教育科技與學習 19 & 20 (2025)：41-71

研究論文

運用主題模式與廣義線性模式分析教育文本資料之研究

林億雄 *

博士後研究員
國立成功大學永續策略發展處
助理研究員
國立高雄大學教學與發展中心
E-Mail: 11408046@gs.ncku.edu.tw

賴建志

專案經理
國立高雄大學教學與發展中心
E-Mail: jjlai@nuk.edu.tw

黃荏渝

專任助理
國立高雄大學教學與發展中心
E-Mail: ksu0001@nuk.edu.tw

張志鴻

特聘教授
國立高雄大學應用數學系
E-Mail: chchang@nuk.edu.tw

摘要

使用主題模式進行教育文本文字資料分析，是自然語言處理一大重要應用領域。隨著教育資源的數位化，透過有效組織、檢索和分析大量的教育文本資料，成為教育研究者面臨的挑戰。主題模式是一種能夠自動提取文本中核心概念的統計方法，能協助教師與教育管理者做出更具數據支持的決策。近年來，隨著人工智慧與機器學習技術的發

展，主題模式分析在教育領域的應用日益受到關注，特別是數位線上學習平台的使用，及線上學習評估盛行等。本研究透過爬蟲程式捉取及分析線上討論平台大學生學習評價相關資料，透過主題模式自動化分析技術減少人工標註的時間成本，提高資料處理效率。潛藏狄利克雷分配（LDA）為一種常用的簡易主題模式，已被廣泛應用於文本分類、情感分析和知識管理等領域。本研究探討簡易主題模式 LDA 在教育文本分析中的應用，並透過對教育文本資料進行主題建模，透過簡易主題模式識別學生回饋、課程評價及學科內容中的主要主題。同時，簡易主題模式亦可與廣義線性模型結合使用，結合文本主題分析與量化學生學習結果之間的關聯性。研究結果顯示，LDA 能夠有效擷取出教育文本的核心內容，並能與廣義線性模型結合使用，更能為教育決策者提供有價值的資訊。最後，本文於附錄處提供本研究 LDA 所使用分析中英文文本資料之 Python 語言程式碼，及廣義線性模式使用之 R 統計軟體程式碼，以利有興趣之讀者可加以使用。

關鍵詞：主題模式、教育文本資料、潛藏狄利克雷分配、廣義線性模式

* 為本文通訊作者

壹、緒論

在當代教育資源日益數位化的趨勢下，如何有效組織、檢索與分析龐雜的教育文本資料，已成為教育研究與實務發展所面臨的重要挑戰。主題模式（Topic Modeling）作為自然語言處理（Natural Language Processing, NLP）的一項核心技術，能夠自動從非結構化文本中抽取潛在語義結構，對於提升教育文本分析的效率與精準度具有關鍵意義。近年來，隨著人工智慧與機器學習技術的迅速發展，主題模式在教育領域中的應用逐漸擴展，尤其是在數位學習平台、線上課程評估及學生回饋分析等方面，展現出高度潛力。其中，潛藏狄利克雷分配（Latent Dirichlet Allocation, LDA）作為一種具代表性的統計式主題建模技術，已被廣泛應用於文本分類、情感分析、知識管理等多元領域。LDA 分析法可透過機率建模方式自動識別文本中潛藏的主題結構，進而促進教育研究的量化分析與決策輔助功能。本研究即以 LDA 為核心分析工具，探討其在教育文本資料分析中的應用效能與理論意涵，尤其聚焦於學生學習歷程中所生成之文本（如學生於公開社群平台紀錄、學習日誌、課程評價與學習回饋等）之自動化分析，藉此挖掘學生在學習過程中的策略、動機與潛在困境。

隨著數位科技日新月異，教育現場的教學模式與學習環境亦正歷經劇烈轉型。各類數位平台（如線上教學系統、學生回饋機制、開放式線上課程 MOOCs 等）大量生成非結構化教育文本資料，使得教師與教育決策者對於資料處理與理解的需求日益提升。然而，該類資料多具結構鬆散、主題分散之特性，傳統人工分析不僅耗時費力，亦難以全面掌握資料中潛藏的趨勢與訊息。在此背景下，主題模式技術提供了一種自動化、客觀且可擴展的文本分析方案，能有效輔助教育現場進行教學內容調整、課程設計優化與學習歷程理解。例如語言學習領域正面臨顯著變革，特別是以學習者為核心的自主學習理念逐漸取代傳統教師主導模式，其學習型態亦因應科技發展而從課堂轉向網路、行動學習與混成學習等多元場域。自主學習強調學習者於目標設定、學習規劃、過程監控與成效評估中扮演主動角色。然而，如 Little（1991）所言，學習者並非天生具備自主能力，特別是在非母語環境中，考試導向文化常使學生依賴教師指導而欠缺主動性。Dias（2000）亦指出，自主學習理念與傳統文化可能產生價值衝突，進而影響學習效果。因此，深入瞭解學習者之動機、策略與實際表現之間的互動關係，已成為語言學習研究的重要課題。目前，臺灣針對語言自主學習的研究多著重於學習平台或工具的開發應用，然而對於學習動機與策略如何影響學習成效，特別是在學習者生成之文本資料中的實徵探究，仍屬不足。有鑑於此，本研究擬結合 LDA 主題模式與自主學習理論，針對大學生自主學習歷程中產出的文本資料進行

系統性分析，探討其學習策略與動機及其可能如何影響學習成效，並藉此建構一套兼具理論深度與實務應用價值之教育文本分析框架。

本研究之主要目的包括：（1）驗證 LDA 模型於教育文本分析中的適用性與準確性；（2）透過主題建模技術對學生學習日誌、課程回饋等文本進行主題分類與語意萃取，辨識學生關注的學習議題與潛在學習困難；（3）探索大學生英文自主學習中學習策略、動機與學習成效之間的相互關係；（4）建構一套結合自然語言處理與教育理論的研究架構，支援個別化教學決策與教學資源配置的智慧化發展。為達成上述目標，本研究以公開社群網站中大學生對課程的評價、修課心得與學習紀錄作為研究文本，並透過網頁爬蟲技術進行資料擷取。資料經過前處理（如斷詞、去除停用詞與標準化處理）後，使用 LDA 進行主題建模與文本分析。藉由具體案例實證，檢驗 LDA 在教育文本資料處理中的效能與適切性。

整體而言，本研究的學術貢獻可從以下層面加以闡述：（1）方法層面上，提出一套基於 LDA 的教育文本分析流程，適用於大規模非結構化資料之自動化處理；（2）應用層面上，協助教育人員掌握學生真實需求與反饋，提升教學回饋機制的實用性與效率；（3）實踐層面上，深化對學生學習歷程的理解，提供具體教學與輔導策略之依據；（4）理論層面上，回應自主學習與數位教育融合的研究趨勢，拓展主題模式於教育研究領域的應用範疇。本文架構如下：第二節為文獻探討，概述主題模式技術與教育文本分析相關研究；第三節進行資料蒐集與分析方法之說明，實作 LDA 於教育文本中的應用案例；第四節則針對研究發現進行綜合討論，並提出對教育文本資料分析之未來建議與發展方向。

貳、文獻探討

主題模式（Topic Modeling）為一種常見的非監督式機器學習技術，主要應用於大量非結構化文本資料中潛藏主題的自動識別與分類。在眾多主題模型中，潛藏狄利克雷分配（Latent Di- richlet Allocation, LDA）因其理論基礎簡潔與實作穩定性，廣泛應用於教育文本分析。LDA 基於機率生成模型，假設每篇文件為多個主題的混合，而每個主題則由一組關鍵詞以不同機率分佈構成。透過貝氏統計推論與吉布斯抽樣（Gibbs Sampling）演算法，LDA 能有效推估文本中的潛在主題分佈，進而揭示文本語意結構與詞彙間的隱含關聯。

此外，廣義線性模型（Generalized Linear Model, GLM）作為一種延伸傳統線性回歸的統計分析工具，亦在教育研究中扮演關鍵角色。GLM 具備

處理不同類型應變數（如二元、類別與計數資料）之彈性，常被用於量化潛在變項與教育結果之間的關聯。將主題模式與 GLM 技術結合，研究者可先藉由 LDA 從學生學習日誌、課程回饋或訪談紀錄中萃取核心主題，再將潛在主題作為 GLM 之自變項，以預測學生學習成效或休退學風險。例如，若 LDA 分析顯示學生頻繁提及「課程難度」、「學習焦慮」與「資源不足」等主題，GLM 便可進一步檢驗此等主題是否顯著預測其求助行為次數或是否有休學傾向。此一整合架構不僅提升文本分析的量化深度，亦可用以整合教育文本資料質性與量化之研究，為教育機構提供可操作的數據決策依據。

在高等教育脈絡中，學生休退學問題長期被視為影響教育品質與高教機構永續經營的關鍵議題。既有研究指出，學生的持續就學行為受多重因素影響，包括學業適應、社交整合、經濟壓力與心理健康等。透過 LDA 對學生歷程檔案、輔導紀錄、問卷開放題與學術交流平台資料進行主題建模，可自動化發掘潛藏於文本中的關鍵議題，如學生對就讀科系的認同程度、學習動機是否內化、入學動機是否與課程期待一致等。這些結果有助於辨識潛在高風險學生，並依據其主題特徵提供客製化支持，如強化學習策略指導、開設適性輔導服務或提供經濟協助，以降低休退學率並提升整體學習成效。

在教學法層面，專題導向學習（Project-Based Learning, PBL）作為一種以學習者為中心的建構式教學法，強調學生透過實作專案主動建構知識。根據邱貴發（1996）與 Thomas（2000）之定義，PBL 融合建構主義、合作學習與情境學習理念，教師角色由知識傳遞者轉化為學習促進者，學生則須在具體情境中完成一項具產出導向的任務。PBL 允許學生依其興趣與學習目標自由選擇專題，並在分組與任務實作中發展主體性與解決問題能力。此類學習模式能有效提升學生學習動機與學習成效，尤適合高等教育中強調批判思維與創造力培養之課程設計。

在學習策略自我認同設計上，本研究進一步援引 James Marcia（1991）所提出之自我認同發展理論，以探究學生在專題學習過程中認同動態對其學習參與之影響，如圖 1 所示。Marcia 將青年認同分為四種狀態，根據其在「承諾」（Commitment）與「探索」（Exploration）兩構面上的表現加以區分，包括：早閉型認同（Foreclosure）、未定型認同（Moratorium）、定向型認同（Identity achievement）與迷失型認同（Identity diffusion）。在自由分組的專題課程設計下，學生可藉此理論架構反思自身與組員的學習風格與動機傾向，進而促進組內合作與認同建構歷程。例如，對早閉型認同學生而言，經由實際任務參與與自我反思，可能觸發其對現有價值與目標的再檢視，並推進至更高階的認同狀態。此種理論應用不僅有助於分組策略的優化設計，也提供教師評估學生心理動態與學習適應的有效工具。

綜上所述，整合主題模式、統計建模與認同理論的研究設計，為理解學生學習歷程與提升教育支持系統提供嶄新視角。藉由技術與理論的交叉應用建構一套具高度實證性與應用價值的教育資料分析框架，期能深化對學生學習行為之理解，並提供教育決策者以數據驅動教育介入策略與政策建議。

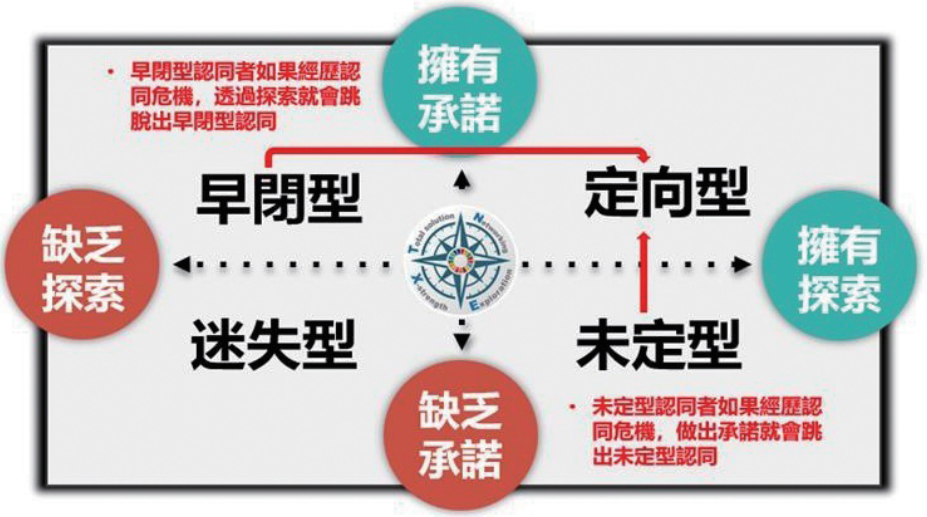

圖 1 Marcia 四種認同動態

此外，專題導向學習通常跨足不同學科領域，本研究亦鼓勵學生將多學科知識應用於專題研究中，以解決複雜的問題。學生在專題研究中通常會遇到具體的問題或挑戰，學生需要使用批判性思考、研究和分析技能來解決這些問題，這有助於培養問題解決能力。進行專題導向學習方法，學生需要設定目標、制定學習計劃、研究相關文獻，並進行研究。專題導向學習也涉及到實際的項目或研究，學生需要將所學應用到實際情境中，這有助於將理論知識轉化為實際技能。在師生互動方面，老師在專題導向學習中扮演導師的角色，提供指導、支持和評估。專題導向學習教學方式強調師生之間的互動和合作，其有助於培養學生的研究能力、批判性思考、問題解決和自主學習技能。Hadgraft 等人（2003）認為「專題導向適合使用於數學、物理或工程學等領域。由於這些學科是具有階層知識結構的，如果缺少一個基本的部分，會導致學習其他概念的失敗，必須依照一定的學習順序，因此在知識與技能的提供上，需要教學者較多的主導」。綜合上述，可知專題導向學習的特色有：（1）以學習者為中心；（2）設計學生在真實世界的情境中學習；（3）強調實作評量；（4）學生需主動設計與自己興趣相符的作品、參與問題解決，並持續增進自己的表現；（5）適合使用於具階層知識結構的學科領域。

故此，主題模式與廣義線性模型在高等教育研究中具有重要應用價值，特別是在探討學生休退學議題時，能夠有效揭示影響因素並進行預測分析。透過結合文本挖掘與統計建模技術，研究人員可更全面地理解學生學習歷程與學業持續性之間的關係，進而為政策制定者與教育管理者提供具體建議，以降低休退學率並提升教育品質。

一、教育文本分析相關研究

美國知名教育心理學家布魯姆（Benjamin Bloom）其重要教育貢獻包含有教育目標分類學和掌握學習理論。布魯姆將教育目標由低階至高階分成：「記憶、理解、應用、分析、評估、創造」。記憶與理解是臺灣教育花最多時間培育的能力，應用與分析近年來漸受重視。然而，記憶很容易被電腦或機器「一鍵搜尋」替代；理解方面，機器在邏輯分析推算方面能力遠超越人類。然若一位大學畢業生只擁有最低階的記憶與理解能力，則很容易在職場上被淘汰或取代。大學教師想拉高教育目標的層次，勢必要突破教學方法。

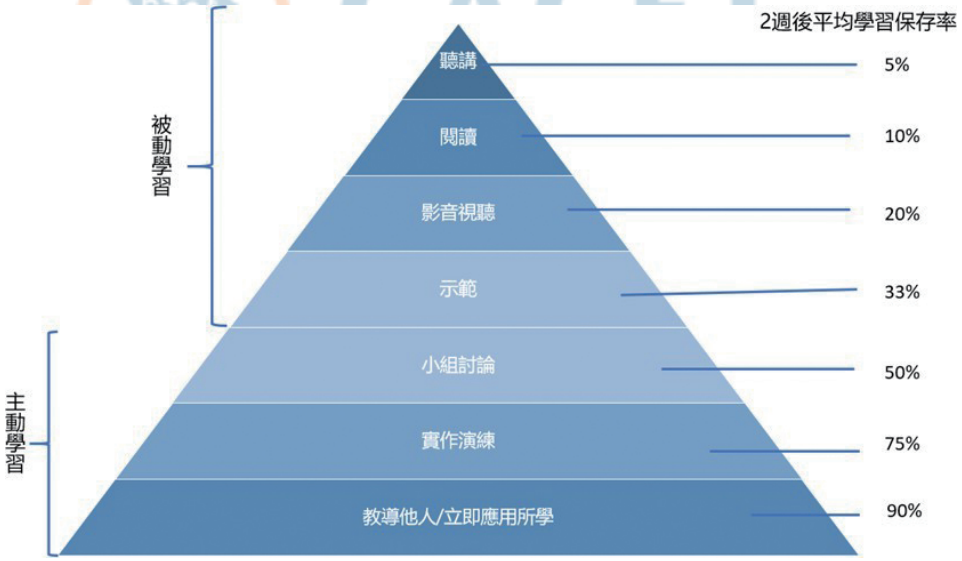

圖 2 學習金字塔 - Cone of learning by Edgar Dale (1969)

學習金字塔是學習上頗為人知的理論主要討論「學習方法」與「記憶保留率」。此理論模型在教學上非常實用，學習金字塔在學習的方式上做了很細緻的分類拆解，從純聽講到主動分享教學。學習金字塔說明在不同學習方式下，學習者在兩週後還能記住的內容有多少，由美國教育學家

Edgar Dale 於 1969 年提出，並由美國國家實驗室驗證的研究結果，如圖 2 所示。美國國家實驗室實驗發現，利用單向講述，學生的學習效率只有 5%，至於閱讀能保留 10%、視聽有 20%、示範是 33%、小組討論 50%、實作 75%、教導別人則高達 90%。此結果顯示學生參與越深，學習效果越好。研究結論說明主動學習的效率大於被動學習的效率，學習新知識後，找機會傳授給他人的效率是最高的。學習金字塔顯示，學生能主動思考、互動討論、實際操作，純熟度就越高。應用專題導向學習教學方法也正是投注在這些能力上，學生必須將聆聽講述、自主閱讀、觀看影片的任務視為課程準備，透過專題導向學習練習示範、討論、實作、教導別人，循序漸進培養分析、評估、創造的高階教育目標技能。

知名教育學家約翰．杜威（John Dewey）認為「最有效的學習是從解決問題中獲得，而解決問題須經過反省思考的歷程」。杜威認為學校課程也應該以實際問題為發展基礎，他提出了一些條件，用來界定有效的問題，以促進學習和思考的發展。根據杜威的教育理論，問題應具備（1）合理性：問題應當是真實且具體的，而不僅僅是抽象的學術練習。它應當在學生的生活和經驗中有實際應用，使學生能夠看到問題的重要性和現實意義。（2）需求感：問題應當引發學生的需求感，激發他們的好奇心和動機，以解決問題或尋找答案。這種需求感應當是個人的，而不僅僅是為了滿足課程或老師的要求。（3）調查：問題應當引導學生進行調查和探究，以尋找答案和解決問題。這要求學生參與主動的學習，培養他們的批判性思維和問題解決能力。（4）解決性：問題應當促使學生參與解決問題的過程，而不僅僅是提出問題。通過思考、實驗和反思，學生可以發展解決問題的能力。（5）社會性：問題應當具有社會性，意味著它應當能夠激發學生進行合作和討論，促進知識的共建。杜威認為，社交互動是學習的重要組成部分。根據杜威的教育理論，問題應當具備上述這些條件，以確保學生們在教育過程中能夠達到其潛在的教育效益，並促進學生的思考和學習。這種方法被稱為「問題導向學習」，強調學生參與、探究和解決實際問題的重要性。這種教學法提供給學生的不僅僅是知識，而是一種終身難忘的經歷，是學生把教學內容與理解過程相結合的經歷。後續，也有大量學者的研究和實踐，證實了問題解決教學法的重要教育意義。

以解決問題為目標的教學模式可分為問題導向與專題導向，二者均強調在真實世界的環境中學習，以學生成效為依歸。兩者相異之處在於：專題導向學習是較為行動傾向的，學生需主動設計與自己興趣相符的作品、並持續增進自己的表現；問題導向學習則要求學生專注在一個待解決的問題上。Bouhuijs 等人（2000）認為「專題導向學習比問題導向學習更需要

知識的應用」。授課教師若能在專題進行中，能隨時監督學生狀況，並給予其支持、引導其針對眼前問題作分析，搜尋解決的方法，將能有助於學生在專題導向課程中，習得經驗與問題解決技巧，運用在往後的學習與日常生活之中。

Polya 發展的問題解決模式：了解問題、想出計畫、執行計畫、回顧。程式設計是運用程式語言撰寫程式以解決問題。程式設計包括四個主要的步驟：了解問題的需求、擬定解題的計畫、撰寫程式碼，以及測試與除錯，與問題解決模式類似。學習者能依循這樣的模式，有系統地思考、規劃、執行與檢視結果來解決問題，方能增進自己的概念認知。程式設計搭配遊戲設計方面，張玟慧等人（2016）、Kiili（2007）及 Ahlers 等人（2002）認為「遊戲是由許多小問題，隨意組合而成的大問題，他們的研究指出學習者在遊戲中所經歷的學習歷程，發現其在遊戲時所經歷的形成遊戲策略、實驗遊戲策略、反思過程，能幫助學習者將理論性的知識實際應用，並達到更高的理解與擴增領域知識」。綜合以上討論，本研究發現解決問題的模型與程式設計的解題歷程類似，因此本研究期望能透過學生於公開平台上發表對於學習心得來探討提升其解決問題的能力。

在課程內容分類上，根據主題模式對教育內容進行自動分類，提高檢索與組織效率。同時，在學生回饋分析與學習趨勢分析上，運用 LDA 模型可識別學生對教師授課方式、教材難易度的主要意見，及可追蹤不同時期的學習趨勢，提供教學改進的依據。在過去的研究中，Griffiths 和 Steyvers（2004） 提出如何利用 LDA 來識別科學論文的主要主題，此技術也逐漸被應用於教育研究（Blei et al., 2003）。同時，Newman 等人（2010） 亦探討評估主題模型的品質，確保主題分類結果的可解釋性與準確性。在教育領域，主題模式已被廣泛應用於分析學生的學習日誌、課程評價、教學反饋等文本資料。例如，研究者可以利用 LDA 分析學生在線上論壇中的發言，識別學生在學習過程中遇到的主要問題；也可以分析課程評價中的文本反饋，了解學生對課程內容和教學方法的意見。近年來，隨著深度學習技術的發展，BERT 等預訓練語言模型也被應用於教育文本分析，進一步提升了文本理解和主題建模的效能（Devlin et al., 2019）。此大語言模型（Large Language Model, LLM）能夠捕捉文本語義資訊，提高主題模型準確性和可解釋性。相關研究包括：主題模型在教育內容推薦應用（Liu et al., 2016）；主題模型在學生情感分析應用（Wang et al., 2017）；主題模型應用於線上學習社群資料之分析（Cheng et al., 2014）。

二、學生休退學與相關因素探討

在高等教育研究領域中，學生休學與退學問題長期以來受到高度關注。此現象不僅攸關學生個人學習歷程與生涯發展，亦對高等教育機構的教學品質、資源配置與整體辦學績效產生深遠影響（Tinto, 1993）。Tinto 所提出的「學生離校模型（Student Departure Model）」指出，學業整合與社會整合為影響學生持續就學的重要因素。當學生在學業表現或人際互動層面無法獲得足夠支持與認同感時，較容易產生中斷學業的傾向。進一步而言，Bean 及 Metzner（1985）擴展 Tinto 模型，納入非傳統學生的休退學行為，強調外部環境變項對離校決策的關鍵作用，包括經濟壓力、家庭責任、工作需求等。相關研究亦指出，特別在亞洲地區，家庭期望與財務壓力為學生教育持續性的重要影響因素（Chen, 2012）。此外，學生對就讀學系的興趣不符、課業負荷過重以及對未來就業前景的疑慮，也常被視為導致休退學的顯著原因（Bean & Metzner, 1985）。

從制度面觀之，高教機構可透過課程設計調整、輔導制度強化與學制彈性化等方式，作為降低休退學率的因應策略（Astin, 1999）。例如，學業輔導、心理諮商、生涯發展諮詢與學習支持服務，皆有助於強化學生的歸屬感與學習動機，進而促進其學業持續性（Kuh et al., 2005）。總結而言，學生休退學現象係由多重因素交互影響而成，包括個人特質、家庭背景、學校制度與社會環境等。透過系統性研究，可更深入理解休退學行為的動態歷程，並據以提出精準有效的預防與介入策略，促進高等教育體系的永續發展。

在臺灣相關研究方面，針對學生學習困擾探討相對有限，且多聚焦於學習困擾與學習態度間的關聯，並以特定學科為分析對象。王天助（2011）、王淑娟（2012）、林國暉（2013）與賈旻暟等人（2016）研究指出，性別與學習困擾存在顯著差異；然而，蔡惠雅（2013）則發現性別在學習困擾上並無顯著影響。類似的分歧也出現在年級與學業成績之間的關聯上，反映學習困擾在不同背景變項下可能呈現多樣性，亦顯示其研究尚有深化與細緻化之必要。黃昌誠（1990）以學習行為與學習困擾問卷調查大學生，發現性別在學習行為上具有差異：男生在學習技巧與讀書習慣上優於女生，而女生則在注意力集中與學習態度方面表現較佳。綜上所述，學生的學習困擾與休退學行為密切相關，兩者皆受到多重背景變項的影響。未來研究宜整合質性與量化方法，深入剖析學習困擾背後的心理歷程與結構性脈絡，並將之納入高等教育中介入機制的設計與評估之中，以提升學習支持系統之實證基礎與成效。

參、研究方法

在進行主題模式分析前，需要先進行數據整合與預處理。研究人員預先整理教育文本資料，包括：（1）文本清理：去除標點符號、特殊字符，及轉換大小寫等；（2）去除停用詞：刪除常見但沒有語義價值的詞，如「你」、「我」、「的」、「是」、「和」等；（3）詞彙標準化：使用詞幹提取（Stemming）或詞形還原（Lemmatization）技術，將詞彙轉換為其基本形式。例如「學習」、「學過」將透過轉換為「學習」；（4）向量化表示：使用詞袋模型（Bag of Words）與詞嵌入（Word Embedding）技術，將文本轉換為數值形式，方便機器學習模型處理。本研究以公開社群平台大學學生對於學校課程評價資料集為例，使用 LDA 分析學生對不同課程學習的反饋，並從中發現文本資料主要評價之主題。本研究數據為透過網頁爬蟲程式，擷取網路線上討論平台之大學學生對課程學習評價、修課心得，及大學生活記錄資料等。資料蒐集經處理後共 8,186 筆（截至 114 年 4 月），資料數據為來自不同教育單位大學學生對課程、教師，及對特定課程的反饋。

一、LDA 模型建構與教育文本資料分析

本研究使用 Gensim 函數工具配置 LDA 模型，關於模型建構與分析的步驟：（1）確定主題數量：選擇最佳主題數量，主題一致性越高，代表模型效果越好；（2）訓練 LDA 模型：將預處理後的文本數據輸入 LDA 模型，進行訓練。使用吉布斯抽樣（Gibbs Sampling）優化主題和詞語的概率分佈；（3）主題可視化與解釋：使用 pyLDAvis 庫將 LDA 模型的可視化，展示主題分佈和詞語關聯。分析每個主題的關鍵詞，解釋主題的實際含義。本研究以爬蟲程式擷取公開性質之社群網路大學學生對於課程學習評價、修課心得，及大學生活記錄資料進行實際分析。該資料集包含學生對不同課程的文本評價，例如：「這門課的老師講解很清晰，內容也很實用」、「作業有點多，但是對鞏固知識很有幫助」、「課程的互動性很強，同學們都很積極參與」、「教材的難度有點高，建議增加一些案例分析」等。本研究使用 Python 函數工具 Gensim 訓練 LDA 模型，由研究者審視主題內容教育意涵並設定主題數量為 5。經模型訓練完成後，由表 1 可知以下 5 個主題，及該分類主題前 3 強之關鍵詞：

表 1 主題模式分析後之 5 個主題分類及其前 3 個關鍵詞

編號	主題分類	該分類主題之前 3 個關鍵詞
1	教學品質	講解、清晰、掌握
2	課程負擔	作業、鞏固、知識
3	互動性	互動、參與、積極
4	教材難度	教材、難度、分析
5	休退學	學校期待、就讀領域、入學方式

本研究透過 Python 語言 pyLDAvis 函數工具將 LDA 模型可視化，如圖 3 展示其主題分佈和關鍵詞之關聯（Python 語言執行 LDA 模型程式碼，請見附錄一、二）。分析結果顯示，學生對教育內容的評價主題為：教學品質、課程負擔、互動性、教材難度，及休退學等 5 個構面。其中，例如：主題 5「休退學」的關鍵詞包括「學校期待」、「就讀領域」，及「入學方式」，此分析結果說明學生是否休退學與學校是否符合期待、對於就讀領域是否有興趣，及不同入學方式等有顯著關係。故，教學機構可根據這些主題，了解學生對課程學習、學習規劃等的主要評價，並針對性地進行輔導介入、行政支援與課程改善等。本文於附錄處，提供本研究於中英文文本進行主題模式分析之 Python 語言程式原始碼（請見附錄一、二），以利有興趣讀者可加以使用。

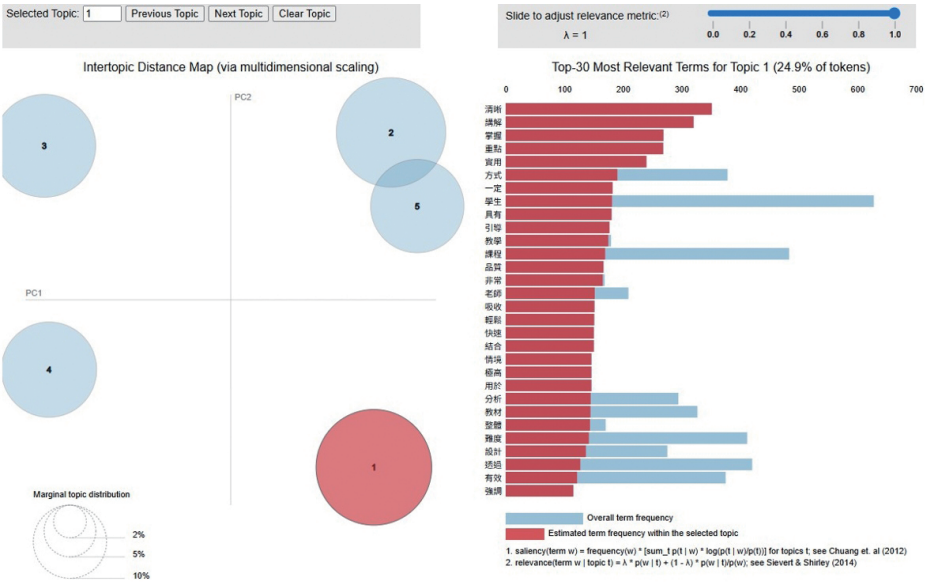

圖 3 主題模式之主題分布和關鍵詞之關聯可視圖

二、學生休退學與相關因素之關聯分析——對數線性模式之應用

學生休退學是高等教育研究中一個重要議題，其影響個人學習歷程及教育資源運用效率。根據本文 LDA 研究發現，學校是否符合學生期待、學生對就讀領域的滿意度，及入學方式可能與其學習持續性密切相關。因此，本研究運用廣義線性模式中的對數線性模式（Log-linear Model）來分析這四個類別變數之間的關聯，並透過建構適切的統計模型探討該變數間的交互作用，以理解影響學生休退學的關鍵因素。本研究以對數線性模式為主要分析方法，該模式適用於列聯表資料，透過對觀察次數取對數來解釋類別變數間的交互作用，並檢驗獨立性假設。在休退學資料方面，以某教學機構 112 學年度之學生休退學數據為研究資料，並以 I 代表學生是否休退學、G 代表學校是否符合學生期待、L 表學生滿意就讀領域，及 S 表該生是否為考試入學。在建模方面，以 R 統計軟體搭配 lmtest 函數庫構建完全獨立模型，假設四個變數彼此獨立，無交互作用；接續建立任兩變項間之交互作用模型，並考慮變數兩兩間的影響，如（I, G）、（I, L） 及（I, S） 等，以探討學生休退學與各因素的關聯性。同時，進一步建立高階交互作用模型，檢視三變數或四變數聯合作用，例如（I, G, L）或（I, G, L, S）的交互影響；此外，本研究透過擬合完整飽和模型，包含所有可能之交互作用，以描述變數間關係。最後，為了確保模型的適切性與穩健性，本研究將以概似比檢定（Likelihood Ratios Test, LR）進行模型選擇，LR 檢定是評估模型效能優劣的指標可避免模型過度擬合，並找出最佳的解釋模型。此外，本研究亦針對變數間之條件獨立性進行假設檢定，以確定是否存在重要的交互作用。例如，若（I, G）交互作用顯著，則代表學校是否符合學生期待可能顯著影響學生休退學的機率；若（I, L, S）三元交互作用顯著，則顯示滿意就讀領域與考試入學的交互影響可能共同作用於休退學決策。透過對數線性模式分析，研究人員可進一步理解學校期待、學科興趣及入學方式對學生學習持續性是否休退學的影響，進而為高等教育體系提供優化招生策略及學生輔導機制的參考依據，降低學生之休退學率，提升教育資源的有效運用，並據此提供學術界與政策制定者更深入的實證基礎。本研究所使用廣義線性模式之對數線性模型，其圖形結構透過適合度檢驗後，選用模型為（GI, GL, IL, IS），模型結構如圖 4 所述（R 統計軟體執行廣義線性模式程式碼，請見附錄三）。

LR 檢定 ： 模型（GI, GL, IL, IS）與 飽和模型為（G*I*L*S）比較
Pr （ > Chisq）= 0.2047 ；（Chisq = 9.7254；DF = 7）

由上式可知，P 值為 0.2047 顯著大於 0.05，故模型（GI, GL, IL, IS）可視為一合適模型。

$$\log \mu_{ijkm} = \lambda + \lambda_i^I + \lambda_j^G + \lambda_k^L + \lambda_m^S + \lambda_{ij}^{IG} + \lambda_{ik}^{IL} + \lambda_{im}^{IS} + \lambda_{jk}^{GL}$$

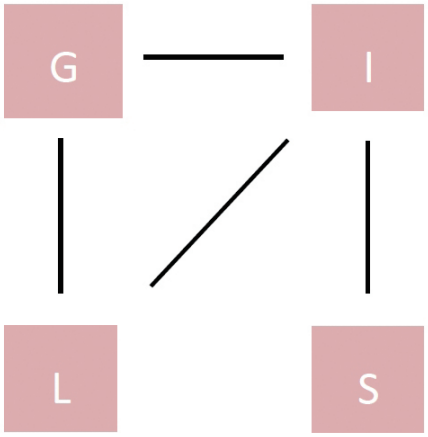

圖 4 學生休退學（I）、學校符合期待（G）、滿意就讀領域（L），及考試入學（S）之關聯圖

表 2 學生休退學（I）、學校符合期待（G）、滿意就讀領域（L），及考試入學（S）列連表及配適值

休退學	學校符合期待	滿意就讀領域	考試入學	對數線性模型					
				GILS 飽和模式	GI, GL, GS, IL, IS, LS	GI, GL, GS, IL, IS	GI, GL, IL, IS, LS	GI, GL, IL, IS	GL, IL, IS, LS
有	是	滿意	考試	7	6.74	6.59	6.72	6.52	7.42
有	是	滿意	非考試	5	3.95	4.09	3.96	4.17	4.38
有	是	不滿意	考試	6	6.82	6.98	6.79	6.90	8.74
有	是	不滿意	非考試	4	4.50	4.34	4.53	4.42	5.83
有	否	滿意	考試	1	2.07	1.99	2.09	2.02	1.38
有	否	滿意	非考試	1	1.24	1.32	1.23	1.29	0.82
有	否	不滿意	考試	11	9.37	9.44	9.41	9.56	7.45

休退學	學校符合期待	滿意就讀領域	考試入學	對數線性模型					
				GILS飽和模式	GI, GL, GS, IL, IS, LS	GI, GL, GS, IL, IS	GI, GL, IL, IS, LS	GI, GL, IL, IS	GL, IL, IS, LS
有	否	不滿意	非考試	6	6.31	6.25	6.28	6.12	4.97
無	是	滿意	考試	207	215.14	211.30	214.76	209.39	214.23
無	是	滿意	非考試	242	235.17	239.01	235.56	240.93	234.97
無	是	不滿意	考試	92	83.30	87.13	82.85	86.34	81.39
無	是	不滿意	非考試	95	102.38	98.56	102.84	99.34	101.03
無	否	滿意	考試	48	39.05	37.50	39.43	38.45	39.96
無	否	滿意	非考試	36	43.63	45.19	43.25	44.24	43.83
無	否	不滿意	考試	58	67.51	69.07	67.96	70.82	69.41
無	否	不滿意	非考試	93	84.81	83.24	84.36	81.49	86.16

由上表 2 之對數線性模型分析結果，可知不滿意目前就讀領域是學生選擇休退學的關鍵因素。

同時，不論「學校期待」，及「入學方式」狀態，本研究計算學生休退學風險之比例如下：

不滿意目前就讀領域學生較滿意目前就讀領域學生採取休退學的比例

= （209.39 * 6.90）/（6.52 * 86.34）

= （240.93 * 4.42）/（4.17 * 99.34）

= （ 38.45* 9.56）/（2.02 * 70.82）

= （ 44.24*6.12）/（1.29* 81.49）

= 2.57.

由此可知，不論學生認為學校是否符合期待，及該生是否為考試入學，在休退學比例上，不滿意目前就讀領域的學生較滿意目前就讀領域學生採取休退學的比例高達 2.57 倍。同時，在休退學比例上，認為學校不符合期待的學生較認為學校符合期待學生採取休退學的比例高達 1.69 倍；在休退學比例上，考試入學的學生較非考試入學學生採取休退學的比例為 1.80 倍。

肆、結論

本研究致力於探討 LDA 與 GLM 模型於教育文本資料分析上的整合應用，並針對學生回饋、課程評價與學科內容進行主題建模與關聯性之探討。結果顯示，LDA 在辨識教育文本潛藏主題方面具有高度的效能，不僅能自動提取出具有意義的主題結構，亦能藉由統計建模方法進一步理解這些主題與學習成效之間的關聯性。綜合來看，本研究具體貢獻可歸納為以下幾個面向：

一、驗證 LDA 於教育文本分析中的適用性與實用性。在教育領域中，文本資料無所不在，涵蓋教學回饋、學習歷程記錄、問卷開放式作答、教師教案及學術交流等非結構化文本資訊。相較於傳統量化研究方法，LDA 提供一種有效轉化與解釋大規模文本資料的手段，使研究人員能更有系統地理解資料中的潛在語意結構。本研究透過實際教育文本案例（如學生課程意見調查），驗證 LDA 不僅能有效辨識出具體且具教育意義的主題，例如「教師互動」、「教材內容難度」與「課程實用性」，更可視覺化呈現主題與關鍵詞的機率分布，提升文本資訊的詮釋深度與應用價值。

二、建立主題模式與課程評價分析之應用架構。LDA 模型不僅能夠處理大量學生課程回饋文字資料，更能將非結構化資料轉化為結構化主題向量，進而作為後續統計分析或預測模型的輸入變數。本研究展示如何將 LDA 萃取出的主題用於理解學生對課程內容的偏好與批評，例如：將「教師教學熱忱」、「課堂互動機會」、「課程設計完整性」等主題作為指標，協助教育機構有系統地識別教學強項與待改善處，從而提出具針對性的課程優化策略。此一方法突破以往僅以量化滿意度評分為依據之限制，實現更具語意深度與動態追蹤能力的評估模式。

三、揭示學科知識之潛藏結構與跨主題關聯。透過對學科內容文本的主題建模分析， LDA 能揭示出學科內部知識結構性與主題間邏輯關聯，為課程模組設計、教材編排與學習地圖繪製提供理論依據。

四、整合 GLM 提升預測與推論力。為進一步量化 LDA 所建構主題與學習成效或學生行為間的統計關係，本研究結合廣義線性模型進行實證分析。結果顯示，當主題向量作為 GLM 的自變項時，能有效預測學生課程滿意度、學業表現，甚至其休退學意向。例如，由模型中發現學生在「學校是否符合期待」及「是否滿意目前就讀領域」與休學風險有顯著關係。此一分析模式為教育單位提供了一種融合語意探測與統計推論的整合性方法，不僅能處理複雜非結構化資料，亦能達成對學生風險預警與課程介入的實務應用。

本研究展示出 LDA 與 GLM 結合應用於教育研究的潛力，然仍存在其研究限制，包含有：

一、資料來源與樣本侷限。本研究所採用之資料主要來自單一教育機構的課程評價系統與開放文本評論，資料類型、文本多樣性及樣本數量均有限，恐限制模型推論結果之外部效度。未來研究可考慮整合來自多校、多平台與跨文化教育環境的文本資料，增加樣本異質性，以提升結果普遍性與應用範圍。

二、LDA 模型參數設定問題。LDA 模型的表現高度仰賴參數設定，主題數過多可能導致語意重疊與過度分割，過少則可能遺漏重要資訊，後續仍需更自動化或資料驅動的方法進行最佳參數選擇。

三、本文 LDA 模型主題數量的選擇，主要由研究者審視主題內容的教育意涵與可理解度決定。然，此參數將影響模型的可解釋性與分析效能。為確定最佳主題數量，未來研究可採取多重評估策略，如以困惑度（Perplexity）衡量模型對資料的擬合度，及一致性指標（Topic Coherence Score）評估主題內部詞彙間的語意連貫性。以此綜合比較與判斷，用以選出兼具統計效度與教育詮釋力之最適合的主題數量。

四、文本預處理之敏感性。文本清理與預處理流程（如去除停用詞、詞幹化、詞彙正規化等）對於 LDA 主題結果具有顯著影響，稍有差異即可能造成語意分類結果迥異。本研究雖遵循一般語言處理規範執行預處理，但仍建議未來可對比不同預處理策略下之主題穩定性，提升模型可靠性。

基於以上限制，未來的研究方向將考慮增加數據來源和規模，並收集更多來自不同教育平台和機構的文本資料，提高研究結果的普遍性。在優化模型參數選擇上，需探索更有效的參數選擇方法，提高 LDA 模型的效能。同時，建議結合人工智慧深度學習技術，將 LDA 與深度學習模型結合，提升文本理解和主題建模的效能。同時，在主題模型建模中，可以加入追蹤學生反饋和學習趨勢的時間變化，分析教學方法的長期影響，提升研究數據的嚴謹性。總體而言，本研究結合主題模式與廣義線性模式為教育文本資料分析提供了一種有效的方法，並為教育領域的數據驅動決策提供更強有力的支持。總結而言，本研究展示 LDA 與 GLM 於教育文本資料分析之整合應用潛力，為教育領域提供了一套可行的數據驅動決策支援架構。透過自動主題建模、統計關聯分析與跨資料串接，不僅拓展了教育研究的工具視野，更為課程設計、學生支持與教育政策規劃帶來實質助益。未來若能結合更先進之語言模型、動態分析技術及多模態資料應用，將有望進一步推動高等教育研究之深化與實踐創新，實現更具精準性與可行性的教育治理與學生支持系統。

本研究探討將 LDA 與 GLM 於教育研究中的整合應用潛力。LDA 能自動化分析大量文字資料，如學生回饋、線上討論與學習歷程紀錄，進而萃取潛在主題與語意特徵；GLM 則可建立統計模型，檢驗這些語意主題與學習成效、課程設計或學習行為之間的關聯。結合兩者，有助於形塑兼具語意理解與統計推論能力的分析框架，彌合質性與量化研究之間的鴻溝。透過此方法，教育研究者能更有效率且系統性地處理龐雜的學習資料，提升資料分析的精確度與可解釋性。同時，本研究的成果亦可為教育實務提供具體啟示，協助教師與決策者洞察學生需求、優化教學設計與學習評估，促進以資料為本的教育創新與持續改進。

參考文獻

王文科、王智弘（2010）。質的研究的信度和效度。**彰化師大教育學報**，*17*，29-50。https://doi.org/10.6769/JENCUE.201006.0029

吳雨桑、林建平（2009）。大學生英語學習環境、學習動機與學習策略的關係之研究。**臺北市立教育大學學報**，*2*，181-222。

張春興（1996）。**教育心理學－三化取向的理論與實踐**。臺北市：東華。

邱貴發（1996）。**情境學習理念與電腦輔助學習—學習社群理念探討**。臺北市：師大書苑。

張玟慧、吳佳娣、陳彤宣（2016）。專題式程式設計教學對國小學童問題解決歷程之研究。**教育科技與學習**，*4*（2），137-162。

黃昌誠（1990）。**空中大學學生學習行為與學習困擾之研究**。未出版之碩士論文，國立高雄師範大學教育研究所，高雄市。檢自：https://hdl.handle.net/11296/fxb7a2

蔡惠雅（2013）。**臺南市國中三年級學生國語文學習困擾來源及其因應策略**。未出版之碩士論文，國立臺南大學教育學系課程與教學研究所，臺南市。檢自：https://hdl.handle.net/11296/p5em3r

賈旻暟（2016）。**新北市國中七年級學生英語文學習困擾與因應策略之相關研究**。未出版之碩士論文，臺北市立大學學習與媒材設計學系課程與教學碩士學位在職專班，臺北市。檢自： https://hdl.handle.net/11296/g6ejgr

Ahlers, R., Driskell, J., & Garris, R. (2002). Games, motivation, and learning: A research and practice model. *Simulation and Gaming, 33*, 441-467.

Astin, A. W. (1999). Student involvement: A developmental theory for higher education. *Journal of College Student Development, 40*(5), 518-529.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485-540. https://doi.org/10.3102/00346543055004485

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Bouhuijs, P.A.J., Perrenet, J.C., & Smits, J.G.M.M. (2000). The suitability of problem-based learning for engineering education: theory and practice. *Teaching in Higher Education, 5*(3), 345-358.

Chen, S. H. (2012). Factors influencing college students' dropout intention: A study in Taiwan. *Journal of Educational Research and Development, 8*(2), 25-42.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirec- tional transformers for language understanding. *In*

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 4171-4186). Stroudsburg: Association for Computational Linguistics.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Hadgraft, R., Ilic, V., & Scott, N. (2003). Engineering education-is problem-based or project-based learning the answer? *Australasian Journal of Engineering Education, 3*, 2-14.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of the 10th European Conference on Machine Learning* (ECML 1998) (pp. 137-142). Berlin: Springer. https://doi.org/10.1007/BFb0026683

Kiili, K. (2007). Foundation for problem-based gaming. *British Journal of Educational Technology, 38*, 394-404.

Kuh, G. D., Kinzie, J., Schuh, J. H., Whitt, E. J., & Associates. (2005). *Student success in college: Creating conditions that matter.* Hoboken: Jossey-Bass.

Marcia, J. E. (1991). Identity and self-development. *Encyclopedia of adolescence, 1*, 529-533.

McCallum, A. K. (2002). *MALLET: A machine learning for language toolkit*. https://mimno.github.io/Mallet/index

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26,* 3111-3119.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 100-108). Stroudsburg: Association for Computational Linguistics.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 1532–1543). Stroudsburg: Association for Computational Linguistics.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition (2nd ed.)*. University of Chicago Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Po- losukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998-6008

附錄一、主題模式分析英文文本資料 Python 程式碼

```
import pandas as pd
from gensim import corpora, models
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk

nltk.download('punkt')
nltk.download('stopwords')

# 一、導入數據
dcard_data = pd.read_csv(' 欲分析檔案 ')# 輸入分析檔案
print(dcard_data.head())

# 二、教育文本英文資料處理
# 文本停用詞列表
stop_words = set(stopwords.words('english'))

# 數據預先處理與分詞
def preprocess_text(text):

# 轉為小寫與分詞
    tokens = word_tokenize(text.lower())

# 移除停用詞和非字母字符
    return [word for word in tokens if word.isalpha() and word not in stop_
    words]
# 對每篇文章進行預先處理
processed_data = dcard_data['Article'].dropna().apply(preprocess_text).tol-
ist()

# 創建辭典與詞袋模型
```

```
dictionary = corpora.Dictionary(processed_data)
corpus = [dictionary.doc2bow(text) for text in processed_data]

# 三、LDA 模型訓練
lda_model =
    models.LdaModel( corpus=corpus,
    id2word=dictionary,
    num_topics=5,
    random_state=42,
    passes=10,
    alpha='auto',
    per_word_topics=True
)

# 輸出每個主題前 15 個關鍵詞
for idx, topic in lda_model.print_topics(num_words=15):
    print(f" 主題 {idx}: {topic}")

# 四、主題分布
# 獲取每篇文章的主題分布
doc_topics = [lda_model.get_document_topics(doc) for doc in corpus]

# 獲取每篇文章的主要主題
dcard_data['Topic'] = [max(doc, key=lambda x: x[1])[0] for doc in doc_top-
ics]

# 輸出结果
print(dcard_data.head())
```

附錄二、主題模式分析中文文本資料 Python 程式碼

```
import pyLDAvis.lda_model
import pyLDAvis
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import LatentDirichletAllocation
import pandas as pd
import jieba
import re
import os

source_csv_path = ' 欲分析檔案 ' # 輸入分析檔案
document_column_name = 'content'
top_words_csv_path = 'top-topic-words.csv'
predict_topic_csv_path = 'document-distribution.csv'
html_path = 'document-lda-visualization.html'
n_topics = 5
n_top_words = 15
pattern = u'[\\s\\d,.<>/?:;\'\"[\\]{}()\\|~!\t"@#$%^&*\\-_=+，。\n《》、？：； “” ‘’ { } 【】（）…￥！—… － ]+'
def top_words_data_frame(model: LatentDirichletAllocation,
                         tf_idf_vectorizer: TfidfVectorizer,
                         n_top_words: int) -> pd.DataFrame:
    rows = []
    feature_names = tf_idf_vectorizer.get_feature_names_out()
    for topic in model.components_:
        top_words = [feature_names[i]
                     for i in topic.argsort()[:-n_top_words - 1:-1]]
        rows.append(top_words)
    columns = [f'topic word {i+1}' for i in range(n_top_words)]
    df = pd.DataFrame(rows, columns=columns)
    return df
```

```
def predict_to_data_frame(model: LatentDirichletAllocation, X: np.ndarray)
-> pd.DataFrame:
    matrix = model.transform(X)
    columns = [f'P(topic {i+1})' for i in range(len(model.components_))]
    df = pd.DataFrame(matrix, columns=columns)
    return df

df = (
    pd.read_csv(
        source_csv_path, encoding='utf-8-sig')
    .drop_duplicates()
    .rename(columns={ document_column_n
        ame: 'text'
    }))

stop_words_set = set([' 你 ', ' 我 ', ' 的 ', ' 是 ', ' 和 '])

df['cut'] = (
    df['text']
    .apply(lambda x: str(x))
    .apply(lambda x: re.sub(pattern, ' ', x))
    .apply(lambda x: " ".join([word for word in jieba.lcut(x) if word not in
    stop_words_set]))
)

tf_idf_vectorizer = TfidfVectorizer()
tf_idf = tf_idf_vectorizer.fit_transform(df['cut'])

lda =
    LatentDirichletAllocation( n_com
    ponents=n_topics, max_iter=1000,
```

```
        learning_method='online',
        learning_offset=500,
        random_state=0)

lda.fit(tf_idf)
top_words_df = top_words_data_frame(lda, tf_idf_vectorizer, n_top_words)
top_words_df.to_csv(top_words_csv_path, encoding='utf-8-sig',
index=None)

X = tf_idf.toarray()
predict_df = predict_to_data_frame(lda, X)

predict_df.to_csv(predict_topic_csv_path, encoding='utf-8-sig', index=None)
data = pyLDAvis.lda_model.prepare(lda, tf_idf, tf_idf_vectorizer)
pyLDAvis.save_html(data, html_path)

os.system('clear')
os.system(f'start {html_path}')

print(' 程式生成文件：',
        top_words_csv_path,
        predict_topic_csv_path,
        html_path)
```

附錄二、主題模式分析中文文本資料 Python 程式碼

```
library(jtools)
library(lmtest)
library(readxl)
WD <- read_excel(" 欲分析檔案 ") # 輸入分析檔案
view(WD)

#
# 模型 1: 休退學 + 學校符合期待 + 滿意就讀領域 + 考試入學
#
Model1 <- glm( 人數 ~ 休退學 + 學校符合期待 + 滿意就讀領域 + 考試入學 , family = poisson(link = "log"), data = WD)
summary(Model1); jtools::summ(Model1, digits = 6, confint = TRUE, exp = TRUE)

with(Model1, cbind(res.deviance = deviance, df = df.residual,
                   p = pchisq(deviance, df.residual, lower.tail=FALSE)))

#
# 模型 2:  學校符合期待 * 滿意就讀領域 + 休退學 * 滿意就讀領域 + 休退學 * 考試入學 + 滿意就讀領域 * 考試入學
#
Model2 <- glm( 人數 ~ 學校符合期待 * 滿意就讀領域 + 休退學 * 滿意就讀領域 + 休退學 * 考試入學 + 滿意就讀領域 * 考試入學 , family = poisson(link = "log"), data = WD)
summary(Model2); jtools::summ(Model2, digits = 6, confint = TRUE, exp = TRUE)

with(Model2, cbind(res.deviance = deviance, df = df.residual,
                   p = pchisq(deviance, df.residual, lower.tail=FALSE)))

#
```

```
# 模型 3: 休退學 * 學校符合期待 + 休退學 * 滿意就讀領域 + 休退學 * 考試入學 + 學校符合期待 * 滿意就讀領域
#
Model3 <- glm( 人數 ~ 休退學 * 學校符合期待 + 休退學 * 滿意就讀領域 + 休退學 * 考試入學 + 學校符合期待 * 滿意就讀領域 , family = poisson(link = "log"), data = WD)
summary(Model3); jtools::summ(Model3, digits = 6, confint = TRUE, exp = TRUE)

with(Model3, cbind(res.deviance = deviance, df = df.residual,
                   p = pchisq(deviance, df.residual, lower.tail=FALSE)))

#
# 模型 4: 休退學 * 學校符合期待 * 滿意就讀領域 * 考試入學
#
Model4 <- glm( 人數 ~ 休退學 * 學校符合期待 * 滿意就讀領域 * 考試入學 , family = pois-son(link = "log"), data = WD)
summary(Model4); jtools::summ(Model4, digits = 6, confint = TRUE, exp = TRUE)

with(Model4, cbind(res.deviance = deviance, df = df.residual,
                   p = pchisq(deviance, df.residual, lower.tail=FALSE)))

newdata = WD
predict(Model1, newdata, type="response")
predict(Model2, newdata, type="response")
predict(Model3, newdata, type="response")
predict(Model4, newdata, type="response")

lrtest(Model1,Model4)
lrtest(Model2,Model4)
lrtest(Model3,Model4)  #          Pr(>Chisq) =       0.2047
```

lrtest(Model2,Model3) # 模型比較：模型 3 誤差項平方和小於模型 2，故選擇模型 3。

A Study on Educational Text Analysis Using Topic Modeling and Generalized Linear Models

Yi-Hsiung Lin*

Postdoctoral Research Fellow, Office of Sustainability and Strategic Development
National Cheng Kung University
Tainan City, Taiwan
Assistant Researcher, Center for Teaching and Learning
National Kaohsiung University
Kaohsiung City, Taiwan
E-Mail: 11408046@gs.ncku.edu.tw

Chien-Chih Lai

Project Manager, Center for Teaching and Learning
National University of Kaohsiung
Kaohsiung City, Taiwan
E-Mail: jjlai@nuk.edu.tw

Ren-Yu Huang

Project Assistant, Center for Teaching and Learning
National University of Kaohsiung
Kaohsiung City, Taiwan
E-Mail: ksu0001@nuk.edu.tw

Chih-Hung Chang

Distinguished Professor
Department of Applied Mathematics
National University of Kaohsiung
Kaohsiung City, Taiwan
E-Mail: chchang@nuk.edu.tw

Abstract

The application of topic modeling in educational text analysis is a critical area of natural language processing. With the digitization of educational resources, effectively organizing, retrieving, and analyzing large volumes of educational text data has become a significant challenge for researchers. Topic modeling is a statistical method that automatically extracts key concepts from texts, assisting educators and administrators in making data-driven decisions. In recent years, with advancements in artificial intelligence and machine learning, topic modeling has gained increasing attention in the educational domain, particularly in the context of digital learning platforms and the widespread evaluation of online educational re- sources. This study employs a web crawler to collect data related to college students' learning evaluations from online discussion platforms. By using topic modeling for automated analysis, it reduces the time cost of manual annotation and enhances data processing efficiency. Latent Dirichlet Allocation (LDA), a widely used probabilistic topic modeling method, has been extensively applied in text classification, sentiment analysis, and knowledge management. This study explores the application of LDA in educational text analysis by constructing topic models to identify key themes in student feedback, course evaluations, and subject content. Furthermore, LDA can be integrated with generalized linear models to examine the relationship between topic analysis and the quantification of student learning outcomes. The findings demonstrate that LDA effectively extracts core content from educational texts. Moreover, when combined with generalized linear models, it provides valuable insights for educational decision-makers, enabling more informed and data-driven decisions. The appendix of this paper provides the Python source code used for topic modeling analysis of the Chinese and English text data, and R source code for Generalized Linear Model, allowing interested readers to utilize it.

Keywords: Topic Modeling; Educational Texts; Latent Dirichlet Allocation; Generalized Linear Model

*** Corresponding author**