國立政治大學商學院統計研究所

碩士學位論文

關鍵詞偵測方法的比較與應用

The Application of Keywords Extraction

指導教授：鄭文惠 博士

余清祥 博士

研究生：許承恩 撰

中華民國 一〇八年七月

DOI:10.6814/NCCU201900953

摘要

近年來由於文本被大量數位化，使得文字探勘（Text Mining）成為熱門研究領域，愈來愈多研究藉由量化技術找出文字涵意，提供專家意見不同角度的語意解讀。文本在經過結構化（Structurization）後，根據不同需求如關鍵詞擷取、尋找潛在文本主題、情感分析、輿情分析等，建立統計及機器學習等數位模型。其中關鍵詞擷取可用於解讀作者想法、提升閱讀效率、掌握寫作風格以及文章出版時空背景的變化。本研究也以決定關鍵詞為研究目標，除了提出一種非監督學習的統計方法，也使用中文文本評估新方法與幾種常見關鍵詞偵測方法，包括網路流行的 TF-IDF (Term Frequency Inverse Document Frequency；詞頻與文本頻率)、統計分析的羅吉斯迴歸（Logistic Regression）、常見的機器學習模型。實證分析採用《人民日報》、《新青年雜誌》兩個白話文的文本，其中《人民日報》為 1971-1989 年與人權有關的 514 篇報導，《新青年》則是第七卷（1919 年）、第八卷（1920 年），這些文本的篇幅大約都介於 40～60 萬字。先由人文學者標記出各文本的關鍵詞，將其視為標準答案，再套用上述三種方法選取可能的關鍵詞，再比較上述方法與專家意見的差異及準確率；另外，我們也將比較人工挑選、自動挑選關鍵詞的差異，並探索兼具兩種方法優點的可能。

關鍵字：文字探勘、關鍵字擷取、數位人文、機器學習、詞頻與文本頻率

DOI:10.6814/NCCU201900953

Abstract

Text Mining has become one of the popular research areas after the IBM proposed the term Big Data in 2010. Since then many texts are being digitalized and more scholars are devoted in developing quantitative tools for giving texts semantic meaning without the help of human experts. This greatly increases the efficiency of reading a hugh amount of texts provided that the texts are properly structurized. The structurization of texts includes quite a few steps, such as keyword extraction and sentiment analysis. The keyword extraction is critical and the keywords can be used to summarize an article and compare two authors' writing styles.

The goal of this study is to propose a new unsupervised method for extracting keywords and compare it to some frequently used methods, including term frequency inverse document frequency (TF-IDF), logistic regression, machine learning models. In the empirical analysis, we considered three modern Chinese texts, one from *People's Daily* (514 articles in 1971-1989) and two from *New Youth Magazine* (volumes 7 and 8 in 1919-1920). The numbers of words in all texts are approximately 400,000 to 600,000. We asked historical scholars to pick up keywords from these three texts and treat them as the true keywords. Then, we applied different keyword extraction methods to these texts and compared their results. We found that the proposed method has the best performance among all supervised methods and it is competitive to the supervised methods.

Keyword: Text mining, Keyword extraction, Digital humanities, Machine learning, Term Frequency Inverse Document Frequency

DOI:10.6814/NCCU201900953

目錄

DOI:10.6814/NCCU201900953

DOI:10.6814/NCCU201900953

圖目錄

DOI:10.6814/NCCU201900953

表目錄

DOI:10.6814/NCCU201900953

第一章 緒論

第一節 研究動機

大數據時代的來臨，使現代人的生活已和資訊及其應用密不可分。隨著電腦等科技的快速成長，近年來資料蒐集及交換傳遞更為普及系統化，分析技術有更大的發揮空間，使得人們更能發掘資訊背後的價值。目前已有許多重要應用奠基於大數據科技的發展，例如：智能家電(Smart Home)或是家庭自動化(Home Automation)指的是家庭中的建築自動化，包括能夠控制燈光、窗戶、溫濕度、及家庭保全，智能家電因為物聯網(Internet of Things)等產業興起得以全面實行，讓人們的生活更加便利。在環境應用上也有基於物聯網提出之小範圍即時監測空氣品質系統（何昱鋒，2019），數據分析在這些應用扮演重要角色，藉由彙總整理龐大的數據，協助決策者滿足我們各類型的需求。

依照資料可否量化，通常將資料分成為結構型資料(Structured data)及非結構型資料(Unstructured data)。結構型資料泛指資料已有固定格式，每一個觀測值以預定格式輸入資料格式，例如在搜集顧客聯絡方式前就已預設需要寫入姓名、電話以及電子郵件三個欄位，分析時需先調整不符合預設格式的資料。非結構化資料無特定格式，依照使用者及使用目的決定，舉凡圖片、影片、聲音記錄以及文字等都是屬於非結構型資料。結構型資料的優點在於資訊明確、整理方便、查詢速度快及儲存空間小，且後續分析處理也較容易。非結構化資料一般較為豐富的訊息，像是書籍、報章雜誌、文宣傳單與信件往返等文字資料，但由於非結構資料的格式不定，需先經過結構化的步驟將轉換為結構型資料才能進行後續分析，這個步驟至今尚無統一的作法。

DOI:10.6814/NCCU201900953

結構型資料的分析技術大致已趨近完善，可以使用的理論與方法較多，面對較不尋常狀況時的因應對策也較為完整，像是遺漏值及極端值處理、資料正規化、模型挑選等。相對而言，非結構化資料分析方法較為侷限，解讀時需要搭配專業領域的知識，一窺隱藏於資料背後的深層意義。不過，近年非結構型資料的分析有顯著發展，得益於神經網路發展迅速，圖像以及語音辨識等領域而有重大突破，以文字分析為例，已有許多逐漸成熟的技術，如自動摘要、文件檢索、機器翻譯、問答系統等。

在這些研究中，關鍵詞擷取(Keyword extraction)扮演了重要角色，像是網路上透過搜尋引擎(Search Engine)找尋目標，輸入適切的關鍵字詞可提高效率，避免耗費查閱不相關網頁的麻煩。文字分析也是如此，關鍵詞除了可用於解讀作者想法、提升閱讀效率之外，也可做為分類文本的依據，掌握寫作風格以及文章出版時空背景的變化。關鍵詞可視為表達文本意義的最小單位，如何擷取可以說是文字分析的第一步，也是極為重要的步驟。然而，以人工選取關鍵字詞十分費時，關鍵詞又與研究目的、文本屬性有關，需要熟稔該領域知識的專家學者協助。或許這也是關鍵詞擷取進行不易的主因。

第二節 研究目的

有鑑於關鍵詞的重要，本文以自動化偵測關鍵詞為研究目標，希冀發展出非監督學習(Unsupervised Learning)的統計方法，再與專家意見結合，促進數位人文的研究。由於關鍵詞取決於語言及文本，本文以中文的白話文為探討目標，選取《新青年》第七卷及第八卷，以及人民日報 1971 年至 1989 年與人權有關的報導，先由人文學者標示關鍵詞，再代入羅吉斯迴歸、機器學習、本文提出方法，經由

DOI:10.6814/NCCU201900953

交叉驗證(Cross-validation)、根據準確率及召回率等準則，比較哪一種方法得出的結果與人文學者較為接近。

文字屬於非結構型資料的一種，中文資料的結構化（資料前處理）過程通常包括以下幾個步驟：去除標點符號與停止詞、斷詞、選擇能夠表達全文大意的關鍵詞（或是關鍵變數）。有不少研究以斷詞為研究目標，如郭益豪(2013)提出基於 N-Gram 的改良式斷詞法，結合潛在語意分析(LSA, Latent Semantic Analysis)建構註解參數集，改善單文件斷詞的限制。謝孟樺(2018)則是藉由建立共現詞詞典，在斷詞時透過共現詞詞典考慮前後文的共現詞關係來提升斷詞的準確度。關鍵詞以往被認為決定單篇文章中最精華的部分，例如在新聞下方由記者標記出關鍵字，使讀者能快速理解新聞內容，亦可透過相同關鍵字搜尋類似的新聞報導資訊。除了標記單篇文章之關鍵詞，多文章文本如何進行關鍵字擷取亦是需要深入研究的部分。

以決定關鍵詞為目的之研究較少，目前較為常見的作法是先考量「結巴」斷詞，再以 TF-IDF(Term Frequency-Inverse Document Frequency)從斷詞中挑選關鍵詞，然而，TF-IDF 只有給出關鍵詞之間的排名，並沒有確切的門檻作為篩選的依據，隨著關鍵詞選取個數增加，真實關鍵字被選中的比例(Recall)也逐漸上升，但與此同時準確率(Precision)卻逐步下降，無法同時兼顧，此外，對於相同性質的系列文本，也容易因為 TF-IDF 去除噪音(noise)的機制使重要關鍵詞的排名順序較低，因此本文希望結合多重群集偵測(multiple cluster detection)以及 alpha 耗費函數(alpha spending function)概念發展非監督式學習方法。本文以探討偵測關鍵詞為目標，提出非監督學習的判斷關鍵詞方法，並與現在常見方法做為比較對象，使用《新青年》及《人民日報》做為實證分析的文本、人文學者挑選關鍵詞

DOI:10.6814/NCCU201900953

作為判斷正確與否依據，找出較佳的關鍵詞偵測方法，並針對各模型以及人文學者所挑之關鍵詞的差異性進行討論。

透過文獻探討與實證分析，本文研究在於提升關鍵詞選取的正確性，透過在詞頻以及文件頻率基礎上發掘更多詞彙的統計數據、建立結構型資料並基於關鍵詞群聚現象發展出非監督式學習的選取方法。本文第二章呈現文字處理的流程、工具與相關的文獻探討；第三章呈現本文研究方法，介紹方法流程及想法；第四章以電腦模擬為輔助討論最佳參數的設定；第五章呈現本次論文使用之實驗文本及不同關鍵詞選取方法比較之結果；最後第六章則會探討本文方法的研究限制與未來研究的發展。

DOI:10.6814/NCCU201900953

第二章 文獻探討

在自然語言以及文字探勘中，擷取文章中的關鍵詞是重要的環節，可用於解讀作者想法、提升閱讀效率、作為分類文本依據、掌握寫作風格與文章出版時空背景的變化等應用。關鍵詞擷取的另一個概念在於資訊含量，即使有少量重要關鍵詞，因資訊含量過少仍無法使讀者了解文章的全貌，因此如何找出足以代表整體文章資訊的關鍵詞亦是關鍵詞擷取的目標。關鍵詞選取方式以人工介入程度可簡單分為人工選取、模型選取以及電腦自動化選取。人工選取是由專家學者在看過所有的文章後挑選出符合該篇文章內容的關鍵詞；模型選取由專家先挑選出部分關鍵詞，再將結構化後的文章搭配人文學者所挑部分關鍵詞一同放入統計模型，由模型找出剩餘關鍵詞；電腦自動化選取則是不加入專家學者所挑關鍵詞，僅以結構化資料進行選取，如表 二-1 所列，人工標記與模型各自優勢且皆須專家學者協助，本文嘗試建立一套關鍵詞自動擷取流程，在不借助人工標記的情況下挑選關鍵詞，並以專家學者所選之關鍵詞作為判斷挑選是否成功之依據。

DOI:10.6814/NCCU201900953

表 二-1 人工標記與模型選取比較表

	人工標記	模型選取
挑選方式	由專家學者看過所有文章後挑出	標記部分關鍵詞再以統計模型找出剩餘關鍵詞
優點	1. 精確判斷關鍵詞 2. 考慮到文本內容	花費時間較少
缺點	1. 人工選取關鍵詞十分費時 2. 需熟稔該領域知識的專家學者協助 3. 挑選標準因人而異	1. 需將資料進行結構化、變數選取沒有固定標準 2. 各模型所挑關鍵詞未必相同
使用模型	無	羅吉斯迴歸、LDA、SVM

第一節 斷詞系統

在中文裡，「字」與「詞」是兩個經常被混合用在一起的概念，黃居仁(2005)談到「字」為中文書寫中最小的單位，在書寫文字時，一次只會寫一個字；而「詞」則是表達語言意義中最小的獨立單位，例如：「公司」在文章中出現時，可以清楚地被讀者理解，但若只有提及「公」或者「司」時，並不能傳達作者想表達的意思。在這個例子中可以清楚區分詞與字的差別，只有當若干個單獨字合在一起能被辨認出意思時，才會稱之為詞。

要解構一段句子時，必須先將句子中的「詞」標記出來才能理解其代表的含義。不同於英文會將每個詞以空格分開，中文在表達一句話時並不會刻意將每個詞標記出來，因此必須借助斷詞系統將句子切割成若干個有意義的詞。然而，詞與詞之間的斷點沒有正確答案，以全台大停電為例，若斷句為全台/大/停電，表示全台灣發生了大規模的停電事件，但如果改成全台大/停電，則表示在台大校園發生了停電的事件。由此可知，不同的斷詞結果對於文章理解也將有所差異。

DOI:10.6814/NCCU201900953

另一個在斷詞中會遇到的問題，系統無法辨認新詞，如網路用語「魯蛇」、「慶記」等，亦或是表情符號「XD」、「：）」，這些都不是曾經出現在字典裡的詞，但因為大眾頻繁使用，使得新的詞彙自然而然融入進生活中，如何辨識為之詞成為另一個中文斷詞中的研究目標，藉由 N-gram 的方法協助整理出現頻率高的詞彙，並由人工判斷是否成詞，而謝博行(2013)提出的局部最長連續共同子序列與新詞組收集等方法則能幫助未知詞擷取的自動化。

第二節 TF-IDF

TF-IDF(Term Frequency–Inverse Document Frequency)為文字分析中常見的關鍵詞選取方法，最早由 Salton (1975)在文章中提出在一文本中(包含多篇文章)，關鍵詞具有「只在特定文章中高頻出現」的特性。而 TF-IDF 就是基於這樣的想法發展出對文字的重要程度進行量化的方法。

TF-IDF 包含了兩個部分：詞頻以及逆向文件頻率。詞頻表示在文章中，詞出現的頻繁程度，若詞出現的頻率高，該詞是關鍵詞也愈高。然而，並非只有關鍵詞會有高頻的特性，日常生活中的常用語通常也會頻繁出現在文章中，例如：「然後」、「就是」等詞。因此，如果只是以高頻詞當作關鍵詞的篩選標準仍不夠準確。逆向文件頻率考慮到上述的狀況，若單一詞彙在愈多文章中出現過，則 IDF 數值愈低。數學式(1)表第i個辭彙出現在第j份文件的頻率，數學式(2)表第i個詞彙在各文章出現比例的倒數，若詞彙在多個文章普遍出現，則 IDF 值愈低。TF-IDF 即是由兩個指標相乘而來，同時考慮到關鍵詞的高頻特性以及排除常用詞。TF-IDF 值量化了詞彙在文章中的重要程度，過數值愈高重要性也愈高，透過訂定數值門檻，則可決定詞彙是否為關鍵詞。

DOI:10.6814/NCCU201900953

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$idf_j = log \frac{文本數}{出現過j詞彙文本數} \quad (2)$$

儘管 TF-IDF 的設計簡單且直覺，透過詞頻找出潛在關鍵詞並利用逆文件頻率降低非關鍵詞的干擾，然而 IDF 的設計對於只考慮詞彙是否曾經出現於文章中，對文章本身的特性沒有一同考慮，例如討論教育的多篇文章中，關於教育的相關詞彙出現頻率高且廣泛出現每篇文章，但由於 IDF 的計算方式，這些詞彙的重要程度將降低許多，使關鍵詞選取的準確率下降。有許多學者針對計算權重的方式進行改良，如 Peng 等人(2014)針對不同性質的文章提出對 IDF 的改進，黃培軒(2018)也考慮了文章長度、IDF 數值門檻等參數對關鍵詞進行篩選。

第三節 t-SNE 降維法

處在資訊爆炸的時代，搜集資料已是生活中不可或缺的一部份，而資訊量隨著科技進步也愈來愈多，不只是觀測量(observation)的增加，特徵(feature)的選取也愈來愈多，雖然資訊量愈多可以幫助我們更了解事情的樣貌，但在資料視覺化上卻愈加困難，例如在學校健康檢查時，會簡單紀錄身高、體重，若只有這兩項變數，很容易可將全校同學的身高、體重以散布圖(Scatter plot)的形式呈現，但當紀錄的數值加上血型、血壓、心跳、視力等其他身體數值，雖然可以更清楚了解該同學的健康狀況，但相對的要將同學們的各項數值視覺化也將更複雜。利用維度縮減的方式，不僅可以克服不易視覺化的問題，同時也能減少資料的儲存空間，加速模型建立速度。

DOI:10.6814/NCCU201900953

t-SNE 是由 SNE 發展而來，由 Hinton 與 Roweis 於 2002 年提出，藉由在高維度以及低維度的數據建立資料間相似機率作為依據，使高維空間中距離較遠的資料點在低維空間中依舊保持著較遠的距離。

令x_i與x_j表示在高維空間中兩相異點，並將高維度上兩點間的歐式距離轉換為這兩點的相似度$p(j|i)$，直觀上可以想像為以x_i為中心，利用常態分配來決定其他點成為x_i鄰近點的機率，當x_i與x_j之間的歐式距離愈遠時，x_j成為x_i鄰近點的機率就愈小。由於只針對相異兩點的相似度進行探討，因此假設$p(i|i) = 0$。考慮到高維資料中x_i周圍區域的分散程度皆不盡相同，σ_i^2的設定也將隨之改變，若x_i的鄰近地區資料點較群聚，選擇較小的σ_i^2將更貼近資料的真實情況。

$$p(j|i) = \frac{\exp(\frac{-\|x_i - x_j\|^2}{2{\sigma_i}^2})}{\sum_{k \neq i} \exp(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2})}$$

為了能夠得到適合的σ_i^2，SNE 引入了困惑度(perplexity)使用。當固定x_i後便能求出剩餘的點與x_i之間的$p(j|i)$值，並形成一個分配Pi與其對應的熵(Entropy)，藉由熵與困惑度之間的關係可知，在給定一個合適的困惑度下，即可計算出合適的σ_i^2。

$$Prep(Pi) = 2^{H(Pi)}$$

其中，$H(Pi)$為分配Pi的熵

$$H(Pi) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

DOI:10.6814/NCCU201900953

另一方面同樣也在低維空間中建立與資料點y_i (對應x_i) 、y_j (對應x_j)，並定義低維空間下資料點的相似度$q(j|i)$。在低維空間中，將所有的σ_i^2設定為$\frac{1}{\sqrt{2}}$，以下是$q(j|i)$的表達式：

$$q(j|i) = \frac{\exp(-\|x_i - x_j\|^2)}{\sum_{k \neq i} \exp(-\|x_i - x_j\|^2)}$$

理論上當高維空間資料完美映射到低維空間時，$p(j|i) = q(j|i)$，換言之，在實際情況中的目標便是盡可能地使P_i與Q_i這兩個分配接近。其中P_i表示在給定x_i的條件下，所有$p(j|i)$集合而成的分配；Q_i表示在給定y_i為中心的條件下，所有$q(j|i)$集合而成的分配。透過 KL 散度（Kullback-Leibler divergence）來表達這兩個分配差異的程度，並以最小化 KL 散度為目標，其目標函數定義如下：

$$C = KL(P\|Q) = \sum_{\mathrm{i}} \sum_{j} p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$

利用梯度下降法(gradient descent)對目標函數求梯度可得：

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

由 KL 散度的計算方式可以發現，KL 散度並不是對稱的，換句話說以低維距離小的點代表高維距離大的點所造成的損失(cost)較大，反之，低維距離大的點代表高維距離小的點所造成的損失(cost)較小，這使得 SNE 較傾向於保留資料中的局部特徵。

Maaten 和 Hinton 在 2008 年提出了 SNE 的改良版 t-SNE，藉由重新定義高維資料點相似度p_{ij}以及改變低維空間中計算相似度q_{ij}為自由度為 1 的 t 分配，使 t-SNE 解決了$p_{i|j} \neq p_{j|i}$的問題，且因為 t 分配的厚尾特性，使資料間的擁擠問題獲得解決。

DOI:10.6814/NCCU201900953

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

使用自由度為 1 的 t 分配重新定義q_{ij}：

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

t-SNE 較 SNE 而言改善了資料不對稱以及降維後造成資料壅擠的問題，藉由引入 t 分配作為低維空間中衡量相似度的指標，不但降低了計算難度，效果也較以前更佳，但缺點為只能將維度縮減至二到三維，且降維後資料點距離本身並不具有意義，因此主要應用僅局限於視覺化。

第四節 文本介紹

本文以《新青年》與《人民日報》作為分析對象，兩者皆為中文報章雜誌類型文本，新青年使用第七、八卷作為主要分析內容，人民日報則是 1971-1989 年間與人權有關之報導，以下將分述各文本特色。

一、 《人民日報》

《人民日報》是中國共產黨中央委員會的機關報，與新華社及中國中央電視台，並列為中國共產黨和中華人民共和國政府官方三大傳媒機構，為中華人民共和國第一大報極具影響力。《人民日報》是中國共產黨黨中央向外界表達其觀點的宣傳工具也是世界各國了解中國的重要窗口。作為和人民溝通的橋樑，人民日報準確地傳達中共當局最新的政令，同時也報導國內外大事，反映人民的意願和需求。網路上已有許多可供下載《人民日報》的網站，但多為部分報導或者掃描

DOI:10.6814/NCCU201900953

文件而非全部報導，本文以網路爬蟲(web Crawler)方式搜集 1979 至 1989 年的報導，並篩選人權相關報導作為本文之實驗文本。

二、 《新青年》

《新青年》是中國一份具有影響力的革命雜誌，在五四運動期間起到重要作用。自 1915 年 9 月 15 日創刊號至 1922 年 7 月共出 9 卷 54 號，起先由陳獨秀在上海創立、群益書社發行並月刊的形式發行，1923 年 6 月由瞿秋白主持，以季刊形式重新出版《新青年》，並成為中共中央機關的刊物，最後於 1926 年 7 月停刊不再更新。該雜誌發起新文化運動，宣傳倡導「賽先生」（科學，Science）、「德先生」，（民主，Democracy）和新文學。《新青年》展現了雜誌語言變化歷程以及轉變特點，潘艷艷(2015)依照《新青年》所使用的字詞分析各卷呈現的風格變遷，前三卷屬白話文萌芽階段，文章內容仍以文言文為主，後期七至九卷已經可以觀察到五四運動所倡導的白話文已被廣泛使用。除此之外亦提及在《新青年》後期因受蘇維埃共產主義論述語言影響的紅色中文使後期文章產生風格上轉變。本文主要以白話文為主要研究對象，因此選用《新青年》中白話文比例較高的七、八卷，並排除偏向紅色中文的第九卷。此外，我們也將各卷出現次數最高的前 500 名辭彙整理並由人文學者挑選符合各卷之關鍵詞。

DOI:10.6814/NCCU201900953

第三章 研究方法

本章中將介紹主流關鍵詞擷取方法的優點與可改進之處，並提出關鍵詞擷取自動化的方法，過程包含資料結構化、分組與降維、關鍵詞選取步驟及介紹其原理和想法，並以常見的分類指標做為評斷模型優劣的依據。

第一節 TF-IDF 優劣探討

TF-IDF 的算法中，藉由高詞頻標準提取出候選關鍵詞，並輔以逆文件頻率降低常用詞的權重。表 三-1 列出以 TF 及 TF-IDF 降冪排序後最重要前十名關鍵詞，標記紅字者表示真實關鍵詞，TF-IDF 能有效排除常用詞使真實關鍵詞取得較高排行。圖 三-1、圖 三-2 呈現了新青年第八卷中詞頻(TF)與文件頻率(DF)分別對關鍵詞及非關鍵詞做盒狀圖，關鍵詞所對應的詞頻較非關鍵詞高，但非關鍵詞中亦有高頻詞彙，若單以詞頻而論無法有效分離。另一個圖中資訊為不論高低頻詞彙皆含有關鍵詞，只是低頻關鍵詞不易偵測、不確定性高。

TF-IDF 對關鍵詞的假設為關鍵詞的特性為詞彙不頻繁地出現在文章中，但當被提及時，使用的次數相較於非關鍵詞來得多，換句話說，當詞彙具有詞頻(TF)高、文件頻率(DF)低時，此詞彙是關鍵詞的機會也就愈大。圖 三-3 為同樣文件下的詞頻、文件頻率散布圖，由此圖可知，雖然在高詞頻部分有少許標記為紅點的關鍵詞與標記為黑點的非關鍵詞明顯分隔，但大多數的詞彙仍彼此纏繞在一起，無法明確的判定是否為關鍵詞。因此除了高頻關鍵詞外，如何尋找低頻關鍵詞亦是本文目標。

何立行等人(2014)提出將生態研究中物種多樣性(species diversity)的概念套入文字分析中，以詞彙比擬為生物物種(species)，本文延續其概念，將詞彙類比

DOI:10.6814/NCCU201900953

物種，資料變數視為對其之描述，而增加對詞彙選取的變數就如同取得物種更多的生活習性等資訊，雖然物種間的各項訊息不盡相同，但仍可透過某些特性將其區隔，如日行性與夜行性動物、草食性與雜食性動物之分。關鍵詞亦是相同的概念，由多項資訊，如平均數、標準差等統計量綜合衡量，並以 t-SNE 降維法描繪出各個詞彙間相似程度，若詞彙表現出群聚現象，則可代表這些詞彙如同物種般有著某些相同的行為模式。並進一步探討群聚的結果是否即為關鍵詞、非關鍵詞區分一致。

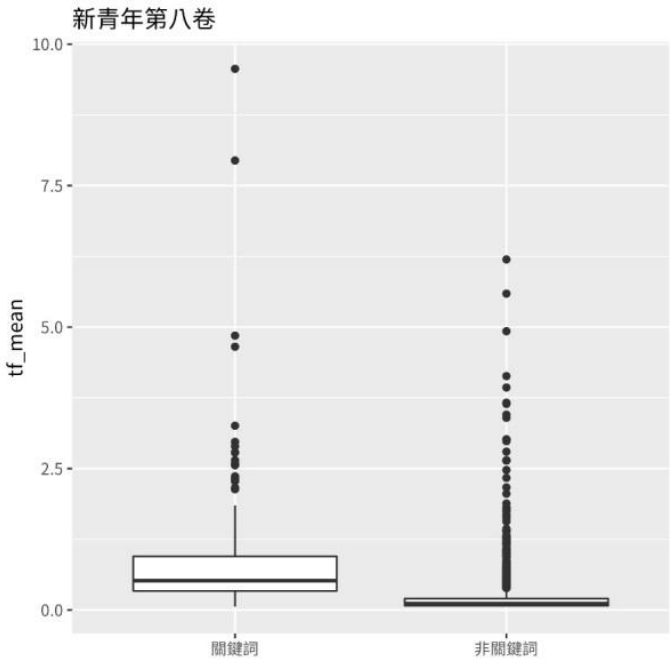

圖 三-1 新青年第八卷 TF 盒狀圖

DOI:10.6814/NCCU201900953

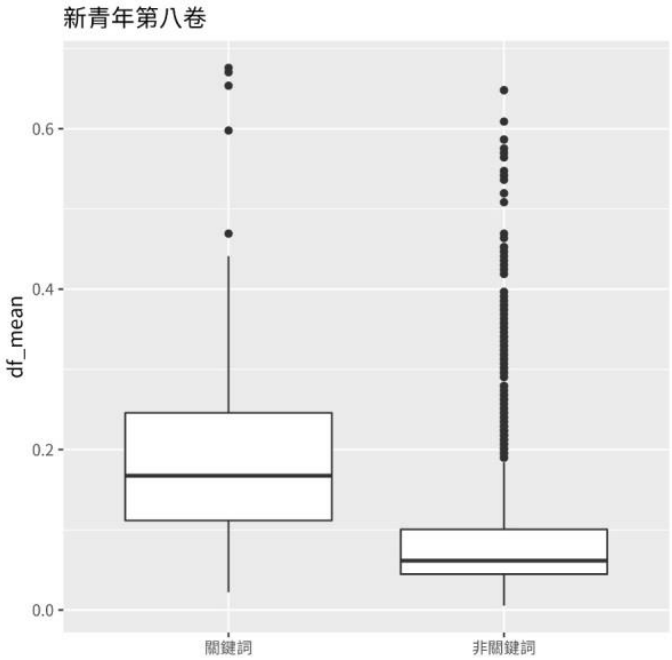

圖 三-2 新青年第八卷 DF 盒狀圖

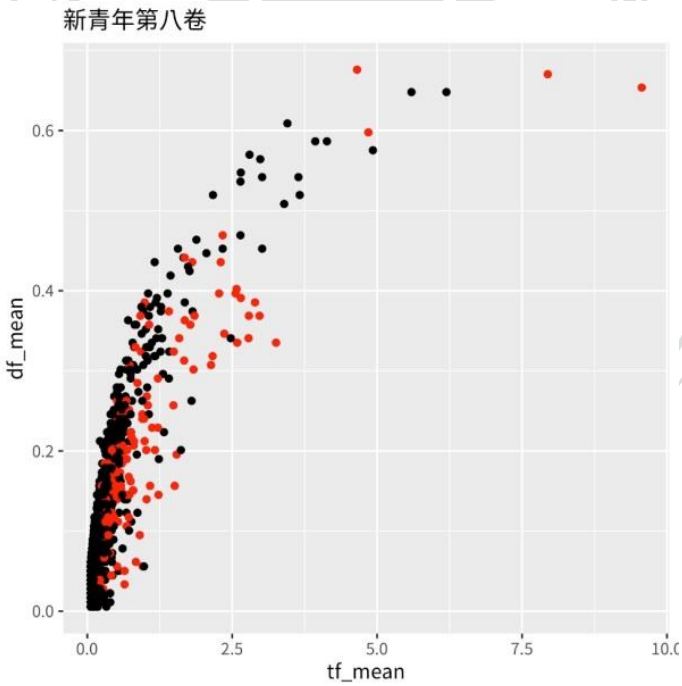

圖 三-3 詞頻對文件頻率散布圖

DOI:10.6814/NCCU201900953

表 三-1 不同排序方法下選取重要詞彙前 10 名

TF 排序		TF-IDF 排序	
他們	對於	他們	聯合會
我們	這種	俄國	勞動
一個	就是	我們	可以
沒有	不能	組織	組合
可以	這個	階級	一個
社會	俄國	國家	社會主義
現在	但是	農民	蘇維埃
所以	因為	工人	政府

第二節 研究方法及流程

本節將詳述本文之研究方法、流程及想法，依序可分為資料結構化、資料分組及降維、關鍵詞選取步驟。圖 三-4 呈現了本研究方法流程圖。

DOI:10.6814/NCCU201900953

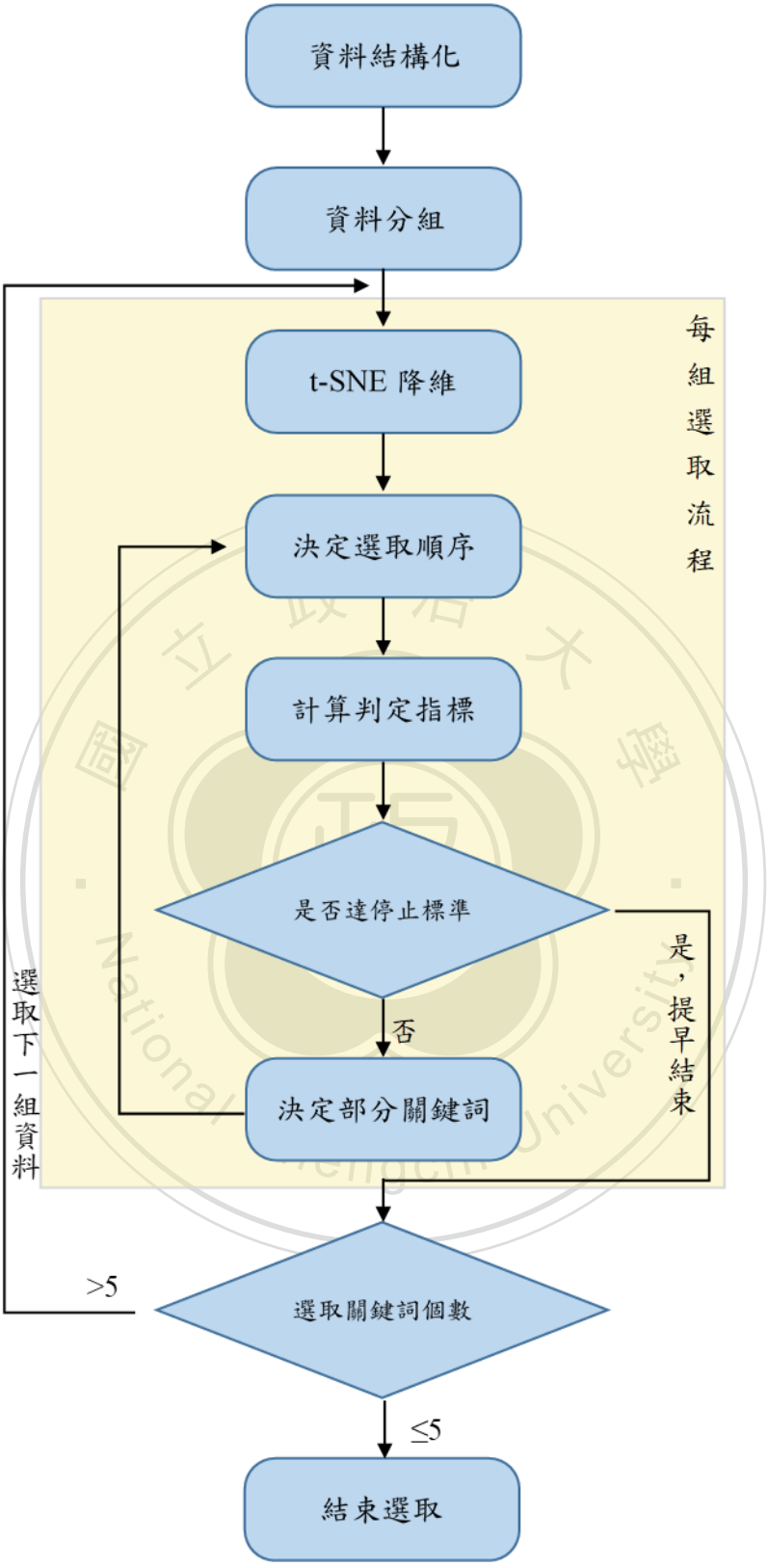

圖 三-4 本研究方法流程圖

DOI:10.6814/NCCU201900953

一、 資料結構化

第一章中提及了結構型資料與非結構型資料的差異，以及一般文字處理的流程，為了使文字資料能夠進行後續的分析，需將文字轉為結構型資料。本節將詳述資料結構化的過程。第二章中提及了 TF-IDF 的相關介紹，主要欄位有詞彙於每篇文章中出現的頻率和一系列文章中出現的比例，本文希望在 TF-IDF 的基礎上找出更多統計量，於是針對詞頻和文件頻率的取得方式作出調整。以 TF 表示詞彙在個文本中出現的次數；DF 表示詞彙是否出現於文章中，若出現則紀錄為 1，反之 0。在這樣的設定下，針對每個詞彙的 TF 進行平均數、標準差、中位數、四分位距、全距、變異係數以及總和的計算。對 DF 的平均數、標準差、中位數、四分位距、變異係數及逆文件頻率(IDF)進行計算，表 三-2 記錄了各欄位名稱以及對應統計量。原始文字在經過斷詞後對個變數進行計算後即可以圖 三-5 的結構形成擁有嚴謹資料格式的表格。

DOI:10.6814/NCCU201900953

表 三-2 變數對照表

詞頻(TF)		文件頻率(DF)	
平均數	tf_mean	平均數	df_mean
標準差	tf_sd	標準差	df_sd
中位數	tf_median	中位數	df_median
四分位距	tf_iqr	四分位距	df_iqr
全距	tf_range	逆文件頻率	df_idf
變異係數	tf_cv	變異係數	df_cv
總和	tf_sum		

word	df_sd	df_mean	df_median	df_iqr	df_cv	df_idf	tf_sd	tf_mean	tf_median	tf_iqr	tf_range	tf_cv	tf_sum	ci_lower	ci_upper	ci_range
他們	0.47714707	0.65363128	1	1	1.36987384	0.41670118	18.1708513	9.56424581	2	10	112	0.526351	1712	6.56498792	12.5635037	5.99851578
我們	0.47138973	0.67039106	1	1	1.42215883	0.39159526	14.6439851	7.94413408	2	8	116	0.54248444	1422	5.5270163	10.3612519	4.83423556
一個	0.47891976	0.64804469	1	1	1.35313836	0.42521187	10.2156952	6.19553073	2	7	60	0.60647176	1109	4.50934091	7.88172055	3.37237964
沒有	0.47891976	0.64804469	1	1	1.35313836	0.42521187	9.19795686	5.59217877	2	8	58	0.60798054	1001	4.07397557	7.11038197	3.0364064
可以	0.49566573	0.57541899	1	1	1.16090131	0.54299491	9.18965462	4.9273743	1	6	63	0.5361871	882	3.41054146	6.44420715	3.03366569
社會	0.49172424	0.59776536	1	1	1.2156516	0.50525458	10.3334646	4.84916201	1	4	72	0.46926778	868	3.14353333	6.5547907	3.41125737
現在	0.46932119	0.67597765	1	1	1.44033056	0.38336476	7.23153407	4.65363128	2	6	46	0.64351924	833	3.46000336	5.84725921	2.38725586
所以	0.49382605	0.58659218	1	1	1.18785184	0.52394671	7.7208248	4.13407821	2	4	51	0.53544515	740	2.85968857	5.40846785	2.54877928
對於	0.49382605	0.58659218	1	1	1.18785184	0.52394671	6.5253161	3.93296089	1	5	43	0.60272343	704	2.85590042	5.01002137	2.15412095
這種	0.50101899	0.51955307	1	1	1.03699278	0.64409102	7.69279842	3.66480447	1	3	71	0.47639419	656	2.39504083	4.93456811	2.53952728

圖 三-5 結構化資料示意圖

二、 資料分組及降維

將結構化完成的資料依照文件頻率平均值(df_mean)降冪排列並以 50 詞為單位分組，前文曾提及關鍵詞中亦有高詞頻低詞頻之分，透過分組的方式能確保不論詞頻高低皆有機會被選取。分組同時也是將詞彙進行初步的分類。本文以文件頻率平均值(df_mean)降冪排序而非詞頻平均值(tf_mean)的原因在於透過圖 三-2 可知，在文件頻率上的關鍵詞非關鍵詞較詞彙頻率差距更明顯，換句話說，在同一個組別下，每個詞彙出現文本的比例相近，此時用詞頻平均值來比較詞彙便可知哪些是詞頻相對高的關鍵詞。

分組後的資料將進入降維階段，本研究方法以 t-SNE 作為降維依據，希望達成高維中相近的資料點在降至二維平面上依然能保持相近，並假設關鍵詞也能再降維後的平面上產生群集現象。圖 三-6 為降至二維後的資料點於平面上之呈現，將資料點以是否代表真正關鍵詞上色，紅點代表真實關鍵詞，黑點代表真實非關

DOI:10.6814/NCCU201900953

鍵詞，在圖形中可看到幾處資料群聚的現象，其中也包含真實關鍵詞群聚，這表示在分組及降維的處理下可將特徵相近的詞彙群聚在一起，然而在沒有顏色標記的情況下，該如何分辨何者為關鍵字群的問題仍需克服，在下一步驟將闡述資料點降維後如何分辨關鍵詞群及選取方法。

三、 關鍵詞選取步驟

為了分辨哪些資料點屬於關鍵詞，本文引入空間統計方法的概念，Kulldorff 於 1997 年提出偵測地區是否有較高疾病發生率的方法，其原理為藉由設定不同區域範圍，以概似比檢定(Likelihood Ratio Test)的方式測試區域內外是否存在顯著差異足以認定區域內存有群聚現象。本研究方法延續其概念，在平面上進行區域的劃分，將資料點分為區域內與區域外兩部分，若區域內的詞頻平均值(tf_mean)與區域外有顯著差異，則將區域內資料點判定為關鍵詞，首先定義判定指標如下：

$$\text{判定指標} = \frac{mean(\text{區域內資料點}tf\_mean\text{值})}{mean(\text{區域外資料點}tf\_mean\text{值})}$$

本文中區域劃定方式以點和點之間的距離決定，首先選取平面上詞頻平均值(tf_mean)最大的資料點並將其設定為區域內，其他點為區域外。之後依序將距離區域最近的點加入區域內並計算判定指標，圖 三-7 顯示了所有平面上資料點被選取的順序，並以此順序逐個計算的判定指標。如圖 三-8 所示，我們希望在判斷指標發生趨勢改變時設立停止點，如圖 三-9 標示，停止點前包含的詞彙即為此輪選取之關鍵詞。實際運作方式為計算判斷指標的變化率，若判斷指標持續穩定的上升或下降則繼續選取，反之，在變化率大於 10%時停止選取。在經過一輪的挑選後，將已選取關鍵詞去除並以剩下的資料點重新計算判定指標，若判定指

DOI:10.6814/NCCU201900953

標的第一個值小於 1.2 則停止此組選取，大於 1.2 則經由上述流程繼續選取下一輪關鍵詞，關於停止選取標準的參數設定將會在第四章更進一步討論。

DOI:10.6814/NCCU201900953

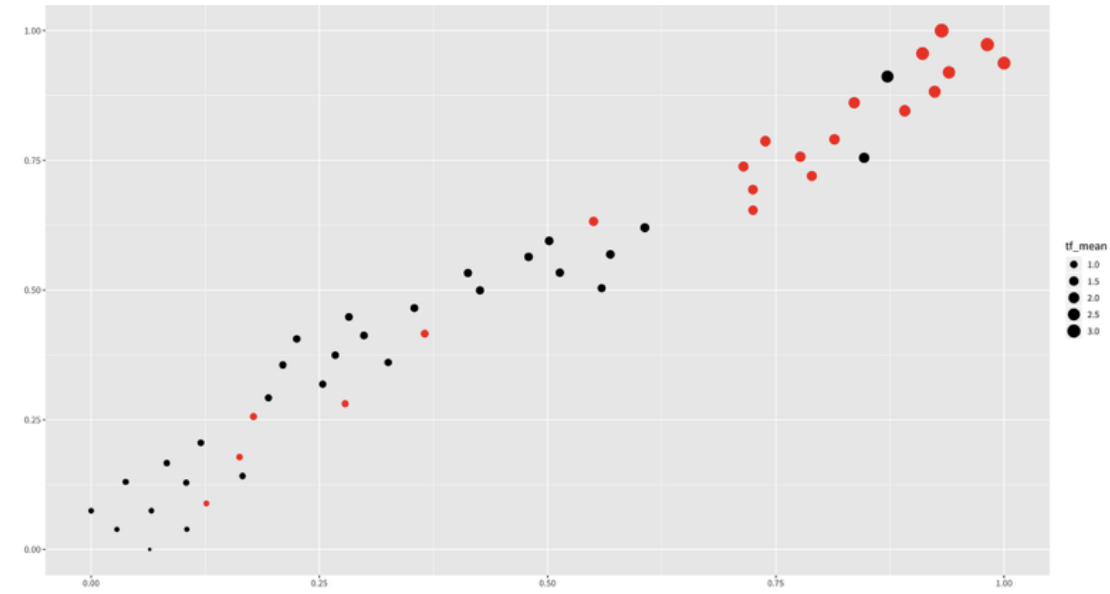

圖 三-6 降維後資料分布狀況

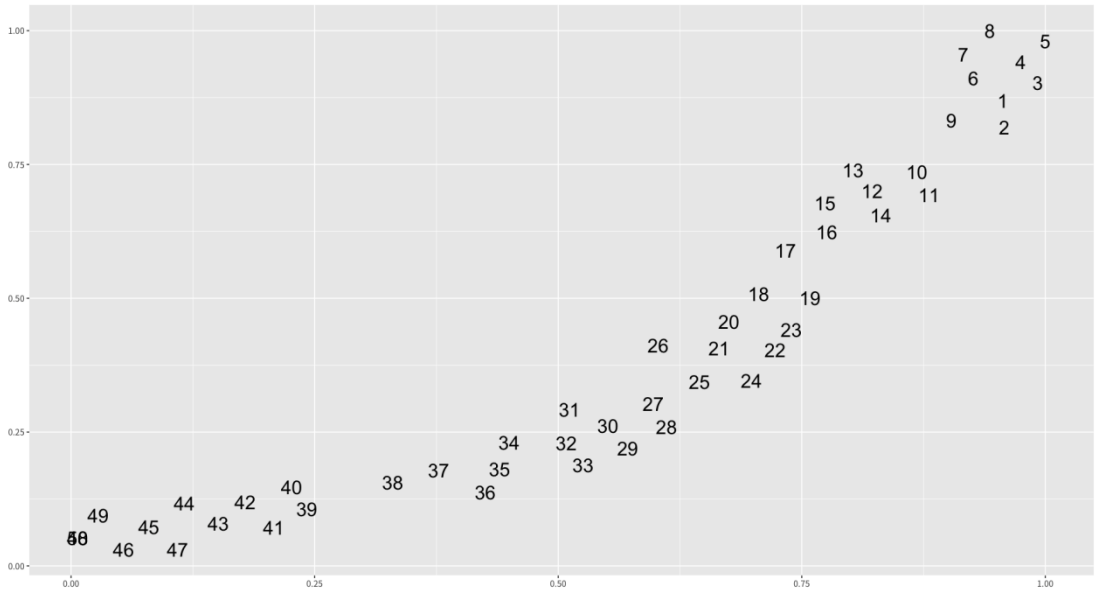

圖 三-7 資料點選取順序標記

DOI:10.6814/NCCU201900953

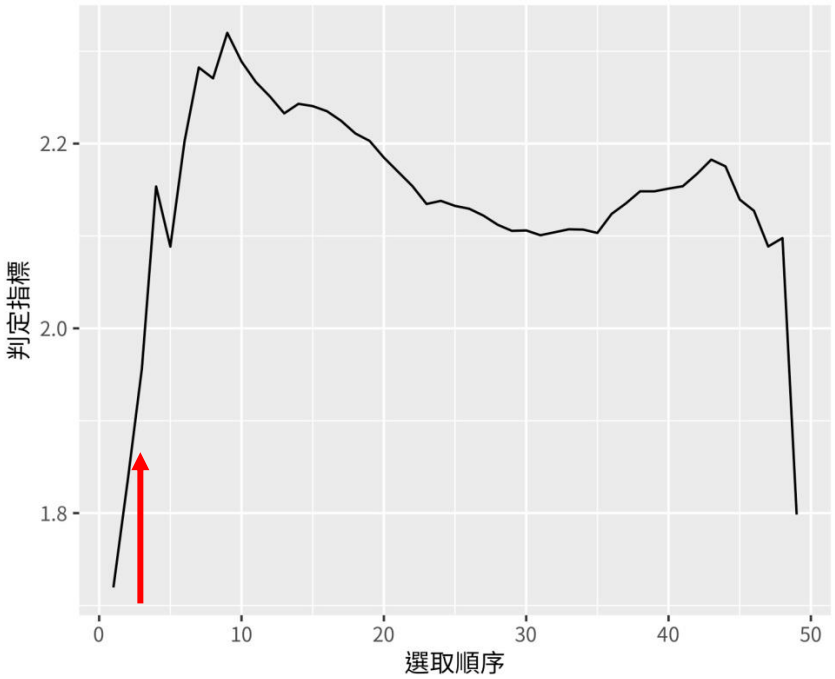

圖 三-8 判斷指標計算

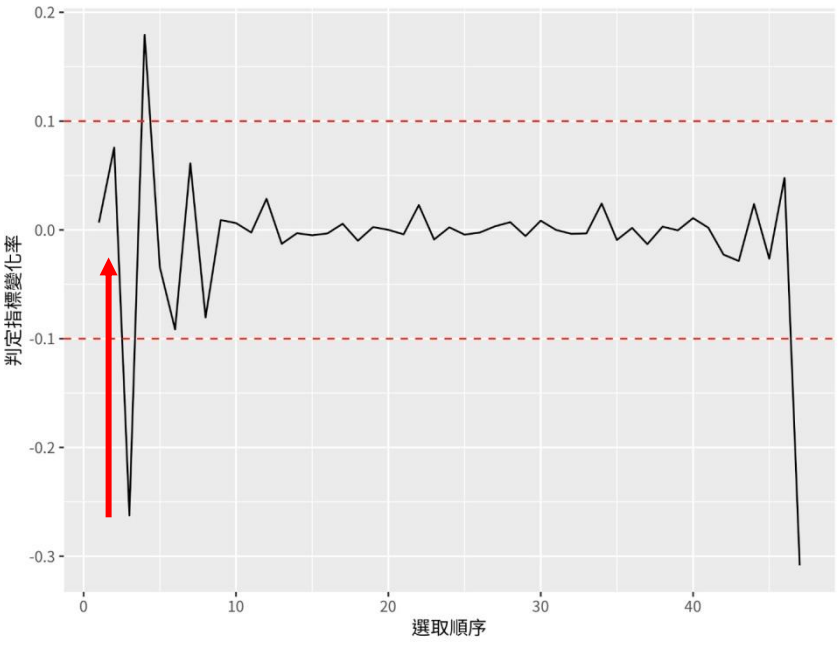

圖 三-9 判斷指標變化率

DOI:10.6814/NCCU201900953

第三節 成效評估

本文主要研究目標為關鍵詞擷取自動化預測文本中哪些詞彙為關鍵詞與非關鍵詞並與其他目前現有的模型比較。判斷是否為關鍵詞屬於二元分類問題(binary case)，為了評估方法間的優劣表現，本文使用了準確率、精確率、召回率、F1 指標等方法進行評估，以下將個別介紹之。首先介紹二元混淆矩陣(confusion matrix)，依照預測結果與真實標記可以將各種情況劃分為真陽性、真陰性、偽陽性、偽陰性，表 三-3 顯示了四種狀況所代表意義。

表 三-3 混淆矩陣定義方法

混淆矩陣		人工標記結果	
		真實關鍵詞	真實非關鍵詞
預測結果	判定關鍵詞	真陽性(TP)	偽陽性(FP)
	判定非關鍵詞	偽陰性(FN)	真陰性(TN)

一、準確率(Accuracy)：$\frac{TP+TN}{TP+TN+FP+FN}$，計算成功預測陽性及陰性佔全體的比例。

二、精確率(Precision)：$\frac{TP}{TP+TN}$，計算系統判定關鍵詞中真實關鍵詞的比例。

三、召回率(Recall)：$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$，計算真實關鍵詞中被成功判定為關鍵詞的比例。

四、F1 指標(F1-Measure)：$\frac{2*\mathrm{Precision}*\mathrm{Recall}}{\mathrm{Precision+Recall}}$，同時考慮了精確率即召回率，當 F1 指標愈大時分類模型愈穩健。

DOI:10.6814/NCCU201900953

第四章 電腦模擬參數設定

本章將延續第三章研究方法進行參數上的調整，第一節討論資料分組，透過移動窗格(Moving Windows)的方式取代原先分組之間不重複的設定。第二節則是探討在每組關鍵詞比例不同的狀況下將如何改變停止選取標準。兩節皆依照第三章中提及之研究流程以電腦模擬的方式隨機抽出 20%的原始資料，共進行 500 次關鍵詞擷取並以此作為參數調整依據。

第一節 移動窗格參數設定

在前一章已介紹本文方法，在預設參數下中，將所有資料依照在文件出現頻率(df_mean)降冪排列，並以 50 個詞為單位分成一組，依序放入 t-SNE 降維並選取關鍵詞，一個關鍵詞只會被分配一個組別，組別之間沒有交疊。然而，這樣的方式可能會使得某些關鍵詞雖然在所屬的組別中不顯著，亦或是在當前的組別中，少數關鍵詞無法有效形成群聚使系統偵測失敗，因此將分組方式調整，希望改善上述所遇到的問題。

藉由引入移動窗格(Moving Windows)的想法，每一個組別中的關鍵詞將重複被放進系統偵測，使真實關鍵詞被選取的機會提升。以設定移動窗格數(move_num) 25 為例，第一組先選取前 50 筆資料，第二組以後將原先第一組文件頻率較高的 25 筆資料去除，並加入新的 25 筆資料，分組情況為 1~50,26~75,51~100,..以此類推。在這樣的設定下，每個候選關鍵詞都會被偵測兩次，且這兩次中分屬組內前 25%和後 25%的部分，每個詞都有機會成為組內詞頻較高的部分。

DOI:10.6814/NCCU201900953

第二節 判斷指標參數設定

本節討論的參數為不同組別之間停止選取判斷指標。以預設參數 move_num=50 為例，第一組選取前 1~50 詞、第二組選取前 51~100 詞…以此類推直到所有的關鍵詞分組完成，由於各組之間所含的關鍵字比例不同，愈前面的組別包含較多的關鍵詞，因此考慮將個組別選取的標準依照線性比例調升，本次模擬比較了在固定以及線性調升判斷指標下的表現比較，探討何種組合將能獲得更好的表現。

透過群內群外詞頻(tf_mean)平均數的比值可以了解這兩群差異的顯著程度。在預設中，每一組的選取標準皆相同，但固定標準的設定情況下，會造成對前段組別標準過高，選不到所有關鍵詞；後段組別選取標準過低，容易選取過量非關鍵詞。如圖 四-1 所示，由於資料表已先將所有候選關鍵詞依照文件出現頻率(df_mean)排序，使得關鍵詞個數也會隨著組別遞增而遞減，因此需要針對這樣的趨勢對選取標準做出調整。

考慮到每個組別之間的關鍵字比例不同，若將所有組別的標準一致，將造成對於前段組別無法選取足量關鍵詞(Recall 低)，後段組別收集過多非關鍵詞(Precision 低)，因此考慮以線性遞增的方式，逐漸提升每個組別的評斷標準，參數設定分為起點(base)以及增量(add)，由不同的參數組合出各種標準，以起點 0.8、增量 0.1 為例，第一組門檻值為 0.8+0.1，第二組之後增加一倍增量，門檻值變為 0.9，以此類推。

DOI:10.6814/NCCU201900953

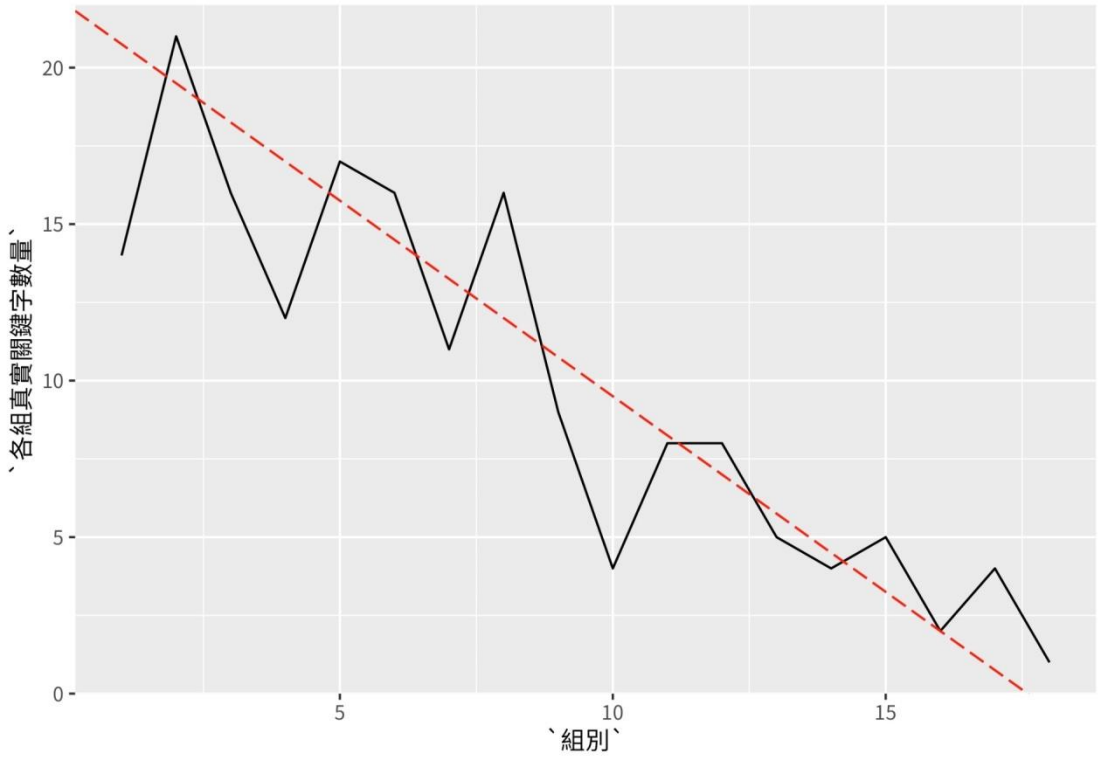

圖 四-1 各組別真實關鍵詞數量

第三節 成果評估

本節將以新青年第八卷為實驗文本，探討個參數組合下的表現優劣。將新青年第八卷的結構化資料隨機抽取 20%作為測試集。在每個參數組合下將重複抽取 500 次並記錄精確率(precision_mean)、召回率(recall_mean)、F1 指標(f1_mean)、系統判定關鍵詞個數(num_mean)作為比較依據。

在電腦模擬中主要改變兩處參數，一是移動窗格參數設定(move_num)，除了預設值 50 以外也增加了 25、15、10、5 共五組設定。二是判斷指標改良，本模擬將使用不同的權重測試，包含固定數值及線性遞增，固定數值採用 1.0、1.1、1.2、1.3、1.4、1.5 五組，線性遞增數值採用起點 0.8、0.9、1.0、1.1、1.2，搭配增量 0.05、0.10、0.15、0.20，共 20 種組合。

DOI:10.6814/NCCU201900953

表 四-1 顯示了固定判斷指標參數與五組移動窗格數設定下之表現，表格中數值高顯示紅色、數值低顯示綠色、兩者之間顯示白色，明顯看出 Precision 和 Recall 的數值剛好完全相反，在選取標準較低如 1.0~1.2 時， 由於選取數量多使得能夠包含較多的真實關鍵詞使 Recall 表現較好，但過度選取的壞處即是無法有效率、精準地選取關鍵詞，也造成了 Precision 較低的原因。選取標準較高如 1.3~1.5 則恰好相反。以個別參數而言，Precision 在固定數值 1.3~1.5 且移動窗格數 10~25 之間時表現最佳，Recall 在固定數值 1.0~1.1 且移動窗格數 5~25 之間時表現最佳，然而在綜合評估兩者的 F1 指標中卻是由移動窗格數 50 搭配固定數值 1.2 取得最佳，顯示在此模擬中仍須在 Precision 與 Recall 中作出取捨，一味追求單一指標最大化並非選取關鍵詞的最佳策略。

表 四-1 選取參數固定下模型表現

		F1					Precision					Recall				
		移動窗格數														
		5	10	15	25	50	5	10	15	25	50	5	10	15	25	50
固定標準	1.0	0.308	0.308	0.312	0.322	0.318	0.209	0.210	0.209	0.222	0.229	0.748	0.723	0.710	0.674	0.557
	1.1	0.338	0.342	0.339	0.346	0.337	0.285	0.275	0.269	0.278	0.269	0.630	0.637	0.618	0.596	0.506
	1.2	0.361	0.369	0.364	0.349	0.381	0.420	0.391	0.367	0.353	0.338	0.440	0.493	0.503	0.483	0.467
	1.3	0.332	0.353	0.364	0.355	0.335	0.477	0.496	0.480	0.451	0.398	0.300	0.337	0.371	0.390	0.362
	1.4	0.291	0.328	0.329	0.331	0.314	0.510	0.527	0.538	0.519	0.468	0.227	0.268	0.273	0.291	0.284
	1.5	0.257	0.302	0.291	0.302	0.287	0.509	0.537	0.548	0.552	0.504	0.185	0.226	0.218	0.232	0.232

DOI:10.6814/NCCU201900953

表 四-6 分別呈現了移動窗格數 50、25、15、10、5 搭配遞增判斷指標共 20 種組合，紅底數值表示該格在指標中排名高，綠色表示該格排名低。綜觀五張表格皆呈現相同趨勢，F1 指標與 Recall 表格中，左上角為數值較高的紅色，右下角為數值較低的綠色，Precision 則與兩者相反。若一開始設立門檻較高，會使系統的決策偏向保守，在比較有把握的情況下才會選取關鍵詞，雖然因此造成關鍵詞的數量較少，但包含了很多真實關鍵詞使 Precision 高，但是相對的選取關鍵詞數量少，真實關鍵詞被找到的機會、Recall 也隨之降低。

圖 四-2、圖 四-3 為圖表之視覺化，在移動窗格數遞增時，各項指標也隨之增加並在移動窗格數 25 時達到最高點。綜合兩種移動窗格數與判斷指標的交互作用，本次模擬認為在設定移動窗格數 25 搭配判斷指標起點(base)：0.8、增量(add)：0.1 的情況下可以有最佳的關鍵詞選取效果，本文將在下一章以此節模擬成果為基礎與其他現行監督式學習模型比較，驗證本文方法之可行性。

DOI:10.6814/NCCU201900953

表 四-2 移動窗格數 50 字的模擬結果

		移動窗格數 50 字											
		F1				Precision				Recall			
		增量											
		0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
起點	0.8	0.357	0.429	0.440	0.420	0.253	0.355	0.435	0.494	0.616	0.558	0.465	0.385
	0.9	0.378	0.427	0.420	0.383	0.288	0.400	0.470	0.524	0.569	0.485	0.399	0.321
	1.0	0.384	0.407	0.373	0.346	0.329	0.443	0.494	0.537	0.497	0.406	0.322	0.269
	1.1	0.375	0.368	0.337	0.299	0.387	0.475	0.525	0.541	0.419	0.329	0.268	0.220
	1.2	0.355	0.336	0.295	0.274	0.452	0.507	0.537	0.551	0.339	0.275	0.219	0.193

表 四-3 移動窗格數 25 字的模擬結果

		移動窗格數 25 字											
		F1				Precision				Recall			
		增量											
		0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
起點	0.8	0.439	0.5	0.466	0.417	0.328	0.47	0.538	0.562	0.688	0.552	0.425	0.341
	0.9	0.465	0.476	0.419	0.371	0.39	0.525	0.561	0.573	0.609	0.459	0.349	0.286
	1.0	0.463	0.428	0.376	0.337	0.449	0.551	0.564	0.583	0.52	0.373	0.295	0.248
	1.1	0.435	0.381	0.33	0.307	0.499	0.563	0.563	0.567	0.424	0.307	0.246	0.218
	1.2	0.384	0.336	0.302	0.269	0.539	0.568	0.579	0.551	0.328	0.254	0.216	0.186

DOI:10.6814/NCCU201900953

表 四-4 移動窗格數 15 字的模擬結果

		移動窗格數 15 字											
		F1				Precision				Recall			
		增量											
		0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
起點	0.8	0.495	0.478	0.412	0.371	0.408	0.513	0.531	0.535	0.645	0.46	0.346	0.292
	0.9	0.492	0.437	0.38	0.349	0.455	0.532	0.538	0.558	0.561	0.385	0.304	0.262
	1.0	0.466	0.392	0.349	0.311	0.503	0.545	0.552	0.547	0.464	0.319	0.265	0.225
	1.1	0.421	0.355	0.302	0.281	0.536	0.545	0.547	0.542	0.374	0.275	0.218	0.199
	1.2	0.348	0.318	0.291	0.264	0.536	0.55	0.55	0.546	0.278	0.236	0.207	0.181

表 四-5 移動窗格數 10 字的模擬結果

		移動窗格數 10 字											
		F1				Precision				Recall			
		增量											
		0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
起點	0.8	0.507	0.441	0.434	0.343	0.462	0.518	0.514	0.527	0.579	0.395	0.386	0.26
	0.9	0.48	0.408	0.399	0.313	0.494	0.533	0.526	0.536	0.485	0.34	0.331	0.228
	1.0	0.448	0.369	0.368	0.294	0.523	0.533	0.536	0.543	0.409	0.29	0.29	0.21
	1.1	0.393	0.331	0.322	0.263	0.535	0.532	0.531	0.537	0.326	0.25	0.241	0.182
	1.2	0.344	0.295	0.294	0.252	0.546	0.536	0.532	0.542	0.268	0.212	0.213	0.172

DOI:10.6814/NCCU201900953

表 四-6 移動窗格數 5 字的模擬結果

		移動窗格數 5 字											
		F1				Precision				Recall			
		增量											
		0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
起點	0.8	0.446	0.364	0.364	0.278	0.488	0.498	0.496	0.475	0.421	0.293	0.293	0.202
	0.9	0.413	0.338	0.333	0.266	0.502	0.502	0.494	0.490	0.362	0.261	0.258	0.189
	1.0	0.372	0.309	0.308	0.244	0.498	0.505	0.500	0.501	0.308	0.231	0.230	0.167
	1.1	0.341	0.283	0.286	0.237	0.512	0.502	0.507	0.518	0.268	0.205	0.208	0.160
	1.2	0.293	0.256	0.255	0.218	0.501	0.502	0.488	0.506	0.218	0.179	0.180	0.145

DOI:10.6814/NCCU201900953

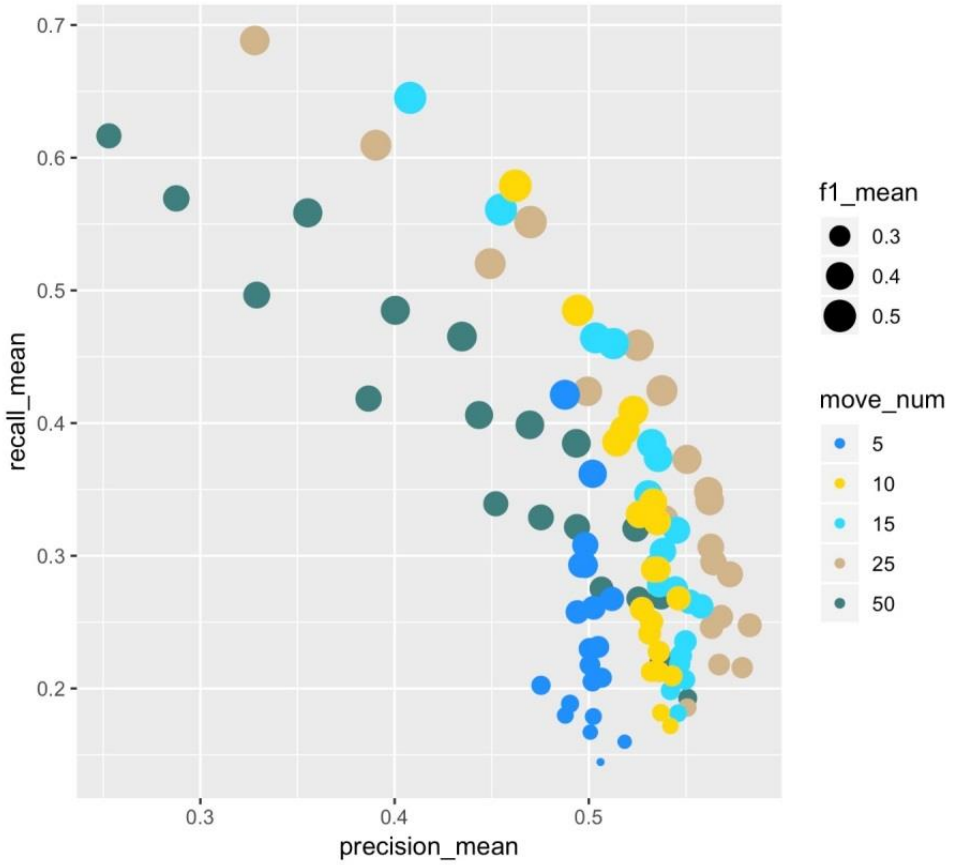

圖 四-2 模擬結果視覺化

DOI:10.6814/NCCU201900953

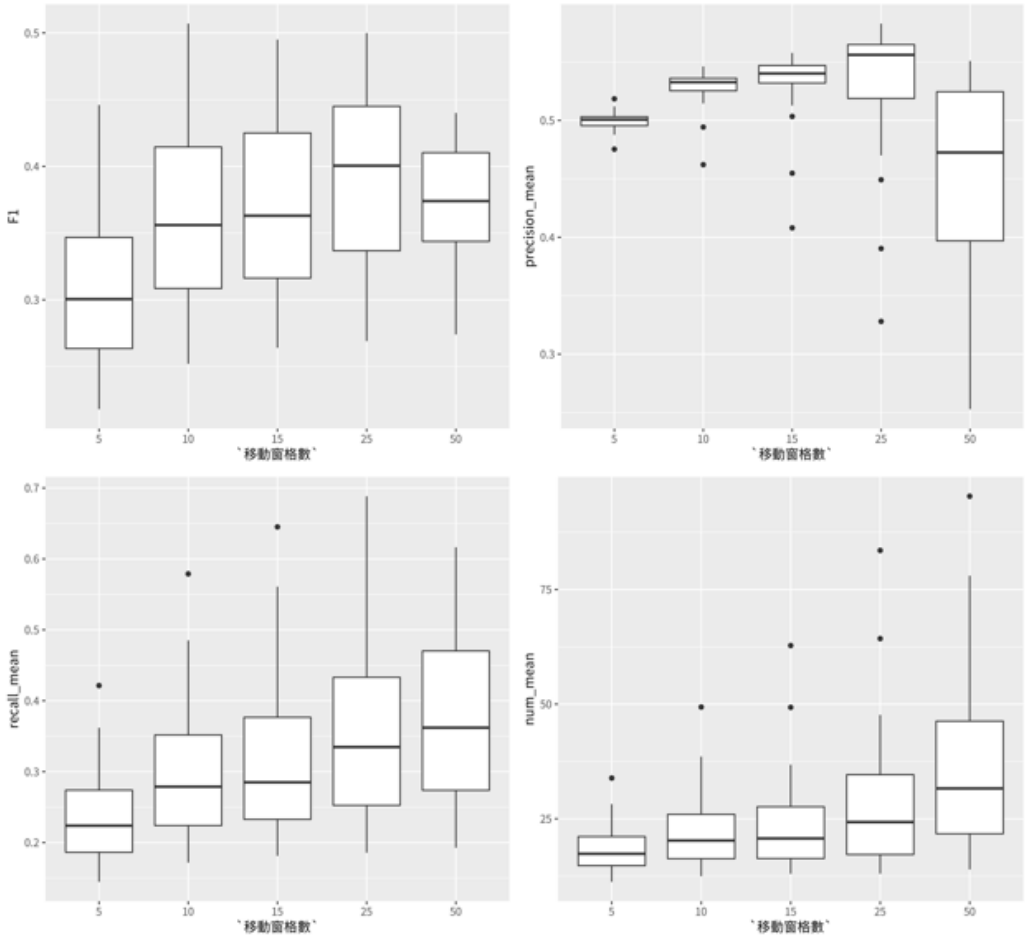

圖 四-3 不同移動窗格下表現

DOI:10.6814/NCCU201900953

第五章 實證分析

本章將以新青年七、八卷及人民日報作為實驗文本，討論各種方法於不同文本下的表現狀況。分為實驗設定說明、多種模型比較以及選取關鍵詞比較。

第一節 模擬設定說明

本節將使用羅吉斯回歸(Logistic Regression)、線性判斷分析(LDA, Linear Discriminant Analysis)、支持向量機(SVM, Support Vector Machine)與本文方法比較，以簡單隨機抽樣(Simple Random Sampling)抽取 80%資料做為訓練集，剩餘 20%資料為測試集。各文本皆模擬 500 次，每次模擬中各模型使用之訓練集、測試集皆相同，其中，由於本文方法屬於非監督式學習模型，事前無需利用訓練集建立模型，因此直接以測試集挑選關鍵詞。

表 五-1 實驗文本為三篇文本的基本資料，其中有些詞彙於整篇文本中出現不足 10 次，因此將其刪除於候選關鍵詞名單中。三篇實驗文本的關鍵詞與非關鍵詞比例約為 5%-9%，屬不平衡資料(Unbalanced Data)，意即資料中的關鍵詞、非關鍵詞比例懸殊，在訓練模型時會傾向將所有資料都判為非關鍵詞，增加正確判斷的難度，為此本模擬嘗試使用過取樣(Over Sampling)的方式使兩種類別的資料重複抽取直至比例相同。

此外，由於羅吉斯回歸在判定結果時需設定閥值，數值設定的過高會使整體決策過於保守，不易判定為關鍵詞；反之數值過低則會使判定關鍵詞的數量大增，使準確率下降。本次模擬以網格搜尋(grid search)的方式，在訓練集先抽取 80%

DOI:10.6814/NCCU201900953

訓練集資料當作調整參數的依據，剩餘 20%訓練集驗證使用不同參數的效果，最終以最佳參數配合訓練集產生模型。

表 五-1 實驗文本的基本資訊

	新青年第七卷	新青年第八卷	人民日報
發行時間	1919~1920	1920~1921	1971~1989
總字數	591,756	465,319	421,899
總篇數	141	179	522
詞彙數	2830	2196	2181
關鍵字數量	200	199	109

第二節 模型比較

表 五-2、表 五-3、表 五-4 分別代表人民日報、新青年第七卷、新青年第八卷的模擬結果。表格中標記紅字者為該指標下分數最高前兩名。圖 五-1 則是顯示各組合間視覺化圖形，橫軸、縱軸、大小、顏色分別代表召回率、精確率、F1 指標、模型選取關鍵詞個數並以訓練及抽取方式分為隨機抽樣、重複抽樣兩區塊；圖 五-2 為各方法挑選的關鍵詞個數，顯示不同模型間挑選關鍵詞策略差異，有些藉由選取少量關鍵詞以獲得較高的精確率其他則是選取較多關鍵詞使召回率有較好表現。從名次的分佈來看羅吉斯回歸與本文方法總體表現最佳，在三篇文本中的 F1 指標與召回率皆取得較高的分數。LDA 在精確率方面表現最優異，SVM 則是在各種指標敬陪末座。

DOI:10.6814/NCCU201900953

前一節提到由於樣本不平衡的問題可能使模型在分類時無法獲得關鍵詞的足夠資訊，因此使用了重複抽樣的方法與隨機抽樣比較。模擬結果顯示重複抽樣的設定並無法提升的表現，反而使 Recall 下降。推測可能因為過擬合(Overfitting)問題造成模型整體表現不佳。透過此次模擬可知本文方法在分類表現上已和監督式學習模型接近，顯示目前所選擇做為資料結構化的特徵隱含了詞彙是否配判定為人工關鍵詞的資訊。

表 五-2 人民日報模擬結果

人民日報					
Model	Accuracy	Precision	Recall	F1	選取個數
LDA	0.942(0.010)	0.395(0.118)	0.286(0.094)	0.323(0.09)	15.8(4.4)
LDA*	0.943(0.008)	0.430(0.107)	0.27(0.083)	0.323(0.078)	23.5(7.5)
Logistic	0.926(0.026)	0.340(0.107)	0.427(0.156)	0.356(0.091)	29.4(15.9)
Logistic*	0.926(0.032)	0.379(0.123)	0.366(0.188)	0.325(0.088)	42.3(34.3)
SVM	0.866(0.016)	0.095(0.039)	0.198(0.080)	0.126(0.049)	45.2(6.8)
SVM*	0.878(0.017)	0.069(0.028)	0.108(0.042)	0.083(0.032)	58.3(12.2)
本文方法	0.914(0.015)	0.287(0.079)	0.476(0.101)	0.353(0.079)	36.5(6.6)

注：*表抽樣方法為重複抽樣(Over Sampling)

DOI:10.6814/NCCU201900953

表 五-3 新青年第七卷模擬結果

新青年第七卷					
Model	Accuracy	Precision	Recall	F1	選取個數
LDA	0.929(0.010)	0.487(0.091)	0.327(0.070)	0.388(0.069)	26.6(5)
LDA*	0.926(0.006)	0.445(0.059)	0.295(0.059)	0.351(0.050)	48.8(10.1)
Logistic	0.900(0.022)	0.386(0.074)	0.629(0.171)	0.460(0.067)	66.9(23.6)
Logistic*	0.912(0.021)	0.411(0.084)	0.463(0.184)	0.404(0.080)	88.6(45.7)
SVM	0.887(0.012)	0.279(0.052)	0.388(0.069)	0.322(0.052)	54.8(6.8)
SVM*	0.899(0.011)	0.278(0.048)	0.292(0.049)	0.283(0.043)	77.8(12.7)
本文方法	0.915(0.014)	0.411(0.079)	0.477(0.113)	0.432(0.073)	46.5(12.3)

注：*表抽樣方法為重複抽樣(Over Sampling)

表 五-4 新青年第八卷模擬結果

新青年第八卷					
Model	Accuracy	Precision	Recall	F1	選取個數
LDA	0.922(0.012)	0.594(0.093)	0.405(0.078)	0.478(0.074)	26.9(5)
LDA*	0.918(0.008)	0.479(0.061)	0.371(0.060)	0.415(0.050)	42.1(7.3)
Logistic	0.902(0.024)	0.490(0.100)	0.576(0.143)	0.509(0.064)	48.9(19.2)
Logistic*	0.903(0.023)	0.437(0.084)	0.510(0.142)	0.450(0.056)	67.2(29.2)
SVM	0.862(0.016)	0.320(0.057)	0.477(0.090)	0.380(0.059)	58.6(8.1)
SVM*	0.880(0.012)	0.298(0.042)	0.386(0.054)	0.335(0.042)	70.1(8.9)
本文方法	0.903(0.015)	0.473(0.084)	0.549(0.090)	0.501(0.067)	46.2(8.7)

注：*表抽樣方法為重複抽樣(Over Sampling)

DOI:10.6814/NCCU201900953

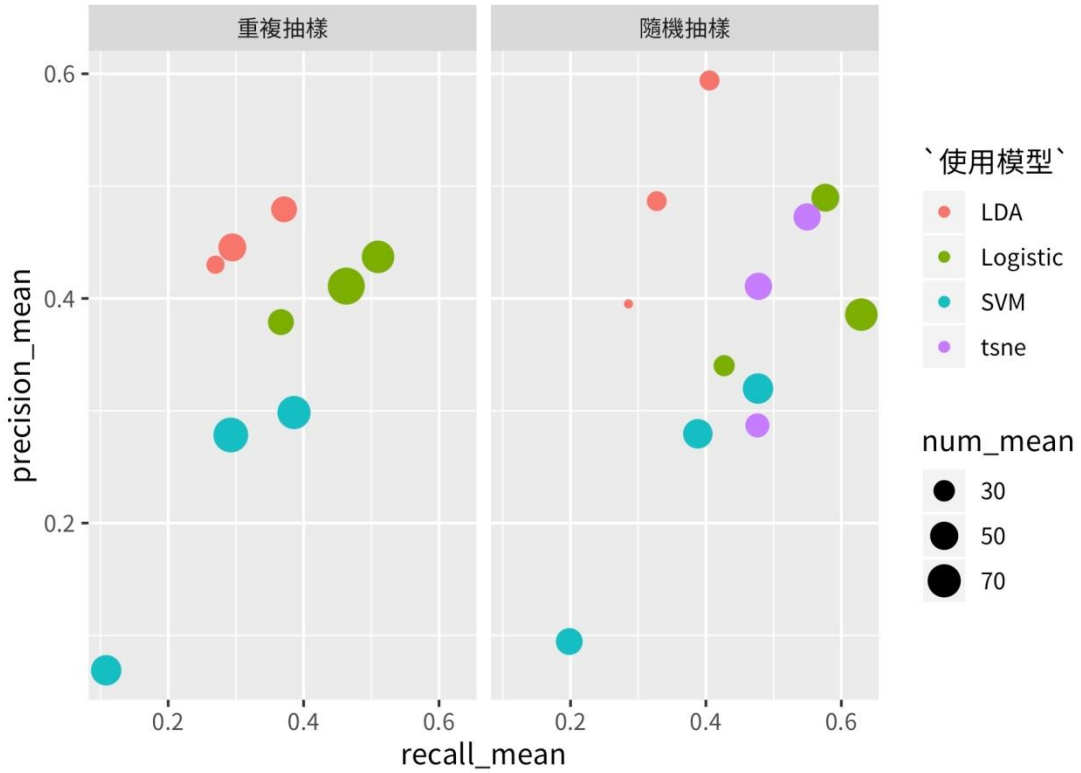

圖 五-1 模型比較結果視覺化

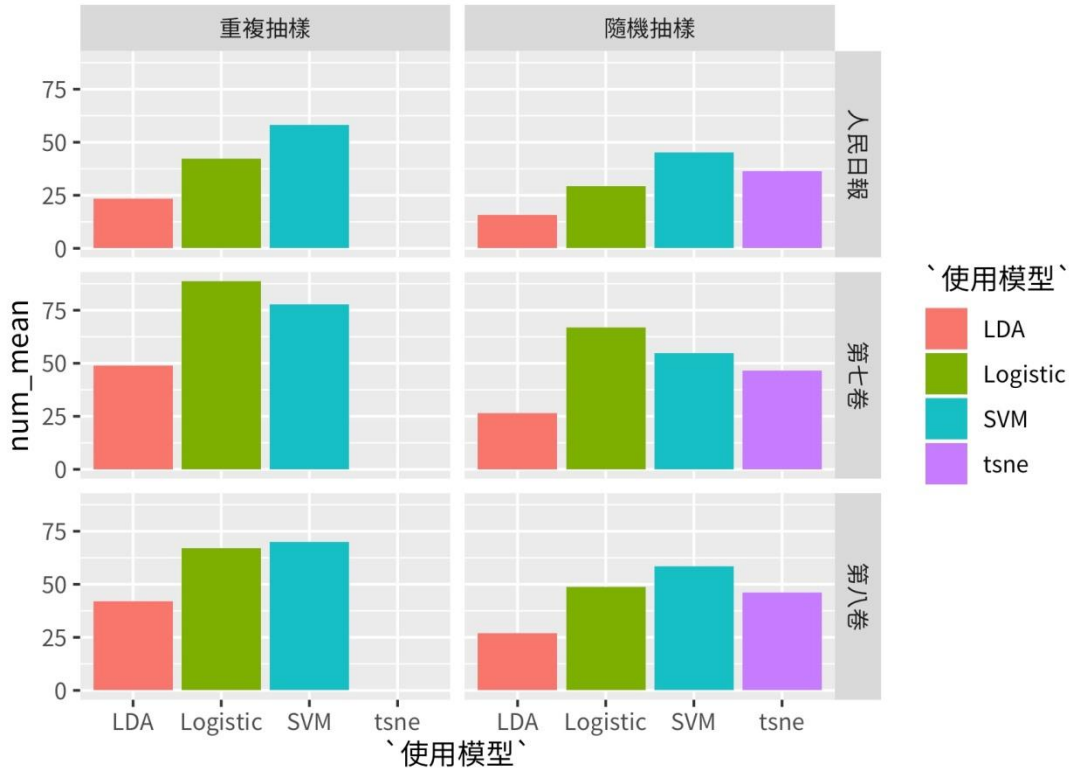

圖 五-2 各模型挑選關鍵詞個數

DOI:10.6814/NCCU201900953

第三節 選取關鍵詞差異比較

本節將討論各模型選取關鍵詞的差異，挑選設定為先將資料隨機分成五等份，以其中四份作為訓練集，剩餘一份作為測試集當作模擬的依據，這樣的設定可以使每一個在資料內的詞彙都有可以被選取的機會，但詞彙是否被選取會依照隨機選取的訓練集所建立之模型來決定，同樣的詞彙在不同隨機決定的訓練集下不一定都有相同的結果。因此，本文將重複上述測試集、訓練集建立方式 500 次，並分別以羅吉斯迴歸、LDA、SVM 以及本文方法選取關鍵詞，若詞彙在 500 次模擬中有超過 400 次被模型選取（選取比例超過八成），則認定該詞彙為關鍵詞。而本文所挑選之人工選取關鍵詞是由文章中先挑取出現次數前500名的雙字詞，再經由專家學者篩選出符合文本的關鍵詞，因此常見的地名、或者三字以上的詞彙將不出現在專家所認定的關鍵詞中，附錄一列出了專家學者在各文本中挑選的關鍵詞。

本節將所選關鍵詞分為兩種狀況討論，第一種為模型分辨正確，四種模型中有三個以上判定該詞彙為關鍵詞且該詞彙為真實關鍵詞；第二種為模型分辨錯誤，四種模型中有三個以上判定該詞彙為關鍵詞但該詞彙不是真實關鍵詞，三種文本所選取之關鍵詞如表 五-5 所列，在新青年部分可以發現少數非雙字詞且與中共相關的詞彙被挑出，如共產主義/共產黨/委員會/社會主義/社會黨/莫斯科/勞動者/資本主義/資本家/聯合會/蘇維埃等詞，也挑出了許多當時提及的國家或地區，如上海/北京/俄國/美國/英國/莫斯科/蘇維埃，另一方面，人民日報在關鍵詞挑選上精確率較差，有許多關鍵詞無法正確挑出，但是在被挑選的詞彙中，和新青年一樣皆挑出了國家或地區的名稱，除此之外從挑選的關鍵詞腫也可以發現石油似乎也是當時熱烈討論的議題之一。

DOI:10.6814/NCCU201900953

表 五-5 各文本所選關鍵詞

使用文本	挑選狀況	選取關鍵詞
新青年第七卷	模型正確判別關鍵詞	人口/人民/人類/女子/女工/工人/工作/工業/工資/工廠/工錢/中國/互助/分配/文明/世界/主義/平均/平和/平等/民族/生活/目的/全國/危險/地方/自由/自治/你們/利益/改造/男女/男子/制度/宗教/法律/社會/青年/政府/政治/科學/限制/革命/個人/原則/哲學/時間/問題/國民/國家/國際/婦人/理想/組合/組織/勞工/勞動/發達/階級/經濟/罪惡/資本/運動/道德/僱主/標準/罷工/戰爭/機關/選舉/壓迫/聯合/競爭/權力
	模型誤判關鍵詞	一天/一方面/一日/一面/三年/上海/土地/小時/公司/夫人/日本/父母/主張/北京/可怕/可是/必定/未必/母親/生出/先生/各國/如果/作工/每天/每日/每月/決定/車伕/事務/事情/或是/所以/東西/法子/狀況/的話/知道/表示/便是/俄國/孩子/美國/英國/要求/食物/時期/根本/假使/規定/這裡/這種/這樣/勞動者/喜歡/普通/然而/發生/結果/傾向/會員/經理/裡面/資本家/態度/需要/增加/數目/衝突/調查/學說/曉得/聯合會/願意/警察/鐵路/變遷
新青年第八卷	模型正確判別關鍵詞	人民/人們/人類/女子/工人/工作/工廠/反對/支配/文化/文學/他們/平民/民族/生活/生產/全國/全體/共產/危險/多數/自由/你們/我們/改革/男子/事實/供給/兒

DOI:10.6814/NCCU201900953

		童/制度/和平/宗教/幸福/法律/物質/知識/俄國/思想/政府/政治/政策/科學/英國/革命/個人/哲學/家庭/時間/財產/國民/國家/教育/理論/產業/組合/組織/創造/勞工/勞動/勞農/無產/發展/進化/階級/勢力/解放/資本/農民/運動/團體/實業/德國/學生/學校/機關/歷史/獨立/選舉/聯合/職業/羅素/競爭/議會/犧牲/權力/權利/觀念
	模型誤判關鍵詞	一方面/一種/一樣/土地/大家/大會/小說/中間/之中/主張/代表/可能/先生/共產主義/共產黨/地主/好像/如果/似乎/但是/利用/形式/決定/那些/事情/委員/委員會/所有/承認/武力/社會主義/社會黨/看見/計畫/時候/做工/現象/理由/莫斯科/這樣/勞動者/然而/發達/進行/會議/當時/資本主義/資本家/農業/過去/說明/曉得/應該/聯合會/關係/蘇維埃/覺得
人民日報	模型正確判別關鍵詞	人民/中國/主義/代表/民主/佔領/法律/侵犯/帝國/政府/政權/美國/鬥爭/問題/國家/婦女/規定/勞動/猶太/越南/集團/幹部/經濟/群眾/種族/罷工/憲法/憲章/選舉/蘇聯
	模型誤判關鍵詞	土地/工人/工作/工會/干涉/中華人民共和國/公司/公民/反映/巴勒斯坦/巴勒斯坦解放組織/日本/他們/代表團/以色列/可以/它們/民族/石油/企業/同志/行為/我們/我國/決議/兩個/委員會/所謂/社會主義/阿拉伯

DOI:10.6814/NCCU201900953

		/非洲/南非/建議/柬埔寨/埃及/敘利亞/組織/這一/提案/發展中國家/進行/意見/當局/農民/領導/學生/總統/聯合/聯合國/舉行/職工

第三章曾提及 TF-IDF 的優缺點，其中 TF-IDF 主要蒐集到高頻關鍵詞，因此提出本文方法使低頻關鍵詞也有被選取的機會。在挑選完關鍵詞後，我們利用秩(Rank)的計算方式來驗證各方法所選取之詞彙有沒有詞頻上的差異。將所有詞彙的詞頻由大至小依序編號，詞頻最大的標記為 1，詞頻次大者標記為 2 接著依序標記，若選取愈多的低詞頻的關鍵詞則相對應的秩也將愈高。表 五-6、表 五-7 分別紀錄了人民日報以及新青年第八卷有關秩的統計量，本文方法以及羅吉斯迴歸皆能選取到詞頻較低的詞彙，同時在 F1 指標的表現也較其他模型好，由此可知本文方法的確能較有效選取高頻及低頻關鍵詞。圖 五-3、圖 五-4 更進一步分別所選的關鍵詞中哪些是真實關鍵詞，圖中以 tsne 代表本文方法、log 代表羅吉斯迴歸並將所選關鍵詞分為真實關鍵詞（以模型方法_1 標記）、真實非關鍵詞（以模型方法_0 標記）以及兩者合併（以模型方法_all 標記）。本文方法和羅吉斯迴歸皆能選到秩在 200 至 300 間的關鍵詞，雖然跟人工標記的真實關鍵詞仍有些差距，但已可突破只有選取到高頻關鍵詞的困境。

DOI:10.6814/NCCU201900953

表 五-6 人民日報各模型秩的統計量

	本文方法	Logistic	LDA	TF-IDF	人工標記
第一四分位數	40	53	25	38	71
中位數	145	113	75	76	191
第三四分位數	243	183	121	124	555
平均數	162	128	86	93	362
選取數量	144	123	81	150	109
F1	0.381	0.389	0.329	0.357	-

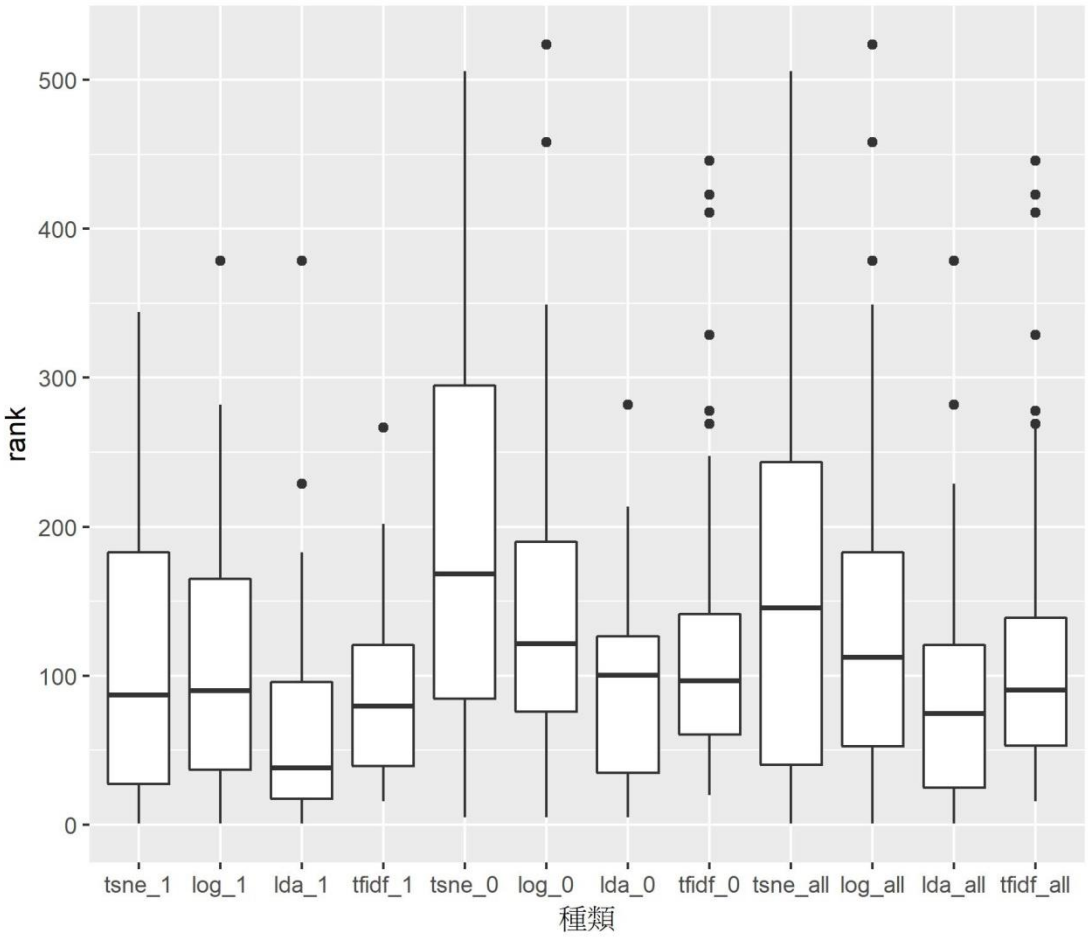

圖 五-3 人民日報各模型秩的盒狀圖

DOI:10.6814/NCCU201900953

表 五-7 新青年第八卷各模型秩的統計量

	本文方法	Logistic	LDA	TF-IDF	人工標記
第一四分位數	77	104	58	50	112
中位數	184	214	116	100	234
第三四分位數	297	354	175	169	397
平均數	207	241	124	127	280
選取數量	208	293	131	200	199
F1	0.506	0.488	0.484	0.471	-

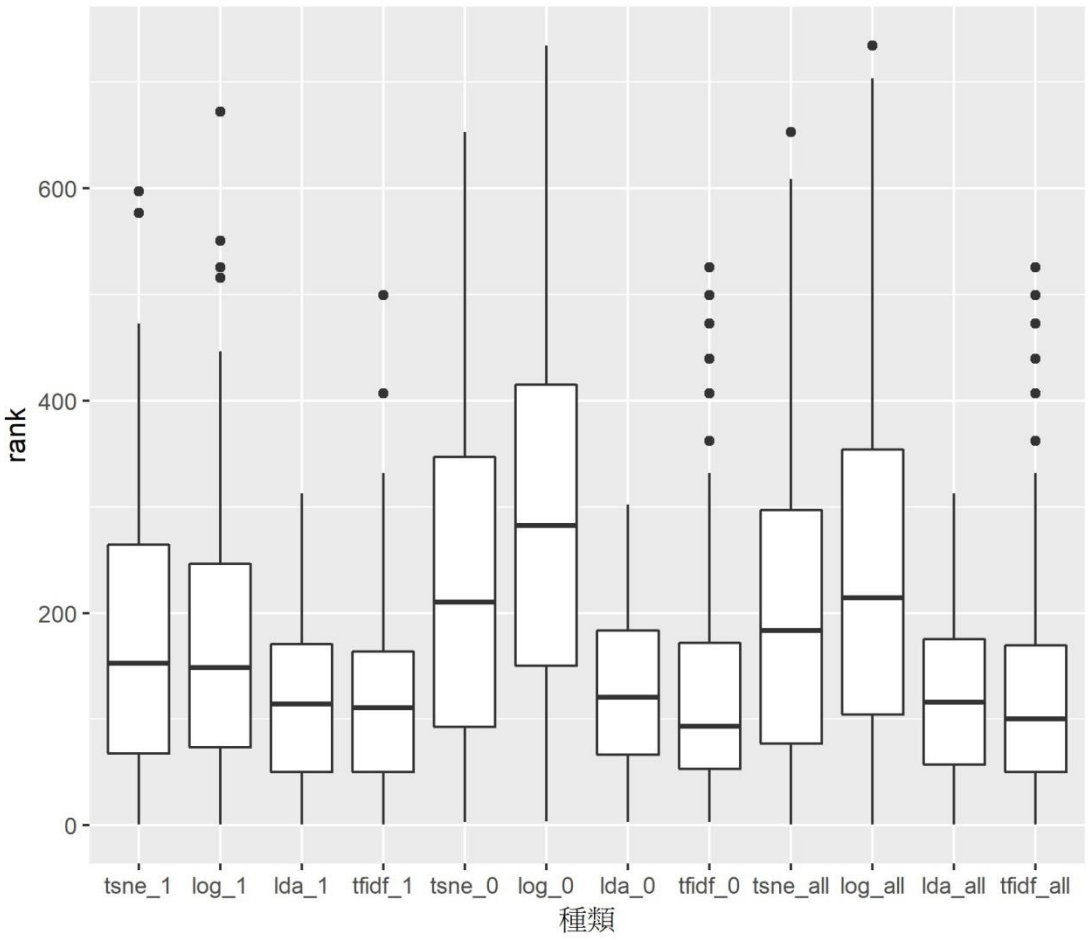

圖 五-4 新青年第八卷各模型秩的盒狀圖

DOI:10.6814/NCCU201900953

第六章 結論與建議

第一節 結論

隨著紙本電子化的普及，許多書籍得以使用科學的方式進行分析，過往文字必須依賴專家學者們對文本進行解析，人工判讀往往曠日費時效率較差，且結果容易因個人主觀意見影響。藉由文字分析的技術能以較客觀且科學的方式分析文本，找出人工判讀不易發現的盲點。關鍵詞擷取是文字分析中重要的一部份，從關鍵詞中可以快速了解一篇文章中最核心的概念，也能從多篇文章中掌握作者注重的議題。藉由關鍵詞擷取自動化在大量文字中篩選出重要的詞彙，為數位文本普及的時代開拓出另一種解析文本的方法。

本文提出了關鍵詞擷取的非監督式學習模型，由文本結構化、資料分組、偵測詞彙等過程找出各詞頻中的潛在關鍵詞，並藉由實證分析比較了羅吉斯回歸、LDA、SVM 及本文方法的分類表現。在《新青年》及《人民日報》中，屬於監督學習的羅吉斯回歸及 LDA 皆有著穩定且準確的表現，而屬於非監督學習的本研究方法，在沒有預先使用資料訓練模型的設定下亦與上述方法有著十分接近的表現。透過數據模擬的方式，本文提出選取關鍵詞的最佳參數，包含加入移動窗格概念的資料分組方法以及每組資料進行選取時的停止標準，使選取關鍵詞的結果在精確率、召回率上取得平衡，讓 F1 指標能有最佳表現。除此之外，藉由調整參數亦能使本文方法決定關鍵詞選取數量，當停止標準較嚴苛，選取關鍵詞的策略將偏向保守，挑選的詞彙數量比較少但模型精確率高，能迅速抓取代表文本的重要關鍵詞。反之，當停止標準較寬鬆，挑選出來的詞彙數量也將增加，能夠抓取多數真實關鍵詞，適合作為初步篩選非關鍵詞的工具。

DOI:10.6814/NCCU201900953

在過去經驗中，出現次數高的詞彙被認定為關鍵詞的機會比較大，因此發展出由詞頻與文件頻率組合而成的 TF-IDF 來選取關鍵詞，然而透過人工標記的方式了解關鍵詞中也包含了低頻關鍵詞，由於出現的次數不多，不容易被模型所挑選。於是提出了本文方法，希望能透過資料分組的方式增加選取低頻關鍵詞選取機會。在比對各模型所挑詞彙以及相對應的秩(rank)，驗證了本文方法能選到比其他模型更低頻的關鍵詞。

第二節 研究限制與未來建議

模型參數設定可直接影響選取關鍵詞成效，除此之外，找出具有代表性的詞彙統計量也是決定選取結果的重要步驟。本文使用了詞頻及文件頻率的相關統計量作為資料結構化的依據，雖然成效已經可以與一般模型達到相同水準，但仍有很大的進步空間，在後續的研究上可以考慮將其他關於文本的統計量加入，例如: 文章發表順序、詞性等，使詞彙能有更多不同的切入點進行分析。

在關鍵詞選取結果中可發現有許多相同概念的詞彙，例如《新青年》中的女人/女子/工人/工作/工廠/公司/男子/勞工/勞動/勞動者/勞農，談論這些詞的背景很可能與中國當時的勞動有關。可透過共現詞建立詞與詞之間的關係，進而分析關鍵詞之間是否有群聚的現象，以及探討這些群聚詞彙是否可回溯至相同概念。

t-SNE 為本研究方法的核心技術之一，主要透過資料間的相似度為依據，使高維資料得以在低維空間中呈現，透過這樣的方法能將複雜且多維度的資料進行降維，然而在加入新的資料點時必須重新計算，無法沿用過去的降維結果，使分

DOI:10.6814/NCCU201900953

組過後的資料在計算上花費的時間較久。本文所使用的詞彙約兩千餘字，資料並不算多，但未來在面對更龐大的資料量時，過長的運算時間仍須想辦法克服。

除了單一模型選取之外，透過不同模型間的配合如羅吉斯回歸與本文方法對文本進行關鍵詞選取，以較寬鬆的停止選取標準能將明顯不為關鍵詞的資料排除，並作為羅吉斯迴歸的資料集；利用較嚴苛的停止選取標準選取出少量且精確率高關鍵詞作為模型的訓練標籤。以多模型配合的模式亦是未來研究可嘗試的方向。

DOI:10.6814/NCCU201900953

參考文獻

一、中文文獻

1. 何昱鋒(2019)，「基於物聯網之即時環境監測系統」，碩士論文，國立臺灣海洋大學電機工程學系。
2. 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*，第七期，頁 427-454。
3. 金觀濤、梁穎誼、姚育松、劉昭麟(2014)，「統計偏離值分析於人文研究上的應用」，*東亞觀念史集刊*，第六期，頁 331-366。
4. 黃居仁(2005)，「漢字知識表達的幾個層面：字、詞與詞義關係概論」，漢字與全球化國際學術研討會論文集，頁 77-88。
5. 郭益豪(2013)，「以改良式 N-Gram 斷詞法結合潛在語意分析進行以改良式 N-Gram 斷詞法結合潛在語意分析進行網頁影像加註」.，碩士論文，國立雲林科技大學資訊管理系。
6. 謝孟樺(2018)，「考量上下文字詞共現關係之短文斷詞研究」，碩士論文，國立中興大學資訊科學與工程學系。
7. 梁家安(2016)，「從國共內戰到改革開放：人民日報風格變遷之量化研究」，碩士論文，國立政治大學統計研究所。
8. 謝博行(2013)，「局部最長連續共同子序列與新詞組收集」，碩士論文，國立清華大學統計學研究所。
9. 潘豔豔(2015)，「探索性資料分析方法在文本資料中的應用—以《新青年》雜誌為例」，碩士論文，國立政治大學統計研究所。

DOI:10.6814/NCCU201900953

二、英文文獻

1. Demets, D.L. and Lan, K.G. (1994). “Interim analysis: the alpha spending function approach.” *Statistics in Medicine, 13*(13-14): 1341-1352.
2. Hinton, G.E. and Roweis, S.T. (2003). “Stochastic neighbor embedding.” *Advances in neural information processing systems*, 857-864.
3. Kulldorff, M. (1997). “A spatial scan statistic.” *Communications in Statistics-Theory methods, 26*(6): 1481-1496.
4. Pocock, S.J. (1977). “Group sequential methods in the design and analysis of clinical trials.” *Biometrika, 64*(2): 191-199.
5. Salton, G., Wong, A., and Yang, C.S. (1975). “A vector space model for automatic indexing.” *Communications of the ACM, 18*(11): 613-620.
6. van der Maaten, L. and Hinton, G. (2008). “Visualizing data using t-SNE.” *Journal of machine learning research, 9*(Nov): 2579-2605.

DOI:10.6814/NCCU201900953

附錄

附錄一、人文學者於各文本所挑關鍵詞

文本	人文學者挑選關鍵詞
新青年 第七卷	我們/他們/社會/勞動/工人/人口/問題/生活/政府/中國/組織/工作/自己/主義/青年/個人/人民/政治/制度/自由/地方/組合/自然/經濟/人類/戰爭/城市/國家/資本/時間/運動/國人/聯合/世界/法律/工廠/勢力/思想/科學/利益/階級/精神/生產/機關/工業/時代/希望/工資/發達/教育/女工/權利/目的/哲學/危險/勞工/文明/道德/你們/選舉/能力/國民/女子/發展/同盟/將來/合會/自治/自殺/罷工/互助/團體/人生/地位/理想/進步/歷史/限制/全國/競爭/心理/幸福/多數/事業/革命/價值/工錢/男子/機器/標準/壓迫/世紀/改造/生計/做工/工讀/分配/利害/生存/國際/男女/文化/全體/勞農/改良/財產/職業/平等/犧牲/女人/生命/責任/平均/農政/覺悟/民治/物質/保護/家庭/權力/觀念/公共/職工/金錢/意志/民族/宗教/平和/行政/罪惡/批評/婦人/憲法/文學/原則/秩序/婦女/報酬/進化/羅素/義務/解放/干涉/結婚/職務/收入/身體/供給/救濟/道理/工藝/失業/民國/生殖/法則/厚生/思潮/創造/衛生/獨立/人道/僱主/中央/支配/運命/預防/工場/本能/風俗/產業/工價/分子/成功/和平/是非/過去/正義/機械/人力/不幸/政策/人工/規則/慾望/歐戰/中華/失敗/信仰/消滅/貧窮/少數/貧困/擴張/知識/力量/攻擊/奴隸/侵略/共同/同業
新青年 第八卷	他們/我們/社會/主義/勞動/政府/資本/階級/政治/組織/工人/教育/國家/自由/革命/聯合/世界/經濟/人類/人民/個人/共產/生產/工作/戰爭/兒童/法律/精神/科學/運動/國人/團體/勢力/職工/組合/文學/女子/選舉/權利/知識/工業/國民/無產/進步/哲學/勞農/改造/創造/產業/幸福/價值/勞工/競爭/供給/物質/發展/民主/民族/理想/工會/平等/農民/藝術/權力/共和/利益/罷工/中央/和平/文化/國際/生存/家庭/財產/文明/宗教/改革/人生/公共/同盟/青年/人們/利害/解放/觀念/專制/心理/建設/民治/平民/強迫/獨立/責任/意志/保護/工團/支配/男子/進化/鄉村/罪惡/農人/

DOI:10.6814/NCCU201900953

	你們/政黨/女工/赤軍/衝突/快樂/攻擊/憲法/互助/婦女/風俗/宇宙/道德/黨員/犧牲/掠奪/開發/學生/少數/生活/俄國/制度/問題/地方/中國/自然/歷史/方法/思想/德國/羅素/學校/多數/事實/研究/希望/時間/英國/機關/工廠/時代/管理/事業/議會/全體/實業/職業/意見/政策/地位/機會/基礎/同業/成功/批評/失敗/危險/性質/男女/美國/小孩/努力/食物/理論/少年/世紀/分配/全國/範圍/經驗/集會/機器/軍隊/新聞/國內/行動/法國/宣傳/上海/城市/孩子/痛苦/提倡/歐洲/生物/智識/感情/聯盟/記者/工資/恐怖/消費/時期/製造/現在/反對/從前
人民日報	人民/中國/主義/美國/權利/國家/政府/合法/帝國/美帝/鬥爭/恢復/反動/和平/民主/殖民/種族/解放/會議/獨立/自決/發展/越南/中東/佔領/正義/領土/蘇聯/自由/猶太/黨委/主權/階級/侵犯/人權/無產/立法/保障/責任/不法/正當/犯罪/憲法/集體/法律/義務/基本/改革/代表/問題/集團/世界/要求/陰謀/社會/國際/群眾/革命/侵略/政權/婦女/統治/鎮壓/罷工/經濟/戰爭/公報/憲章/大國/資本/資源/危機/勞動/剝奪/幹部/生產/書記/法院/規定/中央/公安/違法/教育/選舉/運動/紀律/政治/自治/行使/全國/建設/侵害/原則/利益/秩序/衝突/兒童/法制/宣傳/外交/律師/檢察/殘疾/計劃/任務/依法/公約/法規

附錄二、《新青年》關鍵詞選取結果

使用模型	模型挑選關鍵詞
本文方法	一九/一方面/一個/一條/一種/人民/人們/土地/大會/大學/女人/女子/小說/工人/工作/工廠/不是/不能/不過/中央/中間/之中/什麼/公司/分配/反對/支配/文化/文學/方法/他們/代表/召集/可以/可是/可能/布爾塞維克/平民/必要/民族/生存/生活/生產/先生/全世界/全國/全體/共產/共產主義/共產黨/危險/同時/因此/因為/地主/多數/好像/如此/如何/如果/有產階級/自己/自由/自然/似乎/但是/佈告/作用/你們/兵士/利用/形式/快樂/我們/改革/決定/沒有/男女/男子/身體/供給/

DOI:10.6814/NCCU201900953

	兒童/制度/和平/委員/委員會/宗教/幸福/性質/或者/所以/所有/承認/於是/武力/法律/物質/知識/社會/社會主義/社會黨/俄國/俄羅斯/信仰/思想/政府/政治/政策/甚麼/科學/英國/計畫/軍隊/革命/食物/個人/原因/原則/哲學/家庭/時候/時間/消滅/財產/做工/問題/國民/國家/國際/掠奪/教育/教員/現在/現象/理由/理論/產業/第一次/組織/莫斯科/許多/這次/這個/這種/這樣/創造/勞工/勞動/勞動者/勞農/就是/減少/無政府/無產/然而/發展/發達/進化/進行/階級/勢力/會員/會議/當時/解放/資本/資本主義/資本家/農人/農民/農政/農業/運動/過去/團體/實業/對於/說明/增加/影響/德國/學生/學校/學說/憲法/曉得/機器/機關/歷史/獨立/選舉/聯合/聯絡/聲音/職業/羅素/藝術/關係/競爭/蘇維埃/蘇維埃政府/議員/議會/犧牲/權力/權利/變化/觀念
羅吉斯迴歸	一九/一天/一方面/一句/一次/一定/一般/一部/一部分/一種/一樣/七年/人民/人生/人們/人類/上面/土地/大家/大會/女子/小說/工人/工作/工廠/工錢/不但/不要/不能不/不曾/不會/中國/中間/之中/互助/互相/公共/反對/少年/少數/手段/支配/文化/文學/方面/世界/世紀/主張/主義/他們/代表/可是/可能/外國/失敗/平民/平等/必要/打破/民族/生命/生物/生活/生產/由於/先生/全世界/全國/全體/共同/共產/共產主義/共產黨/危險/各國/各處/各種/地方/地主/地位/多數/好像/如何/如果/成績/而且/自由/至於/似乎/似的/但是/作工/你們/利用/利害/告訴/困難/形式/我們/改革/改造/決定/沒有/男女/男子/身體/那些/那裡/那麼/事件/事情/事業/事實/供給/兒童/兩個/兩種/其中/制度/和平/固然/委員/委員會/宗教/幸福/性質/所有/承認/朋友/武力/法子/法律/物質/知道/知識/社會主義/社會黨/非常/便是/俄國/俄羅斯/保存/卻是/很大/很多/後來/思想/政府/政治/政策/活動/甚麼/看見/研究/科學/英國/要求/計畫/重要/限制/革命/個人/原因/哲學/家庭/效果/時代/時候/時間/真的/真是/秩序/討論/財產/起來/馬克思/做工/問題/國民/國家/國會/國際/推翻/教育/現在/現象/理由/理想/理論/產業/第一個/第二/組合/組織/莫斯科/規定/逐漸/這是/這樣/造成/創造/勝利/勞工/勞動/勞動者/勞農/喜歡/就要/幾個/減少/無產/然而/發生/發表/發展/發達/程度/等到/進化/進行/階級/集合/傳播/勢力/會議/當時/經濟/義務/解放/解釋/資本/

DOI:10.6814/NCCU201900953

	資本主義/資本家/農奴/農民/農業/運動/過去/預備/團體/實在/實行/實現/實業/演說/漸漸/種種/算是/管理/精神/說明/說道/需要/價值/影響/德國/標準/確實/罷工/學生/學者/學校/學說/戰爭/曉得/機會/機關/歷史/獨立/選舉/應當/應該/聯合/聯合會/聯絡/職工/職業/羅素/證明/關係/願意/競爭/蘇維埃/蘇維埃政府/覺得/議會/屬於/犧牲/權力/權利/聽見/變遷/觀念
LDA	一九/一方面/一個/一種/一樣/人民/人類/土地/大家/大會/女子/工人/工作/工廠/工錢/不但/中國/文學/方法/方面/世界/主張/主義/他們/代表/平民/民族/生物/生活/先生/全國/共產/共產主義/共產黨/同業/地主/多數/如果/而且/自由/但是/形式/我們/那些/那裡/事情/事業/事實/供給/兒童/制度/委員/委員會/宗教/幸福/所有/法律/物質/知道/知識/社會/社會主義/社會黨/便是/俄國/思想/政府/政治/政策/看見/科學/英國/重要/革命/個人/哲學/家庭/時候/時間/財產/馬克思/國民/國家/教育/產業/組合/組織/莫斯科/這樣/勞工/勞動/勞動者/勞農/無產/發展/進行/階級/會議/當時/經濟/資本/資本主義/資本家/農民/農業/運動/團體/實行/實業/管理/德國/學生/學校/曉得/機關/歷史/獨立/選舉/應該/聯合/聯合會/職工/職業/羅素/關係/蘇維埃/覺得/議會/犧牲/權力/權利
SVM	一九/一切/一方/一方面/一日/一句/一件/一次/一面/一班/一條/一部/一樣/一點/七月/七年/人生/人們/人類/十一月/三個/上海/大家/大會/大學/女工/小孩/小說/已經/不同/不知/不是/不能不/不曾/不會/不獨/中央/中間/之中/之後/互助/互相/公司/公共/分子/分配/反對/少年/心理/心裡/手段/支配/文化/主人/主張/以上/以及/以外/以前/以後/以為/加入/只要/叫做/召集/可以/可能/外國/失敗/布爾什維克/布爾塞維克/平等/打破/民主/民治/民族/生存/生命/生產/全俄/全國/全體/共和/共產/共產主義/共產黨/危險/各處/同時/地主/地位/多數/好像/如此/如果/存在/成為/有人/有產階級/自己/自由/自從/至於/似乎/似的/何以/你們/別的/利用/利益/努力/困難/完全/形式/快樂/批評/改良/改革/改造/攻擊/決定/男子/見解/赤軍/那些/那時/那種/那樣/事件/事務/事情/事實/例如/供給/兩個/和平/固然/委員/宗教/幸福/忽然/或者/所有/所有權/所說/所謂/承

DOI:10.6814/NCCU201900953

	認/朋友/武力/法國/注意/物質/狀況/狀態/直接/社會黨/保存/保障/保護/信仰/卻是/城市/宣傳/建設/很多/後來/政策/是否/活動/看見/看來/英國/要求/計畫/軍隊/限制/革命/風俗/食物/原則/家庭/容易/差不多/恐怖/時間/根本/特別/真正/真的/真是/耕作/能力/能夠/討論/財產/做工/動物/國會/專制/常常/強迫/得到/掠奪/採用/推翻/教員/現象/理由/理論/產生/眼睛/第一次/第二/組合/莫斯科/規定/設立/這些/這時候/這裡/造成/部分/創造/勞工/喜歡/報酬/就是/就要/幾個/提倡/普通/普遍/普魯士/智識/無政府/然而/痛苦/發達/程度/等等/進化/進步/鄉村/集會/傳播/勢力/意見/會員/極力/當時/當然/經過/經營/罪惡/解決/解放/資本/資本主義/農人/農政/過去/道德/實現/實業/態度/維持/製造/說明/需要/領袖/價值/範圍/罷工/調查/學生/憲法/機器/獨立/辦法/選擇/選舉/壓制/幫助/應用/應當/應該/環境/聯合/聯合會/聯盟/聲音/還有/還要/簡直/雜誌/藝術/證明/贊成/關於/競爭/覺得/議會/黨員/犧牲/權力/聽見/變化/變遷/觀念

附錄三、《人民日報》關鍵詞選取結果

使用模型	模型挑選關鍵詞
本文方法	一月/一個/人民/十月/土地/小時/工人/工作/工會/工資/工廠/干涉/不結盟/中東/中東問題/中國/中華人民共和國/公司/公民/公報/六月/反動/反對/巴勒斯坦/巴勒斯坦人/巴勒斯坦解放組織/巴解組織/支持/日本/主義/他們/代表/代表團/以色列/可以/四人幫/外長/它們/民主/民族/民族團結/生產/石油/示威/企業/同志/安理會/成員國/自己/行為/佔領/我們/我國/決議/兩個/制度/協定/委員會/所謂/法制/法律/社會主義/表決/阿拉伯/阿拉法特/阿爾及利亞/阿爾巴尼亞/青年/非洲/侵犯/南非/宣言/帝國/建設/建議/恢復/政府/政權/柬埔寨/約旦/美帝國/美國/要求/革命/埃及/草案/鬥爭/問題/國家/國際/婦女/控制/敘利亞/教育/第三世界/組織/規定/這一/這個/陰謀/傀儡/勞動/提案/朝鮮/猶太/發展中國家/超級大國/越南/進行/集會/集團/幹部/意見/會議/經濟

DOI:10.6814/NCCU201900953

	/群眾/義務/解放/農民/監督/種族/管理/領導/罷工/蔣介石/黎巴嫩/學生/憲法/憲章/戰爭/機關/選舉/總統/聯合/聯合國/聯合國大會/舉行/職工/雙方/蘇聯/黨委/辯論/驅逐/權利
羅吉斯迴歸	一切/人民/人權/口號/土地/大會/小時/工人/工作/工會/干涉/中國/中華人民共和國/公司/公民/反映/巴勒斯坦/巴勒斯坦解放組織/日本/世界/主義/主權/他們/代表/代表團/以色列/充分/去年/可以/它們/民主/民主權利/民族/民族團結/生產/石油/企業/全國/各國/合法權利/同志/地方/有權/自由/行使/行為/佔領/我們/我國/決議/兩個/制度/委員會/所謂/法律/直接/社會/社會主義/阿拉伯/非洲/侵犯/侵略/南非/宣言/帝國/建設/建議/政府/政權/柬埔寨/美國/軍隊/重申/重要/革命/埃及/討論/鬥爭/問題/國家/堅決/婦女/敘利亞/第三世界/組織/規定/這一/陰謀/勞動/提案/猶太/發展/發展中國家/超級大國/越南/進行/集團/幹部/意見/當局/經濟/群眾/義務/農民/實行/種族/領導/領導人/罷工/學生/憲法/憲章/歷史/獨立/選舉/應當/應該/總統/聯合/聯合國/舉行/職工/蘇聯
LDA	人民/人權/土地/工人/工作/中國/中華人民共和國/公司/公民/反映/巴勒斯坦/巴勒斯坦解放組織/日本/主義/他們/代表/代表團/以色列/可以/它們/民主/民主權利/民族/石油/企業/各國/合法權利/同志/地方/行為/佔領/我們/我國/決議/兩個/委員會/法律/社會主義/阿拉伯/非洲/侵犯/南非/帝國/建議/政府/政權/柬埔寨/美國/要求/埃及/鬥爭/問題/國家/國際/婦女/組織/規定/這一/提案/越南/進行/集團/幹部/意見/會議/當局/經濟/群眾/農民/領導/罷工/學生/憲章/獨立/選舉/總統/聯合/聯合國/舉行/職工/蘇聯
SVM	一定/一般/一票/一項/二十六日/二月/人民代表大會/人民共和國/力量/十一月/十六日/口號/土地/大國/小組/山西省/工會/干涉/才能/不合理/不法/不得/內容/公正/公安/公社/分配/反映/反動/巴西/巴勒斯坦人/巴基斯坦/文化/文件/方面/日內瓦/水準/主任/主要/主張/代表權/加強/包括/北愛爾蘭/去年/可是/四人幫/外交部/外交部長/外長/外國/它們/巨大/布邁丁/平等權利/必要/本國/正式/正常/正義事業/正義鬥爭/民主/生活/由於/申訴/示威/示威遊行/立場/全廠/共

DOI:10.6814/NCCU201900953

	同/各地/合法/合法政府/同志/同意/地位/地區/安全/成員國/有些/有權/自由/自決/自然資源/行使/佔領區/作用/你們/別國/利益/完成/形式/技術/抗議/更加/事件/事業/亞運會/使用/供應/其中/制止/取消/居民/所謂/林彪/法紀/法院/法國/爭取/社員/阻撓/阿拉伯人/阿富汗/阿爾及利亞/保障/保證/南朝鮮/宣傳/政權/研究/美帝/美帝國/美蘇/軍事/重申/剛果/原料/時間/書記/朗諾/泰國/消滅/破壞/記者/起來/偉大/停火/唯一/國家元首/國家機關/基本/專政/強調/採取/接待/控制/推行/推遲/敘利亞/條例/棄權/現代化/現在/產品/統一/處理/責任制/這是/這家/通貨膨脹/部長/部門/勞動/尊重/就是/就業/提出/提供/提高/揭發/猶太/發生/發言/發展中國家/等國/結束/開發/開幕式/階段/集團/黑人/意見/搞好/當局/經濟秩序/經驗/資本主義/資本家/資金/路線/運動員/達到/違反/隔離/團結/實際/種族/管理/認識/增加/審議/影響/數據庫/緩和/罷課/賣國/憲法/戰鬥/機構/機關/選出/錯誤/應當/應該/總理/擴大/擴張主義/壟斷資本/穩定/羅馬尼亞/難民/繼續/蘇聯/警察/議題/黨中央/黨內/黨章/辯論/體育

DOI:10.6814/NCCU201900953