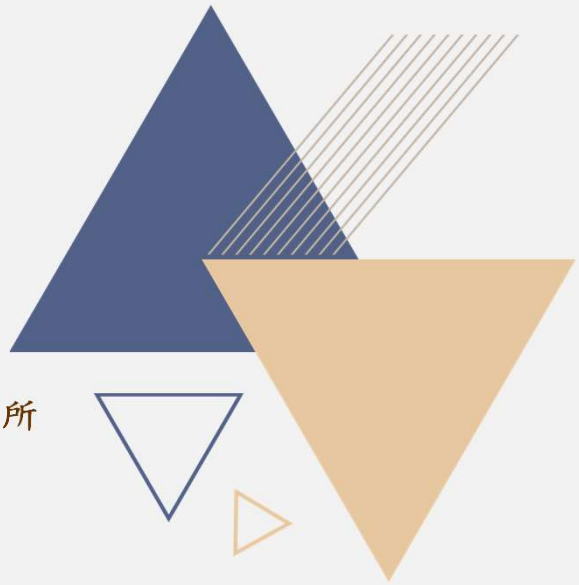
2022年台北國際書展研討會

檔案實踐數位人文研究之探索



國立政治大學圖書資訊與檔案學研究所
教授兼任所長、圖書館副館長
林巧敏

1

目 錄

CONTENT

2

PART.1

數位人文研究概述

- 為什麼需要數位人文
- 數位人文研究目的
- 數位人文研究工具

3

為什麼需要數位人文

數位人文是指以數位工具協助人文研究找到資訊間的脈絡、發掘以往純文本研讀方法難以盡覽內容的研究問題。甚至可進一步發掘原有研究方法難以探知的內容關係。

文本近讀

- 人們(人文學者)較易陷入單一文本中，不易察覺文本/資料之間的關聯。
- 人力對文本屬性的處理終究有極限，海量資料容易失去/隱藏脈絡。

(數位)文本遠讀

- 在合理時間內多工處理海量資料、得出文本脈絡並探索嶄新研究的議題。
- 雖然電腦能發掘文本脈絡/資訊概念，但組織知識的關鍵還是人腦。

4

數位人文研究目的

多工處理海量資料
達到迅速、大量處理資料，並提供交流、傳播目的

將資料結構化
將非結構化資料轉為結構化，符合數位時代研究、典藏、應用需求

自動化作業流程
機器學習自動分類、分群、標註、篩選，可處理高重複性作業

多種應用
結合鏈結資料、知識圖譜、數位策展等運用，發揮資訊組織的價值

研究結果呈現
提供整合空間與時間、直觀且動態的研究結果呈現方式

5

數位人文研究工具

基本工具	• 文書處理、簡報、Excel 、通訊軟體等辦公室生產力工具
資料蒐集工具	• 各種搜尋引擎、檢索資料庫以協助在網路找尋資料
資料分析工具	• 對於找到的資料進行文本探勘，例如：自動斷詞、統計詞頻、內容比對、圖像辨識等
資料呈現與傳播工具	• 視覺化呈現（文字雲、網絡關係）、地理分布、數位出版

整理自林富士（2017）。未來歷史學。人文與社會科學簡訊，18（3），56-62。

6

史學研究的新方法

鳥瞰閱讀	• 細讀 close reading（逐字逐句細讀，詮釋字裡行間意義） → 鳥瞰 distant reading（宏觀人、事、物在時、空的分布，勾勒發展趨勢與重大變遷，尋譯文本結構與關係
社會網絡分析	• 探討特定人物（人群）行為 → 利用社會網絡分析人與周遭環境互動的關係
地理資訊系統	• 將人文思維與空間向度結合，反映空間和地理對於歷史發展的影響
社群協作	• 獨立研究 → 與技術人員合作，協助史學家蒐集整理、分析資料，甚至發展社群共筆史學

整理自林富士（2017）。未來歷史學。人文與社會科學簡訊，18（3），56-62。

7

檔案資料探勘之應用

- 檔案資料探勘目的
- 檔案資料探勘常用方法
- 檔案資料探勘流程
- 相關研究

8

資料探勘

- 「資料探勘」(data mining)是透過各種資料分析技術，擷取資料庫（結構化資料）涵蓋的資訊或知識。
- 但隨著文本內容數位化以及網路資訊越來越多，非結構化的資料比結構化資料發展迅速，有以「文本探勘」(text mining)一詞用於文本內容資訊分析。
- **文本探勘是從文本產生有價值的訊息，分析的資料可以是無規則性、非結構化的原始資料。由文字間發掘核心概念、找出概念之間的關係。**

9

進行檔案內容探勘目的

因應數位檔案館時代

- 增強檔案資料庫/檢索系統/知識庫效能
- 數位典藏、數位策展、檔案推廣、教育
- 原生數位文件/檔案管理程序需求

找出檔案內容脈絡

- 協助判斷檔案層級脈絡
- 發掘離散檔案文件之案件/系列間的關聯
- 輔助編排描述

發揮檔案潛在價值

- 發現與組織檔案內容知識
- 提供多種檔案應用和研究機會

提高檔案管理效率

- 精簡檔案管理行政成本
- 建立自動化/智慧化作業流程
- 降低人為錯誤的風險

10

數位人文工具運用的方向

詞頻分析功能
- 文字雲
- 詞頻分佈統計

文本分析功能
- 情感分析
- 語意分析
- 主題分析

資料探勘功能
- 人物社會網絡探勘
- 事件關聯探勘
- 時間序列探勘

資訊視覺化功能

地理資訊功能

11

檔案內容探勘流程(確定研究主題後)

資料集彙整	自動斷詞/簡易標註	機器學習	探勘結果分析	研究結果呈現
● 文本全文數位化 ✓全文資料庫打包 ✓OCR辨識 ● 資料集清理 ✓ 清除不必要字符 ✓ 探勘前預先諮詢專家意見(預設框架) ※以Excel處理	● 自動斷詞 ✓CKIP Tagger ✓CKIP Transformers ✓Jieba ● 半自動標註 ✓ CKIP Transformers ✓ Docusky-MARKUS ✓ Python	● 準備訓練資料集 ✓確定使用的資料探勘/數位人文分析技術 ✓進行前測 ● 機器學習 ✓ 反覆測試調整必要參數 ✓ WEKA等工具	● 比較探勘結果 ✓邀請專家評估 ✓與既有研究比較 ● 進階結果分析 ✓ 統計、社會網絡等 ✓ Gaphi、VosViewer、各類GIS工具等	● 產製視覺化圖像 ✓知識圖譜/社會網絡 ✓統計圖表 ● 融入資料庫功能 ✓ 索引典、鏈結資料 ✓ 多功能研究輔助 ● 數位策展呈現

12

檔案數位人文研究實例

- 檔案內容斷詞及自動分類
- 檔案事件新聞及網路輿論之情感分析

13

資料探勘運用於檔案管理與內容分析

優化檔案管理作業

- 總裁批簽檔案自動分類
- 文本探勘支援檔案價值鑑定判斷

促進檔案內容分析

- 檔案輿情內容分析及其情感傾向
- 檔案研究文獻內容分析

14

Text Mining for
Subject Classification

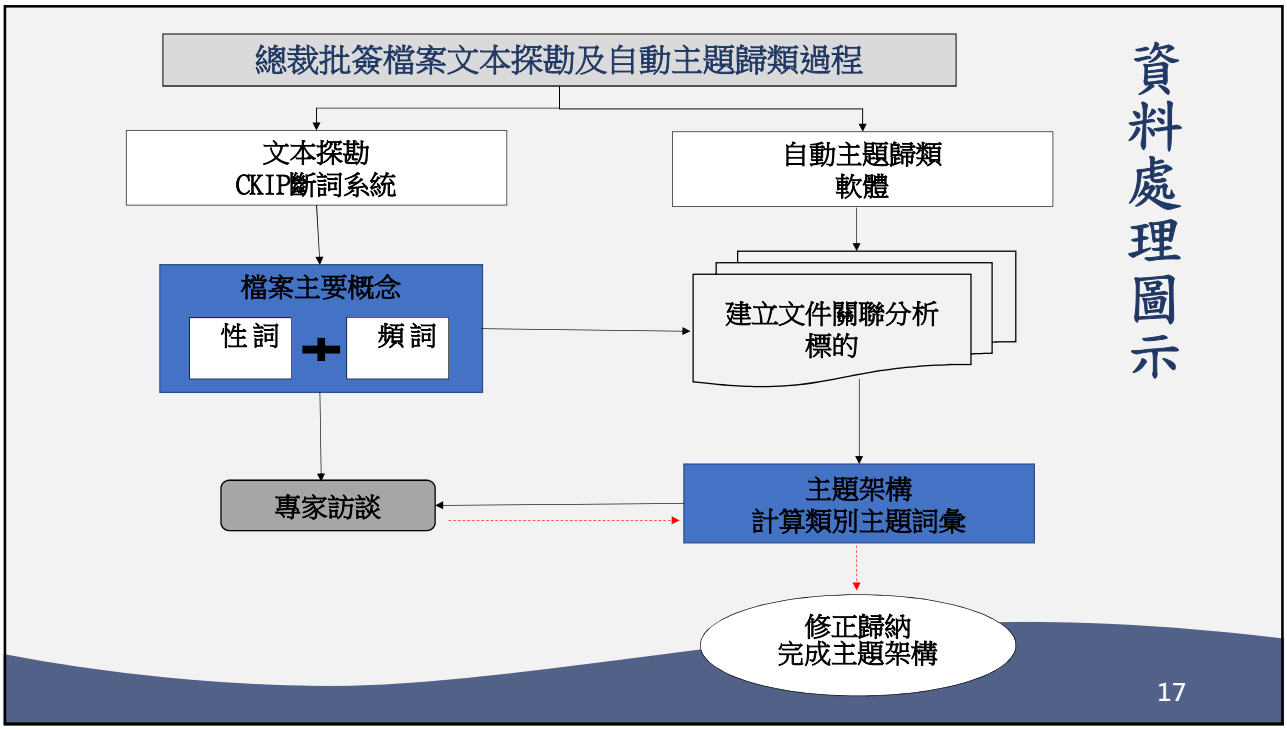
檔案內容斷詞及自動分類

檔案內容自動分類的構想

利用資料探勘分析工具，將檔案標題、內容摘要以及內容涉及人、事、物等含有主題概念欄位中詞彙，結合「相關係數」及「詞頻」排序詞彙的分析，**自動擷取類別特徵詞，以作為類別標題進行檔案文件的分群。**

根據分群之特徵詞，**初擬訂定適當之主題標目，以整理出此批檔案文件系列之主題架構。**

16



自動分類資料探勘對象

- 《總裁批簽》檔案數量共93冊另17卷，檔案原件未予分類，是依時間排序以件為單元建有目錄，案件目錄共4,361筆，已有檔名、檔案內容描述以及檔案內容涉及之人、事、時、地名等metadata欄位

民國64年03月19日	62年12月選出之增額立委，依照「動員戡亂時期選舉條款」第6項每3年改選之規定，應於65年1月期滿前3個月舉行改選，為此，本會與國家安全會議，根據各方對增額中央民代選舉辦法
民國63年12月16日	63年度中央各單位工作績效考核業務經辦理完竣，均係秉承鈞座訓示、中全會通過各案及中常會決定事項認真辦理。附呈各單位年度工作成績名次表、工作考核評語各一份。
民國63年12月07日	第10屆中央評議委員第6次會議于本年11月26日舉行，與會中評委所題各項提案及評議建議意見，經決議於呈請總裁後由中常會分別處理。附呈各中評委提案及評議建議意見之處理意見表
民國63年10月23日	本黨十屆五中全會，以當前青年教育工作至為重要，擬增邀公私立大學暨獨立學院系主任以上本黨同志列席。附呈大學暨學院列席人數統計表一份。
民國63年10月16日	本年11月24日為本黨建黨80週年，海外僑報均將出版特刊，擬恭錄訓詞一則，俾海外僑報刊用。附呈題詞一則。
民國63年10月07日	十屆五中全會已奉核于本年11月24日召開，惟中央委員皮以書病故，擬依黨章由候補委員焦金堂遞補。
民國63年09月12日	依黨章規定，十屆五中全會應於本年11月舉行，有關籌備工作，經研擬初步構想一種，就五中全會舉行之時機、地點及討論的議題進行討論。附呈五中全會籌備工作初步構想一份。
民國63年08月10日	63年雙十國慶，海外各地區僑報均將出版特刊，擬恭錄訓詞兩則，俾製版寄發海外僑報刊用。附呈題詞二則。
民國63年07月30日	立法院院長倪文亞函稱，應韓國國會議長丁一權之邀，擬由倪文亞定期率團前往韓國訪問。經邀約倪文亞、沈昌煥、秦孝儀等研商決定，訪問團團員以立委5人、國代4人、監委2人為宜，
民國63年07月13日	本黨各級黨部62年度保舉最優人員業經審核完畢，計有季峻宮等30人。附呈保舉最優人員簡歷冊、呈報表各一份。
民國63年06月14日	中央委員會海外工作會副主任懸缺擬以江炳倫繼任，另前正中書局總經理李潔服務黨營事業多年，擬改任為文化工作會專任委員。檢附兩人簡歷表。
民國63年05月30日	有感於東南亞政情動盪，為適應未來各地政情演變趨勢，張寶樹擬前往菲律賓視察當地總支部，部署全盤對匪工作，並分訪僑領爭取向心。
民國63年05月27日	黨營文化事業單位64年度股東大會即將召開，茲追具各單位董事及監察人名冊各一份。
民國63年05月25日	正中書局董事長胡健中年逾70，依例應請自退，總經理李潔自感無由展布，請辭現職，擬改由朱建民、黎元譽分任董事長、總經理。附呈兩人簡歷各一份。
民國63年04月16日	張寶樹擬應韓國大統領早餐祈禱會(5月1日舉行)之邀，前往韓國漢城，與會者多為亞洲友好國家民間領袖暨政黨負責人。會後將分訪當地友我人士，回國時將途經日本，並擬訪問日本支持
民國63年04月02日	本黨針對當前革命形勢，為發揮新聞宣傳工作，將於4月7日在台北舉行第4次新聞工作會談，中心主題為如何弘揚三民主義新聞政策、如何動員新聞傳播工具、如何打擊中共政權、如何健
民國63年03月14日	本會婦女工作會秘書方英達、專門委員周靜仙工作努力頗具績效，經該會主任綏劍秋先行報奉指導長同意，擬請調升為該會專任委員。另外本會秘書處秘書陳敬之擔任黨務工作20餘年資
民國63年02月28日	63年青年節海外各地僑報均將出版特刊，擬恭錄訓詞一則，俾製版寄發海外僑報刊用。附呈題詞稿一則。
民國62年11月30日	第10屆中評委舉行第5次會議，會中中評委相繼提出建議案及評議建議意見，經決議，所提各案均呈請總裁核定後，交中常會分別研究處理。附呈各中評委提案及評議建議意見之處理意見
民國62年11月20日	中央各單位62年度工作考核業已完竣，各單位均能秉承鈞座訓示暨中全會通過各案，認真辦理。附呈各單位工作成績評分名次表、及工作考核總評語。
民國62年08月09日	62年國慶，海外各地僑報將出版特刊，擬恭錄所需題詞二則，俾寄發海外僑報。附呈題詞稿二則。
民國62年08月03日	本會文化工作會尚有專任委員懸缺，擬以該會兼任委員殷文俊調任。附呈殷文俊簡歷表一份。

18

總裁批簽檔案特性

採用《總裁批簽》檔案後設資料進行詞彙分析

包含題名、日期、內容描述、典藏號、典藏位置與製作單位等六欄位，主要會採用題名與內容描述兩項具有內容主題性的欄位

採用主題詞彙比檔案原文語意更明確

所謂《總裁批簽》指的是在1950年代中國國民黨透過改造確立由總裁—蔣介石為權力核心的組織結構中，由中央改造委員會與中央委員會評估屬「重要黨務」，並將之上呈予蔣介石批示文書的總稱，時間範圍包含1950年8月中央改造委員會成立至1975年4月蔣介石病逝，一共4361筆。檔案原件未予主題分類，採時間序以件入卷，已建置目錄。

在資料的性質上，由於「重要黨務」的定義，往往取決於主事者個人主觀上的認定，未有一套明確的標準，使得《總裁批簽》的內容可說是紛亂雜陳，缺乏主題分類的情況，不僅對管理者而言難以整理，對使用者來說更是查檢不易。

19

資料探勘工具_CATAR

CATAR（Content Analysis Toolkit for Academic Research）

- 由國立臺灣師範大學曾元顯所研發之CATAR軟體進行文獻內容分析，可以迅速有效地分析文獻的各種計量結果，瞭解內容主題的發展趨勢。因可結合「相關係數」及「詞頻」排序詞彙分析，能自動擷取類別特徵詞，能作為類別標題，進行各檔案文件分群。
- 開發者曾元顯教授「希望不只作者與技術人員，一般人也會使用」之理念，讓非技術人員的研究者也能使用文本探勘工具，讓人文學可跨領域結合資訊技術的應用（曾元顯，2011）。

- 網路上提供有安裝說明
- 參考網站Content Analysis Toolkit for Academic Research（CATAR）
- 網址：http://web.ntnu.edu.tw/~samtseng/CATAR/Readme.html

- 將檔案全宗各案卷內容匯入CATAR之DB Browser（SQLite）資料庫軟體，建立自動主題歸類之分析標的。

20

CATAR自動分類判斷過程

使用者操作：閾值（0～0.1）

↓

自動分類策略：
1. 符合分類原則
2. 利用數位工具，迅速且（有效）

↓

具體實施：
1. 類別數量不超過30類
2. 透過主題地圖擇選主題分布具明確區隔者
3. 結合主題樹衡量主題是否平均

通過兩文件關鍵詞交集個數除以兩文件分別關鍵詞數總和所得到的商，得到全數文件兩兩關聯度之相似度矩陣，並透過「多維縮放」（Multidimensional Scaling，MDS）技術、「層次凝聚歸類法」（hierarchical agglomerative clustering）與多階段主題歸類（multi-stage clustering）的概念進行文件歸類。

21

閾值選擇

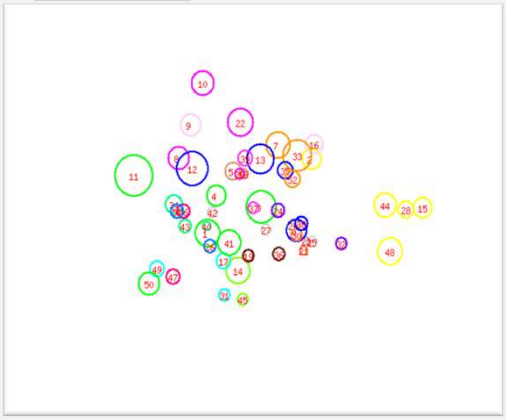
- 閾值預設為0.1
- 閾值的數值越大，進行歸類相對嚴謹，類別與主題樹數量自然便越多
- 自0.06～0.1開始嘗試
- 找尋合適數量者 = 階層四
- 最初僅採用「內容描述」欄位（主題意涵較高），後加入題名欄位

題名及內容描述						
閾值\階層	1	2	3	4	5	6
0.1	489(2728)	137(450)	46(128)	14(40)	2(10)	0(0)
0.09	486(2728)	139(450)	50(132)	14(42)	3(11)	1(3)
0.08	482(2728)	131(445)	43(122)	12(37)	2(10)	0(0)
0.07	478(2728)	126(441)	44(121)	12(40)	3(10)	1(3)
0.06	475(2728)	123(438)	40(117)	12(37)	2(11)	0(0)
內容描述						
閾值\階層	1	2	3	4	5	6
0.1	490(2728)	139(451)	49(132)	11(39)	4(13)	1(3)
0.09	486(2728)	139(450)	52(136)	16(47)	5(15)	1(5)
0.08	482(2728)	131(445)	40(119)	14(37)	4(13)	1(4)
0.07	478(2728)	126(441)	42(122)	12(37)	3(9)	1(3)
0.06	475(2728)	123(438)	38(117)	10(33)	1(6)	無

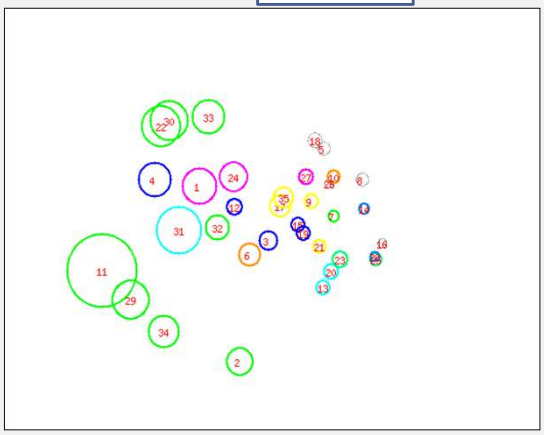
22

主題地圖比較

0.09階層四



0.03階層四



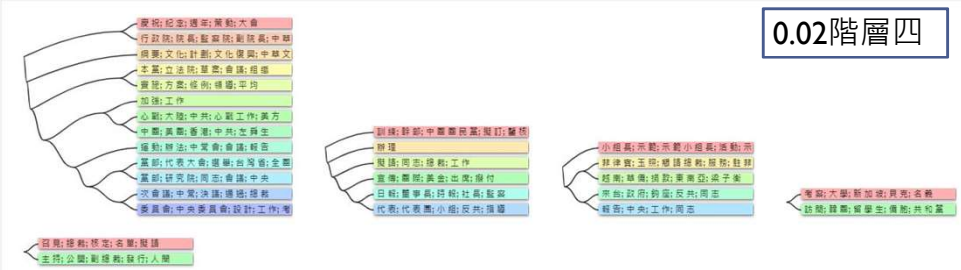
- 重疊程度過高 = 分類成果並不理想（0.09）
- 改變策略：將閾值降低，比較後得出閾值0.02及0.03成果較佳

23

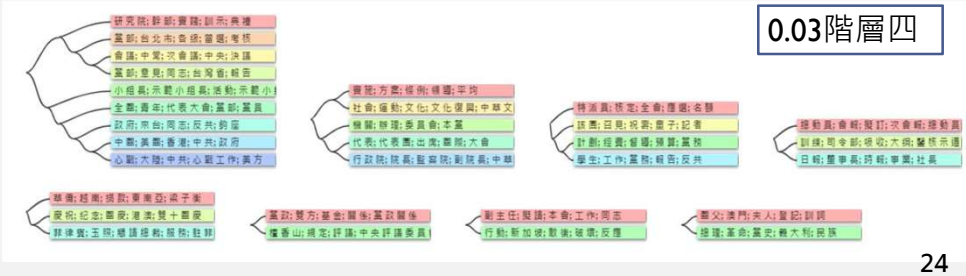
主題樹比較

- 結合主題樹判斷，儘量採取主題分類後各類目涵蓋之主題概念數量分布相對較為均衡的分類
- 閾值0.03，類目數量較平均

0.02階層四



0.03階層四



24

《總裁批簽》自動分類主題表

主題詞代表性

類別內相似度下限值

	類別名稱	擷取主題詞
1	中央黨務	0.0994（中 央:309.0935, 委 員 會:265.9273, 工作:213.1679, 香 港:183.0601, 宣 傳:165.5354）
2	黨組織	0.0870（黨 部:98.5280, 黨 員:56.8174, 青年:51.1386, 中 央:41.9194, 鈞 座:40.5417）
3	人事	0.0694（人 員:46.6553, 小 組:19.3561, 工 作:17.8917, 行 動:16.6111, 單 位:15.9009）
4	選舉	0.2486（選 舉:54.7590, 台 灣 省:48.8408, 地 方:30.2857, 議 員:27.1310, 自 治:21.6620）
5	思想教育	0.1453（復 興:11.1235, 基 金:9.7331, 文 化:9.5899, 運 動:7.2940, 綱 要:5.8121）
6	黨營事業	0.2352（董 事:42.4085, 監 察 人:37.5420, 董 事 長:29.0116, 事 業:25.7505, 電 影:25.6735）
7	中央政務	0.3547（國 民 大 會:15.8276, 總 統:15.5849, 國 大:13.6104, 副 總 統:12.8367, 總 理:8.3793）
8	政黨互動	0.1276（社 黨:40.3229, 徐 傅 霖:29.8945, 蔣 匀 田:29.1993, 民 社 黨:25.6034, 行政院:12.5758）

25

自動分類與人工判斷比較

- 自動類聚結果產出八個類別，每類別附有類別擷取詞彙，每類隨機抽樣2件（共16件），每件有5個詞彙，請10位歷史背景學者專家檢視文本內容後，將自動分類詞彙認為不符合者註記（刪除），註記刪去詞彙越多，表示工具歸類越不準確。計算10位**專家判斷的CATAR歸類正確率為49.2%**。
- **受訪專家認為：數位工具斷詞產生詞彙通常較為籠統、專指性不高，且數位工具無法創造文本沒出現的詞彙，但人工可以根據既有的知識架構擬定主題詞彙。**
- 人工主題分類仰賴專家知識以及相關經驗，人工分類是以學科專家的知識架構為基礎，但人工分類對於各單件主題詞彙的挑選卻沒有很高的一致性，反映出**人工在進行主題分類時的主觀性很高**。

26

自動分類結果討論

學科專家對於數位工具自動分類產生之大類主題可接受，但對於各件歸類的準確程度抱持懷疑（但人工分類同樣有主觀性問題）。

本研究經自動分類及人工修正之分類架構，可提供檔案檢索系統使用者介面以主題瀏覽方式，提供使用者利用此一虛擬主題樹，直接點選瀏覽相關之檔案文件。

將分類產生之主類及次類名稱，批次鍵入後設資料欄位，增加檔案描述欄位之主題檢索詞彙，提高檢索用語可查獲之結果。

Text Mining

News & Forum

檔案新聞及網路輿論之情感分析

探討背景與構想

社會大眾的情感認知是專業領域尋求社會支持的基礎，在傳播媒體揭露的檔案資訊，勢必影響閱聽人對於檔案及檔案機構的認知，民眾對於檔案事業的情感和態度，也會影響到檔案館的社會形象以及資源投入程度。因此，有必要探討媒體所關注之的檔案事件主題及其對於社會大眾情感的影響，作為政策擬定的判斷依據。

那些檔案事件主題經常被媒體及輿論關注？這些事件所議論之檔案管理問題為何？

這些報導和輿論文字呈現的情感傾向為何？那些檔案事件主題報導為正面，那些主題偏向負面？

29

運用的數位工具

中文斷詞系統

- 現今常用於繁體中文斷詞的系統，有CKIP中文斷詞系統及Jieba斷詞
- 研究**採用中央研究院所建置的「中文斷詞系統」（CKIP）**（http://ckipsvr.iis.sinica.edu.tw/）
- 利用電腦前置處理自然語言，將句子中的詞彙以最小「意義」的單位區辨出，主要技術為透過中文分詞語料庫（詞典），比對並標註句中詞性。藉由詞性標註分類，可提供後續進行語言處理的系統，分辨文本詞彙語意

情感辭典

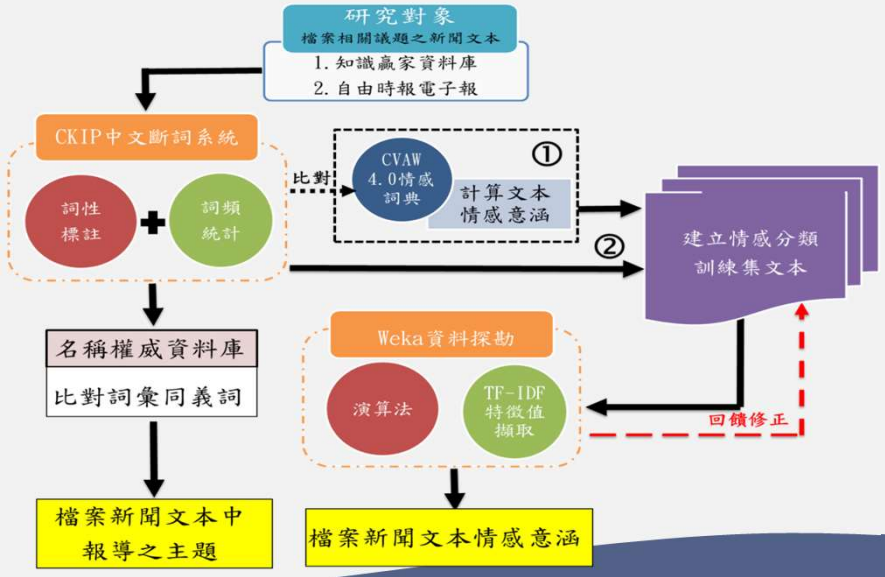
- 對文本進行情感分析的計算，主要分為機器學習以及建立情緒字詞典兩種方法
- **採用CVAW - 中文維度型情感詞典**，CVAW 4.0 於2019年7月16日發布之5,512個中文單詞維度情感分析之數據，每個中文單詞分別標註其情感評價，可表示正面和負面情緒的程度與情感維度，亦可表示平靜和興奮的程度

資料探勘工具Weka

- 資料探勘領域中重要的自由軟體，透過內建相關機器學習之演算法與相關套件框架進行資料處理統計與學習，可進行回歸、分類、聚類、關聯規則之資料探勘分析以及屬性選擇，可將資料分析結果以視覺化方式呈現

30

新聞報導資料分析過程



31

32

文本蒐集

來源名稱	網址	檢索詞彙	檢索筆數	相關筆數
知識贏家（中時報系）	http://kmw.chinatimes.com.autorpa.lib.fju.edu.tw:2048/Index.aspx	1. 檢索欄位：所有欄位（含標題與作者） 2. 時間：1994.1至2019.8.31 3. 檢索詞：「檔案」and「檔案法」；「檔案管理」and「檔案法」；「檔案」or「檔案管理」；「檔案法」。	245	232
自由時報電子報	https://www.ltn.com.tw/	1. 檢索欄位：關鍵字 2. 時間：2005.1至2019.8.31 3. 檢索詞：「檔案」and「檔案法」；「檔案管理」and「檔案法」；「檔案」or「檔案管理」；「檔案法」。	214	202
總　　計			459	434

文本處理

STEP 1

- 434篇檔案相關新聞文本內容，經CKIP斷詞系統斷詞後，根據中研院平衡語料庫詞類標記集之簡化詞類表，統計出58種詞性、16042個詞彙、總詞頻為196,298次。
- 過濾並合併單字詞，針對相關詞彙（如人名、機關名稱），比對中文名稱權威資料庫，進行詞彙合併整理，找出檔案相關新聞文本中主要之主題詞。
- 專有名稱（Nb）的詞彙，多為歷史事件或人物名稱之詞彙，例如：「二二八事件」、「白色恐怖」、「美麗島事件」等歷史事件詞彙，又如「陳水扁」、「施明德」、「陳豐義」、「馬英九」等政治人物姓名。

STEP 2

- 運用CVAW 情感詞典進行情感分析，因CVAW 情感表的數值以1至10為範圍，將數值轉換成正、負面，較容易辨別其情感傾向，參考Frederick（2019）的作法，將原本情感的分數都減掉5，然後再把算完的數值依據詞彙數加以平均得到每篇新聞的情感數值。

STEP 3

- 以機器學習的方式，將上述情感分類訓練集，藉由Weka工具，進行演算法、TF-IDF的計算、特徵值的評估。透過相關評估後，挑選出具影響情感分類之不同詞性或高頻詞詞彙所包含之文本，重新調整情感分類訓練集之文本數，找出最佳情感分類的訓練集文本。最後將434篇原始文本作為測試集，利用上述最佳情感分類的訓練集文本進行情感預測。

33

新聞報導高詞頻之檔案事件詞彙

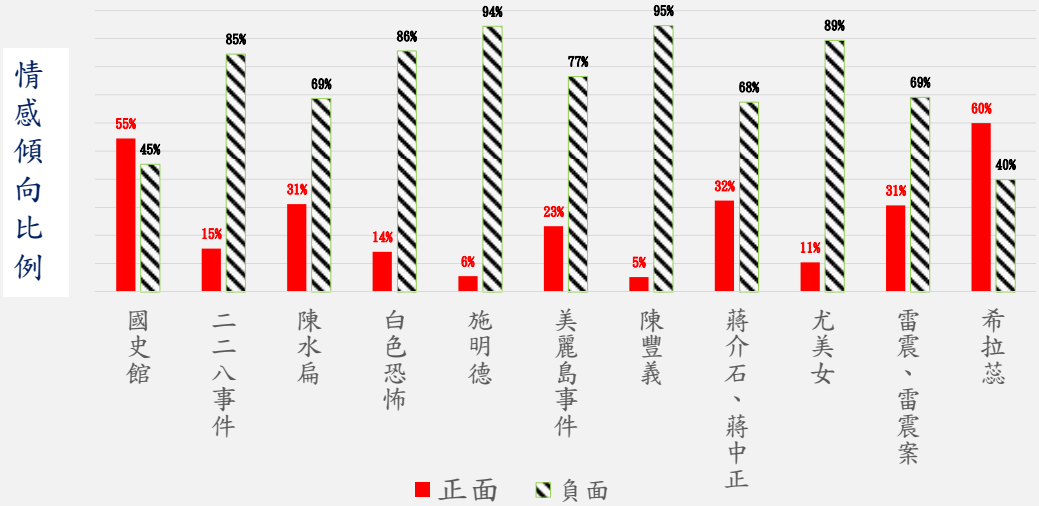
序號	詞彙	詞頻	詞彙出現之文本數	內容涉及檔案管理議題
1	檔案法	713	419	*
2	檔案局	270	79	*
3	國史館	250	55	檔案開放應用
4	二二八事件	160	65	檔案徵集 檔案開放應用
5	陳水扁	130	48	檔案銷毀
6	白色恐怖	106	56	機密檔案管理 檔案開放應用
7	施明德	100	18	個人隱私 檔案開放應用
8	美麗島事件	88	30	檔案開放應用
9	陳豐義	87	19	檔案銷毀
10	蔣介石、蔣中正	86	37	檔案徵集
11	尤美女	48	19	檔案徵集 檔案開放應用
12	雷震、雷震案	48	13	檔案開放應用
13	希拉蕊	36	5	檔案銷毀 機密檔案管理

檔案新聞議論事件

國史館	・高頻詞出現於2016年公告限制大陸地區人民調閱國史館資料，引起探討檔案開放應用政策。
二二八事件	・分布於1997年或2018年的報導內容，皆與強調全面完整徵集二二八事件檔案相關，以盼還原真相。
陳水扁	・高頻詞出現在有關政黨輪替的報導，其擔任總統期間相關承辦的公務文件，應依照國家機密、檔案法妥善保管，並交接予下任總統，與檔案管理作業完整歸檔有關。
白色恐怖	・高頻詞出現於2016年爆發過往白色恐怖時期檔案的外流，魏姓民眾在網路購買資料的事件，該報導為檔案管理作業及民眾對於檔案社會形象問題。
美麗島事件	・出現於2003年美麗島事件檔案展於國父紀念館展出時，展品中出現涉及個人隱私的文件（如：施明德先生在獄中的書信），內容屬於與檔案隱私處理及開放應用政策有關。
陳豐義	・主要出現於2013年，監察院秘書長陳豐義不當銷毀未逾保存年限且須永久保存的案卷，遭監察委員彈劾的事件，大量的報導皆與檔案管理銷毀作業有關。
尤美女	・詞頻出現於2016至2017年間，行政院推動《促進轉型正義條例》、《政黨法》以及推動《政治檔案法》三大草案，輿論重點包括清查所有政府機關的政治檔案避免外流，以及史學家認為許多兩蔣時代關鍵性的檔案文件沒有曝光。
雷震、雷震案	・詞彙集中於2016至2017年期間，由於政府對於促轉型正義的重視，國史館對於「雷震案」等敏感議題檔案的管理應用，提出有別於過往的作法，和檔案開放應用議題有關。
希拉蕊	・主要出現在2016年美國總統大選期間，被爆使用私人電子郵件伺服器進行官方通信，恐涉及國家機密文件不當流通並違反聯邦法規定，主要討論公務電子郵件管理問題。

35

檔案新聞報導各項主題呈現之情感傾向



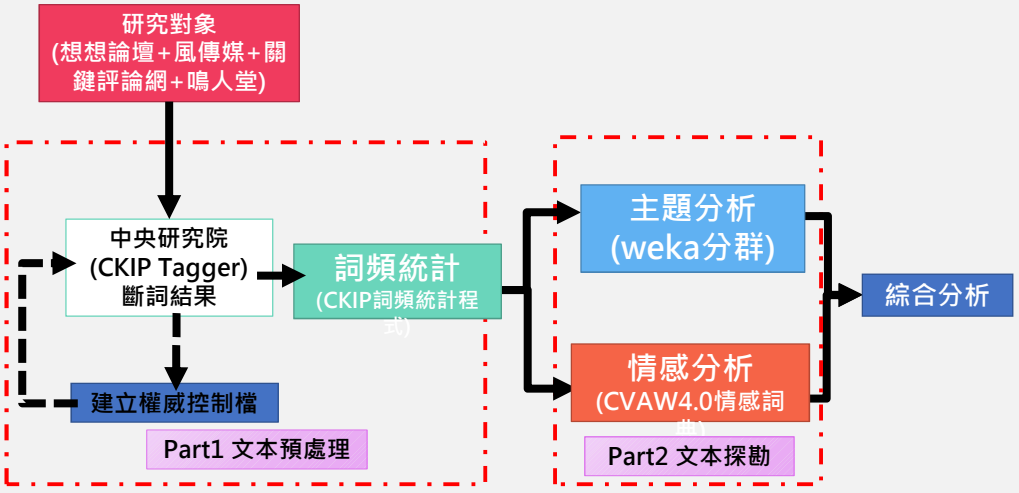
36

新聞報導分析結果討論

- 新聞報導對於檔案解密與開放應用是相對比較關注的議題，不乏議論有關「二二八事件」、「白色恐怖」、「雷震案」等檔案徵集和開放應用問題，建議國家檔案應用政策需要加強宣導，俾於建立社會共識。
- 434篇檔案相關新聞文本的情感傾向，分析結果有61%的文本呈現負面情感傾向，如果以人工檢視檔案新聞主題內容，可知新聞報導對於要求檔案公開、檔案不當管理問題的輿論，影響整體文本的情感傾向，從中可發現「要求」、「問題」、「公開」等強烈情感詞彙，會影響文本負面與正面情感，透過這樣的分析，可運用機器學習預測結果，找出影響情感的重要詞彙，亦可預測其他新聞內容對於大眾情感的影響。

37

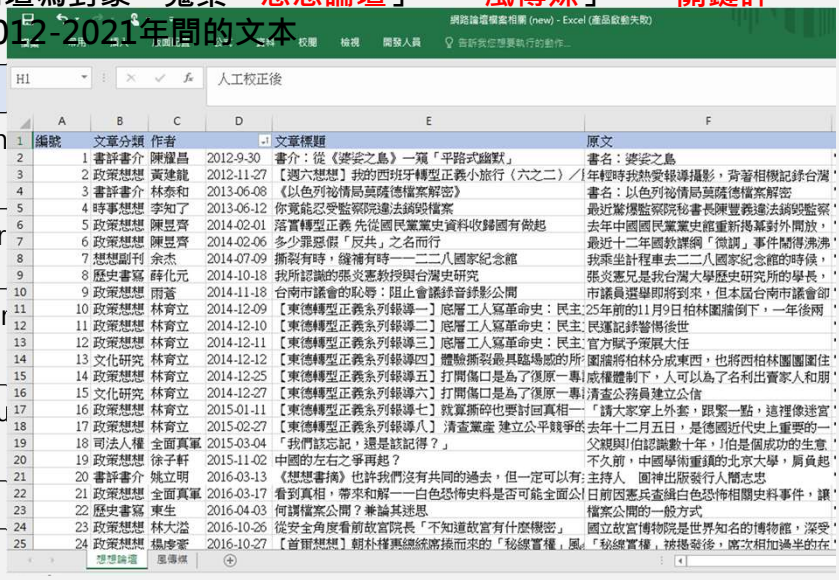
網路論壇資料分析過程



1.文本蒐集整理

以評論時事的長篇議論論壇為對象，蒐集「想想論壇」、「風傳媒」、「關鍵評論網」和「鳴人堂」，2012-2021年間的文本

來源名稱	網址
想想論壇	https://www.thin .com/
風傳媒	https://www.stor
關鍵評論網	https://www.ther om/
鳴人堂	https://opinion.u pinion/index
總計	



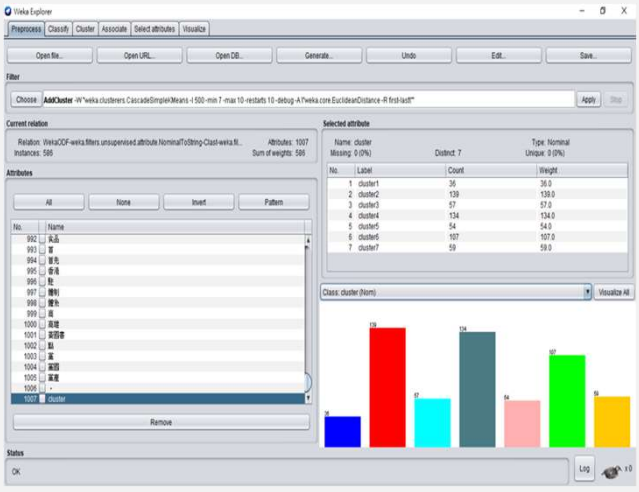
2. 文本斷詞

- 中文不像英文以空格區分詞彙，文本斷詞是進行文本探勘前重要的前置處理，進行最小有意義單位的切割。本研究採用中央研究院所開發的「CKIP Tagger」斷詞系統，進行文本斷詞。

3. 權威控制&校正

- 由於以上斷詞結果，常有專有名詞被切分的現象，例如：「轉型正義」，被分為「轉型」及「正義」，因此需先建立「權威控制檔」，匯入CKIP Tagger的參考詞典，進行回饋修正。

4. 利用Weka進行主題分群



Weka操作頁面

Weka分群分析器

社群論壇議論有關檔案之主題概念

群集	分群筆數	高頻詞(依詞頻高低排序)	檔案管理議題
第1群	224	文件、英國、民進黨、解密、機密、證據、蔡英文、立委、香港、圖片、審查、公文、外交部、保密、法國	**檔案解密公開**
第2群	106	政治、檔案、轉型正義、國家、政府、 社會、工作、歷史、台灣、威權、時期、總統民主、促轉會、國民黨	**轉型正義與政治檔案**
第3群	110	歷史、政府、檔案、政治、社會 國家中國、美國、作為、文化、事件、時期、存在、大學、國際	**歷史研究**
第4群	62	台灣、國家、委員會、檔案、檔案管理、樂活、國家檔案、情報、政府、國發委、檔管局、民國、日本、台北、歷史	**檔案推廣應用**
第5群	84	台灣、中國、政府、蔣介石、政權、政治、日本、人民、軍事、組織、工作、 事件、委員會、統治、美國	**檔案稽憑功能**

43

5. 情感分析

STEP 1

- 將文本斷詞結果58057個詞彙與CVAW4.0情感詞典比對，找出情感詞，利用ACCESS匯入詞彙以查詢功能做比對，共對應出4173個情感意義的詞彙。

STEP 2

- 將文本斷詞結果，利用EXCEL中的COUNTIF公式，計算586篇單一文本中出現詞彙的詞頻數。

STEP3

- 將文本斷詞結果，利用EXCEL中的VLOOKUP公式，與STEP1中4173個帶有情感的詞彙，回推586篇文本全文做比對。

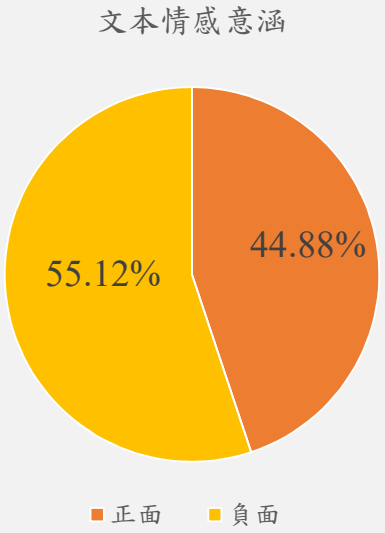
STEP 4

- 將586篇文本各詞彙詞頻與4173個帶有情感的詞彙分數相乘，正面詞彙與負面詞彙分別做計算後相加，即可得出文本情感意涵。

44

網路論壇文本情感分析結果

情感傾向	文本數	百分比(N=586)
正面	263	44.55%
負面	323	55.12%
總計	586	100%



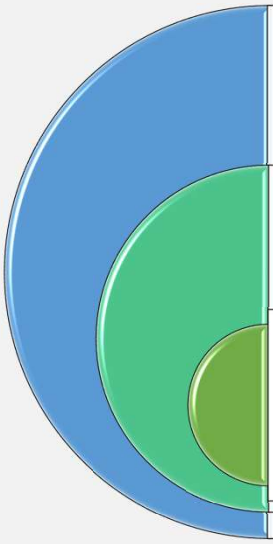
45

網路論壇討論各項主題之情感意涵

主題	正面（篇）	負面（篇）	總數（篇）	該主題文本整體情感
檔案解密公開	105	119	224	負面
轉型正義與政治檔案	41	65	106	負面
歷史研究	48	62	110	負面
檔案推廣應用	49	13	62	正面
檔案稽憑功能	20	64	84	負面

46

綜合分析結果



新聞報導與網路論壇關注之議題偏重在檔案徵集、檔案開放應用以及機密解密等問題，檔案機關如能重視這些大眾關注的議題改善社會觀感，將有助於檔案社會形象的提升。

新聞報導文字以負面情感偏多（批判角度），網路論壇文字呈現之情感亦傾向負面為多。

關於檔案要求公開、檔案不當管理等問題用詞強烈，致使整體文本的情感偏負面。一旦文本有「要求」、「問題」、「無法」等負面詞彙量偏多，容易導致文本呈現負面情感。反之，若為「開放」、「通過」等正面詞彙居多，整體呈現傾向正面。

47

結語

提取高頻詞雖可呈現整體概觀，但為提升斷詞及自動分類結果，仍需要輔以史學者意見調整及修正。

運用資訊工具分析尚無法完全取代專家判斷，但提供初步分析結果可提供整體概觀，也能減少人為主觀認定的印象。

斷詞精確與否影響分析結果→斷詞工具的精進 & 輔以人工過濾合併

中文詞義解釋的多元性→輔以語意分析

他唱歌很好聽　他唱歌好聽嗎？　他唱歌沒什麼好聽

48

未來發展方向

- 以檔案內容探勘技術協助建立全自動化/智慧化檔案管理流程
- 有待探索與解決的問題

未來發展方向

研究目標

全自動化/智慧化檔案管理流程

- 從鑑定、徵集/移轉、登錄、編排描述、提供檢索應用、支援研究、數位策展一系列以全自動化為主之檔案管理和應用流程
- 有待探索

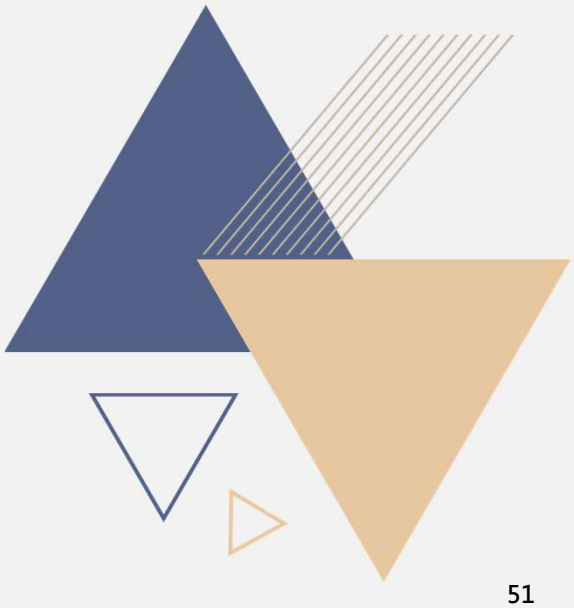
有待解決的問題

專業知能/技術問題

- 非資訊專業需學習資料探勘技術知識
- 如何選擇符合資料分析目的之探勘工具
- 研究結果可能無法適用全部文本
- 需要人類智能輔助的比例

感謝 惠請賜教

THANKS FOR LISTENING



51