese), 'En' (English), '上海例僅人' (Shanghai Example Person), '統計分析' (Statistical Analysis) (which is highlighted in green), '資源管理' (Resource Management), '說明手冊' (User Manual), and '問題建議' (Feedback). Below this, a search bar contains the ID '412479135'. A dropdown menu lists several categories: 公共新聞網 - 劇情劇情(44), 公共新聞網 - 賽點新聞(12112), 公共新聞網 - 以故動事(168168), and 公共新聞網 - 佛典新聞(1) (which is selected). A note below says '他人評論' (Comments from others). Below the search bar are three buttons: '權威詞頻率分析與清潔化' (Authority Term Frequency Analysis and Cleaning), '權威詞共現與計時演化' (Authority Term Co-occurrence and Temporal Evolution), and '地理權威詞頻率分析與清潔化' (Geographical Authority Term Frequency Analysis and Cleaning) (which is highlighted in blue). The main content area displays a large list of geographical names and their frequencies, such as 台灣地名(10), 廣大之陸上人文地理區域(10), 佛典規範地名資料庫(10), etc. At the bottom left, it says '文本處理方式: 合併處理 分別處理'. At the bottom right, there are buttons for '進行分析' (Analyze) and '資料全檢閱' (Full Review). A footer at the bottom of the page includes the text '© 2018-2024 中央研究院數位文化中心所有。TEL: 2652-1985#110 Email: yousheng@gate.sinica.edu.tw'.

這邊地理權威詞選擇「廣大之陸上人文地理區域-佛典規範地名資料庫」、其它權威詞選擇「台灣地名」。

2. 以地圖呈現統計結果

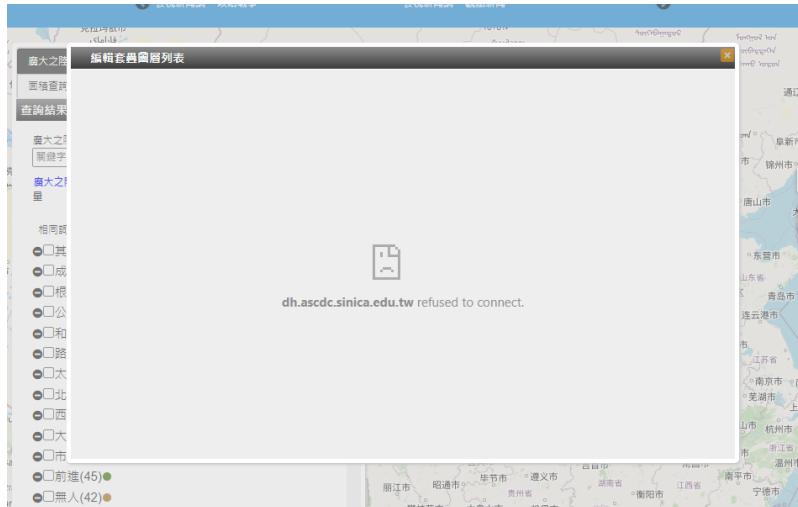


點擊左側的權威詞，地圖上會呈現與該單詞有關的地名。



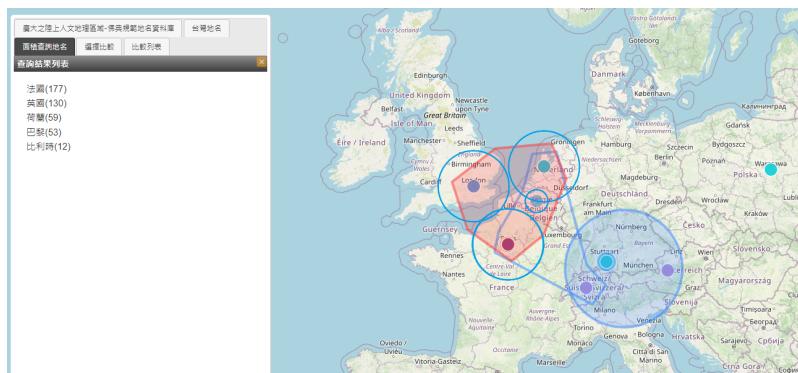
點擊地名節點與權威詞，即可查看所有符合條件原始文本。上圖為檢索條件：「科羅拉多州」AND「消防局」，共找到符合段落 1 段。

3. 套疊歷史地圖

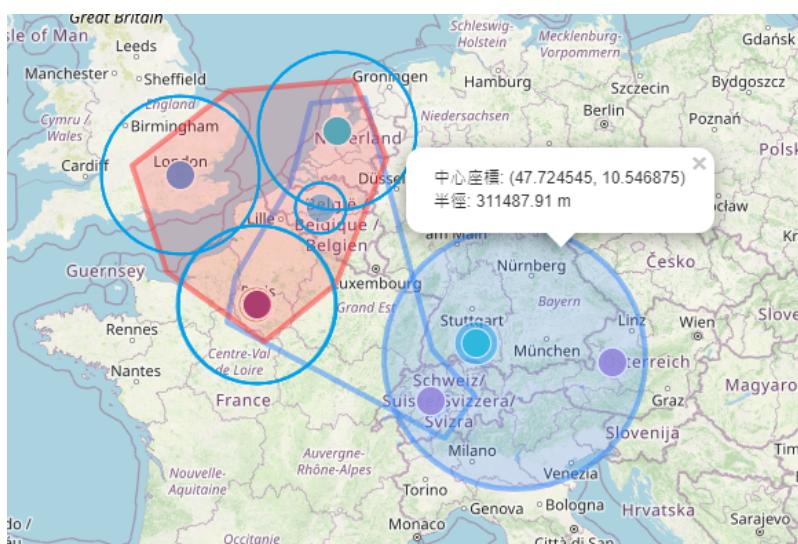


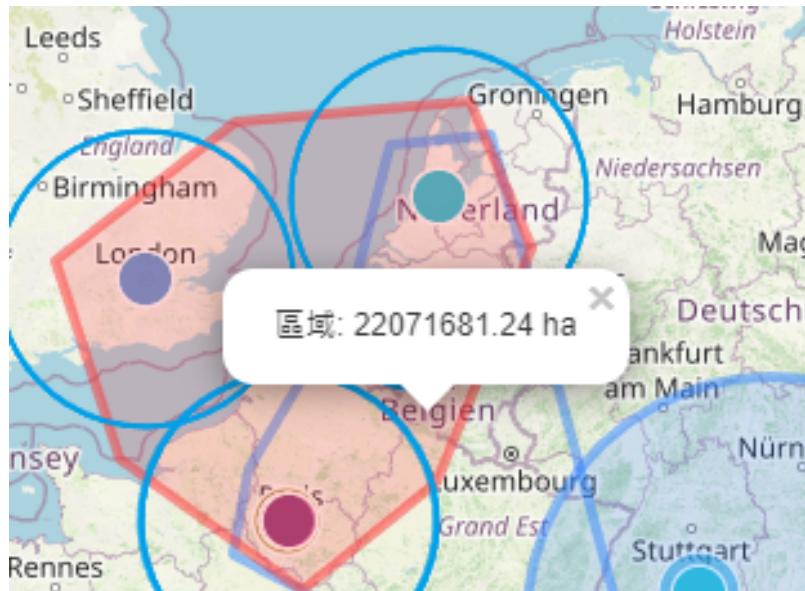
此功能可以在地圖上套疊上古地圖的行政區劃（例如，唐代行政區劃分）來進行統計結果的檢視，不過實際在操作時似乎遇到提供服務端無法連線的狀況，如上圖。

4. 自行繪製圖形



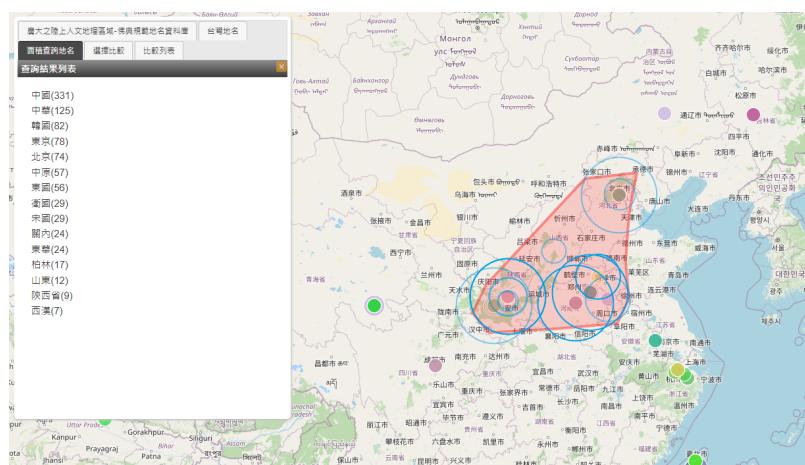
透過右上角的工具可進行圖形的繪製，分為 polyline 繪製、面積繪製與圓形繪製。繪製後也可以進行編輯與刪除。





點擊圖形區域可以查看區域面積、半徑長度等資訊。

5. 面積查詢地名功能



在地圖上繪製面積即可啟用。幾何圖形下需涵蓋我們要統計的地名。



接著點擊左側的列表或是地圖上的圓圈（即該地名在地圖上的範圍），可以看到所有跟此地名有關的權威詞。

點擊畫面中列出的權威詞，即可查看符合檢索條件的原始文本段落，例如上圖中呈現的是「北京」和「亞洲」同時出現的文本，共搜尋到 4 段。

兩文本差異分析

1. 進行文本與權威檔配對

ID	文本	频率	摘要
55571	公視新聞網 - 以強勢爭取	168	100% 重人：
55570	公視新聞網 - 行政院	4	100% 人：
55569	公視新聞網 - 媒體規範	112	100% 重人：
55568	公視新聞網 - 以強勢爭取	168	100% 重人：

先進入「統計分析」中的「管理文本與權威檔配對」，點擊「+」進行文本配對。

2. 兩文本差異分析

進入「統計分析」中的「兩文本差異分析」，選擇要分析的文本（最多 2 個）與權威檔、處理權威詞筆數與頻率差異百分比（這裡以門檻值 10 為例）。

頻率	百分比%	公視新聞網 - 以強勢爭取	公視新聞網 - 媒體規範
88	100	南部 85.2%	北部 60
56	63.6	北部 58.4%	南部 60
55	62.5	其他 45.1%	45
55	62.5	外文類 43.1%	40
52	59.1	和平	35
50	56.8	中心 12.2%	31
50	56.8	大眾 26.3%	25
46	52.3	國民 41.8%	17
44	50	無人 .44.2%	15
42	47.7	新亞 14.9%	15
42	47.7	逢甲 17.2%	15
38	43.2	台師	15
36	40.9	西方 26.1%	14
36	40.9	公司 44.5%	12
35	39.8	根本 25.1%	12
31	35.2	師大	12
29	33	成功 42.2%	12
28	31.8	東吳 12.3%	11
27	30.7	體中	10
25	28.4	卡拉	10
25	28.4	人和	9
21	23.9	教育部	9
20	22.7	大城	8
20	22.7	加拿大 10.7%	8
20	22.7	伊拉	7
18	20.5	路上 23.3%	7
18	20.4	上街	7
18	20.5	海員	6

表格中「頻率」表示出現的次數、「百分比」則從頻率進行換算，例如，第一名頻率為 88，百分比為 100%，第二名頻率為 56，而 56 為第一名的頻率 88 的 0.636 倍，因此第二名的百分比為 63.6%。而在比較兩文本之間差異時，當某一單詞百分比的差超過前一步驟設定的門檻值，頻率較高的一邊會顯示紅色，較低的一邊會顯示綠色，若沒有超過門檻值則不會變色。

3. N 字詞差異分析

在兩文本差異分析頁面，分析方式改為選擇「N 字詞」，依序設定 N 字詞區間（此處設定 2~4）、處理權威詞筆數（此處設定 200）與排名差異名次（此處設定 3）。

點選較高或較低N字詞時，可以標示左右兩詞所在位置。													
排序		公視新聞網 - 線點新聞 2字詞		公視新聞網 - 以油勘爭 2字詞		公視新聞網 - 線點新聞 3字詞		公視新聞網 - 以油勘爭 3字詞		公視新聞網 - 線點新聞 4字詞		公視新聞網 - 以油勘爭 4字詞	
1	台灣	627	以色列	1161	烏克蘭	198	以色列	1160	表本立場	97			
2	中國	459	色列	1160	俄羅斯	150	哈邁斯	533	本茲立場	97			
3	我們	16名	加薩	672	這樣的	125	巴勒斯	362	作者範點	97			
4	美國	18名	305	哈邁	534	不代表	105	斯坦	362	代表本站	97		
5	可能	158名	302	邁斯	533	表本站	97	納坦雅	120	不代表本	97		
6	沒有	111名	288	表示	1172名	403	者範點	97	坦雅胡	120	為作者範	96	
7	一個	1101名	286	斯坦	372	站立場	97	聯合國	118	本文為作	95		
8	可以	1186名	271	邁斯	366	本站立	97	加薩的	102	文為作者	95		
9	因為	143名	265	我們	16名	365	台灣的	97	斯迫人	100	彈道飛彈	34	
10	國家	131名	257	巴勒	362	作為者	97	種是廟	98	監議委分	33		
11	社會	253	人質	275	代表本	275	加薩定	98	狀空母艦	30			
12	公民	122名	250	美國	18名	249	為作者	96	色烈的	83	另一方面	28	
13	就是	169名	221	攻擊	174名	218	本文為	95	色列軍	81	公民投賣	28	
14	關係	218	醫院	214	文為作	95	對加薩	80	連接待臺	27			
15	自己	217	他們	165名	159	菲律賓	87	在加薩	79	安康接得	27		
16	若至	211	衝突	146	自己的	82	發誓人	72	中華民國	27			
17	這些	179名	205	沒有	111名	145	性種族	76	列國防	68	防空飛彈	26	
18	問題	1144名	203	戰爭	140	董事人	71	維爾納	67	烏克蘭的	25		
19	飛彈	199	民眾	1108名	134	的閱覽	67	理納坦	67	普丁政權	24		
20	烏克	198	人權	131	事實上	59	邁斯的	66	戶外活動	24			
21	更變	198	指出	128	中國的	57	加薩地	65	地方政府	23			
22	不島	184	已經	129名	128	是因為	53	色列軍	63	反遙飛彈	23		
23	這樣	182	聯合	127	的情況	190名	52	色列維	63	無人水面	22		
24	因此	182	停火	126	全黑美	50	在以色	63	小馬可仕	21			
25	文件	181	行動	125	被害人	50	黎巴嫩	61	人水氣庫	21			
26	行為	180	以軍	124	進一步	47	國防軍	58	安信音三	20			
27	運動	170	強調	120	重要的	44	列維理	58	一帶一路	20			
28	對於	167	納坦	120	解放軍	44	猶地區	56	菲律賓海	19			

上圖顯示結果中，由左至右依序為二字詞、三字詞與四字詞，並依照出現頻率由高到低排名。若超過設定的門檻值，較高的一邊會顯示紅色、較低的一邊會顯示綠色。圖中可以看到「美國」相差 8 名。

點選較高或較低N字詞時，可以標示左右兩詞所在位置。													
排序		公視新聞網 - 線點新聞 2字詞		公視新聞網 - 以油勘爭 2字詞		公視新聞網 - 線點新聞 3字詞		公視新聞網 - 以油勘爭 3字詞		公視新聞網 - 線點新聞 4字詞		公視新聞網 - 以油勘爭 4字詞	
台灣	627	以色列	1161	烏克蘭	198	以色列	1160	表本立場	97				
中國	459	色列	1160	俄羅斯	150	哈邁斯	533	本站立場	97				
我們	330	1.0 加薩	672	這樣的	125	巴勒斯	362	作者範點	97				
美國	369	1.2 哈邁	534	不代表	105	斯坦	362	代表本站	97				
可能	↑ 302	3.9 邁斯	533	表本站	97	納坦雅	120	不代表本	97				
沒有	288	2.0 表示	↑ 403	5.0 當範點	97	坦雅胡	120	為作者範	97				
一個	286	4.9 斯坦	372	站立場	97	聯合國	118	本文為作	97				
可以	↑ 271	7.1 邁斯	366	本站立	97	加薩的	102	文為作者	97				
因為	↑ 265	3.0 我們	365	1.1 舊開動	97	斯坦人	100	彈道飛彈	97				
國家	257	2.6 巴勒	362	作為者	97	猶地觀	98	監議委分	98				
社會	253	人質	275	代表本	97	加薩走	98	航空母艦	98				
政府	↑ 250	2.3 美國	249	1.0 為作者	96	色烈的	83	另一方面	96				
就是	↑ 221	3.3 攻擊	218	1.9 本文為	95	色列軍	81	公民投賣	95				
關係	218	醫院	214	文為作	95	對加薩	80	連接待臺	95				
自己	217	他們	159	1.4 菲律賓	87	在加薩	79	安康接得	79				
這些	211	衝突	146	自己的	82	發誓人	72	中華民國	72				
這些↑	205	3.4 沒有	145	1.0 性種族	76	列國防	68	防空飛彈	68				
問題↑	203	4.7 戰爭	140	當範事人	71	維爾納	67	烏克蘭的	67				
飛彈	199	民衆	134	1.4 的閱覽	67	理納坦	67	普丁政權	67				
烏克	198	人權	131	事實上	59	邁斯的	66	戶外活動	66				
基美	198	指出	128	中國的	57	加薩地	65	地方政府	65				
不島	184	已經	128	1.0 是因為	53	色列軍	63	反遙飛彈	63				
這樣	182	聯合	127	的情況	↑ 52	2.6 色列維	63	無人水面	63				
因此	182	停火	126	全黑美	50	在以色	63	小馬可仕	63				
女性	181	行動	125	被害人	50	黎巴嫩	61	人水氣庫	61				
行為	180	以軍	124	進一步	47	國防軍	58	安信音三	58				
運動	170	強調	120	重要的	44	列維理	58	一帶一路	58				
對於	167	納坦	120	解放軍	44	猶地區	56	菲律賓海	56				

接著可以嘗試將差異比較方式改為「比例差異」（此處以 1 為例）。此比較方式以頻率較低的一邊作為分母來進行計算，若超過設定的門檻值時，較高的一邊顯示紅色、較低的一邊顯示綠色。

文本的斷詞分析

進入統計分析中的「管理文本與斷詞功能」。針對文本選擇欲使用的斷詞方式，這邊以 CKIP 進行實作。

全編	ID	IF	文本名稱	選擇斷詞方式	移除斷詞
<input type="checkbox"/>	55571		公視新聞稿 - 以始數事	<input checked="" type="radio"/> CKIP[上古] <input type="radio"/> CKIP[加入拆程] <input type="radio"/> jieba[加入拆程] <input checked="" type="radio"/> CKIP <input type="radio"/> jieba、	
<input type="checkbox"/>	55570		公視新聞稿 - 副刊報導	<input checked="" type="radio"/> CKIP[上古] <input type="radio"/> CKIP[加入拆程] <input type="radio"/> jieba[加入拆程] <input checked="" type="radio"/> CKIP <input type="radio"/> jieba、	
<input type="checkbox"/>	55569		公視新聞稿 - 觀點新聞	<input checked="" type="radio"/> CKIP[上古] <input type="radio"/> CKIP[加入拆程] <input type="radio"/> jieba[加入拆程] <input checked="" type="radio"/> CKIP <input type="radio"/> jieba、	
<input type="checkbox"/>	55568		公視新聞稿 - 以始數事	<input checked="" type="radio"/> CKIP[上古] <input type="radio"/> CKIP[加入拆程] <input type="radio"/> jieba[加入拆程] <input checked="" type="radio"/> CKIP <input type="radio"/> jieba、	

批次處理需要處理時間，一次限制10筆，請耐心等候!

接著到首頁點開已經完成斷詞的文本節點，進入斷詞分析，選擇斷詞系統與統計方式，統計界面中可以設定前後的綴詞數（斷詞數）、統計 N 個詞、查詢詞，可向前或向後查詢若干個綴詞，並且可用表格的方式顯示前綴詞在文本中的脈絡，按下「檢視段落」即可查閱原始文本段落，同時可選擇是否顯示斷詞標示。若 N 設定為 2、綴詞數設定為 4，即可查看前 4 個綴詞中 2 個詞連續出現的狀況。同時平台也有提供篩選詞性的功能。

推導查詢

從首頁點選欲分析的文本節點，進入推導查詢。接著設定斷詞系統、統計方式以及是否統計標點符號。下方箭頭代表推導方向，欄位中填入要在幾個詞距離內查詢、查詢幾個連續詞與查詢詞，查詢後即可獲得結果以及文本概況，原始文本查訊時平台會以顏色標記查詢文句，同時可以決定是否顯示所有的詞性。

全編	ID	IF	文本名稱	選擇斷詞方式	移除斷詞
<input type="checkbox"/>	55571		公視新聞稿 - 以始數事	<input checked="" type="radio"/> CKIP <input type="radio"/> CKIP[上古] <input type="radio"/> jieba、	
<input type="checkbox"/>	55570		公視新聞稿 - 副刊報導	<input checked="" type="radio"/> CKIP[上古] <input type="radio"/> jieba、 <input checked="" type="radio"/> CKIP[近代[處理中...]]、	
<input type="checkbox"/>	55569		公視新聞稿 - 觀點新聞	<input checked="" type="radio"/> CKIP <input type="radio"/> CKIP[上古] <input type="radio"/> jieba、	
<input type="checkbox"/>	55568		公視新聞稿 - 以始數事	<input checked="" type="radio"/> CKIP <input type="radio"/> CKIP[上古] <input type="radio"/> jieba、	

批次處理需要處理時間，一次限制10筆，請耐心等候!

如上圖，由於系統持續無法完成斷詞超過 4 小時，我無法完成操作這個部份！！！

心得

數位人文研究平台整體操作起來，表面上確實不容易上手，在觀看說明影片後逐漸了解平台設計的邏輯後操作起來順暢許多，唯有在 CKIP 斷詞處理等待了非常非常非常長的時間（其實根本跑不出來）。在所有功能中我最喜歡權威詞視覺化，這類型的圖表呈現不只在古文、新聞文本分析可以進行 NLP 的詞特徵關聯性比對、甚至近年在一些學術研討會中將議程也進行這樣的分析視覺化，更容易的看要學術與產業關注的議題趨勢。除此之外我認為地理權威詞的功能真的為 NLP 研究與開發者省下許多時間，過往光是資料處理、分箱，再撰寫程式串接地圖視覺化的開源 API 就需要耗費相當大的時間心力，目前平台上將功能完整化，屬實相當實用。本次的操作體驗下來非常新鮮，包括將自己爬取的新聞文本上傳處理後進行間距查詢、推導查詢等，都是過去不曾嘗試的文本分析方式，雖然自己還是比較習慣直接打 code 的方式，不過未來有機會還是想嘗試看看結合平台現有的功能看是否能提升 NLP 模型開發時特徵工程的成效。最後想建議，斷詞除了效能有待加強外，起許日後加入 MONPA 斷詞！