國立政治大學應用數學系
碩士學位論文

以大數據分析影響唐詩流通度之因素

Using big data to analyze the reasons for the popularity of Tang poetry

碩士班學生：黃泰霖　撰
指導教授：宋傳欽　博士
姜志銘　博士

中華民國 107 年 6 月 28 日

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

致謝

不知不覺間我竟已開始草擬致謝的內容，想來總有些不真實感，一時間竟讓我不知如何下筆，這一路走來我總有的某種單純的幻想，總以為我仍是那以閱讀和書寫為本命的少年，但這樣的幻想總在認真的眨幾次眼後扭曲，由鏡中映出有著另一種不同的目光的倒影，但無論如何我想我總是成長著的。在這過程中，我由文字的書寫走向走向程式的刻劃，由抒情的字樣走向精確的符碼，最後由那裡來到了這裡，我想我是有所成長的，儘管我總能不斷的認知到自己的不足。而在這努力蛻變的路上，若是少了大家的參與，我想我是無法努力的讓自己一次又一次的重生。所有陪伴我一起努力成長的人，謝謝你們。

謝謝宋傳欽老師與姜志銘老師的悉心教誨，每一次的討論總使我有新的方向與想法；謝謝高桂惠老師提供中文相關專業的見解，使我在難以解釋的主成分與因子上有如同撥雲見日的進展；謝謝譚克平老師的建議，讓我的論文內容更為周延；謝謝張宜武老師的幫助，讓我的論文有新的思考方向；謝謝陳天進老師的每一餐，讓我食物充足；謝謝一路支持我的家人們，儘管我鮮少回到高雄，但每每思及總使我充滿力量；謝謝我的夥伴，如果你看到這個你會知道；謝謝我的好室友暨師兄顗錞，讓我在不正常的作息中有了那麼一點回到正常的跡象，讓我活得像人類；謝謝我的師姐沁如，讓我的日常煙霧瀰漫，請不要再叫我 debug，拜託；謝謝我的師兄張群，我想我會想念我們一起戰鬥到天亮的日子；謝謝小黑，讓我在某幾個月能吃到早餐；謝謝力夫，讓我吸了不少煙味；謝謝爪宏，讓我在周末時總有人可以說話；謝謝大澤佑，教會了我許多知識，也給了我很多幫助，還有好多咖啡；謝謝治鈞，讓我日常充滿奇妙的歡笑；謝謝振偉，給我在這兩年來的幫助，還有一大堆好喝的咖啡。要感謝的人還有很多，但這篇幅限制了我的文字，讓我無法一一感謝研究室的夥伴們，但你們所有人總是使我感到快樂，這對我意義重大。

最後，僅將此篇論文獻給所有在路上與我相伴的人，若我有任何值得一提的閃耀之處，那也一定是因為你們而點亮。

黃泰霖　2018 年 07 月於政大

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

中文摘要

本研究旨在探討唐詩在流通上的特性與原因，期望能為唐詩詩學研究提供新的研究方向。本文以《唐詩排行榜》所建立的資料作為出發點，並以主成分分析與因子分析為主要的分析方法，萃取出唐詩在流傳上的特性及因素，探討古人與今人在詩文閱覽偏好的不同，並進一步利用詞嵌入法探討詩文內容相似度與主成分分析及因子分析之結果在排序上是否一致。

經過對唐詩排行榜數據的研究，本文發覺主成分分析總結出以下兩項特性：1. **時代性差異** 2. **詩文收錄完整性**，其中**時代性差異**顯示『每一個時代的前理解不同，審美標準自然有明顯落差，因而造成古今閱眾對於詩文的欣賞與偏好有一定程度的差異』；而**詩文收錄完整性**指的是『隨著編纂需求的不同，詩作在流傳上可分為 1. 完整詩文 2. 片段名句 兩種類型』。

而因子分析則總結出兩個影響唐詩流通的原因：1. **歷史性強度** 2. **詩學經典性**，其中**歷史性強度**所代表的是『古今閱眾在詩文內容的喜好上，深受詩文內容的歷史背景所影響』；而**詩學經典性**則顯示『從詩學學術領域的角度出發，可區分詩文是否為一派之經典』

利用詞嵌入法進行詩文文本的相似性研究，發現第一主成分**時代性差異**、第一因子**詩學經典性**以及第二因子**歷史性強度**之結果與其分別對應之詩文相似度排序具有顯著的一致性。

關鍵字：大數據、唐詩、流通性、主成分分析、因子分析、詞嵌入法

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

Abstract

This study aims to explore the characteristics of the popularity of Tang poetry, and hopes to provide new research directions for Tang poetry. First, we use multivariate statistical methods, which include principal component analysis and factor analysis, to analyze the data given by the book *Ranking on Tang Poems*. Based on results of analysis, we extract the characteristics of the popularity of Tang poetry, and compare modern with ancient preferences of reading. Finally, we use word embedding techniques to further analyze the suitability of the results extracted by principal component analysis and factor analysis.

After analyzing the data given by the *Ranking on Tang Poems*, principal component analysis suggests the following two characteristics: 1. **time difference** 2. **poem integrity**. "Time difference" refers to "Having its own pre-understanding, each era has its own aesthetic standard, which makes some differences of poetic appreciation between ancient and modern readers." "Poem integrity" refers to "A poem is selected either in a complete form or in a partial form according to the editing requirements."

Based on factor analysis, we sum up two factors that may influence the popularity of Tang poetry : 1. **history related strength** 2. **poetic classicism**. The "history related strength" refers to "The poem preferences of ancient and modern

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

readers may be influenced by the history related strength of the poem.” The “poetic classicism” indicates that “Poem can be considered to lead a school of thoughts from the academic perspective.”

Using word embedding techniques to study the textual similarity of poems, we find that each of first principal component and two factors has a significant rank correlation with the textual similarity of the top ranking poems based on its corresponding principal component or factor.

Keywords: big data, Tang poems, popularity, principal component analysis, factor analysis, word embedding method

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

目錄

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表目錄

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

圖目錄

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第一章

緒論

第一節　研究背景

詩歌的流傳，一向是文化傳承的重要成分，縱觀各地的文化，鮮少有僅具備語言而無詩歌者，詩歌往往承載且濃縮了一個文化中最為精萃的部分，簡練的構句、華美的詞章、寓意深遠的內涵，在在表現出詩歌一類的文學體裁迥異於其他文類的地方，這樣別樹一格的文體，總使大眾為之著迷。且不論各式詩歌展現在大眾眼中的優劣高低，無論是古時的各類詩體，抑或是現今流行的歌曲，都能在簡練的內容中呈現一個文化的核心，以極盡優美的方式傳遞思想，而唐詩，尤其如此。

唐詩者，唐一代之文學也[1]，自初唐以來，各式詩作在這個詩歌蔚為大成的朝代中流轉、成長、豐富、衰亡，那詩歌是混雜的、敘事的、附庸風雅的、歡樂的、哀痛的，在短短的 289 年間，唐詩在時光流轉間增長，在年代中浮沉，在歷史裡醞釀，在現今為酒，入吾人口中，這樣的陳釀單以《全唐詩》一書所載來說，便足足有五萬首之多，如此大量的作品或直接、或間接的影響後世的所有詩歌創作。

唐詩之所以獨特，表現在許多面向，在此挑幾個不同的的論述為其作解：

1. 詩的格律在唐代完整成型，並產生了大量佳作[2]

[1] 引自王國維《宋元戲曲史》序文『凡一代有一代之文學，楚之騷、漢之賦、六代之駢語、唐之詩、宋之詞、元之曲，皆所謂一代之文學，而後世莫能繼焉者也。』

[2] 見趙義山、李修生 (2010, 79 頁)《中國分體文學史-詩歌卷》第四章古近體詩大備及創作繁榮

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

2. 唐詩為中原文化詩歌發展高峰，且表現出詩美的極致[3]

3. 由於復古之風興盛，唐詩在內容與詞藻之間取得平衡[4]

若要以一較為誇張的句子來形容唐詩之盛，或能由魯迅 (2005) 寫在其書信中的話語作為形容，『我以為一切好詩，到唐已被做完，此後倘若非能翻出如來掌心之齊天大聖，大可不必動手』。[5]

直至今日，人們在日常生活中有許多層面仍然受到唐詩的影響，現代人仍會閱讀唐詩以陶冶性情，有關唐詩的詩學研究也不斷推陳出新，每每有新的觀點與新的解釋在融合了現代的視野與眼光後誕生，隨著新的研究方式、新的思潮加入唐詩的詩學研究中，這片領域也隨之蓬勃發展。

而王兆鵬等 (2011)《唐詩排行榜》一書所提出的影響力公式[6]更是為這個研究領域帶入了新的思路—『唐詩與各類資料的關聯為何?』、『唐詩可否排個高低?』『如何以量化的方式來為唐詩建立排序?』、『如何以數據的眼光來看待唐詩?』諸如此類的有趣問題皆可作為思考的方向，而這正是本研究的起點『在量化的視角下，我們還有什麼指標或因素是可以發覺的?』。

[3]見蔣寅 (2003, 39 頁)《中國古代文學通論-隋唐五代卷》，此處之美指自然妙悟、中和之美
[4]初唐陳子昂以復古為革新的詩歌主張，可見趙義山、李修生 (2010)《中國古代文學通論-隋唐五代卷》
[5]魯迅 (2005, 307 頁)《魯迅全集》第 13 卷一九三四年十二月　致楊霽雲
[6]此書作者以此建立唐詩的名篇排行（在書中亦為影響力、公認度），為《唐詩排行榜》書中所蒐集的各項變數之加權

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第二節　研究目的

在《唐詩排行榜》一書出版之前，有關唐詩的量化研究或較為常見的文學計量研究大多著重在簡單的頻率統計之範圍，而在《唐詩排行榜》一書中，則是將視角擴大，進一步將眼光投注在綜合指標[7]的建構上，進入了單變量統計領域。

而正如前一節所述，《唐詩排行榜》在書中提供了一個影響力公式[6]，以之計算唐詩的流通廣泛程度，然而此公式的建立是基於作者團隊的專業評判，其中權重的選擇屬主觀評定的賦權方式[8]，我們不禁好奇，是否能經由多變量統計分析方法的角度來進行數據的客觀解釋，藉此獲得一些唐詩流通性的指標，或者藉由各式唐詩選集的數據進一步發覺影響唐詩在流通度的原因？而這些指標或原因，又是否和其詩文內容有某種程度的相關性？

本研究將根據《唐詩排行榜》一書所提供的的資料來做進一步的分析研究。首先利用兩種維度縮減的多變量統計方法來增加資料的可解釋性，即 1. 嘗試以主成分分析總結《唐詩排行榜》資料的特性，2. 嘗試利用因子分析研究在背後影響唐詩流通的原因；最後應用詞嵌入法探討各主成份、因子之排序與其分別對應的詩文相似度排序是否一致，以期獲得另一種客觀的視角以為參考。

[7]此綜合指標即前一節所提及之影響力公式
[8]詳細說明可見本文第二章文獻回顧

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第三節　論文架構

本節將簡要介紹各章節的內容，並附上流程圖加以說明。

一、 各章節結構與內容

除第一章外，本研究其餘部份可分為六個章節。由於在研究過程中所使用的數據來源為《唐詩排行榜》，故將於第二章文獻回顧中先簡要介紹《唐詩排行榜》一書，並說明其數據收集方式與相關結果，於第三章研究方法中介紹將使用的各種分析方法，並於第四章、第五章、第六章三個章節中分別應用三種方法對《唐詩排行榜》數據進行分析，並討論其結果之意涵，最後於第七章做出總結。

二、 研究流程圖

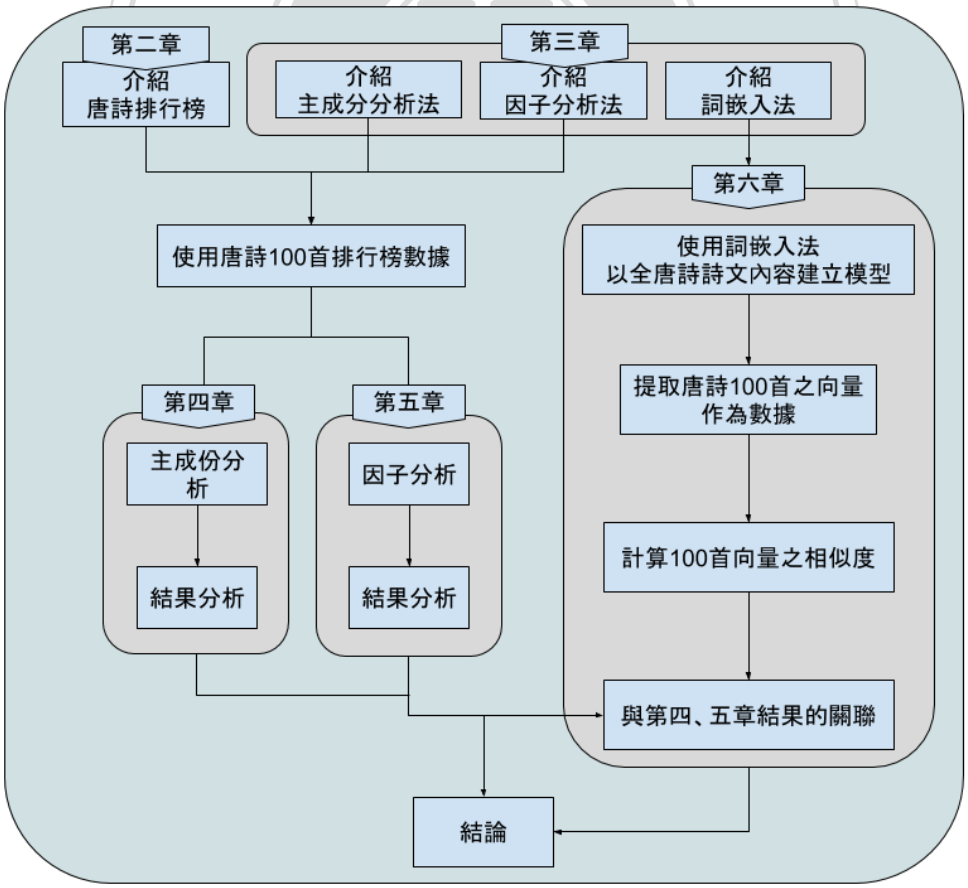


圖. 1.1. 研究流程圖

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第二章

文獻回顧

本章節將介紹本研究中所用到的重要文獻《唐詩排行榜》，於第一節簡要介紹《唐詩排行榜》一書之內容與起源，第二節介紹書中數據的蒐集方式與其意義，並於第三節呈現其研究結果。

第一節　《唐詩排行榜》之簡介

《唐詩排行榜》成書於 2011 年初，本書主旨在於利用所收集的資料為唐詩研究添加一個新穎的研究領域『以定量分析的方式為詩作排序』，其試圖利用所蒐集的數據來解釋唐詩在歷史的流動中有何不同之處，再由這些數據建立一條公式，並給定權重，進一步給出一個較為客觀的指標來定義熱門程度（即流通度廣泛程度）高低，評比在唐詩中哪一首詩是最熱門的詩作，在成書的過程中，本書針對諸多著名的唐詩收集了許多面向的資料，以各式文本類型為依據，共收集了七個面向的數據，雖說本排行結果曾引發爭議，但在唐詩研究上，可說首開先河，極富意義。

第二節　數據收集方式

《唐詩排行榜》主要關注的重點在於唐詩影響力的分析，故其所收集的數據當與影響力強弱相關，在此書中具有影響力的詩作，其意義並非一時的熱門或爆紅的詩作，而是公眾持久認同的作品，而作品的公認度，亦即民意的認同程度，而認同程度也可以理解為受到關注的程度。

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

以下是《唐詩排行榜》一書中對於影響力的解釋：

『公眾對作品的關注度越高,作品的知名度就越高、影響力就越大、名篇指數也越高。』[1]

在此書中作者將關注的方式與讀者分為三類：

1. 對作品的閱讀–消費型的普通讀者

2. 對作品的評論–批評型的專家讀者

3. 對作品的效仿再創作–創作型的作家

再依據此分類分別收集的不同種類的選集、評論作為數據（其中由於第 3 種類型的數據難以客觀計量，故並未收集），以下將分別列出分類與其對應的意義與數據：

對作品的閱讀
消費型的普通讀者

普通讀者，他們傾向於閱讀喜歡的作品，而對於不喜歡的作品予以忽視，故本類型讀者在選擇上是『被動的』[2]，以『無聲的選擇』來表明對文學作品的態度，因此我們可以把『對作品的閱讀』這一面向簡單的看成讀者認同度。而與其對應的數據便是選本類型作品的數量，當讀者閱讀某首詩作之後，對此詩的想法無論是正向或是負向的觀點，都會影響到選本或詩集流通的能力，所以此詩在各式選本所選入的數量，需要考量到讀者群眾的想法，故詩選的編輯者將會依循此想法來選擇詩作，因此選本或選集類的文本將會表現出在一般受眾之中的影響力強度，而據此選入的有古代選本、現代選本兩種數據。

對作品的評論
批評型的專家讀者

專家讀者，評論者，不但閱讀作品，亦會據此發表評論，議論此作品的優劣，而無論是正面或負面的評論，都表明批評者對作品的關注。也會影響到作品的流通，進而影響作品的影響力強度，甚至有

[1]引自《唐詩排行榜》前言，其中名篇指數為其排序依據，可見本章第三節
[2]此處之被動指讀者是被動地挑選已經流通於市面的作品，無法主動決定哪一首詩的流通，只能以選擇、買或不買來進行意見的表達，進而影響流通

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

時詩派與詩派之間的角力，也是看點之一。而與其對應的數據便是評論類型與研究類型作品的數量，這些作品直接的表現出評論者或研究者對於本篇詩作的看法，表現出此詩作在唐詩詩學研究領域中的影響，而據此選入的有歷代評點、論文篇數、文學史全錄、文學史摘錄。

對作品的效仿創作型的作家　作家也是讀者的一種，只是其表現出來的閱讀結果並非只是沉默的觀看或提出批評與建議，作家型的讀者將會借鑑、吸收所閱讀的作品，進一步化為自己創作的養分，無論是一般的創作或是借鑑了相同設定的作品（二次創作），還是格式相仿的詩作，都屬於此類。由於本類型的數據難以直接收集，故實際應用上不予採計。

由於在研究過程中使用到了本書所載之數據，故以下簡介《唐詩排行榜》中七個變數 $(X_1, X_2, \ldots, X_7)$[3]的收集方式與意義，以便作為後續分析之用：

X_1 **(古代選本選錄數)：**

古代選本者，所指為唐詩成詩之後至民國之前，被各個朝代的詩選之選輯者選中進而被收錄的數量，依據年代不同挑選了 31 本詩選作為計量依據，共有

1. 唐人選唐詩　4 本
2. 宋金元人選唐詩　5 本
3. 明人選唐詩　11 本
4. 清人選唐詩　11 本

，某詩作如被幾個選本選錄便會計為幾次，其意義在於計算古人中普通讀者對於某篇詩作的關注度。

如：xxx 詩在本變量上計數為 3，即代表此詩曾三次被歷代詩人選中，錄入籍中。

X_2 **(現代選本選錄數)：**

現代選本所指的是，由作者挑選共計 37 種影響力較大的近代唐詩選本[4]作為計量

[3]分別對應古代選本、現代選本、歷代評點、論文篇數、文學史全錄、文學史摘錄、網路連結總數，共計七種類型

[4]此處所指之近代者，為民國初年以後所創作之唐詩選本

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

基準，統計其中各詩篇出現的數量，其計算方式與古代選本的計算方法相同，某詩作被幾種現代選本選錄即計為幾次，其意義在於計算現代人中普通讀者對於某篇詩作的關注度。

X_3 **(歷代評點選錄數)：**

《唐詩排行榜》以陳伯海《唐詩匯評》[5]一書為依據，每首詩下有多少評論，即計量為有多少歷代評點數量，評論的採計無論正負，只紀錄數量，以此作為在古人中評論型讀者對某詩的關注程度。

X_4 **(論文引用數)：**

《唐詩排行榜》依《20 世紀唐五代文學研究論著目錄檢索系統與定量分析》[6]為來源，以篇數作為計量基準，每首詩曾被多少研究著作提及便計數為多少，本變量代表了在現代詩學研究中，研究型讀者針對某首詩作的關注度。

X_5 **(文學史全詩選錄數)：**

作者選取九種文學史為計量依據，計算各個詩作在文學史中被提及的次數，一首詩凡被全詩引錄即計數為一，也代表了在現代的研究者群中的關注程度。

X_6 **(文學史摘錄數)：**

同文學史全詩部分，但只記錄摘句介紹的部分。

X_7 **(網路連結總數)：**

於搜尋引擎輸入詩作名稱時，所能得到的網路連結數量，此書所指網路連結總數的乃是此書的研究團隊於三個不同時段下，於 Google 和百度兩大搜尋引擎中輸入詩人名稱和詩作篇名搜索後所得加總連結數的平均值，《唐詩排行榜》內提及『由於網路連接總數的資料會隨時間不斷增長、更新快速，故在《唐詩排行榜》書中並無使用此變量作為計算依據』。

[5]陳伯海《唐詩匯評》作為歷代唐詩論評輯要，記錄了唐以來直至清各式文獻中對於唐詩評論的資料，此外內含少數近代人，但仍健在者不錄入

[6]《唐詩排行榜》作者王兆鵬所建立的研究項目

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第三節　影響力公式

根據《唐詩排行榜》作者對於唐詩的理解與研究，其於書中以主觀賦權中的專家評判法的模式總結出一套加權公式用以計算唐詩的熱門程度，此公式即唐詩影響力公式。各變數的權重如表 2.1所列：

表. 2.1. 影響力公式中變數權重表

變數權重表

資料來源	權重	變數名稱	條件權重	最終權重
選本	50%	X_1(古代選本選錄數)	60%	30%
		X_2(現代選本選錄數)	40%	20%
歷代評點	30%	X_3(歷代評點選錄數)		30%
論文篇數	10%	X_4(論文引用數)		10%
文學史	10%	X_5(文學史全詩選錄數)	70%	7%
		X_6(文學史摘錄數)	30%	3%
網路連結總數	0%	X_7(網路連結總數)		0%

此外，本書也對資料內容進行預處理，將每筆資料除以其變數之最大值，如下所示：

$$y_{ik} = \frac{x_{ik}}{\max\limits_{\forall j}(x_{ij})}$$

，其中 x_{ij} 表示在 X_i 變量下的第 j 筆觀察值。

據此，作者提出影響力公式，如下所列：

$$\begin{aligned} z_i &= \left(\frac{x_{1i}}{\max\limits_{\forall j}(x_{1j})}\right) \times 30\% + \left(\frac{x_{2i}}{\max\limits_{\forall j}(x_{2j})}\right) \times 20\% + \left(\frac{x_{3i}}{\max\limits_{\forall j}(x_{3j})}\right) \times 30\% \\ &+ \left(\frac{x_{4i}}{\max\limits_{\forall j}(x_{4j})}\right) \times 10\% + \left(\frac{x_{5i}}{\max\limits_{\forall j}(x_{5j})}\right) \times 7\% + \left(\frac{x_{6i}}{\max\limits_{\forall j}(x_{6j})}\right) \times 3\% \\ &= \left(\frac{x_{1i}}{17}\right) \times 30\% + \left(\frac{x_{2i}}{30}\right) \times 20\% + \left(\frac{x_{3i}}{38}\right) \times 30\% \end{aligned}$$

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

$$+\left(\frac{x_{4i}}{70}\right)\times 10\%+\left(\frac{x_{5i}}{9}\right)\times 7\%+\left(\frac{x_{6i}}{8}\right)\times 3\%$$

$$=(y_{1i})\times 30\%+(y_{2i})\times 20\%+(y_{3i})\times 30\%$$

$$+(y_{4i})\times 10\%+(y_{5i})\times 7\%+(y_{6i})\times 3\%$$

，其中 z_i 為第 i 首詩對應的名篇指數。

依據本公式所建立的名篇指數，《唐詩排行榜》排列出 100 首名篇詩作的影響力排行[7]，藉此反應排行榜內相關詩作的受關注度高低和影響力的大小。

[7]《唐詩排行榜》一書中定義此排行為影響力、综合名篇指數，但保守一點的說法可為流通廣泛程度、詩的綜合熱門程度

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第三章

研究方法

本章節將介紹研究中所應用的方法、效果及其來源，於第一節介紹主成分分析之目的與其方法簡介，第二節介紹因子分析之目的與其方法簡介，第三節介紹詞嵌入法之目的與其方法簡介並給一簡要的例子示範其效果。

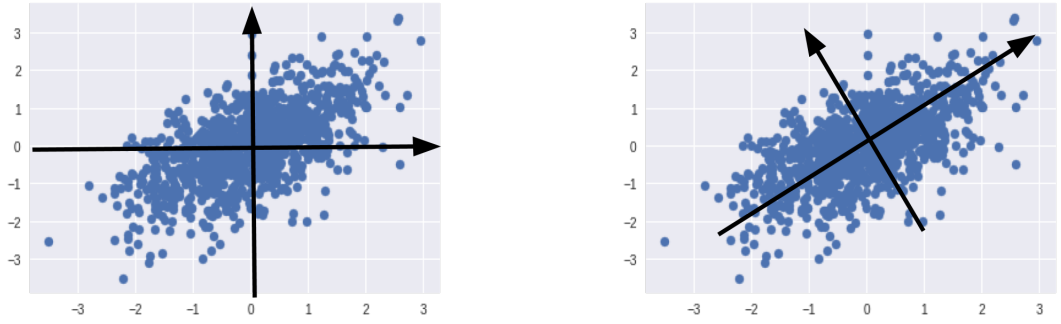
第一節　主成分分析

主成分分析（*principal components analysis*, PCA）常見於數據分析的應用之中，主要作為維度縮減（*dimension reduction*）之用，是一種透過尋找線性組合來進行資料總結的方式，透過所求得的線性組合，主成份分析能夠將原始變量投影到新的空間，而新空間中第一個座標軸[1]會佔有最大的變異，而第二個座標會佔有次大的變異，而在後續的軸上依次遞減，故在前幾大主成份中較容易突顯資料的變化[2]。

在這樣的新空間中，可以只選取前幾大具有代表性的主成分作為代表，而忽視後面佔據變異較小的軸，以此降低資料的維度，減少變量的數量，更可藉此讓新的變量彼此間的相關係數降低至零，更重要的是由於主成分分析在進行維度縮減之時所忽視的是變異較小的軸，故能盡量保持原資料的變化程度，在維度縮減的過程中損失較少的資訊。

[1]即第一主成分軸
[2]可見圖 3.1，取此圖之第一軸即可表現資料主要的變化方向

DOI:10.6814/THE.NCCU.MATH.003.2018.B01



(a) 原始資料座標軸　　(b) 使用主成分分析後座標軸

圖. 3.1. 原始資料座標軸與主成分座標軸比較圖

以下為主成分分析的理論：

令 $X_1, X_2, \ldots, X_p$ 分別表示 p 個變數，

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1p}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}^2 & \sigma_{p2}^2 & \cdots & \sigma_{pp}^2 \end{pmatrix}_{(p,p)}$$ 為變異–共變異矩陣[3]，

若考慮以下之線性變換：

$$\begin{aligned} Y_1 &= \boldsymbol{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= \boldsymbol{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ Y_p &= \boldsymbol{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned}$$

[3]以下之計算亦可以改以相關係數矩陣 $\boldsymbol{\rho}$ 進行，其計算過程相同

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

與其變異數與共變異數：

$$\text{Var}(Y_i) = \boldsymbol{a}_i' \boldsymbol{\Sigma} \boldsymbol{a}_i \qquad i = 1, 2, \ldots, p$$

$$\text{Cov}(Y_i, Y_k) = \boldsymbol{a}_i' \boldsymbol{\Sigma} \boldsymbol{a}_k \qquad i, k = 1, 2, \ldots, p$$

在主成分分析中希望能求得『彼此之間無線性相關』且『變異數越大越好』的線性組合 $Y_1, Y_2, \ldots, Y_p$，且第一個線性組合 Y_1 佔有最大的變異數，第二個線性組合 Y_2 的變異數則小於第一個線性組合的變異數，第三個線性組合 Y_3 再次之，由此而至變異數最小的第 p 個線性組合 Y_p，故可定義 $Y_1, Y_2, \ldots, Y_p$ 如下：

第一主成分 = 使 $\text{Var}(\boldsymbol{a}_1' \mathbf{X})$ 最大的線性組合 $\boldsymbol{a}_1' \mathbf{X}$

且 $\boldsymbol{a}_1' \boldsymbol{a}_1 = 1$

第二主成分 = 使 $\text{Var}(\boldsymbol{a}_2' \mathbf{X})$ 最大的線性組合 $\boldsymbol{a}_2' \mathbf{X}$

且 $\boldsymbol{a}_2' \boldsymbol{a}_2 = 1,\ \text{Cov}(\boldsymbol{a}_1' \mathbf{X}, \boldsymbol{a}_2' \mathbf{X}) = 0$

$\vdots$

第 p 主成分 = 使 $\text{Var}(\boldsymbol{a}_p' \mathbf{X})$ 最大的線性組合 $\boldsymbol{a}_p' \mathbf{X}$

且 $\boldsymbol{a}_p' \boldsymbol{a}_p = 1,$ 對所有 $i, j < p,\ \text{Cov}(\boldsymbol{a}_i' \mathbf{X}, \boldsymbol{a}_j' \mathbf{X}) = 0$

其中 $\boldsymbol{a}_i' \boldsymbol{a}_i = 1$，為限制主成份解為唯一之限制式

根據以上定義，若令 $\lambda_1, \lambda_2, \ldots, \lambda_p$ 為變異–共變異矩陣 $\boldsymbol{\Sigma}$ 的特徵值，且 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$，而 $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ 為對應於 $\lambda_1, \lambda_2, \ldots, \lambda_p$ 的特徵向量，則可證得應取以下

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

之線性組合作為主成分[4]：

$$
\begin{aligned}
Y_1 &= \boldsymbol{e}_1'\mathbf{X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\
Y_2 &= \boldsymbol{e}_2'\mathbf{X} = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
Y_p &= \boldsymbol{e}_p'\mathbf{X} = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p
\end{aligned}
$$

而在此線性組合中，其變異數與共變異數如下：

$$
\begin{aligned}
\text{Var}(Y_i) &= \lambda_i \qquad i = 1, 2, \ldots, p \\
\text{Cov}(Y_i, Y_j) &= 0 \qquad i \neq k
\end{aligned}
$$

在實務上，這些線性組合的係數 e_{ij}，是解釋主成分意義的重要依據，當一主成分的係數皆為正值時，此主成分的解釋應由綜合指標的角度出發，而其係數若是有正有負時，則此主成分的意義通常解釋為變數與變數之間的對立關係[5]。

在實務上，若變數的變異相差過大、變數之意義相去甚遠的情況下，通常建議以相關係數矩陣 $\boldsymbol{\rho}$ 進行主成分分析，其計算過程相同。

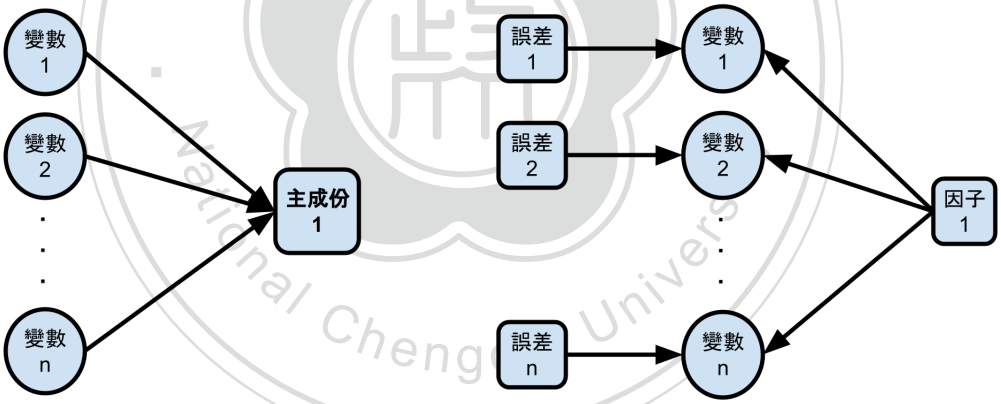
[4]詳細證明過程見 Johnson and Wichern (2007, page. 432) *Applied multivariate statistical analysis*
[5]此處的判準可見 陳耀茂 (1999, 第四章)《多變量解析方法與應用》

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第二節　因子分析

因子分析（*factor analysis*），亦名因素分析，在資料分析中也是維度縮減的一種方法，在某種程度上可以視為主成分分析的擴展，如同主成分分析一般，因子分析所求的也是一種線性組合，但與主成分分析的不同之處在於其建立線性組合的目的不再是進行資料特性的總結，而是去尋找某種不可見的原因。

因子分析假定在資料的群體背後具備一定程度的共同因子，此處的因子所指的是某種無法在原始變量中直接觀察得到，但卻影響了變量變化的『潛在原因』，比方說在一項測驗的結果中，潛藏在背後影響學生成績高低的可能是『智商』、『對某種領域的擅長程度』之類較為抽象的概念，而因子分析的目標便是將此類的共同原因抽取出來，成為一項可以量化的指標。



(a) 主成分與變數關係　　(b) 因子與變數關係

圖. 3.2. 主成分及因子與變數關係比較圖

圖 3.2為主成分與因子之間的差異比較，其中以 $\longrightarrow$ 表示相互影響、因果關係，箭號之起點為因、終點為果，而圖3.2(b)中誤差項所指者為因子所不能解釋的部份，為變量本身的獨特變化。由此比較圖可以直觀的認識到主成分分析為變數的特性總結，因子分析卻是潛藏於變數之後的原因。

以下為因子分析的理論：

與主成分相同，令 $X_1, X_2, \ldots, X_p$ 分別表示 p 個變數，

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

$$\boldsymbol{\rho} = \begin{pmatrix} \rho_{11}^2 & \rho_{12}^2 & \cdots & \rho_{1p}^2 \\ \rho_{21}^2 & \rho_{22}^2 & \cdots & \rho_{2p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1}^2 & \rho_{p2}^2 & \cdots & \rho_{pp}^2 \end{pmatrix}_{(p,p)}$$ 為相關係數矩陣[6]，

其中 ρ_{ij} 表示變數 X_i 與 X_j 間的相關係數

考慮以下之線性變換，作為變數與因子之間關係的描述：

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1p}F_p + \epsilon_1 \\ X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2p}F_p + \epsilon_2 \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pp}F_p + \epsilon_p \end{aligned}$$

或以矩陣表之：

$$\underset{(p\times 1)}{\mathbf{X} - \boldsymbol{\mu}} = \underset{(p\times m)}{\mathbf{L}} \; \underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\epsilon}}$$

其中 ℓ_{ij} 表示第 j 因子在變數 i 上的係數，通稱因子負荷（*factor loading*），F_j 表第 j 個共同原因，稱之為公因子（*common factor*）或潛在因子（*latent factor*），ϵ_i 表第 i 變數所無法被公因子們所解釋的部份，通稱特定因子（*specific factor*）或誤差項（*error term*），而 μ_i 則是各變數的均數。由本式可知，因子分析所求得的共同因子可以用以解釋變數變化原因。

為求得公因子，常見的方法有以下兩種[7]：

(i) 主成分法（*The Principal component Method*）

[6]同主成分分析，以下之計算亦可以改以變異-共變異矩陣 $\boldsymbol{\Sigma}$ 進行，其過程相同
[7]詳見 Johnson and Wichern (2007, chapter. 9) *Applied multivariate statistical analysis*

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

(ii) 最大概似法（*The Maximum Likelihood Method*）

兩種方法都是由相關係數矩陣 ρ 作為出發點，進一步求解公因子 **F** 與特定因子 ϵ，惟最大概似法須假定公因子 **F** 與特定因子 ϵ 服從常態分配。

而在求得原始的因子之後，為了解釋上的便利性，通常會將所得的因子進行轉軸。轉軸法分為正交轉軸（*orthogonal rotation*）與斜交轉軸（*oblique rotation*）兩種，分別表示轉軸後之因子正交與否，其中斜交轉軸由於因子間相互影響，在解釋上較為困難，不常使用。

和主成分分析相似，在解釋因子時，須應用轉軸後所求得的因子負荷來判斷因子所代表的意義，依據 陳耀茂 (1999, 第五章)《多變量解析方法與應用》一書，可將因子分為以下四種類型：

綜合力型 所有的因子負荷之值，均為同號且較大時，可以判斷此因子為綜合型的因子

一極型 某幾個因子負荷大，其餘因子負荷小，表示單一因子

二極型 有較大的正值與負值的因子負荷時，可視為對立的因子

無力型 所有的因子負荷之值接近於零，則此軸無法解釋

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第三節　詞嵌入法

詞嵌入法（*word embedding method*）並非一般意義上的統計方法，而是一種轉換資料的過程，是在機器學習中自然語言處理領域為了進行文本分析而建立的方法，目的是將原本計算上無意義的字或詞編碼為有意義的向量以作為數值計算之用，可以解決早期對字或詞進行編碼時遇上的問題。

在早期的轉化方法中所採行的方式為獨熱編碼（one-hot encoding）[8]，採用此編碼方式轉化向量時，由於文本量通常十分龐大[9]，會導致計算效率低下與向量彼此之間無意義的情況發生。

為解決此問題 Mikolov et al. (2013) 提出 *word2vec*，一種以淺層神經網路為基礎的方法來降低維度，並在降低維度後使詞向量間具有一定程度的相關性，其主旨是透過淺層神經網路的迭代，使原本相互垂直的向量投影到較低的維度中，並在此較低維度的空間中，詞向量與詞向量的差異性能用餘弦相似度的方式描述，接近的詞向量，夾角將越小，餘弦值亦越高。

為了解釋上的便利，以下將以較小的例子來解釋詞嵌入法，而為進一步突顯字詞間的差異性，舉『日』、『月』、『吃』、『飯』四個差異性較大的詞語來作為示例[10]，並假設『吃』、『飯』兩字曾在前後文出現。

首先將文字作獨熱編碼使其成為

$$\boldsymbol{v}_{日}=\begin{pmatrix}1\\0\\0\\0\end{pmatrix},\ \boldsymbol{v}_{月}=\begin{pmatrix}0\\1\\0\\0\end{pmatrix},\ \boldsymbol{v}_{吃}=\begin{pmatrix}0\\0\\1\\0\end{pmatrix},\ \boldsymbol{v}_{飯}=\begin{pmatrix}0\\0\\0\\1\end{pmatrix}$$

[8]獨熱編碼（one-hot encoding），對單一字詞給予一維度為字量總數的標準基底
[9]由於建立在維度為字數總量的標準基底上，若有 1 萬個字或詞就需要維度為 1 萬的向量
[10]本例參考 http://cpmarkchang.logdown.com/posts/773062-neural-network-word2vec-part-1-overview

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

，而在此情況下 word2vec 的網路模型將如下圖 3.3所示：

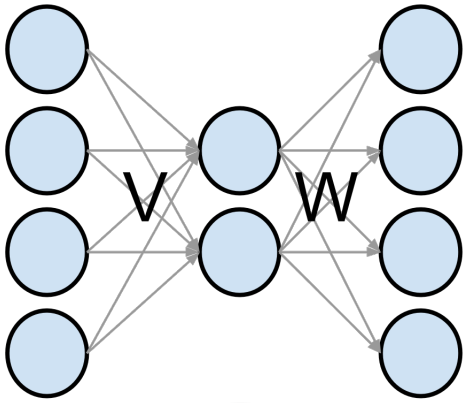


圖. 3.3. *word2vec* 網路圖

其中

$$\mathbf{V}=\begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \\ V_{31} & V_{32} \\ V_{41} & V_{42} \end{pmatrix}$$ 表示輸入層到隱藏層的權重，

$$\mathbf{W}=\begin{pmatrix} W_{11} & W_{12} & W_{13} & W_{14} \\ W_{21} & W_{22} & W_{23} & W_{24} \end{pmatrix}$$ 表示隱藏層到輸出層的權重

假設將 $\boldsymbol{v}_{吃}$ 輸入本網路，將得到如下運算：

$$\left(\boldsymbol{v}_{吃}^{'}\mathbf{V}\right)\mathbf{W}=\begin{pmatrix} V_{31} & V_{32} \end{pmatrix}\mathbf{W}$$

$$=\begin{pmatrix} V_{31}W_{11}+V_{32}W_{21} \\ V_{31}W_{12}+V_{32}W_{22} \\ V_{31}W_{13}+V_{32}W_{23} \\ V_{31}W_{14}+V_{32}W_{24} \end{pmatrix}=\begin{pmatrix} V_3W^1 \\ V_3W^2 \\ V_3W^3 \\ V_3W^4 \end{pmatrix}$$

其中 V_3 表示矩陣 **V** 之第 3 列，其中 W^i 表示矩陣 **W** 之第 i 行

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

而最後輸出的結果在通過 sigmoid 函數[11]得

$$\text{sigmoid}\left(\left(\boldsymbol{v}'_{吃}\mathbf{V}\right)\mathbf{W}\right)=\begin{pmatrix}\frac{1}{1+e^{-V_3W^1}}\\ \frac{1}{1+e^{-V_3W^2}}\\ \frac{1}{1+e^{-V_3W^3}}\\ \frac{1}{1+e^{-V_3W^4}}\end{pmatrix}$$

而由於預設『飯』出現在『吃』的上下文中，而『日』、『月』等不會出現在上下文中，所以要訓練神經網路迭代求解，使其輸出可近似為 $\begin{pmatrix}0 & 0 & 0 & 1\end{pmatrix}'$，如下：

$$\text{sigmoid}\left(\left(\boldsymbol{v}'_{吃}\mathbf{V}\right)\mathbf{W}\right)=\begin{pmatrix}\frac{1}{1+e^{-V_3W^1}}\\ \frac{1}{1+e^{-V_3W^2}}\\ \frac{1}{1+e^{-V_3W^3}}\\ \frac{1}{1+e^{-V_3W^4}}\end{pmatrix}\approx\begin{pmatrix}0\\ 0\\ 0\\ 1\end{pmatrix}$$

若是達成以上之效果，可以發覺 $e^{-V_3W^4}$ 要趨近於零，即表示內積 V_3W^4 越大越好，而 $e^{-V_3W^1}, e^{-V_3W^2}, e^{-V_3W^3}$ 要越大越好，即表示內積 V_3W^1, V_3W^2, V_3W^3 越小越好，故若取 V_3 為 $v_{吃}$ 壓縮後的新向量，則可以發現若是文字出現在前後文中，*word2vec* 將使出現在前後文的詞向量內積大，夾角則小，沒有出現在前後文的詞向量內積小，夾角則大。

故應用此法即可壓縮向量，並保持詞向量與詞向量間的相關性（可用餘弦相似度衡量）。

[11] $\text{sigmoid}(Z)=\frac{1}{1+e^{-Z}}$

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第四章

主成分分析在唐詩排行數據之應用

在本章節中，將應用主成分分析法至《唐詩排行榜》一書所提供的資料[1]，以探討唐詩在流通上的特質。本研究以 SAS 進行主成分分析。

第一節　計算流程與統計報表

在實際應用主成分分析之前，先查看《唐詩排行榜》數據的摘要統計值表，如表 4.1

表. 4.1.《唐詩排行榜》數據摘要統計值表

	古代選本	現代選本	歷代評點	論文引用	文史全詩	文史摘錄	網連總數
均數	8.29	19.33	17.35	6.53	3.52	1.95	97072.61
標準差	3.1150	6.9980	5.8610	11.0586	2.2895	1.5333	81432.3912
最小值	2	1	5	0	0	0	2360
第一四分位數	6	15	13	1	1	1	33975
中位數	8	21.5	16	3	4	2	84100
第三四分位數	10.25	24.25	21	9.25	5.25	3	131625
最大值	17	30	38	70	9	8	377000

可發覺其中的網路連結數量一項變數由於變數的變異較大，若是直接以變異共變異矩陣 **S** 進行主成分分析，將使主成分完全受到網路連結數量一項數據影響，故本章

[1] 見附錄 A 表 A.1

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

節之主成分分析將以相關係數矩陣 **R** 進行。

以下為相關係數矩陣 **R**

$$\mathbf{R} = \begin{pmatrix} 1 & -0.2417 & 0.0536 & -0.4414 & -0.2211 & -0.1175 & -0.3086 \\ -0.2417 & 1 & -0.3292 & 0.3107 & 0.6650 & 0.1399 & 0.4401 \\ 0.0536 & -0.3292 & 1 & 0.0258 & -0.3758 & 0.0829 & -0.0817 \\ -0.4414 & 0.3107 & 0.0258 & 1 & 0.1937 & 0.2875 & 0.5513 \\ -0.2211 & 0.6650 & -0.3758 & 0.1937 & 1 & -0.1536 & 0.2881 \\ -0.1175 & 0.1399 & 0.0829 & 0.2875 & -0.1536 & 1 & 0.1876 \\ -0.3086 & 0.4401 & -0.0817 & 0.5513 & 0.2881 & 0.1876 & 1 \end{pmatrix}$$

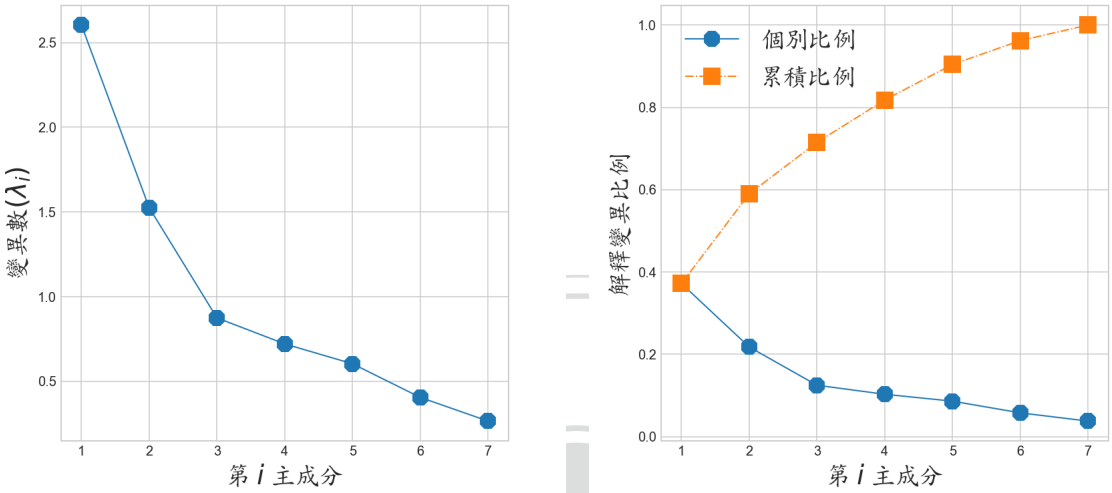
透過對 **R** 進行 eigenvalue 和 eigenvector 的求解，得計算結果如表 4.2與圖 4.1所示。

表. 4.2. 相關矩陣的特徵值與累積變異

相關矩陣的特徵值				
	特徵值	差異	比例	累計
1	2.60389130	1.07869670	0.3720	0.3720
2	1.52519460	0.65037591	0.2179	0.5899
3	0.87481870	0.15363505	0.1250	0.7148
4	0.72118364	0.11725699	0.1030	0.8179
5	0.60392665	0.19846281	0.0863	0.9041
6	0.40546384	0.13994257	0.0579	0.9621
7	0.26552127		0.0379	1.0000

據表 4.2與圖 4.1所示之結果，與Johnson and Wichern (2007) 書中所提及之判斷標準，取兩個主成分進行解釋，表 4.3為其特徵向量。

DOI:10.6814/THE.NCCU.MATH.003.2018.B01



(a) 主成分分析陡坡圖　　(b) 個別、累計解釋變異比例

圖. 4.1. 陡坡圖與累計解釋變異（主成分分析）

表. 4.3. 相關矩陣之特徵向量

	e_1	e_2
古代選本	-.353847	-.223294
現代選本	0.492736	-.232728
歷代評點	-.220831	0.504910
論文引用	0.422071	0.399596
文史全詩	0.423122	-.443463
文史摘錄	0.145565	0.490858
網路連結總數	0.452627	0.209202

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第二節　結果分析

一、 第一主成分

根據前一節的結果[2]，可以發覺第一主成分中的主成分係數有正有負，第一主成分呈現了 古代{古代選本 $(-.353847)$；歷代評點 $(-.220831)$} 與 現代{現代選本 (0.492736)；論文引用 (0.422071)；文史全詩 (0.423122)；文史摘錄 (0.145565)；網路連結總數 (0.452627)} 兩組分別屬於不同時代的文本的對比，可約略看出第一主成分呈現了某種程度的古今對立，而基於原本變數的意義[3]，似乎可將此主成分視為古今喜好的對比，亦即在第一主成分上排行越靠前者越得今人喜愛，越靠後者越得古人喜愛，而排行中段者古人今人之喜好程度相似。

以下將針對主成分值的排序與主成分係數進行進一步的解釋，由時間的角度出發，探討古人與今人的喜好呈現兩極對比的情形，並帶出前理解這一詮釋學的名詞，一步一步解釋古今人的差異，解釋在第一主成份值所呈現的排序有何意義，並為此主成分給出一合理的命名。

(一) 時代的差異

由之前所展示主成分係數的值可以發現，古人和今人在閱覽詩作的選擇上有所差異，而產生這種差異的最根本的原因當為古人和今人的生命經驗和環境有極大的不同，尤其是在近代的歷史當中，由於運用的語言不同（文言對比白話）、環境不同（自然環境對比於都市）、生活方式不同，導致了古人和今人在**認識各式事物上有所差異**，而這些差異所導致的，便是前理解的不同。

(二) 前理解

前理解 (pre-understanding) 是詮釋學的概念，是理解或運用某一事物的先決條件，若是由較為口語化的角度來看，可以直接的看成「進行理解的基礎」，理解這個行為本身，

[2] 見表 4.3
[3] 即各文本的選入數，或其引申意義公認度

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

具有歷史性與開放性，所有的理解者都存在著一定程度的「前理解」，這些前理解來自於生命經驗、生活環境、文化傳統、歷史因素等，這些事物造就了前理解，通過前理解人們得以理解新的事物，將新的視域和舊有的視域相融合，使理解得以發生。

(三) 今人與古人的前理解差異

古人與今人的差異有許多不同，在此舉出三個容易理解的面向：

1. 語言差異導致的語感不同
2. 生命經驗的不同
3. 歷史背景

由於以上三者的差異，導致各個時代的前理解都各不相同，接觸作品時用何種心情、讀者所具備的詮釋能力或是接受作品的能力都有所差異。

(四) 前理解差異在第一主成分排序上的呈現

為更進一步理解此解釋在第一主成分值排序上所表現的特徵，可對第一主成分值所代表的詩作進行觀察。查看以第一主成分值所排序的詩作名單可以發現其呈現了一些對比，以下為一些明確的例子：

1. 第一主成分排序前 15 首詩作：
    - 長篇敘事詩 (長恨歌、琵琶行)
    - 較為平易近人的詩作 (白居易、竹枝詞)
    - 愛情詩 (無題)
    - 描述社會現況 (春望)
2. 第一主成分排序後 15 首詩作：
    - 佛教詩作 (過香積寺)
    - 隱逸詩 (終南別業、歲暮歸南山)
    - 宮怨詩 (春宮怨)

通過這些對比，可以發現在第一主成分上排序較為靠前的 (亦即今人較為喜讀的) 主要從流通的角度出發，所偏好的目標為：

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

- 熟悉感
- 貼近生活
- 關乎愛情
- 口語化
- 朗朗上口
- 無語言上的隔閡與障礙

而古人在閱讀上偏好的目標為：

- 經典的角度
- 強調審美的高度
- 暗藏的美感
- 具有明悟心境

以最為明顯的通俗口語與否角度來看，現代人較易閱讀通俗口語的詩作，如白居易一類老嫗能解的作品，在現代人的語感底下較為容易閱讀，容易理解詩作的內涵與意義，但這在古代來說可能反而不是最好的類型，但古代最好的類型對現代人來說卻障礙太大，這障礙不只是文言白話不易相通，更是意境的難以領會(前理解差異)，如佛家道家那類不食人間煙火的情懷，只談個人一點明悟的作品(如王維的詩)，在和一閃而過的明悟與故事性的張力相較的情況下，現代人較能同理後者的情感。

由於以上所述，可發覺在第一主成分所呈現的差異是時代不同所導致的對比，故可將此第一主成分命名為**時代性差異**。

以下附上第一主成分前後各 15 首詩排序作為參考：

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表. 4.4. 第一主成分前 15 首詩排序

排名	詩名	作者	詩體	時代	第一主成分
1	長恨歌	白居易	七古	中	3.1283880
2	琵琶行	白居易	七古	中	2.5321073
3	無題（相見時難別亦難）	李商隱	七律	晚	1.9231508
4	蜀道難	李白	七古	盛	1.7705154
5	將進酒	李白	七古	盛	1.6322277
6	夢遊天姥吟留別	李白	七古	盛	1.5781794
7	燕歌行	高適	七古	盛	1.4624112
8	竹枝詞（楊柳青青）	劉禹錫	七絕	中	1.3776808
9	春望	杜甫	五律	盛	1.3171208
10	白雪歌送武判官歸京	岑參	七古	盛	1.3089031
11	使至塞上	王維	五律	盛	1.1624317
12	涼州詞（黃河遠上）	王之渙	七絕	盛	1.0859489
13	聞官軍收河南河北	杜甫	七律	盛	1.0606595
14	登鸛雀樓	王之渙	五絕	盛	1.0570674
15	山行	杜牧	七絕	晚	1.0472194

表. 4.5. 第一主成分後 15 首詩排序

排名	詩名	作者	詩體	時代	第一主成分
100	奉和賈至舍人早朝大明宮	岑參	七律	盛	-2.2385086
99	長安春望	盧綸	七律	中	-2.1216878
98	九日藍田崔氏莊	杜甫	七律	盛	-1.9764863
97	行經華陰	崔顥	七律	盛	-1.8544412
96	春宮怨	杜荀鶴	五律	晚	-1.6404981
95	長安秋望	趙嘏	七律	晚	-1.6089993
94	歲暮歸南山	孟浩然	五律	盛	-1.5764149
93	雲陽館與韓紳宿別	司空曙	五律	中	-1.2646284
92	與諸子登峴山	孟浩然	五律	盛	-1.2430792
91	過香積寺	王維	五律	盛	-1.2108974
90	終南別業	王維	五律	盛	-1.1497117
89	丹青引贈曹將軍霸	杜甫	七古	盛	-1.1286345
88	九日齊山登高	杜牧	七律	晚	-1.1193081
87	望薊門	祖詠	七律	盛	-1.1186295
86	晚次鄂州	盧綸	七律	中	-1.1050362

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

二、 第二主成分

根據第一節的結果[4]，可以發覺第二主成分中的主成分係數也有正有負，第二主成分呈現了 $\begin{cases} \text{古代選本} & (-.223294) \\ \text{現代選本} & (-.232728) \\ \text{文史全詩} & (-.443463) \end{cases}$ 與 $\begin{cases} \text{歷代評點} & (0.504910) \\ \text{論文引用} & (0.399596) \\ \text{文史摘錄} & (0.490858) \\ \text{網路連結總數} & (0.209202) \end{cases}$ 兩組原始變數的對比。

在係數上的呈現，若觀察以上兩者，位於左方的文本都是收錄詩文較為完整的作品，而右方的收錄方式則具有較為片段的性質（點評、摘錄等)，依據各類文本本身的意義，或可將此主成分視為詩文收錄完整性的差異，亦即在第二主成分上排行靠前者在各式文本的收集上多較為片段，而在第二主成分排行靠後者，在各式文本收集上多較為完整。

以下將針對主成分值的排序與主成分係數進行進一步的解釋，由詩文收錄完整性的角度出發，探討兩組變數呈現兩極對比的情形，探討此分立的合理性，並為此主成分給出一合理的命名。

(一) 詩文收錄完整性的差異

如前所述，此主成分呈現了各式選集收集詩文文本方式的完整與否，若是僅由文本完整與否為出發點，似乎不易觀察其與詩文流通的關聯，但若由資料本身的特性**選集的編纂目的與讀者需求**出發，則本主分所總結的意義將合理不少。

(二) 出版目的和讀者需求上的不同

此切入點是來自於各個文本本身的編纂目的，事實上對於一般人而言閱讀整首詩的機會並不多，反倒是片段的句子和名句較為容易進入一般人的視角，而對於部份的研究著作而言，也無須將整首詩作引入，更有可能在評論時基於書籍容量的考量而未將詩作完整錄入[5]，亦即這個主成分和編纂者要呈現的切入點、態度與動機有關，對詩作的需求可能為：

片段　　• 讀者只想要有一句美好的名言佳句、欣賞片段詩句的美感

[4] 見表 4.3

[5] 附帶一提，本主成分與詩作長度之相關係數為 0.573222

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

- 對空間的考量，詩作長度較長，無法完整收入
- 研究之用，對引用名句的需求勝於整首詩的完整收入

完整
- 由頭至尾完整的理解整首詩作的意圖、整首詩的意旨
- 對空間的考量，詩作長度較短，可完整收入

基於主成分係數所呈現的結果，可將之命名為**詩文收錄完整性**。

以下附上第二主成分前後各 15 首詩排序作為參考：

表. 4.6. 第二主成分前 15 首詩排序

排名	詩名	作者	詩體	時代	第二主成分
1	長恨歌	白居易	七古	中	3.5911591
2	琵琶行	白居易	七古	中	2.7726496
3	北征	杜甫	五古	盛	2.5178858
4	山石	韓愈	七古	中	2.1375170
5	錦瑟	李商隱	七律	晚	1.7097338
6	羌村三首（崢嶸赤雲西）	杜甫	五古	盛	1.4475133
7	蜀道難	李白	七古	盛	1.3922017
8	蜀相	杜甫	七律	盛	1.1844693
9	登高	杜甫	七律	盛	1.1739883
10	九日齊山登高	杜牧	七律	晚	1.1354542
11	九日藍田崔氏莊	杜甫	七律	盛	1.1177157
12	黃鶴樓	崔顥	七律	盛	1.0267584
13	旅夜書懷	杜甫	五律	盛	0.9510827
14	兵車行	杜甫	七古	盛	0.9103974
15	春宮怨	杜荀鶴	五律	晚	0.9012345

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表. 4.7. 第二主成分後 15 首詩排序

排名	詩名	作者	詩體	時代	第二主成分
100	逢入京使	岑參	七絕	盛	-1.8589729
99	從軍行	楊炯	五律	初	-1.8105401
98	竹枝詞（楊柳青青）	劉禹錫	七絕	中	-1.8090617
97	夜上受降城聞笛	李益	七絕	中	-1.6521207
96	野望	王績	五律	初	-1.6269562
95	從軍行（青海長雲）	王昌齡	七絕	盛	-1.4075329
94	燕歌行	高適	七古	盛	-1.3896635
93	滁州西澗	韋應物	七絕	中	-1.3848110
92	芙蓉樓送辛漸	王昌齡	七絕	盛	-1.2781033
91	送杜少府之任蜀州	王勃	五律	初	-1.2409502
90	石頭城	劉禹錫	七絕	中	-1.1959334
89	涼州詞（葡萄美酒）	王翰	七絕	盛	-1.1463851
88	泊秦淮	杜牧	七絕	晚	-1.0452370
87	次北固山下	王灣	五律	盛	-1.0317499
86	寒食	韓翃	七絕	中	-1.0092336

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第五章

因子分析在唐詩排行數據之應用

在本章節中，將實際應用因子分析法至唐詩排行榜所提供的資料[1]，以探討唐詩流通的因素。本研究以 SAS 進行因子分析，且為了方便比較採主成分法進行因子的抽取，在轉軸上為了解釋上的便利性選取 Varimax 轉軸法進行旋轉[2]。

第一節　計算流程與統計報表

依照表 4.1與第四章所述之理由，本章節將以相關係數矩陣 **R** 為基礎進行因子分析[3]。透過對 **R** 求解，得計算結果如表 5.1與圖 5.1所示。

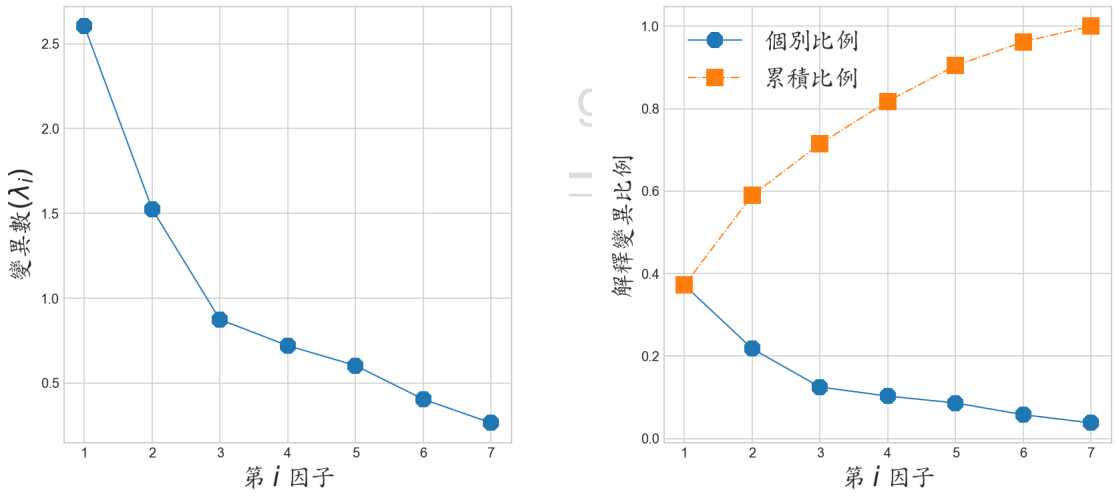
據表 5.1與圖 5.1所示之結果，與Johnson and Wichern (2007) 書中所提及之判斷標準，取兩個因子進行解釋，表 5.2為特徵向量，表 5.3為其因子負荷與解釋變異。

[1]見附錄 A 表 A.1
[2]亦曾以 Harris-Kaiser 轉軸法進行旋轉，並未得到較易解釋之因子，且因子間相關性僅有 0.2，故後續分析仍採 Varimax 轉軸法進行
[3]相關係數矩陣 **R** 與第四章相同

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表. 5.1. 相關矩陣的特徵值與累積變異

	特徵值	差異	比例	累計
1	2.60389130	1.07869670	0.3720	0.3720
2	1.52519460	0.65037591	0.2179	0.5899
3	0.87481870	0.15363505	0.1250	0.7148
4	0.72118364	0.11725699	0.1030	0.8179
5	0.60392665	0.19846281	0.0863	0.9041
6	0.40546384	0.13994257	0.0579	0.9621
7	0.26552127		0.0379	1.0000



(a) 因子分析陡坡圖　　(b) 個別、累計解釋變異比例

圖. 5.1. 陡坡圖與累計解釋變異（因子分析）

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表. 5.2. 相關矩陣的特徵向量

	e_1	e_2
古代選本	-.353847	-.223294
現代選本	0.492736	-.232728
歷代評點	-.220831	0.504910
論文引用	0.422071	0.399596
文史全詩	0.423122	-.443463
文史摘錄	0.145565	0.490858
網路連結總數	0.452627	0.209202

表. 5.3. 因子型式 (Factor Patten)

	第一因子	第二因子
現代選本	0.79511	-0.28742
網路連結總數	0.73039	0.25836
文史全詩	0.68277	-0.54767
論文篇數	0.68108	0.49350
古代選本	-0.57099	-0.27577
歷代評點	-0.35635	0.62356
文史摘錄	0.23489	0.60620
解釋變異	2.6038913	1.5251946

表. 5.4. 最終公因子變異數估計值: 總計 = 4.129086

古代選本	現代選本	歷代評點	論文篇數	文史全詩	文史摘錄	網路連結總數
0.40207323	0.71480356	0.51580641	0.70740644	0.76612473	0.42265740	0.60021414

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表. 5.5. Varimax 旋轉後因子型式 (Varimax-Rotated Factor Patten)

	第一因子	第二因子
論文篇數	0.83886	0.06101
網路連結總數	0.72517	0.27265
文史摘錄	0.57007	-0.31254
古代選本	-0.61443	-0.15670
文史全詩	0.16971	0.85867
現代選本	0.42324	0.73189
歷代評點	0.12890	-0.70654
解釋變異	2.1566091	1.9724768

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第二節　結果分析

一、 第一因子

根據前一節的結果[4]，可以發覺旋轉後的第一因子呈現了　古代選本 (-0.61443)

與 { 論文篇數 (0.83886)；網路連結總數 (0.72517)；文史摘錄 (0.57007)；現代選本 (0.42324) } 兩組變數的對比。

(一) 係數上的呈現

在係數的表現上，可藉由此對比發覺在第一因子前端之詩具有1. 研究者眾多 2. 易進入一般民眾視角 3. 文學史中摘錄較多 4. 古代選本少　幾種特性，但卻不易直接解釋此因子所隱含之意義，因此在解釋此因子時，可先參考其排序。

(二) 排序上的呈現

由排序出發，得知此排序的前 15 首詩作展現出和歷史故事的相關性，一如長恨歌、琵琶行、蜀道難、石壕吏、聞官軍收河南河北、春夜喜雨、北征等詩作，背後都有一個明顯的歷史事件為其背景，此歷史事件即是安史之亂。

安史之亂這一事件對唐朝來說可謂是十分重大的影響，整個大唐國勢自此由盛轉衰，再不復以往榮景，而在唐詩創作的觀點上，可說是由盛唐落入中唐而至晚唐的一大轉折，安史之亂除了國家經歷戰事摧折，家園滿目瘡痍，這類國破家亡的情感與歷史故事之外，還包含了一個美麗的愛情故事[5]。

除了這類的故事外，在第一因子前半部的排序上還有大量的杜甫詩作，杜甫的詩作被稱為詩史[6]，其中有強烈而且豐富的憂國憂民情感，和歷史有強烈的連結，深刻的刻畫了唐代歷史的重要事件。總體而言，第一因子的前端和歷史事件的相關性頗高。

[4] 見表 5.5
[5] 指楊貴妃與唐玄宗間，其相關作品繁多
[6] 杜詩善於描寫當時歷史實況，反映唐代由盛轉衰的現況，有「詩史」之稱

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

(三) 由歷史性推向故事性

從另一角度來說，歷史性高，同時也代表了詩作具備相當程度的故事性。

由故事性的角度來看，若是詩作的背後存在一個故事，對現代讀者而言其可觀性更為強烈。除了故事中美麗的情感、浪漫的情懷能動人心魄外，歷史的沉浸感也是引起讀者共鳴的重要成分，若是詩作中蘊含著故事、歷史中曾發生過的事件，其足以引起讀者沉浸其中並想像其中的情景，亦是一種強烈的共鳴。這般具備了故事性共鳴的詩作，較能進一步引起現代讀者的閱讀興趣。

而若是談論個人一剎那的感動、明悟、閑看花落之類的情感，則現代讀者的共鳴會較少，若是回到第一主成分中的前理解觀點來說，便是由於生活環境，生命經驗的不同，導致了現代人對於此類的情懷共鳴不大。

(四) 敘事群和讀者群相互增長

由上可知，第一因子的前端如同歷史的共鳴箱，其中是一些可以引起讀者共鳴的詩作，這些引起他人共鳴的作品也會被寫為其他不同的文類、戲劇、小說等。而在某故事相關文本眾多等情況下，便產生了一個相關作品的集合，謂之敘事群[7]，在這樣的情況下，作品量和類型會相互增長，讀者群讀的不只是詩，還有小說、戲劇，而其他文類的讀者也會對詩作有所關注，也在無形中推動了詩的傳播與流通，敘事群龐大則讀者群亦增長，讀者群增長，二次創作增多，則敘事群也隨之增大。

由此反推，古代選本較少的原因則可能是由於距離事件發生的時間尚短，相關作品都還未產出，所以古代選本所選中的還不會那麼多，但隨著時間流逝，這個故事相關的二次創作便隨之增長，現在的文本亦由於龐大的敘事群而對這首詩多所關注，敘事群增長，而越近代的文本亦增長。歷史背景豐富、故事性高、敘事群龐大，也將使論文篇數和網路連接的數量上升，前者是由於歷史研究，後者和故事性高相關。而這類描述歷史的作品通常表現形式為長篇的敘事詩，在做文學史相關研究時，可能由於篇幅所限，無法完全收錄進文學史之中，於是文學史摘錄的數量便跟著提升了。

[7]由同一事件、故事所展開的，以不同視角、不同主角、不同立場所發展出來的各式作品

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

(五) 第一因子命名

綜合以上四點，詩作背後的歷史是否豐富，影響了後代文本的創作數量，也影響了現代人是否對此感興趣，所以可將之命名為**歷史性強弱**。

以下附上第一因子前後各 15 首詩排序作為參考：

表. 5.6. 第一因子前 15 首詩排序

	詩名	作者	詩體	時代	第一因子值
1	長恨歌	白居易	七古	中	4.7059394991
2	琵琶行	白居易	七古	中	3.7226704574
3	蜀道難	李白	七古	盛	2.2510738315
4	北征	杜甫	五古	盛	1.8002119920
5	山石	韓愈	七古	中	1.7625884142
6	錦瑟	李商隱	七律	晚	1.6576145672
7	夢遊天姥吟留別	李白	七古	盛	1.6047372363
8	無題（相見時難別亦難）	李商隱	七律	晚	1.4788195372
9	登高	杜甫	七律	盛	1.3431300460
10	石壕吏	杜甫	五古	盛	1.2165129123
11	將進酒	李白	七古	盛	1.2085366687
12	聞官軍收河南河北	杜甫	七律	盛	1.0370952814
13	春夜喜雨	杜甫	五律	盛	0.9703247957
14	兵車行	杜甫	七古	盛	0.9524573381
15	涼州詞（黃河遠上）	王之渙	七絕	盛	0.8977595577

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表. 5.7. 第一因子後 15 首詩排序

排名	詩名	作者	詩體	時代	第一因子值
100	雲陽館與韓紳宿別	司空曙	五律	中	-1.4552492320
99	奉和賈至舍人早朝大明宮	岑參	七律	盛	-1.3361798891
98	逢入京使	岑參	七絕	盛	-1.3058898627
97	行經華陰	崔顥	七律	盛	-1.2655876039
96	長安春望	盧綸	七律	中	-1.2650598142
95	歲暮歸南山	孟浩然	五律	盛	-1.2248156664
94	次北固山下	王灣	五律	盛	-1.1899595667
93	和晉陵陸丞相早春遊望	杜審言	五律	初	-1.1619478326
92	野望	王績	五律	初	-1.1484618336
91	寒食	韓翃	七絕	中	-1.1038831339
90	題破山寺後禪院	常建	五律	盛	-1.0811348579
89	晚次鄂州	盧綸	七律	中	-1.0003301963
88	夜上受降城聞笛	李益	七絕	中	-0.9113436600
87	石頭城	劉禹錫	七絕	中	-0.8853000118
86	從軍行	楊炯	五律	初	-0.8832359518

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

二、 第二因子

根據前一節的結果[8]，可以發覺旋轉後的第二因子呈現了 歷代評點 (-0.70654)

與 $\begin{cases} \text{文史全詩} & (0.85867) \\ \text{現代選本} & (0.73189) \end{cases}$ 兩組變數的對比。

(一) 係數上的呈現

透過觀察因子負荷的結果可以發覺位於正相關一側的文史全詩、現代選本皆能作為教學之用，是作為唐詩教材時很重要的一項知識來源，是能用以建構嚴謹詩學知識的文本，如聲韻學的知識、美學知識、類型學知識、語言學、符號學之類的知識，是一個專業的學門，這和建構一個賞析是完全不同的類型；而位於另一側的歷代點評則是結構較為散亂的個人性書寫，著重在「點」，在創作時為近於賞析，較無系統架構，具有較多的隨機性，是欣賞取向的文本。

(二) 兩方文本的不同之處與排序上的呈現

以此角度出發，教材一類所包含的詩作應是較為經典、典範或核心之作，為教授唐詩所必選的內容，可說是每個詩派或某個詩人最具代表性的作品，以下舉幾首詩作作為範例：

- 《無題》李商隱

  相見時難別亦難，東風無力百花殘，

  春蠶到死絲方盡，蠟炬成灰淚始乾，

  曉鏡但愁雲鬢改，夜吟應覺月光寒，

  蓬山此去無多路，青鳥殷勤為探看。

  這首詩是意識流的書寫，為探討這一門類的典範。

- 《竹枝詞》劉禹錫

  楊柳青青江水平，聞郎江上唱歌聲，

  東邊日出西邊雨，道是無晴還有晴。

  這首詩是劉禹錫所創作的民謠類型作品，是其傳世名篇。

[8] 見表 5.5

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

其他如李白《將進酒》、王之渙《登鸛雀樓》等作品，皆是某一門類的經典作品。而位在排序後半段的詩作具有歷代點評多，而文學史全詩與現代選本少的特性，亦即在詩作的探討上可視為較不是用於建立詩學知識的作品，較非某門某派或某詩人的代表之作，不是詩學必教的典範之作。

(三) 第二因子命名

由於此因子呈現了詩作在各個詩人或者詩派的經典之作、代表名作，而此作品的代表性主要由學術領域出發，來論斷一首詩作是否經典，故將其稱為**詩學經典性**。

以下附上第二因子前後各 15 首詩排序作為參考：

表. 5.8. 第二因子前 15 首詩排序

排名	詩名	作者	詩體	時代	第二因子值
1	竹枝詞（楊柳青青）	劉禹錫	七絕	中	2.2712143204
2	燕歌行	高適	七古	盛	2.0049014254
3	從軍行	楊炯	五律	初	1.6230885981
4	白雪歌送武判官歸京	岑參	七古	盛	1.5969368465
5	從軍行（青海長雲）	王昌齡	七絕	盛	1.5413863854
6	登鸛雀樓	王之渙	五絕	盛	1.4197458186
7	夜上受降城聞笛	李益	七絕	中	1.3923693976
8	山行	杜牧	七絕	晚	1.3619306411
9	送杜少府之任蜀州	王勃	五律	初	1.3379889738
10	逢入京使	岑參	七絕	盛	1.3306643724
11	芙蓉樓送辛漸	王昌齡	七絕	盛	1.3294332158
12	春望	杜甫	五律	盛	1.2848115605
13	春曉	孟浩然	五絕	盛	1.2394958793
14	無題（相見時難別亦難）	李商隱	七律	晚	1.2295265905
15	過故人莊	孟浩然	五律	盛	1.1684842163

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表. 5.9. 第二因子後 15 首詩排序

排名	詩名	作者	詩體	時代	第二因子值
100	九日藍田崔氏莊	杜甫	七律	盛	-2.1278694362
99	奉和賈至舍人早朝大明宮	岑參	七律	盛	-1.8887385590
98	長安春望	盧綸	七律	中	-1.7918214417
97	北征	杜甫	五古	盛	-1.7758487096
96	春宮怨	杜荀鶴	五律	晚	-1.7458896441
95	九日齊山登高	杜牧	七律	晚	-1.5894749888
94	長安秋望	趙嘏	七律	晚	-1.5064190363
93	終南別業	王維	五律	盛	-1.3918265257
92	行經華陰	崔顥	七律	盛	-1.3761724451
91	羌村三首（崢嶸赤雲西）	杜甫	五古	盛	-1.3583409822
90	過香積寺	王維	五律	盛	-1.3175753199
89	山石	韓愈	七古	中	-1.3103535399
88	旅夜書懷	杜甫	五律	盛	-1.2488585516
87	望薊門	祖詠	七律	盛	-1.2297011479
86	獨坐敬亭山	李白	五絕	盛	-1.2063006745

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第六章

詞嵌入法在唐詩排行數據之應用

在計算出主成分和因子後，由於主成分與因子中似乎有某些指標和詩文文本內容具有一定程度的相關性，為探討此一特性是否為真，可試著應用資料科學領域的詞嵌入法，計算詩文文本在歐氏空間的向量，以進一步探討詩文之間的相似性，藉此輔助佐證之前所得結果。

第一節　唐詩 100 首向量之建立

由於詩向量之訓練是基於小型的神經網路，在少量文本的情況下無法取得良好成果，故本文以Gao (2018) 所提供的全唐詩詩文為資料來源，進行以下處理：

1. 將全唐詩之文本內容進行簡繁轉換、文字編碼、通同字與斷字的資料清理

由於網路上所取得之文本常有簡繁轉換、編碼轉換與通同字轉換的問題，這在模型訓練之過程中，易使模型誤判，故在模型訓練之前，需先對文字內容進行處理，並在此同時將詩詞內容切斷為單純的一字一詞[1]，並移除標點符號。

2. 以 Gensim 套件所提供的詞嵌入模型 Doc2Vec[2] 進行文本向量的訓練

Řehůřek and Sojka (2010) 對於詞嵌入法提供了一個簡便的套件 (Gensim) 的解釋及網址，本研究之詞嵌入模型主要基於此套件，程式碼附於附錄，隱藏層節點數為 250 個。

[1] 在此之前曾以斷詞工具對詩詞內容進行斷詞處理，但由於詩詞本身與現代語法相去甚遠，單字所蘊含之意義較多，斷詞後所訓練之向量無法取得良好效果，故以斷字進行後續分析

[2] 見Le and Mikolov (2014) 為普通詞嵌入的進階版本，可以轉換整篇文章為一向量，其原理與詞嵌入相同

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

經前述處理，可以使全唐詩之每一首詩皆轉換為一個與之對應的向量。而在建立完成每一首詩之向量後，自模型中取出《唐詩排行榜》100 首詩對應的向量，以進行後續分析。

第二節　詩間相似度之計算

一、 以一首詩為基準計算相似度

為衡量兩詩所代表之向量是否相似，在此以餘弦相似度[3]計算兩向量之夾角大小，作為詩向量與詩向量之間的相似程度，即詩間相似度。而為了衡量詩間相似度和主成分值與因子值之排序的相關性，可取每一指標的第一首詩作為基準，計算每一首詩之相似度，進而衡量此相似度排序是否和原指標排序相關。

以下以第一主成分為例:

以 $\boldsymbol{p}_{i,j}$ 表示第 i 個指標中排序第 j 首詩之對應向量，其中 $i = 1, 2, \ldots, 4$，分別對應第一、第二主成分、第一、第二因子，例如 $\boldsymbol{p}_{1,1}$ 代表第一主成分中，排序最高者**長恨歌**對應之向量。若以 $\boldsymbol{p}_{1,1}$ 為基準，可計算出向量

$$V_{\boldsymbol{p1,1}} = \begin{pmatrix} \cos(\boldsymbol{p}_{1,1}, \boldsymbol{p}_{1,1}) \\ \cos(\boldsymbol{p}_{1,1}, \boldsymbol{p}_{1,2}) \\ \vdots \\ \cos(\boldsymbol{p}_{1,1}, \boldsymbol{p}_{1,100}) \end{pmatrix}$$

作為 100 首詩相似度排序, 同理亦可得其餘指標上以第一首詩為基準之 $V_{\boldsymbol{p2,1}}, V_{\boldsymbol{p3,1}}, V_{\boldsymbol{p4,1}}$。

二、 以多首詩為基準計算加權相似度

由於第一主成分、第二主成分與第一因子的前幾首詩相同，為了凸顯各指標間的差異可多取幾首詩作為基準，分別計算不同基準之相似度，再進一步取加權，作為詩

[3]在歐氏空間中，兩向量 $\boldsymbol{v}_1, \boldsymbol{v}_2$ 之方向的差異 (夾角 θ 大小) 可以用餘弦值來衡量，即 $\cos(\theta) = \frac{< \boldsymbol{v}_1, \boldsymbol{v}_2 >}{||\boldsymbol{v}_1|| ||\boldsymbol{v}_2||}$，為方便表示，在此以 $\cos(\boldsymbol{v}_1, \boldsymbol{v}_2)$ 表示兩向量 $\boldsymbol{v}_1, \boldsymbol{v}_2$ 間之餘弦相似度

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

的加權相似度。

以下說明以第一主成分為例:

如前所示亦可計算以第 i 首詩為基準之相似度排序向量 $V_{p1,i}$，故以前三首詩作為基準之加權相似度定為 $3V_{p1,1}+2V_{p1,2}+V_{p1,3}$，取前四首詩作為基準之加權相似度定為 $4V_{p1,1}+3V_{p1,2}+2V_{p1,3}+V_{p1,4}$，$\cdots$，依此類推。

第三節　詞嵌入法與主成分分析法及因子分析法結果之相關性

在計算出相似度排序後，為進一步研究其相似性排序是否能和原指標的排序(指前二章所得之主成分與因子之排序，由大到小) 有一致性，以 Spearman 的秩相關係數[4]作為衡量排序相關性的標準，為確認是否存在局部的相關性，分別對各依原指標排序的前 20 首、前 50 首、前 100 首詩與其相似度做秩相關之計算，所得之結果如下:

表. 6.1. 依原指標排序與取前 5 首詩為基準之相似度排序的秩相關

	依原指標之排序取前 n 首詩		
原指標	$n=20$	$n=50$	$n=100$
第一主成分	0.492208*	0.387602**	0.328521**
第二主成分	0.507792*	0.161810	0.099586
第一因子	0.685714**	0.521538**	0.300582**
第二因子	0.511688*	0.295837*	0.349415**

*　表示在 0.05 的信心水準下有顯著相關
**　表示在 0.01 的信心水準下有高度顯著相關

由表 6.1得知，原指標內的第一主成分、第一因子與第二因子分別對加權相似度有顯著的相關性，表示這三者在相似性排序上，都和原指標有一致性，唯有第二主成分不具備顯著的相關性，若是回顧之前在第二主成分的解釋，則可以發覺其主要是和詩文收錄在各式文本的完整性與否有關（且其與詩文長度之相關係數為 0.573222)，和詩文相似性本身較無關係。

[4]rank correlation coefficient, 常用以計算兩排序指標是否具備一致性

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

而第一主成分所代表的是時代性差異，亦即當代與古代閱眾所喜好的文本差異，第一因子代表歷史性強弱，敘事詩一類的作品在內容上偏向淺顯易懂，會有相近的性質，第二因子所代表的經典性在 王宏林 (2012) 所言『不同社會都有特定的審美思潮和意識形態，包括帝王喜好、士人心態和時代風尚，這些均會對唐詩經典的建構產生重要影響。』亦即詩學經典的建構有部分基於群眾的喜好。這三者都屬於與內容相關的指標，故這三者能呈現出與相似性的顯著秩相關並不讓人意外。

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

第七章

結論

依據前幾章所得之結果，我們在唐詩流通傳播上，找出各兩種不同的指標及原因，如**時代性差異**指標，其呈現古、今人喜好唐詩之差異，如**詩文收錄完整性**指標，其呈現唐詩在各式選集的選擇上受到編纂需求及詩文長度的影響，如**歷史性強度**因子，其代表歷史性、故事性因素影響古、今人選讀的唐詩，又或者如**詩學經典性**因子，其呈現的唐詩學術性的強度。亦即唐詩流通度可用時代性差異與詩文收錄完整性兩指標來展現，同時唐詩在流傳散佈上受歷史性強度與詩學經典性的影響。而詞嵌入法如我們預期的結果，呈現了第一主成分、第一因子與第二因子與其分別對應之詩文相似度排序具有顯著的一致性。而第二主成份雖與其對應之詩文相似度排序沒有有顯著的一致性，但由於第二主成份主要與詩文長度有關，故此結果亦與預期相符。

而值得一提的是，無論是在主成分分析或是因素分析所得出之結果都可以發覺所總結的指標都有兩個且都具有對比的性質，亦即單一指標較無法完整探討唐詩的流通性。若改由另外一個角度出發或許可以打破一些迷思，亦即在探討唐詩流通或唐詩的熱門度時，應考慮多個因素的影響，不能簡單地將其組合成單一指標。

而在總結時代性時受限於資料來源與變數特性，在命名時代性時只能粗略的分別為古今之對比，實際上若能有更細緻的資料進行分析，或許可以建立更精細的結論，如呈現各朝代的前理解不同（唐代重抒情、宋代重議論等）。

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

最後，主成分分析與因子分析這一類降維總結、尋找原因的方式，在應用於唐詩以探尋其代表意義時，須仰賴對唐詩有深刻認識的專家提供意見，方能尋得其深藏於數據之下的真義，若無相關背景知識，在實作上可能有一定難度。而在資料的收集上又是另一困難（尤其是古籍的彙整計量數據），所幸現今新出版的文本、書籍、論文等，多半會在網路上留下紀錄，可以用相對人工計量容易的方式進行資料收集，而古籍一類的資料則已有人著手收集，在未來或有機會取得更為全面、更有用的資料進行分析。若在未來能得到更為全面的資料、更多種類的變數，我們所用之多變量分析的方法將能得到更進一步的展現，探索更多有趣的面向。

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

附錄 A

唐詩排行榜數據

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表. A.1. 唐詩排行榜數據

排名*	詩名**	作者	古代選本	現代選本	歷代評點	論文篇數	全錄	摘錄	網路連結	綜合指標
1	黃鶴樓	崔顥	17	24	38	1	5	4	135600	0.8153
2	送元二使安西	王維	13	27	21	3	4	4	108700	0.6256
3	涼州詞（黃河遠上）	王之渙	10	28	17	26	6	3	101000	0.5924
4	登鸛雀樓	王之渙	10	30	15	19	7	1	102700	0.5802
5	登岳陽樓	杜甫	11	26	23	4	2	2	85700	0.5778
6	登柳州城樓	柳宗元	9	25	26	3	4	3	12590	0.5774
7	臨洞庭湖贈張丞相	孟浩然	12	23	20	1	6	1	63580	0.5748
8	題破山寺後禪院	常建	15	16	19	0	5	2	44300	0.5678
9	送杜少府之任蜀州	王勃	12	28	13	10	6	1	118700	0.5658
10	蜀道難	李白	6	25	23	36	6	3	228000	0.5635

表 A.1 接續自前頁

排名*	詩名**	作者	古代選本	現代選本	歷代評點	論文篇數	全錄	摘錄	網路連結	綜合指標
11	次北固山下	王灣	16	23	13	2	2	1	77100	0.5605
12	楓橋夜泊	張繼	13	27	15	11	1	1	203700	0.5551
13	終南山	王維	13	22	18	3	3	2	102400	0.5533
14	長信秋詞（奉帚平明）	王昌齡	14	16	19	1	3	5	33100	0.5472
15	登高	杜甫	7	26	25	10	3	3	236400	0.5431
16	泊秦淮	杜牧	11	26	15	1	6	2	93900	0.5415
17	江雪	柳宗元	7	25	22	6	7	2	152300	0.5344
18	西塞山懷古	劉禹錫	5	23	28	3	8	0	30400	0.5291
19	烏衣巷	劉禹錫	11	25	16	1	3	2	82100	0.5194
20	滁州西澗	韋應物	12	26	10	0	5	2	64900	0.5104
21	夜雨寄北	李商隱	7	27	20	5	4	2	175500	0.5072
22	燕歌行	高適	8	28	12	12	8	0	216400	0.5019
23	琵琶行	白居易	4	22	19	62	4	4	290000	0.5019
24	觀獵	王維	11	22	18	2	1	2	123400	0.501
25	出塞	王昌齡	8	29	13	3	6	3	131600	0.4993
26	過故人莊	孟浩然	10	25	11	9	6	2	104700	0.497
27	長恨歌	白居易	2	22	20	70	4	5	357000	0.4897
28	聞官軍收河南河北	杜甫	3	28	22	23	5	1	72500	0.4888
29	石壕吏	杜甫	7	24	15	31	4	3	88900	0.4886
30	早發白帝城	李白	5	27	20	10	5	2	140400	0.4868
31	靜夜思	李白	9	18	22	7	3	0	220600	0.4858
32	咸陽城東樓	許渾	9	13	26	1	3	1	15030	0.4793
33	山居秋暝	王維	8	25	15	4	5	2	102700	0.4784
34	錦瑟	李商隱	5	18	25	17	4	4	186900	0.476

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表 A.1 接續自前頁

排名*	詩名**	作者	古代選本	現代選本	歷代評點	論文篇數	全錄	摘錄	網路連結	綜合指標
35	寒食	韓翃	12	20	13	0	3	1	41000	0.4748
36	石頭城	劉禹錫	9	22	16	2	5	0	40100	0.4736
37	鹿柴	王維	9	21	15	0	6	1	130000	0.4677
38	春江花月夜	張若虛	7	23	16	10	5	3	135800	0.4676
39	赤壁	杜牧	8	24	16	3	3	3	200100	0.4664
40	黃鶴樓送孟浩然之廣陵	李白	8	23	17	3	3	1	135600	0.4601
41	旅夜書懷	杜甫	10	14	20	2	1	4	62400	0.4533
42	馬嵬	李商隱	10	15	18	3	3	1	37600	0.4499
43	和晉陵陸丞相早春遊望	杜審言	11	15	15	1	4	1	4079	0.4488
44	蜀相	杜甫	4	23	26	10	0	1	108400	0.4472
45	望薊門	祖詠	8	13	26	1	1	1	43200	0.4461
46	古意呈補闕喬知之	沈佺期	7	20	19	0	4	2	2360	0.4455
47	獨坐敬亭山	李白	9	14	22	0	1	3	68300	0.4449
48	九月九日憶山東兄弟	王維	11	16	16	1	1	2	118700	0.4438
49	夢遊天姥吟留別	李白	4	27	13	23	5	4	143600	0.44
50	隋宮（紫泉宮殿鎖煙霞）	李商隱	6	17	25	0	2	2	28200	0.4396
51	奉和賈至舍人早朝大明宮	岑參	12	2	27	0	0	0	3152	0.4383
52	春宮怨	杜荀鶴	10	6	26	0	1	2	11030	0.437
53	望嶽	杜甫	7	22	17	9	1	3	180500	0.4363
54	賦得古原草送別	白居易	7	23	16	1	3	2	79700	0.4354
55	逢入京使	岑參	10	20	10	0	6	0	28100	0.4354
56	春望	杜甫	7	24	12	12	5	0	304200	0.4343
57	九日齊山登高	杜牧	8	11	25	1	1	3	27400	0.4323
58	閨怨	王昌齡	10	18	11	2	4	4	43100	0.4323

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表 A.1 接續自前頁

排名*	詩名**	作者	古代選本	現代選本	歷代評點	論文篇數	全錄	摘錄	網路連結	綜合指標
59	終南別業	王維	12	4	21	0	2	3	135500	0.431
60	無題（相見時難別亦難）	李商隱	2	24	18	18	8	1	260600	0.4291
61	江南春絕句	杜牧	8	22	14	3	1	4	46500	0.4254
62	春曉	孟浩然	7	21	13	6	6	0	196500	0.4214
63	九日藍田崔氏莊	杜甫	8	4	32	0	0	0	6030	0.4205
64	商山早行	溫庭筠	5	18	22	1	4	1	40300	0.4182
65	使至塞上	王維	6	24	14	0	4	2	333900	0.415
66	夜上受降城聞笛	李益	7	23	11	2	6	0	21000	0.4132
67	丹青引贈曹將軍霸	杜甫	7	11	25	2	2	0	15380	0.4126
68	長安秋望	趙嘏	12	7	18	0	0	3	10320	0.4118
69	山行	杜牧	5	27	10	6	6	2	107200	0.4099
70	鳥鳴澗	王維	8	21	12	7	2	2	114200	0.409
71	涼州詞（葡萄美酒）	王翰	9	21	9	2	4	1	85700	0.4076
72	山石	韓愈	3	24	18	8	1	8	51100	0.4043
73	歲暮歸南山	孟浩然	13	6	15	0	1	2	16380	0.4031
74	兵車行	杜甫	4	20	18	4	4	5	93600	0.4016
75	芙蓉樓送辛漸	王昌齡	8	22	7	3	6	2	86600	0.4016
76	從軍行（青海長雲）	王昌齡	6	27	7	3	6	2	33300	0.3996
77	白雪歌送武判官歸京	岑參	5	28	7	10	6	2	122700	0.3986
78	長安春望	盧綸	11	1	25	0	0	0	15740	0.3982
79	晚次鄂州	盧綸	8	12	20	0	2	0	9840	0.3946
80	野望	王績	11	16	5	0	6	2	34200	0.3944
81	賈生	李商隱	6	23	11	1	4	3	51000	0.3898
82	終南望餘雪	祖詠	10	12	14	1	1	3	76530	0.3875

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

表 A.1 接續自前頁

排名*	詩名**	作者	古代選本	現代選本	歷代評點	論文篇數	全錄	摘錄	網路連結	綜合指標
83	將進酒	李白	6	21	10	3	6	3	377000	0.387
84	秋興八首之一（玉露凋傷）	杜甫	7	16	14	10	4	0	105300	0.3861
85	登樓	杜甫	7	9	23	5	1	1	113900	0.3838
86	月夜	杜甫	5	21	16	6	2	1	82500	0.3824
87	北征	杜甫	3	12	24	13	2	6	131700	0.379
88	過香積寺	王維	11	7	16	0	0	3	91900	0.3784
89	竹枝詞（楊柳青青）	劉禹錫	3	26	9	7	9	0	95700	0.3773
90	從軍行	楊炯	7	21	7	0	7	1	32000	0.377
91	與諸子登峴山	孟浩然	10	11	15	0	0	2	21750	0.3757
92	春夜喜雨	杜甫	4	22	15	17	1	2	139500	0.3752
93	送魏萬之京	李頎	8	13	16	0	2	1	18070	0.3735
94	早雁	杜牧	6	16	15	1	5	0	39950	0.3713
95	雁門太守行	李賀	4	20	13	10	6	1	52100	0.3713
96	行經華陰	崔顥	10	4	21	0	0	0	7770	0.3689
97	秋登宣城謝眺北樓	李白	10	9	14	11	0	1	48360	0.3665
98	登金陵鳳凰台	李白	6	12	21	2	1	1	54200	0.3661
99	雲陽館與韓紳宿別	司空曙	11	8	13	0	2	0	28040	0.3656
100	羌村三首（崢嶸赤雲西）	杜甫	6	15	16	10	0	5	23980	0.3652

* 排名依據唐詩排行榜一書所建立之綜合指標排序

** 詩名後括號以部份詩文內容區別同名之不同詩作

DOI:10.6814/THE.NCCU.MATH.003.2018.B01

附錄 B

詞嵌入法程式碼

自原始檔中讀取詩文

```
# -*- coding: utf-8 -*-
import json
import numpy as np
import pandas as pd
folder = 'D:/Morokei-formove_Version_3/字元分析/word/chinese-poetry-master/TANG_Doc/json/'
"""## 排列各類index"""
author       = []
title        = []
author_title = []
paragraphs   = []
for text in np.arange(0,58):
    with open(folder + 'poet.tang.'+str(text*1000)+'.json','r', encoding = 'utf8') as json_file:
        json_data = json.load(json_file)
        for item in json_data:
            author       = author       + [item['author']]
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
            title        = title        + [item['title']]
            author_title = author_title + [item['author'] + '_'
                + item['title']]
            paragraphs   = paragraphs   + [item['paragraphs']]
pd_author_title = pd.DataFrame([author,title,author_title,
    paragraphs]).T
pd_author_title.columns = ['author', 'title', 'author_title', '
    paragraphs']
pd_author_title.to_excel('author_title.xlsx')
"""## 全斷"""
 output = open('TS_seg_all_part.txt','w', encoding = 'utf8')
  for text in np.arange(0,58):
      with open(folder + 'poet.tang.'+str(text*1000)+'.json','r'
          , encoding = 'utf8') as json_file:
          json_data = json.load(json_file)
          for item in json_data:#每一個dict檔
              p =""
              for j in item['paragraphs']:#dict 裡面的句子
                  for k in j:
                      if k not in '，。·.？（《》）1234567890
                          /\-[]{}    「」1234567890':
                          p = p + "␣" + k
              output.write(p +'\n')
 output.close()
```

自原始檔中讀取唐詩 100 首詩文

```
# -*- coding: utf-8 -*-
"""以舊有標題搜尋
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
    1.先找完全一樣的      2.再人工模糊的尋找
"""
import json
import numpy as np
import pandas as pd
folder = 'D:/Morokei-formove_Version_3/字元分析/word/chinese-poetry-master/TANG_Doc/'
DATAS_OR = pd.read_excel("D:/Morokei-formove_Version_3/兩分析結果.xlsx")
author_title = pd.read_excel(folder + 'author_title.xlsx')
def search(author, title):
    author_ = (author_title['author'] == str(author)).values
    title_  = (author_title['title'] == str(title)).values
    mask = np.column_stack([author_, title_])
        return author_title[mask.all(axis=1)]
def search2(author, title):
    author_ = (author_title['author'] == str(author)).values
    title_  = author_title['title'].str.contains(str(title), na=False)
    mask = np.column_stack([author_, title_])
        return author_title[mask.all(axis=1)]
columns = ['ind', 'author', 'title', 'author_title', 'paragraphs']
index = np.arange(0,100)
p100_in_one_line_allinfo = pd.DataFrame(index=index, columns=columns)
count_empty = 0
count_muti  = 0
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
for i in range(0, 100):
#     print(i)
    author = DATAS_OR.loc[i, '作者']
    title  = DATAS_OR.loc[i, '詩名']
    if len(search(author, title)) == 0:
        count_empty += 1
        p100_in_one_line_allinfo.loc[i,['author', 'title']] =
            np.array([author, title])
    elif len(search(author, title)) >  1:
        count_muti += 1
        p100_in_one_line_allinfo.loc[i,['author', 'title']] =
            np.array([author, title])
    else:
        p100_in_one_line_allinfo.loc[i] = np.concatenate((np.
            array(search(author, title).index), search(author,
            title).values[0]))
p100_in_one_line_allinfo[p100_in_one_line_allinfo['ind'].isnull
    ()]
p100_in_one_line_allinfo.to_excel('p100_in_one_line_allinfo.
    xlsx')
```

資料清理

```
# -*- coding: utf-8 -*-
"""1.清理由json檔直接複製的文檔      2.順便製造全斷文檔"""
import pandas as pd
import numpy  as np
folder = 'D:/Morokei-formove_Version_3/字元分析/word/chinese-
    poetry-master/TANG_Doc/'
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
author_title = pd.read_excel(folder + 'author_title.xlsx')
clean_text_list = []
for i in range(len(author_title)):
    clean_text = ''
    for item in author_title.loc[i,'paragraphs']:
        if item not in '\[\',\]，。」，。·.？（《》）1234567890
            /\-[]{}    :「」１２３４５６７８９０':
            clean_text = clean_text + str(item)
        elif item in '，。':
            clean_text = clean_text + '」'
    clean_text_list = clean_text_list + [clean_text[0:-1]]
clean_text_list_allpart = []
for i in range(len(author_title)):
    clean_text = ''
    for item in author_title.loc[i,'paragraphs']:
        if item not in '\[\',\]，。」，。·.？（《》）1234567890
            /\-[]{}    :「」１２３４５６７８９０':
            clean_text = clean_text + str(item) + '」'

    clean_text_list_allpart = clean_text_list_allpart + [
        clean_text[0:-1]]
pd.DataFrame(clean_text_list ,columns=['paragraphs']).to_excel('
    Tangfull_in_one_line.xlsx')
pd.DataFrame(clean_text_list_allpart ,columns=['paragraphs']).
    to_excel('TS_seg_all_part.xlsx')
```

資料清理 2

```
# -*- coding: utf-8 -*-
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
import pandas as pd
import numpy  as np
folder = 'D:/Morokei-formove_Version_3/字元分析/word/chinese-poetry-master/TANG_Doc/'
TS_seg_in_one_line = pd.read_excel(folder + 'word2vec/poem/TS_seg_in_one_line.xlsx')
TS_seg_allpart     = pd.read_excel(folder + 'word2vec/poem/TS_seg_all_part.xlsx')
p100_all_info      = pd.read_excel('p100_in_one_line_allinfo.xlsx')
"""100首詩,乾淨文檔"""
clean_text_list = []
for i in range(100):
    clean_text = ''
    for item in p100_all_info.loc[i,'paragraphs']:
        if item not in '\[\',\]，。␣':
            clean_text = clean_text + str(item)
        elif item in '，。':
            clean_text = clean_text + '␣'
    clean_text_list = clean_text_list + [clean_text[0:-1]]
"""100首詩,全斷文檔(兩種方式)"""
clean_text_list_allpart = []
for i in range(100):
    clean_text = ''
    for item in p100_all_info.loc[i,'paragraphs']:
        if item not in '\[\',\]，。␣':
            clean_text = clean_text + str(item) + '␣'
    clean_text_list_allpart = clean_text_list_allpart + [
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
        clean_text[0:-1]]
clean_text_list_allpart2 = []
for ind in p100_all_info['ind']:
    clean_text_list_allpart2 = clean_text_list_allpart2 + [
        TS_seg_allpart.loc[ind,'paragraphs']]
pd.DataFrame(np.concatenate((np.array(clean_text_list).reshape
    (100,1), np.array(clean_text_list_allpart2).reshape(100,1)),
     axis = 1), columns=['paragraphs','paragraphs_allpart']).
    to_excel('p100_poem.xlsx')
```

建立向量

```
import time
import collections
import pandas as pd
import numpy  as np
from gensim.models import doc2vec
from gensim.models import word2vec
from gensim import models
import logging
import random
import matplotlib.pyplot as plt
folder = 'drive/Colab_Notebooks/meeting/Tang_Doc/'
TS_seg_all_part  = pd.read_excel(folder + 'word2vec/poem/
    TS_seg_all_part.xlsx')
author_title = pd.read_excel(folder + 'author_title.xlsx')
TS_seg = TS_seg_all_part['paragraphs'].values
pre_result = pd.read_excel('drive/Colab_Notebooks/meeting/兩分
    析結果.xlsx')
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
p100_allinfo = pd.read_excel(folder + 'p100_label_allinfo.xlsx')
#label data
LabeledSentence = doc2vec.TaggedDocument
labelized_data = []
for ind in author_title.index:
    label = ind
    if type(TS_seg_all_part.loc[ind,'paragraphs']) != float:
#         gensim.models.doc2vec.TaggedDocument(gensim.utils.simple_preprocess(line), [i])
        labelized_data.append(LabeledSentence(TS_seg_all_part.loc[ind,'paragraphs'].split(),[label]))
    else:
        labelized_data.append(LabeledSentence(str(TS_seg_all_part.loc[ind,'paragraphs']),[label]))
###
text_data = []
for ind in author_title.index:
    if type(TS_seg_all_part.loc[ind,'paragraphs']) != float:
        text_data.append(TS_seg_all_part.loc[ind,'paragraphs'].split())
    else:
        text_data.append(str(TS_seg_all_part.loc[ind,'paragraphs']))
p100_text = []
for i in range(100):
    p100_text.append(np.array(text_data)[p100_allinfo['ind']][i])
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
n  = 250
it = 30
model = doc2vec.Doc2Vec(vector_size=n, min_count=0, epochs=it)
model.build_vocab(labelized_data)
model.train(labelized_data, total_examples=model.corpus_count,
    epochs=model.epochs)
##取出100首詩之向量
p100_docvec = model.docvecs.vectors_docs[p100_allinfo['ind']]
from openpyxl import load_workbook
book = load_workbook(folder + 'p100_doc_vec_final.xlsx')
writer = pd.ExcelWriter(folder + 'p100_doc_vec_final.xlsx',
    engine = 'openpyxl')
writer.book = book
p100_docvec.to_excel(writer,sheet_name = 'p100_docvec')
writer.save()
writer.close()
```

計算相似度

```
from gensim.models import doc2vec
from gensim.models import word2vec
from gensim import models
import time
import collections
import pandas as pd
import numpy  as np
import logging
import random
import matplotlib.pyplot as plt
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
folder = 'drive/Colab_Notebooks/meeting/Tang_Doc/'
p100_allinfo = pd.read_excel(folder + 'p100_label_allinfo.xlsx')
pre_result = pd.read_excel('drive/Colab_Notebooks/meeting/兩分析結果2.xlsx')
p100_docvec_250_55 = pd.read_excel(folder + 'p100_doc_vec_final.xlsx', sheet_name = 'p100_docvec')
mathodofcorr = 'spearman'
token = 15
begin = 0
end = 21
#'pc_scores1'
sim = np.zeros(100)
choose = token
for ind in pre_result.sort_values('pc_scores1',ascending=False).head(token).index:
    dev = choose
    sim = sim + model.wv.cosine_similarities(np.array(
        p100_docvec_250_55.loc[ind,:]),np.array(
        p100_docvec_250_55))*dev
    choose = choose-1
pd.concat([pre_result.loc[:,['pc_scores1', 'pc_scores2','
    fa_varimaxT_scores1', 'fa_varimaxT_scores2','綜合指標']],
           pd.DataFrame(sim)],
          axis = 1).sort_values('pc_scores1', ascending=False).
             iloc[begin:end,:].corr(method=mathodofcorr).loc
             [:,0]
#'pc_scores2'
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
sim = np.zeros(100)
choose = token
for ind in pre_result.sort_values('pc_scores2',ascending=False
    ).head(token).index:
    dev = choose
    sim = sim + model.wv.cosine_similarities(np.array(
        p100_docvec_250_55.loc[ind,:]),np.array(
        p100_docvec_250_55))*dev
    choose = choose-1
pd.concat([pre_result.loc[:,['pc_scores1', 'pc_scores2','
    fa_varimaxT_scores1', 'fa_varimaxT_scores2','綜合指標']],
            pd.DataFrame(sim)],
           axis = 1).sort_values('pc_scores2', ascending=False).
              iloc[begin:end,:].corr(method=mathodofcorr).loc
              [:,0]
#'fa_varimaxT_scores1    '
sim = np.zeros(100)
choose = token
for ind in pre_result.sort_values('fa_varimaxT_scores1',
    ascending=False).head(token).index:
    dev = choose
    sim = sim + model.wv.cosine_similarities(np.array(
        p100_docvec_250_55.loc[ind,:]),np.array(
        p100_docvec_250_55))*dev
    choose = choose-1
pd.concat([pre_result.loc[:,['pc_scores1', 'pc_scores2','
    fa_varimaxT_scores1', 'fa_varimaxT_scores2','綜合指標']],
            pd.DataFrame(sim)],
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

```
48          axis = 1).sort_values('fa_varimaxT_scores1', ascending
                =False).iloc[begin:end,:].corr(method=mathodofcorr
                ).loc[:,0]
49  #'fa_varimaxT_scores2'
50  sim = np.zeros(100)
51  choose = token
52  for ind in pre_result.sort_values('fa_varimaxT_scores2',
        ascending=False).head(token).index:
53      dev = choose
54      sim = sim + model.wv.cosine_similarities(np.array(
            p100_docvec_250_55.loc[ind,:]),np.array(
            p100_docvec_250_55))*dev
55      choose = choose-1
56  pd.concat([pre_result.loc[:,['pc_scores1', 'pc_scores2','
        fa_varimaxT_scores1', 'fa_varimaxT_scores2','綜合指標']],
57              pd.DataFrame(sim)],
58          axis = 1).sort_values('fa_varimaxT_scores2', ascending
                =False).iloc[begin:end,:].corr(method=mathodofcorr
                ).loc[:,0]
```


DOI:10.6814/THE.NCCU.MATH.003.2018.B01

參考文獻

Gao, J. (2018). Chinese-poetry. https://github.com/chinese-poetry/chinese-poetry.

Johnson, R. and Wichern, D. (2007). *Applied multivariate statistical analysis(6th ed.)*. Prentice Hall, Upper Saddle River, NJ.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. *Computing Research Repository*, arXiv:1405.4053.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *Computing Research Repository*, arXiv:1301.3781.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

王兆鵬、張靜、邵大為、唐元 (2011)。*唐詩排行榜（初版）*。北京：中華書局。

王宏林 (2012)。論唐詩經典的基本屬性，建構要素及途徑。*許昌學院學報*，31(4): 54,58。

蔣寅 (2003)。*中國古代文學通論隋唐五代卷（初版）*。遼寧：人民出版社。

趙義山、李修生 (2010)。*中國分體文學史詩歌卷修訂本（2 版）*。上海：上海古籍出版社。

陳耀茂 (1999)。*多變量解析方法與應用（初版）*。台北：五南圖書出版公司。

魯迅 (2005)。*魯迅全集第 13 卷（初版）*。北京: 人民文學出版社。

DOI:10.6814/THE.NCCU.MATH.003.2018.B01