

數位人文與傳統文獻的相互作用：《呂氏 春秋·十二紀》聚類分析與篇章結構新探*

李貴生**

摘要

聚類分析（clustering analysis）能夠通過非監督式機器學習算法（unsupervised machine learning algorithms）揭示文本的隱藏結構，已被廣泛地應用在不同類型的文本探勘（data mining）中。鑑於現時先秦兩漢文獻的研究主要停留在資料庫的單向檢索層次，對人工智能的各種工具還未有足夠的重視，本文嘗試運用聚類分析法重新考察《呂氏春秋·十二紀》的主題和結構這個聚訟紛紜的老問題，藉此展示數位人文與上古文獻之間的相互作用（interplay）和對話的可能性。文章首先回顧近數百年來有關〈十二紀〉內容主題和結構編排的不同意見，繼而介紹聚類分析的原理，以及本文選用的數位工具和參數設定，並把四紀的聚類結果與過去的研究加以比對，說明此一方法的效度。

* 本文的構想源自筆者的教學及行政工作。過去十多年筆者一直為研究生講授文獻學和研究方法，近幾年復因職務需要，主持了過百場博士論文答辯，有機會參與不同科目的前沿討論。假如沒有香港教育大學提供的優厚條件和發展機遇，相信文章不可能如此順利地完成。這篇帶有實驗性質的文章解人難索，幸得《清華中文學報》編輯委員會諸先生延請了兩位專業的匿名審查人給予寶貴的建議，並提示後續研究的方向，謹在此致以由衷的謝忱。又本文初稿成於2020年9月19日，經審查後所作修訂稿成於同年11月29日，此後文中涉及的電腦平台、軟件工具和網上材料時有更新，然而為了避免版本衝突及不必要的混亂，本文保留了送審時的原貌。最後想藉此小文向上個世紀80年代率先籌劃各種古籍資料庫的前輩學者致敬，他們的高瞻遠矚為學界樹立了楷模，提醒後學不要以繼承家業的心態固步自封，而是應該密切留意科技的最新發展，提升古籍研究的水平。

** 香港教育大學研究生院副院長、文學及文化學系副教授。

接著文章會聚焦在異質性文本的元素，分析其獨特的連接功能，揭示〈十二紀〉的結構模式。本文的成果顯示文本探勘確能為古代文獻研究開拓嶄新視野，不但有助發現新問題，亦能夠解決舊有問題，尚有不少空間值得繼續開發。

關鍵詞：《呂氏春秋》、〈十二紀〉、相似度、聚類分析、數位人文

一、引言：發現新問題與解決舊問題

利用電腦技術輔助中國古籍的研究並不是新鮮的事物，經過三十多年的積累，現在已有多個大型古籍資源庫供學者使用。¹ 然而數位科技一日千里，在大數據（big data）的衝擊下，各種資料探勘（data mining）的技術逐漸廣為人知，² 影響遍及人文學科不同範疇。³ 反觀研究先秦兩漢文獻及文化的學者對數位人文（digital humanities）的理解，大多仍停留在資料庫的建立、語料的檢索和統計等層面上，⁴ 只有極少數學者嘗試運用文本探勘（text mining）的方法，⁵ 探討中

¹ 中央研究院歷史語言研究所 1984 年開始建立「漢籍全文資料庫計畫」，至今收錄典籍超過 1300 種，約七億八千一百多萬字，參考漢籍全文資料庫計畫，<http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm>，瀏覽日期：2020 年 7 月 12 日。香港中文大學「漢達古文獻資料庫」始建於 1988 年，收錄商、周至魏晉南北朝傳世及出土文獻，共約 8 千萬字，參考漢達古文獻資料庫，<http://www.cuhk.edu.hk/ics/rccat/database.html>，瀏覽日期 2020 年 7 月 12 日。二者是創立時間較早的資料庫。類似的資源還有很多，蔣經國際學術交流基金會「漢學資源」網頁便介紹了 20 個數位圖書館、36 個古籍文獻庫及 173 個資料庫，參考蔣經國際學術交流基金會，<http://www.cckf.org/zh/sino>，瀏覽日期 2020 年 7 月 12 日。北京愛如生數位化技術研究中心「中國基本古籍庫」則是較為大型的資料庫，收錄書籍一萬餘種，約二十億字，參考愛如生典海數字平臺，http://er07.com/home/pro_3.html，瀏覽日期：2020 年 7 月 12 日。除此之外，中央研究院的「數位人文研究平台」提供了關聯詞分析、N 字詞統計與差異比對等高階功能，免費開放給大眾註冊使用，參考中央研究院數位人文研究平台，<http://dh.ascdc.sinica.edu.tw/member/index.html>，瀏覽日期：2020 年 7 月 28 日。「中國哲學書電子化計劃」也提供一些數位人文的工具，但須訂購始能獲得高階功能，參考中國哲學書電子化計劃，<https://ctext.org/>，瀏覽日期 2020 年 7 月 12 日。

² Mehmed Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, 3rd ed. (Hoboken: John Wiley & Sons, Inc., 2020), pp. 18-22.

³ 有學者仿照大數據之名，把這類研究稱為「大人文學」。Richard J. Lane, *The Big Humanities: Digital Humanities/Digital Laboratories* (London: Routledge Ltd, 2016).

⁴ 如譚國根、梁慕靈、黃自鴻等編，《數碼時代的中國人文學科研究》（臺北：秀威資訊，2018 年）。第二至四章雖然出自不同作者之手，但當中所談到的數位應用不約而同地局限於資料庫的檢索功能上，反映了一種共通的趨向。

⁵ 「資料探勘」的素材包括文字、圖像、音訊等，「文本探勘」則專以文字資料為主。謝邦昌

國古代學術和文化問題。⁶

研究方法滯後的原因，固然與中國文獻學本身的性質有關。須知此一學科擁有源遠流長的傳統，學者要掌握累積千年的文字、聲韻、訓詁、目錄、版本、校勘等科目的方法和成果已非容易的事。⁷ 假如他們能夠嫻熟地運用這些知識，基本上已能自給自足，在所屬領域內從事同儕認可的專門研究。在這樣的氛圍下，願意使用資料庫檢索材料的學者已稱得上是新潮先進，沒有必要學習更前沿的方法。除古籍研究自身的特性外，伴隨新工具而來的學術關懷和問題意識，也是令一般學者卻步的主要原因。數位人文固然能夠帶來新視野，但研究上古文獻的學者卻不容易看到地理資訊系統（GIS）、社會網絡關係（Social networks）等方法與自己從事的課題之間的關係，⁸ 不免會以為新方法主要用來發現或解決時尚的新問題，不見得能夠填補他們關心的中國文獻學實際存在的研究缺口（research gap）。

因此這裏首先要強調，伴隨大數據發展而來的文本探勘（text

編著，《文本探勘》（臺北：元華文創公司，2017年），頁1-5。

⁶ 如森舸瀾（Edward Slingerland）和胡佳佳等曾分別利用主題建模（topic modelling）比較《論語》、《孟子》和《荀子》三書的內容，不過類似研究並不多見。Ryan Nichols, Edward Slingerland, Kristoffer Nielbo, Uffe Bergeton, Carson Logan, and Scott Kleinman, “Modeling the Contested Relationship between Analects, Mencius, and Xunzi: Preliminary Evidence from a Machine-learning Approach,” *The Journal of Asian Studies* 77/1 (2018): 19-57; 胡佳佳、孟琢，〈話題建模在中國古代典籍分析中的應用〉，《數位典藏與數位人文》第4期（2019年10月），頁27-48。

⁷ 關於這方面的問題，可參陳東輝，〈古文獻學研究若干領域的回顧與展望〉，《漢學研究通訊》第27卷第3期（2008年8月），頁5-15。

⁸ 經過學者多年的努力，群體傳記學（Prosopography）、地理資訊系統（GIS）、社會網絡關係（Social networks）一類空間人文學（Spatial Humanities）已逐漸應用到中國歷史文獻的研究中。參考中國歷代人物傳記資料庫，<https://projects.iq.harvard.edu/chinesecbdb>，瀏覽日期：2020年7月14日；王汎森，〈數位人文之可能性及限制——一個歷史學者的觀察〉，收入項潔主編，《數位人文研究與技藝》（臺北：國立臺灣大學出版中心，2014年），頁25-35；通用型古籍數位人文研究平台，<http://ming.ncl.edu.tw/>，瀏覽日期：2020年7月14日。

mining),⁹ 與以往研究者利用資料庫輔助分析的取向並不完全一致。二者最大的區別是,文本探勘技術主要針對當今互聯網上每天不斷產生的巨量數據,因而須要發展出有效的演算法(algorithm)進行半監督式(semi-supervised)以至非監督式(unsupervised)的自動化運算,¹⁰ 以便從龐雜的非結構化材料(unstructured data)中擷取資訊(information),從而建構有意義的知識(knowledge)。¹¹ 儘管先秦兩漢文獻在數量上相對地封閉和有限,不能與之後的朝代相比,但這不表示文本探勘並不適用。原因是經由人工智能各種機器學習法(machine learning)演算出來的成果,不但能夠像資料庫檢索般擷取研究者想要得到的材料,印證已知的事實,還可以推導出研究者本人也未必能夠預期或注意到的資訊,足以拓展視野,激發出新的觀點。

為了展示文本探勘與古代文獻研究之間的相互作用(interplay)與對話的可能性,本文嘗試運用聚類分析法(clustering analysis)重新考察《呂氏春秋·十二紀》的內容結構,看看這個老問題能否在數位時代中找到新的解決方案。誠如祝平次所言:「就實際研究進程中,數位的分析和人文學的探問,常常要來回往覆,不止一次」,¹² 因此

⁹ 陳光華對「文本探勘」涉及的方法有扼要的介紹:「文本探勘經常使用自然語言處理、統計分析、機率模式、機器學習等技術,探討概念擷取(concept extraction)、文件摘要(text summarization)、資訊過濾(information filtering)、命名實體的標註或辨識(named entities tagging or identification)、意見分析(opinion analysis)、關係探索(relation discovery)、情緒分析(sentiment analysis)、文本分類(text classification)、文本分群(text clustering)等議題。」見雙語詞彙、學術名詞暨辭書資訊網, <http://terms.naer.edu.tw/detail/1679014/>, 瀏覽日期:2020年7月12日。本文把clustering譯為「聚類」。

¹⁰ Wang Li-dong, Cheryl Ann Alexander, "Machine Learning in Big Data," *International Journal of Mathematical, Engineering and Management Sciences* 1/2(2016): 52-61.

¹¹ Data(數據、材料、資料)、information(資訊、訊息)和knowledge(知識)在日常生活中有時可以互用,但在資訊科學上卻是意義不同的觀念。Antonio Badia, "Data, Information, Knowledge: An Information Science Analysis," *Journal of the Association for Information Science and Technology* 65/6(2014): 1279-1287.

¹² 祝平次,〈前言〉,《清華中文學報》第22期(2019年12月),頁6。

下文也會反復比對聚類分析結果與《呂氏春秋》原文的關係，以期準確地評估機械學習法的效度和功能。

二、〈十二紀〉內容主題與結構編排的爭議

今天傳世的先秦文獻絕大部分都是經過漢人的整理，始以「書」的型態流通，不過《呂氏春秋》卻是罕有地成書於秦始皇統一天下之前，是名副其實的先秦古籍。該書乃呂不韋（292-235 B.C.）召其賓客著錄所聞，集論而成，¹³ 共分〈十二紀〉、〈六論〉、〈八覽〉三個部分。三者之中以〈十二紀〉的成書年代最為清楚，因為〈季冬紀〉最後一篇〈序意〉明言「維秦八年，歲在涪灘，秋，甲子朔，朔之日，良人請問〈十二紀〉」，¹⁴ 清楚點出始皇八年（239 B.C.）這個年份。雖然「歲在涪灘」¹⁵ 當指始皇六年（241 B.C.），因此有學者以為「八」乃「六」字之訛，而錢穆（1895-1990）則以始皇元年（247 B.C.）為甲寅，推算出始皇七年（240 B.C.）的結論，¹⁶ 但無論如何，一部成於二千二百多年的古籍只有不足兩年的誤差，實在已是極為難得的事。

〈十二紀〉成書年代清楚，亦較受重視。《四庫總目提要》云：

其〈十二紀〉即〈禮記〉之〈月令〉，顧以十二月割為十二篇，每篇之後各間他文四篇。惟夏令多言樂，秋令多言兵，似乎有義，其餘則絕不可曉，先儒無說，莫之詳矣。又每紀皆附

¹³ 〔漢〕司馬遷著，〔南朝宋〕裴駟集解，〔唐〕司馬貞索隱，〔唐〕張守節正義，〈呂不韋列傳〉，《史記》（北京：中華書局，1963年），冊8，卷85，頁2510。

¹⁴ 〔秦〕呂不韋著，陳奇猷校釋，〈序意〉，《呂氏春秋新校釋》（上海：上海古籍出版社，2002年），上冊，頁654。

¹⁵ 〔秦〕呂不韋著，陳奇猷校釋，〈序意〉，《呂氏春秋新校釋》，上冊，頁654。

¹⁶ 參考錢穆，〈呂不韋著書考〉，《先秦諸子繫年》，收入鄭秀蓮編，《錢賓四先生全集》（臺北：聯經出版事業公司，1994年），冊5，卷4，頁560-563。

四篇，而季冬紀獨五篇。末一篇標識年月，題曰〈序意〉，為〈十二紀〉之總論。殆所謂「紀」者猶內篇，而「覽」與「論」者為外篇、雜篇歟？¹⁷

文中以〈十二紀〉為內篇，〈八覽〉、〈六論〉為外、雜篇，確認了〈十二紀〉的優先地位，並且提到〈夏紀〉多言樂，〈秋紀〉多言兵，較早地注意到諸紀的編排「似乎有義」。四庫館臣認為除夏、秋二令似有特定主題外，「其餘則絕不可曉」，但徐時棟（1814-1873）卻別具慧眼，看出貫通四紀的邏輯結構：

呂覽以〈十二紀〉冠本書，故亦僭稱「春秋」。每一月下間以雜論四篇，若無倫緒。然〈孟春紀〉下首以〈本生〉篇，以春之義生，故說養生之道。〈孟夏紀〉下首以〈勸學〉，以夏之義長成，人長則當學，故論為學之要。〈孟秋紀〉下首以用兵，以秋之義肅殺，故說兵戰之事。〈孟冬紀〉下首以節喪，以冬之義閉藏，故講喪葬之法。餘篇則或相連貫，或不相銜接，而其首篇故有意也。¹⁸

他清楚指出四紀中的第一篇文章俱是出於刻意的安排，分別體現春生、夏長、秋殺、冬葬四個主題，與該紀的季節屬性互有關聯。不過他也提到這種安排僅限於「首篇」，其餘文章則「或相連貫，或不相銜接」，沒有強求一致。

徐時棟的說法揭示了一個新方向，啟發了後世眾多學者。余嘉錫（1884-1955）便發揮其旨，並且有更為細緻的補充：

〈提要〉謂夏令言樂，秋令言兵，是也，謂其餘絕不可曉者，

¹⁷ [清]永瑤等，《四庫全書總目》（北京：中華書局，1987年），上冊，卷117，頁1008-1009。

¹⁸ [清]徐時棟，《煙嶼樓讀書志》，收入《續修四庫全書》編纂委員會編，《續修四庫全書》（上海：上海古籍出版社，2002年），冊1162，卷15，頁582。

非也。今以春、冬紀之文考之，蓋春令言生，冬令言死耳。……然則春生而冬死，夏樂而秋刑，其取義何也？曰，此所謂春生夏長秋收冬藏也。¹⁹

文中雖然沒提到徐時棟，取徑卻是如出一轍。二者的分別是，余氏認為四紀中其餘篇章同樣與季節屬性有關，突破了徐氏所述的首篇範圍。如〈春紀〉中除〈孟春紀〉首篇〈本生〉談「生」之外，〈仲春紀〉首篇〈貴生〉和〈季春紀〉首篇〈盡數〉同樣「發凡起例，極言節欲養生之義」；該紀其他篇章如〈貴公〉雖然看似與養生無直接的關係，卻是「示人以修身立命之道，以蘄各遂其生也」，²⁰因而也算與生有關。同樣，〈孟冬紀〉的〈節葬〉和〈安死〉涉及「薄葬送死之義」，同紀中的〈至忠〉、〈忠廉〉雖然看似與喪葬的關係不大，卻是「示人以舍生取義之道，以期善處其死也」，²¹所以他認為這幾篇亦符合冬令言死之主題。

余嘉錫改良並深化了徐時棟等人的說法，令四紀的主題延伸至各紀的其他篇章。他的理論無疑較前人完整，但也有啟人疑竇之處，暴露出此類主題論述本來隱而未發的問題。館臣的立論本來相當謹慎，僅謂「夏令多言樂，秋令多言兵」。²²檢〈仲夏紀〉和〈季夏紀〉八篇文章確與音樂有關，²³〈秋紀〉十二篇文章亦皆涉及兵戰之事。²⁴後世學者基本上接受館臣對夏、秋二紀的主題闡述，如陳奇猷（1917-2006）便斷言〈大樂〉「此下八篇乃陰陽家治樂者即所謂樂家者流之

¹⁹ 余嘉錫，《四庫提要辨證》（北京：中華書局，1980年），冊2，卷14，頁818-820。

²⁰ 余嘉錫，《四庫提要辨證》，冊2，卷14，頁819。

²¹ 余嘉錫，《四庫提要辨證》，冊2，卷14，頁819。

²² 余嘉錫，《四庫提要辨證》，冊2，卷14，頁818。

²³ 八篇言樂的文章分別是〈大樂〉、〈侈樂〉、〈適音〉、〈古樂〉、〈音律〉、〈音初〉、〈制樂〉和〈明理〉。

²⁴ 十二篇言兵的文章分別是〈蕩兵〉、〈振亂〉、〈禁塞〉、〈懷寵〉、〈論威〉、〈簡選〉、〈決勝〉、〈愛士〉、〈順民〉、〈知士〉、〈審己〉和〈精通〉。

言也」，²⁵〈蕩兵〉「下十二篇要旨在言兵之仁義禮讓」，「蓋出於《司馬法》之學歟」云云。²⁶ 可見館臣所述大體上能夠概括夏令和秋令大部分篇章的主題，加上原句中還有表示約數的「多」字，語意周匝，並無明顯漏洞。稍後徐時棟提出春生、夏長、秋殺、冬葬之說，雖然大致上可自圓其說，卻已不及館臣之說穩妥。如徐氏在夏令中標舉「為學之要」，把「當學」視為「長成」，便不免有點牽強，須要有更多的說明，而且〈夏紀〉中論學之文僅限於〈孟夏紀〉四篇，²⁷ 若以代表性而論，反而不及館臣所強調的八篇論樂之文。當然，這些問題並非沒法解決，若按《周易》「屯」、「蒙」二卦的次序，長與學也可以有一定的關連。〈序卦傳〉云：「屯者，物之始生也。物生必蒙，故受之以蒙」，²⁸ 孔疏謂「蒙之為義，利以養正」，《蒙》緊接在代表始生的《屯》卦之後，²⁹ 可見長養亦可以與啟蒙有關，只是須要更多說明；此外，徐氏僅以各紀首篇為立論基礎，承認其餘諸篇或不相連貫，因此他不涉及〈孟夏〉之外論樂諸文，亦算說得通。

從臣館與徐時棟的論述可見，四紀主題之說越是詳細、完整，越容易出現漏洞。余嘉錫的說法無疑較前人更為細密，但由此衍生出來的問題亦較前人為多。他引用《春秋繁露》「喜氣為暖而當春，怒氣為清而當秋，樂氣為太陽而當夏，哀氣為太陰而當冬。……是故春喜，夏樂，秋憂，冬悲」，以及「春，愛志也；夏，樂志也；秋，嚴志也；冬，哀志也」等話，³⁰ 道出四紀與四氣的屬性和關係，誠然發千古之覆，功不可沒。不過他嘗試把各紀其他篇章套進這個框架，卻開

²⁵ 〔秦〕呂不韋著，陳奇猷校釋，〈大樂〉，《呂氏春秋新校釋》，上冊，頁 260。

²⁶ 〔秦〕呂不韋著，陳奇猷校釋，〈蕩兵〉，《呂氏春秋新校釋》，上冊，頁 389。

²⁷ 四篇文章分別是〈勸學〉、〈尊師〉、〈誣徒〉和〈用眾〉。

²⁸ 〔魏〕王弼注，〔唐〕孔穎達疏，〈周易兼義〉，《周易正義》（北京：北京大學出版社，2000年），卷 9，頁 394。

²⁹ 〔魏〕王弼注，〔唐〕孔穎達疏，〈周易兼義上經乾傳〉，《周易正義》，卷 1，頁 45。

³⁰ 余嘉錫，《四庫提要辨證》，冊 2，卷 14，頁 820-821。

了牽強比附的先河。如〈孟春紀〉首兩篇〈本生〉、〈重己〉固然與生有關，但第三、四篇〈貴公〉和〈去私〉主要闡述「聖王之治天下也，必先公」和「天無私覆也，地無私載也」之旨，³¹ 與生無直接關係，為何仍放在〈春紀〉之中？余氏的答案是〈貴公〉諸篇「示人以修身立命之道，以蘄各遂其生也」，³² 因而也符合生的主題。此說看似有理，卻是偷換了觀念，悄悄地擴充了「生」的涵義，因為「養生」與「修身立命」根本上是不同的命題。過去學者都清楚地意識到二者的區別，甚至把它們歸屬到不同的學術流派。如劉咸炘（1896-1932）便認為〈本生〉、〈重己〉「皆多道家說」，³³ 陳奇猷則以為是「陰陽家言養生之要」；³⁴ 至於〈貴公〉和〈去私〉，劉咸炘以為前者源自「儒道二家義」，³⁵ 後者「兼有儒墨二家說」，³⁶ 陳奇猷則謂〈貴公〉乃「伊尹學派之言」，³⁷ 〈去私〉「出於墨家者流之手」。³⁸ 劉、陳二人雖對諸篇的學派歸屬有不同的理解，卻不約而同地把〈本生〉、〈重己〉視為一類，與〈貴公〉和〈去私〉區別開來。他們的論點不同，卻又殊途同歸，正可說明〈孟春紀〉四篇文章其實存在不同的主張。假如像余嘉錫那樣把「修身立命」也納入「生」的範疇，那麼夏令和秋令之文亦可放進春令之內了，因為學習和用兵等也是為了使人「各遂其生」，遵照他的思路，亦算符合「生」此一主題了。

余氏對於冬令的論述也有相近的問題，〈冬紀〉除首兩篇〈節葬〉

³¹ 〔秦〕呂不韋著，陳奇猷校釋，〈貴公〉，《呂氏春秋新校釋》，上冊，頁 45；〈去私〉，《呂氏春秋新校釋》，上冊，頁 56。

³² 余嘉錫，《四庫提要辨證》，冊 2，卷 14，頁 819。

³³ 王利器，〈本生〉，《呂氏春秋注疏》（成都：巴蜀書社，2002 年），冊 1，頁 52。

³⁴ 〔秦〕呂不韋著，陳奇猷校釋，〈本生〉，《呂氏春秋新校釋》，上冊，頁 22。

³⁵ 王利器，〈貴公〉，《呂氏春秋注疏》，冊 1，頁 101。

³⁶ 王利器，〈去私〉，《呂氏春秋注疏》，冊 1，頁 125。

³⁷ 〔秦〕呂不韋著，陳奇猷校釋，〈貴公〉，《呂氏春秋新校釋》，上冊，頁 46。

³⁸ 〔秦〕呂不韋著，陳奇猷校釋，〈去私〉，《呂氏春秋新校釋》，上冊，頁 57。

和〈安死〉與冬藏有關外，其餘多與喪葬無關。王利器（1912-1998）節引《日鈔》云：

〈仲冬紀〉次曰〈至忠〉，次曰〈忠廉〉，次曰〈當務〉，次曰〈長見〉，凡以附〈仲冬〉之後，不曉所謂。³⁹

〈仲冬紀〉四篇固然「不曉所謂」，〈季冬紀〉的〈士節〉、〈介立〉、〈誠廉〉和〈不侵〉亦多談耿介之節，與死喪關係同樣不明確，但余嘉錫卻謂「〈至忠〉、〈忠廉〉以下諸篇，則示人以舍生取義之道，以期善處其死也」，⁴⁰ 所以仍合冬藏之義。他的辯解似能自圓其說，可是若以舍生取義為「善處其死」，因而符合〈冬紀〉主題，那麼〈仲春紀〉中的〈貴生〉亦可以納入〈冬紀〉了，因為該篇提到「迫生」與「死」的關係：

所謂迫生者，六欲莫得其宜也，皆獲其所甚惡者，服是也，辱是也。辱莫大於不義，故不義，迫生也，而迫生非獨不義也，故曰迫生不若死。⁴¹

高誘（?-212）於此句下注云：「急於苟生不仁義，不如蹈仁義死為貴」，⁴² 正是舍生取義之意，但現在〈貴生〉卻隸屬言生的〈春紀〉。由此可見，硬要把紀中所有文章套入該紀的單一主題，很容易做成牽強附會，甚至會弄巧成拙，自相衝突。

徐復觀（1904-1982）很早便注意到各紀文章的主題有不一致的問題，雖然同意四紀確實體現了生、長、收、藏的結構，但「形式太整

³⁹ 王利器，〈至忠〉，《呂氏春秋注疏》，冊2，頁1065。

⁴⁰ 余嘉錫，《四庫提要辨證》冊2，卷14，頁819。

⁴¹ 許維通，〈仲春紀〉，《呂氏春秋集釋》（北京：中華書局，2009年），上冊，頁42。

⁴² 許維通，〈仲春紀〉，《呂氏春秋集釋》，上冊，頁42。

齊了，也太機械了」，不能避免「過於牽強和過於重複的困難」。⁴³ 他歸納〈孟春紀〉四篇文章的內容後，直言：

把一套政治理論安排在「春生」的「生」的觀念之下，以為這樣便會與春之氣相應，自然會流於牽強。並且春有孟春仲春季春，孟、仲、季，各須安排四篇性質相同的東西，更不能不重複。……夏秋冬各紀的政治思想的安排，其用心與春季相同，而其牽強更甚。⁴⁴

徐氏不但指出問題所在，還舉出不少例子佐證其說，如謂〈仲春〉的「〈貴生〉、〈情欲〉，是孟春〈重己〉的重述與發揮」。⁴⁵ 他一方面肯定四紀各有鮮明的主題，從屬該紀的一些文章亦與此主題有一定的對應關係，互有發揮，但他也承認不是所有篇章都能配合該主題，只是被機械化地套進此一結構中。這種說法基本上能夠恰當地描述〈十二紀〉的實際情況，可惜後世不少學者仍好用余嘉錫的解讀模式，企圖把〈十二紀〉所有文章逐一歸屬於生、長、收、藏的序列中。

管宗昌嘗批判地回顧過去有關〈十二紀〉主題的研究，認為有些「『聯繫』並沒有以文本和文獻的事實為依據，屬於主觀的牽扯關聯，顯得較為牽強」，⁴⁶ 有些「解釋方法可謂『完滿』，但是很明顯也陷入了牽強附會之嫌」，⁴⁷ 對這類研究的毛病有清楚的認知。有趣的是，管氏雖然明知這種方法有不少問題，卻仍繼續同類研究，相信只要以「紀」為單位，詳盡具體地解釋每一紀的結構，即可克服比附的問題。然而反複細讀管氏的說法，實在很難看出他與前人的論述有甚麼區

⁴³ 徐復觀，《兩漢思想史》（臺北：學生書局，1976年），卷2，頁33-34。

⁴⁴ 徐復觀，《兩漢思想史》，卷2，頁36-37。

⁴⁵ 徐復觀，《兩漢思想史》，卷2，頁36。

⁴⁶ 管宗昌，《《呂氏春秋》文學研究》（北京：中國社會科學出版社，2016年），頁17。

⁴⁷ 管宗昌，《《呂氏春秋》文學研究》，頁18。

別。仍以〈孟春紀〉為例，大家基本上都同意首兩篇〈本生〉和〈重己〉能夠呼應〈春紀〉的養生主題，但接下來的〈貴公〉和〈去私〉與養生又有甚麼關係？管宗昌自信只要深入到書中提及的故事內容，便可說明這種關係。他注意到〈貴公〉記述了齊桓公因行私而危害自己生命一事，遂謂：

〈貴公〉用大量文字講述齊桓公最終的人生悲劇，與〈孟春紀〉的養生理念又有一定的關聯，二者並非完全疏離、毫不相關。⁴⁸

但正如前面所說，既然故事與「死」有關，把它放在〈冬紀〉之下似乎亦無不可。他論述〈去私〉中墨家鉅子大義滅親之事，尤其能說明這種方法的局限：

這個故事的主題是秉公執法，所昭示的卻是貴生與處死之間的辯證關係。雖然這種啟示溢出本篇的主題之外，卻與〈孟春紀〉的養生理念相關。⁴⁹

姑勿論故事涉及的「死」更配合冬紀的主題，單就文本而言，〈去私〉除了提到墨家鉅子一事外，尚有更大的篇幅講述祈黃羊唯賢是舉的故事。管氏僅選取有利己見的材料，已有斷章取義之嫌。至於所謂貴生與處死的「辯證關係」，更是他本人為了把死與養生連結起來而另行建構出來的關係，他本人也意識到這種關係「溢出本篇的主題之外」，純粹是他強加於文本的詮釋。按照他所說的辯證邏輯，〈冬紀〉論死之文皆可放到〈春紀〉之內，這樣的話，四紀又如何談到上各自有的單一主題？可見他的論述與他所嚴詞批評的附會研究，不見得有太大的分別。

⁴⁸ 管宗昌，《《呂氏春秋》文學研究》，頁 27。

⁴⁹ 管宗昌，《《呂氏春秋》文學研究》，頁 28。

三、聚類分析的原理與操作方法

以上回顧顯示〈十二紀〉的結構主題及其與所屬篇章的關係至今仍是困擾學界的重要問題，同時亦暴露了古代文獻研究過度依賴質性分析（qualitative analysis）的弊端。研究古籍固然須要有深厚的文本細讀（close reading）功夫，始能從中提煉出個人見解，不過純粹依賴學者別具慧眼的研究方法，很容易導致兩個問題：（1）在解釋時以語言雄辯代替邏輯推論，⁵⁰ 令結果流於主觀；（2）在論述過程中隨意取樣，只標舉有利一己觀點的材料，令結果流於片面。以前面提到的〈貴公〉和〈去私〉為例，余嘉錫把養生的涵義擴充至道德修養，勉強地連結二者，便犯了第（1）類毛病；管宗昌僅選取兩篇中有關「死」的故事，並謂生、死有辯證關係，所以仍與養生相應，便先後犯上（2）和（1）的問題了。這類推斷均以質性的觀察為依據，讀者並無客觀可靠的方法判別他們孰為合理。

〈去私〉和〈貴公〉分別只有五、六百字，已可容納如此不同的視角，讓學者能按己意去取材料，要是把分析的範圍擴展至〈十二紀〉合共 60 篇文章，相信分歧只會更為嚴重。筆者當然可以沿用過去的方法，馳騁雄辯，儘量編織一套自圓其說的理論，但前述主觀及片面的問題依舊存在，其他學者可以同意筆者的說法，亦可以覺得當中有比附的成分，因而再提出另一番論述。這種現象表面上看似百家爭鳴、眾聲喧嘩，實際上卻是各說各話、沒完沒了。要突破舊有研究的

⁵⁰ 借用維柯（Giambattista Vico, 1668-1744）的說法，這類雄辯接近西方的論題法（ars topica），旨在疏通文本，彌合矛盾，與笛卡爾（René Descartes, 1596-1650）以降的批判法（ars critica）並不相同，而後者正是現代學術的主流方法。參考 Donald Phillip Verene, “Vico’s Method of Studies in Our Time,” *New Vico Studies* 20(2002): 13-18，以及李貴生，《疏證與析證：清末民初中國文學研究的範式轉移》（北京：中國社會科學出版社，2016年），頁 1-13。

困局，借助以統計學、演算法為基礎的聚類分析，也許能夠提供新的研究思路。鑑於數位人文學尚未普及地應用到先秦兩漢文獻研究中，學者對相關的操作方法不免感到陌生。參考近年類似著述的做法，⁵¹ 以下擬先略述聚類分析的原理、施行步驟，以及參數設定的理據，並且交代本文採用的運算工具和操作過程。

聚類分析屬人工智能的機器學習法，可分為監督式、半監督式和非監督式三類。⁵² 針對〈十二紀〉面對的主觀分析和隨意取樣兩大問題，本文特別選用了非監督式的聚合式階層聚類法（**agglomerative hierarchical clustering**），儘量減低人為的介入和干預，藉此考察電腦演算出來的結果能夠為質性分析帶來甚麼啟發。⁵³ 當然，電腦不能直接處理人類所用的自然語言（**natural language**），要令〈十二紀〉60 篇文章最後以階層式的樹狀圖（**dendrogram**）呈現出來，必須先對原來文本作出一些前處理（**preprocessing**），決定分詞（**word segmentation**）的方法及設定合適的停用詞（**stop words**）。

要獲得較為可靠的分詞，可以直接襲用現成語料庫的成果，如中央研究院「上古漢語標記語料庫」便提供了經專家斷詞及標注的《呂氏春秋》原文，可供瀏覽及檢索。⁵⁴ 不過正如前文提到，本文旨在展示資料庫檢索與文本探勘的分別，希望儘量完整地介紹後者的操作流程，評估其優劣。加上「上古漢語標記語料庫」未及收錄先秦兩漢所

⁵¹ 與慣用量化分析的社會科學論文不同，從事古籍研究的人文學者在展示分析結果之前，多會對研究中涉及的方法、原理及工具有所介紹。參考劉苑如、羅珮瑄、邱琬淳、陳雅琳，〈魏晉南北朝筆記小說疾病文本的細讀與遠讀〉，《清華中文學報》第 22 期（2019 年 12 月），頁 49-115。

⁵² Wang Li-dong, Cheryl Ann Alexander, "Machine Learning in Big Data," 52-61.

⁵³ 關於聚類分析的基本介紹以及本文所用的階層聚類法，可參考 Xu Rui and Donald C. Wunsch II, *Clustering* (New Jersey: IEEE Press, 2009), pp. 1-13, 31-62.

⁵⁴ 參考中央研究院上古漢語標記語料庫，<http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh?ukey=-644530170&qtype=0>，瀏覽日期：2020 年 11 月 27 日。

有文獻，而考古發現的新材料又不斷湧現，研究者若要比對語料庫尚未處理的文本，仍須借助自動化工具，始能在較為一致的參數下與其他學者的成果互相比對，因此以下還是由一般的分詞方法談起。

現時文本探勘的工具主要來自英語世界，如 Python 的 NLTK 和 R 語言的 TM 一類常用套件在前處理階段中多聚焦在詞幹提取 (stemming) 和詞形還原 (lemmatization)。⁵⁵ 這與中文文本所面對的分詞問題並不相同，因為漢語不像英語那樣每一個詞之間都有空格作標記，而且往往一詞多義，搭配靈活，很難準確分詞。⁵⁶ 現時常見的中文分詞工具大多針對現代漢語，處理文言文的效果並不理想，⁵⁷ 因此一些學者便索性根據古漢語中單音詞占絕大多數的現象，以一字為一詞，只識別人名、地名等複音節詞。⁵⁸ 不過《呂氏春秋》屬戰國末期文本，複音節詞彙有 2,017 個之多，幾近 40%，⁵⁹ 若以一字為一詞，或會出現較大誤差。

經過多番思量，筆者特地輸入了張雙棣、殷國光等人所編《呂氏春秋詞典》(下稱《詞典》)的詞條，⁶⁰ 以此取代 Python 的 jieba 套件

⁵⁵ Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (Sebastopol: O'Reilly Media, 2009); Kasper Welbers, Wouter Van Atteveldt and Kenneth Benoit, "Text Analysis in R," *Communication Methods and Measures* 11/4 (2017): 245-265.

⁵⁶ HanLP 套件的作者何晗對此有非常詳盡的解說，參考何晗，《自然語言處理入門》，<https://github.com/NLP-LOVE/Introduction-NLP>，瀏覽日期：2020年8月18日。

⁵⁷ 坊間常見的分詞工具如 jieba、HanLP、SnowNLP、Jiagu 等皆以現代漢語的詞庫為基礎，「甲言」Jiayan 雖然聲稱專為文言文而開發，但筆者試用後發現分詞效果仍未如理想。這些套件均是開源程式，可在網站免費下載 GitHub，<https://github.com>，瀏覽日期：2020年8月18日。

⁵⁸ 如胡佳佳等便謂「所謂分詞，就是將中文文本中連續的漢字序列用空格切分成一個一個單獨的詞語」，見胡佳佳、孟琢著，〈話題建模在中國古代典籍分析中的應用〉，頁 30-31。

⁵⁹ 據張雙棣統計，《呂氏春秋》共有 4,989 詞，見張雙棣，《《呂氏春秋》詞匯研究（修訂本）》（北京：商務印書館，2008年），頁 1。

⁶⁰ 張雙棣、殷國光、陳濤，《呂氏春秋詞典》（濟南：山東教育出版社，1993）。

中原來的詞典檔，並以精確模式作詞典分詞。⁶¹ 這種做法既可以充分利用專家的研究成果，同時又能保留足夠的延伸性，容許其他使用者以相同方式處理其他文本，不受現成資料庫規模的限制。選擇《詞典》，主要是因為該書除詞語釋義外，還會標示詞語的語法功能，以及一些複音節詞的結構形式，令讀者能夠明白編者分詞的理據。而且更為重要的是，張雙棣和殷國光俱是《呂氏春秋》詞彙專家，分別有《《呂氏春秋》詞匯研究》和《《呂氏春秋》詞類研究》行世，⁶² 對詞語的定義和詞類劃分的依據均有充分的說明。如張氏該書首章即介紹「《呂氏春秋》的字和詞」，說明字、詞與義位的涵義；⁶³ 殷著探討《呂氏春秋》詞類前亦先討論字與詞、詞與非詞等分別，⁶⁴ 並且詳盡地說明詞類劃分的標準和依據、詞的歸類和活用，以及《呂氏春秋》詞量和詞類的特色。⁶⁵ 當然，無論定義如何清楚，在實際分析時，詞與短語之間的界線有時並不十分明晰，然而讀者要是覺得《詞典》的釋義和說明仍未足夠，還可翻閱他們的專書，瞭解編者的依據。因此整體而言，《詞典》較同類研究更具參考價值，像香港中文大學「漢達文庫」雖然也列出呂書詞彙，合共約 5,765 條，較《詞典》多出數百條，但「漢達文庫」選詞準則並不清楚。姑以〈孟春紀〉首幾句為例，⁶⁶ 只

⁶¹ 本文所用 jieba 乃臺灣繁體特化版本，除支援繁體中文外，用法及功能與原有版本並無分別。詳參結巴中文斷詞臺灣繁體版本，<https://github.com/APCLab/jieba-tw>，瀏覽日期：2020 年 8 月 19 日。以及結巴中文分詞，<https://github.com/foxsjy/jieba>，瀏覽日期：2020 年 8 月 19 日。替換後的詞典檔的詞頻為該詞在《呂氏春秋》出現的次數，分詞時 HMM 設定為 false。

⁶² 張雙棣《《呂氏春秋》詞匯研究（修訂本）》初刊於 1989 年；殷國光，《《呂氏春秋》詞類研究》則初刊於 1997 年。

⁶³ 張雙棣，《《呂氏春秋》詞匯研究（修訂本）》，頁 15-27。

⁶⁴ 如他具體指出「確定一個字記錄的是詞與否，關鍵是區分複音詞和短語」，而複音詞則是一個穩定的複音組合中「包含兩個有意義的成分」，「這兩個成分之間的意義聯繫是不可分解」。參考殷國光，《《呂氏春秋》詞類研究》（北京：商務印書館，2008 年），頁 2-10。

⁶⁵ 殷國光，《《呂氏春秋》詞類研究》，頁 10-32。

⁶⁶ 本文用作數位分析的《呂氏春秋》電子文本皆來自「中國哲學書電子化計劃」，下同不贅。

要比較依據《詞典》、「上古漢語標記語料庫」（下稱標記）和「漢達文庫」（下稱漢達）的分詞結果，便可看到箇中差異。考慮到空格較難辨認，下表會用下底線作分詞標號。

表一：《呂氏春秋詞典》、「上古漢語標記語料庫」與「漢達文庫」分詞結果

原文	孟春之月：日在營室，昏參中，旦尾中。其日甲乙。其帝太皞。其神句芒。
詞典	孟春_之_月_日_在_營室_昏_參_中_旦_尾_中_其_日_甲_乙_其_帝_太皞_其_神_句芒
標記	孟春_之_月_日_在_營室_昏_參_中_旦_尾_中_其_日_甲_乙_其_帝_太皞_其_神_句芒
漢達	孟春之月_日在營室_昏參中_旦尾中_其_日_甲_乙_其_帝_太皞_其_神_句芒

從表一可見，《詞典》與「標記」的分詞結果高度相似，只是對「甲乙」二字的處理稍有不同。⁶⁷ 反觀「漢達」以「孟春之月」、「日在營室」、「昏參中」等為詞，標準相當寬鬆。尤其令人難以理解的是，「日

參考中國哲學書電子化計劃，<https://ctext.org/>，瀏覽日期：2020年8月19日。

⁶⁷ 就《呂氏春秋》文意而言，《詞典》以「甲」、「乙」為二詞的理由相當清楚，因為「其日甲乙」乃以天干配日，分別指甲日和乙日，而甲、乙於五行中屬「木」，與春季五行屬性相同，所以春季以甲日和乙日為吉。是則甲、乙中間可以加入頓號或連詞，並非結構緊密的一個單位。張雙棣、殷國光、陳濤，《呂氏春秋詞典》，頁513、575；《禮記·月令》「其日甲乙」句下疏云：「孚甲在前，抽軋在後，則應孟春為甲，季春為乙」（參考〔漢〕鄭玄注、〔唐〕孔穎達疏，〈月令第六〉，《禮記正義》（北京：北京大學出版社，2000年），冊2，卷14，頁520。）也是把甲、乙分開處理。「標記語料庫」把此句標示為「其(NH)日(NA5)甲乙(NA5)[+co]」，以「甲乙」為並列的時間詞，礙於體例，語料庫沒有提供釋義，因此也不知道此一標記背後的原因，亦不知道是否因為受到該庫其他語料的影響，須要作出規範化的處理。

在某」是〈十二紀〉紀首中反復出現的套語，〈孟春紀〉的「日在營室」既然算得上是一個詞，其餘諸紀的「日在奎」和「日在胃」等理應作相同處理。可是該詞彙庫除「日在營室」和「日在斗」外，其餘皆未見收錄，⁶⁸ 選詞準則如此不一致，自然難以為據。

機器分詞有時未必能因應上下文義作出恰當的斷詞，但筆者以人工方式抽樣檢查了 12 篇文章（每一紀抽取一篇），發現分詞結果基本上可靠。當然，這類不顧詞類、缺乏標記的分詞方式並未能解決一詞多義的問題，聚類分析能否在此一限制下作出合適的分類，也是值得留意的地方，下文對此會再有討論。交代分詞方法後，接下來要處理的是「停用詞」（stop words），思考如何去除缺乏資訊（non-informative）的詞語（如虛詞、代名詞等）。在一些依賴詞語分析的文本探勘算法中，如關鍵詞提取（keyword extraction）或主題建模（topic modelling）等，設定停用詞是相當重要的工序，不同的停用詞表對聚類分析的結果也有一定的影響。⁶⁹ 然而《呂氏春秋》成於眾人之手，文章內容與寫作風格有頗為複雜的關係，像十二篇紀首便明顯有一套獨特的書寫模式，加上書中引用若干聖王的故事或學者的言論，亦可反映該篇與不同學派的關係，因此貿然刪去某些詞彙，並不是明智的做法。考慮到（1）虛詞也可以顯示作者的身分，有識別風格的功能，⁷⁰ 以及（2）古漢語往往一字多義，難以截然區分某字是否必然與內容無關，⁷¹ 本文採用折衷的方法，保留標點符號以外的所有字詞，但會

⁶⁸ 見香港中文大學「漢達文庫」《呂氏春秋》專書詞彙檢索列表，參考漢達古文獻資料庫，<http://www.chant.org.ezproxy.eduhk.hk/Lexicon/SpSearch.aspx>，瀏覽日期：2020年8月19日。

⁶⁹ 參考官琴、鄧三鴻、王昊，〈中文文本聚類常用停用詞表對比研究〉，《數據分析與知識發現》2017年第3期（2017年3月），頁72-80。

⁷⁰ 參考邱詩雯，〈《史詞》作者數位化研究初探——以三十世家虛字字類為例〉，《數位典藏與數位人文》第2期（2018年10月），頁49-69。

⁷¹ 如Sergey Zinin一文附表所列古漢語停用詞便包括「情」、「生」、「殺」、「身」等，要是去除這些詞語，肯定會對《呂氏春秋》的養生主題分析產生不少影響。Sergey Zinin, "Analysis of

在下一步驟中通過 TF-IDF 算法降低常用詞的權重 (weighting)，在保留虛詞一類詞彙之餘，減少過度的干擾。

完成前處理後，接下來便得把自然語言「向量化」(vectorize)，轉換為電腦能夠運算的文件詞語矩陣 (document-term matrix)。鑑於先秦文獻屬專門領域，材料不算太多，本文選擇以詞袋 (bag of words) 方式進行向量化，⁷² 把〈十二紀〉中出現的所有字詞放進一個詞袋中，逐一臚列每一個詞在各篇中出現的頻度。⁷³ 當然，頻次的多寡雖然在一定程度上可以反映內容的比重，但未必能確切地捕捉文本的主要特徵，如〈情欲〉有「故耳之欲五聲，目之欲五色，口之欲五味」一語，⁷⁴ 三句在內容上最重要的差異源自「耳」與「聲」、「目」與「色」和「口」與「味」，但這六個詞各只出現一次，反而「之」和「欲」分別出現了三次之多。要解決這個問題，最常見和有效的做法是在文本向量化時引入 TF-IDF 算法。⁷⁵ TF (term frequency) 指詞語在文本中出現的頻

Character-frequency Lists of Chinese Classics and its Application to Content Analysis and Genre Attribution,” in Kobzev A, ed., *Scholarly Reports of the Department of China of the Institute of Oriental Studies*, vol. XLVII (Moscow: Russian Academy of Sciences, 2017), p. 291.

⁷² 詞嵌入 (word embedding) 也是常見且有效的詞向量方法，不過這類工具 (如 word2vec) 須要有數量龐大的訓練語料 (training data) 始能導出可靠的結果，加上現時訓練中文的語料庫仍然以現代漢語為主，所以這裏還是選用詞袋之法。參考 Tomas Mikolov, Kai Chen, Gregory S. Corrado and Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space,” paper presented to International Conference on Learning Representations, Scottsdale, 2013.

⁷³ Nitin Hardeniya, *NLTK Essentials: Build Cool NLP and Machine Learning Applications Using NLTK and Other Python Libraries*, 1st ed. (Birmingham: Packt Publishing, 2015), pp. 73-92.

⁷⁴ 〔秦〕呂不韋著，陳奇猷校釋，〈情欲〉，《呂氏春秋新校釋》，上冊，頁 86。

⁷⁵ 曾元顯把 IDF 譯為逆向文件頻率，並謂「在各種資訊檢索的實驗中，使用 IDF 公式在詞彙的權重上，其成效幾乎都比沒有使用時來得高，也比單獨使用詞頻 (term frequency，簡稱 TF) 作為權重來得高。當詞頻的概念與逆向文件頻率的概念以 TF*IDF 的乘積公式計算詞彙權重後，其成效更佳。」參考雙語詞彙、學術名詞暨辭書資訊網，<http://terms.naer.edu.tw/detail/1679000>，瀏覽日期：2020 年 8 月 20 日。

次，IDF（inverse document frequency）指逆向的文本頻次，乃文件總數除以包含該詞語之文件的數目。⁷⁶ 二者相乘後便可降低「之」和「欲」的權重，提升其他詞語的數值了。

有了文件詞語矩陣後，便可利用不同公式計算文件之間的相似性（similarity）或距離（distance），然後按照相近程度逐一把文件叢綴起來。在各種聚類分析中，階層聚類法一般而言較 K means 算法更為優勝，⁷⁷ 樹狀圖的排列在視覺上亦能細緻地呈現各個文件之間的關係。此外以 cosine 算式量度距離亦有較佳的表現，⁷⁸ 而華德連結法（Ward linkage）則是常見的聚合方法。⁷⁹ 這些都是文本探勘中常見且有效的公式，用來分析字數有限的古籍文本已然相當足夠，沒有必要對相關算法作出微調或修訂。

坊間不少程式語言和軟件都能完成上述文字向量化、相似性計算與樹狀圖繪製等工作，當中尤以 Python 和 R 語言較為流行。筆者嘗利用這兩種語言作相同運算，發現結果基本上一致，不過 Python 需要的套件較多，指令也較繁複，⁸⁰ 而且對中文的兼容性亦不及 R 語

⁷⁶ IDF 一般取以 10 為底的對數。關於 TF-IDF 具體的操作及作用，可參考 Shahzad Qaiser, Ramsha Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” *International Journal of Computer Applications* 181/1 (2018): 25-29.

⁷⁷ Michael Steinbach, George Karypis and Vipin Kumar, “A Comparison of Document Clustering Techniques,” paper presented to KDD Workshop on Text Mining, Minneapolis, 2000.

⁷⁸ Anna Huang, “Similarity Measures for Text Document Clustering,” paper presented to Conference Department of Computer Science, Hamilton, 2008; Pranjal Singh and Mohit Sharma, “Text Document Clustering and Similarity Measures,” paper presented to Indian Institute of Technology, Kanpur, 2013; Shouvik Sachdeva and Bhupendra Kastore, “Document Clustering: Similarity Measures,” paper presented to Indian Institute of Technology, Kanpur, 2014.

⁷⁹ 華德法在不同軟件中的分別，可參考 Fionn Murtagh and Pierre Legendre, “Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion?,” *Journal of Classification* 31 (2014): 274-295.

⁸⁰ 例如要製作〈十二紀〉聚類分析樹狀圖，Python 須要 Numpy、Pandas、NLTK、Scipy、Scikit-learning、Matplotlib 等套件，但 R 語言只須用 dist、hclust 和 plot 幾個函數（function）便能

言，⁸¹ 因此若以投入時間的回報而論，R 語言有較大的優勢。若僅著眼於聚類分析，亦可選擇一些現成的軟件，如用史丹福大學 Michael Eisen 編寫的 Java Cluster 3.0 計算矩陣距離，再用 Alok Saldanha 的 Java TreeView 生成樹狀圖。⁸² 這些方法雖然繁簡有別，卻均要求使用者對程式語言有一定認識，並須結合使用不同套件，對習慣在微軟視窗平台上使用一體化軟件的文科學者而言，實在不能算是方便（user-friendly）。使用門檻太高的話，學者難以重檢（re-search）相關數據或作同類研究，不免會以姑妄聽之的態度看待聚類分析的結果，以為可聊備一說，卻未必能夠理解和正視這種方法的特性和價值。

筆者為此特別測試了不同工具，包括文科及社會科學中較為常用的一些電腦輔助質性資料分析軟體（Computer-assisted Qualitative Data Analysis Software），如在臺灣已提倡了近二十年的 NVivo。⁸³ 可是這類軟件雖然間或附帶聚類分析的功能，亦頗為易用，但在參數設定上有很大的限制，⁸⁴ 並且需要付費購買。反復測試多種工具後，筆者覺

完成所有工序。參考 Dipanjan Sarkar, *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*, 2nd ed. (New York: Apress, 2019), pp. 453-517; Bradford Tuckfield and Alok Malik, *Applied Unsupervised Learning with R: Uncover Hidden Relationships and Patterns with K-means Clustering, Hierarchical Clustering, and PCA* (Birmingham: Packt Publishing, 2019), pp. 65-80.

⁸¹ Python 的繪圖套件 Matplotlib 不能顯示中文標籤（label），須要用其他方法修訂（如複製中文字體到 Matplotlib 的字體資料夾），R 語言的 plot 函數則能直接顯示中文。

⁸² Michael Eisen and Michiel de Hoon, *Cluster 3.0 Manual*, <http://bonsai.hgc.jp/~mdehoon/software/cluster/cluster3.pdf>, 瀏覽日期：2020 年 8 月 20 日。

⁸³ 劉世閔很早已介紹這套軟件，並持續以此進行不同研究。劉世閔、吳璟，〈NVivo：新世紀的質性研究電腦輔助軟體〉，《慈濟大學人文社會科學學刊》第 1 期（2002 年 6 月），頁 135-152；劉世閔，〈NVivo 11 Plus 的社交網絡分析在個案研究之運用：以七次國編版國語教科書的負面人物為例〉，《臺灣教育評論月刊》第 7 卷第 1 期（2018 年 1 月），頁 317-335。

⁸⁴ 筆者所用微軟系統 NVivo 12 pro 版便有聚類分析（clustering analysis）功能，操作相當簡單，但此軟件在計算詞語相似性時只能使用 TF，不能選擇 TF-IDF 或 smooth TF-IDF，量度方法雖有三種相關系數（correlation coefficient）模式，卻沒有 Cosine 等選項，樹狀圖亦只能用

得 Orange 最為容易操作，不但包含文本探勘的主要功能，還能容許使用者靈活設定各種公式和參數，甚至可配合 Python 程式碼作進階操作，因此決定以此作為本文的分析工具，闡述觀點之餘，亦希望略作推廣，普及數位人文的應用。

四、春夏秋冬四紀主題的聚類分析

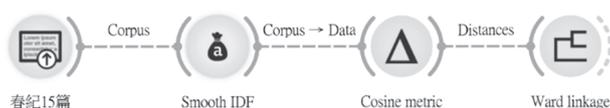
Orange 是由斯洛維尼亞盧比安納大學（University of Ljubljana）生物信息學實驗室（Bioinformatics Laboratory）開發的開源軟件，可以在 Windows、macOS 及 Linux/Source 等不同平台上運作。本文所用的是微軟系統 Orange 3.26.0 版本，⁸⁵ 用家只要在界面工具集（widget）使用以下四個器件，便可快捷地完成文本聚類分析的工作。⁸⁶ 用家只須把完成前處理的文章放在同一檔案夾內，輸入至 Import Documents，然後在 Bag of Words 中設定 TF 和 IDF 等計算方法，即可生成相關詞袋。把詞袋連接至 Distances，設定量度距離的模式後，便可輸出至 Hierarchical Clustering 器件，從中選定樹狀圖連結方法後，即可生成圖表。要明白這個流程如何具體地應用到〈十二紀〉的分析，可以用〈春紀〉諸文作一個先導測試（pilot test）。為了便於閱覽，下表把器件的名稱改為相關內容或參數設定：

complete linkage 方式連結，不能改動。NVivo 12 pro 聚類分析的使用方法及相關公式，可參閱 Nvivo，<https://help-nv.qsrinternational.com/12/win/v12.1.90-d3ea61/Content/vizualizations/how-cluster-analysis-generated.htm>，瀏覽日期：2020 年 8 月 20 日。

⁸⁵ 相關軟件可在官方網站免費下載，參考 Orange，<https://orange.biolab.si/download/#windows>，瀏覽日期：2020 年 8 月 20 日。

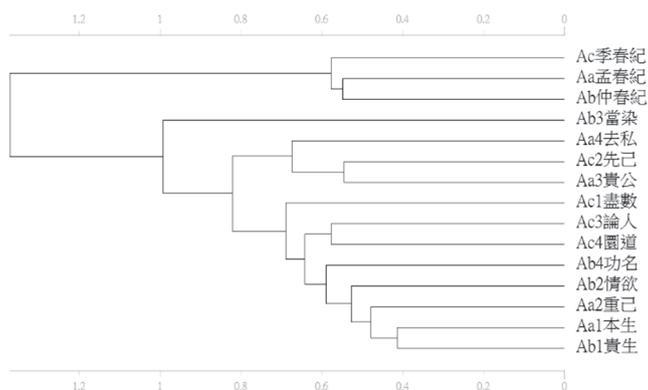
⁸⁶ 表二：Orange 階層聚類分析流程，請見《清華中文學報》電子期刊網頁，<https://cl.site.nthu.edu.tw/var/file/401/1401/img/246322530.png>。

表三：〈春紀〉15 篇階層聚類分析流程



〈春紀〉共有 15 篇文章，生成詞袋時選用了 smooth IDF 算法，於算式中固定加上 1，避免因 0 值而影響到權重的計算；⁸⁷ 接著使用上章所述的 Cosine 計算相似度，再以華德法聚合文件，繪製出以下樹狀圖：

表四：〈春紀〉樹狀圖



表四中篇名前面的大寫 A 表示〈春紀〉，小寫的 abc 分別指孟、仲、季，數字表示該紀中不計算紀首的排列次序，諸紀紀首則不標號，如 Ab1 表示〈貴生〉屬〈仲春紀〉的第 1 篇。樹狀圖的原則是把相似度最高的兩個文件優先聚合起來，然後逐層往外延伸，表四末端的〈貴

⁸⁷ Orange 所用 smooth IDF 的公式與 sklearn 一樣： $\text{idf}(t) = \log[(1+n)/(1+\text{df}(t))]+1$ ，參考 Orange3 Text Mining，<https://orange3-text.readthedocs.io/en/latest/widgets/bagofwords-widget.html>，瀏覽日期：2020 年 8 月 20 日。

生〉與〈本生〉相連，表示它們在內容上最為相似，而二者則又與〈重己〉接近，如此類推。

初步看來，聚類分析確能把相似度較高的文章叢綴起來。以三篇紀首為例，表四清楚顯示〈孟春紀〉與〈仲春紀〉較為相似，二者在下一階層又與〈季春紀〉連在一起；更重要的是，三紀自成一類，與其餘十二篇文章區別開來，沒有從屬關係。這種區分既符合讀者閱讀相關篇章時的印象，也有文獻上的依據，如《禮記·月令》一篇便包含十二紀首的所有文字，把它們當作一個整體。學界對十二紀首與《禮記·月令》的關係向來有兩種相反的看法，第一種以為〈月令〉乃合十二紀首而成，如鄭玄謂〈月令〉「本《呂氏春秋·十二月紀》之首章也，以禮家好事抄合之」，⁸⁸ 徐復觀亦信從此說；⁸⁹ 另一種意見則以為十二紀首乃割裂〈月令〉而來，如楊寬（1914-2005）謂「〈月令〉一篇，當早有成說，呂不韋賓客乃割裂十二月以為〈十二紀〉之首章耳。」⁹⁰ 這裏無意辨析二說孰為正確，只想指出它們的意見雖然看似相反，卻均肯定了十二篇紀首的內容和寫法十分相似，自成一格。因此說十二紀首原屬同一篇文章也好，說它們是後來才被拼合為〈月令〉也好，均顯示十二紀首的內容有一種內在的整體性。聚類分析能夠準確捕捉這種整體性，獨獨把三篇紀首視為一類，正可印證機器學習法的有效性。

接下來要討論的是〈孟春紀〉中「生」這個主題，前面提到余嘉錫認為三春紀首之後的第一篇文章〈本生〉、〈貴生〉和〈盡數〉都是討論養生之道、發凡起例的文章。胡適（1891-1962）同樣肯定「生」的重要，甚至把它視為全書「中心思想」，但他標舉的代表性文章卻與余氏稍有不同：

⁸⁸ [漢]鄭玄注，[唐]孔穎達疏，〈月令第六〉，《禮記正義》，冊2，卷14，頁512。

⁸⁹ 徐復觀，《兩漢思想史》，卷2，頁59-60。

⁹⁰ 楊寬，〈月令考〉，《齊魯學報》1941年第2期（1941年7月），頁27。

《呂氏春秋》的第一紀的第一篇便是〈本生〉，第二篇便是〈重己〉；第二紀的第一篇便是〈貴生〉，第二篇便是〈情欲〉。這都是開宗明義的文字，提倡的是一種很健全的個人主義，叫做「貴生」主義，大體上即是楊朱的「貴己」主義。⁹¹

表四對〈貴生〉諸篇的聚類基本上符合余、胡的觀點，它把〈貴生〉、〈本生〉、〈重己〉、〈情欲〉逐層連結起來，與胡適的主張更是完全一致。由此可見，基於統計學和機器學習的文本探勘能夠模擬人類的工作，對文本作出近乎專家的歸類。

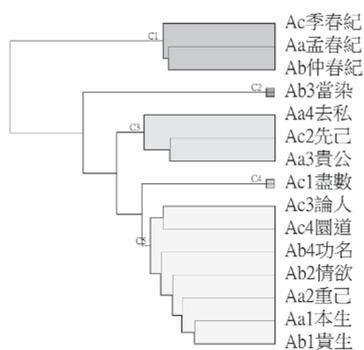
除了驗證已有知識外，文本探勘算法能以有別於人類的方式識別非結構性資料的隱藏模式 (hidden pattern)，有時甚至能夠捕足人類所忽略的細節。⁹² 細看表四，不難發現當中揭示了一些頗為特別的現象，如〈當染〉與其他 11 篇不屬同類，〈盡數〉與〈貴生〉、〈本生〉的關係也顯得有點疏離，不像余嘉錫所說般緊密，這兩點都是饒有意味。我們不必把這些分類視為定論，須知不同算法及參數設定均會對分析結果有細微的影響。然而，我們也不宜完全漠視這些訊息，因為表四既然能夠成功辨識三篇紀首及〈貴生〉諸篇的關係，它對其餘篇章的分析自應有一定的參考價值。因此較為理性的做法乃是按照表四所提示的線索，檢視相關文章的內容，考察當中究竟隱藏哪些元素令得它與其他文章得以聚合起來或是分離出去。在這種情況下，人工智能儼如一位獨立的專家學者，參與了歷來有關〈春紀〉主題和篇章結構的學術對話中，傳統方法與數位技術亦在此得以互動交流，而這正是本文所要強調的「相互作用」(interplay) 的意思。

⁹¹ 胡適，〈讀呂氏春秋〉，《胡適文存三集》，收入胡適著，季羨林主編，《胡適全集》（合肥：安徽教育出版社，2003），卷 3，頁 249。

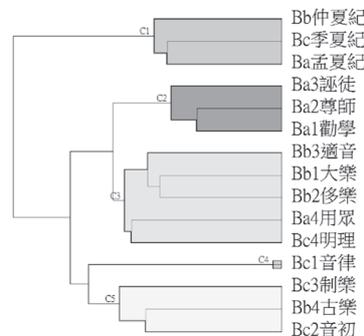
⁹² 這種技術已應用在市場學、醫學、社會關係分析等不同領域，得到廣泛的認同。參考 Robert Nisbet, John Elder and Gary Miner, *Handbook of Statistical Analysis and Data Mining Applications* (London: Academic Press, 2009), pp. 173-195.

〈春紀〉的先導測試初步展現了聚類分析的效度 (validity)，以及這種方法如何有助拓展研究者的思路，接下來便可運用相同的方法，正式分析春、夏、秋、冬各紀的主題。為了方便敘述，以下先把聚類的結果全部列出：

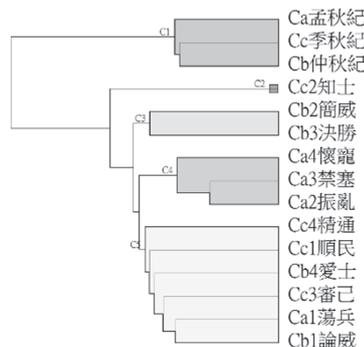
表五：〈春紀〉截枝樹狀圖



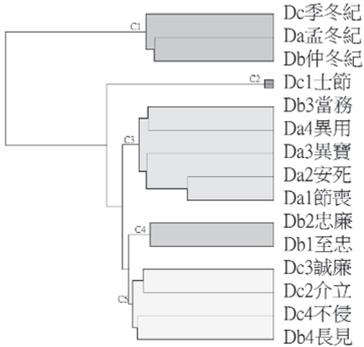
表六：〈夏紀〉截枝樹狀圖



表七：〈秋紀〉截枝樹狀圖



表八：〈冬紀〉截枝樹狀圖



表五與表四都是〈春紀〉15篇的聚類，分析結果完全一樣。二表的分別只是表五刪去了數字刻度，並設定5為截枝 (branch cutting) 值，以不同顏色及編號標示出5個大類，以便閱覽分析。與 K-Means 聚類算法不同，階層聚類法並不需要預設分羣的數值，可以從不同角度

處理分類結果，或是聚焦在不同的細節中。原則上 15 篇文章最多可以分成 15 組，但所有文章分散開來便有違綴合的意義，但類別太少的話，亦不能顯示出文本相互之間的差異。這裏選擇 5 純粹是折衷的做法，其根本目的是要合併諸紀中較為相似的文章，同時又要突顯當中的異質性文本，因為那些異質文本與其他文章不屬同類，極有可能偏離主題，若勉強把它們納入同一個大主題下，便會容易令人產生牽強附會之感。從表五至表八的截枝結果看，分成 5 類基本上已能夠達到預期的效果。

初步看來，四紀的樹狀圖中有幾處地方較為矚目，其一是對十二紀首的處理。表五至表八均能識別出各自所屬的孟、仲、季三篇紀首，把它們歸成 C1 類，與其餘 12 篇文章截然區別過來，這種一致的分類再次確認了聚類分析的有效性。此外《呂氏春秋》書中原來次序相連的篇章，不少被綴合起來，如〈孟春紀〉的〈本生〉與〈重己〉、〈貴公〉與〈去私〉，〈季春紀〉的〈論人〉與〈園道〉；〈孟夏紀〉的〈勸學〉、〈誣徒〉與〈用眾〉，〈仲夏紀〉的〈大樂〉、〈侈樂〉與〈適音〉，〈季夏紀〉的〈音初〉與〈制樂〉；〈孟秋紀〉的〈振亂〉、〈禁塞〉與〈懷寵〉，〈仲秋紀〉的〈簡威〉與〈決勝〉；〈孟冬紀〉的〈節喪〉、〈安死〉和〈異寶〉，〈仲冬紀〉的〈至忠〉和〈忠廉〉，〈季冬紀〉的〈介立〉、〈誠廉〉和〈不侵〉等。如此多的例證很難說是出於偶然，正可說明《呂氏春秋》背後應有一些編訂原則，而聚類分析的結果亦與這些原則有相對應之處。

尤其值得注意的是，各表的截枝分類也可以與專家學者的觀察互相呼應。汪中（1744-1794）是較早討論《呂氏春秋》篇章內容關係的學者，對後世有一定影響，以下試以他的說法為參照系，列出四表中能夠與汪說相應的地方：⁹³

⁹³ 下表引文俱見〔清〕汪中，〈呂氏春秋序〉，《文集》，收入汪中著，蔣秋華、林慶彰編審，王

表九：汪中〈呂氏春秋序〉與聚類結果相應之處

汪中〈呂氏春秋序〉	列表聚類
故〈勸學〉、〈尊師〉、〈誣徒〉（一作〈詆役〉）、〈善學〉（一作〈用眾〉）四篇，皆教學之方。	表六 C2：〈勸學〉、〈尊師〉、〈誣徒〉
〈大樂〉、〈侈樂〉、〈適音〉（一作〈和樂〉）、〈古樂〉、〈音律〉、〈音初〉、〈制樂〉，皆論樂。	表六 C3：〈大樂〉、〈侈樂〉、〈適音〉 表六 C5：〈古樂〉、〈音初〉、〈制樂〉
〈貴生〉、〈情欲〉、〈盡數〉、〈審分〉、〈君守〉五篇，尚清淨養生之術，則道家流也。	表五 C5：〈貴生〉、〈情欲〉
〈蕩兵〉（一作〈用兵〉）、〈振亂〉、〈禁塞〉、〈懷寵〉、〈論威〉、〈簡選〉、〈決勝〉、〈愛士〉七篇，皆論兵，則兵權謀、形勢二家也。	表七 C3：〈簡選〉、〈決勝〉 表七 C4：〈振亂〉、〈禁塞〉、〈懷寵〉 表七 C5：〈蕩兵〉、〈論威〉、〈愛士〉

汪中涉及的篇章範圍不限於〈十二紀〉（如〈審分〉和〈君守〉），因此單就〈十二紀〉而論，表九所列的聚類絕大部分都與汪說相合。這種結果並不令人感到意外，因為在先導測試中，〈春紀〉已有效地把胡適所說的幾篇文章聚合起來。現在既然已確定聚類的結果有一定的參考價值，接下來便可考察春、夏、秋、冬樹狀圖的截枝類別與前人所述生、長、收、藏四個主題的異同了。

清信、葉存芳點校，《汪中集》（臺北：中央研究院中國文哲研究所，2000年），卷4，頁148-149。

〈春紀〉的 C5 包含〈本生〉、〈重己〉、〈貴生〉、〈情欲〉諸文，正是「生」這個主題的代表性篇章；不過要注意的是，〈仲春紀〉首篇〈盡數〉與 C5 的關係並不緊密，自成 C4 類，而 C4 和 C5 又明顯與 C3 的〈貴公〉、〈去私〉等別為一類，反映公、私這個主題有一定的獨立性。過去學者未能正視 C3 的異質性，勉強把它納入養生的主題，結果導致各種附會的問題。

〈夏紀〉多論學和論樂的文章，表六亦能作出有效的分類。其中 C2 三篇文章俱是論學之文，C3 至 C5 則多論樂之文，這兩個主題符合前人所謂長養或樂氣之義，爭議不大。當中較為特別的是，一般被視為論學的〈用眾〉不在 C2 而在 C3，又 C3 雖多論樂之文，卻與 C4 和 C5 不在一類，反而與 C2 相連。

〈秋紀〉的情況更是簡單，該紀以〈蕩兵〉為始，多言用兵之事，應肅殺之氣。表七除〈知士〉較為特別外，餘下 11 篇被歸屬為同一大類，反映主題相當單一，各篇內容高度相似。倘若聚焦該 11 篇在分羣中的組合，C3 二文來自〈仲秋紀〉、C4 三文均屬〈孟秋紀〉，C5 五篇文章中則有三文屬〈季秋紀〉，與《呂氏春秋》原來次序亦有相符之處，再次確認該書編排實有一套內在的原則。

〈冬紀〉以「死」為主題，C3 把〈節喪〉和〈安死〉等文章連結起來，正符合此義，但此一分枝卻與 C4 和 C5 不屬同類，後二者與所代表的死節主題自成一類，相對獨立。這與〈春紀〉言養生之道，卻又夾雜道德性的〈貴公〉、〈去私〉等主題的情況十分相似，不過死節畢竟仍與死亡有關，二者關係尚算密切，不難理解。

總括而言，四紀的聚類分析大體上能夠佐證前人所述生、長、收、藏四大主題，並能細緻地區別出一些相對獨立、偏離主題的篇章。當中〈夏〉、〈秋〉二紀的主題較為明確，與四庫館臣最初的觀察一致，〈冬紀〉有關死節的文章亦能與死、葬的主題相容，只有〈春紀〉中養生與去私兩個主題較難銜接貫通。不過要是繼續深究在樹狀圖中被

單獨標示出來的異質性篇章，可以發現它們在《呂氏春秋》全書結構中承擔了鮮為人知、卻又異常重要的功能，有待認真抉發。

五、異質性文本的元素和功能

聚類分析既然能夠仿效專家學者的分類模式，提出「一家之言」，其分析結果中一些與別不同的地方，便不應該等閒視之了。儘管表五大體上印證了前人所謂「生」乃〈春紀〉的大主題，但在一些細節上仍有值得深思的地方。如徐復觀論及〈春紀〉諸篇關係時提出了三點：（1）「〈貴生〉、〈情欲〉，是孟春〈重己〉的重述與發揮」；（2）「〈盡數〉、〈先己〉，是〈重己〉、〈貴己〉（按當指〈貴生〉）的重述與發揮」；（3）「〈論人〉、〈園道〉，乃〈當染〉的發揮」。⁹⁴ 可是這三點中只有第一點與表五 C5 相符，第二點提及的〈先己〉在表五中隸屬 C3，〈盡數〉則自成 C4 類，與 C5 的〈重己〉和〈貴生〉不屬同類；此外第三點所論〈當染〉亦單獨成 C2 類，與 C5 的〈論人〉和〈園道〉更是有明顯的距離。把文字轉換成數字矩陣後，聚類算法究竟是基於甚麼特徵把〈盡數〉和〈當染〉區別開來？要瞭解這些差異純粹是人工智能的誤判，還是它確實捕捉到人類忽視的文本細節，最簡單和有效的方法自然是細讀這幾篇文章，⁹⁵ 考察當中的內容主題和異質性元素。

先看〈盡數〉，余嘉錫和徐復觀均認為〈盡數〉可與〈貴生〉諸篇互相發明，但胡適論及貴生主義開篇明義的篇章時卻沒有提及此文。這兩種說法均與表五沒有衝突，因為 C4 自成一類，與 C5 諸篇不同，

⁹⁴ 徐復觀，《兩漢思想史》，卷 2，頁 36-37。

⁹⁵ 數位人文中也有各種文本分析工具，如關鍵詞提取（keyword extraction）、文本比較（text comparison）及 n 字詞分析（n-gram analysis）等，可以擷取、標示或計量不同文本之間的詞語差異和關聯。然而本文主要介紹聚類分析法，希望展示量化分析如何有助質性研究，彼此互動，為免令焦點模糊，本節不會涉及其他工具，而是沿用古代文獻學者熟悉的文本細讀及互文比對之法，檢視異質性元素。

與胡適說法相通；但 C4 和 C5 卻又在下一階層中連成一類，較其餘三大類更為緊密，所以余、徐之說亦似有一定的道理。〈盡數〉這種相對獨立的位置顯然與篇中所用的詞語有關，〈本生〉和〈貴生〉極言生之可貴，⁹⁶ 詳述節欲與養生的關係，⁹⁷ 與〈重己〉、〈情欲〉的內容和用語俱十分相似。⁹⁸ 反觀〈盡數〉雖然也談養生去害之旨，但正如陳奇猷指出，此篇所論重點在「飲食得宜，而精氣流動」，⁹⁹ 其中「精氣」一詞正是〈貴生〉諸文從未出現過的字眼，論者以為此觀念乃受稷下學派的影響，並非源自原始道家之說。¹⁰⁰ 〈盡數〉以「精氣」論養生，整體上符合「貴生」之旨，卻又有所突破，擴展了〈貴生〉篇中子華子等人的學說，正可解釋為甚麼 C4 既獨立成類，卻又在下一階層與 C5 相連。「精氣」一詞不但見於 C4 的〈盡數〉，亦出現在 C3 的〈先己〉和 C5 的〈園道〉，¹⁰¹ 三篇文章在表中位置相鄰，後二

⁹⁶ 〈本生〉云：「始生之者，天也；養成之者，人也。能養天之所生而勿撓之謂天子」，參考〔秦〕呂不韋著，陳奇猷校釋，〈本生〉，《呂氏春秋新校釋》，上冊，頁 21；〈貴生〉云：「聖人深慮天下，莫貴於生」，參考〔秦〕呂不韋著，陳奇猷校釋，〈貴生〉，《呂氏春秋新校釋》，上冊，頁 75。

⁹⁷ 〈本生〉云：「是故聖人之於聲色滋味也，利於性則取之，害於性則舍之，此全性之道也」，參考〔秦〕呂不韋著，陳奇猷校釋，〈本生〉，《呂氏春秋新校釋》，上冊，頁 21；〈貴生〉云：「夫耳目鼻口，生之役也。耳雖欲聲，目雖欲色，鼻雖欲芬香，口雖欲滋味，害於生則止。」，參考〔秦〕呂不韋著，陳奇猷校釋，〈貴生〉，《呂氏春秋新校釋》，上冊，頁 75。

⁹⁸ 如〈重己〉「今吾生之為我有，而利我亦大矣。論其貴賤，爵為天子，不足以比焉；論其輕重，富有天下，不可以易之；論其安危，一曙失之，終身不復得」，便是重述生之可貴，參考〔秦〕呂不韋著，陳奇猷校釋，〈重己〉，《呂氏春秋新校釋》，上冊，頁 34；〈情欲〉「天生人而使有貪有欲。欲有情，情有節。聖人修節以止欲，故不過行其情也」，便是重申節欲的重要，參考〔秦〕呂不韋著，陳奇猷校釋，〈情欲〉，《呂氏春秋新校釋》，上冊，頁 86。

⁹⁹ 〔秦〕呂不韋著，陳奇猷校釋，〈盡數〉，《呂氏春秋新校釋》，上冊，頁 140。

¹⁰⁰ 參考部建華、樓宇烈，〈《呂氏春秋》中的「精氣說」〉，《華僑大學學報（哲學社會科學版）》2017 年第 3 期（2017 年 6 月），頁 40-53。

¹⁰¹ 〈先己〉云：「精氣日新，邪氣盡去，及其天年」，參考〔秦〕呂不韋著，陳奇猷校釋，〈先己〉，《呂氏春秋新校釋》，上冊，頁 146；〈園道〉云：「精氣一上一下，園周復雜，無所稽留，故曰天道園」，參考〔秦〕呂不韋著，陳奇猷校釋，〈園道〉，《呂氏春秋新校釋》，上冊，

者也許正是因為這些用語而得以與〈盡數〉連結成一大類，與 C2 的〈當染〉區別開來。

〈仲春紀〉的〈先己〉也有類似的現象，它不與同紀的〈論人〉、〈園道〉相連，而是與〈孟春紀〉的〈貴公〉和〈去私〉組成 C3。這種組合應是由於〈先己〉與 C3 其餘兩篇文章有不少共同用語，如「天下」、「寡人」均為三者所共有，卻不見於〈論人〉、〈園道〉中。驟看之下，這些詞匯似乎與內容主旨無甚關係，但沿著此一方向細心對比，仍可發現一些重要的線索，如「孔子」一詞不見於〈論人〉、〈園道〉，但在〈先己〉卻出現了四次，〈貴公〉、〈去私〉亦各有一次。它們所引用的「孔子」言行雖不見於《論語》，卻也與《論語》所述孔子學說沒有甚麼衝突，屬儒、道共通的主張，反映了《呂氏春秋》融會諸家的傾向，而這兩種思想又因〈先己〉的「精氣」平添了稷下學派的色彩，並通過〈盡數〉而得以與 C5 的貴生主題連結起來，構成一個大類。以養生為主題的 C5 諸文（如〈重己〉）本來只是強調個體（己）生命的重要，但得到〈盡數〉這個中介後，「己」這個關鍵詞卻可以與〈先己〉的「反於己身」間接地貫通起來，增添了〈貴公〉和〈去私〉的道德元素。這種扣連是否成功當然可以再作討論，卻可以解釋為何崇尚個體安危的貴生主義最後沒有推導出遁世隱居一類典型的道家主張，反而與鼓吹大公無私的儒、墨之學串聯起來。

從上述〈盡數〉和〈先己〉的「精氣」、「孔子」、「己」等例，可以得出一個斷言：異質性元素乃由個別字詞所構成，這些字詞能夠發揮橋樑的作用，鈎連相近學說，藉此延伸這些學說的意涵。此一斷言要是成立，不但能夠說明《呂氏春秋》融合各家學說的方法，還能為〈十二紀〉結構的探討開拓新的視角，突破前人由於過分拘執於四紀各自的主題，因而忽視了春生、夏長、秋收、冬藏之間的內在聯繫。

遵循此一思路，〈當染〉在〈春紀〉中所占的特殊位置便有解答的線索。

〈當染〉在表五中獨自構成 C2 一類，與其餘 11 篇截然區別開來。這種編排不難理解，因為〈當染〉全文亦見於《墨子·所染》篇，二者文字高度相似。以下試列出兩篇文章開首幾段文字，分別以方框、間線和圓圈標註出相異、增加及減少的文字，並把斷句、標點略作統一，¹⁰² 令讀者對二者的相似性有具體的認知：

表十：〈當染〉與〈所染〉互文對照

《呂氏春秋·當染》	《墨子·所染》
○墨子○見染素絲者而歎曰：「染於蒼則蒼，染於黃則黃，所以入者變，其色亦變，五入○而以為五色矣。」故染不可不慎也。	子墨子言見染○絲者而歎曰：「染於蒼則蒼，染於黃則黃。所○入者變，其色亦變，五入必，而已則為五色矣。」故染不可不慎也。
非獨染絲然也，國亦有染。舜染於許由、伯陽，禹染於皋陶、伯益，湯染於伊尹、仲虺，武王染於太公望、周公旦，此四王者所染當，故王天下，立為天子，功名蔽天地，舉天下之仁義顯人，必稱此四王者。	非獨染絲然也，國亦有染。舜染於許由、伯陽，禹染於皋陶、伯益，湯染於伊尹、仲虺，武王染於太公○、周公○。此四王者所染當，故王天下，立為天子，功名蔽天地。舉天下之仁義顯人，必稱此四王者。
夏桀染於干辛、 <u>歧踵戎</u> ，殷紂染於崇侯、惡來，周厲王染於 <u>虢公長父</u> 、 <u>榮夷終</u> ，幽王染於 <u>虢公鼓</u> 、 <u>祭公敦</u> ，此四王者所染不當，故國殘身死，為天下僂，舉天下之不義辱人，必稱此四王者。	夏桀染於干辛、 <u>推哆</u> ○，殷紂染於崇侯、惡來，○厲王染於 <u>厲公長父</u> 、 <u>榮夷終</u> ，幽王染於 <u>傅公夷</u> 、 <u>蔡公穀</u> 。此四王者所染不當，故國殘身死，為天下僂。舉天下不義辱人，必稱此四王者。

¹⁰² 〈當染〉據〔秦〕呂不韋著，陳奇猷校釋，〈當染〉，《呂氏春秋新校釋》，上冊，頁 96-97；〈所染〉據〔清〕孫詒讓著，孫啟治點校，〈所染〉，《墨子閒詁》（北京：中華書局，1985 年），上冊，卷 1，頁 11-14。斷句、標點有異者，則斟酌二書略作統一。

〈當染〉大量襲用墨家學派的文字，在內容和風格上自然與崇尚道家的〈貴生〉和〈情欲〉諸文大不相同，別為一類，正是自然不過的事。以道家養生為主題的〈春紀〉忽然羈入〈當染〉一文，本是相當突兀的事，但過去學者卻鮮有正視這些異質性元素，反而致力彌縫二者的差距，如劉咸忻說：「〈春紀〉言本生、貴己是論其本質，然所謂生、所謂己，非夙成也、必有所染；故因及乎染。」¹⁰³ 這類以情理為依據的雄辯並不是嚴格的邏輯推論，並無太大的說服力，因為運用相同方法的話，〈情欲〉篇之後的文章即使不是〈當染〉，也一樣可以說出一番似是而非的道理來，設想〈情欲〉之後的文章是〈勸學〉，只須把劉氏那段話中的「染」字改為「學」，同樣說得通。要是橫說豎說皆可通，這類文字便沒有甚麼解釋力了。若勉強把〈當染〉牽扯到生的主題，謂「〈貴生〉、〈情欲〉主要強調人的自我約束，而〈當染〉篇則關注外部因素對君主生存狀態所起的作用」，¹⁰⁴ 由是連類至「生存狀況」，那便無所不包，比「本質」、「夙成」等話更為牽強了。

〈盡數〉和〈先己〉令我們認識到異質性文本能夠以相近字詞為中介，把不同學說鉤連起來，然則〈當染〉在〈春紀〉中又有沒有，或有甚麼連接作用？單從表五看，C2的〈當染〉與C3的〈去私〉、〈先己〉、〈貴公〉位置相鄰，這四篇文章共通而又有實質意義的字詞正是「孔子」，反映該文同樣不是以道家學說為中心。可是〈當染〉之中又有哪些異質性元素，令得它不能與其餘諸篇併為一類？這個問題的答案可以在該文末段中找到。〈當染〉絕大部分文字與〈所染〉相似，只有末段例外：¹⁰⁵

¹⁰³ 王利器，〈當染〉，《呂氏春秋注疏》，冊1，頁202。

¹⁰⁴ 管宗昌，《《呂氏春秋》文學研究》，頁30。

¹⁰⁵ 參考〔秦〕呂不韋著，陳奇猷校釋，〈當染〉，《呂氏春秋新校釋》，上冊，頁98；〔清〕孫詒讓著，孫啟治點校，〈所染〉，《墨子閒詁》，上冊，卷1，頁18-20。按：因字體問題，「禽滑釐」之「釐」已據畢沅校語改訂。參考〔漢〕高誘注，〔清〕畢沅校，《呂氏春秋新校正》（臺北：世界書局，1955年），頁21。

表十一：〈當染〉與〈所染〉末段原文

《呂氏春秋·當染》	《墨子·所染》
<p>非獨國有染也。孔子學於老聃、孟蘇夔、靖叔。魯惠公使宰讓請郊廟之禮於天子，桓王使史角往，惠公止之，其後在於魯，墨子學焉。此二士者，無爵位以顯人，無賞祿以利人，舉天下之顯榮者必稱此二士也。皆死久矣，從屬彌眾，弟子彌豐，充滿天下，王公大人從而顯之，有愛子弟者隨而學焉，無時乏絕。子貢、子夏、曾子學於孔子，田子方學於子貢，段干木學於子夏，吳起學於曾子。禽滑釐學於墨子，許犯學於禽滑釐，田繁學於許犯。孔、墨之後學顯榮於天下者眾矣，不可勝數，皆所染者得當也。</p>	<p>非獨國有染也，士亦有染。其友皆好仁義，淳謹畏令，則家日益，身日安，名日榮，處官得其理矣，則段干木、禽子、傳說之徒是也。其友皆好矜奮，創作比周，則家日損、身日危、名日辱，處官失其理矣，則子西、易牙、豎刀之徒是也。《詩》曰「必擇所堪，必謹所堪」者，此之謂也。</p>

除第一句外，二者差異如此之大，已很難用簡單的符號標示出來。陳奇猷認為〈當染〉「末一段由本篇作者所增」，¹⁰⁶ 乃呂書編者刻意為之。〈當染〉末段比〈所染〉多出了「學」這個關鍵詞，共出現 11 次之多，但這個重要的字眼卻竟不見於〈春紀〉其餘 11 篇文章中，這恐怕正是它不能與其他文章拼合在一起的主要原因。若說異質性文字有鈎連的作用，〈當染〉中那個無法在〈春紀〉其餘文章中找到的「學」字，又可以有甚麼功能？要回應這個問題，須要突破過去以為四紀各

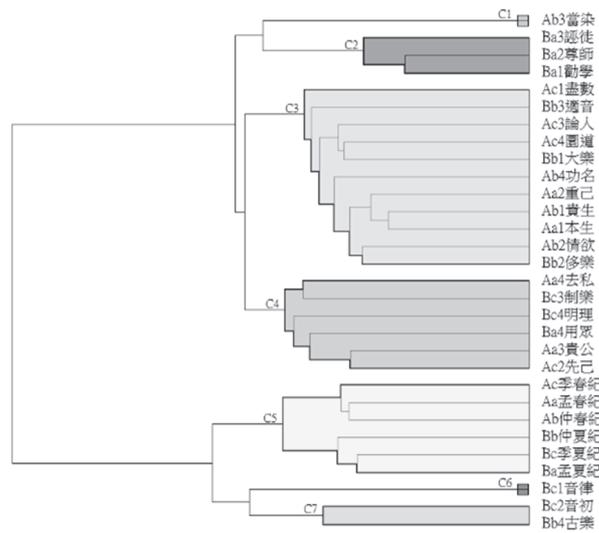
¹⁰⁶ [秦]呂不韋著，陳奇猷校釋，〈當染〉，《呂氏春秋新校釋》，上冊，頁 98。

有單一主題的思路限制，從較為宏觀的角度考察各紀的銜接關係。

六、〈十二紀〉篇章結構關係新探

表五至表八均是以春、夏、秋、冬各紀中的十五篇文章為素材，展示這些文章之間的聚類結果。假如把視線擴大，考慮紀與紀之間的關係，以聚類算法分析〈春紀〉和〈夏紀〉合共 30 篇文章，便可以得出另一幅有趣的圖像：

表十二：〈春紀〉與〈夏紀〉截枝樹狀圖



表十二的截枝數值以 7 為限，原因是分成 8 類的話，C5 中 6 篇紀首便會拆為兩截，也就是說當中所描述的差異已深入至紀首的細微特徵，不免過於瑣碎。因此要說明異質性篇章的作用，分為 7 類已相當足夠。表十二顯示 C1 〈當染〉依然自成一類，但不再孤零零地脫離所有篇章，而是與 C2 〈孟夏紀〉三篇論學文章〈勸學〉、〈尊師〉和

〈誣徒〉組成一大類。究其原因，自是由於四篇文章中均多次出現「學」這個關鍵詞。單看〈春紀〉的話，〈當染〉無疑是異類，但宏觀地看，它卻儼然是〈夏紀〉的伏筆，預告了表六 C2 幾篇文章的學習主題，有連結二紀、承上啟下的作用。

除〈當染〉外，表十二還有不少類別由〈春紀〉和〈夏紀〉的篇章交錯而成，這些組合皆甚有啟發性。如〈仲夏紀〉的〈侈樂〉在表六中本與同紀的〈大樂〉相連，但在表十二中卻與〈情欲〉構成一類，與〈大樂〉反有較遠的距離。這種排列意味著〈侈樂〉和〈大樂〉雖同樣論樂，但相較之下，前者的內容實與〈情欲〉更為接近。翻閱兩篇文章，〈情欲〉一再強調節欲的重要：「聖人修節以止欲，故不過行其情也」，¹⁰⁷〈侈樂〉恰好也否定奢侈之樂，主張「制乎嗜欲」，¹⁰⁸二者確有相通之處。節欲是〈春紀〉論養生時提到的主要方法，反復出現在〈本生〉、〈貴生〉諸文，〈侈樂〉對嗜欲的理解正好呼應〈春紀〉這些篇章，而它本身又以「樂」為主題，符合〈夏紀〉的大方向，因而也發揮了樞紐作用。

〈孟夏紀〉的〈用眾〉同樣有類似的情況。過去學者一般以為此篇與前面〈勸學〉、〈誣徒〉和〈尊師〉同屬一類，乃汪中所謂「皆教學之方」，¹⁰⁹可是表六中〈用眾〉歸入 C3，與〈明理〉同類，反而與〈勸學〉等三篇組成的 C2 有較大的距離。仔細比較〈用眾〉和〈明理〉後，並未能發現二者有甚麼共通的、富有內容意義的關鍵詞，它們被綴合起來，似乎純粹緣於二文的風格特徵與其他篇章有較大的差異，所以才會相對地變為同類而已。然而要是把春、夏二紀連結起來，便可發現〈用眾〉的異質性功能。〈用眾〉在表十二中雖然依舊

¹⁰⁷ [秦] 呂不韋著，陳奇猷校釋，〈情欲〉，《呂氏春秋新校釋》，上冊，頁 86。

¹⁰⁸ [秦] 呂不韋著，陳奇猷校釋，〈侈樂〉，《呂氏春秋新校釋》，上冊，頁 269。

¹⁰⁹ [清] 汪中，〈呂氏春秋序〉，《文集》，收入汪中著，蔣秋華、林慶彰編審，王清信、葉存芳點校，《汪中集》，卷 4，頁 148。

與 C2 論學諸文分離，但它在 C4 中卻與下一階層的〈貴公〉、〈先己〉相連。從內容上看，〈用眾〉仍與學習有關，但其重點是「善學者，假人之長以補其短」，¹¹⁰ 即借用他人之所長，而非像〈勸學〉那樣專論個體的學習，因此末段云：

天下無粹白之狐，而有粹白之裘，取之眾白也。夫取於眾，此三皇、五帝之所以大立功名也。凡君之所以立，出乎眾也。¹¹¹

這番議論涉及天子與眾人的關係，與〈貴公〉的內容頗為類似：

嘗試觀於上志，有得天下者眾矣，其得之以公，其失之必以偏。凡主之立也，生於公。¹¹²

〈用眾〉承接〈孟夏紀〉論學的主題後，擴展至對「眾」的肯定，由是帶出聖王之道，呼應了〈春紀〉貴公、去私的主題。這與〈侈樂〉把〈夏紀〉的「樂」與〈春紀〉節欲的主張連接起來的作用正相一致。它們的異質性雖然沒有〈當染〉那樣明顯，卻同樣有連類不同主題的功能，乃《呂氏春秋》融合眾說的重要手段。

繼續發掘下去，還有不少值得玩味的地方，像前面提到文本探勘主要以詞頻為基礎，沒有特別顧及一詞多義的現象，但聚類分析根據整篇文章的詞頻矩陣計算出來的結果，似乎能夠彌補多義的問題。如〈春紀〉的大主題是「生」，表十二中「生」字出現次數最多的三篇文章分別是〈音律〉（31次）、〈貴生〉（29次）和〈情欲〉（11次）。純就詞頻而論，〈音律〉應與〈貴生〉更為接近，但〈音律〉「黃鐘生林

¹¹⁰ [秦] 呂不韋著，陳奇猷校釋，〈用眾〉，《呂氏春秋新校釋》，上冊，頁 235。

¹¹¹ [秦] 呂不韋著，陳奇猷校釋，〈用眾〉，《呂氏春秋新校釋》，上冊，頁 236。

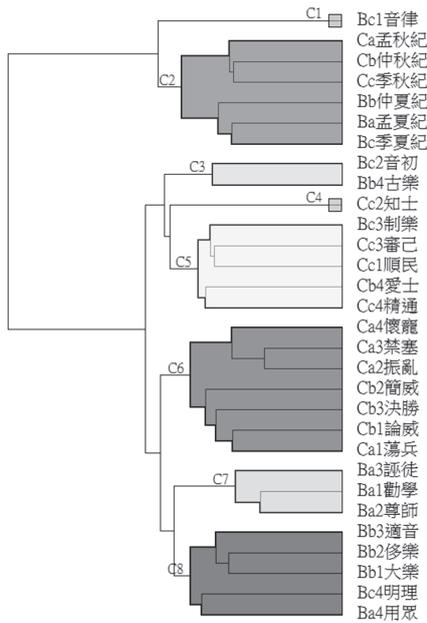
¹¹² [秦] 呂不韋著，陳奇猷校釋，〈貴公〉，《呂氏春秋新校釋》，上冊，頁 45。

鐘」等二十多個「生」字，¹¹³ 無論在意義和用法上，俱與〈貴生〉和〈情欲〉的「生」所強調的生命之意有所不同。令人意外的是，表十二竟能有效地偵測此一關鍵，沒有把〈音律〉和〈貴生〉連類起來，而是把〈音律〉獨立標作 C6，〈貴生〉則在 C3 中與〈情欲〉歸為同一組。

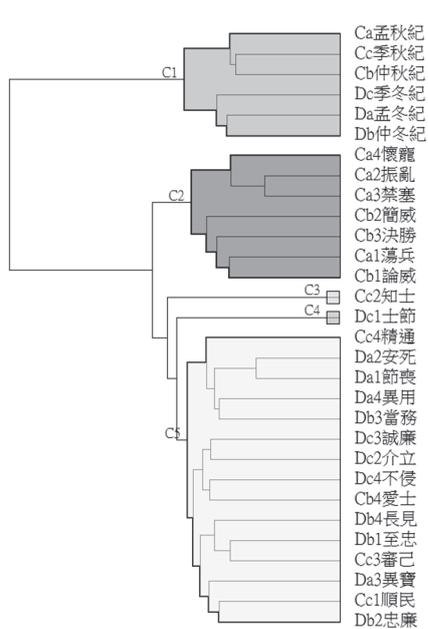
此外，表六中〈音律〉自成 C4 一類，與 C5 〈制樂〉、〈古樂〉和〈音初〉在下一階層連成一類，但在表十二中〈音律〉雖仍是單獨成為 C6 一類，卻在下階層中與 C7 的〈音初〉和〈古樂〉相連，〈制樂〉則不在其中；至於 C6 和 C7 則在下一階層與 C5 的六篇紀首相連，與其餘 21 篇文章區別開來。這些組合亦與文本的特色有關，〈音律〉大談三分損益之法，全文屢見黃鐘、林鐘、太簇、南呂、姑洗、應鐘、蕤賓、大呂、夷則、夾鐘、無射、仲呂等，又有「黃鐘之月」、「大呂之月」等話，與十二紀首開段的某某之月、律中某某一類用語相當類似，因此把它們歸在同一大類，亦有一定的道理。在聚類分析的指引下細讀相關篇章，無疑能夠加深我們對文章的用語和內容細節的理解，學者即使最終找不到異質之處，或是未能同意分類的結果，也可以得到不少啟發。這類細節在各樹狀圖中為數不少，礙於篇幅，無法逐一細表，以下只能承接表十二的分析方向，考察〈夏紀〉與〈秋紀〉、〈秋紀〉與〈冬紀〉之間的關係，略述表七、表八中異質性文章的銜接功能。

¹¹³ [秦] 呂不韋著，陳奇猷校釋，〈音律〉，《呂氏春秋新校釋》，上冊，頁 328。

表十三：〈夏紀〉與〈秋紀〉
截枝樹狀圖



表十四：〈秋紀〉與〈冬紀〉
截枝樹狀圖



表十三和表十四仍是以六篇紀首聚成一類為前提，作最大限度的截枝，結果分別析出 8 類和 5 類。表十三的分類與原書的編排相當接近，如 C7 三文皆來自〈孟夏紀〉，與其相連的 C8 諸文同樣全屬〈夏紀〉；C6 全屬〈秋紀〉，C5 亦只有〈制樂〉屬入其中，兩紀之間間線亦算分明。關於〈音律〉的特性，上文討論表十二時已有涉及，而在沒有〈當染〉的對照情況下，它在表十三中便直接與 C2 的紀首編成一大類，與 C3 的〈音初〉和〈古樂〉分離開來，反映它的音樂術語實與紀首更為接近。

原屬〈季夏紀〉的〈制樂〉納入 C5 亦不難理解，因為該文只有首段兩句提到「樂」：

欲觀至樂，必於至治。其治厚者其樂治厚，其治薄者其樂治薄，亂世則慢以樂矣。¹¹⁴

之後所述故事均與音樂無關。孫鏘鳴（1817-1901）謂「此篇歷引成湯、文王、宋景公之事，與樂制初不相涉，疑必〈明理篇〉文而錯簡在此」。¹¹⁵ 說〈制樂〉是〈明理〉錯簡未必正確，¹¹⁶ 但說此文與樂制不相涉，則是準確地道出了成湯等故事與「樂」的疏離關係。陳奇猷辯說「聲音之道與政通之旨，是以欲制樂者，必先平其政」，¹¹⁷ 其實也是承認了此文並不如〈音律〉那樣直接談論音樂的制作和制度，而是旁涉政治之事，與〈明理〉言「治亂之理」互為表裏。¹¹⁸ 表十三把〈制樂〉安置在〈順民〉和〈審己〉附近，原因是前者與後二者所述故事頗有相近之處。〈制樂〉三個故事都強調國家出現自然異象時，人君應反躬自思，加意為治，這與〈審己〉高誘注所謂「求諸己則存，求諸人則亡」之意正相符合。¹¹⁹ 此外〈順民〉提到湯以身禱告曰：「萬夫有罪，在余一人。無以一人之不敏，使上帝鬼神傷民之命」，¹²⁰ 更是與〈制樂〉所述文王（1152-1056 B.C.）之言「夫天之見妖也，以罰有罪也。我必有罪，故天以此罰我也」等語若合符節。¹²¹ 由是而言，〈制樂〉的異質特徵也有連結的功能，把〈夏紀〉的音樂主題與〈秋紀〉的人君之道聯繫在一起。

相較之下，表十三中獨成 C4 一類的〈知士〉，其作用便不那麼明

¹¹⁴ 〔秦〕呂不韋著，陳奇猷校釋，〈制樂〉，《呂氏春秋新校釋》，上冊，頁 350。

¹¹⁵ 〔秦〕呂不韋著，陳奇猷校釋，〈制樂〉，《呂氏春秋新校釋》，上冊，頁 351。

¹¹⁶ 陳奇猷對此已有詳細辨析，參考〔秦〕呂不韋著，陳奇猷校釋，〈制樂〉，《呂氏春秋新校釋》，上冊，頁 351-352。

¹¹⁷ 〔秦〕呂不韋著，陳奇猷校釋，〈制樂〉，《呂氏春秋新校釋》，上冊，頁 352。

¹¹⁸ 〔秦〕呂不韋著，陳奇猷校釋，〈明理〉，《呂氏春秋新校釋》，上冊，頁 363。

¹¹⁹ 許維遜，〈順民〉，《呂氏春秋集釋》，上冊，頁 209。

¹²⁰ 〔秦〕呂不韋著，陳奇猷校釋，〈順民〉，《呂氏春秋新校釋》，上冊，頁 485。

¹²¹ 〔秦〕呂不韋著，陳奇猷校釋，〈制樂〉，《呂氏春秋新校釋》，上冊，頁 350。

顯。仔細閱讀它與鄰近的〈制樂〉、〈審己〉諸文，仍未能發現當中有甚麼相近的用語或內容。不過把目光轉至表十四，卻可看到〈知士〉雖是獨自構成 C3 一類，卻在下一階層與同樣自成一類的 C4〈士節〉，並與 C5 一大堆文章互相連結。這一分類相當合理，〈知士〉篇幅短小，首段總論「夫士亦有千里，高節死義，此士之千里也」之旨，然後引用「劑貌辨之所以外生樂、趨患難故」一事為證；¹²² C4 的〈士節〉同樣只有一段總論和一個故事，開篇云：「士之為人，當理不避其難，臨患忘利，遺生行義，視死如歸」，¹²³ 接著講述北郭騷自刎之事。二者內容十分相似，呼應了 C5 中死喪和士節的主題。

表十四的 C5 共有 15 篇文章，彼此之間高度相似，即使把表十四的截枝值提高至 6，C5 的組合仍然不變，倒是 C1〈秋紀〉三篇紀首和〈冬紀〉三篇紀首會拆散開來，反映 C5 諸文的關係非常緊密。前面討論表八時已指出，士節是〈冬紀〉的一大主題，C5 便收錄了不少這方面的文章。從〈仲冬紀〉第一篇〈至忠〉提到申公子培、文摯等事，¹²⁴ 到〈季冬紀〉第四篇〈不侵〉所謂「天下輕於身，而士以身為人」，¹²⁵ 皆重申士節以至死節之義，正與 C3 的〈知士〉和 C4 的〈士節〉旨趣相通。〈知士〉在表十三的夏秋聚類中作用不明，原因是它的異質性元素與〈夏紀〉無關，但在表十四中，它卻是連接〈秋紀〉與〈冬紀〉的橋樑。可見〈知士〉與〈春紀〉中的〈當染〉、〈夏紀〉中的〈制樂〉一樣，都有綴合不同學說、貫通不同主題的功能。繼續檢查表十二至表十四，仍可發現不少異質性文本，但大體而言，它們的功能可以通過表十五簡單地展示出來。

¹²² [秦] 呂不韋著，陳奇猷校釋，〈知士〉，《呂氏春秋新校釋》，上冊，頁 496-497。

¹²³ [秦] 呂不韋著，陳奇猷校釋，〈士節〉，《呂氏春秋新校釋》，上冊，頁 629。

¹²⁴ [秦] 呂不韋著，陳奇猷校釋，〈至忠〉，《呂氏春秋新校釋》，上冊，頁 584-585。

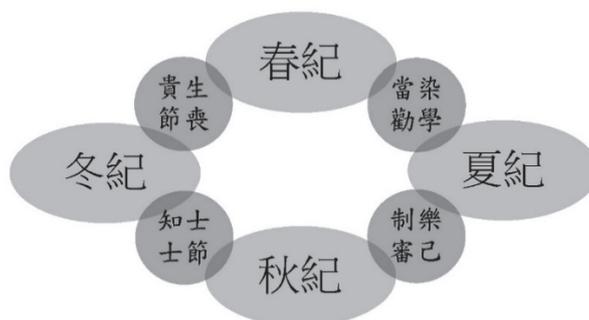
¹²⁵ [秦] 呂不韋著，陳奇猷校釋，〈不侵〉，《呂氏春秋新校釋》，上冊，頁 646。

表十五：異質文本結構功能示意圖



通過環環相扣，書中各紀得以一一連結。這種現象正符合《呂氏春秋》融會諸家學說之旨，四紀主題若是毫無關連，斷為數截，反倒有違成書之意了。最近李貴生以「焊接」形容這種連結方式，並且具體指出書中如何通過生、欲、理、義、死等關鍵詞貫串〈十二紀〉。¹²⁶ 他認為〈春紀〉中〈貴生〉諸篇與〈孟冬紀〉有呼應關係，這種說法可以在下文表十七中找到佐證。¹²⁷ 結合此一分析，表十五的線性形態可以轉化為環狀結構，構成一個首尾呼應的完整系統：

表十六：〈十二紀〉焊接結構示意圖



¹²⁶ 李貴生，〈《呂氏春秋》貴生思想的意涵與詮釋效率：兼論〈十二紀〉的「焊接」結構〉，《臺大中文學報》第 71 期（2020 年 12 月），頁 1-52。

¹²⁷ 下文表十七 C9 中，〈本生〉、〈貴生〉和〈重己〉歸屬為同一類。這三篇文章內容、用語皆甚相似，串聯在一起亦是不難理解。緊接此一階層的卻是隸屬〈孟冬紀〉的〈節喪〉和〈安死〉，可見聚類分析確能有效識別〈節喪〉中生、死、義諸字與〈貴生〉諸文的關係，與學者的質性分析若合符節。

從內容上看，各紀的銜接或許有不夠順暢之處，予人突兀之感，但此書有意構成一個整體，卻是顯而易見的事。無論我們對這種連結方法有甚麼評價，都應先對作此一現象有客觀的認識。

七、結語：數位人文與古代文獻的互動空間

整體而言，四紀中 60 篇文章大體上符合生、長、收、藏之義，但這四個主題並不是機器化地切割開來，互無關連。本文通過聚類分析重新確認了過去說法的合理之處，並且展示了諸紀雖然各有主題，但各紀之中尚夾雜一些異質性文本。這些文本通過相同字詞或內容，與另一紀的篇章連結起來，在承接自身主題之餘，預示或呼應了其他主題。異質性文本既有連接其他主題的功能，當中自然蘊含溢出於該紀的元素，若是輕率地、生硬地把它們套進所屬季節中的生、長、收或藏的單一主題框架，強為之說，便很容易會出現過去學者常犯的牽強比附問題。因此正視這些異質元素的樞紐作用，不但有助說明四紀主題之間的關係，還能增進我們對《呂氏春秋》全書組織的理解。

以上論述主要是在各個樹狀圖的提示下，檢閱相關文本而得出來的結果，足以說明嶄新的數位人文學與傳統的文獻研究實有互相對話、彼此印證的空間。這類印證突破了過去學者檢索電子資料庫的單向性，發揮出人工智能的長處。機器學習法以有別於人類的閱讀模式，揭示文本背後的隱藏結構，學者可以把相關資訊視為對話的對象，複檢現有說法。本文正是通過諸紀的不同組合和聚類分析，回應了近數百年文獻學者關注的〈十二紀〉主題和篇章結構等問題。當然，聚類分析是否適用於先秦兩漢其他文獻，仍須要通過其他實證研究始能得到較為可靠的答案，而這也是值得大家思考和發掘的後續方向。

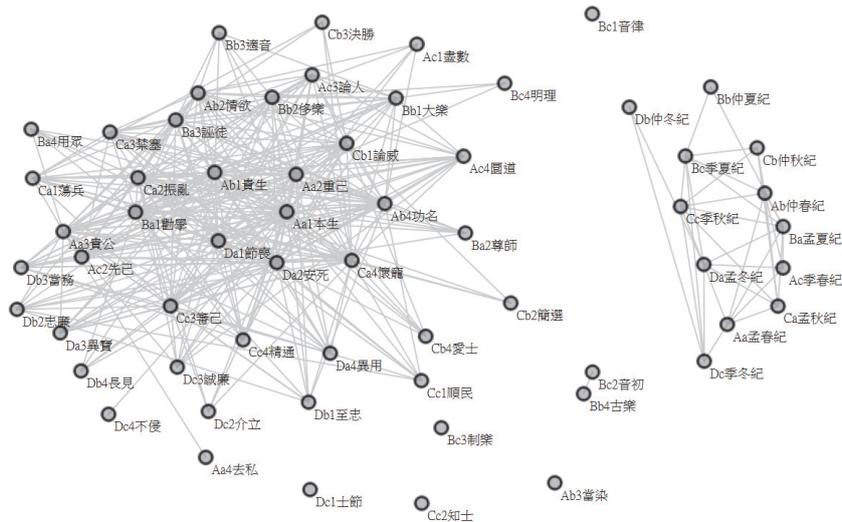
最後想要補充的是，聚類分析的應用範圍並不僅僅限於諸紀的個別分類（表五至表八）或是兩大紀之間的併合考查（表十二至表十

四)，相同的方法和設定還可以不受四紀線性順序的限制，直接比對〈春紀〉和〈秋紀〉或〈冬紀〉之間的關係，作不同搭配，甚至可以用來分析〈十二紀〉全部 60 篇文章，繪製出表十七。¹²⁸ 表十七仍是以紀首自成一類為界線，分成 10 大類。這個大表從另一角度確認了表十五和表十六所描述的連環結構，不過當中提示的資訊量極多，細緻程度已遠遠超越過去學者有關〈十二紀〉篇章關係的各種論述。如 C6 底部〈夏紀〉的〈制樂〉與〈秋紀〉的〈審己〉仍舊連成一小類，但當中卻夾雜了〈冬紀〉的〈長見〉和〈至忠〉，儼然融合了表十三和表十四的相關部分，反映〈審己〉亦有連結〈冬紀〉的作用。當然，這些分類究竟是程式算法的偏見還是高見，仍然有待進一步核實，但不得不承認的是，一般學者很難運用同樣的方式考察 60 篇文章的相似度，更遑論根據細微的數值差異分門別類了。

其實別說繪製相同的圖表，單是消化表十七所提供的資訊已不是容易的事，為了掌握這些複雜的資訊，人類有時須要反過來借助文本探勘中各種降維（dimensionality reduction）工具，如主要成分分析法（Principle Component Analysis，簡稱 PCA）、多維標度（Multidimensional Scaling，簡稱 MDS）等，把相關結果簡化起來。其中 MDS 同樣能夠作相似度結構分析（Similarity Structure Analysis），Orange 軟件亦提供此項功能，用家只須在計算文本的距離後選擇 MDS，便可生成 60 篇文章的散佈圖（Scatter plot）：

¹²⁸ 表十七：〈十二紀〉截枝樹狀圖，請見《清華中文學報》電子期刊網頁，<https://cl.site.nthu.edu.tw/var/file/401/1401/img/704009523.png>。

表十八：〈十二紀〉60 篇相似度結構分析散佈圖



散佈圖能把相似的兩篇文章連成一線，從另一個角度展示 60 篇文章的關係。表十八同樣有許多資訊，不少地方可與表十七的大表互相對照，參詳比較。當然，這類工具已溢出本文所論階層聚類分析的範圍，此處略作演示，旨在重申文本探勘還有不少工具有助揭示文本的隱藏結構，可以提供新的角度考察古代文獻。倘能善加利用，不但能夠發現新問題，亦能夠解決舊問題，實在有很大的發展空間。

（責任校對：郭思捷）

引用書目

一、傳統文獻

- 〔秦〕呂不韋著，陳奇猷，《呂氏春秋新校釋》，上海：上海古籍出版社，2002年。
- 〔漢〕司馬遷著，〔南朝宋〕裴駙集解，〔唐〕司馬貞索隱，〔唐〕張守節正義，《史記》，北京：中華書局，1963年。
- 〔漢〕高誘注，〔清〕畢沅校，《呂氏春秋新校正》。臺北：世界書局，1955年。
- 〔漢〕鄭玄注，〔唐〕孔穎達疏，《禮記正義》，北京：北京大學出版社，2000年。
- 〔魏〕王弼注，〔唐〕孔穎達疏，《周易正義》，北京：北京大學出版社，2000年。
- 〔清〕永瑢等，《四庫全書總目》，上冊，北京：中華書局，1987年。
- 〔清〕汪中著，蔣秋華、林慶彰編審，王清信、葉存芳點校，《汪中集》，臺北：中央研究院中國文哲研究所，2000年。
- 〔清〕徐時棟，《煙嶼樓讀書志》，收入《續修四庫全書》編纂委員會編，《續修四庫全書》，冊1162，上海：上海古籍出版社，2002年。
- 〔清〕孫詒讓著，孫啟治點校，《墨子閒詁》，北京：中華書局，1985年。

二、近人論著

- 王汎森，〈數位人文之可能性及限制——一個歷史學者的觀察〉，收入項潔主編，《數位人文研究與技藝》，臺北：國立臺灣大學出版中心，2014年，頁25-35。
- 王利器，《呂氏春秋注疏》，成都：巴蜀書社，2002年。
- 李貴生，《疏證與析證：清末民初中國文學研究的範式轉移》，北京：

中國社會科學出版社，2016年。

_____，〈《呂氏春秋》貴生思想的意涵與詮釋效度：兼論〈十二紀〉的「焊接」結構〉，《臺大中文學報》第71期，2020年12月，頁1-52。

余嘉錫，《四庫提要辨證》，北京：中華書局，1980年。

官琴、鄧三鴻、王昊，〈中文文本聚類常用停用詞表對比研究〉，《數據分析與知識發現》2017年第3期，2017年3月，頁72-80。

邱詩雯，〈《史詞》作者數位化研究初探——以三十世家虛字字類為例〉，《數位典藏與數位人文》第2期，2018年10月，頁49-69。

祝平次，〈前言〉，《清華中文學報》第22期，2019年12月，頁5-8。

胡佳佳、孟琢，〈話題建模在中國古代典籍分析中的應用〉，《數位典藏與數位人文》第4期，2019年10月，頁27-48。

胡適著，季羨林主編，《胡適全集》，合肥：安徽教育出版社，2003年。

郜建華、樓宇烈，〈《呂氏春秋》中的「精氣說」〉，《華僑大學學報（哲學社會科學版）》2017年第3期，2017年6月，頁40-53。

殷國光，《《呂氏春秋》詞類研究》，北京：商務印書館，2008年。

徐復觀，《兩漢思想史》，臺北：學生書局，1976年。

陳東輝，〈古文獻學研究若干領域的回顧與展望〉，《漢學研究通訊》第3期，2008年8月，頁5-15。

許維遜，《呂氏春秋集釋》，北京：中華書局，2009年。

張雙棣，《《呂氏春秋》詞匯研究（修訂本）》，北京：商務印書館，2008年。

張雙棣、殷國光、陳濤，《呂氏春秋詞典》，濟南：山東教育出版社，1993年。

楊寬，〈月令考〉，《齊魯學報》1941年第2期，1941年7月，頁1-36。

管宗昌，《《呂氏春秋》文學研究》，北京：中國社會科學出版社，2016

年。

劉世閔，〈NVivo 11 Plus 的社交網絡分析在個案研究之運用：以七次國編版國語教科書的負面人物為例〉，《臺灣教育評論月刊》第 1 期，2018 年 7 月，頁 317-335。

劉世閔、吳璟，〈NVivo：新世紀的質性研究電腦輔助軟體〉，《慈濟大學人文社會科學學刊》第 1 期，2002 年 6 月，頁 135-152。

劉苑如、羅珮瑄、邱琬淳、陳雅琳，〈魏晉南北朝筆記小說疾病文本的細讀與遠讀〉，《清華中文學報》第 22 期，2019 年 12 月，頁 49-115。

錢穆著，鄭秀蓮編，《錢賓四先生全集》，冊 5，臺北：聯經出版事業公司，1994。

謝邦昌編著，《文本探勘》，臺北：元華文創公司，2017 年。

譚國根、梁慕靈、黃自鴻編，《數碼時代的中國人文學科研究》，臺北：秀威資訊，2018 年。

Badia, Antonio. "Data, Information, Knowledge: An Information Science Analysis." *Journal of the Association for Information Science and Technology* 65/6 (2014): 1279-1287.

Bird, Steven, Klein, Ewan, and Loper, Edward. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol: O'Reilly Media, 2009.

Hardeniya, Nitin. *NLTK Essentials: Build Cool NLP and Machine Learning Applications Using NLTK and Other Python Libraries*. 1st Edition. Birmingham: Packt Publishing, 2015.

Huang, Anna. "Similarity Measures for Text Document Clustering." Paper presented to Conference Department of Computer Science. Hamilton. 2008.

Kantardzic, Mehmed. *Data Mining: Concepts, Models, Methods, and*

- Algorithms*. 3rd Edition. Hoboken: John Wiley & Sons, Inc., 2020.
- Kastore, Bhupendra, and Sachdeva, Shouvik. "Document Clustering: Similarity Measures." Paper presented to Indian Institute of Technology. Kanpur. 2014.
- Lane, Richard J. *The Big Humanities: Digital Humanities/Digital Laboratories*. London: Routledge Ltd, 2016.
- Mikolov, Tomas, Kai Chen, Corrado, Gregory S. and Dean, Jeffrey. "Efficient Estimation of Word Representations in Vector Space." Paper presented to International Conference on Learning Representations. Scottsdale. 2013.
- Murtagh, Fionn, and Legendre, Pierre. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?." *Journal of Classification*, 31(2014): 274-295.
- Nichols, Ryan, Slingerland, Edward, Nielbo, Kristoffer, Bergeton, Uffe, Logan, Carson, and Kleinman, Scott. "Modeling the Contested Relationship between Analects, Mencius, and Xunzi: Preliminary Evidence from a Machine-learning Approach." *The Journal of Asian Studies*, 77/1(2018): 19-57.
- Nisbet, Robert, Elder, John, and Miner, Gary. *Handbook of Statistical Analysis and Data Mining Applications*. London: Academic Press, 2009.
- Qaiser, Shahzad, and Ali, Ramsha, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents." *International Journal of Computer Applications*, 181/1(2018): 25-29.
- Sarkar, Dipanjan. *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*, 2nd Edition. New York: Apress, 2019.
- Singh, Pranjal, and Sharma, Mohit. "Text Document Clustering and

- Similarity Measures.” Paper presented to Indian Institute of Technology. Kanpur, 2013.
- Steinbach, Michael, Karypis, George, and Kumar, Vipin. “A Comparison of Document Clustering Techniques.” Paper presented to KDD Workshop on Text Mining. Minneapolis. 2000, 1-20.
- Tuckfield, Bradford, and Malik, Alok. *Applied Unsupervised Learning with R: Uncover Hidden Relationships and Patterns with K-means Clustering, Hierarchical Clustering, and PCA*. Birmingham: Packt Publishing, 2019.
- Verene, Donald P. “Vico’s Method of Studies in Our Time.” *New Vico Studies*, 20(2002): 13-18.
- Wang, Li-dong, and Alexander, Cheryl A. “Machine Learning in Big Data.” *International Journal of Mathematical, Engineering and Management Sciences*, 1/2(2016): 52-61.
- Welbers, Kasper, Van Atteveldt, Wouter, and Benoit, Kenneth. “Text Analysis in R.” *Communication Methods and Measures*, 11/4(2017): 245-265.
- Xu, Rui, and Wunsch, Donald C. *Clustering*. New Jersey: IEEE Press, 2009.
- Zinin, Sergey. “Analysis of Character-frequency Lists of Chinese Classics and its Application to Content Analysis and Genre Attribution.” In Kobzev A, ed., *Scholarly Reports of the Department of China of the Institute of Oriental Studies*. Vol. XLVII. Edited by Kobzev A. Moscow: Russian Academy of Sciences, 2017, pp. 254-299.

三、網路資源

中央研究院上古漢語標記語料庫，<http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh?ukey=-644530170&qtype=0>，瀏覽日期：2020年11月27日。

中央研究院數位人文研究平台，<http://dh.ascdc.sinica.edu.tw/member/index.html>，瀏覽日期：2020年7月28日。

中國哲學書電子化計劃，<https://ctext.org/>，瀏覽日期：2020年8月19日。

中國基本古籍庫，http://er07.com/home/pro_3.html，瀏覽日期：2020年7月12日。

中國歷代人物傳記資料庫，<https://projects.iq.harvard.edu/chinesebdb>，瀏覽日期：2020年7月14日。

何晗，《自然語言處理入門》，<https://github.com/NLP-LOVE/Introduction-NLP>，瀏覽日期：2020年8月18日。

通用型古籍數位人文研究平台，<http://ming.ncl.edu.tw/>，瀏覽日期：2020年7月14日。

結巴中文分詞，<https://github.com/fxsjy/jieba>，瀏覽日期：2020年8月19日。

結巴中文斷詞臺灣繁體版本，<https://github.com/APCLab/jieba-tw>，瀏覽日期：2020年8月19日。

愛如生典海數字平臺，http://er07.com/home/pro_3.html，瀏覽日期：2020年7月12日。

漢達古文獻資料庫，<http://www.chant.org.ezproxy.eduhk.hk/Lexicon/SpSearch.aspx>，瀏覽日期：2020年8月19日。

漢籍全文資料庫計畫，<http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm>，瀏覽日期：2020年7月12日。

蔣經國國際學術交流基金會，<http://www.cckf.org/zh/sino>，瀏覽日期

2020年7月12日。

雙語詞彙、學術名詞暨辭書資訊網，<http://terms.naer.edu.tw/detail/1679014/>，瀏覽日期：2020年7月12日。

Eisen, Michael and De Hoon, Michiel. *Cluster 3.0 Manual*，<http://bonsai.hgc.jp/~mdehoon/software/cluster/cluster3.pdf>，瀏覽日期：2020年8月20日。

GitHub, <https://github.com>，瀏覽日期：2020年8月18日。

Nvivo, <https://help-nv.qsrinternational.com/12/win/v12.1.90-d3ea61/Content/vizualizations/how-cluster-analysis-generated.htm>，瀏覽日期：2020年8月20日。

Orange, <https://orange.biolab.si/download/#windows>，瀏覽日期：2020年8月20日。

Orange3 Text Mining, <https://orange3-text.readthedocs.io/en/latest/widgets/bagofwords-widget.html>，瀏覽日期：2020年8月20日。

The Interplay of Digital Humanities and Traditional Documentology: A Clustering Analysis and Structural Exploration of the Twelve Chronicles in the *Lüshi Chunqiu*

Kwai-Sang Lee*

Abstract

Clustering analysis, a method of revealing the hidden structure of texts through unsupervised machine learning algorithms, has been widely used in different types of data mining. In view of the fact that the study of pre-Qin and Han texts has largely remained at the level of one-way searches of databases and paid little attention to the various tools of artificial intelligence, this article attempts to use clustering analysis to re-examine the themes and structure of the “Twelve Chronicles” in the *Lüshi Chunqiu*, which are long-standing and controversial issues; it further seeks to demonstrate the possibility of interplay and dialogue between the digital humanities and ancient texts. We first revisit different views concerning the main themes and structural organizations of the “Twelve Chronicles” over the past few hundred years, and then introduce the concept of cluster analysis as well as the digital tools and parameters used in this article. To illustrate the validity of this approach, a comparison between the clustering results of four of the “Chronicles” and past studies is carried out. We next

* Associate Dean, Graduate School and Associate Professor, Department of Literature and Cultural Studies, The Education University of Hong Kong.

focus on the heterogeneous elements of the texts and analyze their unique connecting functions so as to demonstrate the structural patterns of the “Twelve Chronicles.” The results of this study show that text mining can indeed open up new horizons for the study of ancient Chinese texts. Text mining not only helps to identify new problems, but also makes it possible to solve old ones. Given this, we conclude that there is still much room for further development.

Key words: *Lüshi Chunqiu*, Twelve Chronicles, similarity, clustering analysis, digital humanities