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

數據唐詩

王姿嵐 陳騏業 林奕均

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

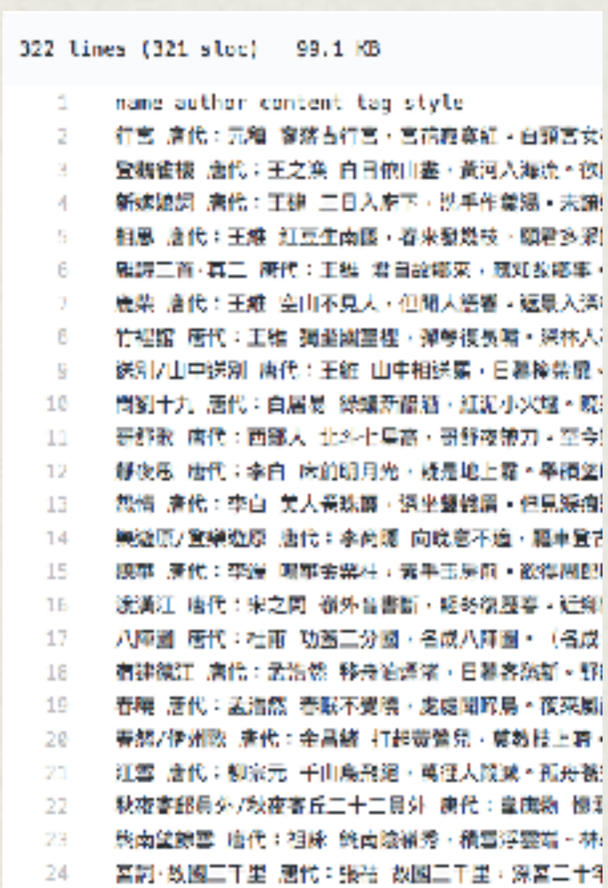
STEP 1

資料集

✓ 唐詩文本蒐集

✓ 轉換類別型態、列出

· 唐詩三百首



· 詩人集（含詞曲）

· 整體分類（篇名/詩人/內文）

STEP 2

程式內容

- 清理檔案
- 切詞，刪去符號、英文

困難解決

- 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

STEP 3

發佈至 SHINY

作品成果

#抒情
#自然
#浪漫
#詠物
#奇險
#懷人
#思鄉
#主題
#豪放
#邊塞
#寫實
#社會
高適
李賀
李白
王維
元稹
白居易
岑參
王之渙
韋應物
孟浩然
杜牧
旅
故國
詩人
悵惘
深處
江南
雨
詞彙
淚
病
愁
闌珊
長安
州
書
帳望
馬
登高
酒
友
笑
明日

呈現方式

資料視覺化

【文字雲】【長條圖】【PCA/K-means】

資料分析

- 詞彙
- 主題 #TAG
- 主題與詞彙
- 詩人與詞彙
- 詩人之間
- 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

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。

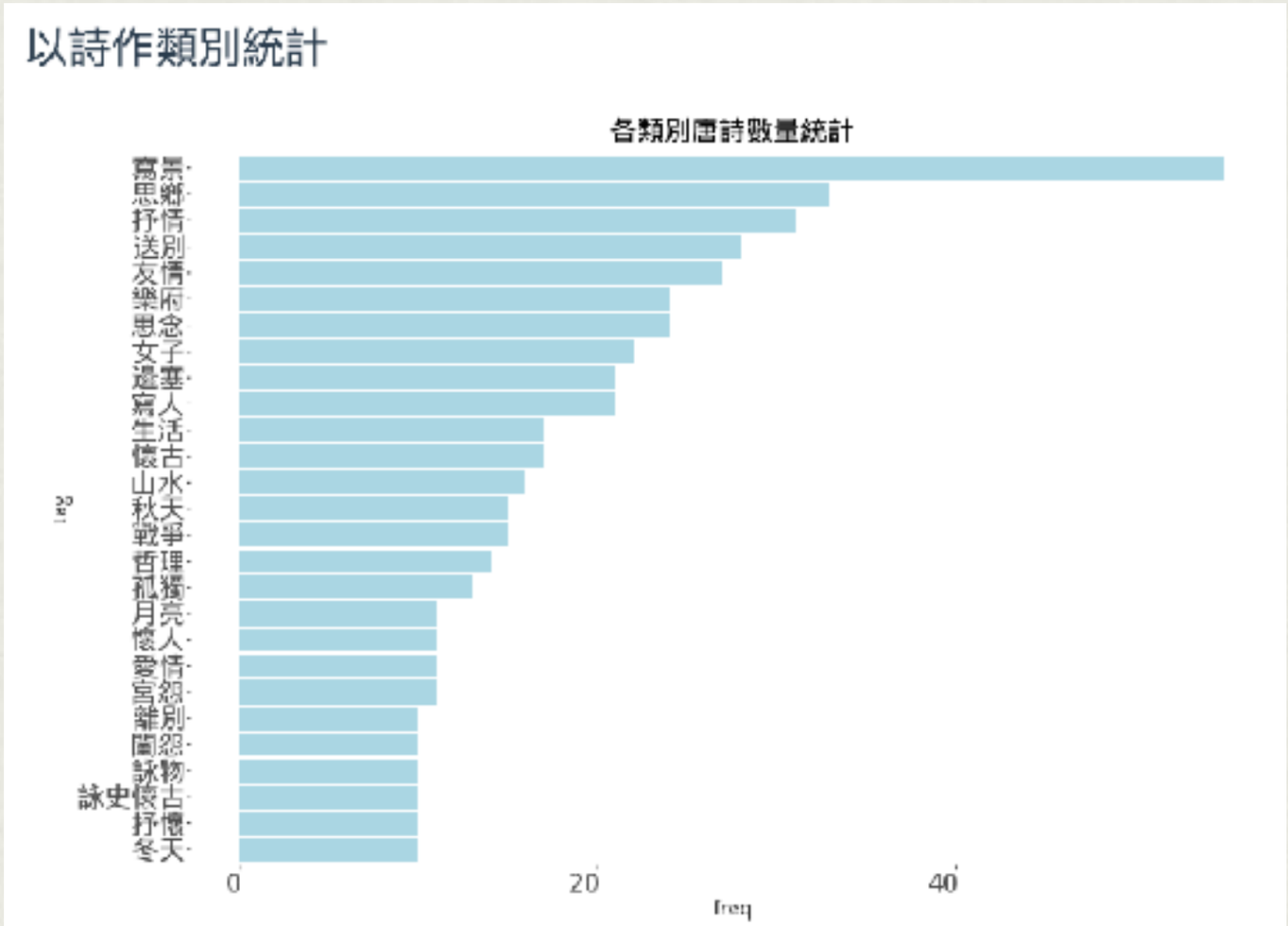
長條圖

主題統計

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。

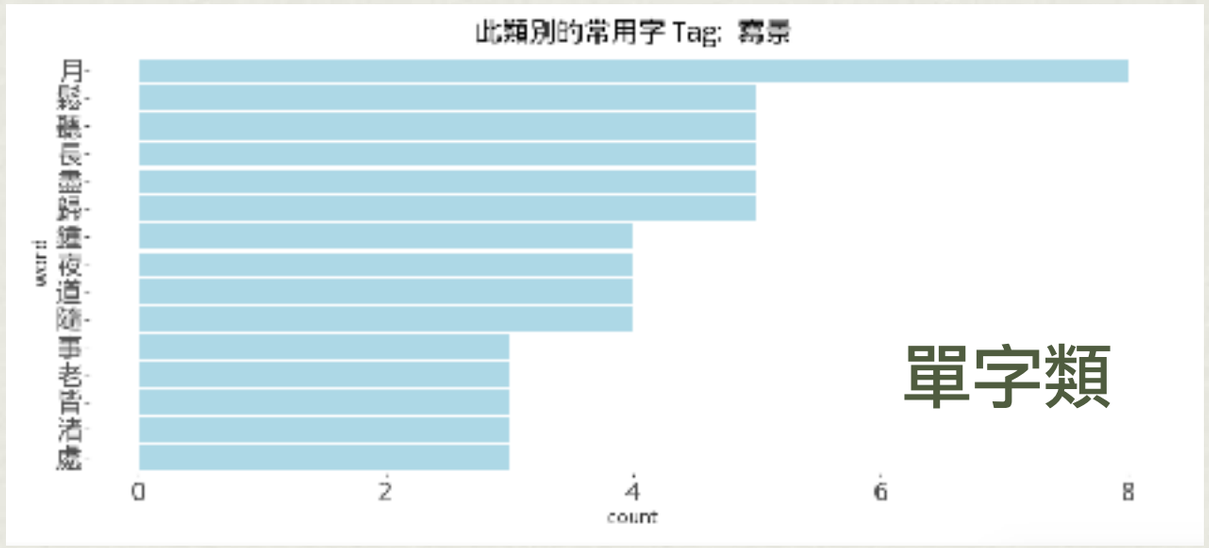


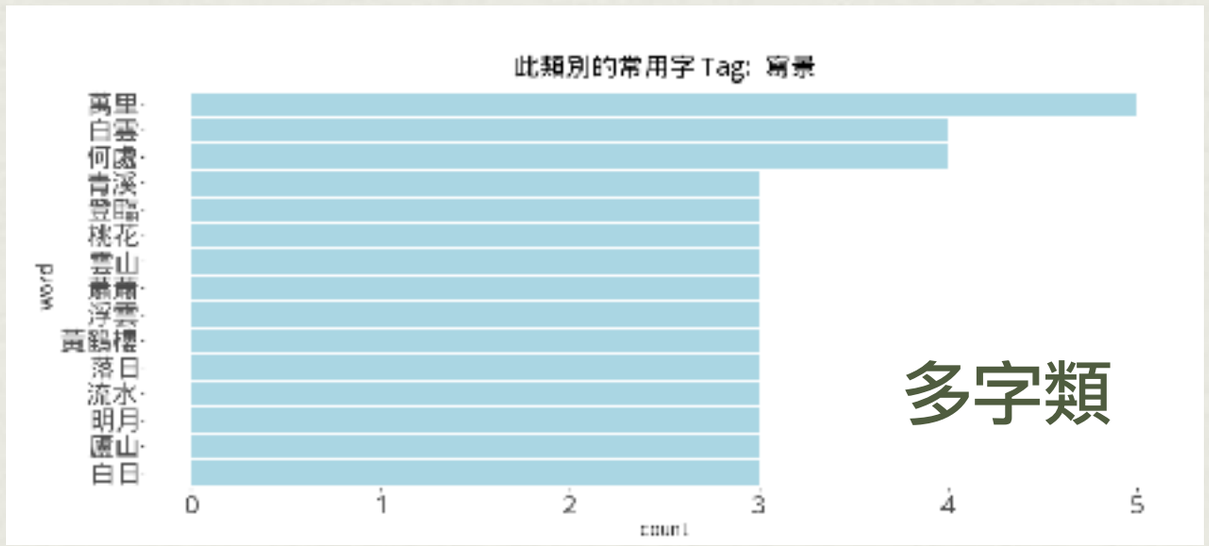
長條圖

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 #主題 分類 (1)



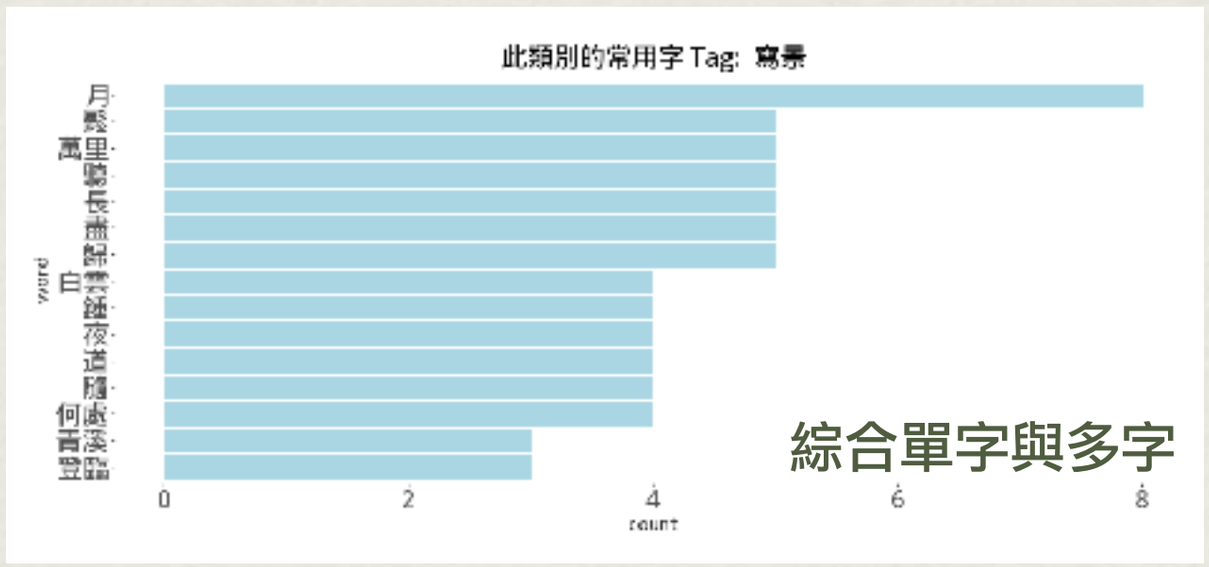


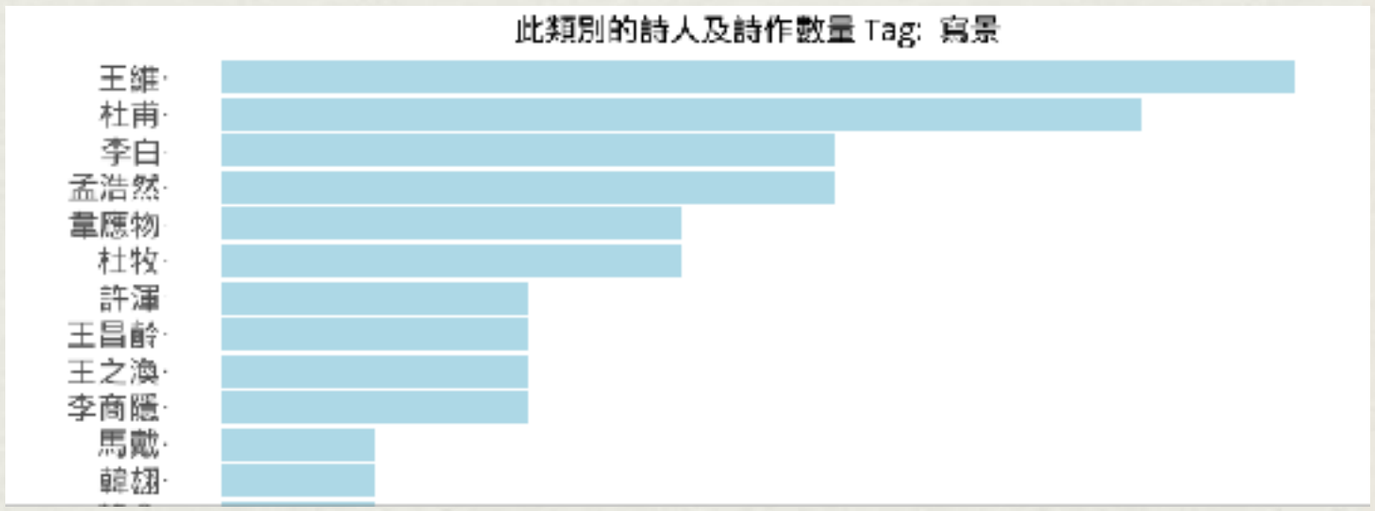
長條圖

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



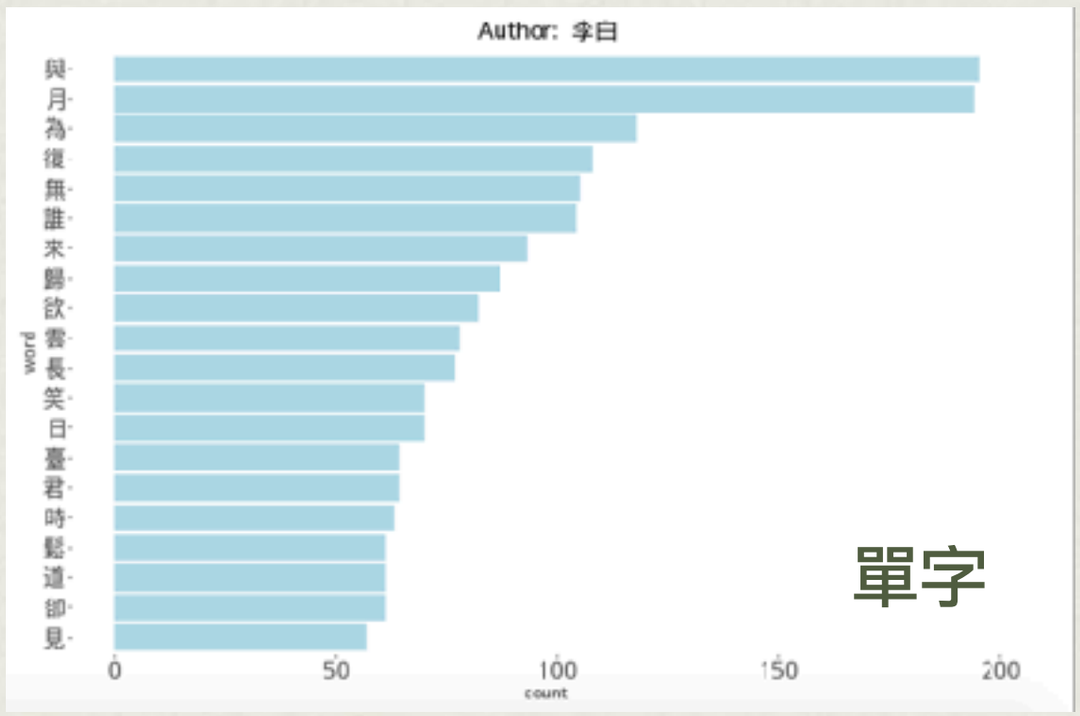


長條圖

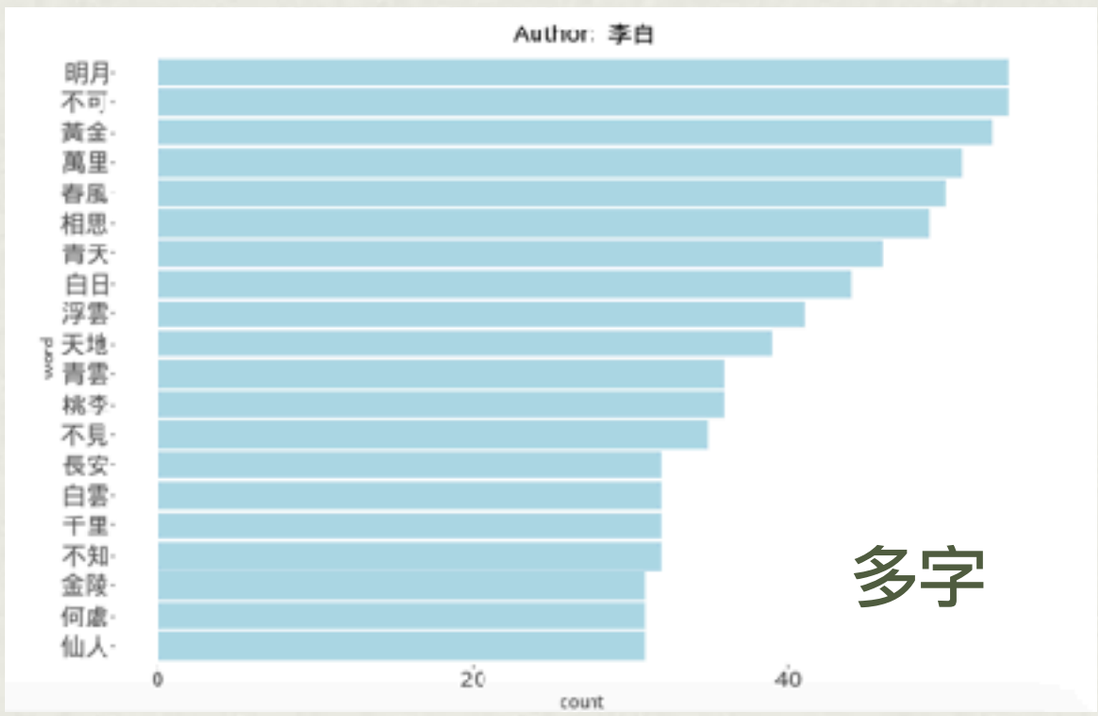
詩人與詞彙

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：



單字



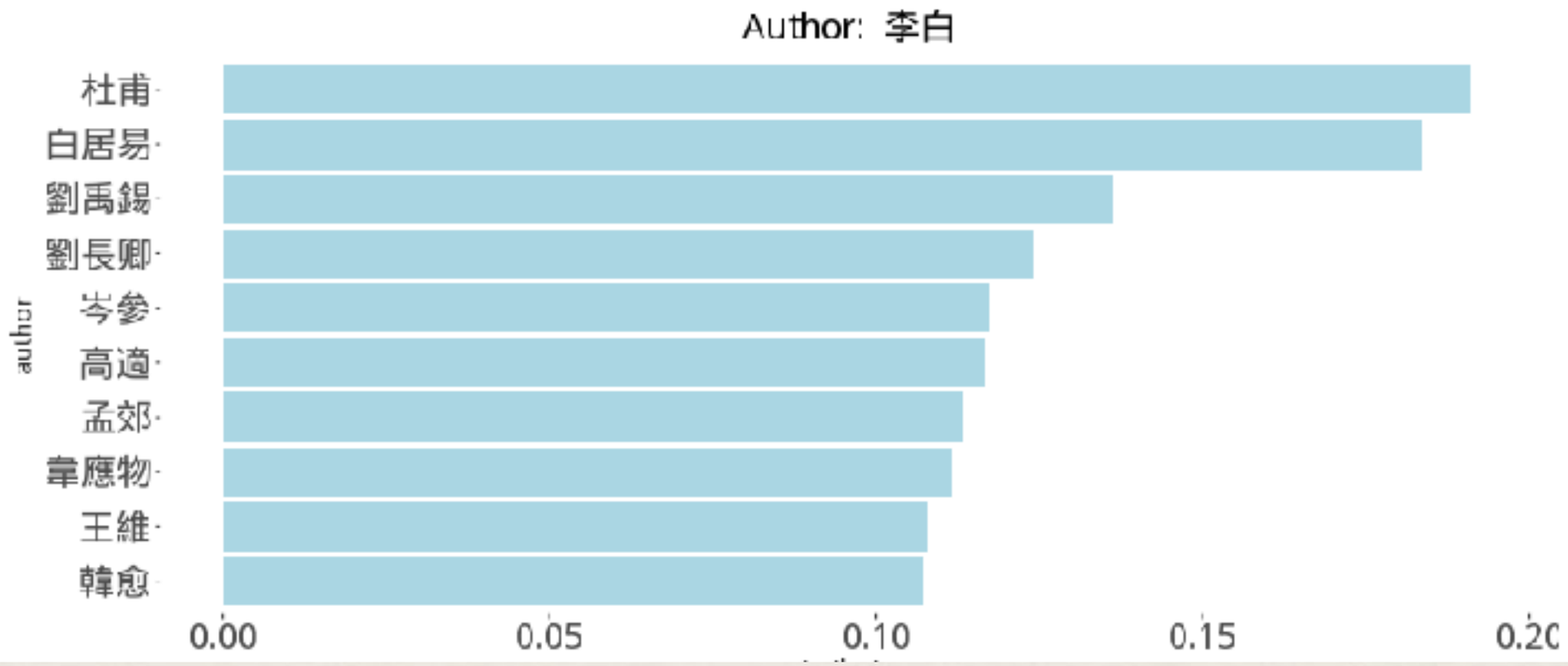
多字

長條圖

詩人之間

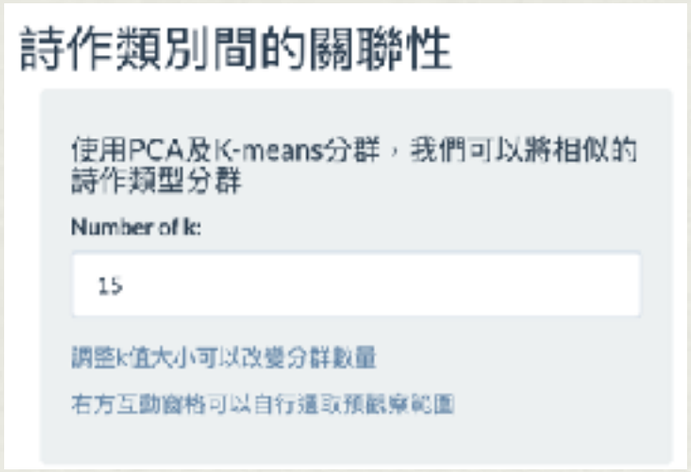
輸入詩人，即顯示其他與此詩人在運用詞彙方面 最為相似者。
以李白為例：

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（今甘肅省天水市秦安縣），先世西涼武昭王李暠之後，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」。被賀知章驚呼為「天上謫仙」。



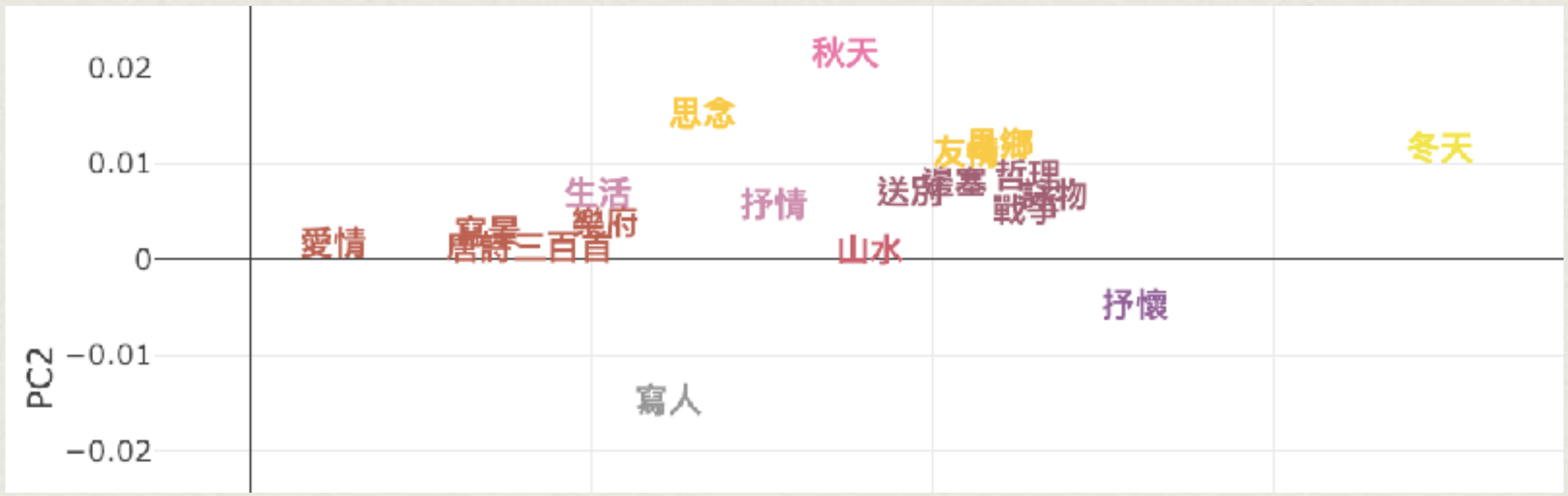
PCA/K-means

主題之間



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。

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

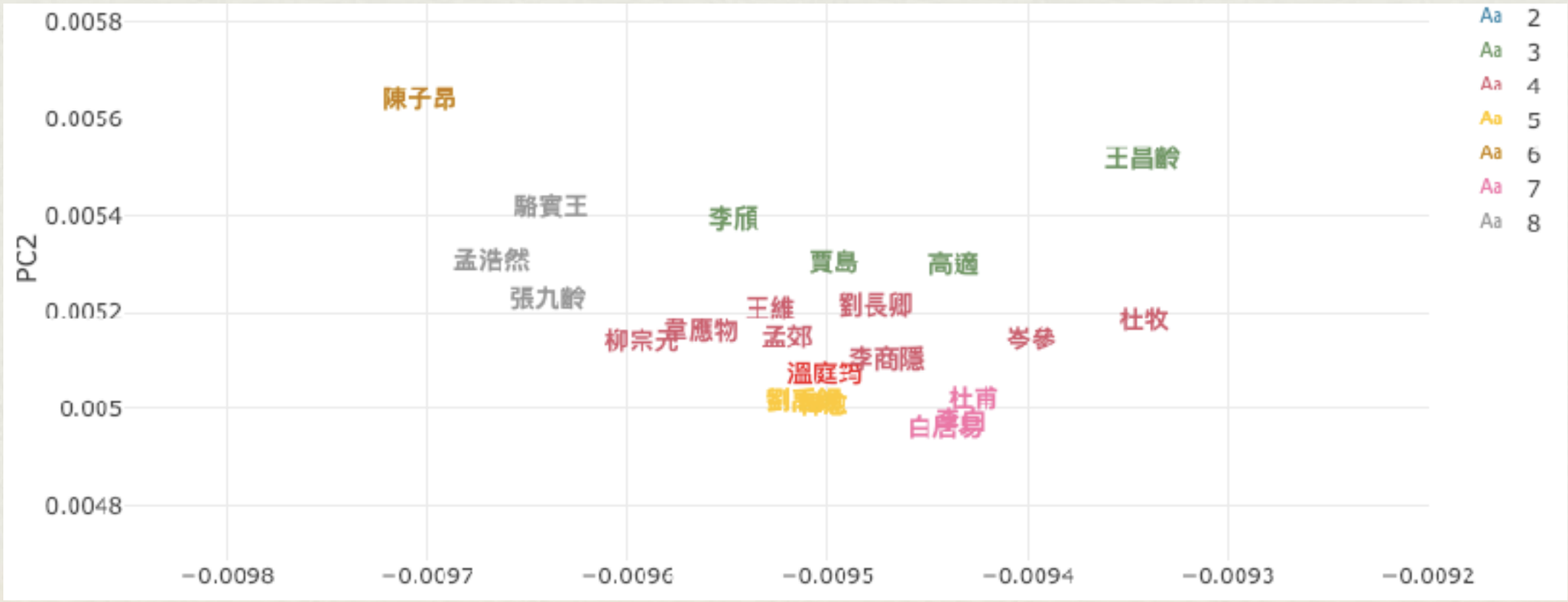


PCA/K-means

詩人之間

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/