

以文風指標分析《紅樓夢》的作者爭議問題

陳佳琳^a 楊昌彪^{a*} 曾國尊^b

^a國立中山大學資訊工程學系

^b國立高雄海洋科技大學-航運管理系暨研究所

摘要

關於《紅樓夢》的作者是否僅出自一人的議題，自西元1750年以來即備受討論。在本篇論文中，我們參考前人研究，針對不同特徵值（字數、詞頻、句類、變動點等），使用機器學習進行作者分類實驗。我們利用中央研究院CKIPTagger進行斷詞與詞性標註，將CKIPTagger提供的61種詞性，整理成我們的45類文風句類。文風句類能正確地分辨不同作者之書籍，然同作者之書籍則難以分辨，此結果代表我們的文風句類能正確分辨不同作者之寫作風格及習慣。此外，為檢驗文風句類之有效性，我們將文風句類與其他特徵值，利用SVM分類器與Tanimoto相似度，對同作者之相同類型書籍進行實驗，結果得文風句類較其他特徵值來得穩定，表示文風句類適合作為特徵值。我們針對42本對照組小說進行同作者同本書、同作者不同書、不同作者共三種實驗，得到SVM正確率與Tanimoto相似度之值域範圍作為文風指標。最後，我們將《紅樓夢》分為前80回與後40回，以文風指標為特徵，進行作者分類實驗。所得結論為，無法證明《紅樓夢》僅一位作者，也無法證明有多位作者。

我們將本論文的方法，開發為文風相似度比對之網頁應用程式，以供有興趣者使用。網址如下：<http://par48.cse.nsysu.edu.tw:3000>

關鍵詞：紅樓夢、作者歸屬、分類問題、句類、支援向量機、Tanimoto相似度。

1 簡介

《紅樓夢》全書共一百二十回，前八十回的作者被公認為曹雪芹，而後四十回的作者存在爭議，截至目前，尚未有完全定論，其主要原因乃是作者的手稿有遺失的情形。《紅樓夢》依復刻技術分抄本（手抄）與刻本（活字版印刷）二種。其中抄本著名的有脂硯齋的批語本（甲戌本、己卯本與庚辰本），及其他抄本（如夢覺

本、楊本、舒本、鄭本、卞藏本等）；而著名的刻本有程高本（如程甲本、程乙本）。

1791年，由程偉元花重金網羅各地《紅樓夢》抄本及殘稿，再經高鶚手抄後，以活字印刷付梓發行成一百二十回之程高本，程高本是二百多年來我們所認知的《紅樓夢》面貌。在其書出版之後的二百餘年間，時有學者（如：周汝昌、張愛玲、俞平伯）認為《紅樓夢》後四十回文采不及前八十回，認為現在所看到的《紅樓夢》應為曹雪芹與高鶚的共同創作，而非全為曹雪芹之作。

雖然《紅樓夢》的書籍相當多，但能被我們所用的電子文字檔卻相當少。在缺乏程高本電子文字檔的情況下，本文選擇使用羅鳳珠教授生前整理的庚辰本。這雖然不是更有爭議的程高本，但胭脂齋的庚辰本也存在作者爭議問題。

針對《紅樓夢》的作者，文學家林語堂、高揚、王蒙、白先勇等人和統計學家陳炳藻、李國強和李瑞芳等人，認為《紅樓夢》僅一位作者；而文學家胡適、俞平伯、周汝昌、王佩璋、張愛玲等人和統計學家余清祥[2]、張運良[3]等人、張凱和張明允[4]、蕭天久和劉穎等人[6]，則認為有多位作者。

由於機器學習分類器在各領域表現優秀，近年來關於《紅樓夢》是否僅一位作者問題大部分是使用機器學習分類器，且幾乎所有論文的言論，都是前80回與後40回能分類清楚。但能分類清楚這些的情況，可能僅是由於機器學習的分類能力強。我們使用前人使用過的特徵值，搭配不同種機器學習分類器，得到的結果，不僅前80回與後40回能分類正確，前80回內部分前後、後40回內部分前後，居然也可分類正確，顯示這些模型其實並不能分辨寫作風格差異。

為幫助讀者閱讀本論文，我們羅列本研究的相關設定，如下：

1. 假設《紅樓夢》前八十回為同一整理。本論文不討論曹雪芹是否為《紅樓夢》前八十回作者。
2. 實驗時使用之《紅樓夢》原文，來源為國立臺灣大學圖書館。其版本為經羅鳳珠教授整理過的庚辰本，並已取得該館同意使用。本文所使用的資料集，共有42本對照組小說，包含兩本明清章回：《儒林外史》、《三

*通訊作者信箱: cbyang@cse.nsysu.edu.tw (楊昌彪)。

國演義》，以及40本現代言情小說(共計10作者，如瓊瑤、席絹等，每位作者有4本)，其中2本明清章回小說來源為維基百科，而40本現代言情小說為一般公開網路搜尋而來。為便於研究，我們將每本小說，每約6000字整理成一回，換言之，回與回的斷點跟原著的章節斷點有所不同。

3. 中文詞斷詞與詞性標註，使用中央研究院開發的CKIPTagger，已取得中研院同意使用。
4. 本論文使用到SVM (support vector machine，支援向量機) 之實驗，其SVM模型皆使用Python之sklearn套件。參數設定為 $kernel = rbf$ ，而參數C與 γ 則進行各種組合測試， $C \in \{0.1, 1, 10, 100, 100\}$ 和 $\gamma \in \{1, 0.1, 0.01, 0.001, 0.0001\}$ 。最終，我們設定 $C=1$ 且 $\gamma=0.01$ ，為我們實驗的最佳參數。

2 我們的文風句類分析

我們主要參考1993年中研院《中文詞類分析(三版)》[1]、《句結構樹中的語意角色》[5]及學者陳淑芬《現代漢語否定詞「不」和「沒」的句法、語意和言談/語用特點及其教學應用》[7]等文獻，再使用中研院CKIPTagger套件，將文本斷句與標注詞性(共61種)後，將其歸納成能反映文風差異之45個句類，以下稱之文風句類。

以範例說明我們如何設計文風句類，如下：

1. 附著標記的給予：

在詞與句子中，有些詞需附著在其他詞上才有意義，此類有如狀聲詞：叫一聲['VA', 'Neu', 'Nf']。有人習慣使用「叫」，例：「孩子來，叫媽媽」；有人則習慣使用「叫一聲」，例：「孩子來，叫一聲媽媽」。此特性可適度地體現作者文風，故我們把它當作文風句類的第一種句類。

2. 陳述句與雙重否定句：

有些人習慣使用雙重否定句替代肯定句，例：「我會來。」與「我不會不來。」：

- 我會來。['Nh', 'D', 'D', 'PERIODCATEGORY']
- 我不會不來。['Nh', 'D', 'D', 'VA', 'PERIODCATEGORY']

3. 陳述句與反問句：

有些人習慣使用反問句替代陳述句，例：「身為學生，應該寫作業。」與「身為學生，難道不應該寫作業嗎？」：

- 身為學生，應該寫作業。['VG', 'Na', 'COMMACATEGORY', 'D', 'VC', 'Na', 'PERIODCATEGORY']

- 身為學生，難道不應該寫作業嗎？['VG', 'Na', 'COMMACATEGORY', 'D', 'D', 'VC', 'Na', 'T', 'QUESTIONCATEGORY']

4. 主動句與被動句：

相同的語意，有人習慣使用被動句替代主動句，例：「我可以打你嗎？」與「你可以被我打嗎？」：

- 我可以打你嗎？['Nh', 'D', 'VC', 'Nh', 'T', 'QUESTIONCATEGORY']
- 你可以被我打嗎？['Nh', 'D', 'P', 'Nh', 'VC', 'T', 'QUESTIONCATEGORY']

基於上述原則，我們整理出45類文風句類。至此我們可以分析在一個章回中，每種文風句類各使用的次數，從而產生一個45維度的向量 $X = \langle x_1, x_2, \dots, x_{44}, x_{45} \rangle$ ，其中 $x_i, (1 \leq i \leq 45)$ 表示第*i*種句類於該章回中出現之次數。請注意，一個片語可能同時為不同的句類，因此一個片語被計數可能超過一次。接著我們利用X代表該章回之特徵值，進而對各章回進行分類。

3 Tanimoto相似度

Tanimoto相似度可以展現空間的兩個向量之相似程度，常見於生物、化學領域[8]。2008年，亦有紐西蘭的學者Huang將其用於比對文件間的相似程度[9]。Huang單純將兩文件間的文字分割成為單詞(terms)，並根據各單詞的使用次數做為文件表示向量，然後依公式計算其相似度。此方式雖可鑑別文件相似度，卻由於只看單詞使用量而無法鑑別文風差異。

我們使用代表兩個章回之文風句類向量，套上Huang之計算公式，以求兩章回間之相似程度，計算方式詳見公式1。

$$Tanimoto = \frac{\vec{\alpha} \cdot \vec{\beta}}{\|\vec{\alpha}\|^2 + \|\vec{\beta}\|^2 - \vec{\alpha} \cdot \vec{\beta}} \quad (1)$$

將二文本之文風句類以 a_i 、 b_i 表示， $1 \leq i \leq 45$ 。 $\vec{\alpha} = \langle a_1, a_2, \dots, a_{45} \rangle$ 以及 $\vec{\beta} = \langle b_1, b_2, \dots, b_{45} \rangle$ 代表標準化前的文風句類向量， $\vec{\alpha}$ 和 $\vec{\beta}$ 不可同時為 $\vec{0}$ 。

依Tanimoto相似度公式，若有二文本完全相似，則 $\vec{\beta} = k \cdot \vec{\alpha}$ ， k 為常數，則 $Tanimoto(\vec{\alpha}, \vec{\beta})$ 不一定為1，此不符事實：

$$\begin{aligned} Tanimoto(\vec{\alpha}, \vec{\beta}) &= \frac{\vec{\alpha} \cdot \vec{\beta}}{\|\vec{\alpha}\|^2 + \|\vec{\beta}\|^2 - \vec{\alpha} \cdot \vec{\beta}} \quad (2) \\ &= \frac{\vec{\alpha} \cdot (k \cdot \vec{\alpha})}{\|\vec{\alpha}\|^2 + k^2 \cdot \|\vec{\alpha}\|^2 - \vec{\alpha} \cdot (k \cdot \vec{\alpha})} = \frac{k}{1 + k^2 - k} \end{aligned}$$

表 1: 「我會來。」對「我不會不來。」之文風句類分佈

句類編號	1	2	3	4	5	6	7	8	9	10
句1	0	0	0	0	0	0	0	0	0	0
句類編號	11	12	13	14	15	16	17	18	19	20
句1	0	0	0	0	0	0	0	0	0	0
句類編號	21	22	23	24	25	26	27	28	29	30
句1	0	0	0	0	0	0	0	0	0	0
句類編號	31	32	33	34	35	36	37	38	39	40
句1	0	1	0	0	0	0	0	0	0	1
句類編號	41	42	43	44	45					
句1	0	0	0	0	0					
句類編號	1	2	3	4	5	6	7	8	9	10
句2	0	0	0	0	0	0	0	0	0	0
句類編號	11	12	13	14	15	16	17	18	19	20
句2	0	0	0	0	0	0	0	0	1	1
句類編號	21	22	23	24	25	26	27	28	29	30
句2	0	0	0	0	0	0	0	0	0	0
句類編號	31	32	33	34	35	36	37	38	39	40
句2	0	0	1	0	0	0	0	0	0	1
句類編號	41	42	43	44	45					
句2	1	0	0	0	0					

故在計算相似度前應先將文風句類標準化，標準化公式如下：

$$\hat{a}_i = \frac{a_i - \bar{a}}{\sigma_a} \quad (3)$$

\hat{a}_i 、 \hat{b}_i 是標準化後的 a_i 、 b_i ， $\hat{\alpha}$ 、 $\hat{\beta}$ 代表二標準化後文風句類向量。 $b_i = k \cdot a_i$ ， k 為常數，帶入相似度公式，得

$$\hat{b}_i = \frac{k \cdot a_i - k \cdot \bar{a}}{k \cdot \sigma_a} = \hat{a}_i \quad (4)$$

接著可得

$$Tanimoto(\hat{\alpha}, \hat{\beta}) = Tanimoto(\hat{\alpha}, \hat{\alpha}) = 1 \quad (5)$$

因此，整理得標準化後的Tanimoto相似度公式如下：

$$Tanimoto(\hat{\alpha}, \hat{\beta}) = \frac{\hat{\alpha} \cdot \hat{\beta}}{\|\hat{\alpha}\|^2 + \|\hat{\beta}\|^2 - \hat{\alpha} \cdot \hat{\beta}} \quad (6)$$

相似度值介於0至1，愈接近1代表二文本愈相似。

舉一個同義但寫作風格不同的例子說明如下：

陳述句與雙重否定句：

句1：「我會來。」

我(Nh)會(D)來(D)。(PERIODCATEGORY)

句2：「我不會不來。」

「我(Nh)不會(D)不(D)來(VA)。」

(PERIODCATEGORY)」

兩者文風句類分佈如表1。

Tanimoto值為0.607，代表文風差異大。

4 演算法步驟

我們的演算法的流程圖，如圖1所示，其主要5個步驟如下：

1. 設計文風句類：首先歸納出影響文風之45種文風句類。

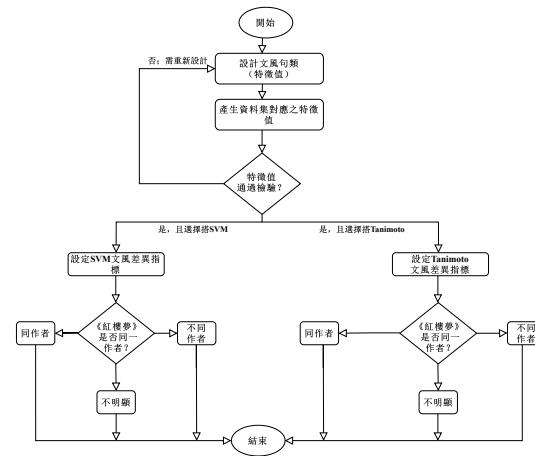


圖 1: 流程圖

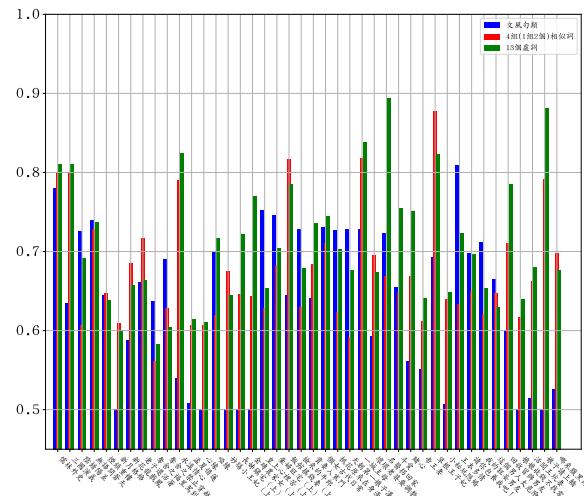


表 2: 文風句類特徵的SVM分類正確率之範圍

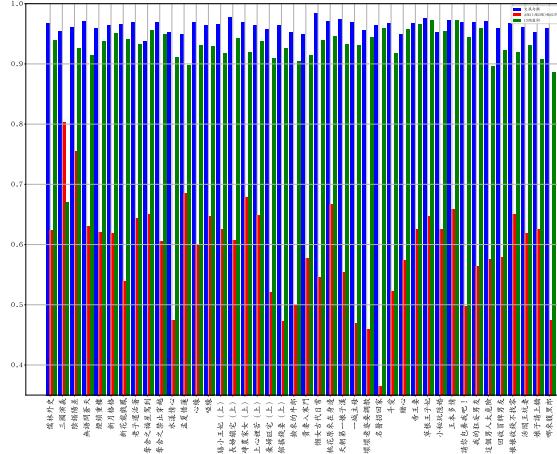


圖 3: 不同特徵值在同作者同本書分類實驗的Tanimoto相似度；橫軸：42本小說

2. 產生資料集及對應之文風句類特徵值：
根據資料集，將資料集中每一本小說，每一回的文字轉為45維的文風句類。

3. 檢驗特徵值：

我們又嘗試將不同特徵值進行同作者分類實驗，亦即將同一本書分成前後二類，理論上SVM分類準確度應接近0.5，亦即無法正確分類（也就是分2類隨機猜中的機率），而且Tanimoto相似度理應接近1（同本書前後文風接近）。但是，根據實驗結果，我們發現，SVM使用不同特徵值，會得到顯著性差異，我們將其畫成圖2及圖3。我們期待能找到一個穩定的特徵值，亦即在所有小說的分類效果大約相同，SVM分類正確率的值域不可太大，Tanimoto相似度的值域亦不可太大。舉一分類正確率之不穩定的例子，若某特徵值在某一本書的前後二類的分類正確率為0.5，但在另一本書卻為0.9，表示此特徵值無法適用於所有小說，因為其分類正確率值域過大。為快速汰除不合適的特徵值，我們以值域大小做為篩選特徵值之依據。

圖2與圖3之實驗，皆以同樣方式（將每一本小說等量分前後兩本）進行實驗。其中SVM經100次5-fold交叉驗證。Tanimoto相似度則是同本小說，前後半各取一回計算其Tanimoto值，最後再算得全部的平均值，成為Tanimoto相似度。除了文風句類特徵值之外，我們使用另外兩組對照特徵值：前人研究[2]所提之4組相似詞（「嗎、麼」、「給、與」、「都、多」、「我們、咱們」）、13個虛詞（在、著、嗎、麼、與、兒、了、的、給、都、多、我們、咱們）。

實驗種類	最小	第25百分位數	中位數	第75百分位數	最大	標準差
同作者同本書分類	0.5	0.529	0.645	0.724	0.809	0.095
同作者不同本書分類	0.594	0.656	0.720	0.787	0.896	0.078
不同作者分類	0.6	0.95	0.971	1	1	0.52

SVM分同作者的值應該要接近0.5，愈遠離0.5表示該特徵值理解文風能力愈差。如圖2所示，除了藍線（文風句類），其他實線均有異常高的波峰表示其理解文風能力差。藍線的SVM正確率值域為[0.5, 0.809]；綠線的值域為[0.582, 0.893]；紅線的值域為[0.560, 0.877]。故僅文風句類較合適作為特徵值。

Tanimoto相似度分同作者的值應該要接近1，愈遠離1表示該特徵值理解文風能力愈差。如圖3所示，除了藍線（文風句類），其他實線均有異常低的波谷表示其理解文風能力差，且整體值域較文風句類值域大。藍線Tanimoto相似度的值域為[0.937, 0.985]；綠線的值域為[0.671, 0.973]；紅線的值域為[0.365, 0.802]。故僅文風句類較合適作為特徵值。

由圖2與圖3可得，特徵值如[2]所提之4組相似詞與13個虛詞皆不適合作為判斷文風的特徵值。與之比較，我們設計的文風句類較適合作為特徵值。

4. 文風指標：

檢驗文風句類後，以SVM與Tanimoto相似度，分別對42本小說進行三種實驗：(1) 同作者同本書分類(每次取1本小說等量分前後兩類)；(2) 同作者不同本書分類(每次取2本相同作者但不同本書)；(3) 不同作者分類(每次取2本不同作者之小說)。¹ 與步驟3檢驗特徵值不同之處是，因檢驗特徵值需考慮噪音 (noise)，故選擇值域；而文風指標則盡量排除噪音，故選擇箱距（第25百分位數和第75百分位數之差值）。

以文風句類特徵進行SVM分類：

每次實驗以5-fold交叉驗證之（以其中隨機4/5的資料作為訓練集，其餘作為測試集），進行100次。實驗結果紀錄於表2。

以文風句類特徵向量計算Tanimoto相似度：

假設欲計算相似度的二本小說分別有m回與n回。二本小說各取一回，計算此二回的Tanimoto值，則此二本小說共有 $m \times n$ 個Tanimoto值。我們以這 $m \times n$ 個Tanimoto值的平均作為此二本小說之Tanimoto相似度。

¹ 註：我們曾試著再將不同作者分類實驗細分成二類：(1) 不同作者不同本書分類及 (2) 不同作者同本書分類。在搜集 (2) 的過程時，因鮮少有電子書屬多位作者且標明分工，故我們轉向搜集名著與其續作。我們一共搜集了四本書（含原作《鏡花緣》與《西遊記》與續作《續鏡花緣》與《續西遊記》）。依結果而論，(2) 的值接近(1)，說明即便是不同作者寫同本書（原作與續作），兩者寫作風格仍有明顯不同，故將兩者合併統稱為不同作者分類實驗。

表 3: 文風句類向量的
Tanimoto相似度之範圍

實驗種類	最小	第25百分位數	中位數	第75百分位數	最大	標準差
同作者同本書分類	0.937	0.957	0.964	0.969	0.985	0.009
同作者不同本書分類	0.925	0.939	0.946	0.956	0.974	0.011
不同作者分類	0.628	0.882	0.912	0.930	0.962	0.051

表 4: 以文風句類特徵對《紅樓夢》進行SVM分類

編號	實驗方式	實驗描述	分類正確率	判定
1	實驗組	第1到第80回為一類， 第81到第120回為另一類， 進行5-fold交叉驗證100次	0.77	無法判斷
2	對照組	第1到第40回為一類， 第41到第80回為另一類 (公認為同作者)， 進行5-fold交叉驗證100次	0.66	小於0.787， 判定為同作者

我們歸納得出，文風句類特徵的SVM分類正確率，若小於0.787，可視為來自同作者（文風差異小）；否則視為不同作者（文風差異大）。文風句類特徵向量的Tanimoto相似度，若小於0.882，可視為來自不同作者（文風差異大）；否則視為同作者（文風差異小）。

5. 《紅樓夢》是否同一作者之實驗：

我們將《紅樓夢》的前八十回稱為《紅樓夢A》，後四十回稱為《紅樓夢B》。我們以文風句類特徵，對《紅樓夢A》與《紅樓夢B》進行SVM分類實驗，結果紀錄於表4。我們也以文風句類特徵向量，計算《紅樓夢A》與《紅樓夢B》Tanimoto相似度，結果紀錄於表5。

為了進行比較，我們亦將《三國演義》拆成前八十回與後四十回，稱為《三國演義A》、《三國演義B》。接著，我們將同作者、不同作者、《紅樓夢》、《三國演義》之SVM分類正確率、Tanimoto相似度的箱型圖分別繪製於圖4與圖5。

由表4與表5之數值得，我們的結論是，無法斷定《紅樓夢》是否僅有一位作者，亦無法論斷其有多位作者。進一步的證據，由圖4，可以看到《紅樓夢A》與《紅樓夢B》的SVM分類正確率，介於相同作者與不同作者之間。由圖5，可以看到《紅樓夢A》與《紅樓夢B》的Tanimoto相似度，也介於相同作者與不同作者之間。此外，我們特別將《紅樓夢》120回彼此之間的Tanimoto值呈現於圖6。由圖可見，對角線為最深（2個同一回的Tanimoto值為1），除此之外沒有特別深淺處，也印證無法斷定《紅樓夢》是否僅一位作者或有

表 5: 以文風句類特徵向量對《紅樓夢》計算Tanimoto相似度

編號	實驗方式	第一類	第二類	Tanimoto相似度	判定
1	實驗組	第1到80回	第81到第120回	0.933	無法判斷
2	對照組	《紅樓夢》共120回	《三國演義》共120回	0.844	小於0.882， 判定為不同作者

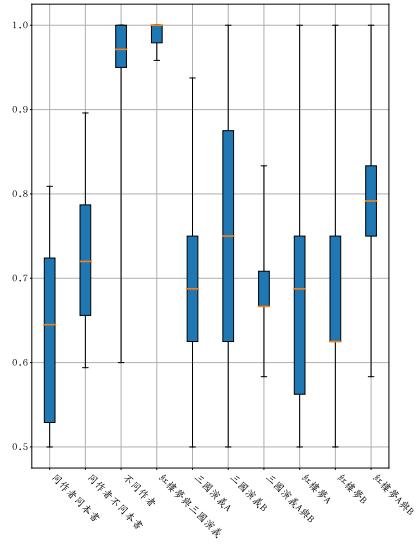


圖 4: SVM分類正確率之箱型圖

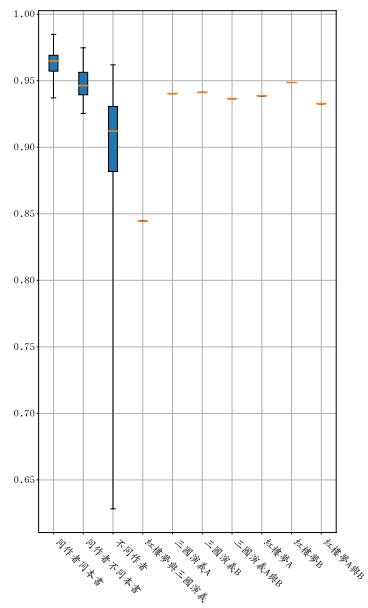


圖 5: Tanimoto相似度之箱型圖

參考文獻

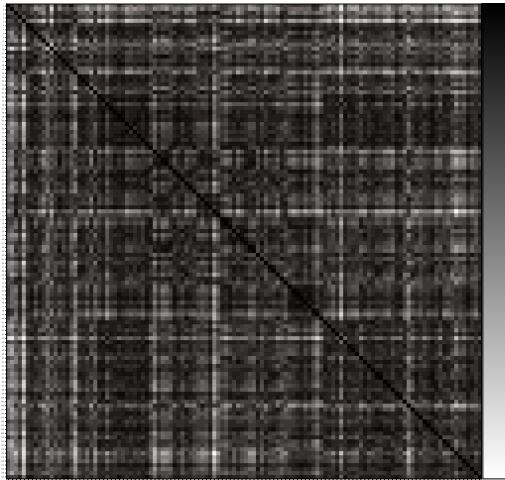


圖 6: 《紅樓夢》120回彼此之間的Tanimoto值

多位作者。

5 結論

我們由中研院CKIP小組制定的61種詞性，依我們的文學知識歸納出45類文風句類，並檢驗文風句類是否合適作為特徵值。接著使用SVM分類器與計算Tanimoto相似度，進行三種類型實驗：同作者同本書、同作者不同書、不同作者，統計其值域，並作為文風指標，判斷《紅樓夢》之作者爭議問題。由實驗結果，我們認為《紅樓夢》全書是否有超過一位作者之爭議無法定論，故無法推定僅一位作者，亦無法推定有多位作者。

我們實驗的資料集，除古籍書典外、亦加入現代小說作為實驗資料集，並驗證文風指標之正確性。此外，我們的實驗設計，考量了世界上所有類型的書（同作者同本書、同作者不同本書、不同作者同本書、不同作者不同本書），可惜的是資料集並不多。期待在未來，搜集更多資料集，以優化我們的文風指標。我們的文風模型將來亦可能應用於以下領域：歌詞是否抄襲、一本書是否僅一位作者、寫作自動評分、機器翻譯、自動分詞、問答等。

- [1] 詞庫小組, 中文詞類分析(三版). 中央研究院資訊科學研究所, 1993.
- [2] 余清祥, “統計在紅樓夢的應用(註),” 國立政治大學學報, Vol. 76, p. 303, 1998.
- [3] 張運良、朱禮軍、喬曉東、張全, “基於句類特徵的作者寫作風格分類研究,” 計算機工程與應用, Vol. 45, No. 22, pp. 129–131, 2009.
- [4] 張凱、張明允, “基於svm的《紅樓夢》寫作風格研究,” 貴陽學院學報 (自然科學版) , Vol. 6, No. 1, pp. 55–57, 2011.
- [5] 詞庫小組, 句結構樹中的語意角色. 中央研究院資訊科學研究所, 2013.
- [6] 肖天久、劉穎, “《紅樓夢》詞和n元文法分析,” 現代圖書情報技術, No. 4, pp. 50–57, 2015.
- [7] 陳淑芬、陳力綺, “現代漢語否定詞「不」和「沒」的句法、語意和言談/語用特點及其教學應用,” *UST Working Papers in Linguistics (USTWPL)* , 2017.
- [8] D. Bajusz, A. Rácz, and K. Héberger, “Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations?,” *Journal of Cheminformatics*, Vol. 7, No. 1, pp. 1–13, 2015.
- [9] A. Huang *et al.*, “Similarity measures for text document clustering,” *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, Vol. 4, pp. 9–56, 2008.