國立政治大學商學院統計學系

碩士學位論文

運用文字探勘分析人民日報的風格變遷

A Study of Writing Style of *The People's Daily*

指導教授：陳麗霞 博士

余清祥 博士

研究生：陳庭偉 撰

中華民國 110 年 07 月

DOI:10.6814/NCCU202101184

摘要

大數據發展促使各類型資料的數位化，文字探勘更是當中典範，在不同領域都可看到相關應用，寫作風格是常見議題之一。然而，文章風格容易受到議題的影響，即便是同一作者或文本，文字使用可能因為時空背景等因素而產生差異。以中國共產黨機關報刊《人民日報》為例，內容及題材不僅呈現當代特色，也會顧及官方立場與目的，該報的特色變化可反映中共建國至今的政治及社會變遷。因此本文以《人民日報》的風格變化為研究目標，藉由比較各年度的遣詞用字差異，透過統計方法及分群劃分不同時期；另外，本文也運用多種關鍵詞偵測指標，篩選各時期的代表詞作為分類的解釋變數，希望能夠兼顧準確率、運算速度、解釋性。

本文以《人民日報》1949～2019 年頭版報導為研究素材，因為頭版內容大多涉及全國性及國際等重大事務，避免某些地方性事務造成用詞的異質性。本文先考量探索性資料分析，包括字、詞以及字詞的 Jaccard、Yue 相似指標，挖掘《人民日報》的文字基本特性；接著套用群集分析近年中國分成數個時期，再與專家的分期結果比較。研究發現：透過雙字詞更能看出各時期的差異，如果以雙字詞或相似指標進行分群，《人民日報》可分為四個時期（或可命名為「建國」、「文化革命」、「改革開放」、「現代化」），不同分群方法的分析結果相當一致，而各時期的用詞風格有明顯差異。另外，分類解釋變數的挑選以本文提出的代表詞偵測指標最佳，無論是準確率、運算速度、解釋性三者的結果，都優於卡方指標或維度縮減等方法。

關鍵詞：寫作風格、風格變遷、群集分析、關鍵詞、挑選變數

DOI:10.6814/NCCU202101184

Abstract

Big data enhances the quantitative analysis in all kinds of data and text mining is one of them. Identifying authors' writing style is one popular topic of text mining. However, the writing style can be affected by, for example, the theme and language of articles. Take *the People's Daily*, official newspaper of the Central Committee of the Chinese Communist Party, as an example. The Chinese Communist Party attaches great importance to the *People's Daily*, and has given strong guidance to the work of the *People's Daily* in all periods of revolution, construction and reform. In order words, through the text analysis of *the People's Daily*, we may find the changes of political/social environment of Chinese Communist Party, and we want to know if it is possible to differentiate different periods of China (1949~2019) via text analysis of the articles in *the People's Daily*.

We first conduct exploratory data analysis, including characters, words, Jaccard and Yue's Index. Then we use cluster analysis to divide modern China into several periods, and then compare with the results of experts' research. The research found that the differences between the periods can be more clearly seen through the two-character words. If the two-character words or similar indicators are used to cluster, the *People's Daily* can be divided into four periods. Besides, we use multiple keyword indicators to select representative words in each period, and we select these representative words as explanatory variables to classify. Whether in terms of accuracy, calculation speed, or explanatory performance, it is better than chi-square indicators or dimensionality reduction methods.

Keywords: Writing Style, Style Change, Cluster Analysis, Keyword, Variable Selection

DOI:10.6814/NCCU202101184

目次

DOI:10.6814/NCCU202101184

DOI:10.6814/NCCU202101184

表次

DOI:10.6814/NCCU202101184

圖次

DOI:10.6814/NCCU202101184

DOI:10.6814/NCCU202101184

DOI:10.6814/NCCU202101184

DOI:10.6814/NCCU202101184

第一章　緒論

第一節　研究動機

由於大數據的推波助瀾，幾乎所有專業領域都引進了量化分析，藉由電腦科技彌補人力的不足。以往資料分析大多侷限在數字類型的紀錄，現在則可應用於文字、影像、聲音等非數字型態的資料，這種沒有固定格式的資料又稱為非結構資料（Unstructured Data），根據 IDC 的預測，到 2025 年時全球八成的數據將是非結構資料[1]。非結構資料蘊含的資訊通常比較豐富，以文字資料為例，透過用字遣詞等文字風格可以大略猜出作者，像是《哈利波特》作者 JK 羅琳曾經以筆名羅勃・蓋布瑞斯，隱藏自己身分撰寫偵探小說，但很快地就因為用字習慣被識破身份。[2]

非結構化資料中發展最早者應該是文字，文字分析（又稱為文字探勘，Text Mining）可追溯自 1980 年代（或更早），早期多以用於文學，包括寫作風格的變化、判斷是否為同一作者等。近年文字分析的應用範圍更廣，Google 以搜尋引擎預測流行性感冒的盛行時間（Google Flu Trends）[3]，就是廣為人知的範例，最近兩年也有學者整合 COVID-19 相關文獻，藉此擷取有用的防疫及治療的重要資訊。寫作風格及文本分類在目前仍是主要應用之一，協助人文學者整理及判讀大量文字資訊，從更寬廣的面向剖析問題。例如：《紅樓夢》是否全由曹雪芹所寫，還是前八十回是由曹雪芹所寫，而後四十回由旁人續寫，這類問題在過去常常因為解讀角度不同而有不一致的結果[4]，而文字探勘則是套用量化技術，從不同角度解讀和分析。

---

[1] 參考 Solutions Review，作者為 Timothy King，在 2019 年提出。
[2] 參考《The Bestseller Code》，作者為 Jodie Archer and Matthew L. Jockers，本文在此引述該書第四章的內容。
[3] 參考 FluBreaks: early epidemic detection from Google flu trends，作者為 Pervaiz, F.等人。
[4] 參考《紅樓夢考證》，作者為胡適，提出前八十回和後四十回的作者並非同一人。

DOI:10.6814/NCCU202101184

在進行文字分析時，中文和英文在前處理上有很大的區別，英文是以字根為基本分析單位，但古代和現代的中文則有明顯不同。中文過去的書面文體是文言文，現代書寫大多屬於白話文為主，其中白話文通常是以雙／多字詞為基本單位，因此在對分析現代中文文本時需要先斷詞，同時配合標點符號、虛字等使用習慣，作為區隔作者寫作風格的判斷依據。

但中文分析在選擇變數時仍無定論，有人認為選擇常見字詞，也有人認為也要加入標點符號、虛詞等資訊，也有人認為只要考量常見／關鍵字詞。因為無法確定最合適的變數，目前許多研究都是直接代入全部字詞當作，但這樣的方法不但運算時間會花費很長，同時也無法瞭解哪些變數與研究目標的關係最密切。有鑑於此，本文比較幾種變數選取方式，包括倍數指標、卡方指標等方法，林志軒（2020）提出相關的概念，以維度縮減方法只挑選部分變數，希冀能找出以較少變數、而且不犧牲解釋能力的方法。

另外，我們認為文字風格能夠反映個人情緒、國家發展的軌跡，藉由比對同一文本的長期資料，大略可以發現風格變化與當事人／國家事件間的契合痕跡。本研究選擇中國共產黨的機關報《人民日報》頭版報導作為研究對象，觀察中華人民共和國 1940 年代末期建國至今的文字風格，並以上述字詞變數代入常見的群集分群（Cluster Analysis）模型，將過去 70 餘年《人民日報》報導分成數群。接著再將分群結果與近年重大事件比對，探究中國不同時期的變化能否與《人民日報》群集結果一致，亦即驗證文字風格與中國發展兩者同步。

DOI:10.6814/NCCU202101184

第二節　研究目的

本文以寫作風格分類為研究目標，探討不同時期《人民日報》頭版報導的風格變遷，是否與中國的發展息息相關。為了確定兩者是否一致，本文先以群集分析將《人民日報》分成數個時期，比較單字或（雙字）詞何者更適合作為變數，以及不同群集方法的差異。此外，本文也提出維度縮減的方法，透過關鍵詞偵測指標等篩選重要變數，選出各時期的代表詞作為分類解釋變數，以此和代入所有字詞當變數的結果比較，同時進一步與過去的卡方指標、倍數指標等方法去比較，希望能夠兼顧準確率、運算速度、解釋性。

本文選用《人民日報》頭版報導為研究對象，主要因為頭版新聞大多為全國性及國際事務，比較不像地方新聞版面，容易因為報導地區不同而產生主題、用詞的巨大變化，且因為頭版新聞具有代表性，因此更能反映出中國不同時期的背景特色。另外，本文也將比較中國不同時期的用字風格，探討不同事件對《人民日報》帶來的影響。本文從旁觀者的角度進行風格分析，寫作風格可以代表一個人或一個個體的特性，從各年度人民日報頭版報導的變化觀察中國這些年的變化，本文通過用字遣詞的變化區分中國的不同時期，與專家學者認定的時期比較，研究是否文字風格可以反映中國的轉變。

本文在第二章整理過去的相關研究與文獻和介紹使用的資料，第三章介紹本文使用的研究方法，有相似度指標、用於對年分的分群方法、代表詞偵測指標、用來比較的降維方法還有運用在分類的統計學習模型和機器學習模型。第四章是對人民日報頭版內文進行分析，從探索性資料分析開始，對年代分群，分群結束後用關鍵詞偵測指標篩選出各時期的代表詞，最後運用篩選出的代表詞當作變數進行文本分類，同時與其他降維方法，有倍數指標、卡方指標等等進行比較。第五章整理和總結本文的研究發現和結果，同時提出未來的研究方向和建議。

DOI:10.6814/NCCU202101184

第二章　文獻探討

由於中文及英文的特性及表達方式差異很大，本章第一節先整理中文和英文分析的相關研究，介紹有哪些地方需要特別注意，中文分析與英文分析有非常大的不同。另外，變數選取在中文分析往往是很重要的一步，說明常見的選取標準，包括 TF（Term Frequency，譯為詞頻）及 IDF（Inverse Document Frequency，譯為逆向文件頻率）等等，以及本文提出的方法。

第一節　文獻回顧

過去關於寫作風格分析，無論是在中文還是英文都有許多研究，但兩種語言有很大的差異，陳鳳芝（2003）提出中英文兩種語言的寫作風格在思維方式、篇章結構、語言特點等方面都有許多不同。除了語言的差異之外，根據特定時間或事件也有不同的分析方法，Puglisi（2006）針對 1946～1997 年的《紐約時報》新聞，分析現任總統和美國國會的議題；王宇（2012）曾探討食品安全的報導，資料是以《人民日報》近 10 年的報導為例，從報導主題、消息來源、報導領域等方面進行分析，去了解《人民日報》食品安全議題的形成和變化。

而在文本分類的研究中，變數的選取是很關鍵的一步，孫曉明等人（2001）提出了使用文本中的虛詞頻率分佈作為解釋變數，研究結果顯示使用虛詞頻率分佈作為解釋變數是有效的，而張運良等人（2009）曾提出運用句類特徵當作解釋變數，認為不同的作家在詞彙、句型、修辭手法等方面都擁有自己的習慣與特色，因此，他們運用了句類特徵來進行文本分類。Zhai 等人（2018）提出了運用卡方統計進行變數選擇後，再進行文本分類，作者認為變數選擇是文本分類中關鍵的一步，除了會影響分類的準確度，同時也可以提高解釋性。因此，變數選擇會對分類的結果有很大影響。

DOI:10.6814/NCCU202101184

過去在變數選取的部分，也有使用單一指標去找出關鍵字詞或信息，姚興山（2009）運用詞頻統計的方法對文本進行分類，於韜等人（2018）提出使用 TF-IDF 提取文本的信息，TF-IDF 相較於詞頻來說，因為 IDF 是用來處理常用字詞的問題，可以避免選到常用字詞當作關鍵字詞，而夏天（2013）運用對字詞位置加權的 TextRank 方法去抽取關鍵詞，同時考慮了字詞的位置和 TextRank，這些方法都獲得了不錯的結果。

綜合以上研究可以發現，過去在寫作風格分析的研究中，無論是中文還是英文，大多都是針對特定時間或者特定主題，而本文使用《人民日報》1946～2019 年的頭版報導，研究在不同時空背景下的文本風格變化，從單／雙字詞、標點符號、虛字這三個角度出發，接著對雙字詞和相似度指標進行多種分群方法，把《人民日報》分成不同時期。而過去在文本分類的研究中，變數的選取中，有研究選取的變數是選取虛字、句類特徵等等，而也有研究選取字詞變數，主要是考慮卡方統計、詞頻等方法。本文認為字詞較能反映出文本的背景特色，但過去研究的考慮較為單一，因此本文提出的方法是運用詞頻、TF-IDF、TextRank 三個指標選出代表詞當作變數，與過去的卡方指標、倍數指標、PCA 這些降維方法去比較。

第二節　資料介紹

本研究使用的資料是《人民日報》1949～2019 年頭版報導，《人民日報》為中國共產黨中央委員會的機關報，是在 1949 年 8 月 1 日從華北中央局機關報升格為中國共產黨中央委員會的機關報，同時《人民日報》也是中國第一大報，有廣泛的影響力，而《人民日報》頭版報導主要是報導全國及國際事務，除了頭版報導之外，《人民日報》其它版有周刊、專版和理論版等等。頭版報導具有代表性且多為重大事件，因此相比其它版更能反映出中國在不同時空背景下的特色。

DOI:10.6814/NCCU202101184

74 年來，人民日報積極宣傳黨的理論和路線方針政策，積極宣傳中央重大決策部署，及時傳播國內外各領域信息。因此，《人民日報》可以反映出中國自 1940 年代至今的社會變遷。

黃秋林等人（2009）研究 1978～2007 年《人民日報》兩會社論，探討經歷改革之後的中國隨著形勢的發展，《人民日報》也改變各類概念隱喻的構成和使用形態，運用關鍵詞定位檢索和詞頻統計找出關鍵詞後，排除非隱喻用法的詞，發現在旅行隱喻、建築隱喻和植物隱喻的使用沒有發生太大的變化，而在戰爭隱喻、航海隱喻和家庭隱喻的使用發生了變化。《人民日報》在改革開放後的這段時間，在某些領域中，隱喻的使用確實是有改變。

徐超（2017）認為《人民日報》作為中國共產黨中央委員會的機關報，了解《人民日報》的社論詞彙就可以全面的了解中國的社會歷史。透過對《人民日報》的社論詞彙進行統計分析，計算篇數和高頻詞等等，並且結合當時的時代背景，了解近代中國的變化以及對社會的影響。《人民日報》在中國的不同時期皆有指標性的意義，許多重大事件的發展皆與《人民日報》息息相關，例如：文化革命、改革開放等等。

中國中央黨史和文獻研究院院長曲青山（2021）所著的《中國共產黨百年輝煌》一書中將中國共產黨百年歷史劃分為不同時期：從 1949 年 10 月至 1978 年 12 月黨的十一屆三中全會召開，是社會主義革命和建設時期；從 1978 年 12 月至 2002 年黨的十六大前，是改革開放時期；2002 年黨的十六大後，是社會主義現代化建設新時期，中國在這段時間經歷了建國初期、文化大革命、改革開放、現代化等重大事件。綜上所述，本研究對《人民日報》頭版報導進行文字探勘，希望從各年度《人民日報》頭版報導的文章風格變遷了解近代中國在不同時期下的變化。

DOI:10.6814/NCCU202101184

第三章　研究方法

本章節將簡介本研究所會使用的方法，其中包括 Jieba 斷詞、探索性資料分析、不均度指標 Entropy、兩大相似度指標 Jaccard、Yue index。再來是介紹多種分群方法，有 Hierarchical Clustering、K-means、K-medoid、PCA、T-sne。而代表詞偵測使用了詞頻、TF-IDF、TextRank 這三個指標來進行偵測，因此會介紹這三個指標。最後介紹多種分類方法，有 Logistic Regression、Random Forest、XGBoost 這三大模型。

第一節　Jieba 斷詞

Jieba 斷詞主要是結合規則斷詞和統計斷詞，規則斷詞主要是透過事先建好的詞典對句子進行斷詞，會將句子中的每個字與詞典中的詞去進行匹配，匹配成功則斷詞。統計斷詞則是主要看相連的字在文本中出現的次數越多，就會推斷這相連的字很有可能是一個詞，因此若相連的字出現頻率高於一個臨界值時，就會進行斷詞。因此規則斷詞會基於前綴詞典進行詞圖掃描，生成所有成詞情況的有向無環圖，前綴詞典是指詞典中的詞按照前綴包含的順序排列。例如詞典中出現了「台」，之後以「台」開頭的詞都會出現在這一部分，例如「台北」，進而會出現「台北市」，從而形成一種層級包含結構，最後再透過動態規劃查找最大機率路徑。而統計斷詞則是應用隱藏式馬可夫模型(Hidden Markov Model, HMM)去找出未記載於詞典的詞。

第二節　探索性資料分析

探索式資料分析（Exploratory Data Analysis, EDA）最早是在西元 1977 年由 Tukey 所提出，主要概念是透過敘述性統計，統計視覺化等方法去了解資料的特

DOI:10.6814/NCCU202101184

性，從各種面向去了解資料。探索式資料分析不僅可以提早發現資料品質問題，找出資料的離群值等，也可以發掘出重要變數。探索性資料分析也可以從資料中去發掘未發現的議題，良好的探索式資料分析可以在之後深入分析資料時提供更明確的方向。

當拿到一筆不熟悉的資料時，我們可以運用探索性資料分析而了解資料，如果在不了解資料的情況下就進行建立模型的步驟，即使預測的準確度很高，但卻不知道為什麼預測結果良好。在分析資料時，很重要的就是知其然且知其所以然，這樣才能使分析結果具有價值。

第三節　群集分析

一、Hierarchical Clustering

階層式分群法（Hierarchical Clustering）是 Hastie et al.在 2002 年提出的，透過一種階層架構的方式，將資料層層反覆地進行分裂或聚合，以產生最後的樹狀結構，常見的方式有兩種，一種是聚合，一種是分裂。本文使用的方式是聚合，該方法會由樹狀結構的底部開始層層聚合，一開始會將每一筆資料當作一個群集(cluster)。首先各筆資料自成一個群集，再找出所有群集間距離最接近的兩個群集，合併兩個群集成為一個新的群集，經過(n-1)次的合併後，便能將所有資料合併成一個群集。研究者可以依據最後的樹狀結構來決定合適的分群數量。常見的群集之間距離計算方式有單一連結法(Single Linkage)、完全連結法(Complete Linkage)、平均連接法(Average Linkage)、中心法(Centroid Method)、沃德法(Ward's Method)。本研究在內文的分析中採用沃德法(Ward's Method)，其群集間的距離定義為在將兩群集合併後，各點到合併後的群集中心的距離平方和，其群集間距離定義如下:

DOI:10.6814/NCCU202101184

$$d(C_i, C_j) = \sum_{a \in C_i \cup C_j} \|a - \mu C_i \cup C_j\|$$

二、K-means

K-means 是 Lloyd 在 1957 年提出的，原本是訊號處理中的一種向量量化方法，現在流行於資料探勘領域。K-means 的目的是把 n 筆資料分類到 k 個群集中，使得每個資料都會屬於離它最近的群集中心對應的群集，則群集內的資料會有最大的相似性，不同群集間會有最大的相異性，但是 K-means 易受異常值或極端值的影響且不能處理類別變數資料。K-means 的作法是先決定群集數 k 群，並且隨機選 k 個點做群集中心。接者將每筆資料分類到距離其最近的群集中心，衡量方式為歐式距離，再以各群集的平均值作為各群集中心點，重複此過程直到群集內之元素和群集中心不再改變為止。

三、K-medoid

K-medoid 是一種比 K-means 更加穩健的方法，是 Kaufman 和 Rousseeuw 在 1987 年提出的。K-medoid 的目的和 K-means 一樣，但作法卻有一些差別，在使用 K-medoid 時，先決定群集數為 k 群，並且隨機由資料點中選 k 個點做群集中心。接者將每筆資料分類到最接近的群集中心所對應的群，衡量方式是曼哈頓距離，且以各群集中絕對誤差最小的樣本點當作群集中心，因此較不易受離群值所影響，重複這個過程直到群集元素和群集中心都不變為止。

DOI:10.6814/NCCU202101184

四、PCA

PCA（Principal Component Analysis）由 Pearson 在 1901 年發明，是一種線性降維的方式，目的是希望在特徵空間找到投影軸，資料經投影後可以得到這組資料的最大變異量。PCA 的作法是將 n 維特徵投影到 k 維特徵向量上，這 k 維是全新的正交特徵也被稱為主成分。首先先將資料標準化，接著建立共變異數矩陣，求出特徵值與特徵向量，將特徵值由大到小排列後，選取前 k 個特徵值和特徵向量，最後將原本的資料投影到特徵向量上。

五、t-SNE

t-SNE（t-distributed Stochastic Neighbor Embedding）是一種非線性降維方法，由 Maaten 和 Hinton 於 2008 年提出，是由 SNE(Stochastic Neighbor Embedding）發展而來。本研究會先介紹 SNE 的原理，在進一步推廣到 t-SNE。

SNE 先計算高維中樣本點之間的歐式距離（Euclidean distances），將其轉換為表示相似度的條件機率$p_{i|j}$，當常態分配中心為x_i時，x_i與x_j為鄰居的機率:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/(2\sigma_i^2)\right)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/(2\sigma_i^2))}$$

其中σ_i^2是以x_i為中心之常態分配的變異數，且$p_{x|x} = 0$。

同理，SNE 在低維度同樣建構一個常態分配的條件機率$q_{j|i}$，x_i和x_j分別對應到y_i和y_j，標準差設為$1/\sqrt{2}$:

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq i}\exp(-\|y_i - y_k\|^2)}$$

同理，其中$q_{x|x} = 0$。

DOI:10.6814/NCCU202101184

在歐式距離轉成條件機率建構常態分配時，需要選擇每一個x_i對應的σ_i，因為資料點的密度通常不平均，所以 SNE 以二分搜尋法找尋P_i，P_i要符合預先設定的困惑度（Perplexity）：

$$\mathrm{Perp}(P_i) = 2^{H(P_i)}$$

$H(P_i)$是P_i以 bit 度量的夏農熵（Shannon entropy）：

$$H(P_i) = -\sum_j P_{j|i} \log_2 P_{j|i}$$

通常將困惑度設定為 5 至 50，可將其視為平滑過後的有效鄰居數。

目標是在高維和低維中的條件機率分佈盡可能的相似$p_{j|i} \sim q_{j|i}$，因此 SNE 要最佳化所有樣本點間的 KL 散度（Kullback–Leibler Divergence, KLD）總和:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log\left(\frac{p_{j|i}}{q_{j|i}}\right)$$

將上面的式子作為成本函數（Cost Function），可以利用隨機梯度下降法（Stochastic Gradient Descent, SGD）求解:

$$\frac{\delta C}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

KL 散度為兩個機率分布差別的非對稱性度量，當作 SNE 的價值函數時有保留局部特徵的特性，也就是說在高維度中接近的兩個樣本在低維度中也要接近，算出來的 cost 較大，但在高維度中較遠的樣本並不一定也要較遠，算出來的 cost 較小。

為了避免陷入局部最佳解和加速最佳化過程，使用梯度下降法的時候會加入一個會衰減的動量項:

DOI:10.6814/NCCU202101184

$$\Upsilon^{(t)} = \Upsilon^{(t-1)} + \eta\frac{\delta C}{\delta \Upsilon} + \alpha(t)\left(\Upsilon^{(t-1)} - \Upsilon^{(t-2)}\right)$$

其中$\Upsilon^{(t)}$為第 t 次迭代時的解，η為學習率，α(t)為第 t 次迭代時的動量。

因為維度災難（Curse of Dimensionality）高維資料中的距離關係不能完整在低維空間中保留，各個分群會有聚集在一起無法區分的擁擠問題（Crowding Problem），為了解決擁擠問題，t-SNE 做了兩項改進來達到更好的降維效果:

第一項改進是使用 Symmetric SNE，將映射方式更改成聯合機率分佈以簡化梯度公式，實驗表明 Symmetric SNE 和 SNE 有一樣好的結果。因此公式變為：

$$p_{j|i} = \frac{\exp\left(-\|X_i - X_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp(-\|X_i - X_k\|^2/2\sigma_i^2)}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

第二項改進是低維空間時改用 t 分佈，為了解決擁擠問題，在 t-SNE 中低維空間改用 t 分佈而非原本的常態分配，以 t 分佈表達的好處是 t 分佈比常態分配更偏重長尾，使得高維度下中近的距離的樣本點在映射後能夠有一個較大的距離，且 t 分佈受異常值影響更小，在樣本數較少時仍可以模擬分佈情形而不受雜訊影響。因此使用 t 分布時聯合機率分佈 q_{ij} 公式變為：

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

第四節　TF-IDF

TF（Term Frequency）指的是字詞在文本中出現的頻率。這個數字是對字詞出現次數的歸一化，以避免同一個詞語在長文本裡可能會比在短文本裡有更高的字詞出現次數，而不管該字詞是否重要。TF 的公式為:

DOI:10.6814/NCCU202101184

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$n_{i,j}$是字詞 i 在文本d_j中的出現次數，而分母則是在文本d_j中所有字詞的出現次數之和。

IDF(Inverse Document Frequency)指的是逆向文件頻率，用來處理常用字詞的問題，由總文本數目除以包含該字詞的文本數加上一，再將獲得的比值取以十為底的對數。包含該字詞的文本數加上一的原因是為了避免分母為零，也就是所有文本都不包含該詞。IDF 的公式為:

$$\mathrm{idf}_i = log\frac{|D|}{|\{j: t_i \in d_j\}| + 1}$$

|D|是語料庫中的文本總數，分母是包含字詞t_i的文本數目，也就是$n_{ij} \neq 0$的文本數），如果字詞不在文本中，分母會為零，因此通常會加上一。

TF-IDF(Term Frequency-Inverse Document Frequency）就是 TF 乘上 IDF，可以看出 TF-IDF 和字詞在文本中的出現次數成正比，與字詞在整個語料庫中出現次數成反比。

第五節　TextRank

TextRank 的主要原理是受到 Google 公司發展的 PageRank 演算法所啟發，而後者原先是用在計算網頁的相關性與重要性，作為對其搜尋引擎的搜尋結果中的網頁進行排序的依據。TextRank 的概念是單一篇文本中字詞與字詞之間的關聯性會以網路做呈現。每一個字詞為一個節點(Vertices)，而字詞與字詞之間會相連，連接的權重則為字詞之間的相似程度。

TextRank 的演算法為:

DOI:10.6814/NCCU202101184

$$h(V_i) = (1-d) + d \times \sum_{V_j} \frac{\omega_{ji}}{\sum_{V_k \in Out(V_j)} \omega_{jk}} h(V_j)$$

V_i為節點，$h(V_i)$為某個節點的 TextRank 分數，d 為阻尼係數，為一個定值且介於 0 到 1 之間。

由以上的演算法可以發現 TextRank 分數主要可以分成兩個部分，在 d 已經被確定的情況下，第一部分是很單純的常數(1-d)，而第二部分則為從其他節點所分配到的重要程度。第二部分的意義為針對每一個除節點 i 外的節點 j ，計算節點 i 與此節點 j 的權重所占比例，再乘上節點 j 的 TextRank 分數。代表其他節點所貢獻的重要程度，節點 j 本身越重要或是彼此之間的連結權重比例越高，提供的數值就越高。實際計算上，先初始化每個節點的 TextRank 分數，並依照演算法迭代更新節點的 TextRank 分數直到$h(V_i)$收斂。

第六節　變數選擇方法

一、代表詞指標

在白話文中，多以雙字詞和多字詞來敘述，因此會以雙／多字詞進行分析，而本研究以雙字詞進行分析，並經由 TF、TF-IDF、TextRank 三者對雙字詞排序，每個時期都取出各指標的前 n 名雙字詞之交集，而 n 的值會依變數數量和模型分類準確度而決定。因此，本研究偵測每一時期代表詞的方法，乃同時考量三個指標，每個時期挑選出的代表詞再取聯集當作解釋變數。

二、倍數指標

倍數指標的概念是選取不同的文本與文本之間詞頻有差距的字詞來當作解釋變數。本文運用倍數指標的方法是該字詞的詞頻必須大於 1000 次且該字詞在

DOI:10.6814/NCCU202101184

不同的文本與文本之間詞頻的比值大於 n 或小於 $\frac{1}{n}$ ，會考慮變數數量和模型分類準確度決定 n 的值。所以運用倍數指標篩選解釋變數可以去除常用詞且可以篩選出各文本之間有差異的字詞。但倍數指標的問題在於很重要的字詞雖在不同的文本之間皆出現卻未被選到，而使得解釋性下降。

三、卡方指標

先將資料進行標準化，再使用卡方指標，公式如下圖:

$$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$$

O_i和E_i代表字詞 i 出現在 A 文本和 B 文本的數量。

在顯著水準 0.05 的情況計算出在自由度為 1 的卡方分配下的 p-value，給予不同的門檻值進行篩選。在只有兩個文本時，則以上述公式計算，而如果有多個文本時，則兩兩文本為一組計算，只要於任何一組通過門檻值，就選入該字詞。

第七節　分類模型

一、Logistic Regression

Logistic Regression 目的是在處理分類問題，通常目標變數 Y 是兩類別之變數。Logistic Regression 的想法是希望能將資料點透過迴歸線分隔開，以達成分類的效果。Logistic Regression 的迴歸式是對勝算值(Odds)取自然對數:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

固定其它解釋變數下，當 X_i 增加時，若$\Delta Odds > 1$，則表示當 X_i 增加時，事件

DOI:10.6814/NCCU202101184

Y 發生的勝算會提高，若$\Delta Odds < 1$，則表示當 X_i 增加時，事件 Y 發生的勝算會降低。

二、Random Forest

在機器學習的模型中，Random Forest 是一種監督式學習的分類模型，此術語是由貝爾實驗室的何天琴在 1995 年提出的隨機決策森林（Random Decision Forests）而來的。隨機森林是由 Leo Breiman 在 2001 年提出。其概念是結合多個決策樹，並且加入隨機選取的訓練資料。也就是由多顆不同的決策樹所組成的一個學習器，這種方法稱為 Ensemble Method，經由結合多個弱學習器以建構出一個強學習器，成為更強的模型。當處理單一數據集時，為形成多個具差異性的決策樹以進行 Ensemble Method，就需要產生不同的數據集，才能產生多個具有差異性的決策樹。因此，Random Forest 使用了 Bagging(Bootstrap Aggregation)方法，而 Bootstrap 的作法是從原有的數據集以抽出放回的方法重覆抽樣而產生新的數據集，所以使用 Bootstrap 可以從原始數據集中生出多組數據集。這種方法會從訓練資料集中取出 n 個樣本，再從這 n 個樣本訓練出 n 個分類器，也就是 n 個決策樹。每次取出的 n 個樣本都會再放回原有的訓練資料集，因此這些抽出的數據集之間會有部份的資料重複，但是由於每個決策樹的樣本還是不同，因此訓練出來的決策樹之間是有差異性的，而最後用投票的方式得到結果。

Random Forest 的步驟做法為 ：

1. 從數據集中隨機選取 n 個資料，取完後放回。

2. 從選取的 n 個資料中，訓練出決策樹。對每一個節點隨機選取 k 個特徵和使用特徵分割該節點

3. 重複上述步驟 1~步驟 2

DOI:10.6814/NCCU202101184

4. 結合所有決策樹的預測，以多數決的方式，來決定分類結果。

三、XGBoost

在機器學習的模型中，XGBoost 是一種監督式學習的模型，可以應用在分類和迴歸上，使用的方法也是 Ensemble Method，但是它使用的 Boosting 方法。Boosting 與 Bagging 類似，差異為提高舊分類器分類的錯誤資料權重，加強訓練先前預測錯誤的地方，而訓練出新的分類器，如此新的分類器可學習到錯誤分類資料的特性，進而提升分類結果。XGBoost 在建構一顆樹時所進行的特徵劃分，採用的方法為 exact greedy algorithm 和 approximate algorithm，在建構樹的每一層的過程中，來優化目標。

DOI:10.6814/NCCU202101184

第四章　頭版報導內文分析

本研究對高頻詞和相似度指標進行多種分群方法，透過投票的概念去決定分群結果，這樣同時考慮了用詞種類和用詞頻率，而使用多種分群方法讓結果更為穩健，將分群結果與專家學者定義的時期去比較。因為分群是非監督式學習，是從資料本身出發，因此並沒有主觀意識的影響。接者運用多種關鍵詞指標去偵測各大時期的代表詞，把偵測出來的代表詞當作分類變數，與將所有雙字詞當作變數去比較，建立文本-詞頻矩陣（Document Term Matrix，DTM）後，因為《人民日報》是在 1949 年 8 月 1 日才從華北中央局機關報升格為中共中央機關報，可能因為這個原因，文章風格與其它年分很不一樣，進行文章風格分類時容易分類錯誤，因此拿掉這段時間進行文章風格分類，依照年分將資料隨機抽取 56 年當作訓練集，15 年當作測試集，並且進行模擬 500 次的分類，計算平均準確率。同時，進一步去比較其它維度縮減方法：PCA、倍數指標、卡方指標。最後進一步將代表詞指標推展到月份分類，圖 4-1 為本文的分析流程圖。

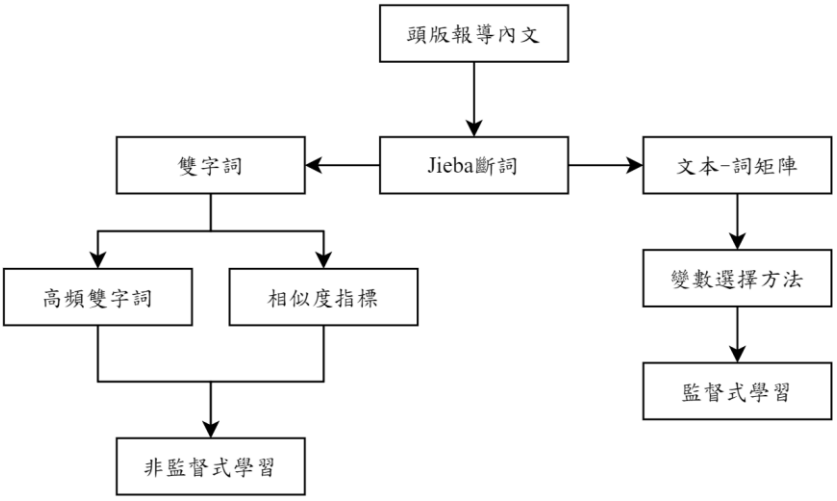

圖 4-1、頭版內文分析流程圖

DOI:10.6814/NCCU202101184

第一節　探索性資料分析

本研究先將文本去除數字、英文字母、標點符號，然後進行 Jieba 斷詞，計算了篇數、總字數、相異字個數、相異雙字詞數、相異詞個數、最常見的前 500 字字數比例、前 500 雙字詞比例和前 500 所有字詞比例的平均數、中位數、標準差。

表 4-1、人民日報頭版字彙、雙字詞、多字詞統計（1946～2019 年）

	篇數	總字數	相異字個數	相異雙字詞數	相異詞個數	前 500 字字數比例	前 500 雙字詞比例	前 500 所有字詞比例
平均數	3094	2589980	4194	33898	60058	82.89%	54.86%	45.80%
中位數	3190	2645217	4261	34781	62219	82.52%	53.93%	44.38%
標準差	1003	494706	246	6876	11893	1.32%	4.07%	4.29%

分析各年的篇數，可發現在 1950 年中期後篇數開始急速下降，直到 1970 年代中期後才開始逐漸上升，到了 1990 年代後篇數達到了最高，之後又開始逐漸下降至今。

DOI:10.6814/NCCU202101184

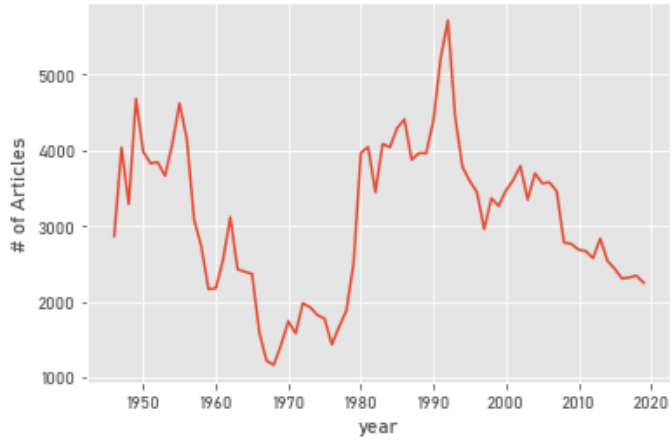

圖 4-2、人民日報頭版篇數數量（1946～2019 年）

分析各年的總字數，可發現跟篇數的趨勢有點相似，但最高點出現在 1950 年代，之後開始急速下降，也是到 1970 年代中期後才開始逐漸上升。跟篇數較不一樣的地方在於，篇數在 2010 年後是不斷的下降，但總字數在 2010 年後有個急遽的下降，但又突然有了回升。儘管篇數下降，但總字數卻有了上升，也就是說，敘述的主題與事件變少，但敘述的篇幅卻開始變長了。

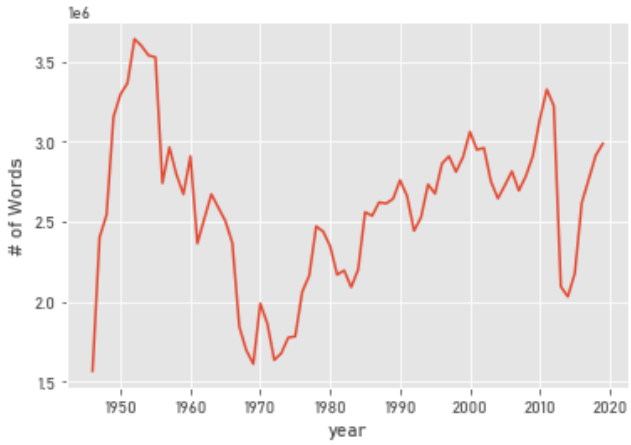

圖 4-3、人民日報頭版總字數數量（1946～2019 年）

DOI:10.6814/NCCU202101184

因為篇數多通常字數也會較多，而篇數少通常字數也會較少。因此去分析各年的總字數除以篇數，可以看出 1946 到 1970 中期的趨勢跟 1990 年代以後的趨勢有相似之處，而 1980 年代到 1990 年則都處於低谷期。

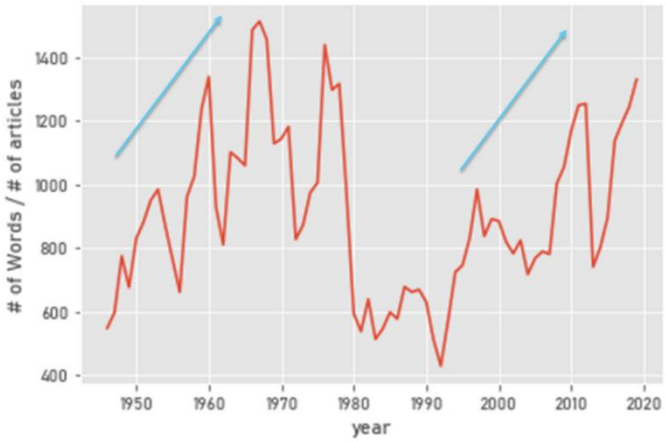

圖 4-4、人民日報頭版總字數除以篇數（1946～2019 年）

分析各年的相異字數量跟相異詞數量，可看出兩者的趨勢極為相似，且都在 1960 年代有急遽的下降，最低點都在 1960 年代中後期，之後又有迅速的回升。同時也發現在人民日報創刊初期，有較高的相異字數量跟相異詞數量。

DOI:10.6814/NCCU202101184

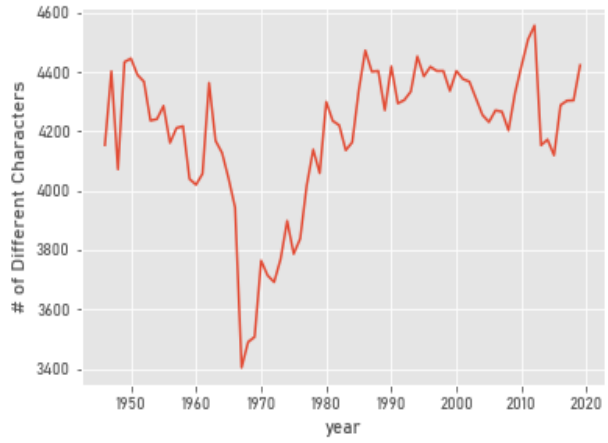

圖 4-5、人民日報頭版相異字個數（1946～2019 年）

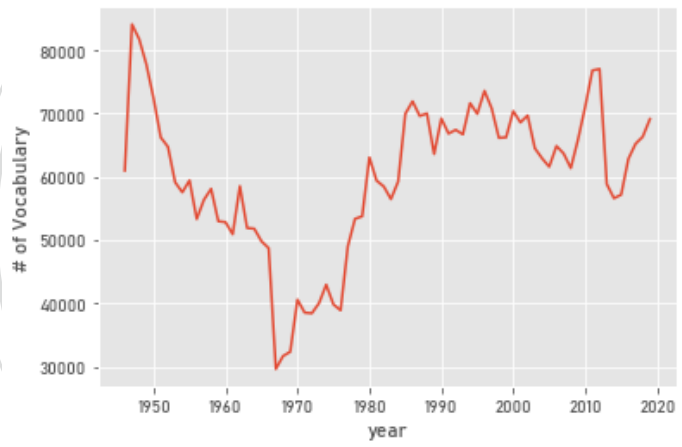

圖 4-6、人民日報頭版相異詞個數（1946～2019 年）

從相異字數量跟相異詞數量可分析字詞的豐富度，但有可能是因為篇數的影響，影響了相異字數量跟相異詞數量的多寡。因此進一步去分析各年最常見的前 500 字字數比例和前 500 雙字詞詞數比例，可看出兩者的趨勢也是極為相似的。在 1960 年代中期到 1970 年代後期，最常見的前 500 字字數比例和前 500 雙字詞詞數比例都是最高的。同時，在人民日報創刊初期，最常見的前 500 字字數比例

DOI:10.6814/NCCU202101184

和前 500 雙字詞詞數比例都是最低的。

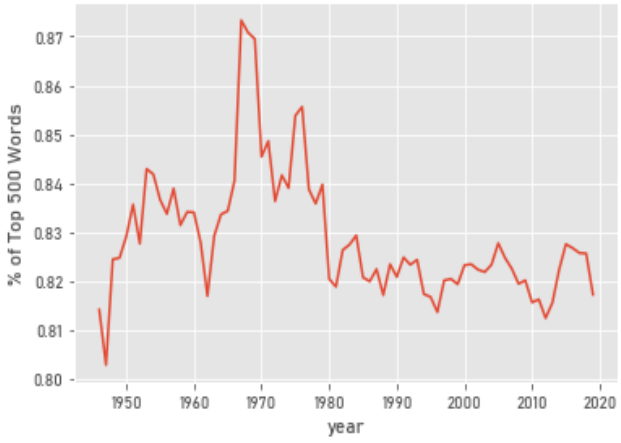

圖 4-7、人民日報頭版前 500 字字數比例（1946～2019 年）

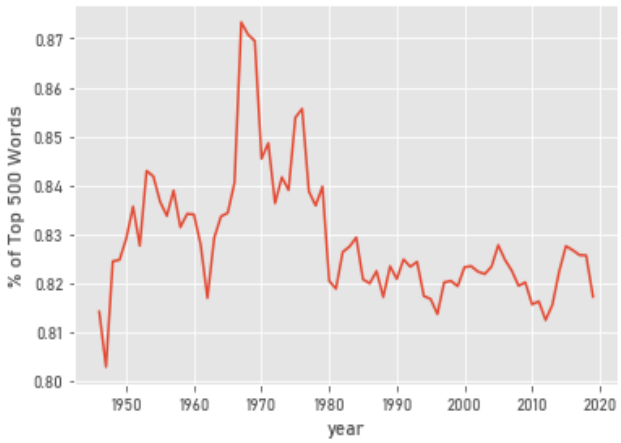

圖 4-8、人民日報頭版前 500 雙字詞比例（1946～2019 年）

從各年的相異字數量、相異詞數量、最常見的前 500 字字數比例和前 500 雙字詞詞數比例來看，1960 年代中期到 1970 年代後期的字詞豐富度是比較低的，

DOI:10.6814/NCCU202101184

而人民日報創刊初期的字詞豐富度是比較高的，因此進一步分析字彙豐富度的另一種衡量方式，常見的統計量為 Type-Token Ratio (TTR)。Type-Token Ratio 的概念是文本是否一遍又一遍地使用相同的字詞，還是使用多種不同的字詞。Type-Token Ratio 的作法是給定文本中的相異字詞總數除以總字詞數量，所以 Type-Token Ratio 的比值越接近 1，則該文本的字詞豐富度越高。本研究使用 Type-Token Ratio 的作法是對每一年的文本進行 1000 次模擬，每次抽 10 萬字，計算出中位數、97.5%的上界、2.5%的下界。從圖 4-8 可看出與各年的相異字數量跟相異詞數量的趨勢極為相似，在 1960 年代中期到 1970 年代後期字詞豐富度是比較低的，而人民日報創刊初期的字詞豐富度是比較高的。

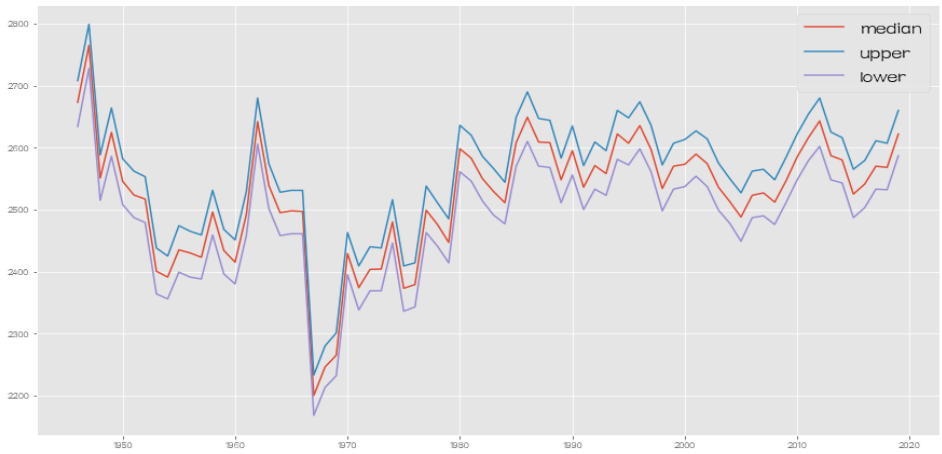

圖 4-9、人民日報頭版 Type-Token Ratio（1946～2019 年）

接者從不均度的角度出發，我們運用 Entropy 這個指標，Entropy 原先是使用在化學和熱力學中，是一種測量值在動力學不能做功的能量總數。Entropy 也被用在計算一個系統混亂的程度。Entropy 是由 Clausius 在 1865 年所提出。在本研究中使用的是 Information Entropy，概念就是 Entropy 越高，則資訊越多。也就是會越混亂。公式為:

DOI:10.6814/NCCU202101184

$$\text{Entropy} = -\sum_{i} p_i \log(p_i)$$

p_i代表某篇文章中第 i 個字詞的出現比例。

從圖 4-9、圖 4-10 可以看出，單字的 Entropy 和雙字詞的 Entropy 兩者的趨勢是非常相似的，從 1950 年代中期到 1970 年代中期都是比較低的，而在人民日報創刊初期，單字的 Entropy 和雙字詞的 Entropy 兩者都非常高，也就是說代表越混亂，也就是資訊量越多。結合豐富度和不均度的結果，在人民日報創刊初期豐富度跟不均度都是較高的，而在 1960 年代中期到 1970 年代後期的豐富度跟不均度都是較低的，其它時期則是較為穩定。

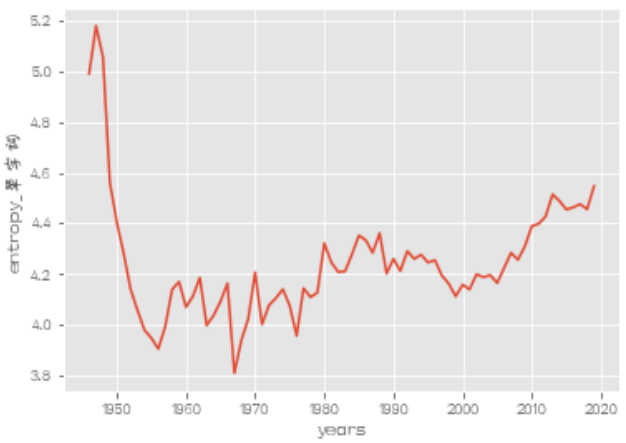

圖 4-10、人民日報頭版單字 Entropy（1946～2019 年）

DOI:10.6814/NCCU202101184

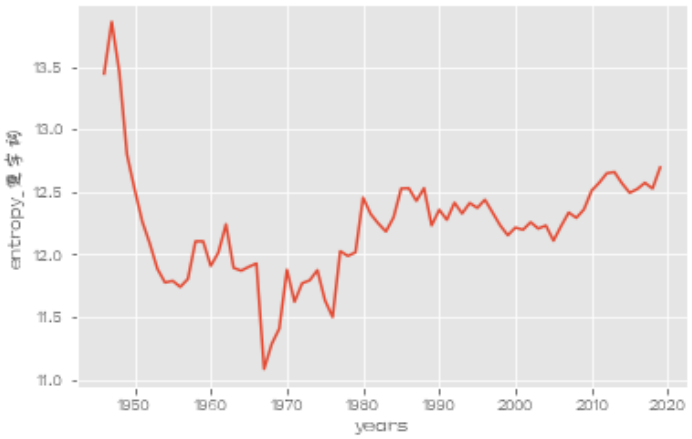

圖 4-11、人民日報頭版雙字詞 Entropy（1946～2019 年）

接者從趨勢變化這個角度出發，圖 4-11、圖 4-12 是最常見的前 500 名單字跟雙字詞的各年重疊率，淺色代表重疊率高，深色則代表重疊率低。從圖中可看出運用前 500 名單字的各年重疊率較沒有群集的跡象，而運用前 500 名雙字詞的各年重疊率大致可分成四群，1946～1948、1949～1965、1966～1978、1979～2019。

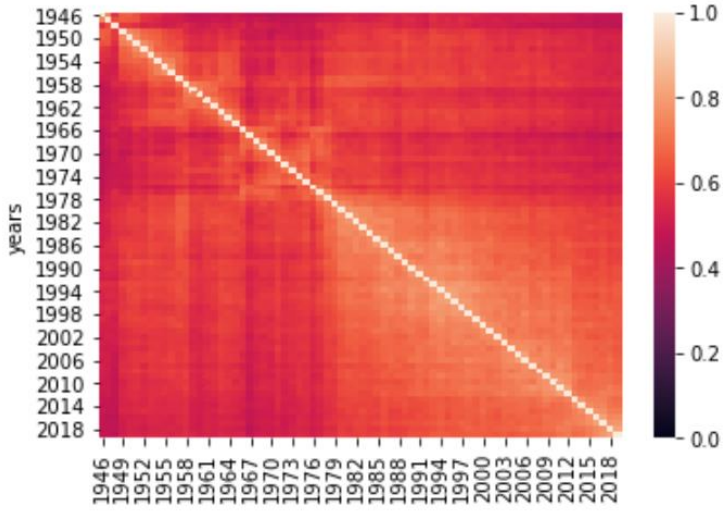

圖 4-12、人民日報頭版前 500 名單字各年重疊率（1946～2019 年）

DOI:10.6814/NCCU202101184

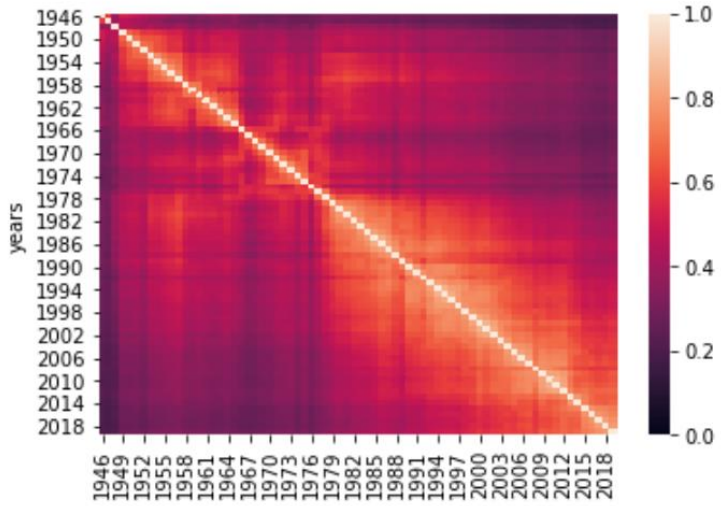

圖 4-13、人民日報頭版前 500 名雙字詞各年重疊率（1946～2019 年）

進一步分析相似度指標，相似度指標使用的有 Jaccard Index(Real, 1996)和 Yue's Index(Yue and Clayton, 2005)，Jaccard Index 是用來比較兩組樣本之間的相似性與差異性，考慮不同字詞數，其定義為兩組樣本的交集(intersection)的元素個數除以聯集(union)的元素個數:

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

其中S_1、S_2分別為兩組樣本的個數，S_{12}為兩組樣本的交集。該指標數值越高代表兩組樣本有越多相同元素，表示兩組樣本的用詞種類相似性越高。本研究利用該指標衡量不同年分的文本用詞種類相似程度。

Yue's Index 相較於 Jaccard Index，考慮字詞使用的頻率，其定義為:

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

其中p_i和q_i第 i 個元素在兩組樣本的出現頻率。該指標數值越高代表兩組樣本中的元素組成比例越相似，表示兩組樣本的用詞組成比例相似性越高。本研究利用

DOI:10.6814/NCCU202101184

該指標衡量不同年分的文本用詞組成比例相似程度。

圖 4-13、圖 4-14 是各年度的最常見的前 500 名單字跟雙字詞的 Jaccard Index，淺色代表 Jaccard Index 較高，深色則代表 Jaccard Index 較低。Jaccard Index 代表年與年之間，使用的字詞的重疊數，但是與重疊數不同的地方在於有做一個類似標準化的動作，避免了重疊數多是因為這兩年的字詞也多的現象。從圖中可看出運用前 500 名單字的各年 Jaccard Index 較沒有群集的跡象，而運用前 500 名雙字詞的各年 Jaccard Index 大致可分成四群，1946～1948、1949～1965、1966～1978、1979～2019。

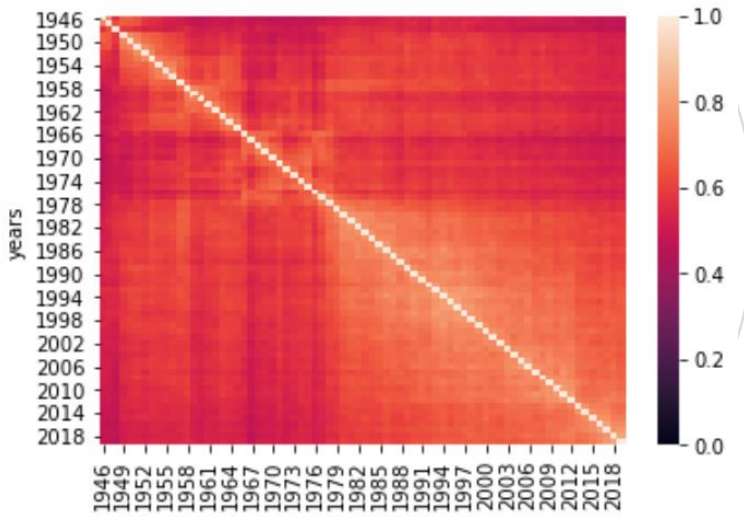

圖 4-14、人民日報頭版前 500 名單字 Jaccard Index（1946～2019 年）

DOI:10.6814/NCCU202101184

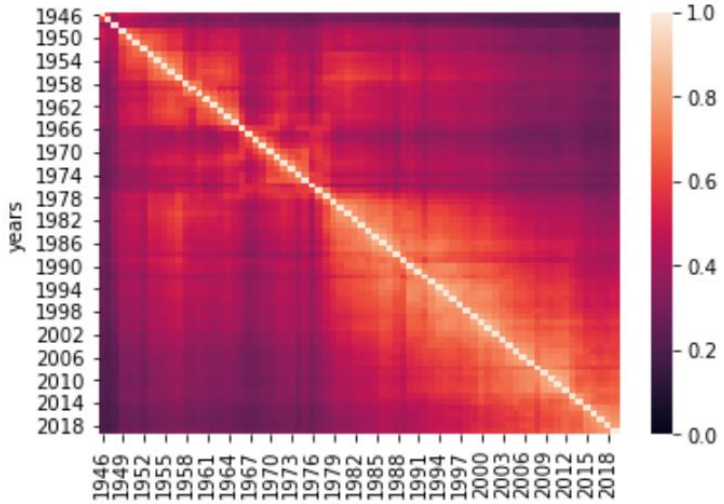

圖 4-15、人民日報頭版前 500 名雙字詞 Jaccard Index（1946～2019 年）

圖 4-15、圖 4-16 是各年度的最常見的 500 名單字跟雙字詞的 Yue's Index，淺色代表 Yue's Index 較高，深色則代表 Yue's Index 較低。該指標相較於 Yue's Index，考慮字詞使用的頻率。從圖中可看出運用前 500 名單字的各年 Yue's Index 較沒有群集的跡象，而運用前 500 名雙字詞的各年 Yue's Index 大致可分成四群，1946～1948、1949～1965、1966～1978、1979～2019，且在前 500 名雙字詞的 Yue's Index 中 1979～2019 有被分為兩群的現象，分成 1979～2003 跟 2004～2019。

DOI:10.6814/NCCU202101184

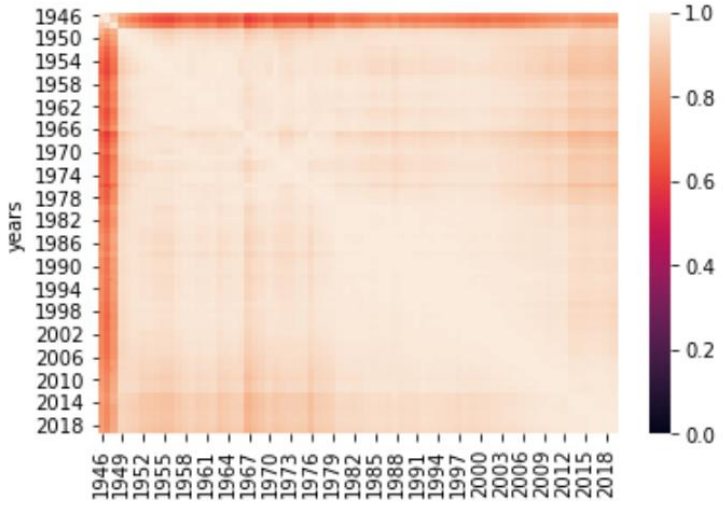

圖 4-16、人民日報頭版前 500 名單字 Yue's Index（1946～2019 年）

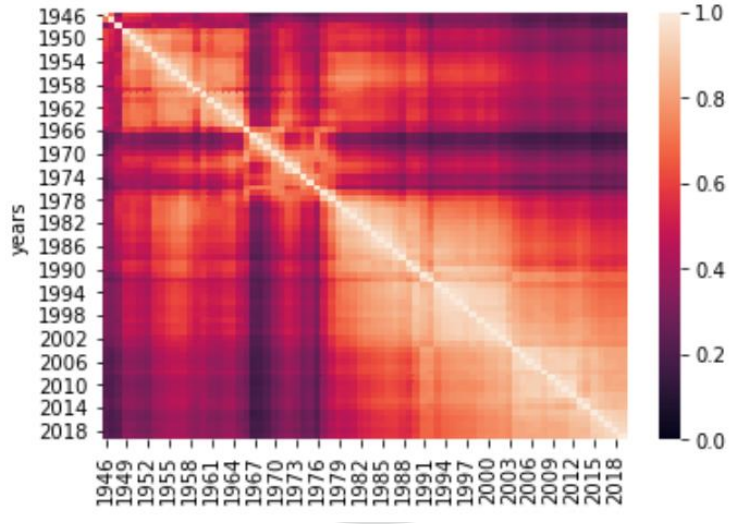

圖 4-17、人民日報頭版前 500 名雙字詞 Yue's Index（1946～2019 年）

從上述可看出 1946 到 1948 年自成一個群集，可能是因為《人民日報》在 1949 年 8 月 1 日從華北中央局機關報升格為中共中央機關報，因此跟其它年分的用字遣詞的風格有很大的不同。

DOI:10.6814/NCCU202101184

選取四群的交界處的年分，計算最常見的前十大單字和前十大雙字詞，去看不同群之間用詞遣字的趨勢變化，從前十大單字來看，1948 到 1950 年，期間有出現了民、國等新的前十大單字，1965 到 1967 年，期間有出現了革、命等新的前十大單字，因此，可以看出不同群之間都有出現上述的趨勢變化。1978 到 1980 年，期間雖然有出現了有、工這兩個新的前十大單字，但可以看出相較於其它群，不同之處較少，2003 到 2005 年，期間有出現發、會這兩個新的前十大單字，而 2005 年和 2019 年的前十大單字是一樣的，只是排名不一樣。

表 4-2、四群交界處年分和 2019 年前十大單字

Rank	1948	1949	1950	1965	1966	1967	1978	1979	1980	2003	2004	2005	2019
1	的	的	的	的	的	的	的	的	的	的	的	的	的
2	一	人	国	国	一	命	一	一	一	国	国	国	国
3	了	民	人	人	主	革	人	国	国	中	中	中	中
4	不	国	民	和	人	大	大	和	和	和	和	和	一
5	十	一	一	一	大	主	和	人	人	一	一	发	和
6	人	中	中	民	是	一	了	了	了	人	人	会	人
7	地	会	会	中	国	产	国	是	在	要	会	一	大
8	有	工	在	主	了	级	主	大	有	大	发	人	发
9	在	在	和	在	和	毛	是	在	是	会	在	大	在
10	是	十	工	了	们	阶	在	工	大	在	了	在	会

DOI:10.6814/NCCU202101184

從前十大雙字詞來看，1948 到 1950 年，出現了人民、中國、美國、朝鮮等新的前十大雙字詞，1965 到 1967 年，出現了群眾、偉大、同志、道路等新的前十大雙字詞，1978 到 1980 年，出現了國家、經濟等前十大雙字詞，2003 到 2005 年，出現了加強、國家、重要等前十大雙字詞。因此，可以看出不同群之間都有出現上述的趨勢變化。1978 到 1980 年，期間雖然有出出現了國家、經濟等前十大雙字詞，但可以看出相較於其它群，不同之處較少，2003 到 2005 年，期間有出現了加強、國家、重要等前十大雙字詞，但是不同之處也較少。

從前十大雙字詞跟前十大單字來看，因為人民日報屬於白話文，多以雙字詞和多字詞來表達，所以運用雙字詞除了能看出趨勢變化之外，還能了解到一些當代的資訊，例如，在 1950 年出現的美國、朝鮮這兩個雙字詞也反映了韓戰這個事件，而這是從單字中無法看出的。在本研究中，運用雙字詞分析會是較好的選擇。

DOI:10.6814/NCCU202101184

表 4-3、四群交界處年分和 2019 年前十大雙字詞

Rank	1948	1949	1950	1965	1966	1967	1978	1979	1980	2003	2004	2005	2019
1	群众	人民	人民	人民	我们	革命	我们	我们	我们	发展	发展	发展	发展
2	生产	中国	我们	中国	人民	我们	工作	工作	工作	工作	中国	中国	中国
3	工作	我们	中国	我们	革命	他们	主席	生产	生产	建设	建设	建设	工作
4	干部	工作	工作	越南	他们	人民	他们	同志	发展	中国	工作	工作	建设
5	领导	代表	代表	他们	同志	群众	人民	问题	问题	重要	经济	合作	国家
6	组织	工人	美国	美国	中国	伟大	同志	发展	同志	经济	合作	经济	我们
7	中农	民主	朝鲜	总理	学习	斗争	问题	人民	人民	我们	重要	我们	经济
8	他们	解放	生产	革命	思想	同志	发展	他们	国家	人民	加强	加强	推动
9	我们	生产	会议	进行	群众	道路	革命	国家	他们	思想	我们	国家	合作
10	问题	他们	和平	斗争	一个	路线	生产	建设	经济	群众	国家	重要	人民

針對這四群的交界處的年分，本研究運用齊夫定律（Zipf's law）去觀測字詞出現的頻率是否正常，齊夫定律是由 Zipf 於 1949 年發表的實驗定律。在自然語言的語料庫裡，一個字詞出現的頻率與它在頻率表里的排名成反比。所以，頻率最高的字詞出現的頻率大約是出現頻率第二名的字詞的 2 倍，而出現頻率第二名的字詞則是出現頻率第四名的字詞的 2 倍。齊夫定律是在描述詞頻分佈的規律，但是因為它是經驗定律，所以此處運用齊夫定律做為參考是否在這四群交界處年分的常用字詞是否符合詞頻分佈規律。

對這四群的交界處的年分個別選取最常見的前 500 名的單字和雙字詞，從圖

DOI:10.6814/NCCU202101184

4-17、圖 4-18 可以看出，不論是單字還是雙字詞，都符合齊夫定律，也就是從齊夫定律的角度出發是符合詞頻分佈規律的。

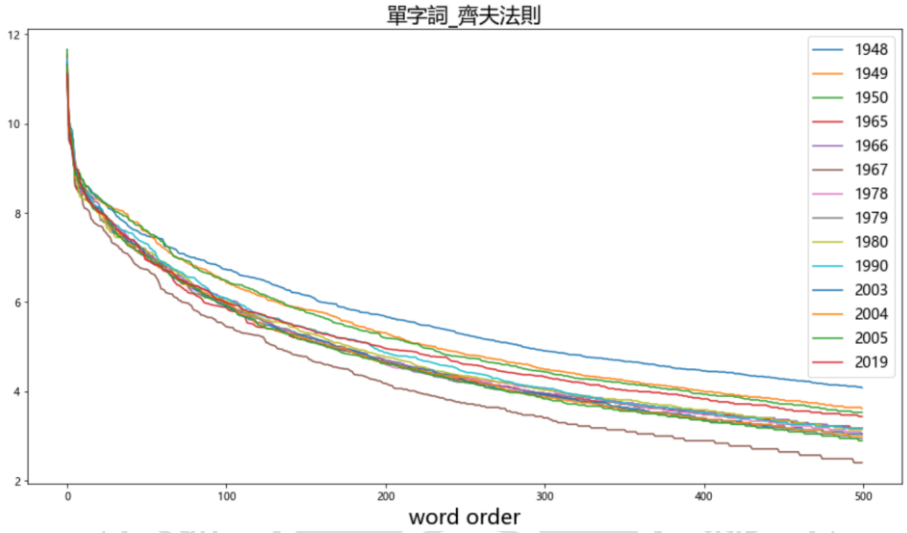

圖 4-18、人民日報頭版前 500 名單字齊夫法則（1946～2019 年）

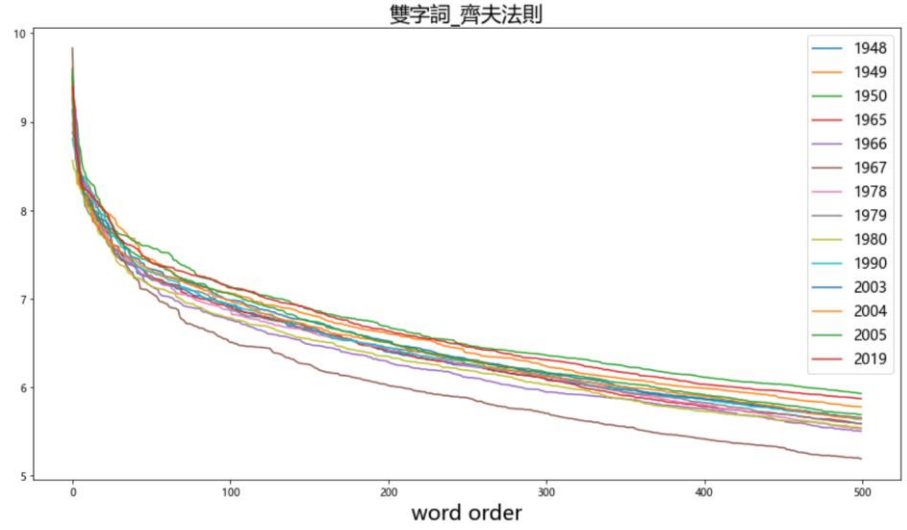

圖 4-19、人民日報頭版前 500 名雙字詞齊夫法則（1946～2019 年）

DOI:10.6814/NCCU202101184

接下來從標點符號的角度出發，表 4-4 和表 4-5 是人民日報每年使用標點符號的數量的平均數、中位數和標準差，同時也計算了百分比。標點符號選取常用的句號、逗號、問號、驚嘆號、分號、冒號、引號“”、頓號。可以看出逗號超過一半的比例，再來則是逗號。

表 4-4、人民日報頭版標點符號統計（1946～2019 年）

	句號	逗號	問號	驚嘆號	分號	冒號	引號“”	頓號
平均數	55518	121926	1104	1275	3703	5639	12667	38783
中位數	58406	125887	998	1059	3968	4657	11321	38981
標準差	11108	18535	492	668	1068	2775	5161	10243

表 4-5、人民日報頭版標點符號百分比統計（1946～2019 年）

	句號%	逗號%	問號%	驚嘆號%	分號%	冒號%	引號“”%	頓號%
平均數	23%	50.95%	0.47%	0.55%	1.51%	2.34%	5.20%	15.98%
中位數	22.99%	50.35%	0.42%	0.41%	1.58%	2.07%	4.70%	15.88%
標準差	2.02%	2.42%	0.20%	0.34%	0.26%	1.04%	1.64%	2.42%

DOI:10.6814/NCCU202101184

圖 4-19 是各年標點符號使用的比例，但其實各個標點符號的趨勢變化都較為穩定，沒有太大的變動。

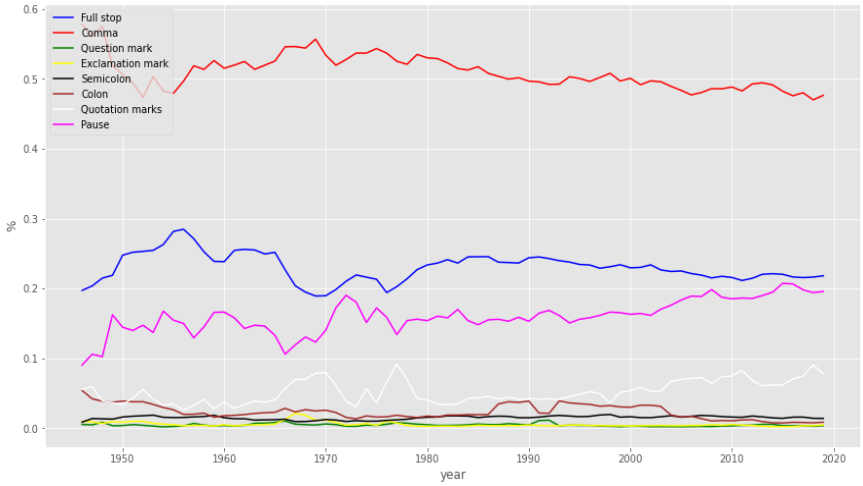

圖 4-20、人民日報頭版標點符號百分比（1946～2019 年）

最後從虛字的角度出發，表 4-6、表 4-7 是人民日報每年使用虛字的數量的平均數、中位數和標準差，同時也計算了百分比。虛字選取常用的的、了、是、和、着、们、个、吗、吧、么。可以看出的超過一半的比例，再來則是和。

表 4-6、人民日報頭版虛字統計（1946～2019 年）

	的	了	是	和	着	们	个	吗	吧	么
平均數	86120	18506	10802	22352	1692	1386	987	145	77	11
中位數	85534	18850	11106	21707	1647	1137	884	122	65	8
標準差	22924	4301	2535	6567	562	774	768	89	59	12

DOI:10.6814/NCCU202101184

表 4-7、人民日報頭版虛字百分比統計（1946～2019 年）

	的%	了%	是%	和%	着%	们%	个%	吗%	吧%	么%
平均數	60.46%	13.14%	7.69%	15.67%	1.20%	0.94%	0.73%	0.10%	0.06%	0.01%
中位數	60.38%	12.86%	7.89%	15.81%	1.19%	0.82%	0.72%	0.08%	0.05%	0.01%
標準差	2.59%	1.82%	1.18%	2.94%	0.29%	0.36%	0.58%	0.06%	0.05%	0.01%

圖 4-20 是各年虛字使用的比例，但其實各個虛字的趨勢變化都較為穩定，沒有太大的變動。

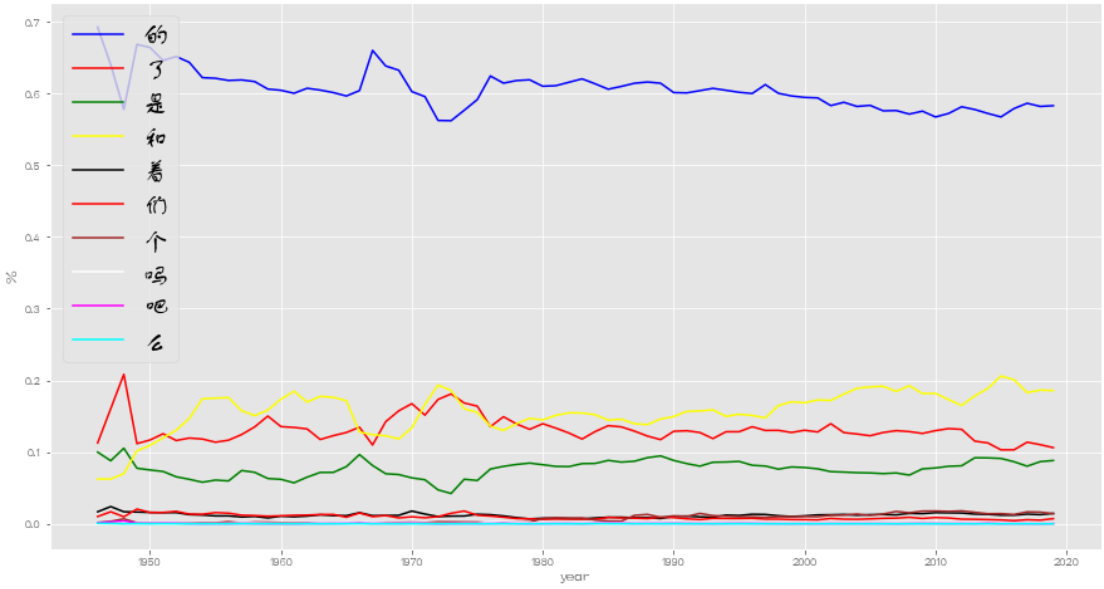

圖 4-21、人民日報頭版虛字百分比（1946～2019 年）

DOI:10.6814/NCCU202101184

綜上所述，各年的用字遣詞確實是有差異的，而在標點符號和虛字的部分則較為平穩。在用字遣詞的部分，因為人民日報是白話文，所以使用雙字詞進行分析會是較好的選擇。

第二節　風格分群

從前面的研究可以發現各年的用字遣字確實是有差異的，而在白話文中使用雙字詞進行分析會是較好的選擇。因此，本研究對各年詞頻前 500 名雙字詞、各年詞頻前 500 名雙字詞的 Jaccard Index、各年詞頻前 500 名雙字詞的 Yue's Index 進行多種分群，這樣同時考慮了用詞種類和用詞頻率，而使用多種分群方法讓分群結果更為穩健。

首先，先對各年詞頻前 500 名雙字詞進行分群，第一個使用的方法是 Hierarchical Clustering。運用 Elbow method 和 Silhouette method 決定最佳的分群數。Elbow method 在選擇分群數時，選擇的分群數小於最佳的分群數時，分群數每增加 1，cost 值就會大幅的減小，當選擇的分群數大於最佳的分群數時，分群數每增加 1，cost 值的變化就不會那麼明顯。因此，最佳的分群數就會出現在這轉折點處。從圖 4-21 可以發現運用 Elbow method 的最佳分群數為 4 群。

DOI:10.6814/NCCU202101184

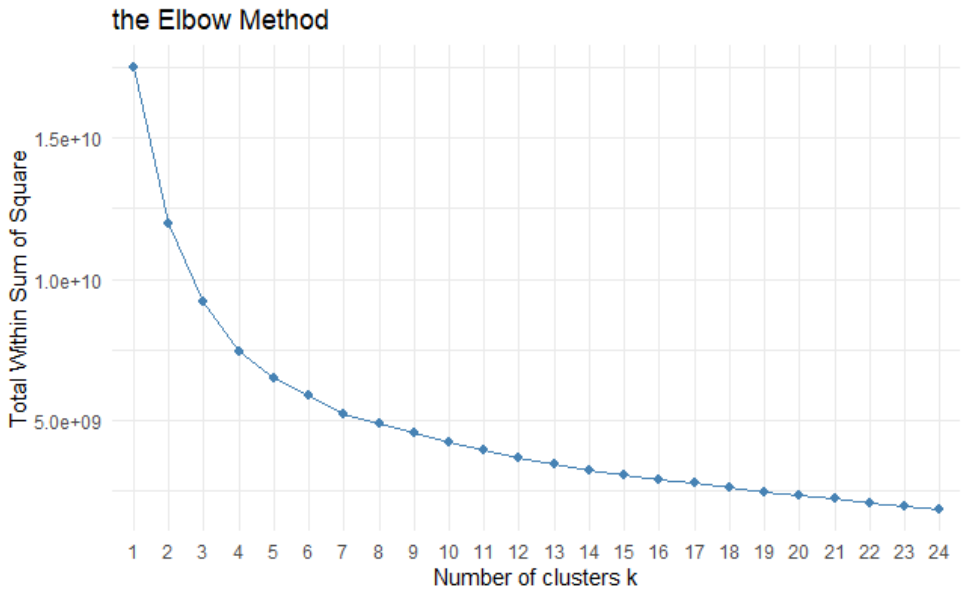

圖 4-22、Hierarchical Clustering 運用 Elbow method 尋找最佳分群數

而運用 Silhouette method 在選擇分群數時，會衡量樣本和所屬的群之間的相似度，會透過 silhouette 值來衡量，silhouette 值的範圍在-1 到 1 之間。silhouette 值越接近 1，代表樣本與所屬的群之間較為相似，而越接近-1，代表樣本與所屬的群之間較不相似。從圖 4-22 可以發現運用 Silhouette method 的最佳分群數為 3 群。

DOI:10.6814/NCCU202101184

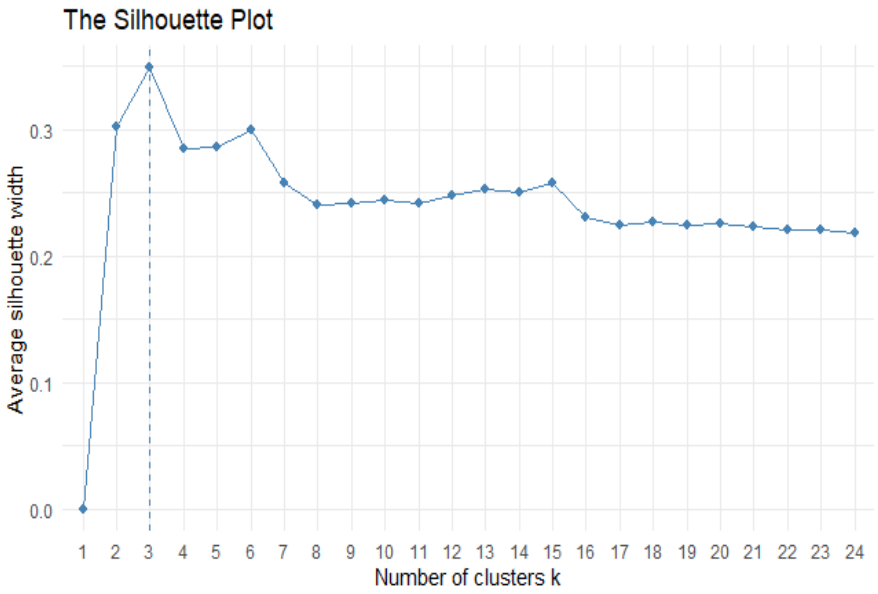

圖 4-23、Hierarchical Clustering 運用 Silhouette method 尋找最佳分群數

綜合 Elbow method 和 Silhouette method 的結果，選擇的最佳分群數為 4 群。對各年詞頻前 500 名雙字詞進行 Hierarchical Clustering，運用歐式距離和 Ward's Method。如圖 4-23 所示，可以看出每一群的年分都是連續的，除了 1946～1948 之外。四群分別為：第一群：1949～1965、第二群：1946～1948、1966～1978、第三群：1979～2003、第四群：2004～2019

DOI:10.6814/NCCU202101184

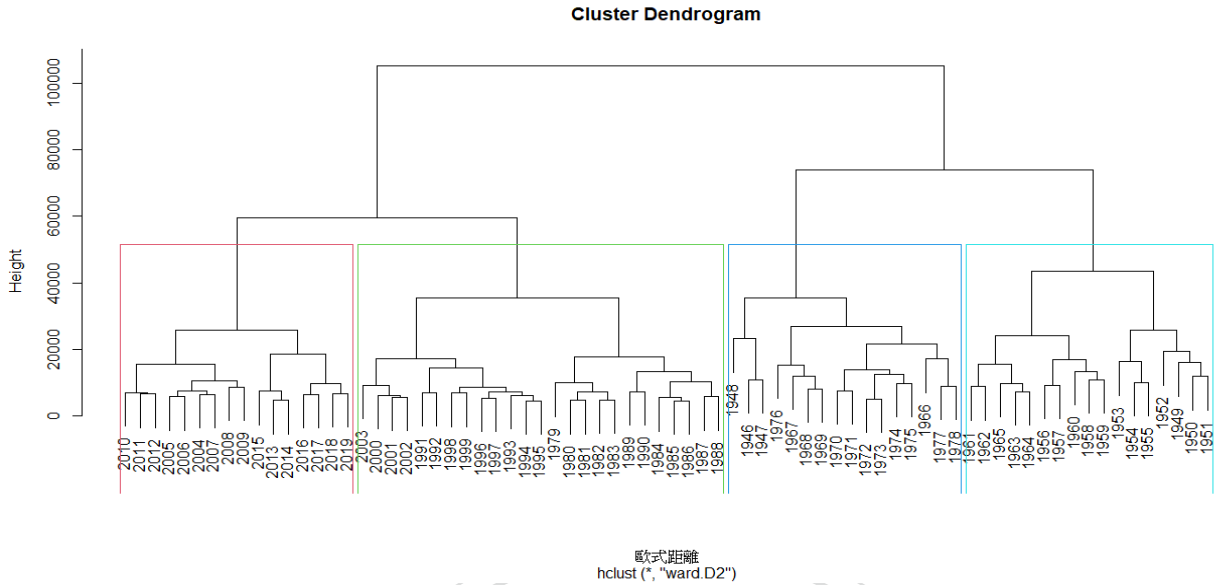

圖 4-24、Hierarchical Clustering 分四群結果

接著進行 K-means 分群，綜合 Elbow method 和 Silhouette method 的結果，選擇的最佳分群數為 4 群。

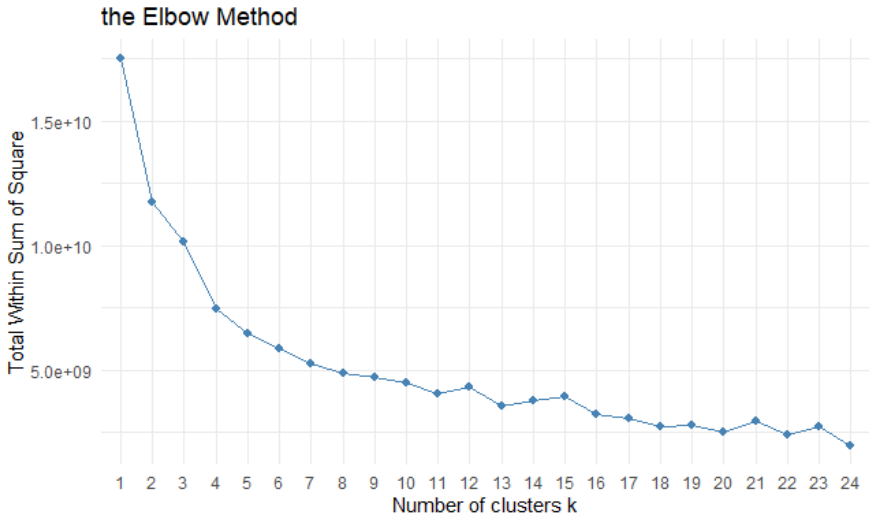

圖 4-25、K-means 運用 Elbow method 尋找最佳分群數

DOI:10.6814/NCCU202101184

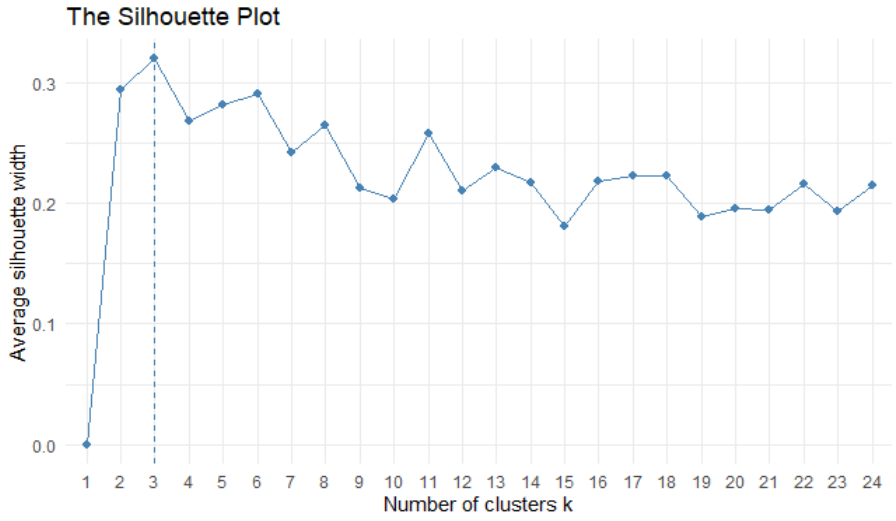

圖 4-26、K-means 運用 Silhouette method 尋找最佳分群數

對各年詞頻前 500 名雙字詞進行 K-means 分群，如圖 4-26 所示，可以看出每一群的年分都是連續的，除了 1946～1948 之外。四群分別為：第一群：1949～1965、第二群：1946～1948、1966～1978、第三群：1979～2001、第四群：2002～2019。從分群結果可看出相較於 Hierarchical Clustering 的結果是很相似的。只有在第三群和第四群的年分有些微的差異。

DOI:10.6814/NCCU202101184

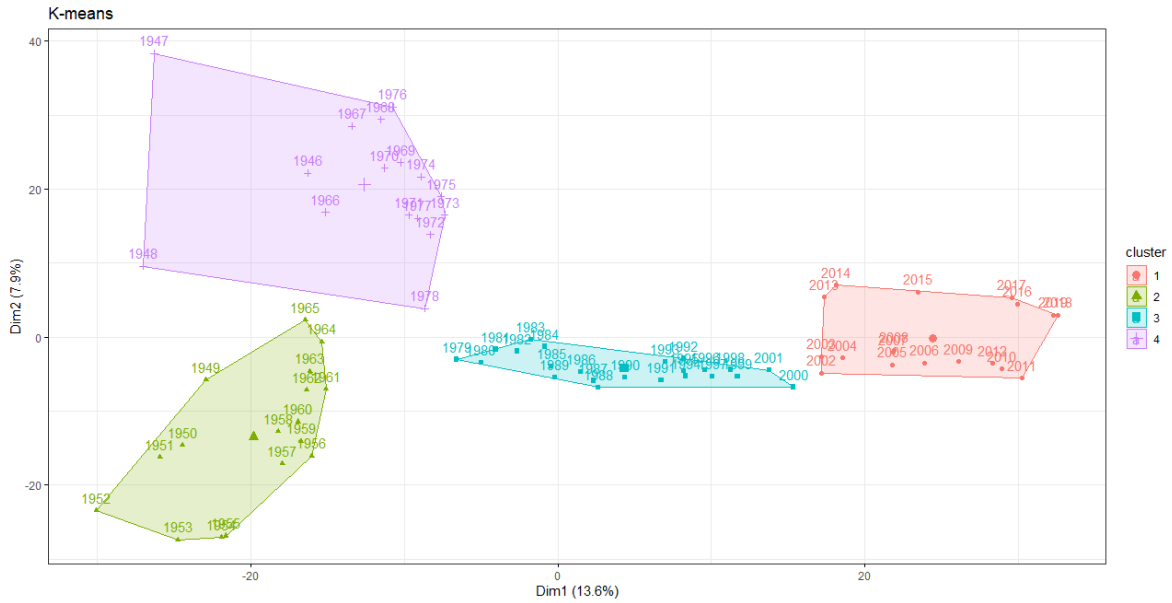

圖 4-27、K-means 分四群結果

接著進行 K-medoid 分群，綜合 Elbow method 和 Silhouette method 的結果，選擇的最佳分群數為 4 群。

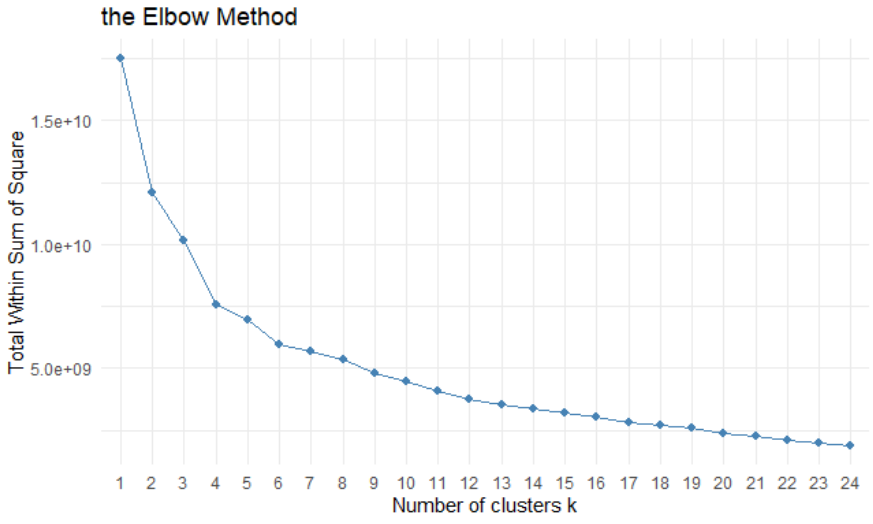

圖 4-28、K-medoid 運用 Elbow method 尋找最佳分群數

DOI:10.6814/NCCU202101184

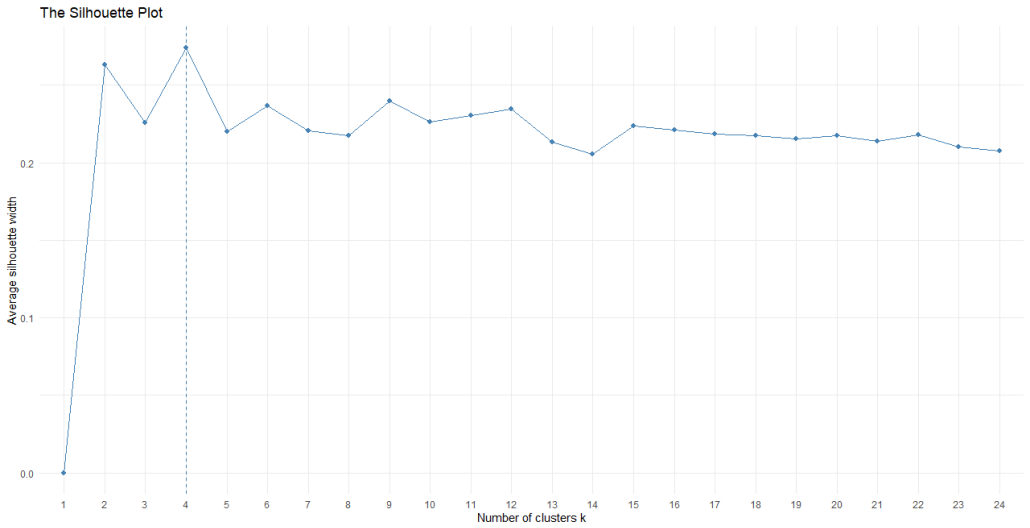

圖 4-29、K-medoid 運用 Silhouette method 尋找最佳分群數

對各年詞頻前 500 名雙字詞進行 K-medoid 分群，如圖 4-29 所示，可以看出每一群的年分都是連續的，除了 1946～1948 之外。四群分別為：第一群：1949～1965、第二群：1946～1948、1966～1978、第三群：1979～2002、第四群：2003～2019。從分群結果可看出相較於 K-means 和 Hierarchical Clustering 的結果是很相似的。只有在第三群和第四群的年分有些微的差異。

DOI:10.6814/NCCU202101184

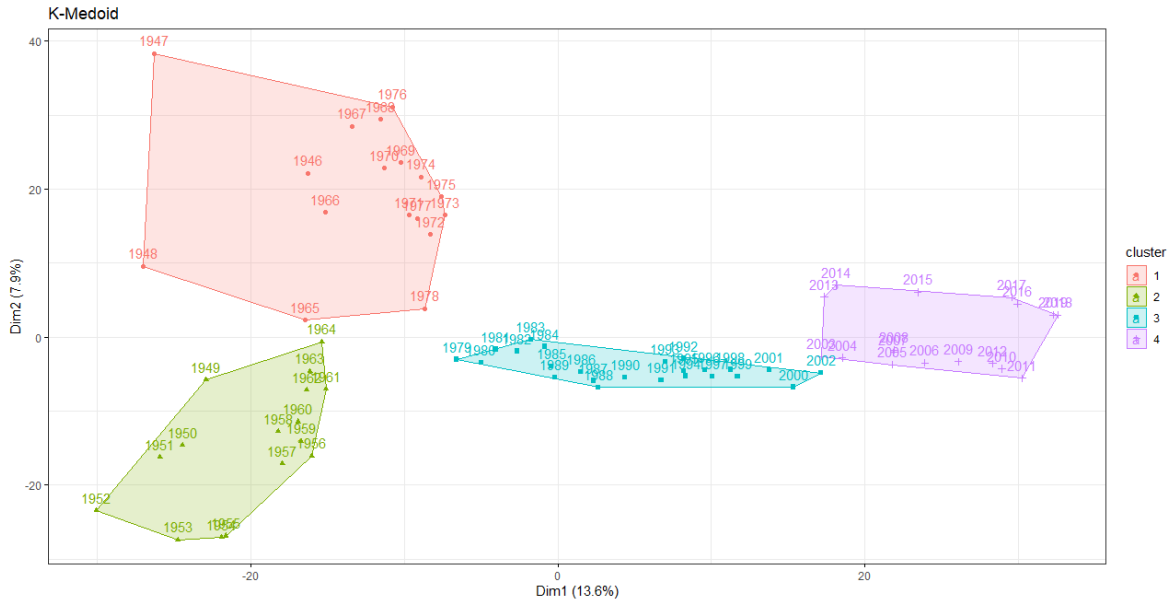

圖 4-30、K-medoid 分四群結果

接著對各年詞頻前 500 名雙字詞進行 PCA，將資料以 0 做中心化，以單位方差進行標準化之後，進行 PCA，如圖 4-30 所示，當分為 4 群時，可以看出每一群的年分都是連續的，除了 1946～1948 之外。四群分別為：第一群：1949～1965、第二群：1946～1948、1966～1978、第三群：1979～2001、第四群：2002～2019。從分群結果可看出跟 K-means 的結果是一樣的。

DOI:10.6814/NCCU202101184

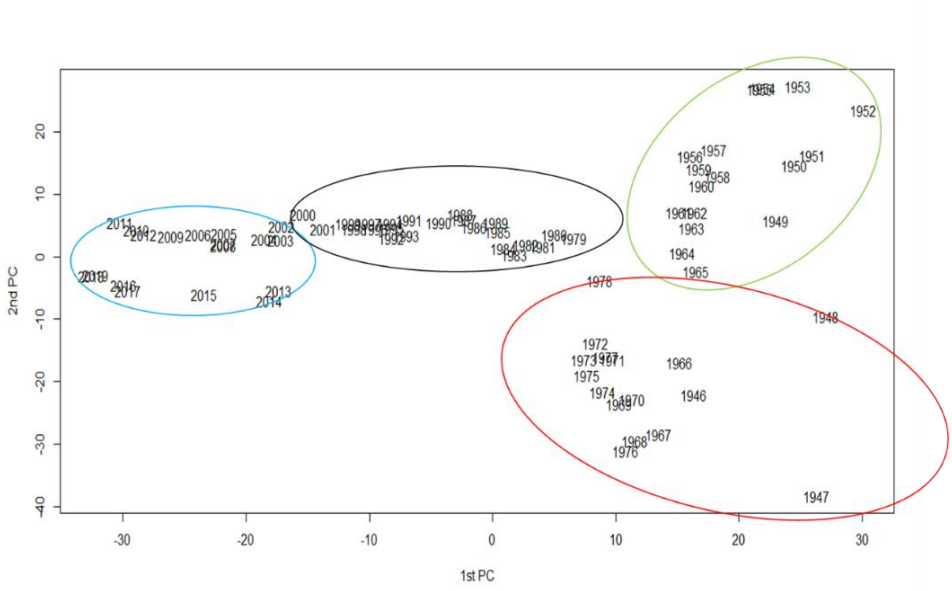

圖 4-31、PCA 分四群結果

最後對各年詞頻前 500 名雙字詞進行 t-SNE，如圖 4-31 所示，當分為 4 群時，可以看出每一群的年分都是連續的，除了 1946～1948 之外。四群分別為：第一群：1949～1965、第二群：1946～1948、1966～1978、第三群：1979～2003、第四群：2004～2019。 從分群結果可看出跟 Hierarchical Clustering 的結果是一樣的。

DOI:10.6814/NCCU202101184

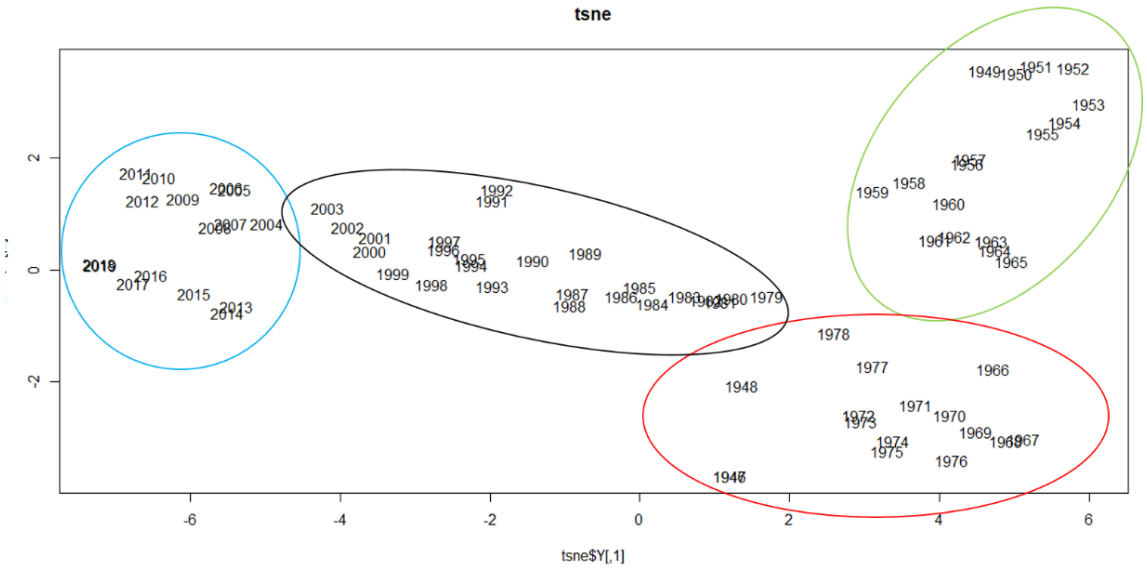

圖 4-32、t-SNE 分四群結果

對各年詞頻前 500 名雙字詞進行多種分群後，接著對各年詞頻前 500 名雙字詞的 Jaccard Index 進行分群，第一個使用的方法是 Hierarchical Clustering。運用 Elbow method 和 Silhouette method 決定最佳的分群數。綜合 Elbow method 和 Silhouette method 的結果，選擇的最佳分群數為 3 群。

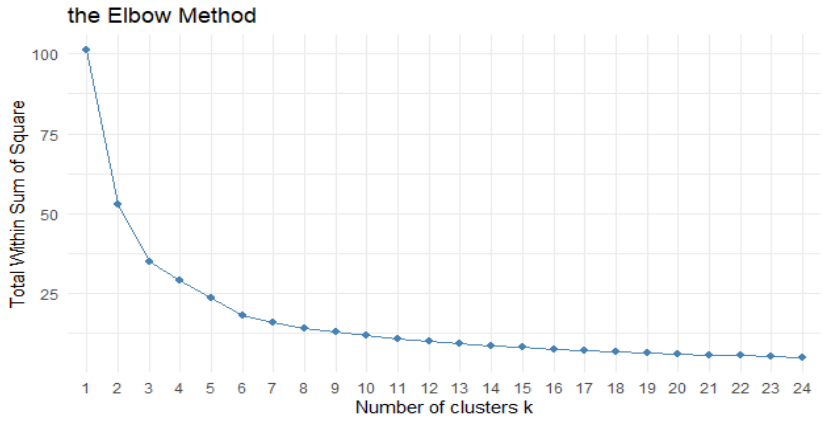

圖 4-33、Hierarchical Clustering 運用 Elbow method 尋找最佳分群數

DOI:10.6814/NCCU202101184

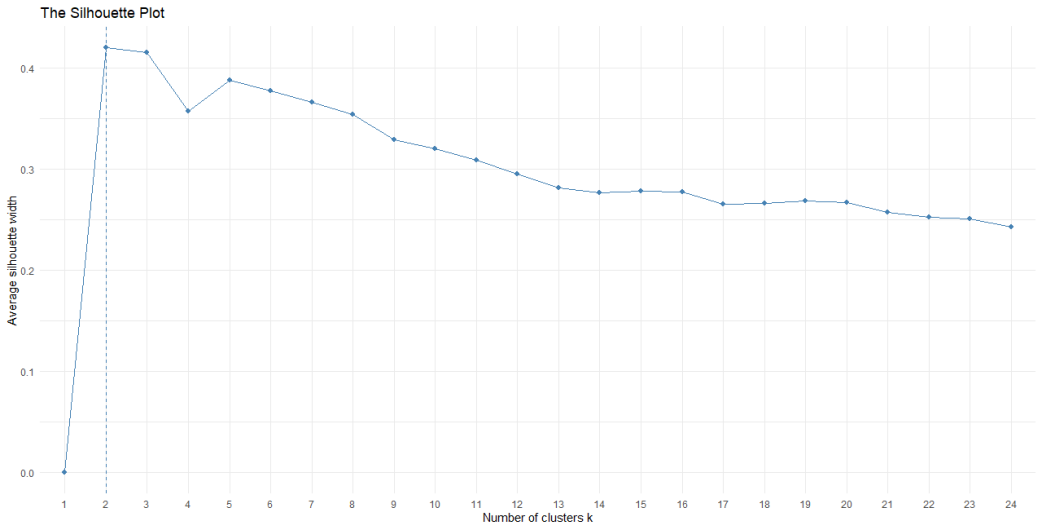

圖 4-34、Hierarchical Clustering 運用 Silhouette method 尋找最佳分群數

對各年詞頻前 500 名雙字詞的 Jaccard Index 進行 Hierarchical Clustering，運用歐式距離和 Ward's Method。如圖 4-34 所示，可以看出每一群的年分都是連續的。三群分別為：第一群：1946～1977、第二群：1978～2002、第三群：2003～2019。

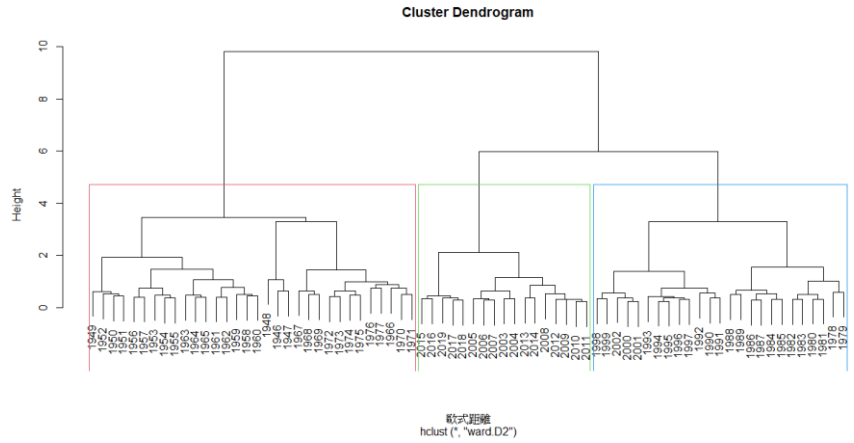

圖 4-35、Hierarchical Clustering 分三群結果

DOI:10.6814/NCCU202101184

接著進行 K-means 分群，綜合 Elbow method 和 Silhouette method 的結果，選擇的最佳分群數為 3 群。

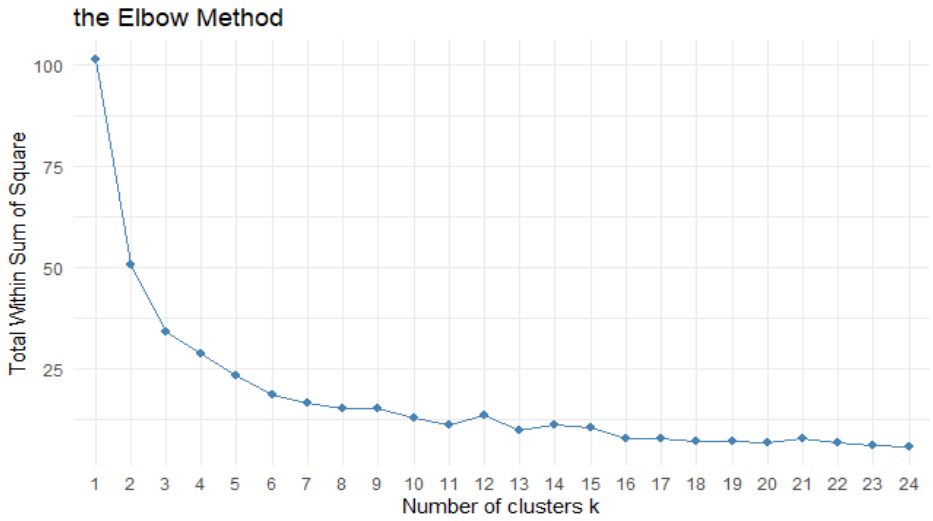

圖 4-36、K-means 運用 Elbow method 尋找最佳分群數

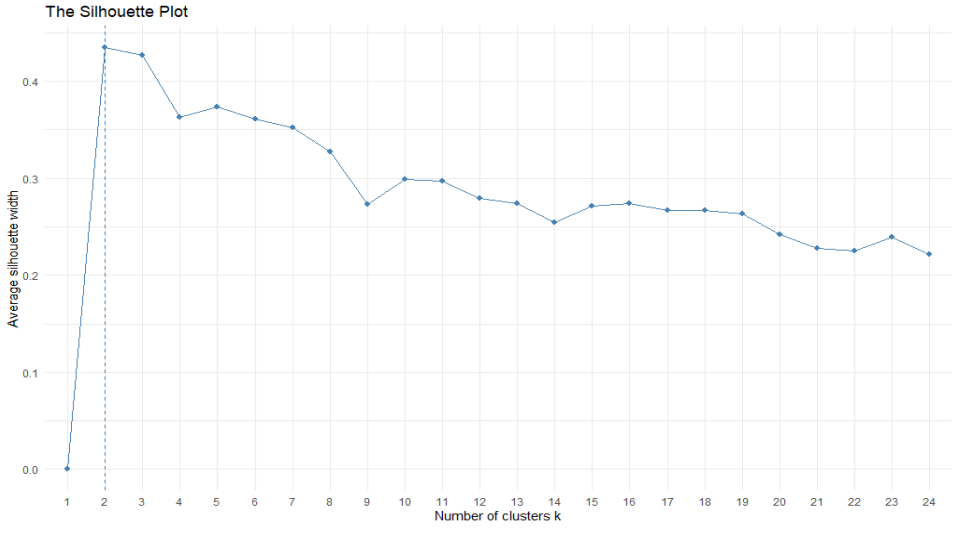

圖 4-37、K-means 運用 Silhouette method 尋找最佳分群數

DOI:10.6814/NCCU202101184

對各年詞頻前 500 名雙字詞的 Jaccard Index 進行 K-means 分群，如圖 4-37 所示，可以看出每一群的年分都是連續的。三群分別為：第一群：1946～1978、第二群：1979～2001、第三群：2002～2019，跟 Hierarchical Clustering 的結果相似。

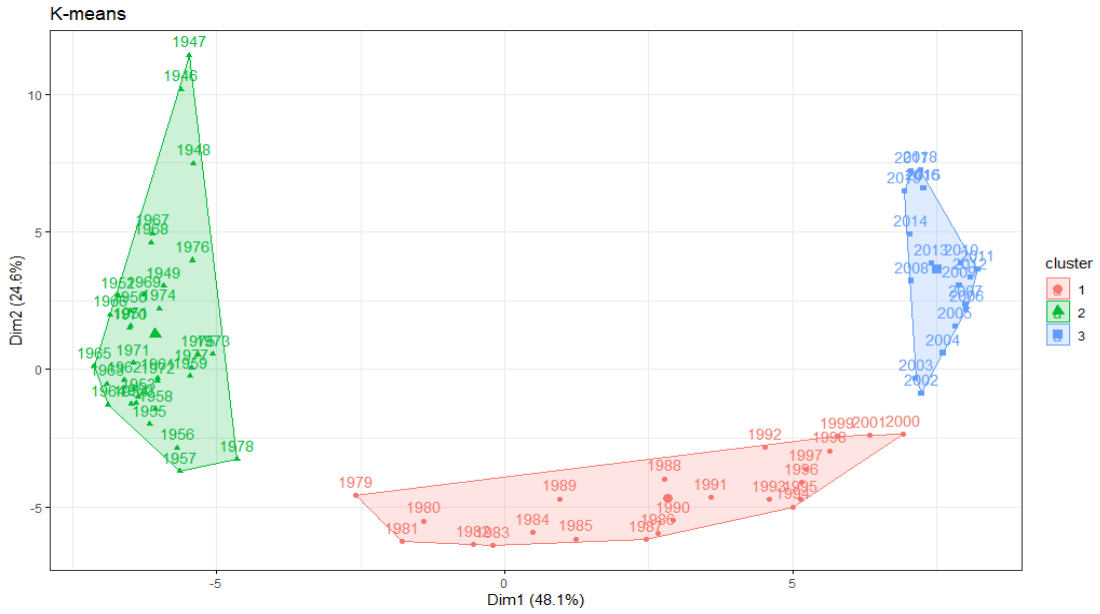

圖 4-38、K-means 分三群結果

接著進行 K-medoid 分群，綜合 Elbow method 和 Silhouette method 的結果，選擇的最佳分群數為 3 群。

DOI:10.6814/NCCU202101184

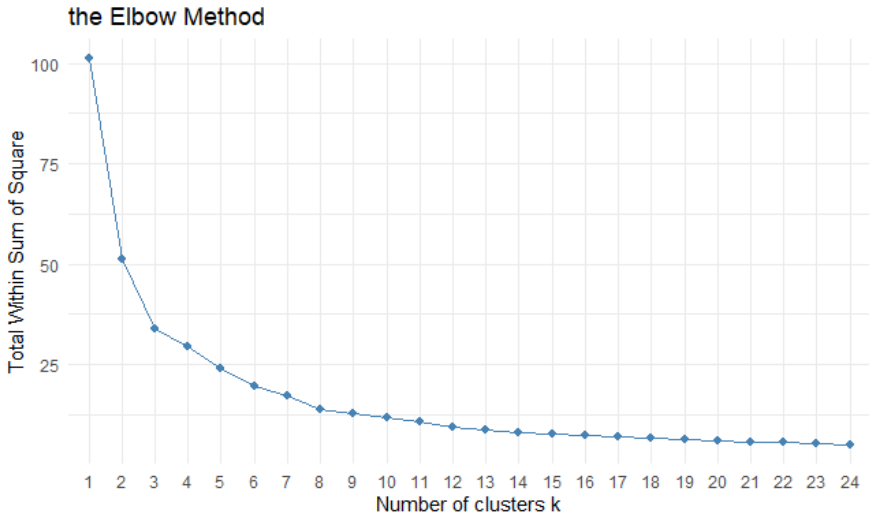

圖 4-39、K-medoid 運用 Elbow method 尋找最佳分群數

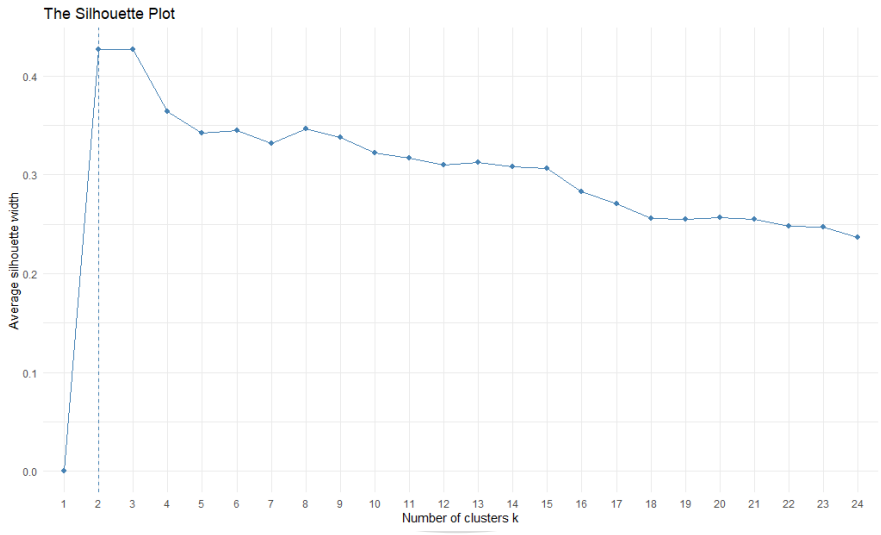

圖 4-40、K-medoid 運用 Silhouette method 尋找最佳分群數

對各年詞頻前 500 名雙字詞的 Jaccard Index 進行 K-medoid 分群，如圖 4-40 所示，可以看出每一群的年分都是連續的。三群分別為：第一群：1946～1978、第二群：1979～2001、第三群：2002～2019，跟 K-means 的結果一樣。

DOI:10.6814/NCCU202101184

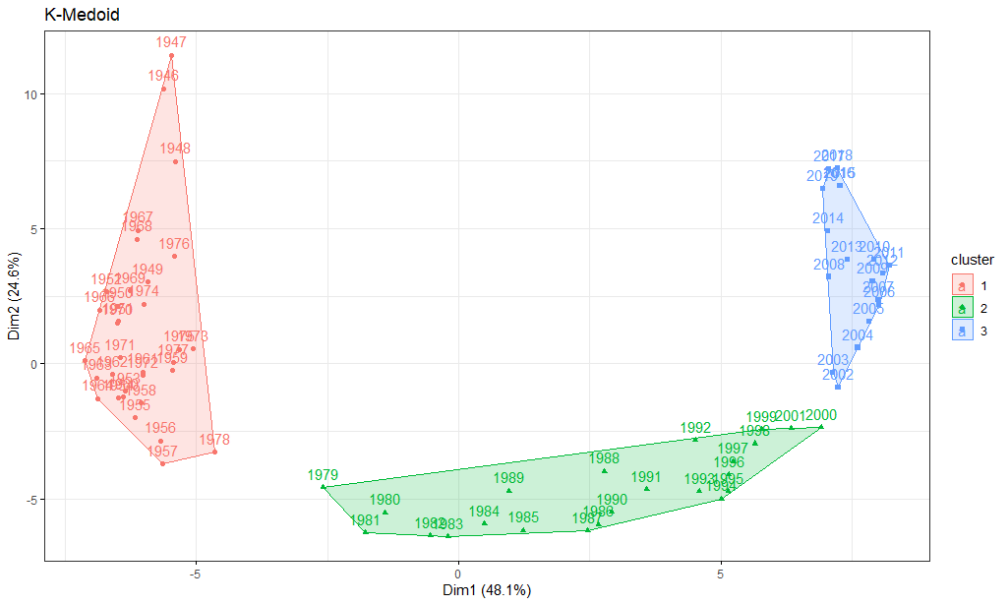

圖 4-41、K-medoid 分三群結果

接著對各年詞頻前 500 名雙字詞的 Jaccard Index 進行 PCA，將資料以 0 做中心化，以單位方差進行標準化之後，進行 PCA，如圖 4-41 所示，當分為 3 群時，可以看出每一群的年分都是連續的。三群分別為：第一群：1946～1978、第二群：1979～2003、第三群：2004～2019。跟 K-medoid、K-means 和 Hierarchical Clustering 的結果相似，只有在群跟群之間的交界年分有所不同。

DOI:10.6814/NCCU202101184

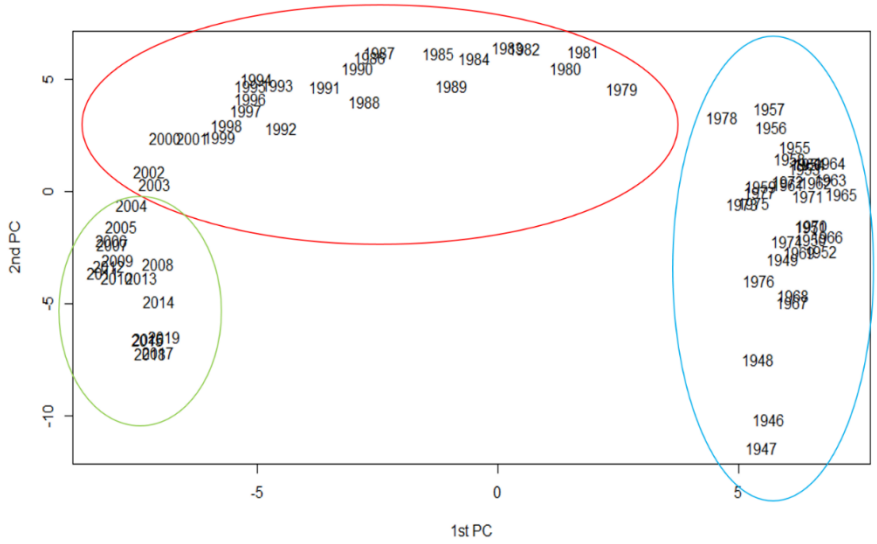

圖 4-42、PCA 分三群結果

最後對各年詞頻前 500 名雙字詞的 Jaccard Index 進行 t-SNE，如圖 4-42 所示，當分為 3 群時，可以看出每一群的年分都是連續的。三群分別為：第一群：1946～1978、第二群：1979～2003、第三群：2004～2019。跟 PCA 的結果一樣。

DOI:10.6814/NCCU202101184

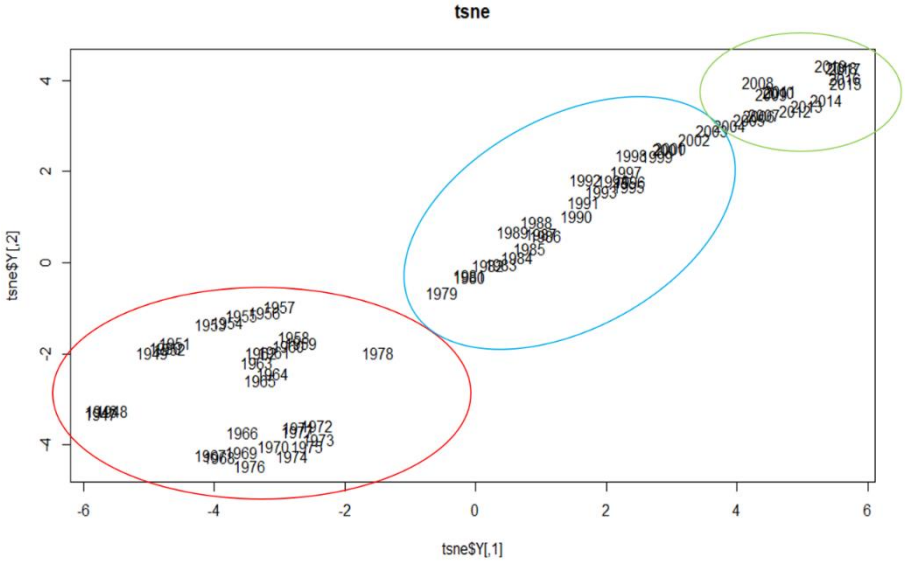

圖 4-43、t-SNE 分三群結果

對各年詞頻前 500 名雙字詞和各年詞頻前 500 名雙字詞的 Jaccard Index 進行多種分群後，接著對各年詞頻前 500 名雙字詞的 Yue's Index 進行分群，第一個使用的方法是 Hierarchical Clustering。運用 Elbow method 和 Silhouette method 決定最佳的分群數。綜合 Elbow method 和 Silhouette method 的結果，選擇的最佳分群數為 4 群。

DOI:10.6814/NCCU202101184

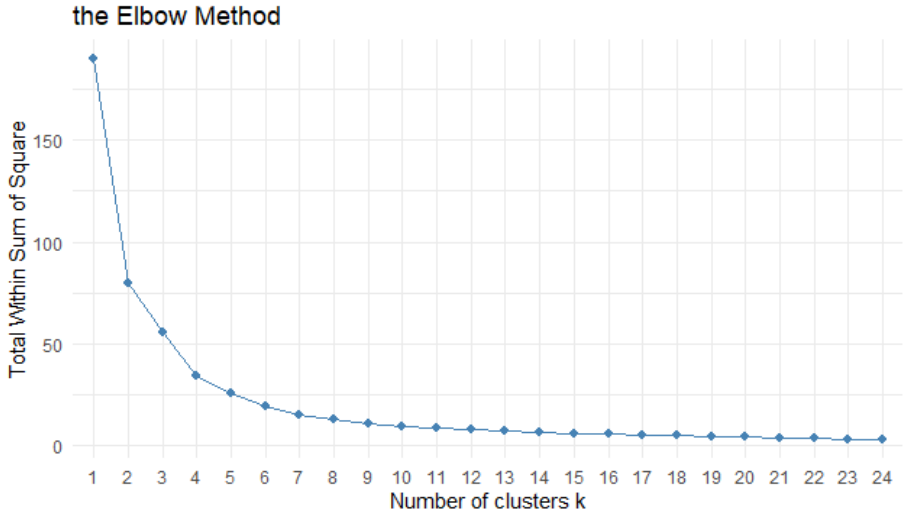

圖 4-44、Hierarchical Clustering 運用 Elbow method 尋找最佳分群數

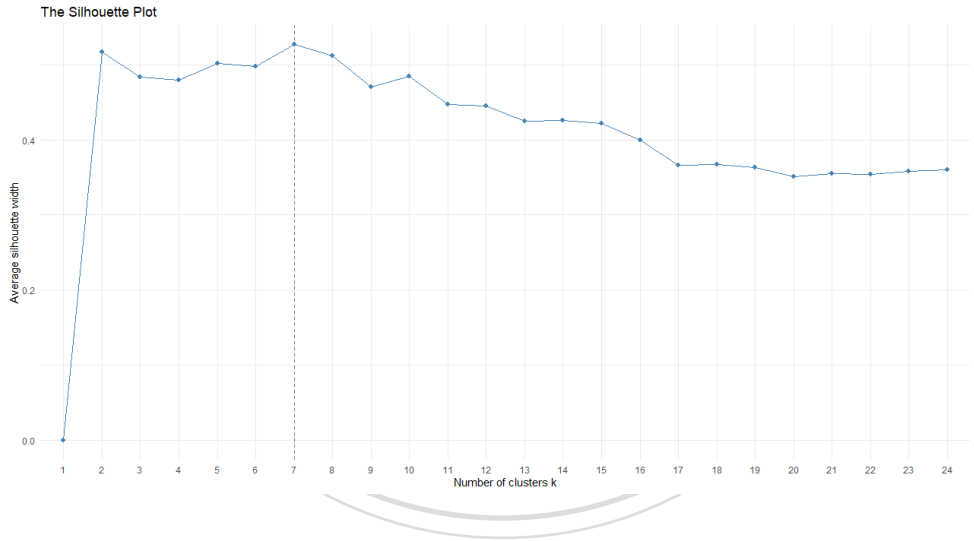

圖 4-45、Hierarchical Clustering 運用 Silhouette method 尋找最佳分群數

對各年詞頻前 500 名雙字詞的 Yue's Index 進行 Hierarchical Clustering，運用歐式距離和 Ward's Method。如圖 4-45 所示，可以看出每一群的年分都是連續的。四群分別為：第一群：1949～1965、第二群：1946～1948、1966～1977、第三群：1978～2003、第四群：2004～2019。

DOI:10.6814/NCCU202101184

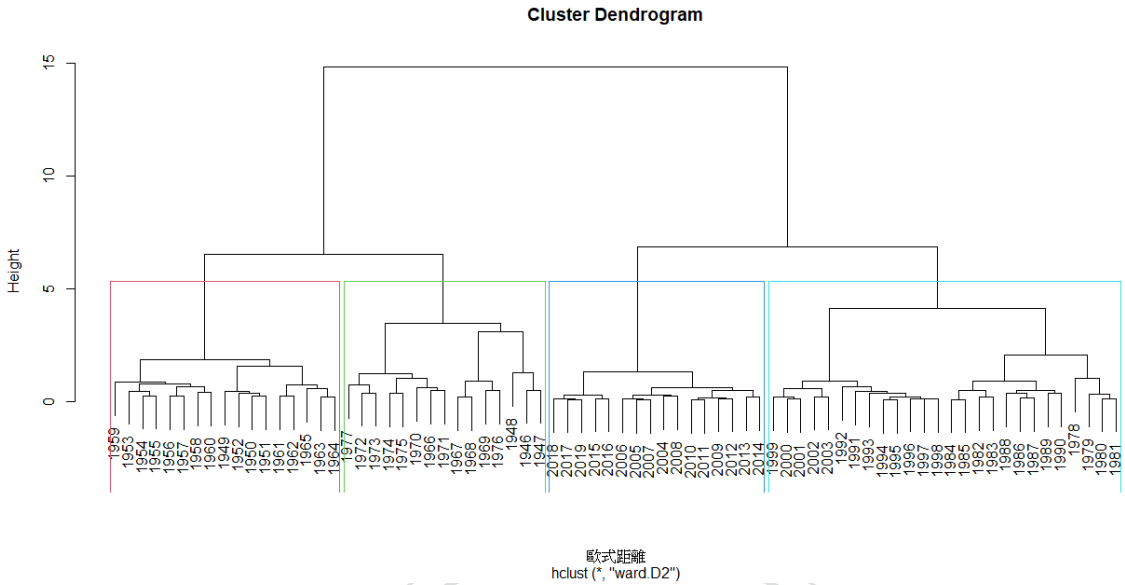

圖 4-46、Hierarchical Clustering 分四群結果

接著進行 K-means 分群，綜合 Elbow method 和 Silhouette method 的結果，選擇的最佳分群數為 4 群。

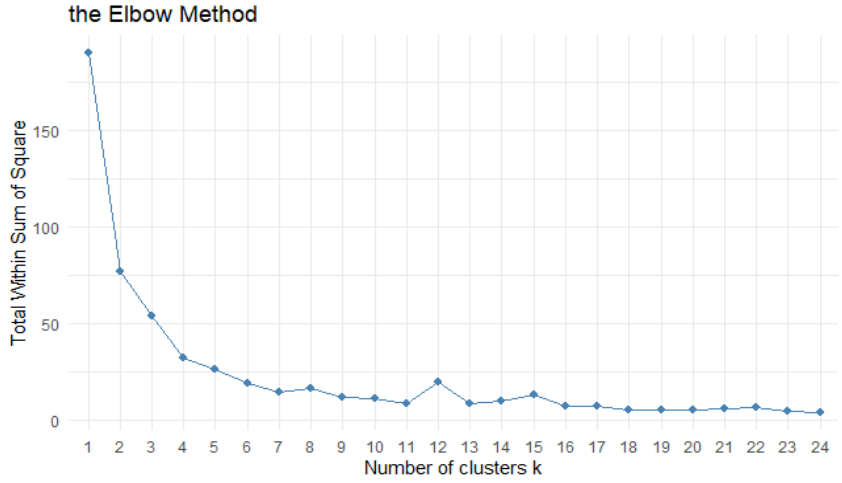

圖 4-47、K-means 運用 Elbow method 尋找最佳分群數

DOI:10.6814/NCCU202101184

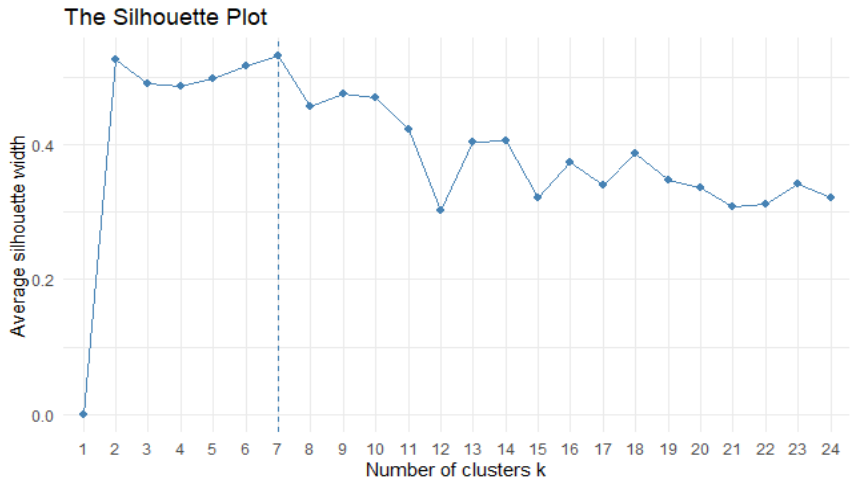

圖 4-48、K-means 運用 Silhouette method 尋找最佳分群數

對各年詞頻前 500 名雙字詞的 Yue's Index 進行 K-means 分群，如圖 4-48 所示，四群分別為：第一群：1949～1965、1972、1978、第二群：1946～1948、1966～1971、1973～1977、第三群：1979～2001、第四群：2002～2019。在第一群和第二群出現了一些不連續的年分。

DOI:10.6814/NCCU202101184

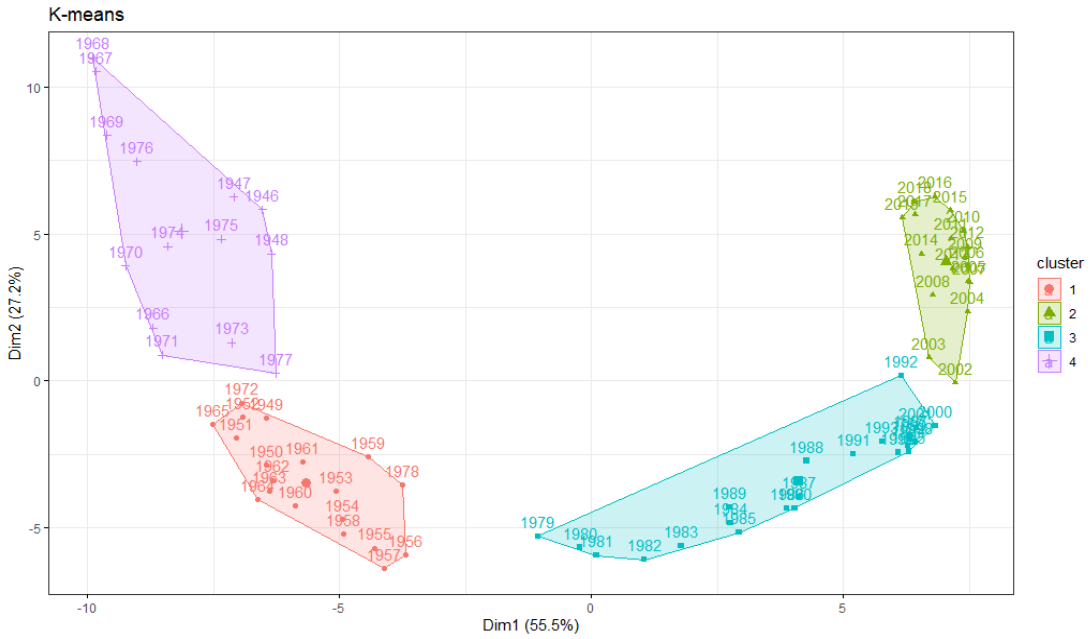

圖 4-49、K-means 分四群結果

接著進行 K-medoid 分群，綜合 Elbow method 和 Silhouette method 的結果，選擇的最佳分群數為 4 群。

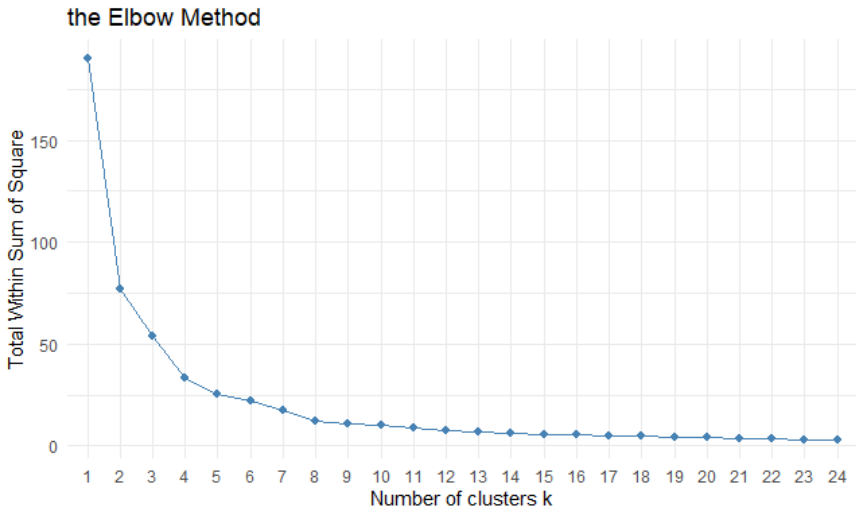

圖 4-50、K-medoid 運用 Elbow method 尋找最佳分群數

DOI:10.6814/NCCU202101184

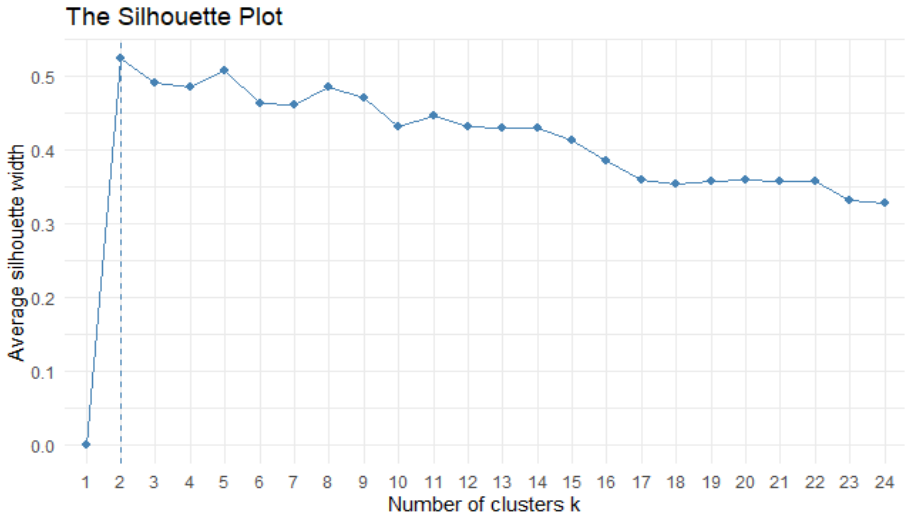

圖 4-51、K-medoid 運用 Silhouette method 尋找最佳分群數

對各年詞頻前 500 名雙字詞的 Yue's Index 進行 K-medoid 分群，如圖 4-51 所示，四群分別為：第一群：1949～1965、1978～1981、第二群：1946～1948、1966～1977、第三群：1982～2003、第四群：2004～2019。在第一群和第二群出現了一些不連續的年分。

DOI:10.6814/NCCU202101184

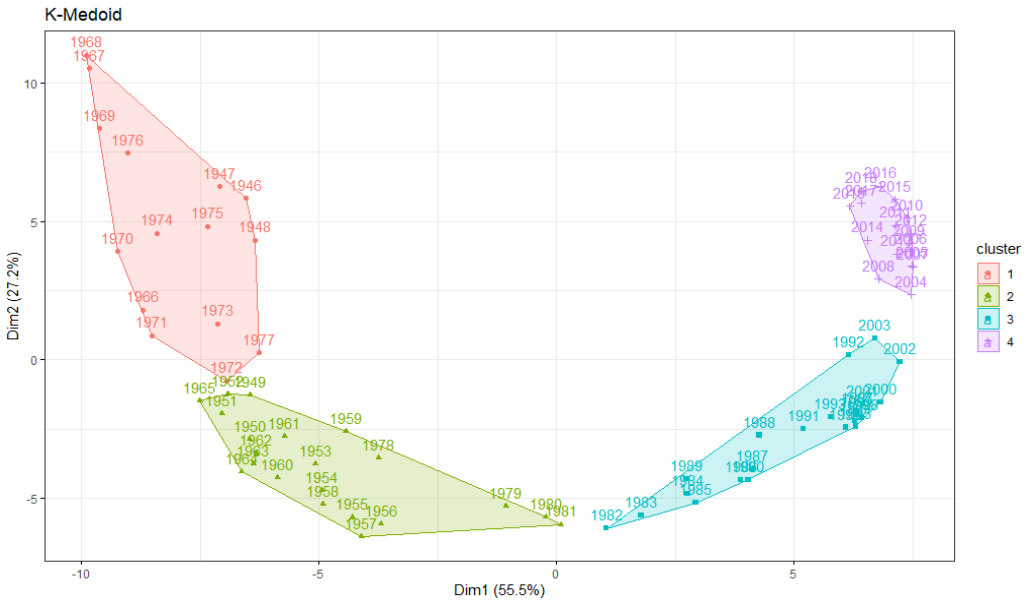

圖 4-52、K-medoid 分四群結果

接著對各年詞頻前 500 名雙字詞的 Yue's Index 進行 PCA，將資料以 0 做中心化，以單位方差進行標準化之後，進行 PCA，如圖 4-52 所示，當分為 4 群時，可以看出每一群的年分都是連續的。四群分別為：第一群：1949～1965、1972、1977、1978、第二群：1946～1948、1966～1971、1973～1976、第三群：1979～2003、第四群：2004～2019，在第一群和第二群出現了一些不連續的年分。

DOI:10.6814/NCCU202101184

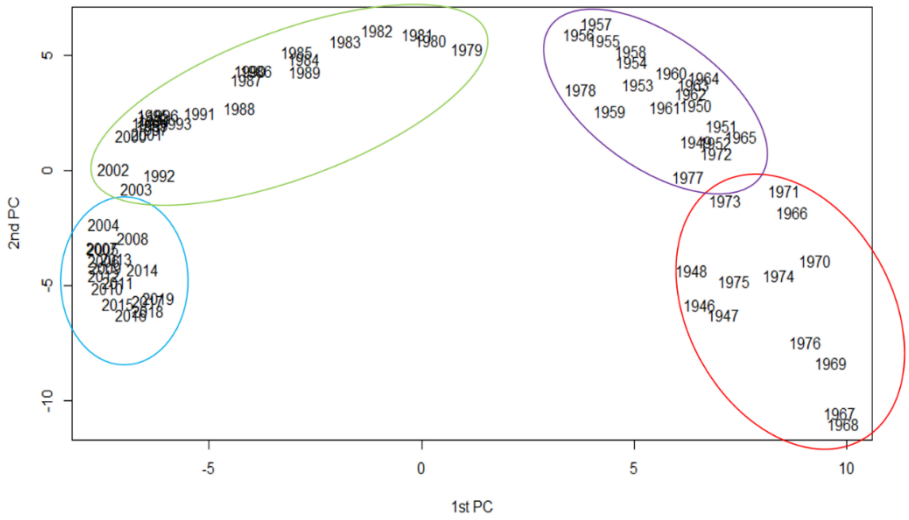

圖 4-53、PCA 分四群結果

最後對各年詞頻前 500 名雙字詞的 Yue's Index 進行 t-SNE，如圖 4-53 所示，當分為四群時，第一群：1946-1948、1966-1977、第二群：1949-1965、1978、第三群：1979-2003、第四群：2004-2019。

DOI:10.6814/NCCU202101184

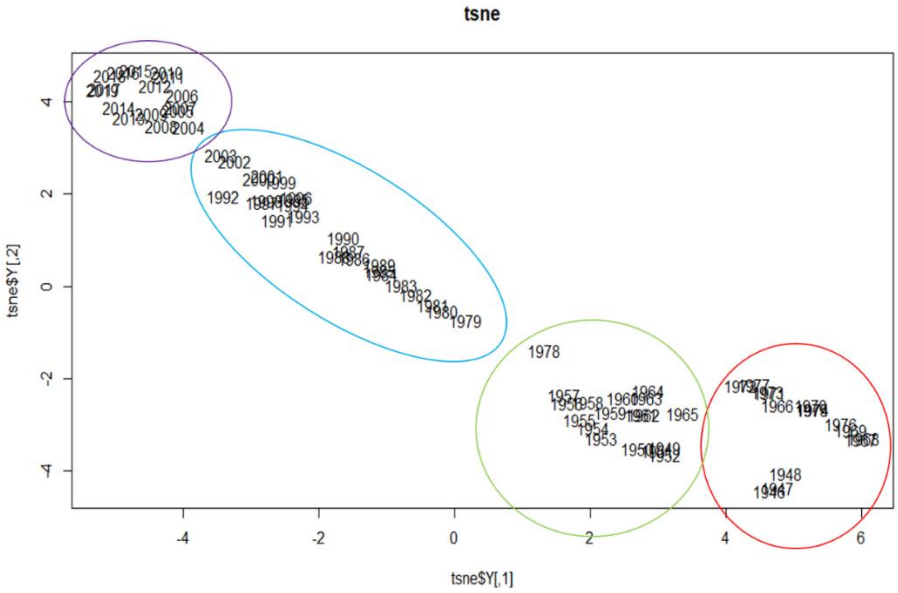

圖 4-54、t-SNE 分四群結果

綜合對各年詞頻前 500 名雙字詞、各年詞頻前 500 名雙字詞的 Jaccard Index、各年詞頻前 500 名雙字詞的 Yue's Index 進行多種分群的結果，這樣同時考慮了用詞種類和用詞頻率，而使用多種分群方法讓分群結果更為穩健。將人民日報分為四大時期，第一時期：1949～1965、第二時期：1946～1948、1966～1978、第三時期：1979～2003、第四時期：2004～2019。

在前面有提到在白話文中，使用雙字詞進行分析會是較好的選擇，所以在這裡只使用了各年詞頻前 500 名雙字詞、各年詞頻前 500 名雙字詞的 Jaccard Index、各年詞頻前 500 名雙字詞的 Yue's Index 進行多種分群。本文同時也有對各年詞頻前 500 名單字進行分群，但分群結果表現並不良好，因此，也呼應了在白話文中使用雙字詞進行分析會是較好的選擇。

DOI:10.6814/NCCU202101184

第三節　代表詞偵測

將人民日報分為四大時期後，希望能找出關鍵分類變數來進行文章風格分類，因此，運用代表詞偵測來找各時期的代表詞，此處的代表詞是雙字詞，透過將各時期的代表詞當作關鍵分類變數來進行良好的預測。

本研究運用三大指標在代表詞偵測上，三大指標有 TF、TF-IDF、TextRank。選用 TF 的原因是因為當一個字詞是關鍵字時，出現的次數會較多，以各時期前十名 TF 的雙字詞為例子：

表 4-8、四大時期 TF 前十名雙字詞

Rank	第一時期	第二時期	第三時期	第四時期
1	人民	革命	发展	发展
2	我们	我们	中国	中国
3	中国	人民	工作	建设
4	工作	他们	我们	工作
5	生产	群众	建设	经济
6	他们	生产	经济	合作
7	国家	工作	问题	国家
8	进行	学习	国家	我们
9	会议	同志	人民	加强
10	问题	斗争	企业	问题

DOI:10.6814/NCCU202101184

而選用 TF-IDF 的原因是當一個字詞是關鍵字時，詞頻會較高，但同時許多虛字等等詞頻也會很高，因此 TF-IDF 同時考慮了這個詞在其他文件中不出現的逆頻率 IDF。此處以各時期前十 TF-IDF 為例子：

表 4-9、四大時期 TF-IDF 前十名雙字詞

Rank	第一時期	第二時期	第三時期	第四時期
1	人民	革命	发展	发展
2	我们	人民	工作	建设
3	工作	群众	建设	工作
4	生产	我们	我们	合作
5	和平	同志	人民	中国
6	中国	斗争	同志	加强
7	苏联	学习	中国	国家
8	会议	他们	经济	人民
9	国家	干部	问题	群众
10	他们	路线	国家	推进

最後運用 TextRank，TextRank 的概念是給予每個詞一個重要程度並以此作為代表詞的選取依據，此時字詞之間的關聯程度則常以字詞的共現性作為指標，會以移動窗格的方式做計算。此處以各時期前十 TextRank 為例子:

DOI:10.6814/NCCU202101184

表 4-10、四大時期 TextRank 前十名雙字詞

Rank	第一時期	第二時期	第三時期	第四時期
1	人民	革命	发展	发展
2	中国	人民	工作	中国
3	生产	群众	中国	建设
4	工作	生产	建设	工作
5	进行	工作	经济	经济
6	国家	学习	企业	国家
7	美国	干部	国家	合作
8	会议	同志	问题	加强
9	代表	中国	人民	企业
10	问题	进行	生产	问题

經由每個時期的 TF、TF-IDF、TextRank 三者對雙字詞排序，每個時期都取每個指標的前 n 名雙字詞取交集，會考慮變數數量和模型分類準確度決定 n 的值。這是本研究偵測每一時期代表詞的方法，能同時考量到三個指標，每個時期挑選出的代表詞再取聯集當作變數。

DOI:10.6814/NCCU202101184

第四節　四大時期風格分類

本研究提出透過代表詞偵測篩選變數進行四大時期風格分類，第一步是必須先決定 n 的值，因此運用不同的 n 值，使用 Logistic Regression 進行四大時期風格分類，在不同的 n 值上都會進行 500 次模擬過程。此處選擇 Logistic Regression 的原因是希望四大時期風格分類的準確度高的原因是因為代表詞偵測篩選變數，而不是透過模型，因此選用較簡單的模型。使用的訓練集是將人民日報頭版新聞，依照年分將資料隨機抽取 59 年當作訓練集，15 年當作測試集。從圖 4-54 可看出當 n 等於 100，也就是代表詞偵測篩選變數的數量到達 128 個的時候，平均準確率有達到 98.47%。

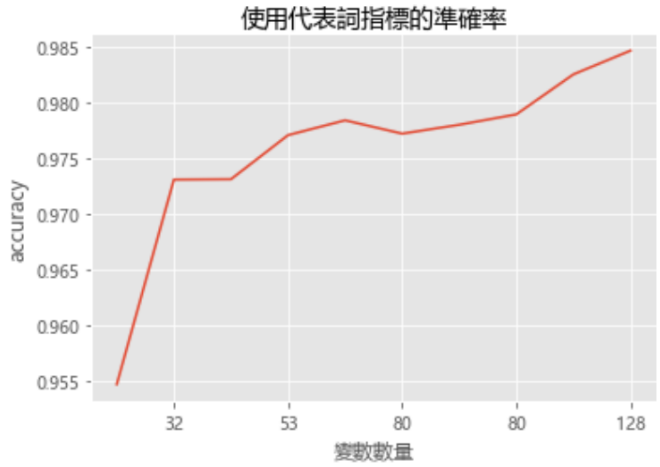

圖 4-55、使用代表詞指標的準確率(1946 年～2019 年)

進一步去察看錯誤預測的時期，圖 4-55 是當 n 等於 100，也就是代表詞偵測篩選變數的數量選取 128 個的時候，使用 Logistic Regression 進行四大時期風格分類，進行 500 次模擬的結果。可看出預測錯誤的都在相鄰兩個時代。

DOI:10.6814/NCCU202101184

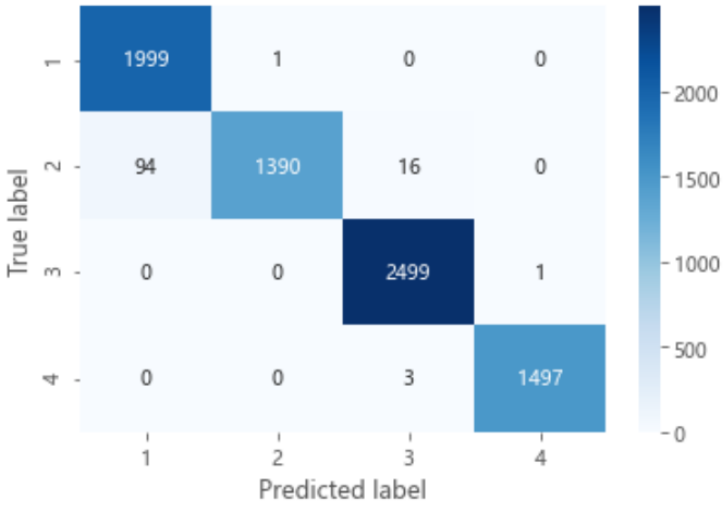

圖 4-56、使用代表詞指標的預測結果(1946 年～2019 年)

從預測錯誤的年分來看 1946 年預測錯誤 82 次、1978 年預測錯誤 16 次、1947 年預測錯誤 12 次、2004 年預測錯誤 3 次、2003 年預測錯誤 1 次、1949 年預測錯誤 1 次。可看出幾乎都是在相鄰兩個時代的交界年分，除了 1946 年、1947 年。

從先前的分析來看，1946～1948 年的用詞與其它年分較不相似，本文認為可能是因為《人民日報》是在 1949 年 8 月 1 日才從華北中央局機關報升格為中共中央機關報。因此，文章風格與其它年分很不一樣，進行文章風格分類時容易分類錯誤，因此拿掉這段時間再次進行文章風格分類，依照年分將資料隨機抽取 56 年當作訓練集，15 年當作測試集，使用 Logistic Regression 進行四大時期風格分類，在不同的 n 值上都會進行 500 次模擬過程。從圖 4-56 可看出當 n 等於 100，也就是代表詞偵測篩選變數的數量到達 128 個的時候，平均準確率有達到 99.40%。除了 Logistic Regression 之外，也有使用機器學習模型：Random Forest 和 XGBoost 去模擬 500 次，而 Random Forest 的平均準確率只有 87.99%，比 Logistic Regression 低了不少，而 XGBoost 的平均準確率雖有 95.43%，但還是比 Logistic Regression

DOI:10.6814/NCCU202101184

差上一些。從上述來看，如果能挑選到關鍵變數，簡單的模型也能表現優良，而較為複雜的模型可能會有過擬合的現象。因此，以下在使用各種降維方法和指標時，會以 Logistic Regression 為模型去比較且運用代表詞偵測篩選變數的變數數量 128 個。

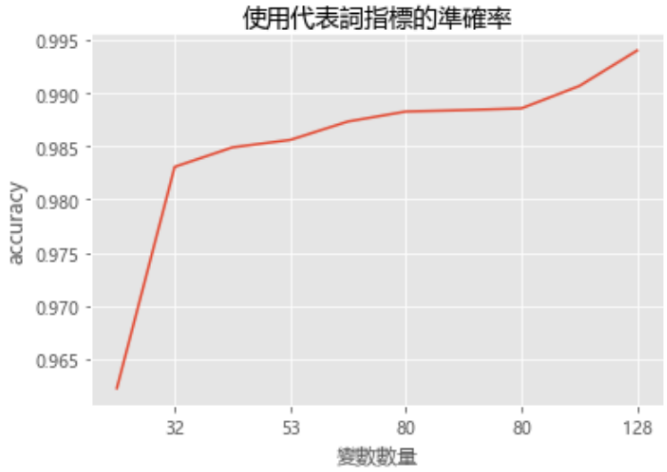

圖 4-57、使用代表詞指標的準確率(1949 年～2019 年)

從上述結果可看出代表詞偵測篩選變數可以有很好的表現，因此跟放入全部雙字詞當作變數去比較，放入全部雙字詞當作變數，總共有 348971 個變數，使用資料同樣是人民日報 1949 年～2019 年，依照年分將資料隨機抽取 56 年當作訓練集，15 年當作測試集，使用 Logistic Regression 進行四大時期風格分類，模擬 500 次，平均準確率可達 99.51%。可看出運用代表詞偵測篩選變數跟放入全部雙字詞當作變數，兩者的準確度差不多，但是放入全部雙字詞的變數數量遠大於代表詞偵測篩選變數的變數數量，這影響到了運算速度，放入全部雙字詞當作變數的運算速度也會遠遠大於代表詞偵測篩選變數。且在解釋性上，本研究提出的方法，代表詞偵測篩選變數也更具有解釋性，因為是偵測各時期的代表詞，因

DOI:10.6814/NCCU202101184

此，可以反映出各時期的時代背景，例如：第一時期的代表詞有出現朝鮮、美國、和平、勝利、侵略等等，反映出了韓戰，而第三時期代表詞有出現發展、建設、經濟等等，反映出了改革開放。而在放入全部雙字詞當作變數時，就不具有這種解釋性。因此，代表詞偵測篩選變數是同時具有運算速度、模型準確度、解釋性三大優點，接下來進一步去跟 PCA 降維、倍數指標、卡方指標等降維方法去比較。作法皆是使用資料是人民日報 1949 年～2019 年，依照年分將資料隨機抽取 56 年當作訓練集，15 年當作測試集，使用 Logistic Regression 進行四大時期風格分類，模擬 500 次

首先比較的方法是 PCA 降維，運用 PCA 降維會受限於訓練集數量，最多只能選擇 56 個主成分。從圖 4-57 跟表 4-11 可以看出準確率最高就出現在選擇 56 個主成分時，平均準確率可達到 97.89%，且累積解釋變異已經達到 100%。相比代表詞偵測篩選變數，準確度較低、降維運算時間較長且難以解釋降維後的主成分。

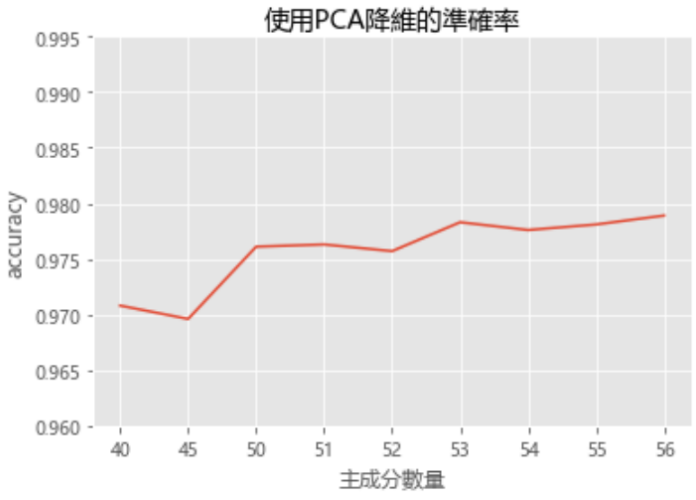

圖 4-58、使用 PCA 降維的準確率

DOI:10.6814/NCCU202101184

表 4-11、使用 PCA 降維的主成分數量、準確率和累積解釋變異

主成分數量	準確率	累積解釋變異
40	97.08%	83.45%
45	96.96%	90.35%
50	97.61%	96.21%
51	97.63%	97.09%
52	97.57%	97.93%
53	97.83%	98.74%
54	97.76%	99.41%
55	97.81%	100%
56	97.89%	100%

接下來比較的方法是經由倍數指標篩選變數，從圖 4-58 跟表 4-12 可以看出隨著倍數變高，變數越少，則準確率越低，而準確率最高就出現在選擇倍數 30 倍的時候，平均準確率可達到 98.01%，但變數多達 1297 個。相比代表詞偵測篩選變數，準確度較低且變數多上許多，運算時間較長且倍數指標的問題在於可能會去除很重要的字詞但它不同的文本之間皆有出現，會使得解釋性下降。

DOI:10.6814/NCCU202101184

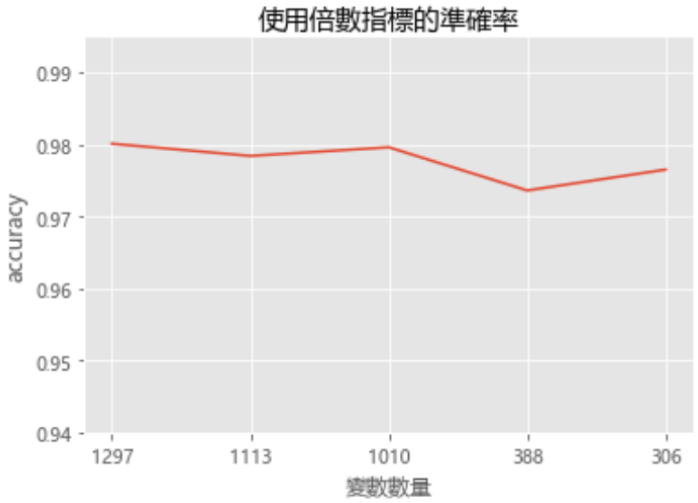

圖 4-59、使用倍數指標的準確率

表 4-12、使用倍數指標的倍數、變數數量和準確率

倍數	變數數量	準確率
30	1297	98.01%
40	1113	97.84%
50	1010	97.96%
500	388	97.36%
1000	306	97.65%

接下來比較的方法是經由卡方指標篩選變數，從圖 4-59 跟表 4-13 可以看出不論 p-value 小於 0.01、0.001、0.0001、0.00001、0.000001 這幾種情況，準確率表現得都差不多，而準確率最高就出現在選擇 p-value 小於 0.000001 的時候，平

DOI:10.6814/NCCU202101184

均準確率可達到 99.38%，但變數多達 135 個。相比代表詞偵測篩選變數，準確度較低且變數也較多，運算時間較長且卡方指標的問題跟倍數指標一樣，在於可能會去除很重要的字詞但它不同的文本之間皆有出現，會使得解釋性下降。

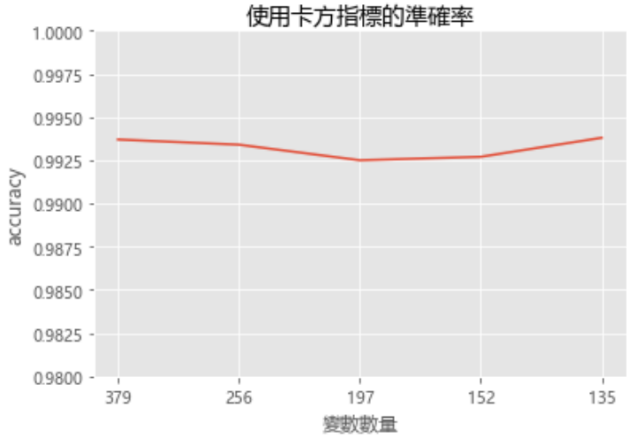

圖 4-60、使用卡方指標的準確率

表 4-13、使用卡方指標的 p-value、變數數量和準確率

p-value	變數數量	準確度
0.01	379	99.37%
0.001	256	99.34%
0.0001	197	99.25%
0.00001	152	99.27%
0.000001	135	99.38%

DOI:10.6814/NCCU202101184

綜上所述，代表詞偵測篩選變數在準確性、解釋性、運算速度三個方面都有更良好的表現，同時能以篩選出的變數去描述當下時代的背景特色。

進一步將代表詞偵測篩選變數運用在月份分類，使用的訓練集是將 1949～2019 年的人民日報頭版新聞，依照月份將資料隨機抽取 681 個月當作訓練集，171 個月當作測試集，運用 Logistic Regression 進行分類，準確率接近 0.93，也有很好的表現。

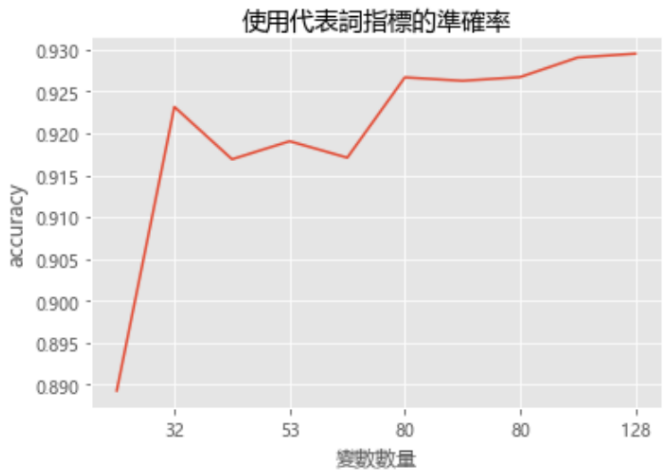

圖 4-61、使用代表詞指標的準確率(1949 年～2019 年月份分類)

DOI:10.6814/NCCU202101184

第五章　結論與建議

第一節　結論

本研究以探討在不同時空背景下同個文本的文本風格變化和運用新的變數選擇方法進行文本風格分類為研究目標，而本研究以《人民日報》1949～2019 年頭版報導為分析的文本。本研究從單／雙字詞、標點符號、虛字這三個角度出發，發現在不同時間下的用字遣詞有差異，而標點符號和虛字在各年之間並沒有太大的差異。雖然在單字和雙字詞的分析中，不同時間下的用字遣詞有差異，但因為人民日報是白話文，所以使用雙字詞進行分析會是較好的選擇，因為白話文多以雙字詞和多字詞來表達。

接著對各年詞頻前 500 名的雙字詞、Jaccard Index、Yue's Index 進行多種分群後，透過投票的概念去決定分群結果，這樣同時考慮了用詞種類和用詞頻率，而使用多種分群方法讓分群結果更為穩健，分群結果顯示人民日報分為四大時期：第一時期：1949～1965、第二時期：1946～1948、1966～1978、第三時期：1979～2003、第四時期：2004～2019。從分群結果可看出跟中國的變化是相符的，這四個時期或可命名為建國、文革、改革開放、現代化。本研究從旁觀者的角度出發，運用群集分析將人民日報劃分成不同時期，與專家學者認定的時期一致。因此，本研究認為文字風格可以反映中國的轉變。

在運用變數選擇方法進行文本風格分類的部分，本研究提出了一種新的變數選擇方法，經由 TF、TF-IDF、TextRank 三個指標去偵測代表詞，當作解釋變數。經由此方法可篩選出各大時期的代表詞當作解釋變數進行分類，結果顯示運用 Logistic Regression 有良好的表現，準確率達到 99.40%，相比之下，Random Forest 和 XGBoost 的表現較差，挑選到關鍵的變數，簡單的模型也能表現優良。本研究提出的代表詞指標，與過去的 PCA、倍數指標、卡方指標相比，在準確性、解

DOI:10.6814/NCCU202101184

釋性、運算速度三個方面都有更良好的表現，同時能以篩選出的變數去描述當下時代的背景特色。本文從人民日報四大時期年分分類進一步推展至人民日報四大時期月份分類，都有良好的表現。

第二節　建議

本研究是運用文字探勘去分析《人民日報》1949～2019 年的頭版報導，找出在經過時空背景的變化後，同一個文本的文本風格的變化，而在研究的過程中有去比較《人民日報》與台灣的四大報(蘋果日報、自由時報、中國時報、聯合報)之間的差異，例如《人民日報》和台灣的四大報在引號上的使用就有所不同。因此在後續的研究上可以進一步去比較《人民日報》與台灣的四大報(蘋果日報、自由時報、中國時報、聯合報)之間在用字遣詞、標點符號、虛字等等的的差異。

本文是以《人民日報》用字遣詞的的變化進行風格分析，而在後續可以加入字詞之間的相似程度、字詞在文章中的位置和字／詞意進行風格分析。每一時期相似字詞使用的轉變、字詞在文章中位置的變化和字／詞意能帶來更多的資訊去研究《人民日報》寫作風格的轉變。

本文提出的維度縮減的方法是運用多種關鍵詞偵測指標去篩選變數，選出各時期的代表詞作為分類的解釋變數，而在後續對此方法的改進可以加入更多指標去選擇變數，例如卡方值、RAKE 等等，如此可能可以更好的選出代表詞，使準確率和解釋性都再度提高。本文最後運用代表詞指標篩選出的變數，可以使用生態變遷的想法，對各時期代表詞進行分群，或可區分為常用、新生、滅絕雙字詞三種類別。

DOI:10.6814/NCCU202101184

參考文獻

一、中文文獻

1. 王宇（2012）。「框架視野下的食品安全報導——以《人民日報》近 10 年的報導為例」，《*現代傳播: 中國傳媒大學學報*》，*34*(2)，頁 43-47。

2. 曲青山（2021）。《中國共產黨百年輝煌》。北京市: 人民出版社。

3. 余清祥、葉昱廷（2020）。「以文字探勘技術分析臺灣四大報文字風格」，《*數位典藏與數位人文*》，6，頁 69-96。

4. 於韜、王洪岩（2018）。「基於 TF-IDF 算法的文本信息提取」，《*科技視界*》，16，頁 117-118。

5. 林志軒（2020）。「維度縮減於文本風格之應用研究」，*政治大學統計學系學位論文*，頁 1-51。

6. 林晏辰（2020）。「中文關鍵詞偵測的探討」，*政治大學統計學系學位論文*，頁 1-62。

7. 胡適（2016）。《紅樓夢考證》。北京市: 北京出版社。

8. 姚興山（2009）。「基於詞頻的中文文本分類研究」，《*現代情報*》，*29*(2)，頁 179-181。

9. 孫曉明、馬少平（2001）。「基於寫作風格的作者識別」，《*中國中文信息學會第五屆全國會員代表大會暨成立二十週年學術會議論文集》，北京：清華大學出版社。*

10. 陳鳳芝（2003）。「中西方思維差異與寫作風格對比分析」，《*三峽大學學報： 人文社會科學版*》，*25*(3)，頁 95-97。

11. 夏天（2013）。「詞語位置加權 TextRank 的關鍵詞抽取研究」，《*現代圖書情*

DOI:10.6814/NCCU202101184

報技術》，9，頁 30-34。

12. 徐超（2017）。「《人民日報》社論詞彙統計與分析」，《*采寫編*》，(3)，頁 144-145。

13. 張運良、朱禮軍、喬曉東、張全（2009）。「基於句類特徵的作者寫作風格分類研究」，《*計算機工程與應用*》，*45*(22)，頁 129-131。

14. 黃秋林、吳本虎（2009）。「政治隱喻的歷時分析——基於《人民日報》(1978—2007) 兩會社論的研究」，《*語言教學與研究*》，(5)，頁 91-96。

二、英文文獻

1. Archer, J. and Jockers, M.L. (2016). The Bestseller Code, New York: St. Martin's Press.

2. Beliga, S. (2014). "Keyword extraction: a review of methods and approaches." University of Rijeka, Department of Informatics, Rijeka, 1-9.

3. Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005). "Text classification using machine learning techniques." WSEAS transactions on computers, 4(8), 966-974.

4. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). An Introduction to Statistical Learning: With Applications in R, Berlin: Springer

5. Liu, F., Pennell, D., Liu, F., and Liu, Y. (2009). "Unsupervised approaches for automatic keyword extraction using meeting transcripts." In Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics (pp. 620-628).

6. Matsuo, Y., and Ishizuka, M. (2004). "Keyword extraction from a single document

DOI:10.6814/NCCU202101184

using word co-occurrence statistical information." International Journal on Artificial Intelligence Tools, 13(01), 157-169.

7. Puglisi, R. (2006). "*Being The New York Times: Thepolitical Behaviour Of A Newspaper* (No. 20)." Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.

8. Pervaiz, F., Pervaiz, M., Rehman, N. A., & Saif, U. (2012). "FluBreaks: early epidemic detection from Google flu trends." Journal of medical Internet research, 14(5), e125.

9. Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). "Automatic keyword extraction from individual documents." Text mining: applications and theory, 1, 1-20.

10. King, T., "80 Percent of Your Data Will Be Unstructured in Five Years.", Retrieved June 15, 2021, from: https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/

11. Zhai, Y., Song, W., Liu, X., Liu, L., and Zhao, X. (2018). "A chi-square statistics based feature selection method in text classification." In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 160-163). IEEE.

DOI:10.6814/NCCU202101184