文獻回顧

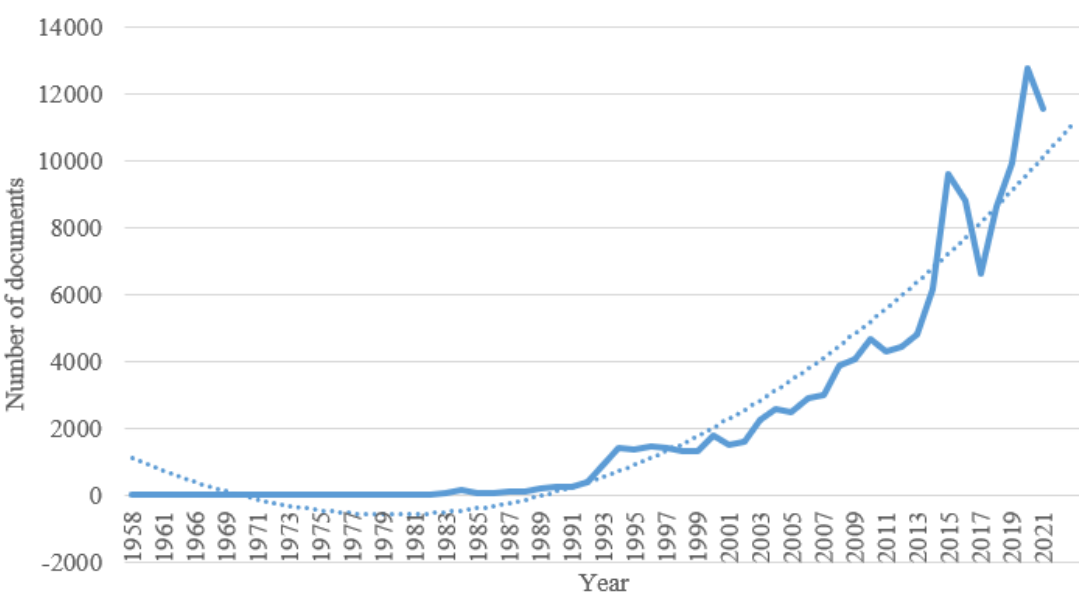
第三節、文本探勘研究回顧

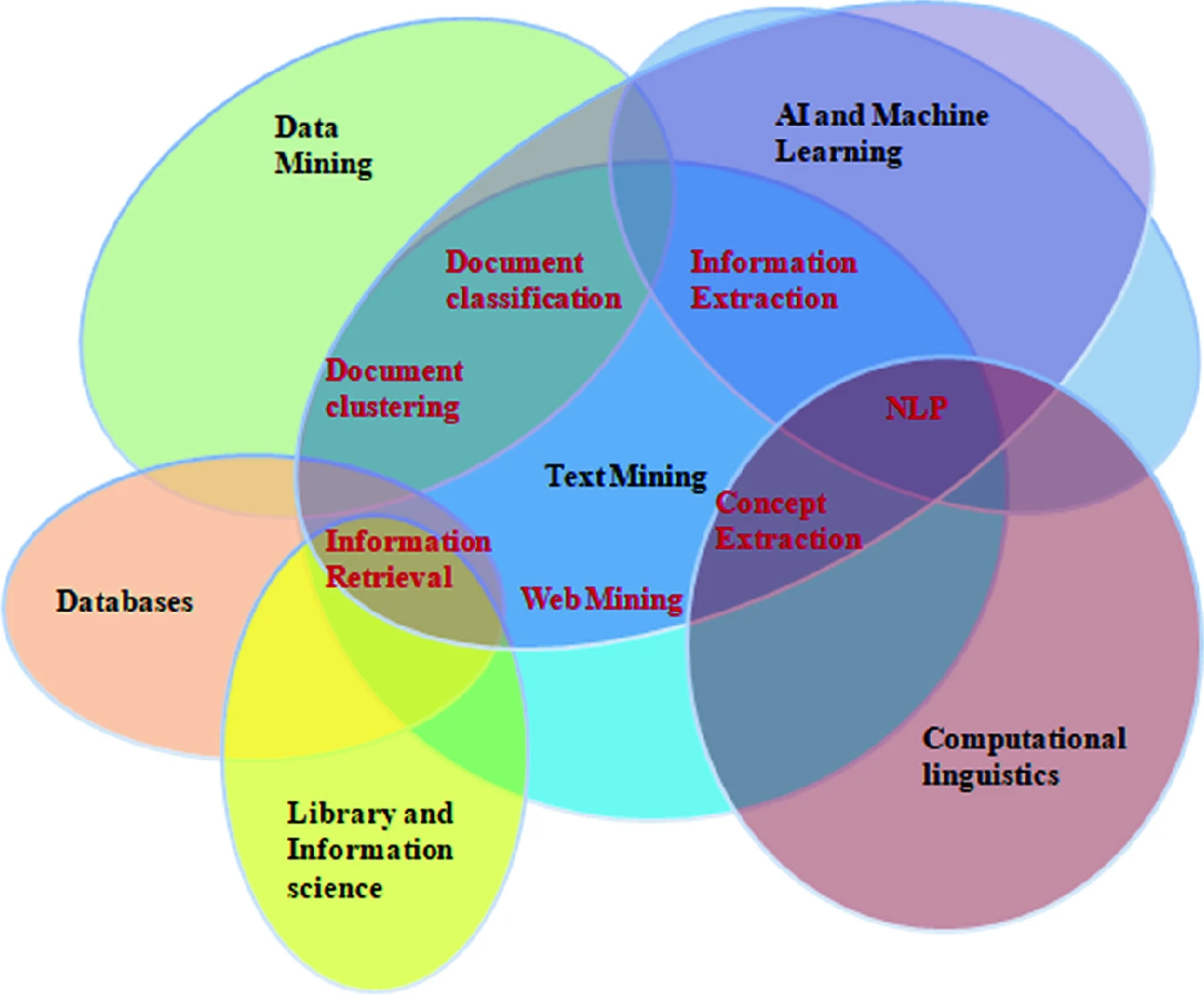
延續前節文獻整理，本節爬梳文本探勘應用在社會科學領域文獻，整理媒體框架與隱含立場的探勘研究，以及台媒現有的探勘研究與應用。

一、社會科學探勘研究

文本探勘研究除了保有量化研究的大量運算、可重複操作實驗再現性、串接並比較不同作用演算法等良好性質，文本資料亦可做為次級資料做質性研究，其演算法實驗結果也擁有質性研究的結構性分析。機器從文本數據向上提供資料特性，研究者從結果與先驗知識向下理解現象，深化質量研究混和的解釋能力。為使研究「有質有量有多聞」，據紐文英(2017)<研究方法與論文寫作>的結合量化與質性方法之研究章節提到，需考量研究問題結合目的與特徵，從研究方法面得知社會科學研究適合宏觀層面分析，相當符合「大量-大質設計」(QUAN+QUAL)的研究方法，亦即先採量化資料蒐集資料並做量化分析且交錯質性分析，最後給出整合解釋。不止於此，量化研究中的文本探勘領域更適合社會科學研究，係因原始資料多以文本留存數位資料，演算法可接受個體訪談紀錄、複數文稿紀錄、異質文本資料等僅需資料以文字結構留存即能應用，也伴隨自然語言處理技術演進，更強大算力機器與開源開放預訓練模組來輔助探勘。(Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. Qualitative Report, 16(3), 730-744.、McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. (2018). What doesn’t get measured does exist: Improving the accuracy of computer-aided text analysis. Journal of management, 44(7), 2909-2933.、Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. R&D Management, 50(3), 329-351.)

近年計算機算力大幅提升，借助機器幫助的研究方法興起迅速。文本探勘與自然語言處理高度相關也一同蓬勃發展，Gao等人(Gao, Y., Dligach, D., Christensen, L., Tesch, S., Laffin, R., Xu, D., ... & Afshar, M. (2022). A scoping review of publicly available language tasks in clinical natural language processing. Journal of the American Medical Informatics Association, 29(10), 1797-1806.)計數從1958年至2021年的自然語言處理論文出版數，如下圖出版趨勢的斜率依舊增加當中，如下圖由Rajendra等人(Roul, R. K., Varshneya, S., Kalra, A., & Sahay, S. K. (2015). A novel modified apriori approach for web document clustering. In Computational Intelligence in Data Mining-Volume 3: Proceedings of the International Conference on CIDM, 20-21 December 2014 (pp. 159-171). Springer India.)所繪製文字探勘包含許多計算機技術，這些技術包括人類主觀給予機器答案訓練，也包括機器自行彙整規則與結構，也包括人類答案與機器歸納之間再次整理。





本文使用不少技術完成，文本資料也由筆者自行爬蟲取得，Information retrival技術與知識幫助拆解段落、語句、詞彙之間關係，Document clustering在大量文章借助機器幫助下歸納結構，尤關本文注重的內容分析則由自然語言處理(Natural language process)、Information extraction以及Concept Extraction完成。內容分析途徑在Neuman(2006)與紐文英(2017)與潘淑滿(潘淑滿. (2022). 質性研究: 理論與應用. 心理.)提到資料不論文字、意義、圖畫、觀念、主題或其他任何紀錄訊息，可明確描述資料內容的特徵並再應用一些規則輔以定義這些特徵，最後給予分類與其架構進一步解釋。由此可知人類短期無法閱讀大量訊息限制下，借助計算機與演算法歸納特徵並找出規則與結構，進一步作出解釋尤為適合。(Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications and issues. International Journal of Advanced Computer Science and Applications, 7(11).、Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. R&D Management, 50(3), 329-351.)

在社會科學借助文字探勘的研究不少，整理自Lazer等人(Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., ... & Van Alstyne, M. (2009). Computational social science. Science, 323(5915), 721-723.)與Kudyba等人(Kudyba, S., & Kudyba, S. (2014). Big data, mining, and analytics. Boca Raton: Auerbach Publications.)與Roiger(Roiger, R. J. (2017). Data mining: a tutorial-based primer. Chapman and Hall/CRC.)與Antons等人(Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. R&D Management, 50(3), 329-351.)的文獻回顧，應用包含選舉調查、政策與民意探勘、經濟情勢探勘、網民意見調查、數位化歷史文件探勘、多模態(Modality)應用(幫我腳註文字配合其他資料如影像與聲音一同訓練模組)、大量次級資料探勘、媒體探勘等族繁不及備載，僅列出與社會科學高度相關與顯著使用部分。上述使用文字探勘技術有幾項特性，包含探究母體真實意見、增加決策品質與選擇手段、減少人力介入或少量樣本造成偏誤、樣本或調查成本低廉等特性。

儘管上文提到文字探勘帶來許多好處，仍需注意技術限制與過分推論可能。Bock提到蔚為風行的文字雲在建立圖文僅憑出現頻率，將可能導致錯誤推論(Bock, M. A. (2009). Impressionistic content analysis: Word counting in popular media. The content analysis reader, 38-41.)。Kononova等人認為(Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E. A., & Ceder, G. (2021). Opportunities and challenges of text mining in materials research. Iscience, 24(3), 102155.)文字探勘面臨不少挑戰，尤其語言複雜度與累加偏誤(Language complexity and error accumulation)將帶來更大問題，理由在於資料處理與資料清洗有不同手法，配合多元且多種演算法再搭配多種參數，種種誤差累積加上原文與結果重新解釋時，無法保證推論一致性。陳世榮((陳世榮 (2015)。〈社會科學研究中的文字探勘應用: 以文意為基礎的文件分類及其問題《人文及社會科學集刊》，27(4):683-718〉)提到語用學的「言外之意」以及「語言行為」（speech act）等細微且複雜的表達，儘管機器歸納收斂規則，也不易連結社會科學注重的論述與溝通部分。謝舒凱等人(謝舒凱 & 曾昱凱 (2019). 深度詞庫: 邁向知識導向的人工智慧基礎. 中華心理學刊, 61(3), 231-247.)也認為當前自然語言處理與文字探勘的基礎詞庫是建構在「形式」與「意義」的配對關係上，語言知識在過於專注在工程模式(Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.)，蕭惠貞等人也認為(蕭惠貞, 詹士微, & 陳瀅妤. (2022). 人工智慧學習平台之教學應用反思——以法律華語文本為例. 臺大華語文學習與科技, 2(1), 107-143.)語法知識在建構模組與演算法相對貧脊。人類閱讀過程當文字形式改變其意義也隨之改變，但是傳統機器評量基準多以XY(自變數與應變數)對應決定高低，如翻譯任務驗證效度模組透過BLEU(bilingual evaluation understudy)分數表達，而BLEU成績高低是透過兩組序列重現程度而定，以致在一文多義的豐富語言知識下，機器被迫記憶一對一對應關係而不易學習多義能力。

二、媒體探勘相關研究

文本探勘應用於媒體至今已有不少案例與研究。早在兩千初期網路興起年代，已有新聞分類媒體(Agrawal, R.; Rajagopalan, S.; Srikant, R.; Xu, Y. Mining newsgroups using networks arising from social behavior. In Proceedings of the 12th International Conference on World Wide Web(WWW2003),Budapest,Hungary, 20 May 2003; pp. 529–535, doi:10.1145/775152.775227)、政治新聞情緒分類(Young, L.; Soroka, S. Affective news: The automated coding of sentiment in political texts. Polit. Commun.2012, 29, 205–231, doi:10.1080/10584609.2012.671234)、多語言摘要(. Evans, D.K.; Klavans, J.L.; McKeown, K.R. Columbia newsblaster: Multilingual news summarization on the web. In Proceedings of the Demonstration Papers at HLT-NAACL, Boston, MA, USA 2–7 May 2004.Available online: https://www.aclweb.org/anthology/N04-3001)等應用，到了2010時期發展相當活絡，連同大數據帶動發展的處理速度，不僅帶動更多新聞探勘應用，也深化至專項領域的文本(Tsirakis, N.;Poulopoulos, V.; Tsantilas, P.; Varlamis, I. Large scale opinion mining for social, news and blog data. J. Syst. Softw. 2017, 127, 237–248, doi:10.1016/j.jss.2016.06.012.、Za’in, C.; Pratama, M.; Lughofer, E.; Anavatti, S.G. Evolving type-2 web news mining. Appl. Soft Comput. 2017, 54, 200–220, doi:10.1016/j.asoc.2016.11.034、Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. Big Data and Cognitive Computing, 4(1), 1.)，最常見探勘領域包括股票、管理、社群媒體、媒體情感、旅遊網站評論等(Hagenau, M.; Liebmann, M.; Neumann, D. Automated news reading: Stock price prediction based on financial news using context-capturing features. Decis. Support Syst. 2013, 55, 685–697, doi:10.1016/j.dss.2013.02.006、Chen, W.; Yeo, C.K.; Lau, C.T.; Lee, B.S. Leveraging social media news to predict stock index movement using RNN-boost. Data Knowl. Eng. 2018, 118, 14–24, doi:10.1016/j.datak.2018.08.003、Iglesias, J.A.; Tiemblo, A.; Ledezma, A.; Sanchis, A. Web news mining in an evolving framework. Inf. Fusion 2016, 28, 90–98, doi:10.1016/j.inffus.2015.07.004.、Chen, M.-Y.; Chen, T.-H. Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena. Future Gener. Comput. Syst. 2019, 96, 692–699, doi:10.1016/j.future.2017.10.028、Loureiro, S. M. C., Guerreiro, J., & Ali, F. (2020). 20 years of research on virtual reality and augmented reality in tourism context: A text-mining approach. Tourism management, 77, 104028.)，由於新聞與媒體作為重要資訊媒介，不僅許多小至民生水電大至國家政治，以及經常被閱聽人關注的假新聞現象(Chen, Y.; Conroy, N.J.; Rubin, V.L. Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, Seattle, WA,USA,1 August 2015; pp. 15–19.、Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2), 211-236.、Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. Advances in neural information processing systems, 32.)。近年2020區間由人工智能再次興起一波熱潮，文本相關領域研究不僅搭配機器學習技術也借助預訓練模組，提升以往過去表現不好的任務，也增加更多落地案例。依據Bettina Berendt(Berendt, B. Text mining for news and blogs analysis. In Encyclopedia of Machine Learning and Data Mining; Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2017; pp.1247–1255,doi:10.1007/978-1-4899-7687-1)與Hassani(2020)等人整理文字探勘與機器學習研究領域使用在新聞媒體的技術，首先是適用重點信息擷取(Information extraction、Arguments Extraction)(Goudas, T.; Louizos, C.; Petasis, G.; Karkaletsis, V. Argument extraction from news, blogs, and social media. Int. J. Artif. Intell. Tools 2015, 24, 287–299, doi:10.1142/S0218213015400242.、Yang, S., Feng, D., Qiao, L., Kan, Z., & Li, D. (2019, July). Exploring pre-trained language models for event extraction and generation. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 5284-5294).)可以做到大量文本的重點抽離、名詞網絡圖關係、知識網絡圖、關鍵字擷取等項目。其次非監督機器學習技術應用於社群媒體與部落格，例如推特(Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and research: A systematic literature review through text mining. IEEE access, 8, 67698-67717.)與臉書(Text Mining and Analytics: A Case Study from News Channels Posts on Facebook, https://doi.org/10.1007/978-3-319-67056-0\_19)、Youtube(Can, U., & Alatas, B. (2019). A new direction in social network analysis: Online social network analysis problems and applications. Physica A: Statistical Mechanics and its Applications, 535, 122372.)等各大社群平台的官方推文經常是研究者分析熱點，其相應爬蟲與反爬蟲與軟爬蟲技術也是近年高度關注的技術之一（Chen, S., Cao, Y., Sarparast, M., Yuan, H., Dong, L., Tan, X., & Cao, C. (2020). Soft crawling robots: design, actuation, and locomotion. Advanced Materials Technologies, 5(2), 1900837.），透過降維、分群、主題分析等歸納目的的作業，可容易歸納出經常出現的主題，或是歸納經常出現的詞彙，藉此在大量文本中萃取一些隱含在文章的背後意圖。

儘管有大量研究與文獻探勘新聞與社群媒體文本，但仍然要注意發展帶來問題。Coletto等人(Coletto, M., Garimella, K., Gionis, A., & Lucchese, C. (2017). Automatic controversy detection in social media: A content-independent motif-based approach. Online Social Networks and Media, 3, 22-31.)認為網路有著多樣多量資訊可以利用，但也過分自由而容易出現爭議性主題，因此社群媒體與新聞平台多有自己一套的爭議性發言或爭議性主題偵測。前文提到機器學習能力，也加強了機器造假能力(Westerlund, M. (2019). The emergence of deepfake technology: A review. Technology innovation management review, 9(11).、Pulido, C. M., Ruiz-Eugenio, L., Redondo-Sama, G., & Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in health. International journal of environmental research and public health, 17(7), 2430.、Rocha, Y. M., de Moura, G. A., Desidério, G. A., de Oliveira, C. H., Lourenço, F. D., & de Figueiredo Nicolete, L. D. (2021). The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. Journal of Public Health, 1-10.)，甚至是刻意攻擊社群網站，Zhou等人（Zhou, Z., Guan, H., Bhat, M. M., & Hsu, J. (2019). Fake news detection via NLP is vulnerable to adversarial attacks. arXiv preprint arXiv:1901.09657.）提到三種攻擊機器學習模組的方式，包括失真事實、錯置名詞、混淆原因等，使得模組學習到錯誤資訊並誤導。感覺不好寫入，你斟酌放入吧，或是穿插前文

國外研究成果豐碩，相對於國內也有文字探勘應用於新聞媒體。羅珮怡等人(2020)運用BERT探究台灣新聞媒體如何看待美國選舉，文字探勘（林顯明，2015）

過去對於新聞文本內容，多利用研究者事先定義議題的框架，近年來越來越多研究（羅珮怡、蔣旭政，2022；）使用文字探勘技術，對新聞文本進行自動化分析。余清祥、葉昱廷（2019）以文字探勘技術分析台灣四大報《蘋果日報》、《自由時報》、《聯合報》、《中國時報》的報導風格，

Jiang, Y. (2021, April). Diachronic Lexical Features of Chinese Government Work Reports: A Text Mining Approach. In 2021 6th International Conference on Social Sciences and Economic Development (ICSSED 2021) (pp. 735-738). Atlantis Press.

邵軒磊, & 曾元顯. (2018). 文字探勘技術輔助主題分析—以 [中國大陸研究] 期刊為例. 問題與研究, 57(1), 29-62.)

所以在計算機快速發展的近年，許多研究領域受益機器輔助，仍舊注意

研究設計

本文分析對象為各大網媒新聞文本，以文字探勘為本文主要分析工具，依循常見文字探勘研究流程與演算法應用，採用Python程式語言作為主要開發語言。本章第一節資料處理描述文本蒐集過程、爬蟲細節以及文本前處理，第二節研究流程與架構說明本文分析流程，以及對應分析工具的預期結果，第三節研究方法說明為何使用研究工具。

第一節、 資料處理

資料來源是由網路爬蟲而來，下載來源鎖定於重點台媒的新聞文本。相對紙媒更有傳播力的網媒，本文善用各大網媒網站的搜尋功能，搜尋內文一旦提及「維吾爾」、「新疆」兩詞彙的文本皆下載。以下本節敘述爬蟲細節與其限制，以及文本進入分析工具前的資料清洗。

本文蒐錄新聞文本有數點前提與限制，如下列條：

一、關鍵字搜尋

本文將搜尋關鍵字僅定調於維吾爾與新疆兩個詞彙，係因各新聞網站搜尋入口的搜尋條件不一，以及非特指意義的關鍵字會帶來更多非關維吾爾人的資料。各新聞網站搜尋大多僅可接受單詞輸入，若採用「棉花」與「抵制」兩詞取空白隔開並搜尋，網站可能回饋交集搜尋結果或是聯集搜尋結果，並且各家網站進階搜尋功能不盡相同，為了取得所有網站一致性而採單個關鍵字下載。

本文採單關鍵字並輔以多次爬蟲盡搜完整相關文本，但限制在於「棉花」與「抵制」分開將過分蒐集不相干文本，故無法使用組合搜尋。反之「七五事件」、「東伊運」、「疆獨」、「烏魯木齊大火」等詞相當具有單獨意義，儘管可以逐次爬蟲蒐錄，但本文不帶主觀條件設定更多關鍵字避免樣本偏誤，例如「東伊運」、「疆獨」倆關鍵字將涉及獨立運動、反恐運動、反分裂主義等其他非關文本。

二、以2017年後發稿時間為限

儘管筆者盡可能將各媒體文本不分時間全部蒐錄，其中發現早期關於新疆的新聞文本皆以風俗民情、地方建設、地方官員上下任、特色料理等，並非本文關注的媒體框架與媒體立場。近年來新疆與維吾爾人事件不斷，越是受世界關注，媒體越是更多投入與報導產出，故本文聚焦近年新聞文本，特別是在聯合國開始關注維吾爾人相關議題之時。

三、新聞文本附帶資訊

舉單篇新聞網站版面為例，各網站提供附加功能也不盡相同，提供互動功能諸如留言功能、留言底下再次留言功能、按讚與表情符號、分享新聞閱讀網址、連結主流社群媒體、其他相關新聞連結推薦等。反之有些新聞版面相對精簡，例如文章編責未出現內文結尾、僅文本無任何附加功能、僅影片介紹無文字、僅廣告影片無文字等情況。網站提供服務不一情況下，筆者取最大共同附帶資訊下，單篇文本將包含新聞標題、發稿時間、新聞內文、新聞連結等四項資訊。

在上述三點爬蟲限制之下，筆者使用Selenium(腳註https://www.selenium.dev/documentation/webdriver)、BeautifulSoup(腳註https://beautiful-soup-4.readthedocs.io)、Requests(腳註https://requests.readthedocs.io)等套件執行爬蟲作業，盡可能下載所有新聞貼文。網站大多具有搜尋功能，由使用者點擊興趣新聞是透過搜尋後網站回傳新聞標題與部份內文，使用者點擊連結並閱讀內文，故本文採取同樣模式先搜尋關鍵字並取得時間、標題、連結，再隨著連結下載內文。網頁資料是採用html(HyperText Markup Language)標記語言所呈現，並搭配CSS(Cascading Style Sheets)階層式樣式表敘述的風格介面語言強化視覺效果，也搭配JavaScript互動網頁語言與用戶做互動，所以需要Selenium執行人類動作，輔以BeautifulSoup擷取靜態html資料，完成擷取動作。

本文研究多家媒體，任一網站將有多種新聞排版，連帶html與CCS也會不同，代表爬蟲必須各自獨立處理排版。筆者不一一陳列所有網站如何定位到指定欄位並取得資料，僅概括陳述步驟。首先筆者採用selenium與webdriver(腳註https://chromedriver.chromium.org/downloads)模擬使用者在瀏覽器上執行的鍵盤與滑鼠動作，經使用者滑鼠點擊或鍵盤輸入搜尋字後，網頁透過JavaScript進一步將資訊由使用者用戶端傳送到伺服器後端資料庫做查詢，再將查詢結果傳回使用者用戶端網頁展示，最後筆者拆解用戶端取得網頁資料並解析，透過BeautifulSoup取得含有資料的html，依序得到每篇報導網址、標題、發稿時間。最後筆者輪循所有報導網址，下載內文的同時，刪除廣告與非內文雜訊。

儘管爬蟲有跡可循，仍需耗費不同媒體與不同風格各自完成，亦需處理反爬蟲作業。筆者在selenium執行數個人類閱讀習慣，如隨機秒數等待、隨機按下F5網頁更新、按下page down按鍵、按下scoller、按下end按鍵、滾動滑鼠wheel、滑鼠隨機hover於網頁某一處等操作。筆者也需處理不同網站的反爬蟲機制，諸如報導網址加密、被網站偵測為機器人(單一IP訪問次數過多遭禁止)、是否機器人詢問視窗、混亂html標籤、iframe網站設計、拉長回應時間等，需要多次客製的反反爬蟲手段解決。

所幸本文意圖分析媒體網站皆能克服下載問題，為求取得更佳分析結果，本文將統一處理所有新聞內文格式，資料清洗包括如下。

一、全形轉半形

二、大寫轉小寫

三、regular expression(腳註:https://docs.python.org/3/library/re.html)處理

四、斷詞處理

五、停用字處理

六、同義字處理

七、隨機抽取檢查前處理結果

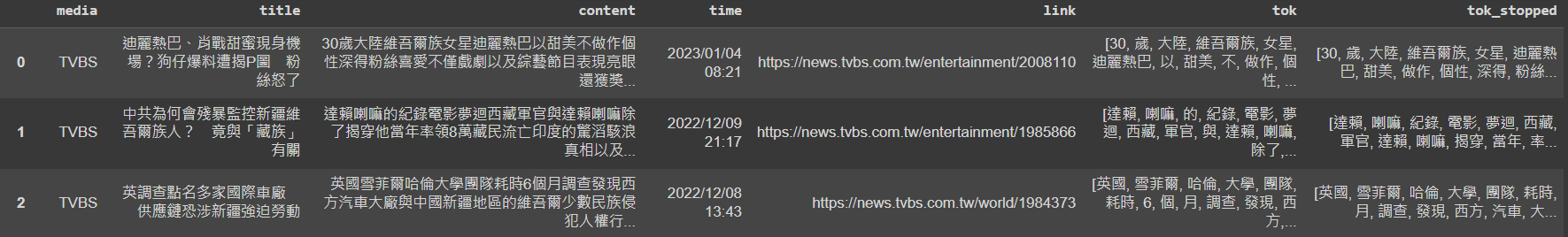
資料清洗的一至三項為字元級別處理，係因英文與標點符號會影響後續斷詞處理，影響後續計數為重的聲量分析，亦影響後續使用主題分析。當文章已經執行過轉半形以及小寫處理後，為求通順，透過Python套件中最常使用正規化處理的re套件，再次將語句整理。

三之一、整理所有連續兩個以上的跳脫符號(腳註說明是電腦處理排版格適用的特別符號\r 、 \n 、 \t)為一個空白

三之二、整理所有連續兩個以上的空白符號(腳註一些特別空白\xa0、&nbsp)為一個空白

三之三、整理所有連續兩個標點符號為一個中文逗號(，)

延續說明資料清洗四到七。係因研究工具所需，將中文句子內詞彙以空白斷開，形成英文語句書寫形式，並盡可能將此斷詞處理的結果優化。筆者自定義添加中英文同義字以及中英文停用字去蕪存菁，借助colab(腳註:https://colab.research.google.com/)定額免費GPU(Graphics Processing Unit)的幫助加速斷詞。最後步驟七，如下內文部份截圖顯示，筆者隨機顯示數則新聞檢查，依據斷詞後結果持續增添同義字與停用字，以及去掉對本研究無意義語句(腳註:https網址、不代表新聞台立場、新聞台關心您、遺漏廣告詞等)。經過多次瀏覽後決定斷詞參數以及同義詞以及停用字，如下表為統計本文在資料處理的簡易敘述統計表。



圖、內文部份截圖

表、資料清洗前後統計數據

多少媒體各自篇數、時至2023年1月20日、總計新聞數19146篇、??個中英文同義字以及3885個中英文停用字、詞彙數

第二節、 研究流程與架構說明

本節輔以圖示說明本文研究流程與架構，並逐段說明預期效果。

你畫一下: 資料處理(含清洗+斷詞) -> 聲量分析(有沒有統計哩?) -> 分析工具(bertopic(主題時間分析、主題分類分析、主題抽象意義)) -> 推論結果

(請參考我的論文22頁，抓感覺就好，內容差別不少)

文本已於前節說明斷詞套件的使用與其結果，承接說明聲量分析。(這邊雖然要看看建實老師的反應才有後續，我的設想是三種狀況，其一沒有統計，老師放手給你，那就是發文量時間分析。其二，老師確實如我所想?加時序分析統計?這邊就不知道老師怎麼出招了，港墘時候跟你講的k方似乎太過粗淺。其三可能混和? 變成某某(關鍵字)複雜的統計時序分析?超脫我想像?)

還好我有熬夜，靈機一動，新聞標題也可以搞事，但我目前沒想法~可以討論看看

承上段時序分析，本文預期得到各大媒體在不同時間注重議題，僅能藉由「次數」解釋注重程度，故本文進一步執行主題分析探勘媒體對維吾爾相關報導的背後框架。本文借助主題分析在有限多篇文章中得到多個主題，這些主題是隱含透露出撰文者可能帶有基本假設寫作或單憑現實書寫報導。這些隱含主題所附帶的詞彙，帶有機器運算意義或統計顯著意義，將幫助本文從「隱含主題」重定義為研究者立意的「主觀主題」。

經過重定義為研究者立意的主觀主題後，有數個進階研究工具可繼續分析。主題的時間分佈可以觀察主題的時間波動，藉此觀察各家媒體注重的主題是否不同消長，配合紀事表確認媒體在事件發生前中後已無不同。主題的類別分佈可以觀察特定主題在不同媒體的比重，確認媒體是否有側重特定主題報導，甚至是沒有敘述。重點文章(我是看到你在研究問題有強調長文章是媒體強調報導(這要假定過)，是不是可以變成一個更細緻研究工具來說故事?火災解釋立場?人權宣傳立場?)的主題佔比可以觀察媒體在重點文章上講述那些主題。

第三節、 研究方法

以下介紹研究方法，本節先行介紹文字探勘意義，並依次介紹各個研究工具的核心演算法，包含斷詞、時序分析(聲量?或統計?)、主題分析等。

一、文字探勘意義

文字探勘是結合大數據、語言學、統計、演算法、機器學習、自然語言處理(natual language process)等技術的概括多種工具的計量方法，特別是應用處理人類難以負荷的大量文字訊息，將非結構資料(意味著不符合預先資料定義格式)按照一定規則轉換為結構化資料，進一步將大量結構化資料經歷一系列處理得到結果。傳統量化分析也能處理大量結構化資料，但優勢在於後進技術機器學習可以處理更大量資料，並且多出更多分析角度，劣勢在於傳統量化分析著重在客觀數據與統計模組，強調上而下(top-down)的結構分析，容易令模組收斂數據與結果解釋囿於一隅。為此，本文透過多種分析工具不僅保有客觀數據建立演算法模組，同時借助預訓練模組(pretrained model:研究者在其他研究釋出的研究模型)的研究成果交織更多分析結果，更透過主題分析的主題重定義的研究流程，引入文獻分析的學者成果，帶出研究者主觀解釋。(Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. Journal for Language Technology and Computational Linguistics, 20(1), 19-62.、Tu, H. T., Phan, T. T., & Nguyen, K. P. (2017, July). An adaptive latent semantic analysis for text mining. In 2017 International Conference on System Science and Engineering (ICSSE) (pp. 588-593). IEEE.、(Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. Big Data and Cognitive Computing, 4(1), 1.))、(Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-Supervised Representation Learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine, 39(3), 42-62.)

除上述演算法處理資料量越發龐大之外，文字探勘與自然語言處理相較其他領域需特別處理之處，包括文字前處理、中文語句斷詞準確性、維度災難(curse of dimensionality)(Systematic review of clustering high-dimensional and large datasets. Pandove et al.,2018)、理解文義與語意等。文字前處理與中文斷詞準確性已在前文說明不再贅述，維度災難代表越高維度(腳註一下影像255、推薦預測取決欄位數、文字探勘跟自然語言處理原罪)越需要運算能力，可用數據在計算矩陣中變得很稀疏。稀疏性對於任何優化或收斂問題皆不易擁有正確或可信結果，若隨著維數提高需要更多樣本數量，問題也越難優化或收斂。

延續討論中文文義與語意問題，前文利用同義字處理多重文義，以及利用停用字處理雜訊及降低計算量，仍然無法處理句子歧義(因上下文理解而不同意義)，以及文章背後的蘊含意義，以及多篇文章背後的脈絡( Word identification for Mandarin Chinese sentences K. J. Chen & S. H. Liu,1992)。為此許多學者與先進提出許多方法解決，其中之一是常見深度學習(deep learning)模型bert的高階層數的特徵向量(Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (pp. 4171-4186).)。更深入解釋高階層數的特徵向量，其物理意義僅是處理低階網絡層數的特徵資訊無法處理(一般如bert等NLP是收斂問題處理)而逐漸往高階網絡層數的特徵調整網絡神經參數，而高低階特徵向量代表是越接近輸入層資訊，同時也越容易令人類閱讀，越高階特徵向量代表越接近輸出層，也越是令人類無法理解(CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization)(腳註https://github.com/poloclub/cnn-explainer)。所以在神經網絡未達過適(overfitting)之前，越是訓練與收斂越使模組學習到抽象意義。

機器學習到的抽象意義是稱作representation(幫腳註可能稱呼有多種，有時與embedding混用意思則是強調在母體上嵌入一個新的空間，稱feature特徵則表示資料經由特徵工程或是數值轉換後的新特徵，稱representation「表徵」強調一種新的表示，稱vector則突顯數列意義而不具有衍伸意義)來表達，representation是透過嵌入空間(embedding space或稱向量空間vector space)取樣而得，正因嵌入空間建立於研究樣本，其representation才有代表性，也才有比較價值。儘管研究者與模組開發者也無法輕易解釋的抽象意義為何而來，機器自動執行特徵工程令研究者直接探勘機器分析結果，其representation結果可以建立在字詞之間(word embedding)、段落之間(sentence embedding)、文章之間(document embedding)等，乃至研究者指定自製的嵌入空間。所幸深度學習模型發展快速且支援性高，不論各種NLP任務皆有預訓練深度模型開放使用，並且中文系統使用者眾，語言模組相對英文以外的語言容易尋找。

二、斷詞

中文斷詞在中文文本探勘是核心技術，亦是影響後續分析的基礎作業，持續有許多學者與先進們努力強化斷詞效果。早期斷詞發展基於統計機率以及詞庫技術完成，至今而有機器學習與深度學習技術相輔相成，按斷詞歷史演進有以下發展。

二之一、以中文文法結構斷詞，並搭配專有詞彙或自定義詞庫

按字詞詞性與語法結構輔助斷詞，再由詞長度進行切割分詞，並搭配詞庫含自定義詞庫(再區分建議數值自定義詞庫與引申詞與絕對斷詞詞庫)、停用字、同義字等，斷詞品質仰賴研究者與專業領域專家悉心維護。如中文知識庫「廣義知網知識本體架構（E-HowNet）」為發展語言學習資料庫，建立中文句結構樹資料庫（Sinica Treebank），結構樹（Treebank）為擷取語法規則與語法發生機率（Chen & Hsieh, 2004）幫助斷詞。如 Articut (卓騰語言科技)，不使用機器學習與資料模型，使用現代白話中文語法結構與規則，加上詞庫維護仍有相當高效斷詞結果。

二之二、斷詞如何決定取決於分詞期間的條件機率

相左於語法結構與各類詞庫，借助單一演算法與統計分佈求算詞彙在前後文的條件機率。此階段斷詞器依據斷詞器曾經餵入的大量文本並建構詞彙的統計分佈，斷詞器優劣必然受到演算法侷限、文本代表性、數量等因素影響。

二之三、預訓練與斷詞器配合

如今斷詞器已閱讀相當大量文本，並建立詞彙的多種統計分佈，也使用多種演算法，如何令斷詞器經歷大量文本以及多種演算法混用是重要關鍵。如中文最為廣泛jieba斷詞器使用詞庫維護品質，未出現的新詞則透過Hidden Marcrov學習大量文本改善。近年台灣中研院長期致力繁體中文系統的數位化與智能化，ckip(中文詞知識庫小組)提供ckipTagger與ckip Transformers兩種對中文友善的斷詞器，兩個套件皆使用深度學習模組閱讀大量文本，前者引入LSTM且後者引入Transformers。更新穎斷詞器Monpa(Yeh, W. C., Hsieh, Y. L., Chang, Y. C., & Hsu, W. L. (2022, November). Multifaceted Assessments of Traditional Chinese Word Segmentation Tool on Large Corpora. In Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022) (pp. 193-199).)是由臺北醫學大學自然語言處理實驗室與ckip一起合作，使用更大型語料庫與更有算力GPU完成自然語言處理。現今斷詞也早已進入「以量取勝」的時刻，一般公司行號或是小型研究組織也難以負荷重新訓練成本。

二之四、特定文本領域優化與GPU強化斷詞器閱讀量

計算機設備算力與大量文本不可能無限上綱，為了模組推斷精確性，斷詞器不追求所有領域文本的泛化能力，而是針對特定領域的Domain Generalization(Zhou, K., Yang, Y., Hospedales, T., & Xiang, T. (2020, April).)。誠如醫療領域不僅用詞與影像具有資料機敏性與醫療專業，PadlePadle推行並聚焦於醫療領域套件使用，在github與huggingface兩大開放平台可以發現針對推薦系統、古文閱讀、法條閱讀、商用評論等族繁不及備載多元斷詞器。

本文最後選用中研院ckip的ckip-transformers(quote <https://ckip.iis.sinica.edu.tw/> <https://github.com/ckiplab/ckip-transformers> )件完成斷詞動作。鑒於數個考量使用ckip，包括ckip訓練的文本皆為繁體中文、本文研究文本也皆為新聞報導、支援GPU加速斷詞、中文斷詞效果名列前茅(Yeh, W. C., Hsieh, Y. L., Chang, Y. C., & Hsu, W. L. (2022, November). Multifaceted Assessments of Traditional Chinese Word Segmentation Tool on Large Corpora. In Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022) (pp. 193-199).)，不須刻意微調(fine tuning)等優勢。

三、時序分析

等到確定方向再來寫~

四、主題分析

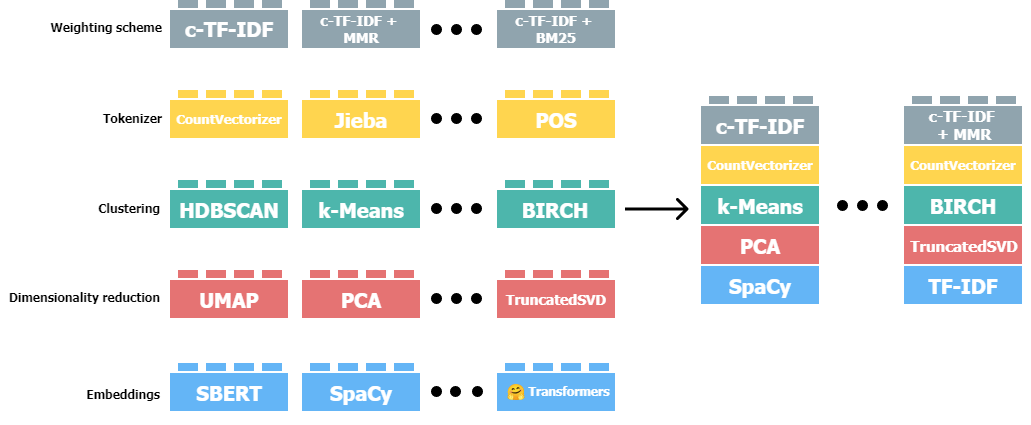
文字探勘領域有許多供決策與分析的演算法，本文盡可能採用不需大量人力介入的研究工具，並選擇主題分析為主要分析工具。粗略可將技術領域區分非監督式(unsupervised learning)、監督式(supervised learning)、自監督式(self-supervised learning)、半監督式(semi-supervised learning)、強化學習(reinforcement learning)等技術，特別適合探索性研究的技術當屬非監督、半監督以及自監督，其中非監督式無人類主觀標籤的標記資料，半監督則有部分標記資料與部分未標記資料，自監督則借助預訓練模型自行產出虛擬標記資料並自行學習。監督式學習主觀給予標籤將限定研究者未知資訊，所以不需人力介入以及預訓練模組兩項優點下，本文使用的主題分析歸屬非監督與半監督式的演算法類型，相對更有解釋空間。(Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. Big Data and Cognitive Computing, 4(1), 1.)、(Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-Supervised Representation Learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine, 39(3), 42-62.)。

本文借助機器歸納大量文章內常見字詞以及相似文章，進一步分析撰文者背後想要表達的隱含觀點，不僅主題分析演算法可以達到目標，其他機器自動歸納的常見演算法包含降維(dimension reduction)與分群(clustering)兩項。降維演算法的主要目的是付出些許資訊壓縮讓高維度資料轉換到低維度資料，此舉帶來大幅降低計算機算成本、降低矩陣稀疏性、降至二維或三維繪圖利於研究者觀察等，所以壓縮資訊剩下關鍵字詞或關鍵文章，藉此幫助理解隱含觀點，又可使後續演算法降低更多計算成本。分群演算法具有排他性，強迫分群任一篇文章分群至一個類別，在有限且大量文章內分群多少種類別，便是研究者關注的撰文者所有觀點。

傳統主題分析以LDA(Latent Dirichlet Allocation)演算法最常使用也討論最多。LDA 分出主題的概念是文章由詞彙組合，撰文者帶有主觀意識拼湊詞彙形成文章，既代表一篇文章具有「字詞→主題→文章」的三層架構(機器學「習」：以文字探勘法探索\_習近平時期之大外宣戰)。LDA計算文章內所有斷詞出現頻率，並最大概似「字詞組合」可能，並使用貝式組合機率（Bayesian probability）計算，從而概似出「主題」(Blei, Ng and Jordan 2003; Blei and Lafferty 2007)。經概似所有文章的字詞後，字詞映射某種分佈亦代表某種主題，不似分群演算法強迫將文章推向「最佳分群」，保有一篇文章中含有多種主題的機率分佈，換言之更有衍伸解釋的空間。

不過LDA演算法有一些限制與缺點待改進，已有不少研究青出於藍(Chauhan, U., & Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. ACM Computing Surveys (CSUR), 54(7), 1-35.)。傳統LDA需要輸入指定主題數量，而且僅計算詞彙頻率，不僅存在很多無法發揮作用的低頻率出現詞彙，也沒有考慮字裡行間或文章與詞彙關係。不過近年主題分析不斷出新並引入其他技術，如本文採用的BERTopic套件(BERTopic: Neural topic modeling with a class-based TF-IDF procedure)、整體演算法架構相似BERTopic並且使用者也多的Top2vec(Top2Vec: Distributed Representations of Topics)、注重上下文CTMs(Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual Contextualized Topic Models with Zero-shot Learning. EACL)、引入文章層級資訊的STM(Roberts, Stewart and Tingley. Navigating the Local Modes of Big Data: The Case of Topic Models" In Data Analytics in Social Science, Government, and Industry. New York: Cambridge University Press. 2016.)等。之所以選擇BERTopic不僅在主題分析領域使用者最多，演算法包含降維、分群、預訓練模型處理，令結果更便於解釋，同時作者公開釋出Python套件於Github，也釋出更多進階工具，如主題在時間分佈、主題在不同類別的分佈、找出離群主題、自行切換預訓練模型等。(A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts)

BERTopic之所以能權衡並混用分群、降維、預訓練是因為不同階段考量不同資訊。參考下圖BERTopic在不同階段使用不同的演算法，先是由下而上的第一階處理，Maarten Grootendorst在Emeddings層處理Document層級的representation，將每一篇文章倒入文章層級的預訓練模型取得文章的representation。接著第二Dimensionalilty reduction層減少特徵向量維度長度(代表減少文章representation的資訊量)。接著第三clustering層分群所有文章，主要目的在於區分哪些文章是離群文章，後續將進一步蒐集這些離群文章統一為一個離群主題。既然文章已被分群亦是被分為數個主題，但是研究者不明白主題代表抽象意義為何，所以透過最後兩層作業，將字詞轉向量，再陳列主題內重要性高的字詞，由研究者主觀賦以該主題意義。



圖、引自BERTopic github

BERTopic處理中文的流程還需要一些變化，包括第一層必須選擇中文預訓練模組或多國語言預訓練模組，以及第四層字詞轉向量要改為中文斷詞並加上空格的英文書寫形式。預訓練模組選擇理由相當直覺，不過字詞對電腦而言僅是一個符號紀錄在記憶體，必須轉換為機器可計算的數值，其數值可以是所有出現斷詞的序號，也可以是斷詞對應文章的相關矩陣。字詞轉向量(或稱特徵化vectorization)常見演算法如Tfidf(Thorsten Joachims. 1996. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science)、TextRank (TextRank: Bringing Order into Texts)、word2vec(Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. arXiv preprint arXiv:1712.09405.)皆是以英文語言處理為主。如前節斷詞所敘，還需同義詞與停用詞協助提升斷詞品質。特別的是第一層中文預訓練模組則不需斷詞處理，係因深度學習模組訓練時直接以原文訓練，不過還是需要去除雜訊，如廣告詞、混入html標籤、責編敘述等無關語句。

\*參數設定以及模組選用在寫final時的分析步驟的每章節前面講

維吾爾人長久居於中國大陸新疆一帶，其文化與地理位置相左華夏民族。原為綠洲游牧國家的西域，自西漢神爵二年並在張騫兩次出使西域，令西域都護府首次在歷史上處於漢人政權的附屬國地位，並隨著中原王朝更替興衰，西域長期於遊牧民族與中原王朝交替統治。西域儘管多次更替政權，在西元六世紀新疆人種和語言逐漸回鶻化，而在西元十一世紀自中東而來的伊斯蘭教傳播，開始融合穆斯林民族。當新疆開始伊斯蘭化後 ，宗教包含伊斯蘭教、佛教、道教、基督宗教等，至今已是多種宗教並存的多民族(可以找主要民族列腳註或直接這邊)地區。也因為其地理位置的重要性，在自古西域是連通中亞和西亞的古絲綢之路的重要通道，而今新疆地處歐亞大陸腹地，周邊接壤國家有俄羅斯、哈薩克、吉爾吉斯、塔吉克、巴基斯坦、蒙古、印度、阿富汗等八國。所以不僅地理位置遠離漢人農耕社會，宗教多元且以伊斯蘭教為眾，使用語言與人種也與華夏民族存在差異。

國際非政府組織人權觀察（Human Rights Watch）曾於2019年發文表示，維吾爾議題已經成為國際議題(要註腳)。從早年(哪一年開始?)的東突厥斯坦議題、反恐治理的相關辯論，到近期中西(不能概括，可講美德義大利多國)在新疆人權爭論不休並相互制裁，新疆從未淡出傳媒與社會的輿論場。(這邊生硬沒有串接)王力雄（2007）在《我的西域，你的東土》中提及，「新疆問題」在某種程度(沒有某種)上超過「西藏問題」，成為北京最頭疼的民族問題(太口語了吧XD)。

中國大陸對人權紀錄自2008年北京奧運以來爭議不斷(應該有紀錄，擴增一下ref)。近期2019年11月，《紐約時報》在頭版刊出獨家報導，表示取得超過400頁中文內部文件(出處)，揭露新疆「再教育營」大規模拘禁的政治決策和運作方式。消息一出，引發各界討論聲浪，面對國際媒體的報導，中國外交部發言人耿爽在2019年11月18日的例行記者會上，回應「新疆事務純屬中國內政，涉疆問題不是宗教、民族、人權問題，而是反暴恐和反分裂問題」，並強調新疆持續保持繁榮穩定，民族團結，社會和諧，就是最有力的回擊。中國一方面批評這樣的報導「粗暴干涉中國內政」，國際社會卻也未因中國施壓而停止關注新疆情勢。(一個段落盡量講一個敘述)

(這段也滿硬的，你想敘述媒體有責任就直言，看看台媒怎麼講)隨著媒體揭露新疆維吾爾自治區的狀況越多，維吾爾人權不僅(改不僅有連帶意思，不論是二分意思)在台灣以及(不該用或二分)國際都受到比以前更多的關注。在關注度上升的同時，在新聞自由的台灣面對維吾爾議題亦有不同的觀點，體現台灣新聞多元價值。(時間跟新聞要腳註)國民黨前主席洪秀柱曾於德國之音專訪中問及維吾爾議題看法時，反質問記者「你講的問題都是正確的嗎？是確實的嗎？你新疆那個棉的事都是真實的嗎？殺維吾爾族人是事實嗎？」，引發國內討論聲浪。

本文無法論定維吾爾人權紀錄真實與否，卻引起研究者研究媒體的動機之ㄧ之。隨著網絡發展，媒體經營需仰賴流量與討論度賴以生存，如何從更及時訊息傳遞以及吸睛新聞內容獲得收益是必須的。同時媒體不僅民眾接收資訊的重要管道，肩負傳播真實資訊、解釋教育、監督權力及維護人民權益等功能(這邊要ref一下書或腳註)。不僅媒體對事件帶有立場播報，長期報導的主題也帶有框架以及馴化(幫我quot一下)。民眾連續一段時間接觸維吾爾相關報導，連帶提及中國的族群治理、經濟社會秩序維護和人權保障等議題內容，民眾被媒體馴化並認知維吾爾人處境與情勢。故本文欲探勘媒體背後是否存有利益、宣傳、監督、解釋、關注等考量。

在新聞傳媒蓬勃發展的台灣，也高度關注維吾爾議題，多家媒體長期跟蹤報導。儘管新疆與台灣相距甚遠，風土民情也充滿歧異，然而台灣和中國大陸(中國算以前封建舊稱，可以統一中共或其他名稱)自國共內戰以來的關係複雜，自帶許多獨立狀況、維持現況、統戰、和平統一(幫我quot自然獨啥的)等深刻政治發展。在中國大陸也積極宣達反第三勢力與反分裂主義相關宣傳，並頒布反分裂國家法作為法條實行依據，諸如疆獨、藏獨、蒙獨、港獨等主張言論視為違法，同樣台獨宣言亦是中國大陸不可容許。正因此本文研究動機之二，台灣不能忽視來自新疆地區的重大事件新聞，不能忽視來自維吾爾人的相關處境報導，而後本文探勘各大媒體文本，並進一步深究內容分析。

(下面這句建議放到文獻回顧講拉，或是腳註，太硬惹)就人權報導的觀點來看，許多研究均指出人權議題的報導具有多種效應，包括教育民眾、設定政治議題、影響政府政策等(Caliendo, Gibney, & Payne, 1999; Iyengar & Kinder, 1987)。

(這邊要收尾惹，不太應該開展洪秀柱案例，可拉到前面)新聞媒體對國際議題的報導，有形與無形會影響在地閱聽人對國際議題的看法。