數位典藏與數位人文 6 期（2020 年 10 月）
Journal of Digital Archives and Digital Humanities No. 6, pp. 69-96 (October, 2020)
DOI:10.6853/DADH.202010_(6).0003

以文字探勘技術分析臺灣四大報文字風格

余清祥 [1,*]、葉昱廷 [2]

摘要

媒體對同一主題的報導經常有明顯差異，本文以報紙報導風格為研究目標，比較臺灣四大報《蘋果日報》、《自由時報》、《聯合報》、《中國時報》的報導風格。避免報紙立場及題材造成的干擾，本文選取每天頭版頭條新聞，資料期間為 2012 年至 2018 年，其中頭條標題涵蓋四大報，但頭條內文僅有《蘋果日報》、《自由時報》。本文分析常用詞彙、句長、標點、虛字，發現四大報頭條標題及內文的用字風格不完全相同，其中標題的 Jaccard 指標將同時期報導放在同一群，Yue 指標將四大報分成三群；內文分析發現《自由時報》傾向政治新聞，《蘋果日報》傾向社會新聞，以能夠代表兩報的常見詞彙為解釋變數，統計及機器學習模型都能準確地區隔《自由時報》或《蘋果日報》。

關鍵詞：寫作風格、臺灣四大報、探索性資料分析、相似指標、關鍵詞

投稿日期：2020 年 3 月 13 日；通過日期：2020 年 6 月 5 日。
[1] 國立政治大學統計學系教授。
[2] 國立政治大學統計學系碩士。
[*] 通訊作者：余清祥，Email: csyue@nccu.edu.tw

壹、引言

新聞媒體大約從 19 世紀後半開始，對世界各國發生重要影響，引領輿論走向及群眾思維。新聞媒體是大眾認知的「第四權」，有別於行政權、立法權、司法權三權分立，以報導、評論等方式探索問題及根源，藉由發表公共見解引領討論及關注，發揮非官方的制衡力量（Boyce, Curran, & Wingate, 1978）。歷史上有不少因為新聞報導而發生影響的案例，較為知名者包括 1972 年美國水門案導致尼克森總統辭職，1989 年六四天安門事件（坦克車鎮壓）使得西方全面制裁中國。由於強大的媒體力量，為避免公共廣電媒體受到商業團體或外力干擾，不少國家（如德國、日本）規定家戶繳交廣電費，以維持公共媒體的獨立營運。

即使不少國家設置公共廣播媒體之類的組織，外力仍透過不同方式試圖操縱及控制媒體，假借大眾公益的名義，藉由媒體的獨立客觀性達到不正當目的。例如：「假新聞」（fake news，或譯為假訊息）是近年與媒體操控有關的議題，自從 2016 年美國選舉後成為全球熱門話題，由於資訊不對稱等因素，很難區分新聞內容哪些是事實、哪些是報社及記者的主觀想法（如：美國兩大新聞媒體 Cable News Network〔 CNN 〕及 Fox News 的立場鮮明）（Lazer et al., 2018）。如何從報導中分辨哪些是捏造的假消息並不容易，一般讀者很難追蹤消息來源，不少國家將分辨訊息來源列為必要的國民教育，以打擊假消息的傳播。像是與俄國交界長達 1,000 公里以上的芬蘭，為了面對強鄰的潛在威脅，提供國民分辨假新聞的方法，宣導假訊息帶來的傷害，認定這是未來資訊戰的一部分（Horn & Veermans, 2019）。

臺灣人口數及土地面積不算大，文化及族群卻相當多元，因為崇尚自由、民主，對不同思維的容忍度非常高，經常可看到風格迥異的言論，因此更難追蹤新聞報導的訊息來源。以臺灣的報紙為例，發行量較多的前四大報紙為《蘋果日報》、《自由時報》、《聯合報》、《中國時報》，這幾家報社的題材偏好、風格立場有明顯不同，讀者或許能夠從版面編排及報導角度等分辨報紙，但也因選材不同，更難找到共同標準驗證報導的真偽。

原先本文目標為分析假新聞、一般新聞的報導風格，希冀研究發現可

作為大眾判斷資訊真偽的參考，但鑑於偵測假訊息牽涉層面較多，單從文字使用、報導方式及內容很難判斷真偽，本文將研究目標訂為探討上述四家報紙的文字風格，以量化角度解讀各報的報導特色。由於報紙選材差異頗大，為了避免抽樣偏差、樣本數不同等因素造成的干擾，本文選擇四大報每天的頭版頭條新聞（包括標題、內文報導），探索各報紙標題及內文的寫作風格，資料期間為 2012–2018 年。頭版頭條除了每日僅有一則，各報紙有相同樣本數外，一般頭條新聞也是當日最重要的新聞，多半各報會選擇相同議題，更方便於比較四大報報導風格的異同。另外，由於本研究以網路方式下載資料，受限於各報於網路揭露訊息的規定，《聯合報》和《中國時報》只有頭條標題，內文比較僅涵蓋《蘋果日報》、《自由時報》兩家報紙。

本文與一般文字分析最大不同在於採取非監督學習（unsupervised learning）的想法，亦即先不設定研究標的，基於資料驅動（data driven）的原則，透過探索性資料分析（exploratory data analysis, EDA）找出各大報頭條的文字特性。這些特性並不侷限於常見字詞及其統計分布，也包括句長、標點、虛字等測量值，這些分析項目看似不起眼，但往往可區隔作者寫作風格的關鍵資訊。除了探索寫作特性外，本文也考量以統計方法選取重要解釋變數，作為分類報紙的依據，這部分將以《蘋果日報》、《自由時報》的頭條內文為分析對象，分類模型則為常見的統計學習、機器學習方法。

由於中文為單位相同的方塊字，很適合將每個字或詞視為事件（event），套用機率的概念；另外，本文的 EDA 引入生物學及生態學的概念，將每個中文字視為不同物種（species），套用生態系（ecology system）及物種多樣性（species diversity）的想法描述報導特性。本文將在第二節先介紹 EDA 和物種多樣性的基本想法，之後再繼續說明實證分析的執行方式，示範如何將之應用在寫作風格的分析。

另外，由於篇幅限制使得頭條標題無法像一般報導，必須在一句或兩句話內摘要出新聞重點，類似一本書的書名、文章章節名稱，預期頭條標題用字特性與頭條內文明顯不同。本文將在第三節以 EDA 找出標題、內文的用字差異，除了常用字詞的總類及其分布外，也會比較上述提及的其

他寫作特性，像是句長、標點符號、白話文虛字等。不過，本文分析不會涉及議題（如：主題模型），或是報紙立場與議題設定（或新聞選擇），因為這多半會與擷取關鍵詞（keyword extraction）有關，雖然找出能夠代表文章主題的關鍵詞是重要的研究議題，但這並非本文的研究方向。

貳、研究方法

本文報導風格的分析可分為：EDA、驗證性資料分析（confirmatory data analysis, CDA），此與大數據分析的三種分類略有不同（Evans & Lindner, 2012）。這三種分析為敘述性分析（descriptive analytics, DA）、預測性分析（predictive analytics, PA）、處方性分析（prescriptive analytics），目標各為「發生了什麼事」（What has happened）、「未來會如何」（What would happen）、「我們如何調整」（What should we do）。EDA 與 DA 的角色較為類似，將資料的主要特性以較易理解的指標（或新變數）或圖表呈現，而 CDA 與 PA 較為接近，著重於引進模型估計及預測目標變數的未來趨勢。

EDA 由統計大師 Tukey（1977）提出，主旨在於挖掘變數的基本特性（如：平均數、變異數）以及變數間的關聯（如：相關係數），並以圖形表格等視覺化工具呈現，讓研究者能夠一目了然地洞悉資料特性。本文的分析對象是中文，建議 EDA 分析範圍包括總字（詞）數、不同字（詞）彙數、不均度指標、相似指標等。將文字對照於物種多樣性，可以把字（詞）彙視為不同物種，藉此探討字（詞）彙的豐富度及不均度，該指標數值越高代表生物多樣性越高，越低則代表有明顯的優勢物種（dominant species）存在，在文字分析領域代表用詞越不均勻，或是用詞較侷限於某些特定詞彙。

其中不均度指標計有字（詞）彙的熵（entropy，又稱為 Shannon index）及辛普森指標（Simpson index），用於描述字（詞）彙的豐富程度（species richness），判斷文章字（詞）彙使用的基本特性（Yue & Clayton, 2005）。令 p_i 為第 i 種字（詞）彙的出現機率，則 entropy = $-\sum_i p_i \log(p_i)$ 可用於測量不均度、Simpson index = $\sum_i p_i^2$ 測量常見字（詞）彙等之集中程度。這兩個指標的方向剛好相反，entropy 數值愈大、分布愈平均，而 Simpson 數值

愈大則代表資料愈集中，因為兩者通常顯示類似的結果，本文將只顯示其中一個指標的結果，

本文也考量 Jaccard 及 Yue 兩個相似指標（similarity index），透過用字（詞）頻率等測量四大報紙間的關聯，作為後續統計分析的測量值，想法類似進行迴歸分析前先計算相關係數（correlation coefficient）。相似指標數值越高、兩組樣本的組成越相似，或是兩個生態系的優勢物種組成比例相近，用於詮釋文本的特徵，代表兩組文本的常用詞彙非常相似，意謂用詞類型及比例接近或是報導採材角度類似。Jaccard 指標與 Yue 指標通常用於生物學及生態學，兩者各有特色及使用限制，主要差異在於前者考慮不同字詞數、後者加入字詞使用的頻率，兩者的範圍都介於 0 與 1 間（Yue & Clayton, 2005）。Jaccard 指標只考量字（詞）彙總數，定義為：

$$\theta_J = \frac{S_{12}}{S_1 + S_2 - S_{12}} \tag{1}$$

其中 S_1、S_2 分別為兩份報紙的所有字（詞）彙數，S_{12} 為兩份報紙共同使用的字（詞）彙數。Yue 指標考慮字（詞）的出現機率，定義為：

$$\theta_Y = \frac{\sum_i p_i q_i}{\sum_i p_i q_i + \sum_i (p_i - q_i)^2} \tag{2}$$

其中 p_i 和 q_i 第 i 個詞彙在兩份報紙的出現頻率，相似指標的 $\sum_i p_i q_i$ 和 $\sum_i (p_i - q_i)^2$ 分別量測兩個年度相似、相異程度。

上述的相似指標可用於集群分析（cluster analysis）、社會網絡（social network）等分類方法，比較各大報的差異。另外，本文也採用資料縮減方法 *t*-distributed stochastic neighbor embedding（*t*-SNE）降低資料的維度，透過廣義相關圖（generalized association plots, GAP）以視覺化描述分類結

果，並以時間數列方法分析趨勢變化。廣義相關圖是由中央研究院統計科學研究所陳君厚博士所提出非降維的 EDA 視覺化方法（Chen, 2002），透過變數關係矩陣、樣本關係矩陣，以階層式集群分析（hierarchical cluster analysis）整合並繪製成充分統計圖（sufficient data map）、沉澱矩陣圖（sediment map）、條件矩陣圖（matrix condition map）等多解釋層面的視覺化呈現。本次研究採用階層式集群分析和充分統計圖進行分析及詮釋。

本文也考量時間數列（time series）模型測量相似指數的變化趨勢，在此使用整合移動平均自迴歸模型（autoregressive integrated moving average model, ARIMA）。ARIMA 模型是最常見的時間數列模型，其中 "I" 用於當資料不屬於平穩（stationary）數列，像是期望值、變異數不為定值時。四大報的相似指標都為平穩數列，僅需考慮 AR(p) 及 MA(q)：

$$Y_t = \mu + (\phi_1 Y_{t-1} + \cdots + \phi_{t-p} Y_{t-p}) + (\theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}) \quad (3)$$

其中 Y_t 為相似指標數列、e_t 為白噪音（white noise）。

參、研究結果

本文分析的臺灣四大報頭條新聞，來自於中央通訊社網站收錄臺灣四大報的每日報紙頭版的開放電子紀錄，[1] 但這個網站只有各報頭版紀錄，沒有頭條標題及其內文的電子檔。本文的頭條報導資料是以 python 程式軟體的網路爬蟲（Web Crawling）蒐集各大報電子報，時間為 2012 年 1 月至 2018 年 12 月（共 84 個月），其中電子報的來源分別為：「蘋果新聞網」網站的《蘋果日報》（https://tw.appledaily.com/daily）、「自由時報電子報」網站的《自由時報》（https://news.ltn.com.tw）、「中時電子報」網站的《中國時報》（https://www.chinatimes.com/newspapers/2601?chdtv）、「聯合新聞網」網站的聯合新聞（https://udn.com/news/index）。原則上，各大報有 2,557 則頭條新聞，但有時報紙會以廣告取代頭版版面，某些報紙的樣本數會略小於 2,557。

1　範例見中央通訊社（2018）。

電子資料下載後先進行資料比較與整併，對應（mapping）中央通訊社頭條標題與電子報標題，比對發現電子報與實體報紙的頭條標題有六成以上用字完全相同，主要差別為用詞調整或將中文數字改成阿拉伯數字，通常不會影響標題的原意。接著繼續進行資料前處理，去除英文、數字及空白等雜訊後，再對處理後的中文單字進行分析。對於雙字詞等多字詞的分析，則以結巴斷詞系統（Jieba）進行斷詞處理，使用 R 軟體中的 jiebaR 套件，並加入 N 元語法（N-gram）及隱馬可夫模型（hidden Markov chain）改善斷詞效果。

本文使用 R 軟體中的 jiebaR 套件進行文本斷詞，然而結巴斷詞系統有一定的限制，不易判斷人名、地名等專有名詞，而且無法根據文本特性加入新詞，因此加入 N 元語法及隱馬可夫模型改善斷詞效果。N 元語法考慮了長詞優先法則（maximum matching），先針對文本進行第一次斷詞，接著篩選合適的高頻詞並結合結巴斷詞系統的原詞庫，建立一個臺灣四大報頭版標題專用的詞庫，再將隱馬可夫模型導入結巴斷詞系統進行斷詞，改善專有名詞及新詞的偵測效果。以 2018 年《蘋果日報》資料為例，先彙整該報 2018 年每一天的標題斷詞結果，再以一年為單位，統計每個單字或每個詞彙出現之頻率，以利後續進一步分析。

為了方便讀者理解分析的方向及步驟，四大報文字 EDA 大致分成兩個部分：一階動差（first degree moment）及二階動差（second degree moment）。一般的探索性分析通常會計算一階及二階動差相關的統計量，包括樣本數、平均數、中位數（以上為一階動差）及變異數、相關係數（以上為二階動差）。以下將分項整理文字分析的一階統計量（一階動差）動差及二階統計量（二階動差）。

一、一階統計量

除了字詞種類、常見字詞及其比例外，本文一階動差的 EDA 也考量句長、標點、虛字。加入這些測量值是因為能夠有效鑒別某位作者作品的關鍵，經常是看似無關緊要的細節，像是介係詞、代名詞等的使用習性，以英文文本分析為例，冠詞 “the”、所有格 “his” 和 “her”，或是介係詞 “of” 的使用頻率，都是區隔暢銷書、非暢銷書的重要特徵（茱蒂．亞契、馬修．

賈克斯，2016/2016，頁 119）。首先整理臺灣四大報頭條標題及內文的字詞基本統計數值（表 1），其中內文僅有《蘋果日報》及《自由時報》。標題、內文的總字數差異很大（70 倍以上），雙字詞數也有不小的差異（6–8 倍），但字彙數的差別卻不到兩倍，這個現象頗為符合歷來的白話文研究結果（例如：何立行、余清祥、鄭文惠，2014）。白話文是以日常口語為基礎的書面語，通常藉由雙字或多字詞表達意義，有別於文言文的一字多義，常用字彙數僅約四千餘字（林樹，1972）[2]，因此頭條標題及內文的字數差異非常大，但總字彙數卻沒有等比例增加。另一個現象也能印證白話文以多字詞呈現：最常見的前 500 個單字累計約占所有字數的七成以上，但最常見的前 500 個雙字詞累計不到所有雙字詞的四成，亦即單字字彙多樣性較低，但雙字詞詞彙有較高的多樣性。若以字彙及雙字詞總數、前 500 大所占比例判斷，平均而言《蘋果日報》的字彙及雙字詞多樣性較高。此外，《蘋果日報》和《自由時報》頭條內文的單字、雙字詞分布接近齊夫法則（Zipf's law），但四大報標題的單字與雙字詞都不服從齊夫法則，限於篇幅不列出結果。

接著觀察字彙及雙字詞多樣性的時間趨勢，限於篇幅僅展示頭條標題的結果。《蘋果日報》頭條標題的不同單字種類最多，除了《聯合報》字彙數近年有逐年下降的傾向，其他三家報紙的字彙多樣性從 2015 年後穩定上升（圖 1）。若以 N-gram 搭配結巴斷詞系統進行斷詞，四大報的歷年

表 1　臺灣四大報頭條標題及內文的字彙、雙字詞統計（2012–2018 年）

	內文		標題			
項目	蘋果日報	自由時報	蘋果日報	自由時報	聯合報	中國時報
文章篇數	2,557	2,556	2,557	2,556	2,550	2,548
單字個數	**4,963**	**4,512**	2,693	2,462	2,426	2,598
雙字詞數	**34,935**	**27,653**	4,456	4,283	4,255	4,625
前 500 字字數比例	2,573,206 (72.6%)	1,755,408 (**76.2%**)	23,042 (**67.6%**)	25,074 (71.6%)	24,967 (71.9%)	24,460 (69.4%)
前 500 雙字詞比例	337,645 (**35.1%**)	237,396 (38.0%)	3,449 (39.7%)	3,712 (41.5%)	3,751 (41.5%)	3,459 (**38.1%**)

資料來源：作者自行整理。

註：網底及紅色數字者為較為明顯不同的結果。

2　教育部公布的常用字為 4,808 字（國家教育研究院，n.d.）。

統計結果如圖 2，其中《蘋果日報》的雙字詞種類仍然較高，且與單字相同在 2015 年後雙字詞種類快速增加，而《中國時報》及《自由時報》的字彙及雙字詞種類位居第二名及第三名。字彙及雙字詞多樣性較低的《聯合報》，近幾年呈現下降趨勢，2018 年的字彙與雙字詞數都比《蘋果日報》少用了約 300 種（約兩成）左右，這是非常明顯的差別，背後原因值得深入研究。

寫作風格並不侷限於常見字詞，本文也考量句長、標點、虛字（function word）等特徵，其中考量白話文常見的十個虛字：「的」、「是」、「和」、「個」、「了」、「們」、「著」、「麼」、「嗎」、「吧」（何立行等人，2014）。由於標題字數有限，通常會以空格作為語氣轉換（類似現在的網路文章），不會使用標點符號，因此每篇報導的句長、總字數大致相當，四大報的標點符號使用比例非常低（《聯合報》最高）。這個現象和內文報導非常不同，無論《蘋果日報》或《自由時報》，內文的標點使用率相當高，而這兩大報的句長也很接近（表 2）。另外，內文白話文虛字的使

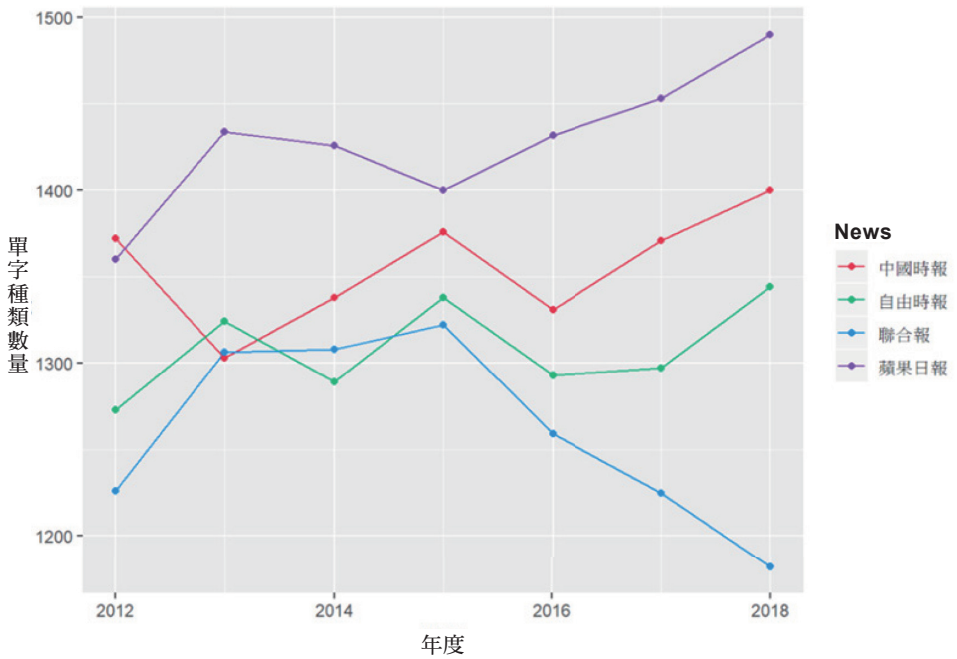

圖 1　四大報每年頭條標題的單字種類

資料來源：作者自行整理。

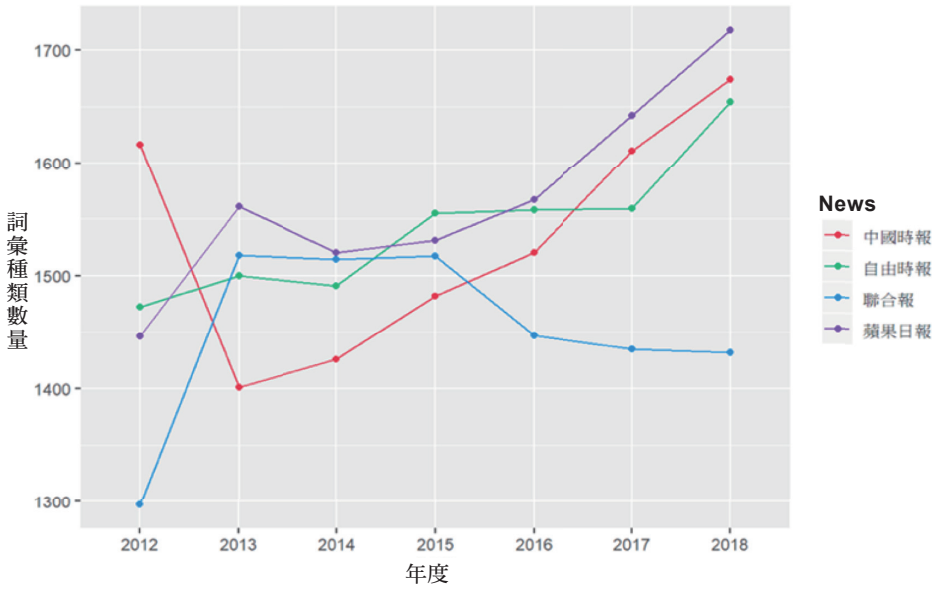

圖 2　四大報每年頭條標題的雙字詞彙種類

資料來源：作者自行整理。

表 2　臺灣四大報頭條標題及內文的標點、句長、虛字統計（2012–2018 年）

	內文		標題			
項目	蘋果日報	自由時報	蘋果日報	自由時報	聯合報	中國時報
標點比例	**0.1259**	**0.1121**	0.0338	0.0357	**0.0548**	0.0355
每篇字數	1701.2	1042.7	16.7	16.4	15.9	16.1
句長	14.3	14.7	16.5	15.7	14.7	15.2
虛字比例	**0.0285**	**0.0295**	0.0062	0.0054	0.0104	0.0092

資料來源：作者自行整理。

註：網底及紅色數字者為較為明顯不同的結果。

用比例明顯高於標題，其中「的」差異甚至高達 10 倍左右，這和標點符號的使用情形類似，猜測頭版標題字數的限制，壓縮了標點符號、白話文虛字的空間。

除了總字詞數及不同字詞數外，也進一步觀察常見的單字及雙字詞，除了可比較各大報的特性外，也能觀察出頭條標題及內文的用字差異。標題以較少字數概要的說明新聞內容，平均每則頭條標題約 16 個字，有別於每則頭條內文報導超過 1,000 字，由於字數少、更需精簡選擇用字，因此比較不會出現代名詞、虛字等功能性用字，「的」、「在」、「是」是

內文常見單字前十名可作為證據（表 3）。另一個比較有趣的現象是《蘋果日報》的單字前十名，「死」、「女」、「萬」、「殺」不是其他報紙標題或內文的前十名，判斷這幾個字與社會新聞有關，通常不會出現在國際或國家重要新聞，此與《蘋果日報》的市場定位與報導風格有關。附錄列出使用「死」、「女」、「萬」、「殺」的幾則頭條標題，這些報導應與社會新聞有關，藉此可推論《蘋果日報》選擇頭條新聞的角度與其他報紙確實不同。

頭條標題及內文的前十名雙字詞顯示類似結果（表 4）。頭條內文的前十名雙字詞中有幾個與文章報導的用詞有關（「表示」、「報導」、「發現」、「指出」），這類型用詞沒有出現在頭條標題的前十名。《蘋果日報》與其他報紙的差異從雙字詞也可看出，除了該報傾向於將社會新聞放在頭條（黃底雙字詞），而且總統及兩岸相關用詞都是其他三大報的前十名雙字詞，卻不是《蘋果日報》頭條標題的前十大雙字詞。對於兩岸關係的用詞，《中國時報》及《聯合報》傾向於使用「大陸」、「兩岸」而非「中國」；另外，這兩家報紙的用詞似乎很接近，標題中的前十名雙字詞中有一半相同，但與其他兩家報紙僅有兩個相同。[3]

表 3　臺灣四大報頭條標題及內文的前十大字彙（2012–2018 年）

	內文		標題			
排序	蘋果日報	自由時報	蘋果日報	自由時報	聯合報	中國時報
1	的	的	死	國	大	台
2	人	一	台	中	台	大
3	不	國	女	台	人	國
4	一	人	人	大	年	人
5	有	不	萬	年	國	不
6	在	中	大	人	不	年
7	是	台	年	馬	會	中
8	年	會	不	不	中	一
9	大	在	殺	會	一	民
10	中	是	國	美	馬	全

資料來源：作者自行整理。

註：網底及紅色字彙為較為不同的結果。

3　《蘋果日報》與《自由時報》在標題前十大用詞中有三個相同。

表 4　臺灣四大報頭條標題及內文的前十大雙字詞（2012–2018 年）

	內文		標題			
排序	蘋果日報	自由時報	蘋果日報	自由時報	聯合報	中國時報
1	表示	表示	中國	中國	總統	兩岸
2	警方	中國	死傷	總統	兩岸	總統
3	蘋果	記者	小時	政府	英文	大陸
4	相關	總統	總統	立委	大陸	台北
5	萬元	政府	千萬	立院	立院	政府
6	報導	報導	蘋果	明年	民調	民調
7	發現	台北	離譜	學生	經濟	市長
8	資料	指出	曝光	下台	公投	共識
9	新聞	美國	學生	國安	共識	小時
10	可能	立委	女友	起訴	國道	道歉

資料來源：作者自行整理。

註：網底及有顏色的雙字詞為較為不同的結果。

二、二階統計量

除了字詞種類、常見字詞及比例外，EDA 通常會加上變異數、相關係數，亦即為隨機變數的二階動差，彌補字詞種類、比例等這類型的一階動差統計量。文字資料的二階動差可分為單一母體或兩母體間，entropy 和 Simpson 指標屬於單一母體，測量母體內部各種類的分布狀態，但兩個指標的方向相反，Simpson 指標愈小表示分布愈均勻，entropy 愈大愈均勻，通常也隱含多樣性較高。因為套用至四大報時，兩者顯示相同結果，僅以頭條標題的雙字詞 Simpson 指標說明（圖 3）。雙字詞種類數與 Simpson 指標間未必存在必然趨勢，種類愈多不見得指標值愈大（或愈小），以《中國時報》和《聯合報》為例，兩報在 2017–2018 年 Simpson 指標都是遞增，但雙字詞總類數呈現不同趨勢，前者上升、後者下降。同理，《蘋果日報》雙字詞種類數大致隨時間遞增，但 Simpson 指標卻沒有絕對遞增或遞減的現象，似乎雙字詞種類數與分布均勻間的關聯不高。

本文考慮的兩母體間二階動差包括 Jaccard 和 Yue 指標，其特性類似相關係數，可用於描述兩個母體的關聯性（association），其數值介於 0 與 1 之間，愈接近一代表兩個母體的用字（詞）愈相似。這類型的二階動差測量兩母體間的相似程度，通常稱為相似指標（similarity index），也可視為量測兩個觀察值間的「距離」，作為分類（classification）及群集（clustering）的依據，本文將在下一節詳加說明。在此展示相似指標的其

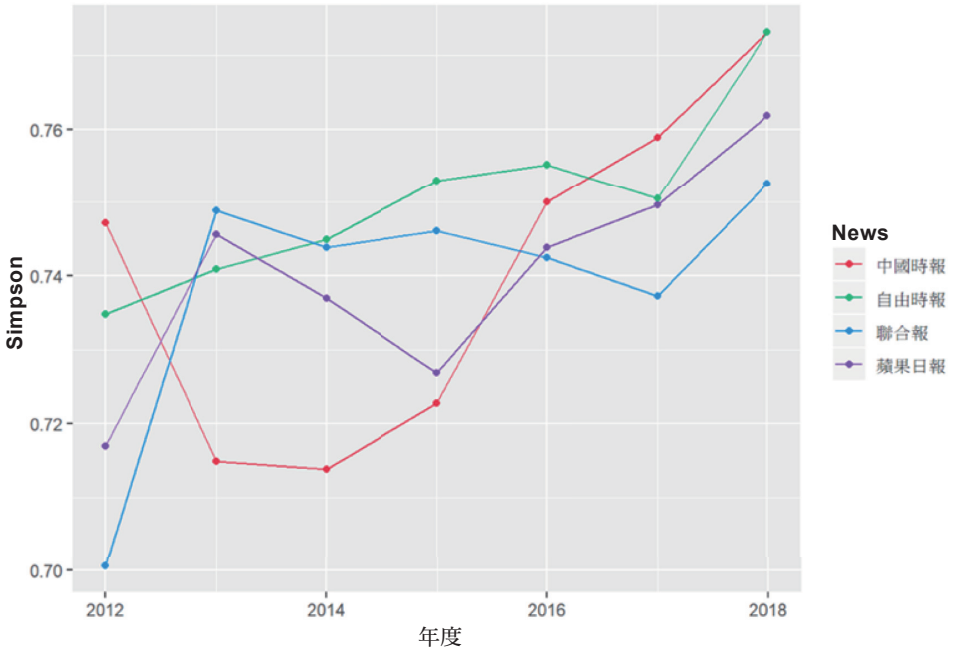

圖 3　四大報頭條標題的雙字詞 Simpson index

資料來源：作者自行整理。

他用途，可透過相鄰兩個月間的指標描述歷年四大報的用詞時間延續性，僅以 Yue 指標說明（圖 4）。Yue 指標測量相鄰兩個月的用字相似性，數值愈大、代表相鄰兩個月的用字比例愈相似，某些特定事件時指標稍微高一些，像是 2014 年反服貿、2016 年美國總統選舉，在這些大眾較為關注的特定事件發生時，各家新聞媒體通常會持續報導，因此相鄰兩個月的標題用字較為類似，使得相似指標數值較高。平均而言，《自由時報》的相鄰兩個月用字相似指標比較大，顯示標題用字具有時間延續性，而《蘋果日報》的指標則相對較低。

相鄰兩月的相似指標也可套入時間數列模型，探討是否存在系統性時間趨勢，可用於驗證前一段的觀察結果（表 5）。《聯合報》的 AR(1) 代表相似指標具有時間趨勢，相鄰兩個月的 Yue 指標間有關聯；《自由時報》及《中國時報》的 Yue 指標也具有時間趨勢，MA(1) 顯示相似指數的模型誤差有關聯。這三家報紙頭條標題用字大致與前一個月有些關聯，《蘋果日報》用字不具時間特性，模型分析結果與上一段圖 4 的結果一致。頭條內文相鄰兩月的相似指數也可仿造上述方式，在此省略不再討論。

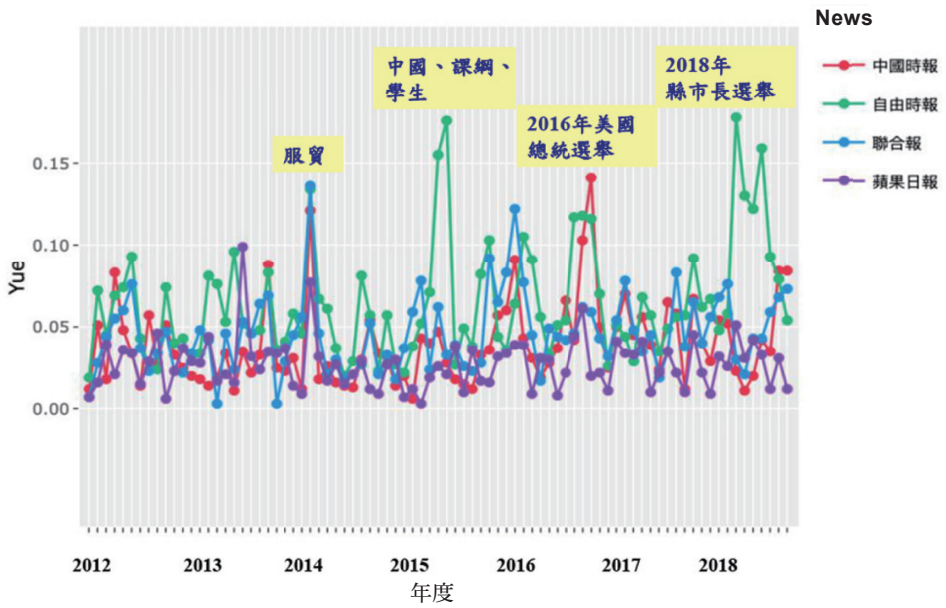

圖 4　四大報頭條標題的相鄰兩月雙字詞的 Yue 指標

資料來源：作者自行整理。

表 5　四大報頭條標題相鄰兩月相似指數的時間數列趨勢

報紙名稱	Jaccard 指標		Yue 指標	
	AR(p)	MA(q)	AR(p)	MA(q)
自由時報	0	1	0	1
蘋果日報	0	0	0	0
中國時報	0	1	1	0
聯合報	1	0	1	0

資料來源：作者自行整理。

註：AR = autoregressive models; MA = moving average models。

肆、群集與分類

延續前一節的 EDA，本節以驗證性資料分析為目標，透過合適的統計模型比較四大報頭條標題，並以常用詞作為分類《蘋果日報》及《自由時報》頭條內文的依據報導來源。前一節相似指標可描述相鄰兩月的用詞延續性，比較各大報在不同時間的特色，在此將相似指數套入集群分析。首先考量 Jaccard 指標的年度分群，以個別媒體報紙一整年的頭版標題為

基本單位，透過計算兩個年度間或兩家報紙間 Jaccard 指標，之後建立相似度矩陣作為分群用測量值，經過 GAP 呈現後更易觀察差異（圖 5）。經過集群分析處理後，似乎分群單位與時間有關係，與媒體報紙沒有關係，同一年度的資料通常被分到同一群內，而且 2012–2014 年較為相像、2015–2018 年距離接近。猜測 Jaccard 指標只考量用詞的種類，而用詞的種類又受到事件題材的限制，由於同一年有相同的重大事件，因此四大報頭版標題會使用相同詞彙。

Yue 指標的分類結果略為不同（圖 6）。Jaccard 指標傾向於以時間為分群單位，而 Yue 指標則以媒體為分群單位，例如：2012–2018 年的《蘋果日報》被歸類為一群，2013–2018 年的《自由時報》可歸類為另一群，而有趣的是《中國時報》與《聯合報》在分群上較難以區隔，或是這兩家報紙的用詞特性接近，此與前一節這兩家報紙前十大常用詞有半數相同頗為一致。Jaccard 指標著重於用詞類別，而 Yue 指標強調用詞頻率，兩者提供不同角度的詮釋，實證分析或可結合兩個指標。

由於頭條標題的資料長度限制，無法全然呈現文字表達的特色，前一

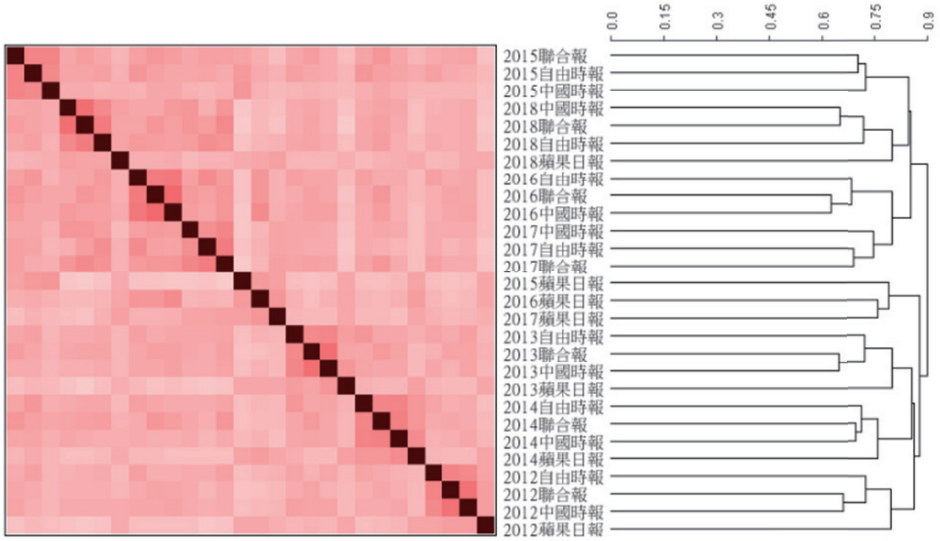

圖 5　臺灣四大報頭條標題的 GAP 群集分類結果（Jaccard 指標）

資料來源：作者自行整理。

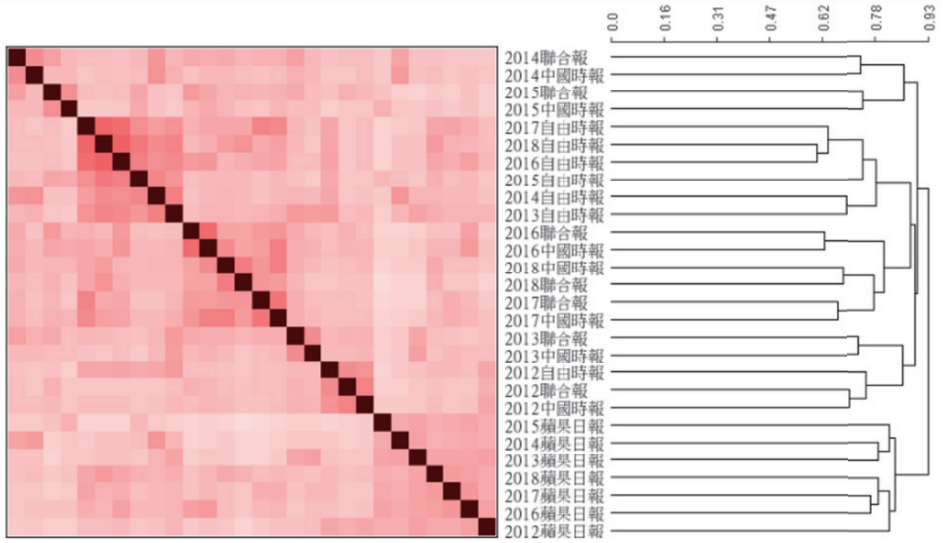

圖 6　臺灣四大報頭條標題的 GAP 群集分類結果（Yue 指標）

資料來源：作者自行整理。

節十大常見字彙沒有虛字就是其中一例，頭條內文的字數較多，能夠探討的層面當然更廣，諸如檢視用詞是否符合齊夫法則[4]或其他自然語言常見的特性。前面各項 EDA 分析可直接套用至頭條內文，本節將考量另一種分析方式，引進詞向量與文本向量的方法，對頭條內文的用詞進行分群，再以社會網絡（social network）探討媒體報紙的用詞布局。最後，以分類《蘋果日報》及《自由時報》的頭條內文報導為目標，選擇適合的詞彙為解釋變數，透過交叉驗證（cross validation）比較支援向量機（support vector machine, SVM）、決策樹（decision tree）、隨機森林（random forest）等方法的準確率。

頭條內文的篇幅較長，先以文本向量技術量化，透過 Word2vec 量化後的詞向量處理，取得代表該詞彙的語意屬性後，再計算 2012–2018 年之間任兩篇頭條內文間的餘弦相似度（cosine similarity），衡量兩篇文章的關聯性，作為後續文本集群分析的分析依據。詞向量間的餘弦相似度可用於探討詞向量的特性，作為後續分析的參考依據，例如：餘弦相似度可視為詞彙間的關聯性，透過 R 軟體中的 visNetwork 套件建立詞彙之間的社會

4　限於篇幅，在此不另外說明齊夫法則的分析結果，大致而言，頭條標題的用字不服從齊夫法則，但頭條內文的用字很接近齊夫法則。

網路結構，作為詞彙分群的依據。本研究著重於高頻詞的網路關係結構，也就是《蘋果日報》或《自由時報》中較常出現的詞彙，觀察兩大報紙在頭版的題材上大致分成幾個領域，以及在這些領域的布局情形是否有明顯的風格差異，作為用詞布局風格的比較。

由於文本向量在某種程度上可代表文本的語意屬性，透過語意屬性將內文分成數個層面，就能將頭條內文的詞彙進行分群。以《蘋果日報》的頭條內文為例，從集群分析的樹狀圖結果大致呈現 5 大群（或 12 個小群），經過 *t*-SNE 變換的視覺化呈現可觀察出 12 個小群的差異（圖 7），小群間重疊區域不大，顯示劃分 12 群的界線明顯。進一步觀察 12 小群中每一群的高頻詞，可嘗試為各群高頻詞的所屬領域命名，以第 5 群為例，前十大高頻詞包括「總統、臺灣、中國、川普、決戰、兩岸、王金平、蔡英文、關說、民主」，代表該群探討主題偏向國際與兩岸的政治議題。以同樣方式提供各群議題，第 1、2、3 群偏向感情糾紛的社會新聞，第 8 群偏向災害、第 10 群偏向經濟類犯罪、第 11 群則為一例一休的勞基法報導。

限於篇幅省略《自由時報》頭條內文報導的分群過程，最後可分成 6 大群與 12 小群，其中 12 個小群也都有明確議題，像是第 1 群為臺灣與中

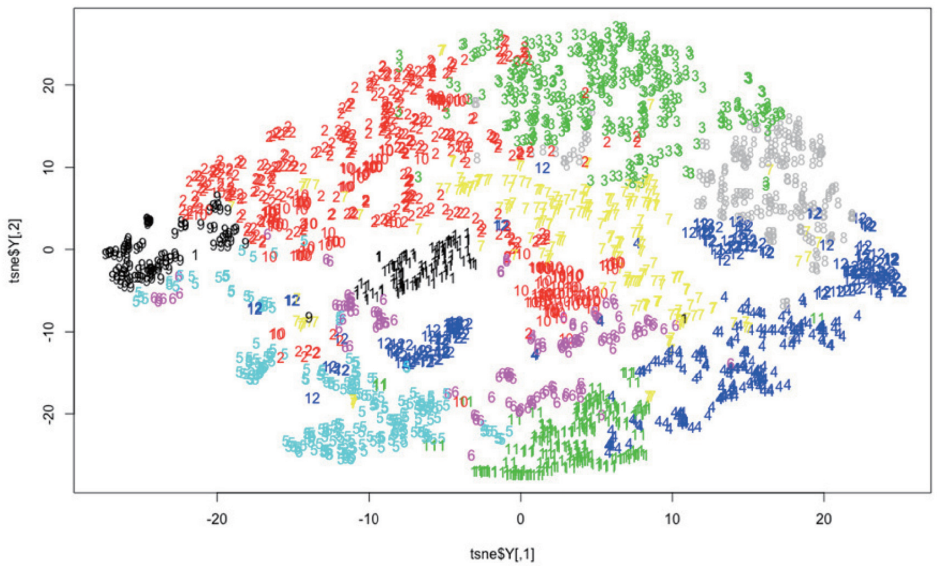

圖 7　《蘋果日報》頭條內文的 *t*-SNE 視覺化分群結果

資料來源：作者自行整理。

國之間的經濟報導，第 4 群聚焦於黨產問題，第 5 類偏向臺灣天災報導。比較特別的是，無論在大群及小群的分類中，「中國」是最常出現的高頻詞（6 大群中有 5 大群），而「台灣」一詞也經常出現，推測《自由時報》的頭條較常著眼於與臺灣、中國有關的議題。相對而言，《蘋果日報》頭條報導是以社會新聞為主，但很難從各小群的常用詞找到交集，也就是難以清楚地定義各小群的議題。由於各群高頻詞可為該群給定確切議題，我們認為只要能找到具有鑑別力的變數（如某些高頻詞），區隔出《自由時報》與《蘋果日報》的頭條報導應該不會太難。

正因為《自由時報》與《蘋果日報》的頭條內文在分群後，議題的詮釋及常見詞彙非常不同，本文選擇常見於頭條內文的常見詞彙，挑出滿足「出現次數不少於 50 次」、「兩報紙間的差距倍數為 2 以上」者為解釋變數。這兩個標準可透過卡方檢定（Chi-square Test）詮釋，假設 O_i 及 E_i 為某個詞彙在《蘋果日報》及《自由時報》的出現次數，則這兩份報紙使用這個詞彙是否相同大致與 $\frac{(Q_i - E_i)^2}{E_i}$ 及 χ_1^2 有關。代入出現至少 50 次、差距 2 倍，統計檢定量 $\frac{(Q_i - E_i)^2}{E_i} \geq 4 \times 50 = 200$ ，即便有 5,000 個字彙滿足上述兩個條件、顯著水準 0.05，經過多重檢定（Multiple Testing）的 Bonferroni 顯著水準的調整，其檢定值仍然大於卡方檢定的拒絕門檻值 $\chi_1^2(1 - 0.05 / 5000) = \chi_1^2(0.99999) = 19.51$。滿足上述這兩個條件的《蘋果日報》及《自由時報》共有 3,316 個詞彙可作為分類變數，圖 8 左邊（或右邊）顯示經常出現在《蘋果日報》（或《自由時報》）、較少出現在《自由時報》（或《蘋果日報》）的詞彙，換言之，這些詞彙能夠區隔兩家報紙的寫作習慣，可視為重要解釋變數。其中，《蘋果日報》及《自由時報》分別偏向於社會、政治議題。

本研究以 *K*-fold 驗證分類準確率，以 2012–2017 年的六年資料為訓練集（training data），2018 年資料為測試集（testing data），檢視 random forest、SVM、decision tree 三種方法的分類效果。[5] 訓練集的準確率以 Random Forest 的準確率 100% 最佳，而 SVM 及 decision tree 的準確率分

5　隨機從 2012–2018 年抽取六年為訓練資料、留下一年為測試資料，分類的結果大致相同，準確率仍以 SVM 最高。

圖 8　滿足兩個條件的《蘋果日報》及《自由時報》頭條內文常見詞彙

資料來源：作者自行整理。

別為 97.51% 及 89.79%。測試集的準確率也相去不遠，2018 年預測準確率以 SVM 表現最好（高達 95.35%），random forest 降至 93.52%，decision tree 預測準確率則為 82.68%。亦即以具鑑別力的詞彙為解釋變數，都能有效區隔《蘋果日報》及《自由時報》的頭條新聞報導，無論是統計或機器學習模型的預測準確率至少都有八成以上，機器學習方法的分類準確度比較高。

伍、結論與討論

新聞報導如同作者的寫作風格，即使主題相同，由於切入角度、用詞鋪陳等因素，各報紙的新聞經常有明顯差異，不難從文章編排、用字遣詞中判斷來自於哪一個媒體。臺灣民眾對於發行量較多的前四大報紙《蘋果日報》、《自由時報》、《聯合報》、《中國時報》多少有些既定印象（或偏見），認為這些報紙的立場及風格不盡相同，一般會根據自己的偏好及政治傾向選擇新聞媒體。本研究主旨在於探討臺灣四大報的頭條報導，透過 EDA、驗證性資料分析等方法，基於資料驅動的原則比較用字風格，驗證質性意見及統計分析的一致性，研究發現各報的頭條標題及內文的用詞不盡相同。藉由 EDA 除了可確認哪些文字特性具有鑑別力，作為解釋變數、用於建立分類模型外，也可協助我們瞭解報紙報導風格不同之處，提

供實質詮釋及後續研究的參考，像是圖 8 可區隔《蘋果日報》及《自由時報》的常見詞彙，可用於建立主題模型（topic model）。

有別於其他數位人文的研究，本文除了考量群集、分類等方法及模型外，也著重於 EDA，示範如何應用 EDA 方法論於文字分析。過去有不少研究提到 EDA 的重要性，也大略提到 EDA 的方向（Allen, Sui, & Akbari, 2018; Martinez, Martinez, & Solka, 2017），本文更進一步結合生物學的概念，引進物種多樣性、生態學的想法描述報導風格。EDA 較偏向非監督學習，以資料驅動的原則探索資料特性，盡量避免先入為主的帶來的限制，以期挖掘出潛藏於表面以外的重要資訊。例如：本文分析顯示四大報的字詞多樣性及豐富度不同，從常見字詞及其分類等分析結果，也可看出頭條報導的議題選擇及其時間持續性，這些資訊可協助研究者以更寬廣的視野瞭解文本特性，不侷限在既有的思考框架，根據現況釐出創新的可能性。

整體而言，四大報的頭條標題用詞上確實存在差異，Jaccard 和 Yue 相似指標提供不同角度的分群結果，前者傾向於將同時期的四大報放在同一群，後者則是將四大報分成三群。其中《蘋果日報》較為特別，兩種相似指標的頭條標題都是自成一群，主要原因可能是《蘋果日報》的社會新聞用詞較多。[6] 藉由字數較多的頭條內文更能看出差異，《蘋果日報》同樣偏好社會新聞用詞，《自由時報》則多為政治題材用詞。若選擇出現較為頻繁（50 次以上）、《蘋果日報》與《自由時報》使用次數超過兩倍以上的詞彙，透過常見統計或機器學習方法區隔《蘋果日報》與《自由時報》的報導，其預測準確率都相當高，SVM 的準確率甚至高達 95% 以上。

受限於資料取得，本研究以臺灣四大媒體的電子報網站代替實體報紙，雖然比對兩個版本後發現大多數的頭版標題都相同，但限於時間，我們沒有一一檢核頭條內文報導。由於實體報與電子報的受眾族群不同，電子報偏向於常用網路的年輕族群，實體報的主要客群為年長者，各家媒體在不同媒介的用詞或許會以受眾作為導向。後續可比較實體報紙與電子報之間的差異，探討兩者的常用字詞，或是將研究素材擴大至頭版其他新聞、甚至是各報不同版面的文章報導，更完整地勾勒出每家報紙的市場定位。

6　《蘋果日報》有近 35% 的頭條標題用詞屬於「女友、少女、性侵、恐怖、雙亡、大生、曝光、冷血、落網、情侶」等社會新聞用詞，並且內文詞彙的社會網路也發現社會新聞用詞為明顯分支，頗為符合不少人認為《蘋果日報》常用腥羶色等社會題材作為頭版標題的印象。

另外，本文雖然將詞彙進行量化，但僅止於描述字詞的豐富性及相似性（或重複程度），尚未進行其他類型的文字分析（如：輿情分析）。我們覺得詞彙分群的精神類似物種的生態聚落，可引進扮演維繫生態平衡的基石物種（keystone species），找出比較特別或特殊的字詞，進一步發展可有效區別不同媒體的關鍵詞彙（亦即關鍵詞擷取，keyword extraction）。而本文也發現相似指數有時在相鄰兩月間會有劇烈震盪，例如：2015 年 7、8 月分的《自由時報》Yue 指標會有驟升的現象發生，代表這些月分的頭條標題不斷強調某些詞彙，使用頻率維持高檔，後續可再進一步探討哪些詞彙延續使用，再與當時發生的重大事件連結，加入事件相關學者的專業知識，結合數位分析、專家意見後嘗試更深層的剖析。

參考文獻

中央通訊社（2018）。107 年 11 年 29 日臺灣各報頭條速報。*中央通訊社*。取自 https://www.cna.com.tw/news/firstnews/201811295001.aspx

何立行、余清祥、鄭文惠（2014）。從文言到白話：《新青年》雜誌語言變化統計研究。*東亞觀念史集刊*，*7*，427-454。doi:10.29425/JHIEA.201412_(7).0011

林樹（1972）。*中文電腦基本用字研究*。新竹：交通大學工學院計算與控制學系。

茱蒂・亞契（Archer, J.）、馬修・賈克斯（Jockers, M. L.）（2016）。*暢銷書密碼：人工智慧帶我們重新理解小說創作*（The bestseller code: Anatomy of the blockbuster novel kindle edition）（葉妍伶譯）。臺北：雲夢千里。（原著出版年：2016）

國家教育研究院（n.d.）。編輯說明。取自 https://dict.variants.moe.edu.tw/variants/rbt/page_content.rbt?pageId=2982208

Allen, T. T., Sui, Z., & Akbari, K. (2018). Exploratory text data analysis for quality hypothesis generation. *Quality Engineering*, *30*, 701-712. doi:10.1080/08982112.2018.1481216

Boyce, G., Curran, J., & Wingate, P. (1978). *Newspaper history from the 17th century to the present day*. London, UK: Constable.

Chen, C.-H. (2002). Generalized association plots for information visualization: The applications of the convergence of iteratively formed correlation matrices. *Statistica Sinica*, *12*, 7-29.

Evans, J. R., & Lindner, C. H. (2012). *Business analytics: The next frontier for decision sciences*. Retrieved from http://www.cbpp.uaa.alaska.edu/afef/business_analytics.htm

Horn, S., & Veermans, K. (2019). Critical thinking efficacy and transfer skills defend against 'fake news' at an international school in Finland. *Journal of Research in International Education*, *18*, 23-41. doi:10.1177/1475240919830003

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096. doi:10.1126/science.aao2998

Martinez, W. L., Martinez, A. R., & Solka, J. L. (2017). *Exploratory data analysis with MATLAB* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Yue, J. C., & Clayton, M. K. (2005). A similarity measure based on species proportions. *Communications in Statistics—Theory and Methods*, *34*, 2123-2131. doi:10.1080/STA-200066418

附錄　《蘋果日報》頭條標題出現「死」、「女」、「萬」、「殺」範例

字彙	日期	標題
死	2016 年 1 月 31 日	全身 33 彈孔 警遭轟死
	2016 年 4 月 2 日	報復勸架 縱火害 6 死 惡男太狠毒
女	2016 年 3 月 25 日	翁坦承 贈女兒資金 買浩鼎股票
	2016 年 4 月 4 日	女募萬件胸罩救雛妓
萬	2016 年 4 月 17 日	5 百年強震毀熊本 41 死 10 萬人驚逃 千棟房屋倒
	2016 年 4 月 26 日	冷血醫師 補 7 千萬炒股虧損 狂賣牛奶針害 3 死
殺	2016 年 3 月 29 日	天地不容 人魔斬殺 4 歲童
	2016 年 4 月 24 日	一中香腸王涉殺女友

資料來源：作者自行整理。

數位典藏與數位人文 6 期（2020 年 10 月）
Journal of Digital Archives and Digital Humanities No. 6, pp. 69-96 (October, 2020)
DOI:10.6853/DADH.202010_(6).0003

A Study of Text Mining on Taiwan's Newspapers

Ching-Syang Jack Yue[1,*], Yu-Ting Ye[2]

Abstract

Like an author's writing style, every newspaper has its own opinion and narrative methods. In this study, our goal is to explore the reporting styles of Taiwan's four major newspapers (*Apple Daily*, *Liberty Times*, *United Daily News*, and *China Times*) and to compare their differences. We chose the headline news for analysis to prevent the influence of nuisance factors, such as differences in political positions and target audience. The newspaper headlines considered are between 2012 and 2018. The titles of headlines are considered for all four newspapers but the content of headlines is available only for *Apple Daily* and *Liberty Times*.

We first applied the methods of exploratory data analysis (EDA), such as Jaccard index and Yue index, for the word frequencies and word types to evaluate the similarities between four newspapers. In addition, we also compare the length of sentences, punctuation, and function words of four newspapers. Finally, we used statistical and machine learning methods to distinguish if the news articles are from *Apple Daily* or *Liberty Times*. The analysis results showed that there are significant differences between four newspapers. For the headline titles, the Jaccard index grouped titles by time and the Yue index grouped titles by the media (i.e., three groups). For the headline contents, we found that the articles of *Liberty Times* emphasize political terms and those of *Apple Daily* focus on social affairs and crime problems. All classification methods have high accuracy if we chose frequently used two-word phrases as the independent variables.

Keywords: writing style, Taiwan's newspaper, exploratory data analysis, similarity index, keywords

Manuscript received: March 13, 2020; Accepted: June 5, 2020
[1] Professor, Department of Statistics, National Chengchi University.
[2] Master, Department of Statistics, National Chengchi University.
[*] Email: csyue@nccu.edu.tw

Extended Abstract

Big Data has become a popular research topic since IBM (International Business Machines) proposed this terminology in 2010. It seems that all kinds of information can be digitalized and analyzed, including texts and pictures which are referred as unstructured data. Texts are particularly high-profile among all unstructured data. A considerable amount of techniques have been developed and integrated across many fields such as linguistics, computer science, and statistics. The process of applying these methods to texts for extracting useful information are generally denoted as text mining. Authorship identification is one application of text mining. A famous example is that Robert Galbraith was identified as the pen name of J. K. Rowling (author of *Harry Potter*) in 2013.

Like an author's writing style, every newspaper has its own opinion and narrative methods, and it can be easily distinguished just by reading its articles. We think that text mining can also used to identify which newspaper are the articles from. In this study, our goal is to explore the reporting styles of Taiwan's four major newspapers (*Apple Daily*, *Liberty Times*, *United Daily News* and *China Times*) and to compare their differences. We chose the headline news for analysis to prevent the influence of nuisance factors, such as differences in political positions and target audience. The study year of newspaper headlines is between 2012 and 2018. The titles of headlines are considered for all four newspapers. But the reports of headlines are available only for *Apple Daily* and *Liberty Times* due to the data availability.

The use of exploratory data analysis (EDA) is the main difference between our study and other studies. Basically, there are two parts in data analysis: EDA and confirmatory data analysis (CDA). The role of EDA is to figure out the essence of data and to develop possible research hypothesis, while the role of CDA is to examine evidence and test hypothesis. EDA was promoted by the famous statistician John W. Tukey in 1970's. He thought that more emphasis should to be placed on using data to construct research hypotheses and statistical models. Data driven is the idea behind EDA and the errors of building inappropriate statistical models can be reduced if we know the data better.

We should use the authorship identification as an example to demonstrate the importance of EDA. Surprisingly, the variables which can effectively determine that Robert Galbraith and J.K. Rowling are the same person are not verbs or nouns. Instead, it is possessive (such as "his" and "her") or preposition

(such as "of") playing the critical role. Many previous studies showed similar results. Texts are unstructured data, and there are no standard ways for defining variables. In this situation, EDA can help us explore the important features of texts. Note that, in addition to adapting the idea of EDA, Chinese words can be treated as species and thus we introduce measures of species diversity (or species richness) to distinguish the differences between the headlines of four newspapers.

The EDA tools we use in this study can be classified into first moment statistics and second moment statistics. The measurements of first moment statistics include numbers of words and two-word phrases, top 10 words and 10 two-word phrases, function words, punctuations, and average number of words in a sentence. The measurements of second moment statistics consist of species diversity indices (e.g., entropy and Simpson's index) and species similarity indices (e.g., Jaccard's index and Yue's index). We should use these first and second moment measurements to explore the attributes of headline reports from four newspapers. Other than these EDA tools, we also apply time series techniques to model the similarity indices and classification methods to divide four newspapers into different groups.

We first applied the methods of EDA. The distributions of words and two-word phrases behave quite differently for headline titles and headline contents. Because the title space is very limited (about 15 words per title), usually there are few function words or punctuations in headline titles. For example, the proportions of function words are about 1% and 3% in headline titles and headline reports, respectively. On the other hand, the proportions of punctuations are about 4% and 12% in headline titles and headline reports, respectively. Moreover, the distributions of words and two-words phrases in headline reports satisfy the Zipf's law, but those in headline titles do not satisfy the Zipf's law. In general, *Apple Daily* has the largest number of words and *China Times* has the largest number of two-word phrases.

We also calculated Jaccard's index and Yue's index for the word frequencies and word types to evaluate the four newspapers. The species similarity indices between two consecutive months suggest that there is a time trend between the words used in two consecutive months for all newspapers except for *Apple Daily*. The similarity indices can be used to classify four newspapers for headline titles, and it seems that the results are measurement dependent. First, if Yue's index on headline titles, then we can divide four newspapers into three groups: *Apple Daily*, *Liberty Times*, and *United Daily News* & *China Times*. On the other hand,

if we use Jaccard's index on headline titles, then all newspapers in the same year tend to be classified into the same group (i.e., there are seven groups).

Regarding headline reports, we have more choices on the analysis models because of more words (i.e., sample sizes). We first applied dimension reduction techniques (e.g., *t*-SNE, *t*-distributed stochastic neighbor embedding) and then used cluster analysis to separate frequently used two-word phrases in *Apple Daily* or *Liberty Times*. We found some obvious differences by examining the groupings of two-word phrases in the two newspapers. The headline reports of *Liberty Times* emphasize political terms, and those of *Apple Daily* focus on social affairs and crime problems. This result is very interesting. The headline news should be the most important news in a day, and we would expect all newspapers would choose same topics and probably similar wordings. Apparently, the analysis results do not support our conjecture. The choices of topics and wordings in headline reports are different in *Apple Daily* or *Liberty Times*.

The different wordings motivated us to propose a rule for choosing two-word phrases. We picked up those phrases whose occurrences are at least 50 times and the number of occurrences in one newspaper is at least twice as much as the other newspaper. Similar to the results of cluster analysis, the phrases chosen from *Liberty Times* and *Apple Daily* are related to politics and crime issues, respectively. Furthermore, we use these frequently used phrases as classification variables to distinguish where the articles are from, either *Liberty Times* or *Apple Daily*. The classification models considered are decision tree, random forest, and support vector machine (SVM). In addition, we applied the *k*-fold cross-validation to evaluate the classification results, by assigning 6/7 articles as training data and leaving 1/7 articles as testing data. We used accuracy of testing data to evaluate classification methods, in order to avoid over-parameterization. There is high accuracy (at least 82%) in all classification methods if we choose frequently used two-word phrases as the independent variables. Among all models, SVM performed the best (over 95%).