

友 李賀 病 奇險 江南
自然 浪漫 詠物 州 元稹

數位人文大數據競賽

國立臺灣科技大學 電資學士班 王姿嵐
國立臺灣科技大學 電機工程系 陳騏業
國立政治大學 財務管理學系 林奕均

愁 白居易 長安 王之渙 書 抒情 深處
旅 高適 笑 韋應物 登高

友 李賀 病 奇險 江南
自然 浪漫 州 元稹 明日
邊塞 酒 詠物 淚 豪放
岑參 闌珊 王維 雨 懷人 李白

數據唐詩

王姿嵐 陳騏業 林奕均

孟浩然 惆悵 馬
故國 社會 悵望 杜牧
思鄉 書 抒情 深處
長安 王之渙 笑 韋應物 登高
愁 白居易 旅 高適

前言

數據科學（資料科學）近年被廣泛應用於取得有價值的資料部分來生產數據產品。包括應用數學、統計、模式識別、機器學習、數據可視化、數據倉庫以及高性能計算等等。

然而，我們卻甚少能看見數據科學應用於人文學科上。透過本次競賽，我們希望能激盪一些文學與資料處理的火花。

動機

詩詞曲向來被視為文學上的瑰寶；
本次專題透過採礦資料分析找出【詩詞曲】中：
【主題】、【詩人】、【詞彙】之間的相關聯性，
以探討此三元素的彼此的交互作用。

問題定義

本次專題的問題定義，我們設定在「架設人文科學與數據資料的橋樑」，採取過去很少應用於數據科學的人文資料，於採礦中呈現。

在人文的元素，我們選擇古典詩詞、唐詩三百首此類文學作品作資料數據分析；欲探討詩人之間，與詩人的文學應用、寫作主題有何種相關及鏈結。

挑戰

不同於過去學習資料處理的經驗，本此專題我們選取「唐詩三百首」樣本，故因為「中文字詞」的特性而產生了一些挑戰。例如：

無法判斷詞語的斷接

產生結果為助詞（無意義）的字

資料處理

STEP 1

資料集



唐詩文本蒐集

· 唐詩三百首

· 詩人集 (含詞曲)



轉換類別型態、列出

· 整體分類 (篇名/詩人/內文)

322 lines (321 slots) 99.1 KB

1	name	author	content	tag	style
2	行宮	唐代：元稹	宮殿古行宮，宮前春草紅。白頭宮女		
3	登鸛雀樓	唐代：王之渙	白日依山盡，黃河入海流。欲窮千里目，		
4	新秋感興	唐代：王維	二日入東下，思年作舊愁。未識		
5	相思	唐代：王維	紅豆生南國，春來發幾枝。願君多采		
6	觀鵝二首·其二	唐代：王維	君自故鄉來，願知收鵝事。		
7	鹿柴	唐代：王維	空山不見人，但聞人語響。返景入深		
8	竹裡館	唐代：王維	獨坐幽篁裡，彈琴復長嘯。深林人		
9	送別/山中送別	唐代：王維	山中相送罷，日暮掩柴扉。		
10	聞郎上九	唐代：白居易	綠蟻新醅酒，紅泥小火爐。晚來天欲		
11	哥舒歌	唐代：西鄙人	北斗七星高，哥舒夜帶刀。至今		
12	靜夜思	唐代：李白	床前明月光，疑是地上霜。舉頭望		
13	怨情	唐代：李白	美人卷珠簾，深坐蹙眉眉。但見淚		
14	楓橋夜泊/登樂遊原	唐代：李商隱	向晚意不遠，驅車置古		
15	送李儋	唐代：李商隱	君去秋來秋，君去玉房前。欲得		
16	送李儋	唐代：宋之問	嶺外音書斷，經冬復歷春。近鄉		
17	八陣圖	唐代：杜甫	功高三分國，名成八陣圖。(名成		
18	春望	唐代：杜甫	移舟泊煙渚，日暮客愁新。野		
19	春望	唐代：杜甫	移舟泊煙渚，日暮客愁新。野		
20	春望	唐代：杜甫	移舟泊煙渚，日暮客愁新。野		
21	江雪	唐代：柳宗元	千山鳥飛絕，萬徑人踪滅。孤舟		
22	秋夜寄邱少府/秋夜寄丘二十二弟	唐代：韋應物	桂子落		
23	秋夜寄邱少府	唐代：韋應物	桂子落		
24	高郵/故園二千里	唐代：張籍	故園二千里，梁		

孟浩然.txt

孟郊.txt

張繼.txt

李世民.txt

李商隱.txt

李煜.txt

李白.txt

李賀.txt

杜牧.txt

杜甫.txt

柳宗元.txt

溫庭筠.txt

王勃.txt

王昌齡.txt

1	name	author	content
2	行宮	元稹	宮殿古行宮，宮前春草紅。白頭宮女在，
3	登鸛雀樓	王之渙	白日依山盡，黃河入海流。欲窮千里目，
4	新秋感興	王維	二日入東下，思年作舊愁。未識
5	相思	王維	紅豆生南國，春來發幾枝。願君多采
6	觀鵝三首·其二	王維	君自故鄉來，願知收鵝事。未識
7	鹿柴	王維	空山不見人，但聞人語響。返景入深
8	竹裡館	王維	獨坐幽篁裡，彈琴復長嘯。深林人
9	送別/山中送別	王維	山中相送罷，日暮掩柴扉。未識
10	聞郎上九	白居易	綠蟻新醅酒，紅泥小火爐。晚來天欲
11	哥舒歌	西鄙人	北斗七星高，哥舒夜帶刀。至今
12	靜夜思	李白	床前明月光，疑是地上霜。舉頭望
13	怨情	李白	美人卷珠簾，深坐蹙眉眉。但見淚
14	楓橋夜泊/登樂遊原	李商隱	向晚意不遠，驅車置古
15	送李儋	李商隱	君去秋來秋，君去玉房前。欲得
16	送李儋	宋之問	嶺外音書斷，經冬復歷春。近鄉
17	八陣圖	杜甫	功高三分國，名成八陣圖。(名成
18	春望	杜甫	移舟泊煙渚，日暮客愁新。野
19	春望	杜甫	移舟泊煙渚，日暮客愁新。野
20	春望	杜甫	移舟泊煙渚，日暮客愁新。野
21	江雪	柳宗元	千山鳥飛絕，萬徑人踪滅。孤舟
22	秋夜寄邱少府/秋夜寄丘二十二弟	韋應物	桂子落
23	秋夜寄邱少府	韋應物	桂子落
24	高郵/故園二千里	張籍	故園二千里，梁

STEP 2

程式內容

- ✓ 清理檔案
- ✓ 切詞，刪去符號、英文
- ✓ 建立TDM文本矩陣，統計辭彙出現數量
- ✓ 由TDM轉TFIDF，檢視圖表
- ✓ PCA + K-means 尋找相關性

困難解決

STEP 3

發佈至 SHINY



The screenshot shows a web browser window with the URL https://lanw368.shinyapps.io/poetry-analysis_2018-big-data-contest/. The navigation bar includes links for 數據唐詩, 簡介, 文字雲, 唐詩三百首, 以詩作類別分析, 類別間的關聯性, 詩人常用字, 詩人相似度, and 詩人間的關聯性. The main content area is titled 簡介 and contains the following text:

詩詞曲向來被視為文學上的瑰寶；

本次專題透過資料分析找出【唐詩】中：

【詞彙之間】、【詞彙與詩人之間】的相關性

以探討詞彙之間的關聯程度，

及作者生平、性格與常用字詞的關係。

作品成果

#抒情

#奇險

#自然

#浪漫

#詠物

#懷人

#思鄉

#主題

#邊塞

#豪放

#寫實

#社會

李白

王維

高適

李賀

詩人

元稹

白居易

王之渙

韋應物

岑參

孟浩然

杜牧

旅

故國

惆悵

深處

詞彙

雨

淚

病

江南

愁

闌珊

書

長安

州

友

笑

悵望

明日

馬

登高

酒

呈現方式

資料視覺化

資料分析

【文字雲】

- 詞彙

【長條圖】

- 主題 #TAG
- 主題與詞彙
- 詩人與詞彙
- 詩人之間

【PCA/K-means】

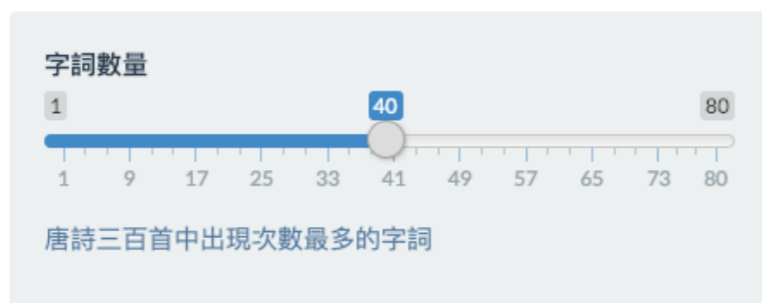
- 類別之間
- 詩人分群

* 備註：詞彙於本次專題中，又分有單詞/多詞及綜合。

文字雲

詩人最常用的詞彙

唐詩三百首文字雲



調整詞彙數量，
顯示不同範圍的文字雲。



古人好詠月，
夜裡，與友人盡吐心聲，為了誰而有夢、有欲。
他們思考生死、思考身在何時、何處，望月，想人生。

長條圖

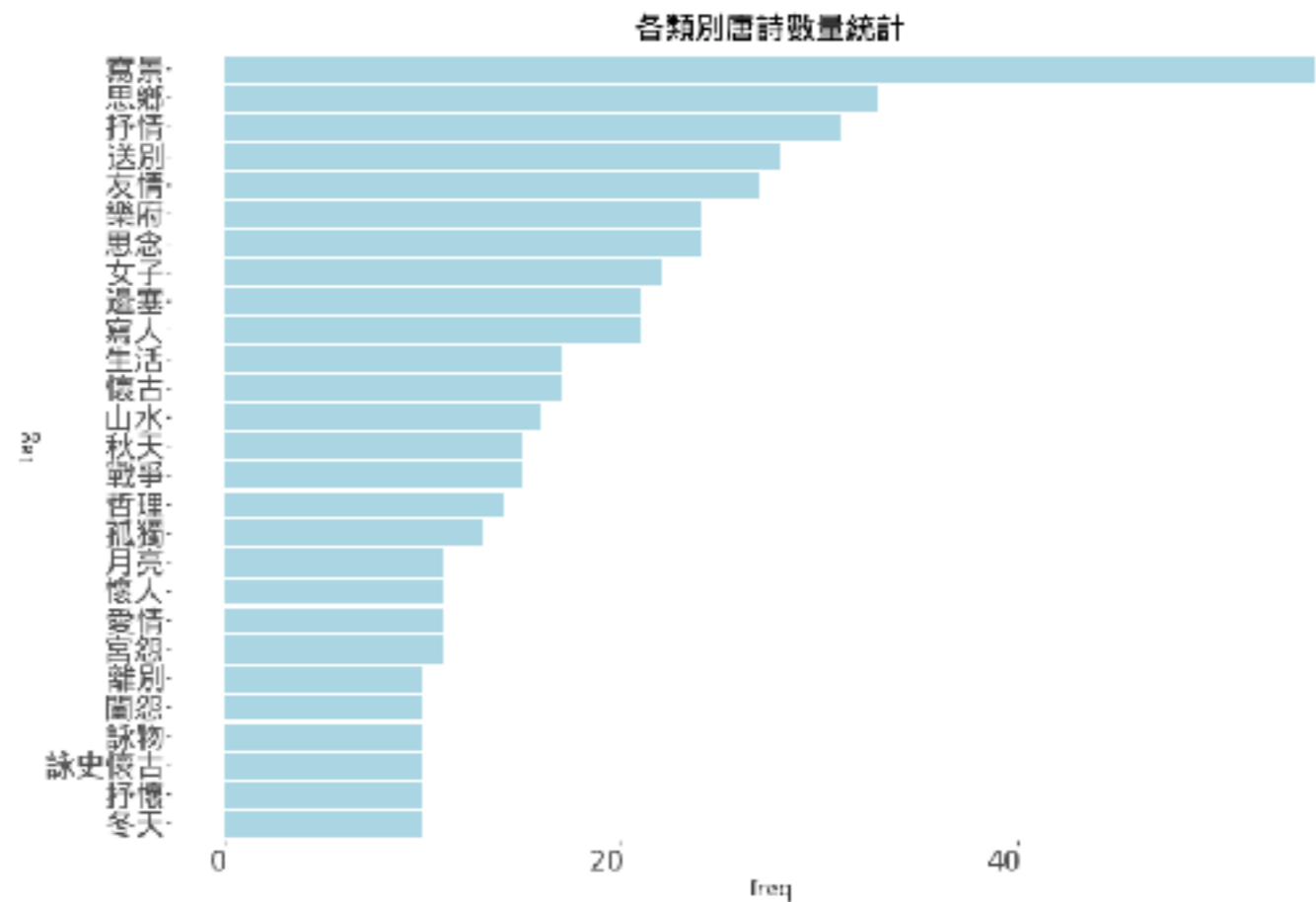
本次專題中，
我們統計唐詩三百首中
最常出現的主題。

以時下流行的#HASHTAG
標籤詩人作品主題。

由圖表可知，
最常見的#TAG為
寫景、思鄉、送別。

主題統計

以詩作類別統計



長條圖

我們統計各個主題中出現最多的詞彙（包含單、多字）

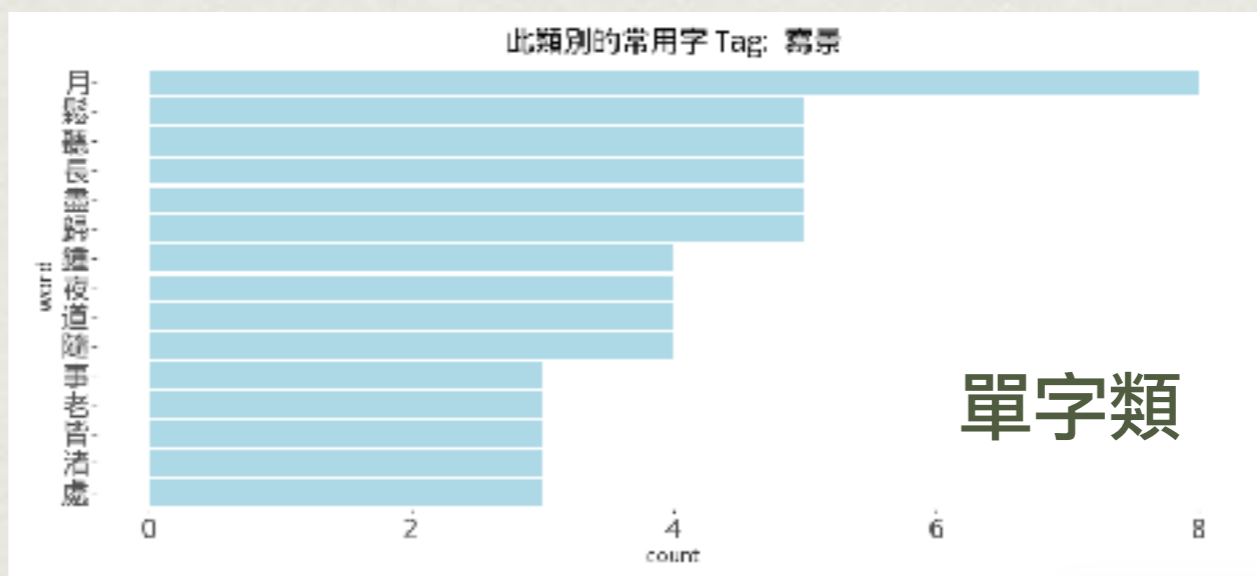
以 #寫景 為例：

以詩作類別統計

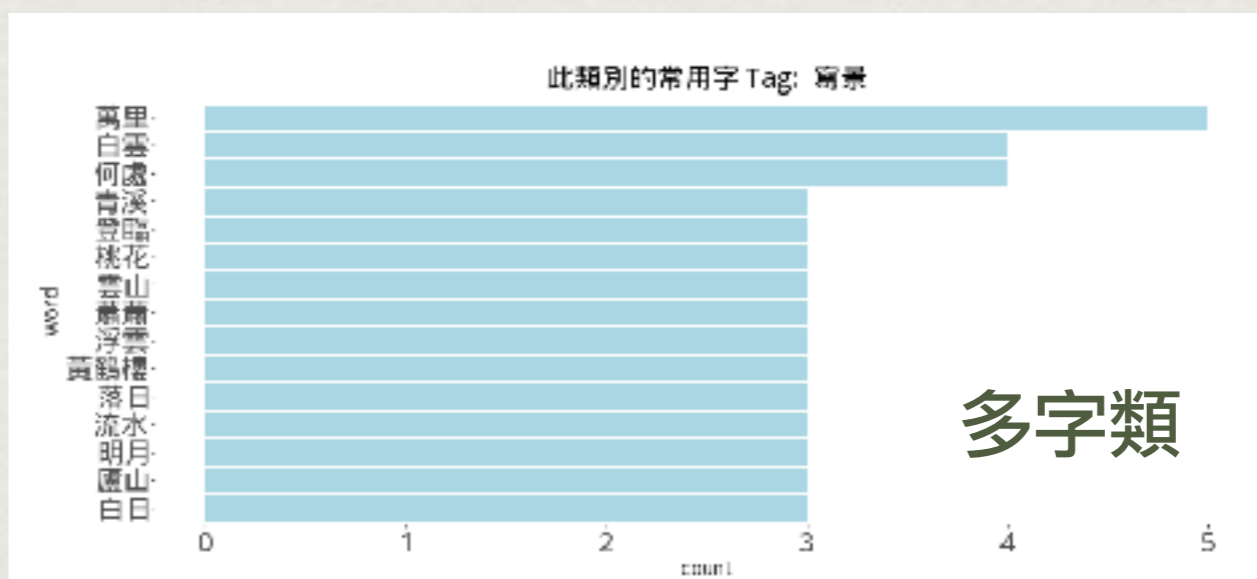
詩作類別:

- 寫景
- 寫景
- 思鄉
- 抒情
- 送別
- 友情
- 思念
- 樂府
- 女子

以 #主題 分類 (1)



單字類



多字類

長條圖

以 #主題 分類 (2)

續以 #寫景 為例：

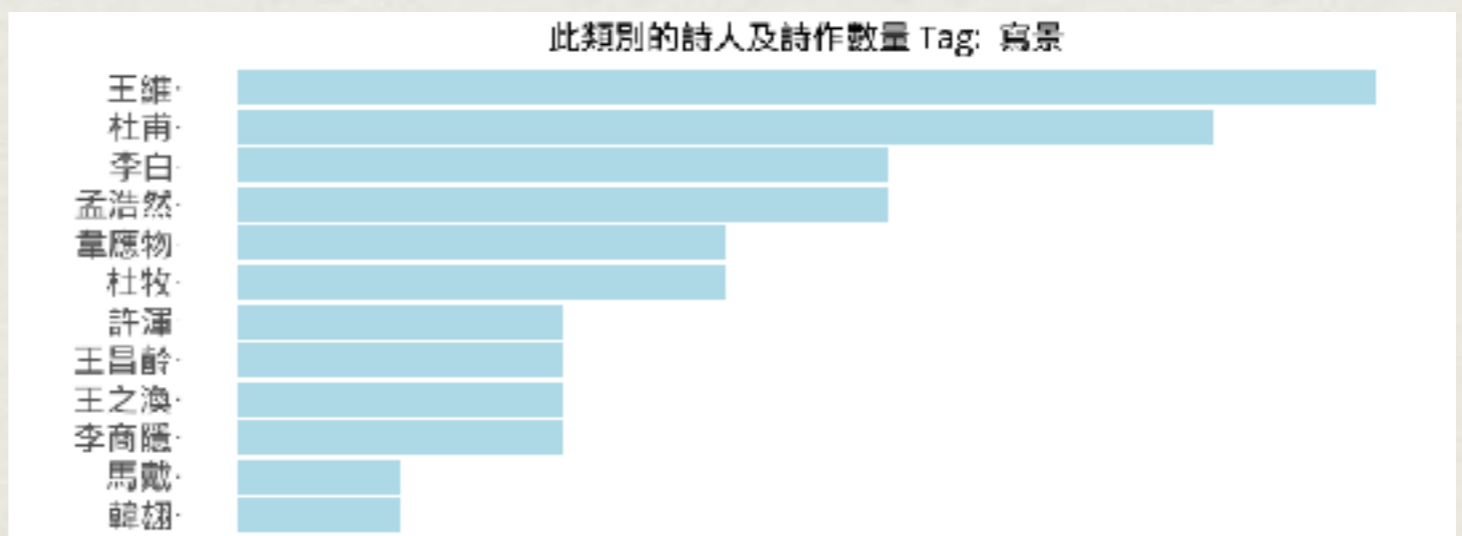
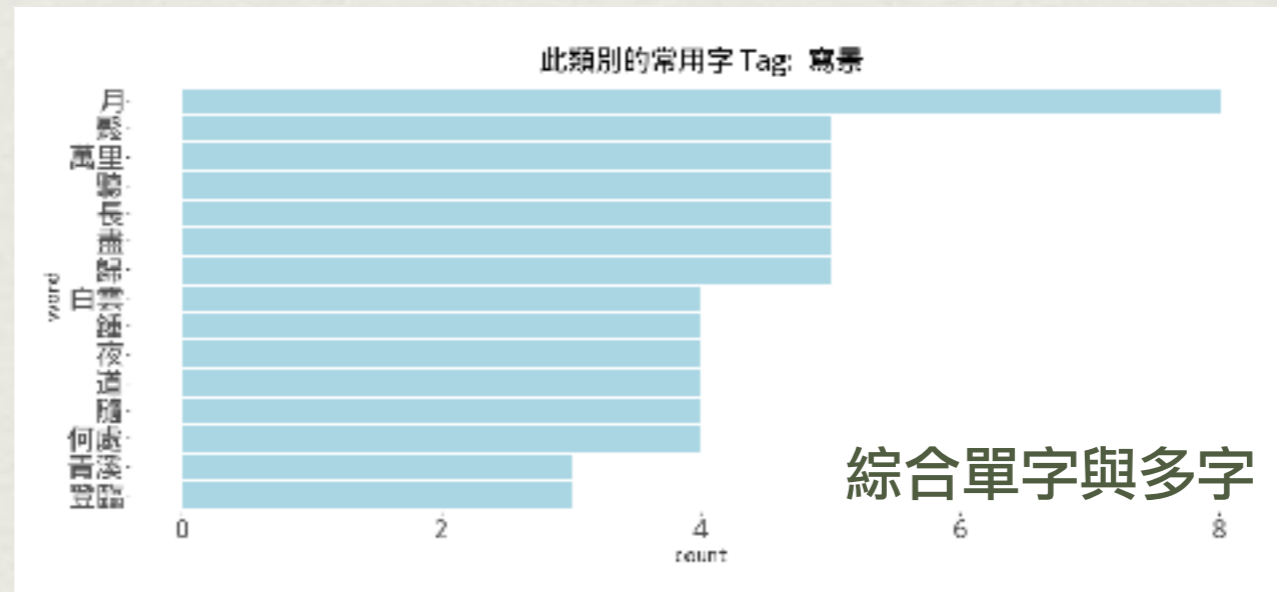
還有「綜合」包含單字與多字
以及 最常創作此主題的詩人

以詩作類別統計

詩作類別:

寫景

- 寫景
- 思鄉
- 抒情
- 送別
- 友情
- 思念
- 樂府
- 古詩



長條圖

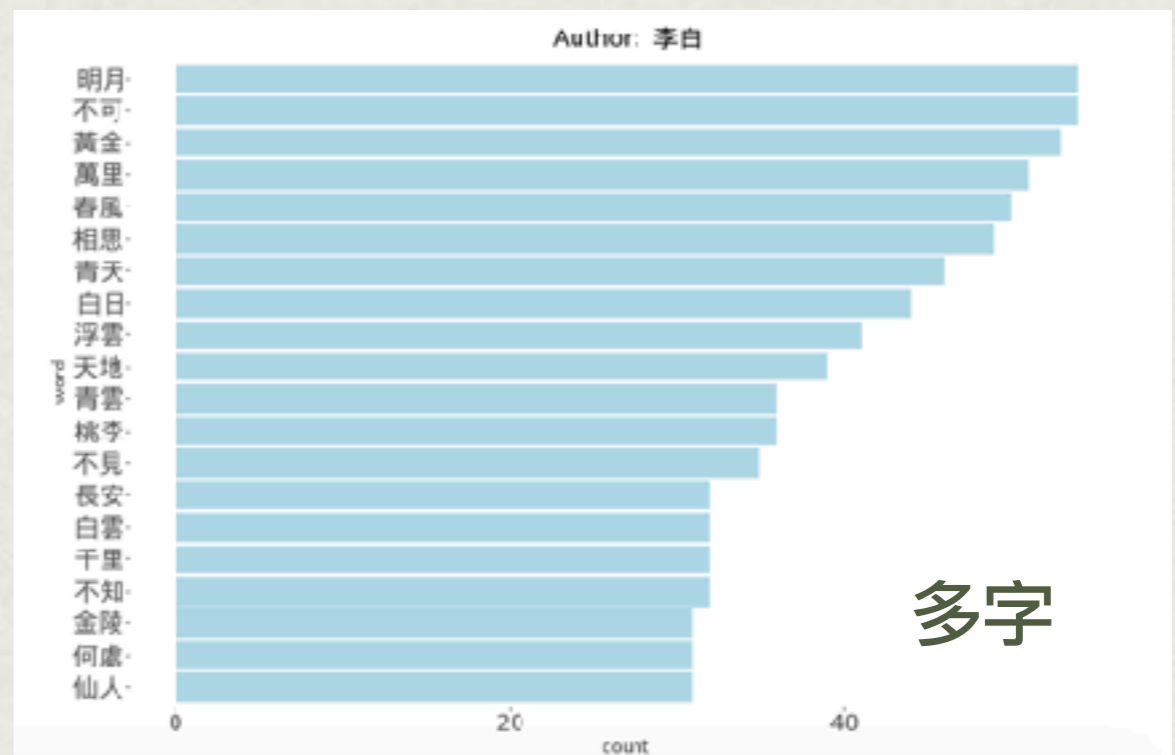
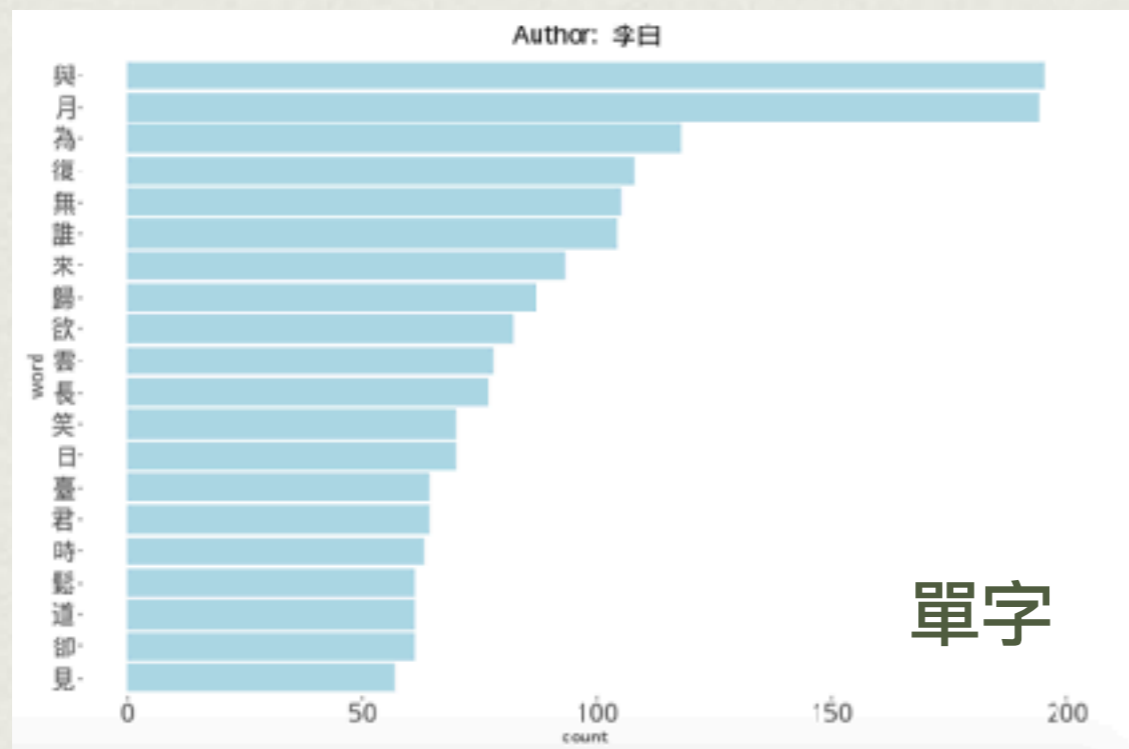
詩人與詞彙

選擇詩人，
並顯示此詩人最常使用的詞彙
(含單字、多字)。

以李白為例：

詩人常用字統計

詩人：
李白
李白
岑參
張九齡
王維
白居易

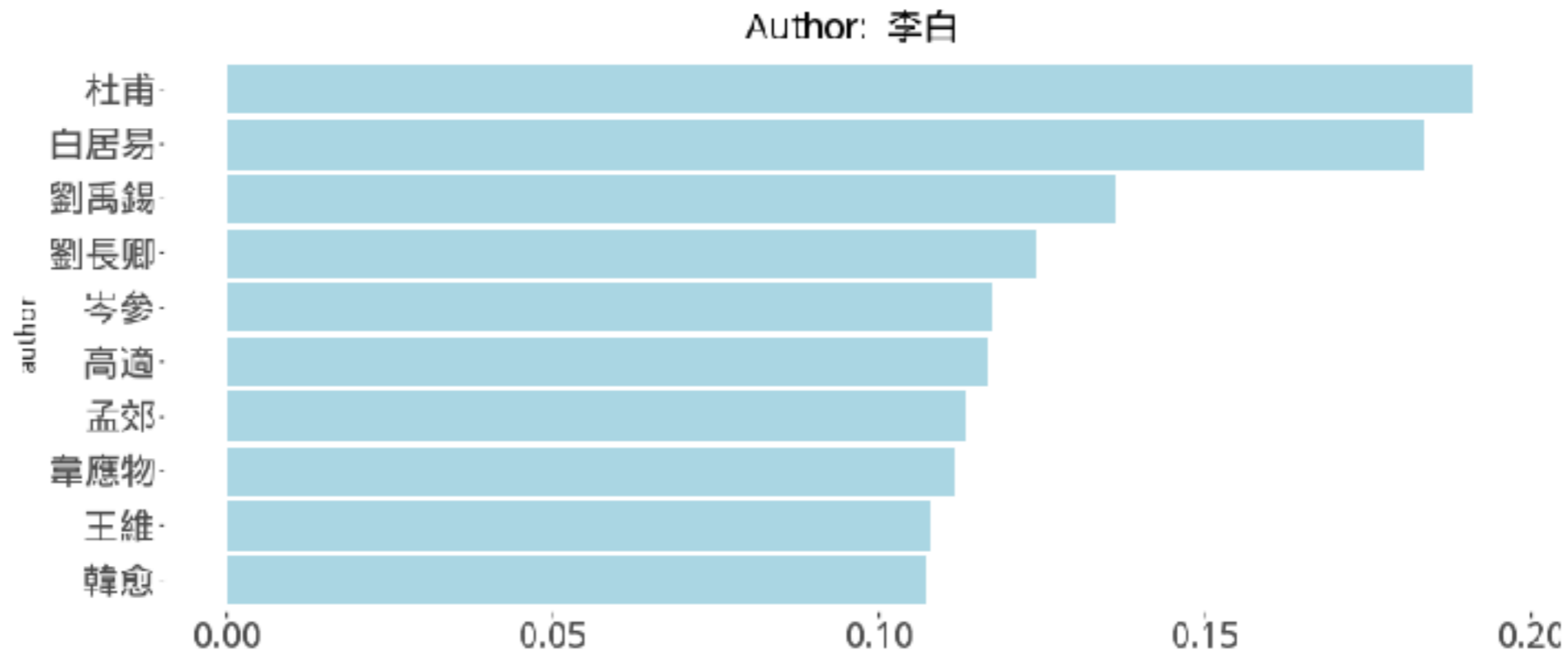


長條圖

詩人之間

輸入詩人，即顯示其他與此詩人在運用詞彙方面 最為相似者。
以李白為例：

李白（701年－762年），字太白，號青蓮居士，中國唐朝詩人，自言祖籍隴西成紀（今日肅省天水市秦安縣），先世西涼武昭王李嵩之後，與李唐皇室同宗。幼時內遷，寄籍劍南道綿州（今四川省江油昌隆縣）。另郭沫若考證李白出生於吉爾吉斯碎葉河上的碎葉城，屬唐安西都護府（今楚河州托克馬克市）。有「詩仙」、「詩俠」、「酒仙」、「謫仙人」等稱呼，活躍於盛唐，為傑出的浪漫主義詩人。與杜甫合稱「李杜」。被賀知章譽為「天上謫仙」。



PCA/K-means

主題之間

詩作類別間的關聯性

使用PCA及K-means分群，我們可以將相似的詩作類型分群

Number of k:

15

調整k值大小可以改變分群數量

右方互動窗格可以自行選取預觀察範圍

調整k值，以選擇主題分群數量，進行各主題之間的相關分析。

孤獨
女子

例如：
孤獨與女子相緊存在座標另端
可發掘詩人與歷史的思想型態



PCA/K-means

詩人之間

詩人間的關聯性

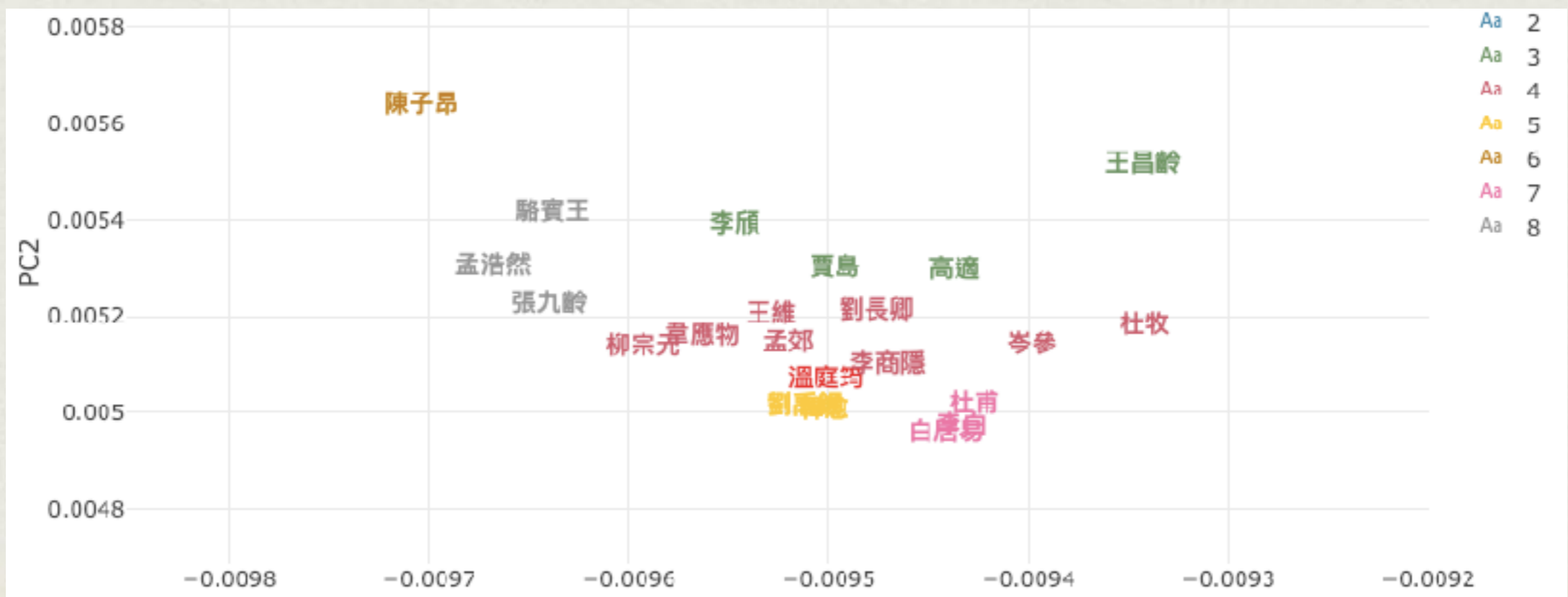
使用PCA及K-means分群，我們可以將風格類似的詩人們分群

Number of k:

8

調整k值，以選擇詩人分群數量，進行各詩人之間的相關分析。

此外，我們也發現位置相近的詩人處在相去不遠的朝代也有類似的境遇。



創新價值

本次專題，我們的創新價值在於探討了過去不曾應用於「文學作品」此類資料。主題、詩人、詞彙，三者的交互關係，存在者比「相關與否」之間更有趣的連結；包括歷史脈絡、生長背景、作者境遇等等。

人文學科與資料科學的碰撞是精彩且有趣的，資料顯示的成果理性地存在更多值得我們探究的思考。

資料來源

· 詩詞名句網 <http://www.shicimingju.com/>

唐诗三百首(306首)

查询到306首诗词

■ 新窗口打开

全部 唐(306)

全部 李白(33) 韩翃(3) 岑参(7) 张九龄(5) 王维(30) 权德舆(1) 白居易(6) 祖咏(2) 杜甫(35) 李商隐(22) 马戴(2) 李颀(7) 王建(1) 王翰(1) 刘方平(2) 张籍(1) 杜牧(10) 陈子昂(1) 高适(2) 许浑(2) 王勃(1) 杜秋娘(1) 张继(1) 温庭筠(4) 韩愈(4) 孟浩然(14) 韩偓(1) 孟郊(2) **贺知章(1)** 綦毋潜(1) 元结(2) 唐玄宗(1) 杜审言(1) 刘禹锡(4) 三湾(1) 钱起(4) 李益(3) 司空曙(3) 张乔(1) 崔涂(2) 柳宗元(5) 韦庄(2) 王之涣(2) 僧皎然(1) 崔颢(3) 骆宾王(2) 崔曙(1) 皇甫冉(1) 刘慎虚(1) 卢纶(3) 薛逢(1) 王昌龄(9) 秦韬玉(1) 刘长卿(11) 宫建(2) 韦应物(12) 李端(1) 张祜(4) 贾岛(1) 西鄙人(1) 邱为(1) 张旭(1) 柳中庸(1) 朱庆余(1) 戴叔伦(1) 郑畋(1) 陈陶(1) 宋之问(2) 益嘉运(1) 杜荀鹤(1) 元稹(2) 朱庆馀(1) 沈佺期(2) 张泌(1)