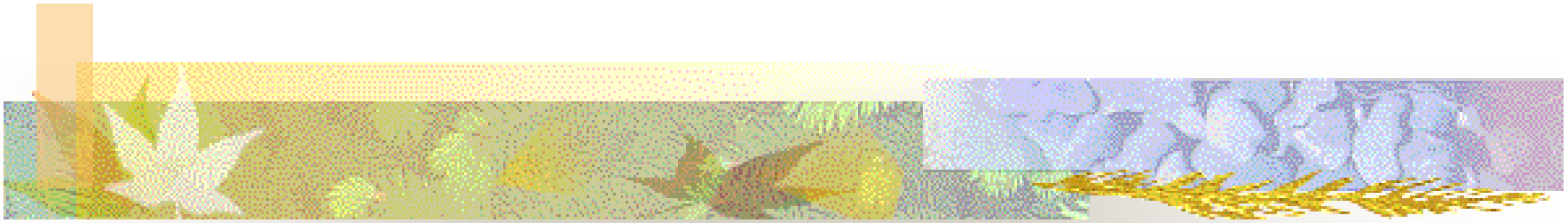
從文言到白話：《新青年》雜誌語言變化統計研究



余清祥：政治大學統計系

2018年12月4日

《新青年》雜誌

- 由陳獨秀等人創刊於1915年，是五四運動前後重要的中文刊物。

→一般認為《新青年》是文言文到白話文轉變的代表性刊物。

→文白轉變從何時開始？轉變過程中有何特點、何時完成文白語言的轉變？

- 《新青年》總共十一卷，有人只承認一至九卷，因為後兩卷為中共發行。

文言文vs.白話文

- 現代漢語與古代漢語的區別之一在於書面用語，或古代的文言文及現代的白話文。
- 五四運動是白話文取代文言文的關鍵，其中倡導白話文最力的莫過於《新青年》雜誌，出版時間剛好跨越五四運動前後，可作為驗證五四運動時期的文體變化。

文字資料的前置分析

- 處理文字資料時，統計分析包含以下步驟：
- Data Collection
- Text Parsing and Transformation

→摘錄、清理、由NLP定義變數等，包括斷句、篩選相關資料段落、定義關鍵字詞。

- Text Filtering

→挑選合適關鍵字詞。

- Text Mining

→Clustering, Classification, Association, and Link Analysis.

文言文、白話文的比較

比較	文言文	白話文/現代中國語文
長短	言簡意賅	較長篇
出處用法	書面語為主	「我手寫我口」為主，亦經修飾
語感	古雅精煉	通俗易明
文法詞組次序	彈性較大	次序明確
用詞1	單字已有獨立意思	二字詞為主
用詞2	一字多用	異字異用
用詞3	之	的
句末助語詞	已、矣、乎、也.....	了、吧、啊、嗎.....
標點	標點少而簡，句讀為主	標點繁多
經典例	《桃花源記》、《醉翁亭記》、《庖丁解牛》、《出師表》、《六國論》.....	魯迅《吶喊》自序、朱自清《綠》、冰心《紙船》、舒乙《香港：最貴的一棵樹》.....
流傳	限於曾學習文言的人，須有一定傳統文學修養，但可於東亞通行	一般中小學生也能看懂，廣傳於華文世界
習法	背誦為主，輔以字詞拆解	字詞拆解為主，文法分析輔助

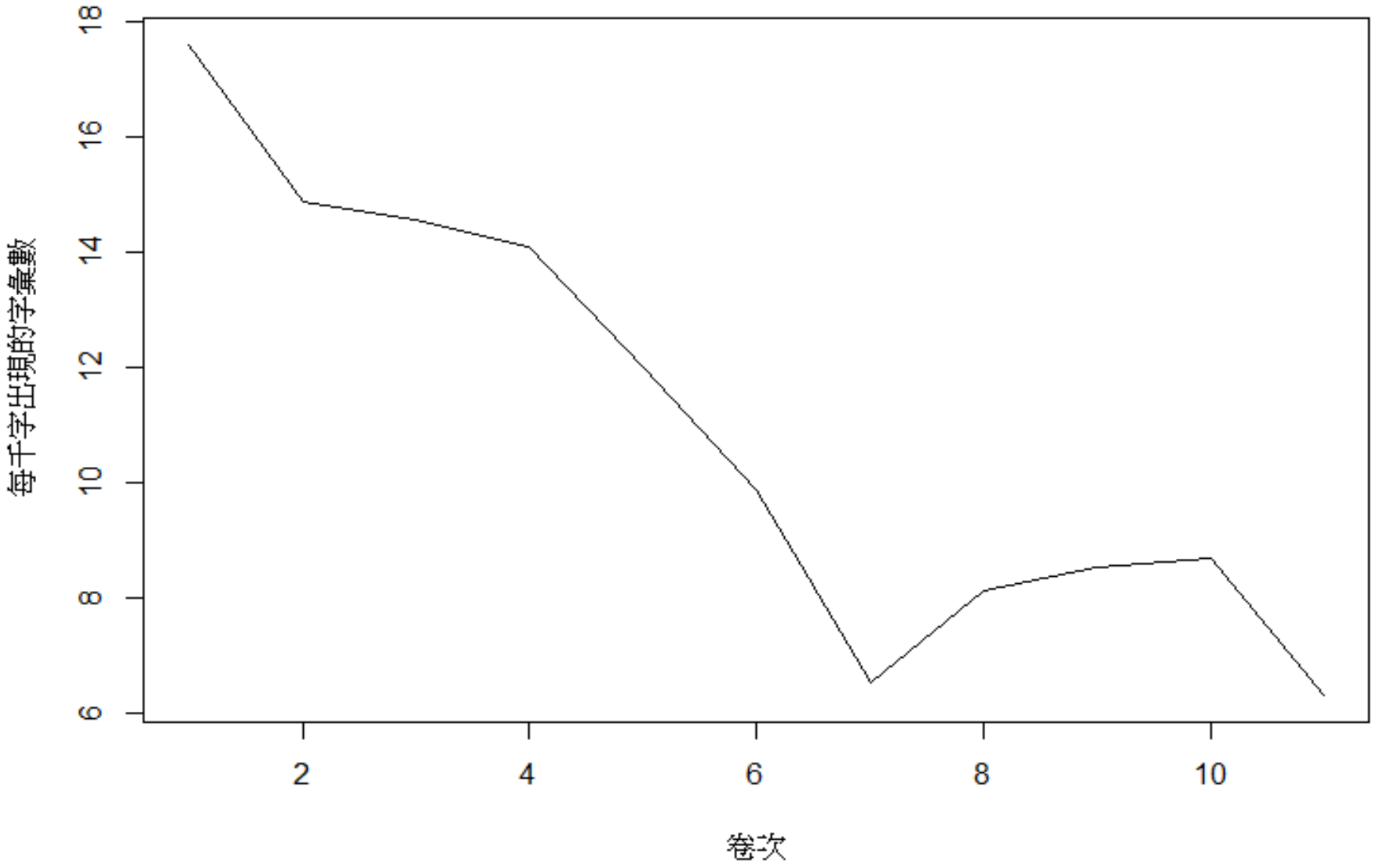
本研究的文字分析方向

- 除了常見的探索性分析，本研究也加入常見的虛字、各卷的常見字詞、每個句子的字數作為輔助變數。
- 採用文言文、白話文各十個常用虛字：

→矣、乎、焉、歟、哉、耳、豈、之、乃、無（文言文）；

→的、是、們、個、了、和、麼、著、嗎、吧（白話文）。

	總字數	字彙數	Simpson Index	Entropy
第 1 卷	**248,833**	**4,379**	**0.004568**	**6.654036**
第 2 卷	291,848	4,344	0.004500	6.649539
第 3 卷	290,038	4,227	0.004954	6.541824
第 4 卷	305,020	4,298	0.004172	6.539378
第 5 卷	343,519	4,125	0.004672	6.461579
第 6 卷	389,407	3,848	0.005749	6.348547
第 7 卷	**586,942**	**3,850**	**0.006053**	**6.328604**
第 8 卷	461,731	3,753	0.006035	6.320355
第 9 卷	437,748	3,745	0.005574	6.322103
第 10 卷	342,778	2,980	0.005700	6.177278
第 11 卷	**489,223**	**3,093**	**0.005712**	**6.212699**



《新青年》各卷每千字出現的新字彙數

Standardized Type/Token Ratio

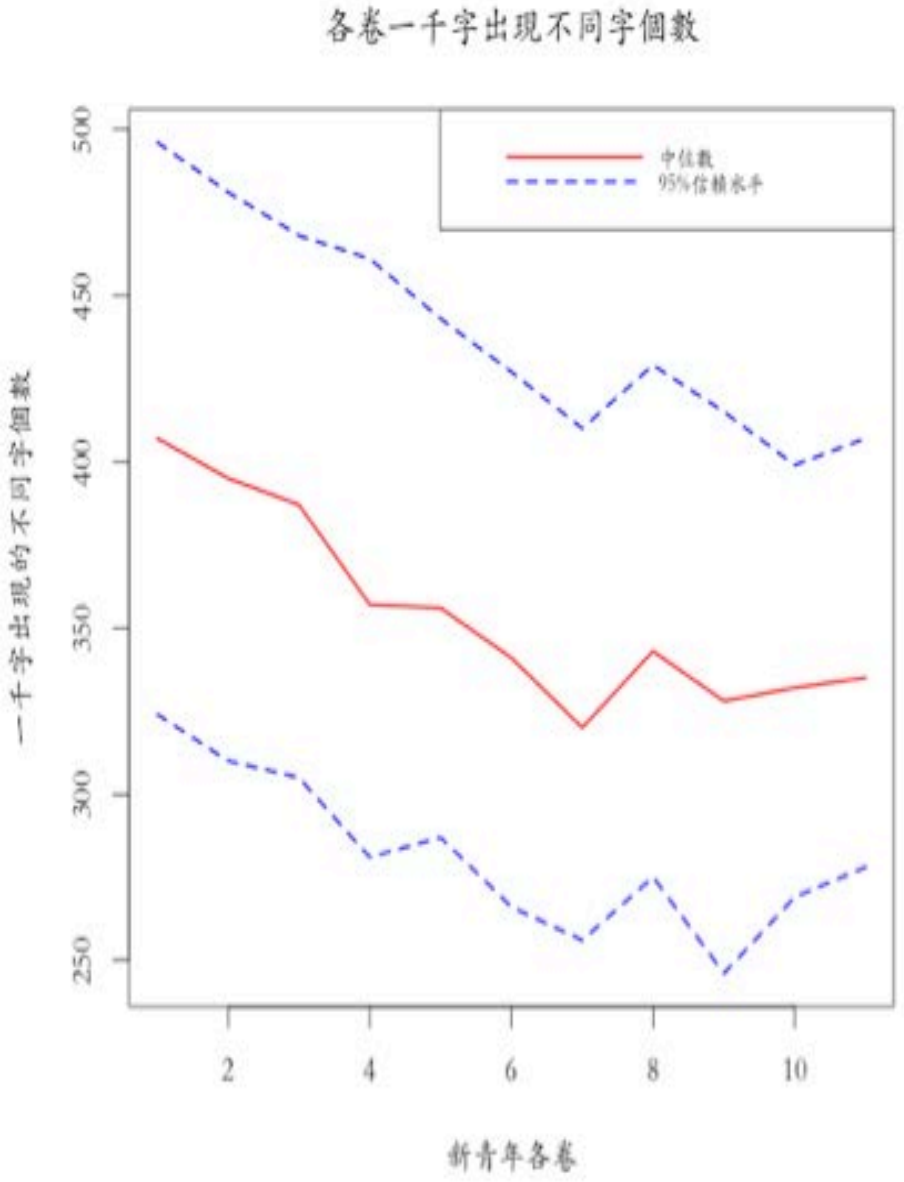
rank	word	freq	rank	word	freq	rank	word	freq	rank	word	freq
1	we	6	17	asleep	1	33	impressions	1	49	seem	1
2	and	5	18	at	1	34	instant	1	50	so	1
3	them	5	19	beliefs	1	35	is	1	51	the	1
4	are	3	20	brief	1	36	just	1	52	then	1
5	can	3	21	but	1	37	later	1	53	things	1
6	they	3	22	call	1	38	metaphor	1	54	thinking	1
7	to	3	23	coming	1	39	mull	1	55	this	1
8	again	2	24	concepts	1	40	notions	1	56	thoughts	1
9	as	2	25	described	1	41	occasions	1	57	times	1
10	in	2	26	endure	1	42	opinions	1	58	values	1
11	on	2	27	fall	1	43	other	1	59	variously	1
12	a	1	28	going	1	44	over	1	60	views	1
13	act	1	29	handle	1	45	perceptions	1	61	well	1
14	all	1	30	have	1	46	put	1	62	what	1
15	an	1	31	however	1	47	refer	1	**TOTAL**		**87**
16	aside	1	32	ideas	1	48	return	1			

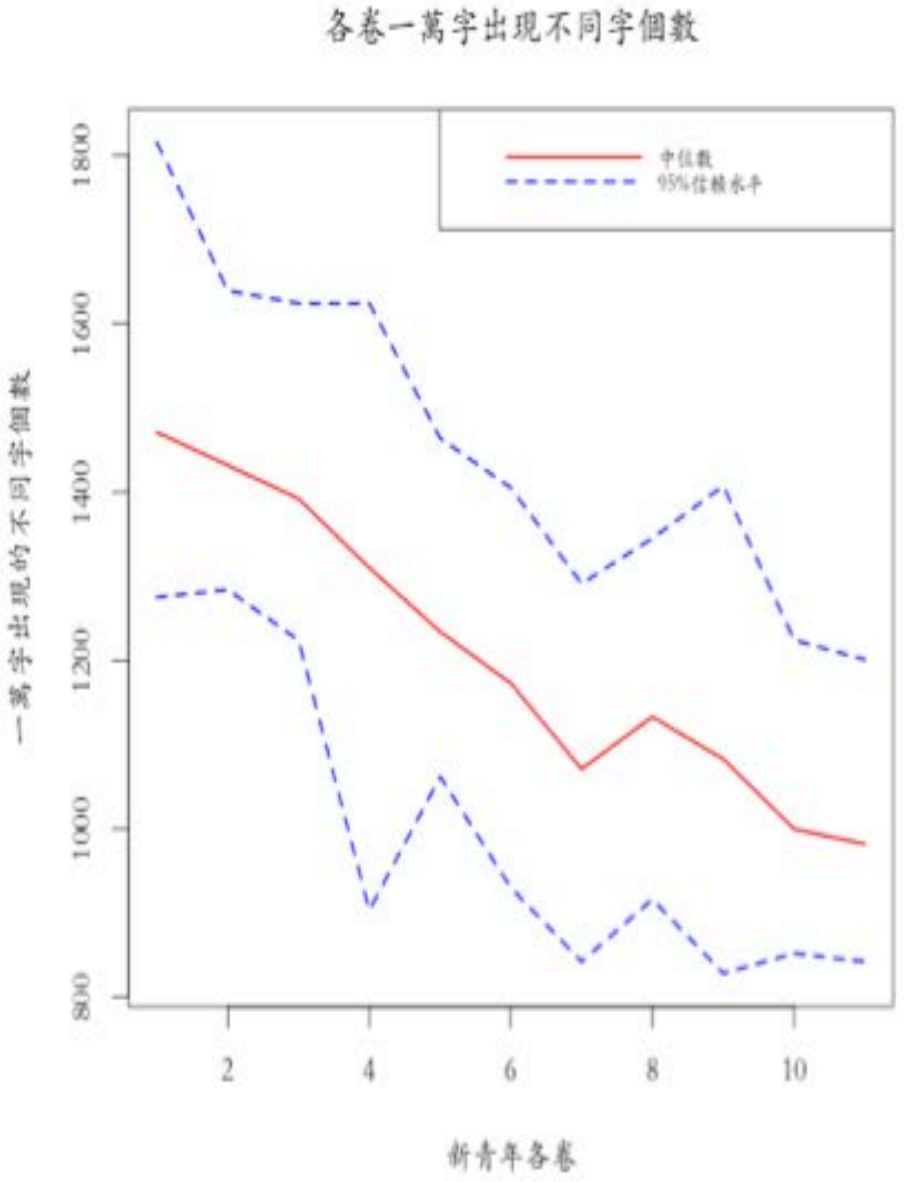
We see, then, that of the total of 87 tokens in this text there are 62 so-called *types*. The relationship between the number of types and the number of tokens is known as the *type-token ratio (TTR)*. For Text 1 above we can now calculate this as follows:

type-token ratio = (number of types/number of tokens) * 100

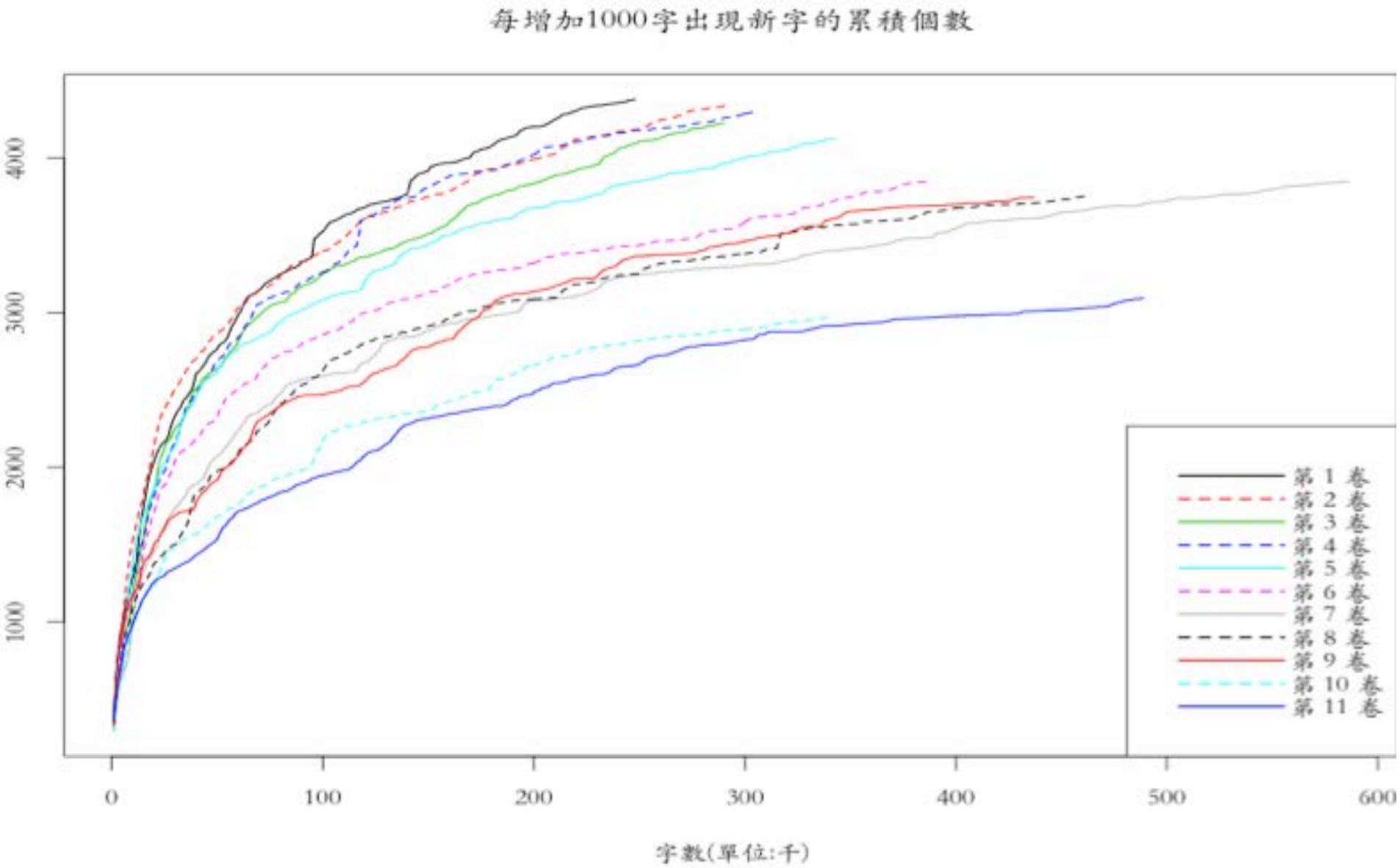
= (62/87) * 100 = **71.3%**

《新青年》初探：文章字彙數變化





「新青年」初探：新字出現頻率



關連性分析

- 關連性分析可提供詞彙之間的關係，輔助辨別文章特性和文章間的關係。

→列連表、用有字詞的交集與聯集等；

→相關係數、相似指數（若為連續資料）。

- 關連性分析可加上文意，或是單純評估字詞間的連結程度。

→文言文、白話文的虛詞(Functional Words)；

→根據專家意見選擇重要關鍵詞，並計算這些詞彙間的關連，以及與研究目標間的關連。

《新青年》初探：各卷相同單字詞

卷別	1	2	3	4	5	6	7	8	9	10	11
1	10	9	9	6	6	4	3	3	3	4	3
2	9	10	9	7	7	5	4	4	4	5	3
3	9	9	10	6	6	4	3	3	3	4	3
4	6	7	6	10	10	8	7	6	7	6	4
5	6	7	6	10	10	8	7	6	7	6	4
6	4	5	4	8	8	10	8	8	9	6	4
7	3	4	3	7	7	8	10	8	8	5	5
8	3	4	3	6	6	8	8	10	9	5	5
9	3	4	3	7	7	9	8	9	10	5	5
10	4	5	4	6	6	6	5	5	5	10	7
11	3	3	3	4	4	4	5	5	5	7	10

各卷十個常用字相同數

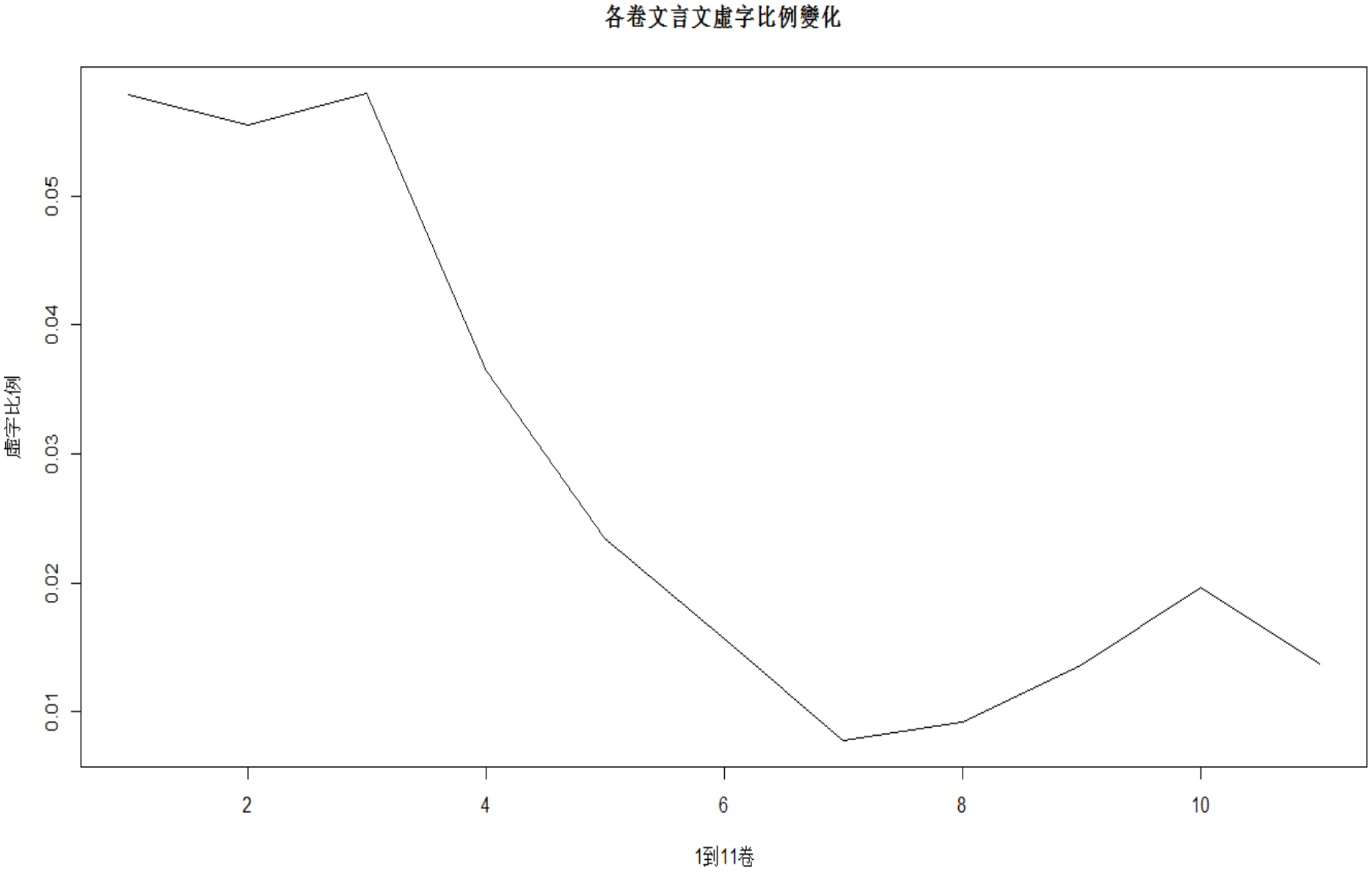
《新青年》初探：各卷相同雙字詞

卷別	1	2	3	4	5	6	7	8	9	10	11
1	10	5	3	1	1	1	1	2	2	1	1
2	5	10	3	1	3	2	1	3	2	2	2
3	3	3	10	4	5	4	1	2	1	1	1
4	1	1	4	10	8	7	4	4	4	2	2
5	1	3	5	8	10	8	4	5	5	3	3
6	1	2	4	7	8	10	4	5	5	3	3
7	1	1	1	4	4	4	10	5	5	3	3
8	2	3	2	4	5	5	5	10	8	5	5
9	2	2	1	4	5	5	5	8	10	6	6
10	1	2	1	2	3	3	3	5	6	10	8
11	1	2	1	2	3	3	3	5	6	8	10

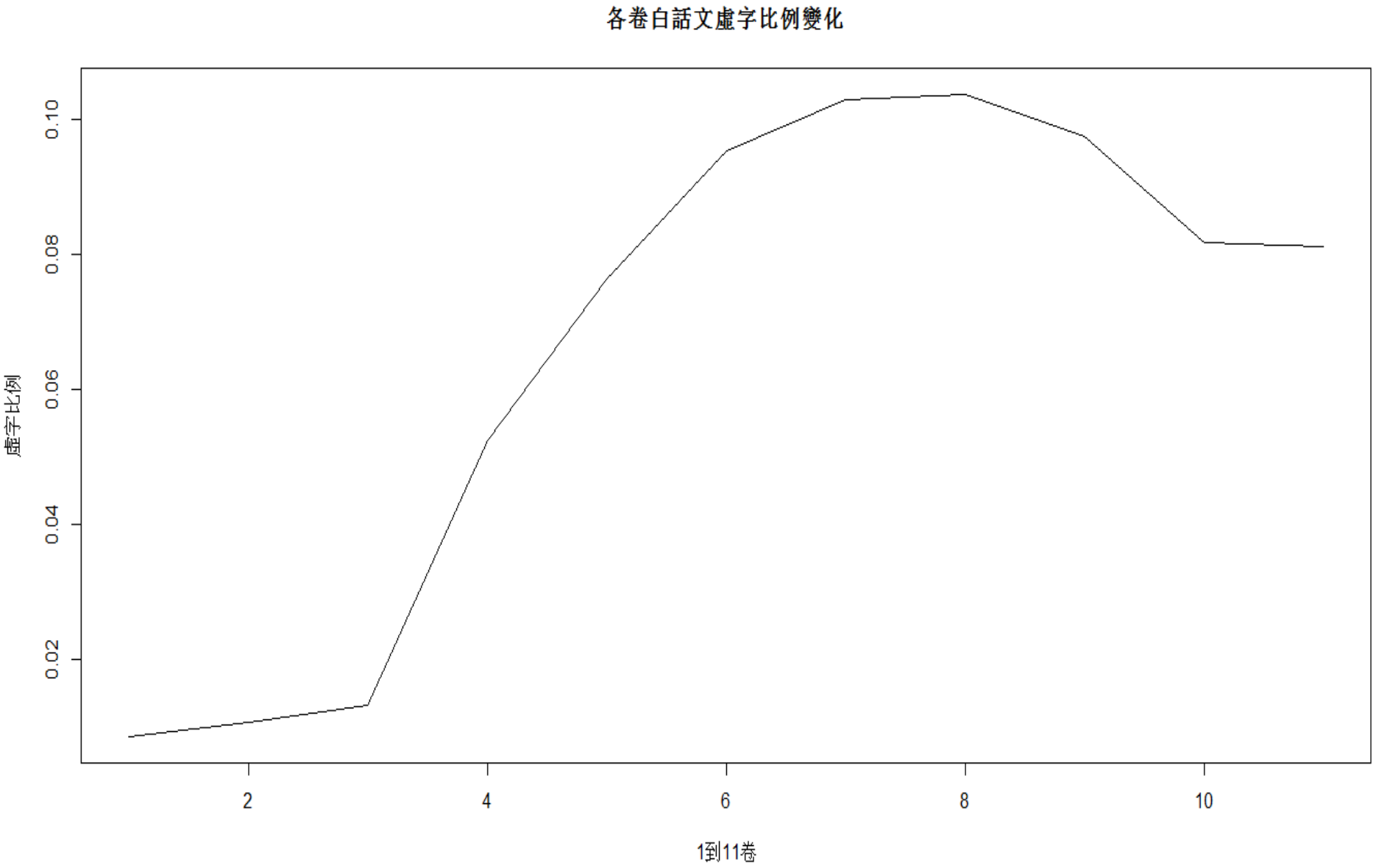
各卷十個常用雙字詞相同數

文言文、白話文的虛字及其出現比例

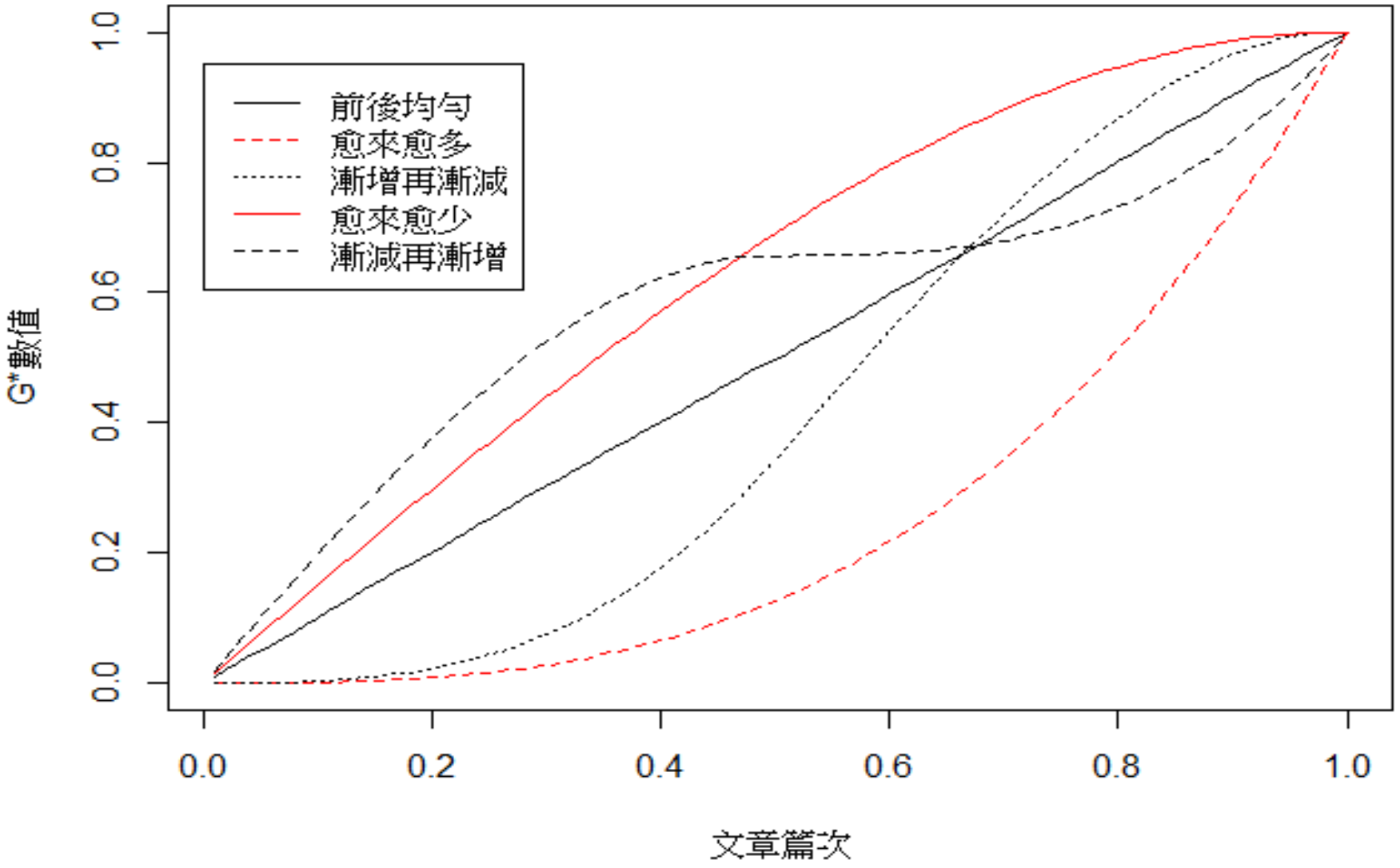
	Classical Chinese	Modern Chinese
Words	矣乎焉歟哉耳豈之乃無	的是們個了和麼著嗎吧
Volume 1 Proportion	3.6%	0.7%
Volume 7 Proportion	0.5%	8.8%
Total Proportion	2.4%	7.3%



文言文虛字出現比例變化

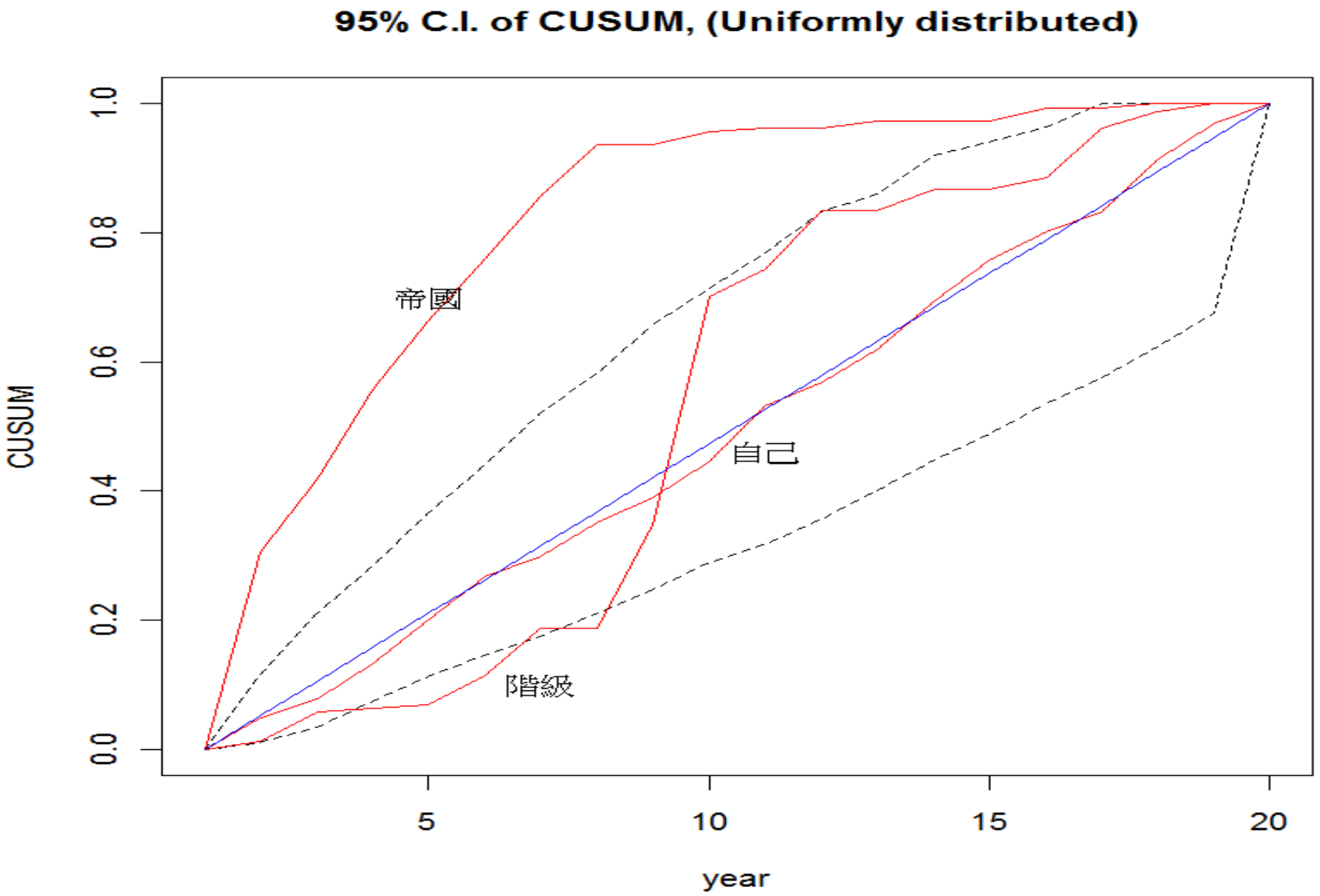


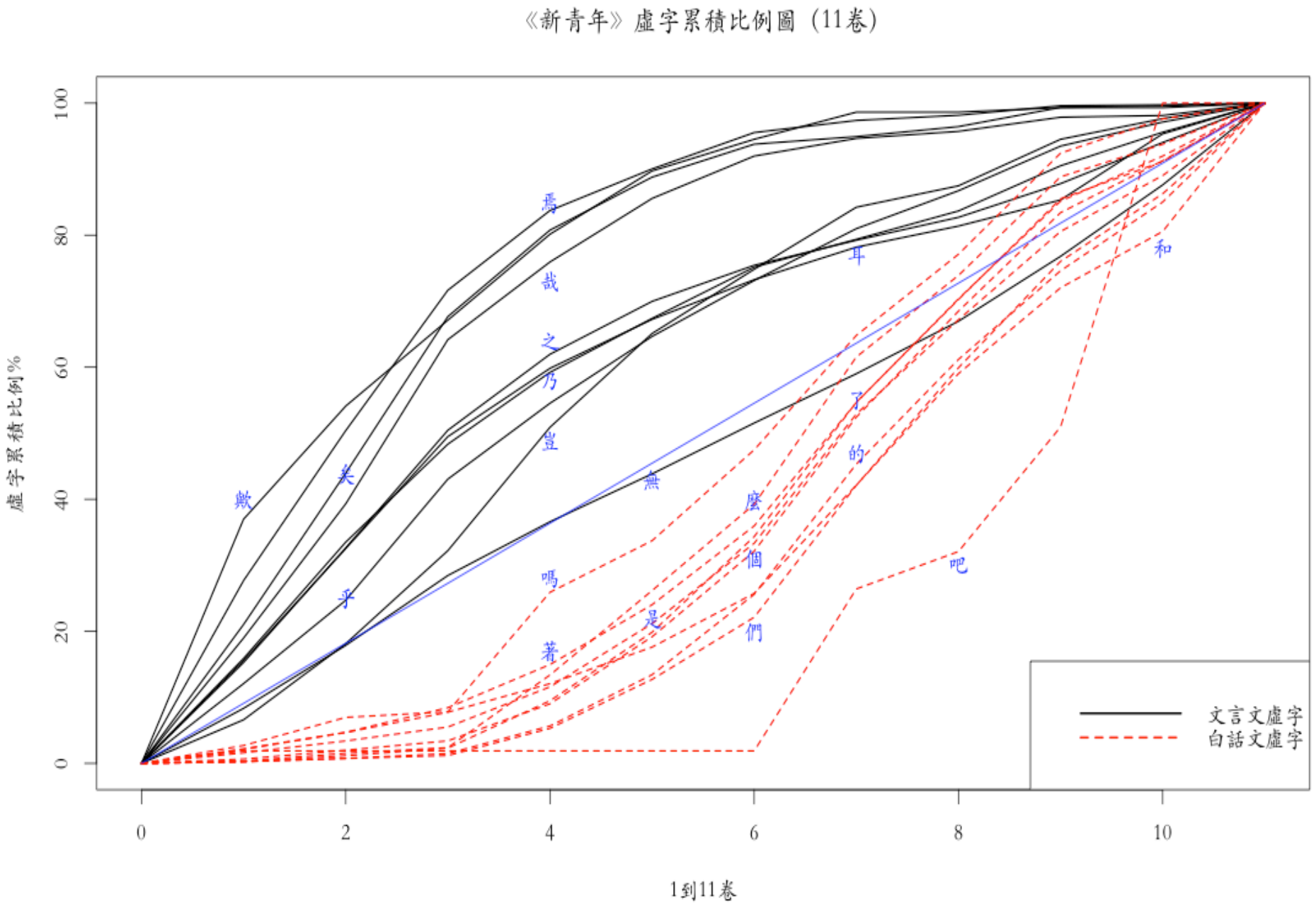
白話文虛字出現比例變化



不均度曲線類型代表的意義

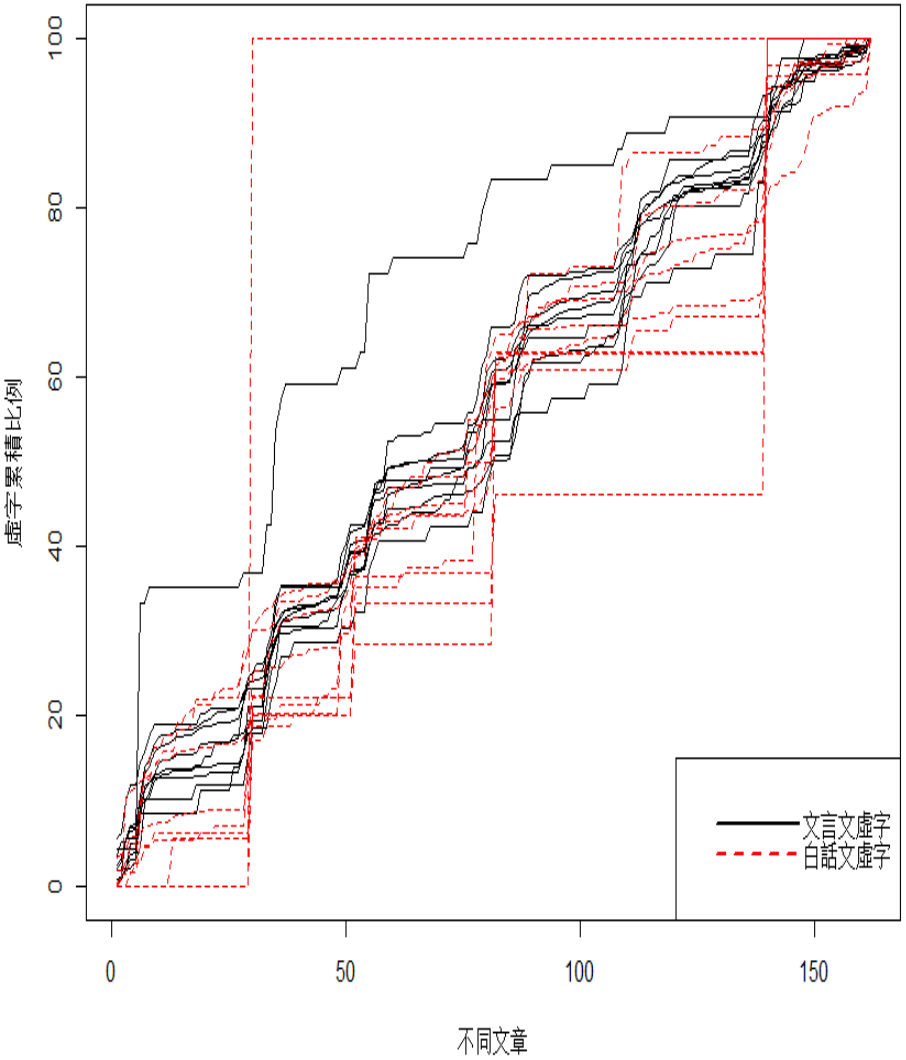
不均度的範例（人民日報）


95% C.I. of CUSUM, (Uniformly distributed)
CUSUM
0.0
0.2
0.4
0.6
0.8
1.0
帝國
自己
階級
5
10
15
20
year

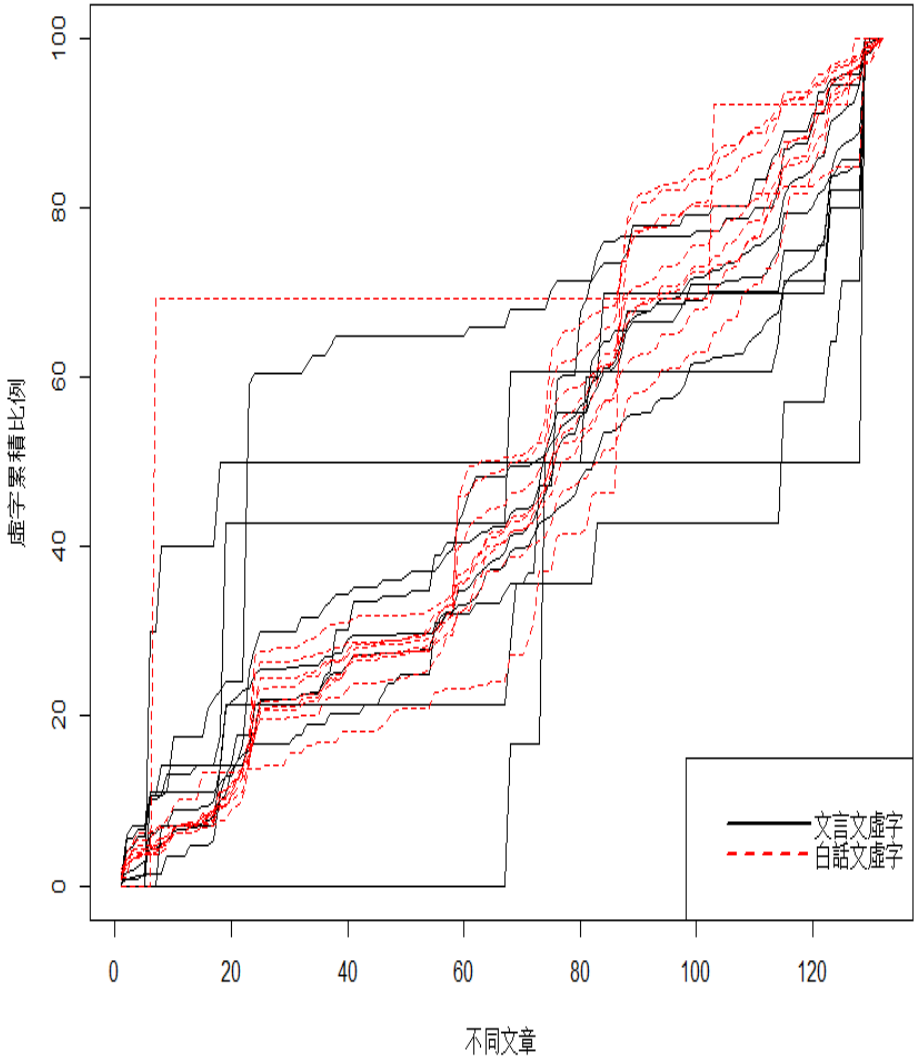


文言、白話虛詞的趨勢變化（不均度）

第一卷



第七卷



黑、紅各為文言、白話虛字，接近對角線為常見字。

（第一、七卷各偏向文言文、白話文）

虛字在各卷的吉尼係數變化趨勢

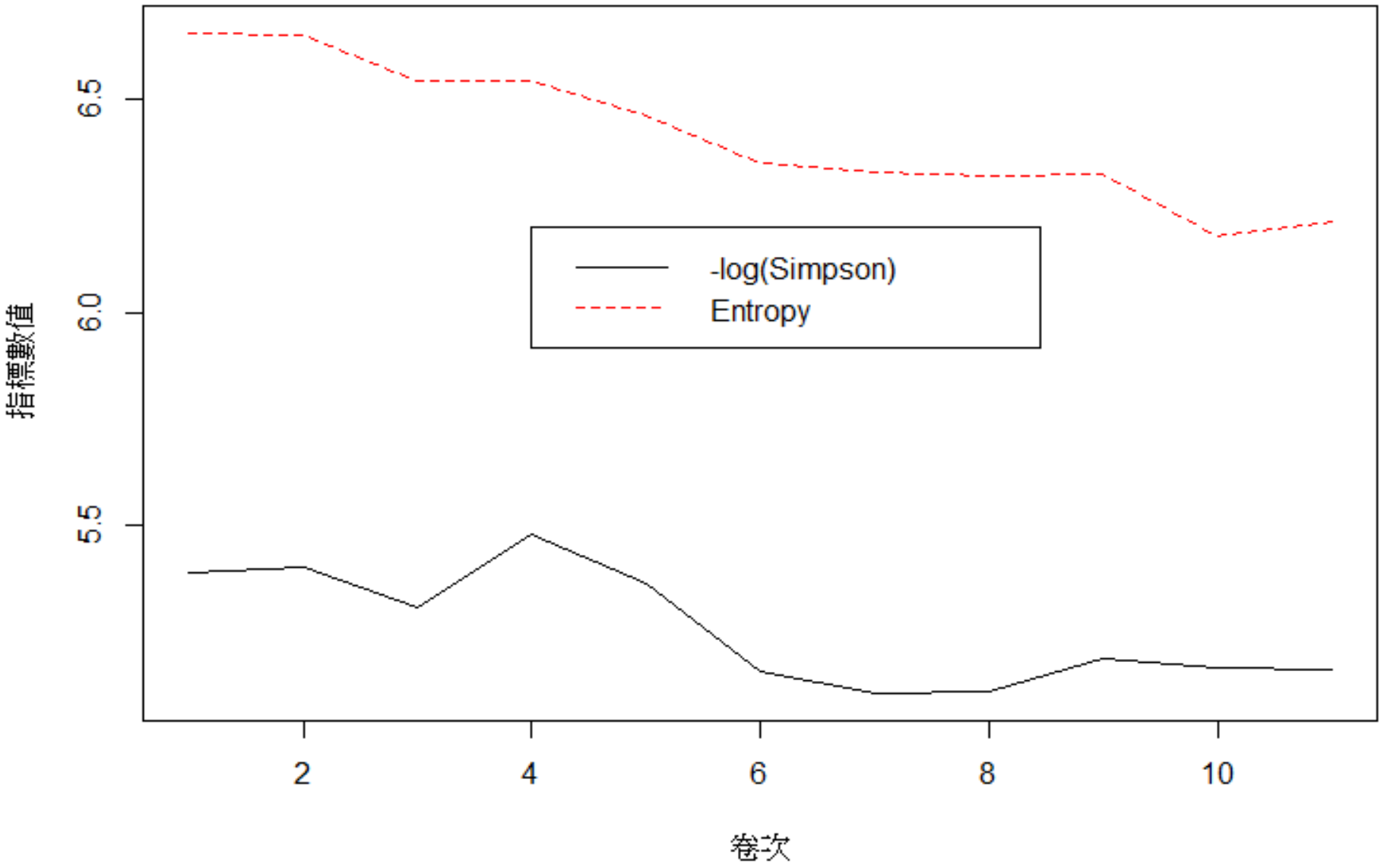
第1到11卷的變化	文言虛字個數	白話虛字個數
明顯上升	6	0
明顯下降	0	9
上升、下降交錯	4	1

單字（詞彙）的探索性分析

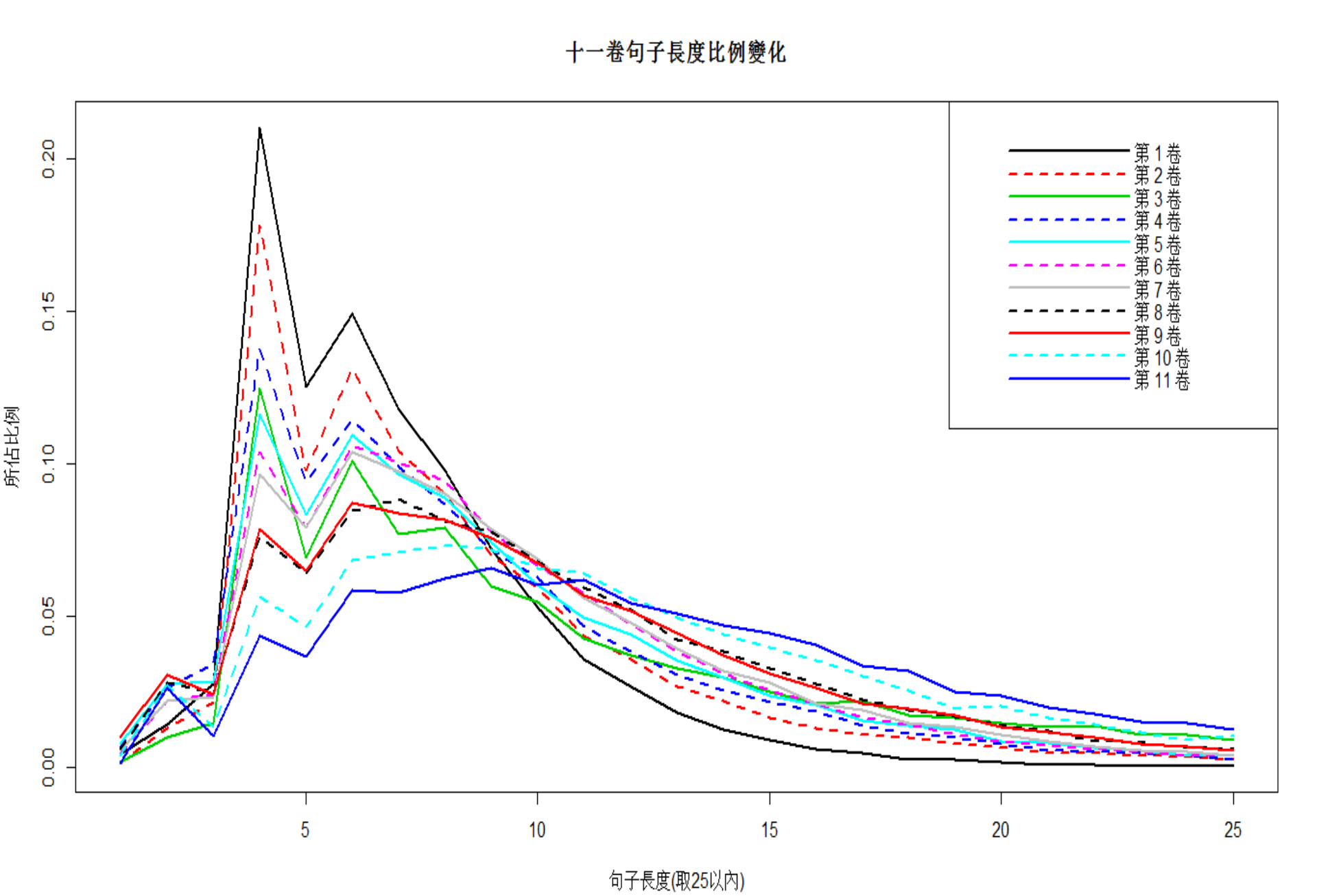
- 單字、詞彙的基本特性整理，包括單字及詞彙種類（即字彙數）及其出現機率（多項分配），以目視(EDA)可觀察出基本特性與差異，做為導引後續分析的參考。
- 不均度（或物種豐富度、生物多樣性）等之測量值，也可用於描述單一母體的特性、或是比較兩個母體間的關連。

→熵(Entropy) $= -\sum_i p_i \log(p_i)$ 、 $\theta_S = \sum_i p_i^2$ 。

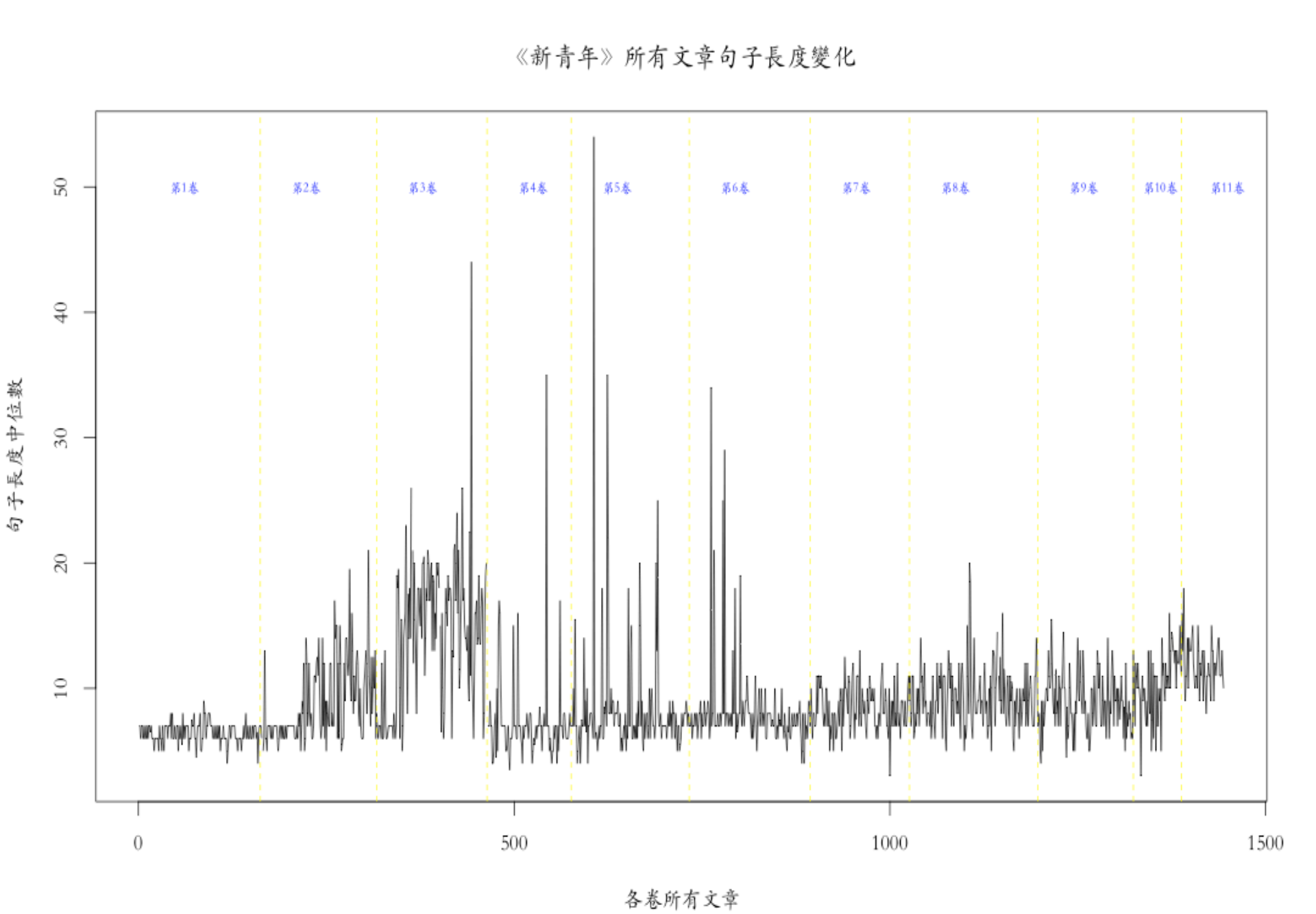
註：可參考生物、經濟相關指標（吉尼係數）。



《新青年》各卷多樣性趨勢圖



《新青年》各卷每句字數分佈圖



《新青年》每篇文章中位數字數分佈圖

研究小結

- 《新青年》雜誌用字趨向於集中（多樣性降低），或不需要認識許多字就能閱讀，符合白話文的推廣精神。
- 除了虛詞、字彙總數及其使用比例、句子長短之外，尋找其他潛在客觀指標，可反映出語言表徵。

→以資料驅動定義關鍵字詞；

→套入生態變遷、物種演化？

- 也可使用分類方法分析各卷間的風格變化。

研究方法

- 主成份分析

→透過線性變換有效縮減變數維度，並保持各主成份變數間彼此獨立。

- 羅吉斯迴歸模型

→適用於二元目標變數，選擇適合的自變數估計、預測分類結果：

$$\log\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k$$

文言文和白話文的分類分析

- 訓練樣本：第1卷「1」、第7卷「2」

→測試樣本：第4卷

- 分析步驟：

1、計算各卷變數數值（34個變數）

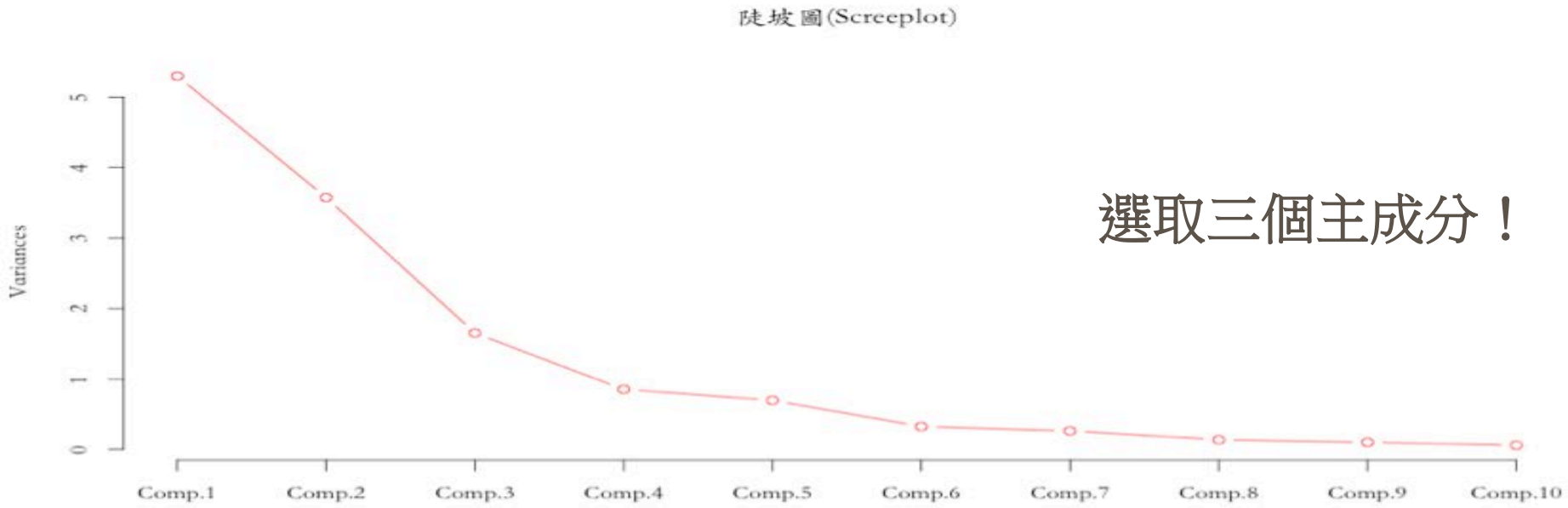
2、主成分分析提取變數主成分

3、對第1、7卷進行羅吉斯迴歸（訓練模型）

4、以訓練模型對第4卷進行預測

主成分分析

主成分	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th
標準差	2.3019	1.8908	1.2859	0.9223	0.8337	0.5683	0.5095	0.3672	0.3148	0.2414	0.1681	0.1228	0.094
變異數所佔比例	0.4076	0.275	0.1272	0.0654	0.0535	0.0248	0.02	0.0104	0.0076	0.0045	0.0022	0.0012	0.0007
累積比例	0.4076	0.6826	0.8098	0.8752	0.9287	0.9535	0.9735	0.9839	0.9915	0.996	0.9982	0.9993	1



選取三個主成分！

羅吉斯迴歸（訓練樣本）

	係數估計值	標準差	Z 值	P 值
截距	-0.7467	0.5772	-1.294	0.195783
第 1 主成分	3.3461	0.7095	4.716	2.40E-06
第 2 主成分	1.3619	0.3853	3.534	0.000409
第 3 主成分	1.8529	0.659	2.812	0.00493

$$log\frac{p}{1-p} = -0.7467 + 3.3461 * \text{第 1 主成分} + 1.3619 * \text{第 2 主成分} + 1.8529 * \text{第 3 主成分}$$

羅吉斯迴歸的估計結果（表列）

		預測	
		白話	文言
標記	白話	129	3
	文言	2	160

註：準確率98.30% ！

分類錯誤文章：

所屬卷別	文檔名稱	語言類別	標記類別	機率預測值	預測類別
第一卷	Y0003.000083.txt	白話	1	0.277571652	0
第一卷	Y0003.000017.txt	白話	1	0.050371456	0
第七卷	Y0003.000931.txt	文言	0	0.83390559	1
第七卷	Y0003.000944.txt	文言	0	0.900878947	1
第七卷	Y0003.000960.txt	白話	0	0.900878947	1

註：若加入工判斷（區隔文言文、白話文），模型只判錯一篇文章，準確率高達99.66%！

- 為避免過度配適，採用十次的十折交叉驗證。

	模型預測準確率	
	平均值	標準差
訓練集	96.10%	0.07%
測試集	95.95%	0.31%

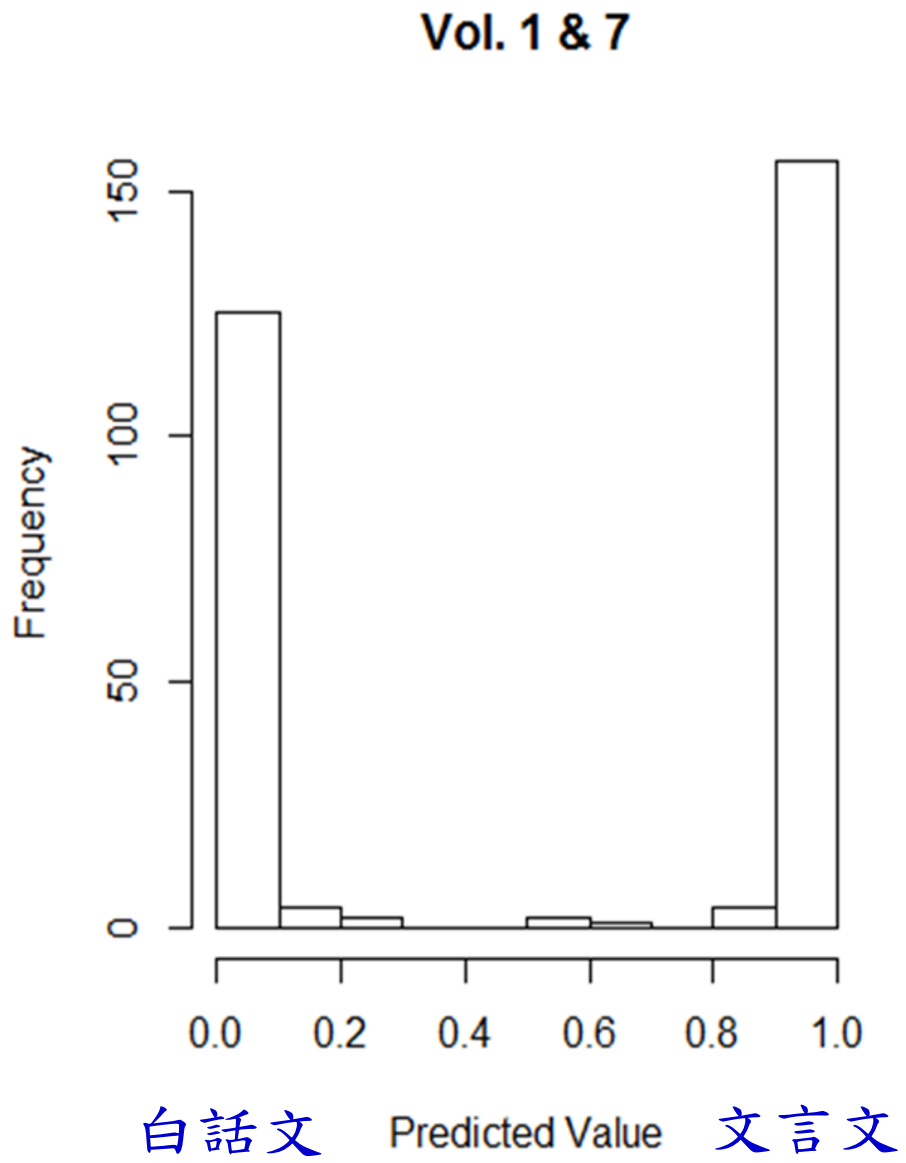
註：模型可視為穩健、可靠。

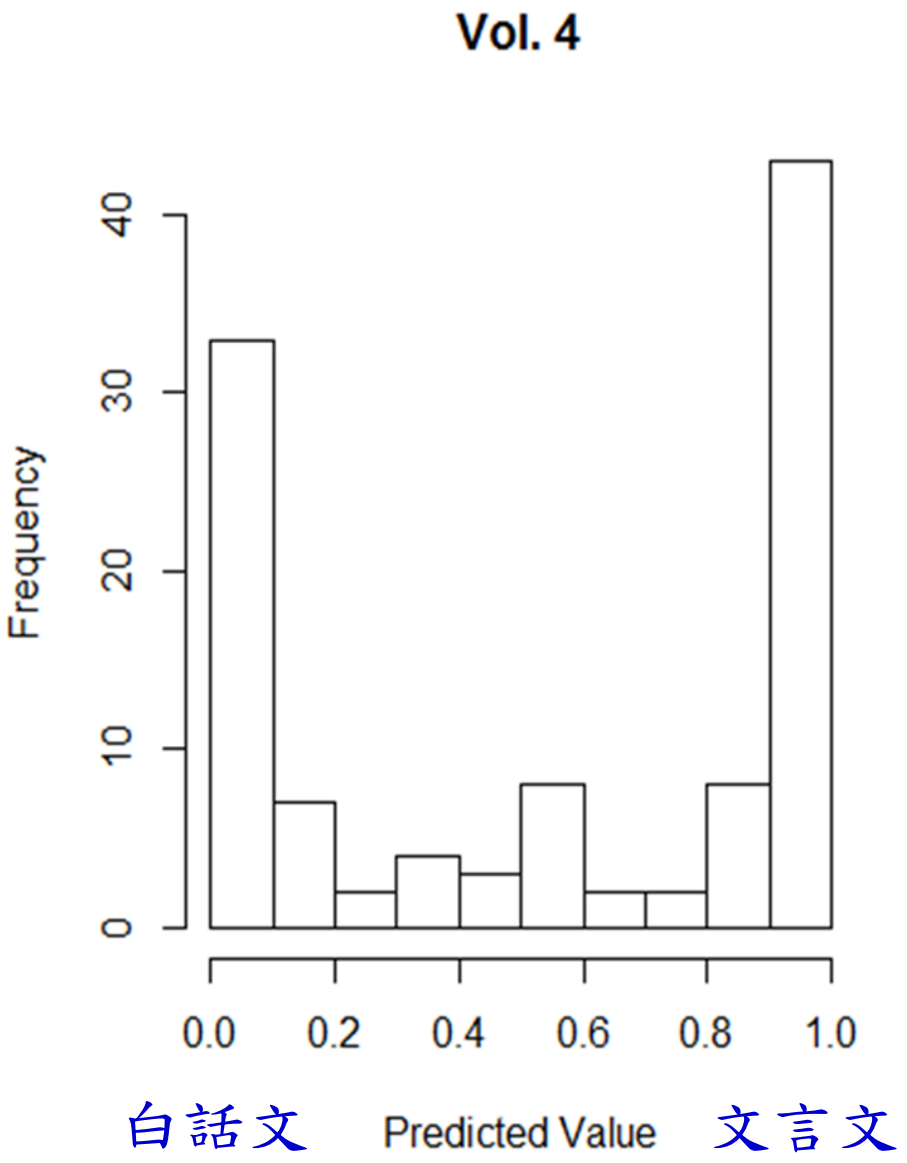
羅吉斯迴歸的預測結果（表列）

		預測	
		白話文	文言文
真實	白話文	34	0
	文言文	13	32

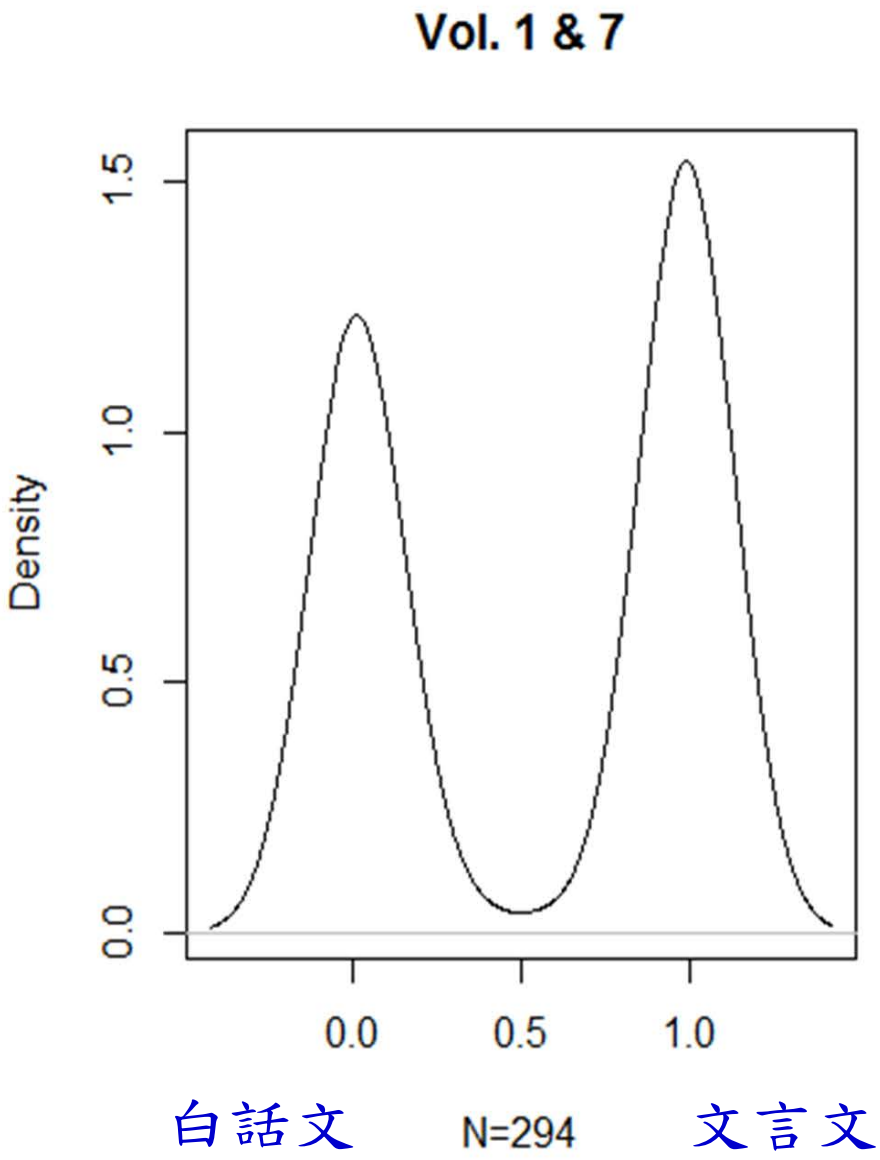
註：準確率83.54% ！

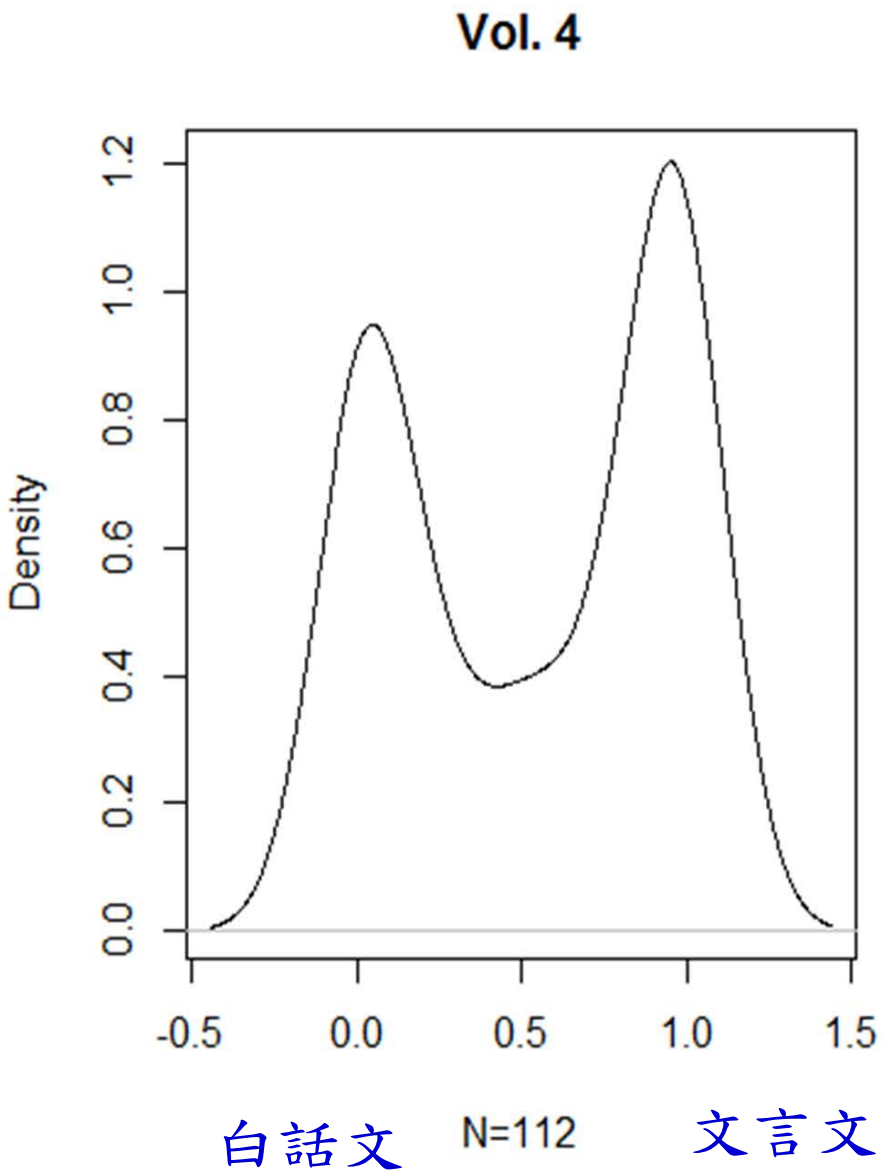
羅吉斯迴歸的分類結果

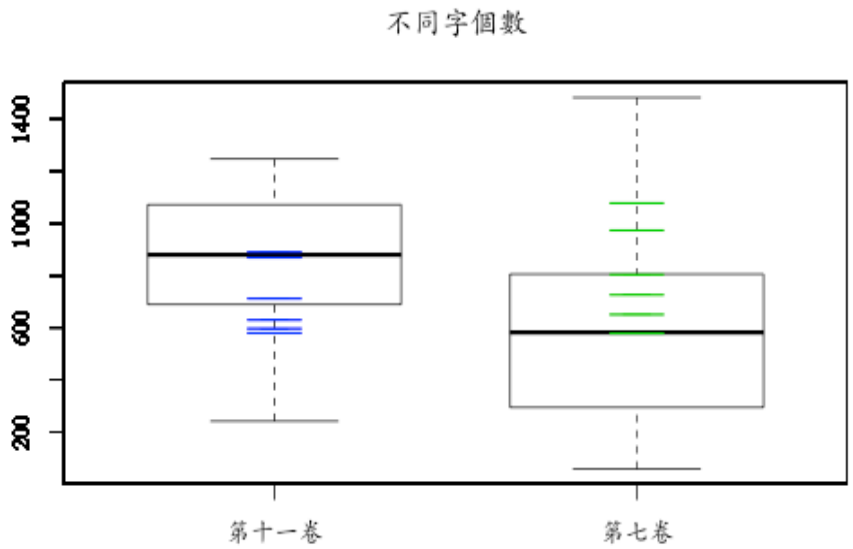

Vol. 1 & 7
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
白話文
Predicted Value
文言文

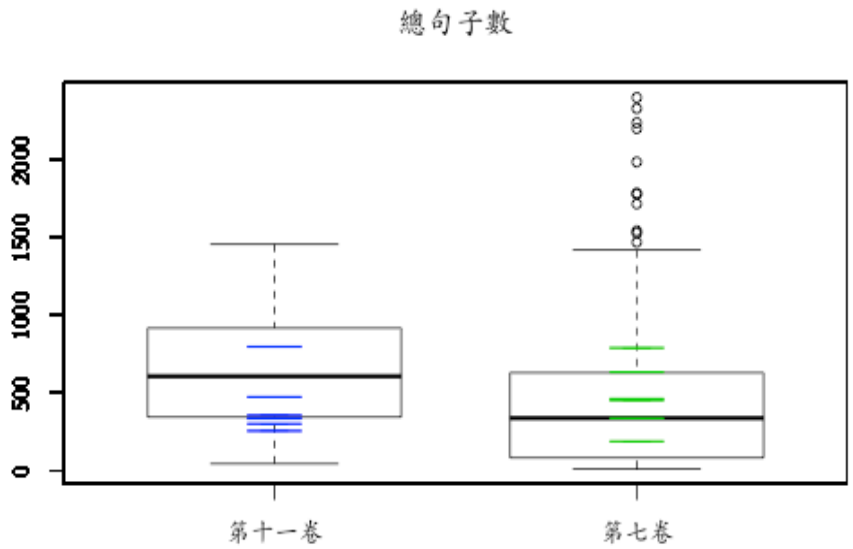

Vol. 4
Frequency
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
白話文
Predicted Value
文言文

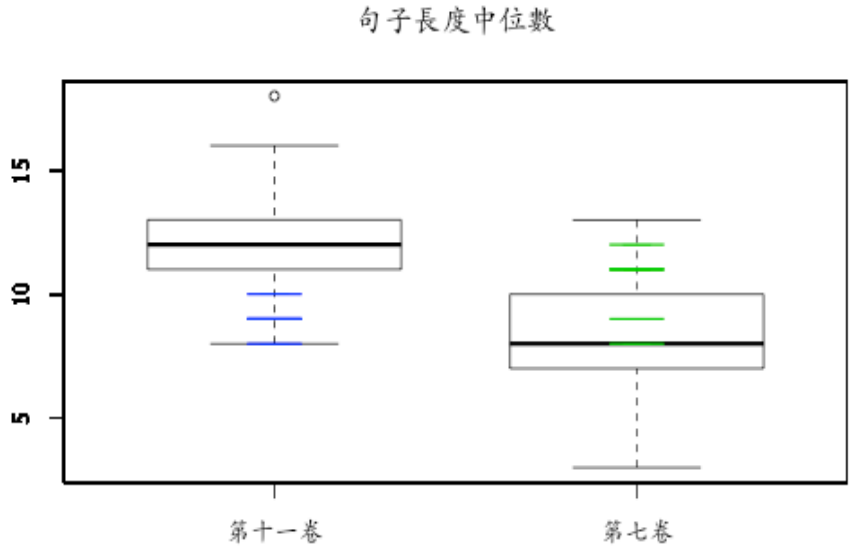
羅吉斯迴歸的平滑化結果

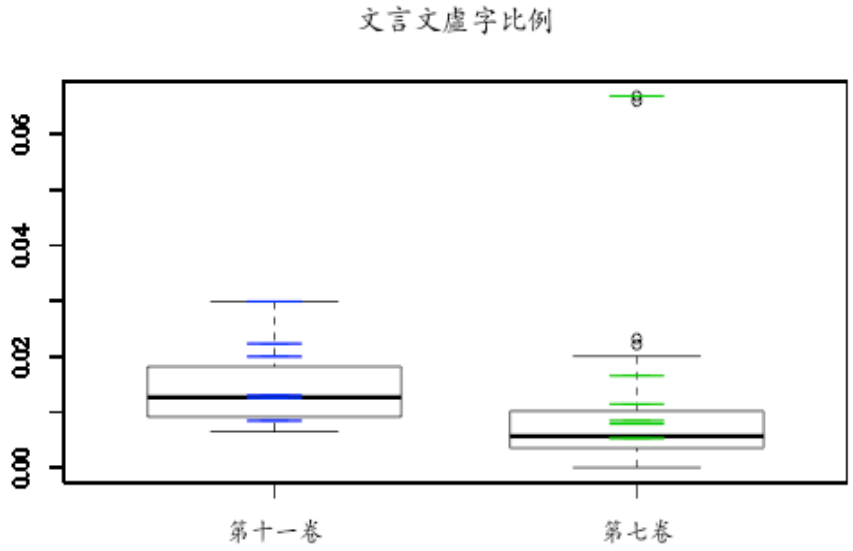












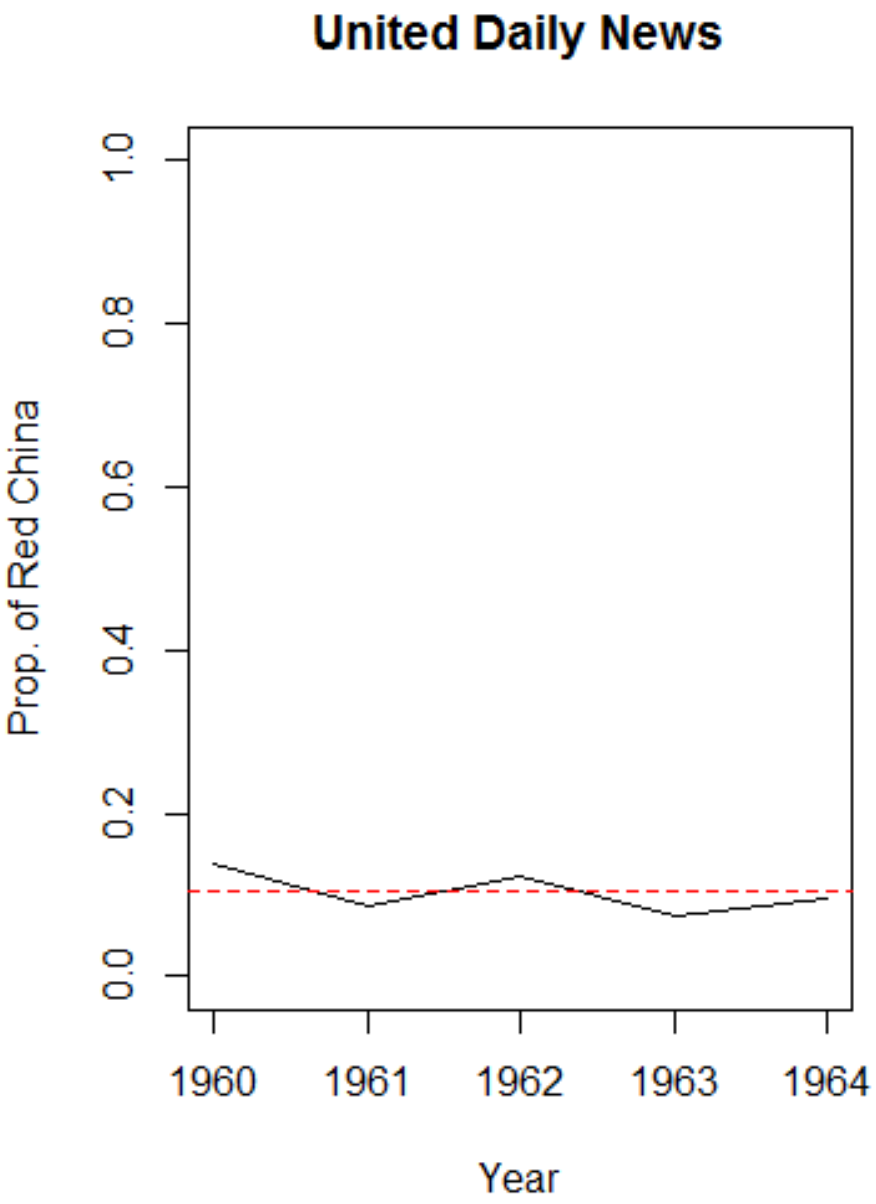
第7、11卷四個文本特徵變數的比較

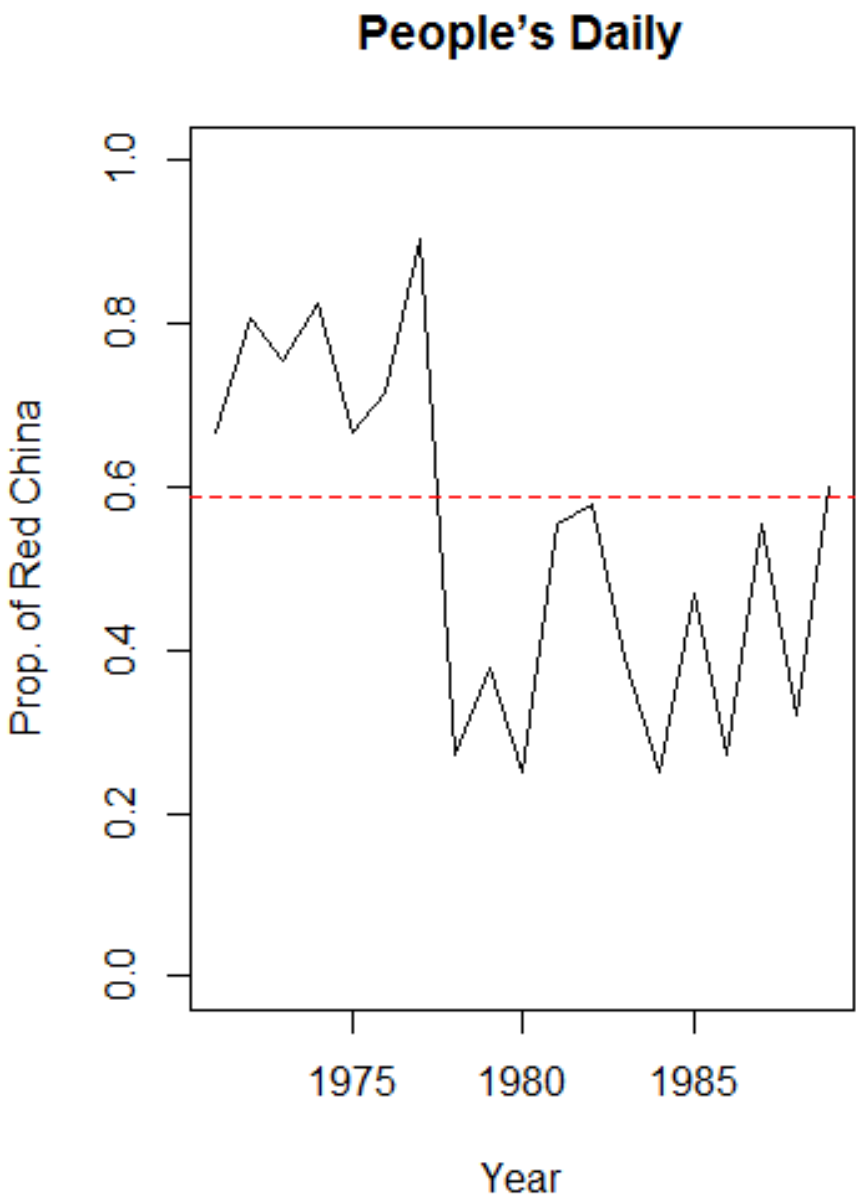
第7卷、第11卷分類結果

		Fit	
		第7卷	第11卷
True	第7卷	126	6
	第11卷	6	50

第7卷、第11卷交叉驗證

	估計結果	
	平均數	標準誤
Training	93.20%	0.10%
Testing	92.17%	0.67%

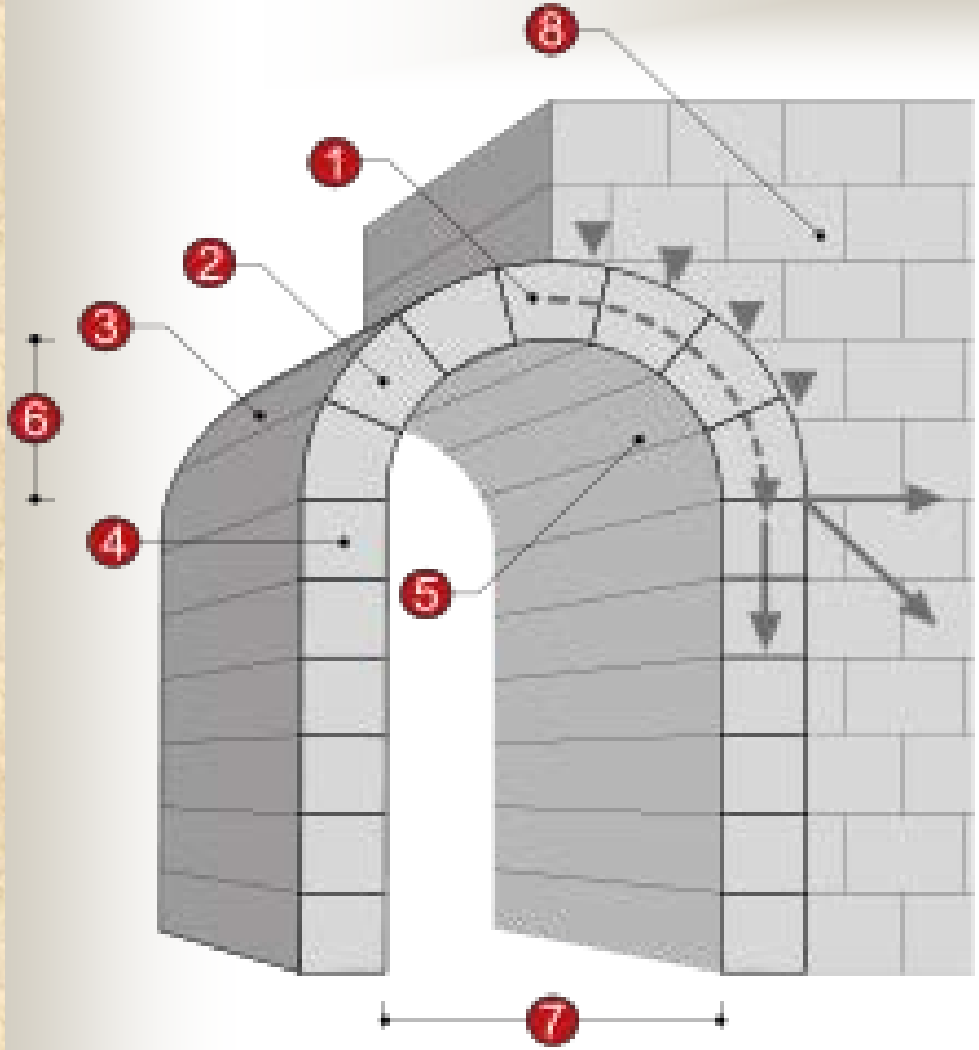




聯合報、人民日報的各年度分類結果

生態變遷、基因圖譜

- 建立核心關鍵字詞的分析技術之後，套入生物棲息地的概念，以物種多樣性的角度判斷寫作風格、觀念等之特性是否隨時間變動。
- 除了探索物種間的競爭與合作、棲息地的演變對物種存活的影響，也可考慮社交網絡(Social Network)，分析棲息地之間的關連（例如：搜尋引擎Google的連結；或是關連性Association）。



(1)基石（或拱心石）

- 基石平衡門上石塊應力，使整個拱門結構堅實，少了基石會使拱門崩垮。

→知名基石物種包括海獺、馬糞海膽（後壁湖）、美洲豹。

註：基石物種或稱為關鍵物種。

關鍵詞與統計分析

- 關鍵詞(Keywords)猶如統計分析的變數(Variable)，根據研究目的及問題定義，先從原始觀察值找出適當的變數，可提高數量化分析的效率和準確性。
- 關鍵字詞的研究大致可分為三個方向：

→偵測及判斷「潛在關鍵詞」

→篩選潛在關鍵詞：(「常用關鍵詞」、「核心關鍵詞」)

→關鍵詞間的關聯（基石關鍵詞？）

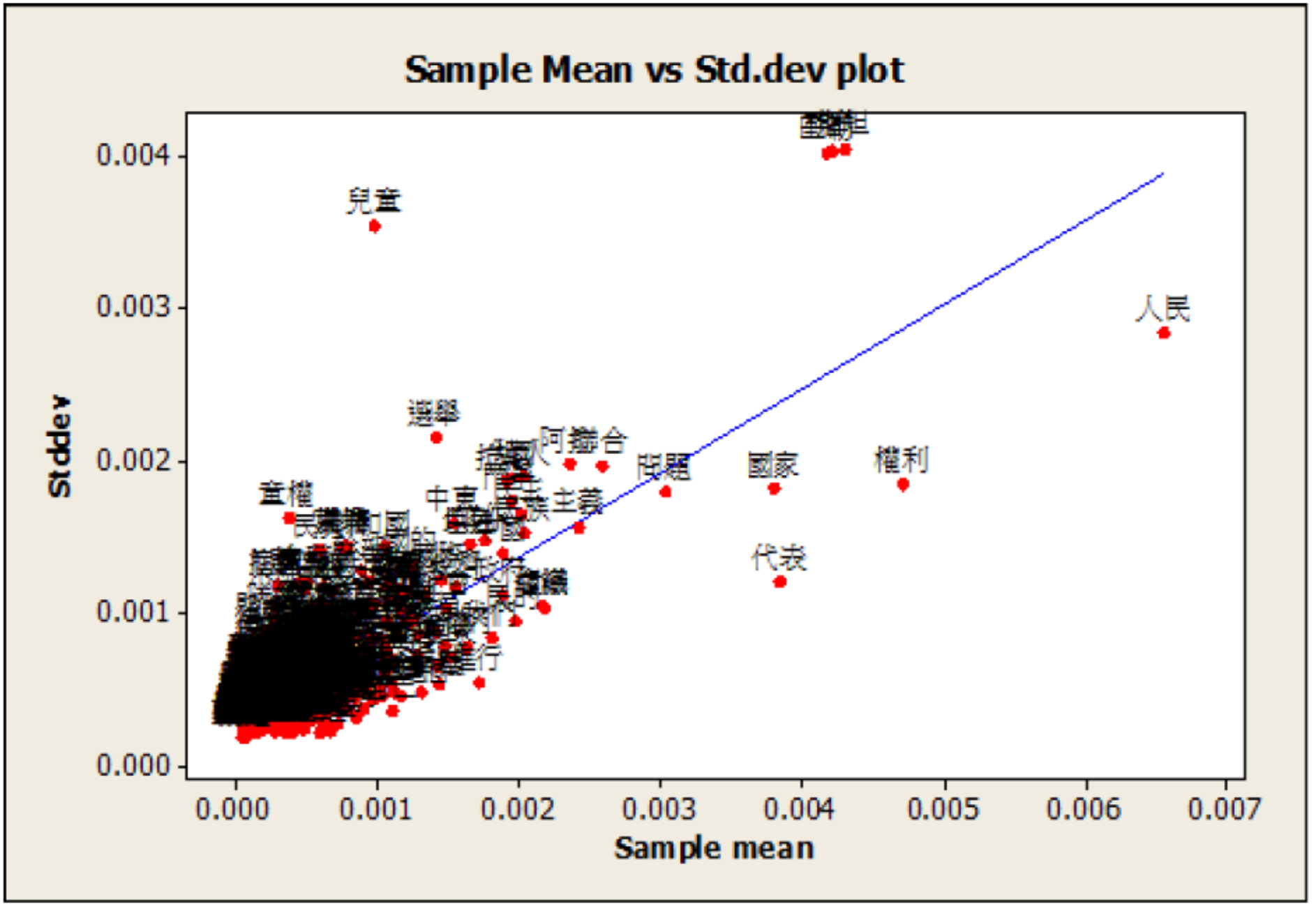
研究流程圖(Float Chart)

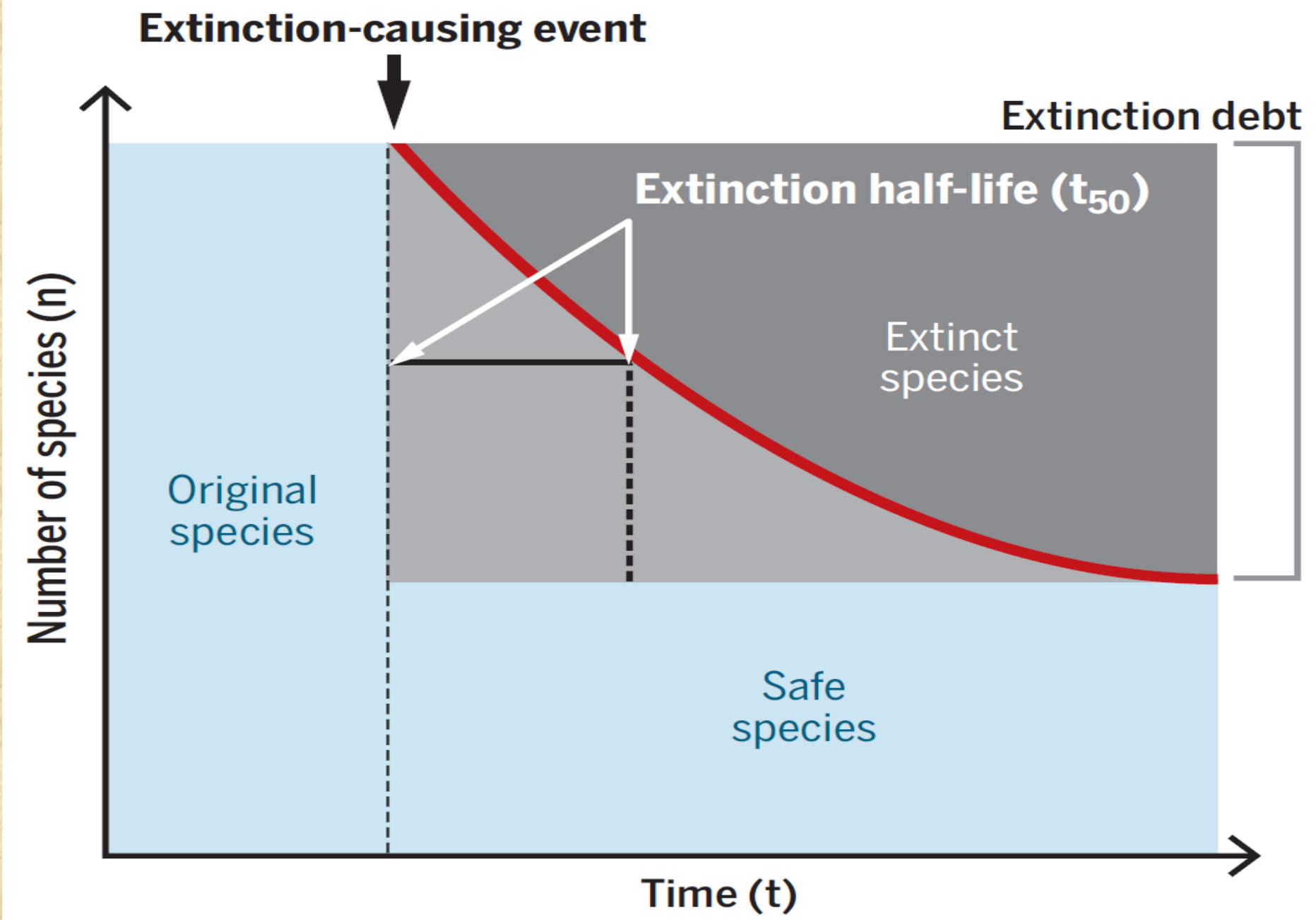
龐大字詞數量
（原文）

1. 以數位方法決定準關鍵字。
2. 人文學者協助去除不相關的關鍵字。
3. 引入生物多樣性、物種演化概念，進一步篩選出核心關鍵字詞。
4. …… (Iteration…)

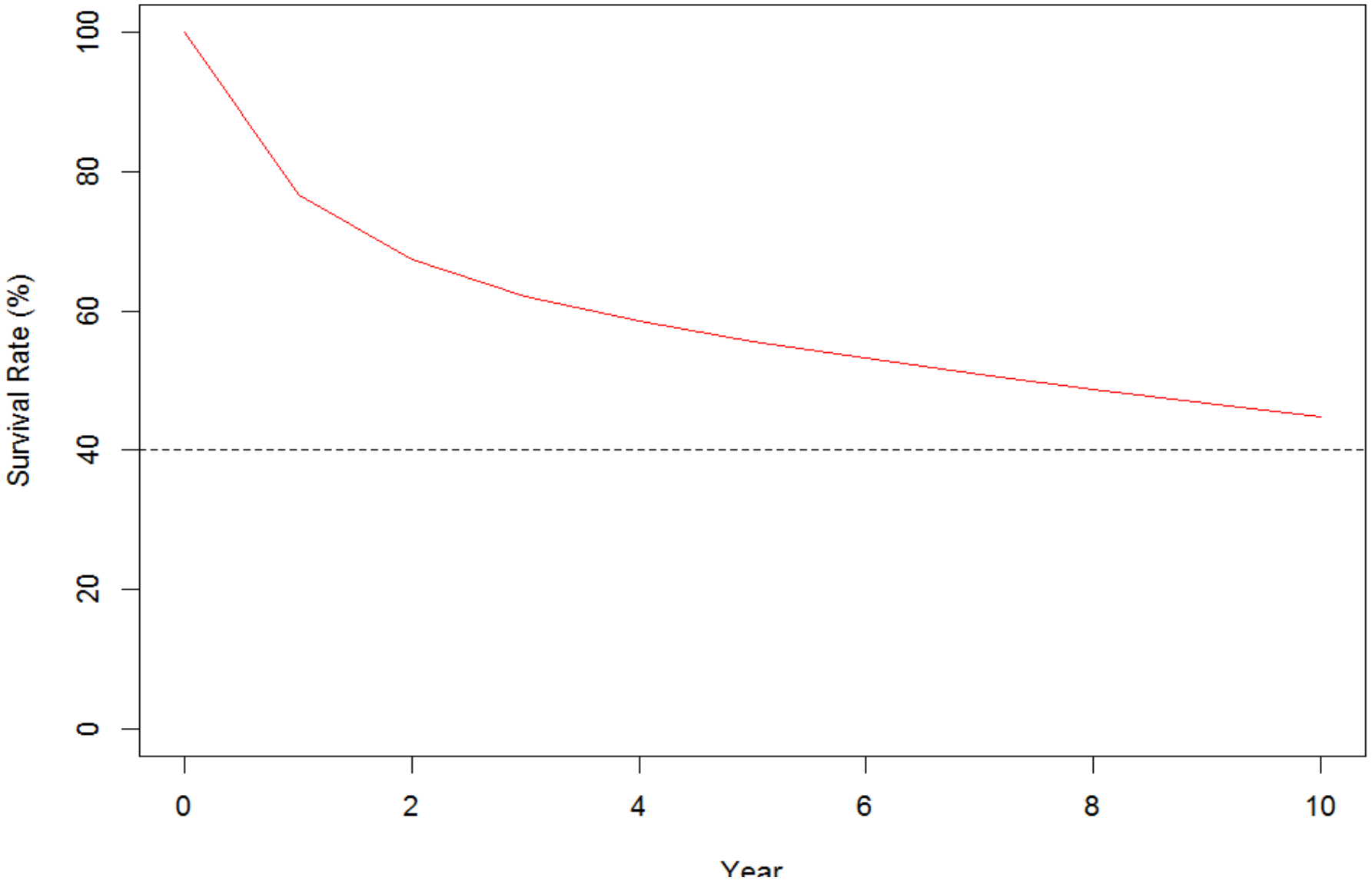
核心關鍵字詞

人民日報(1971-1989)幾個常見關鍵詞之分佈



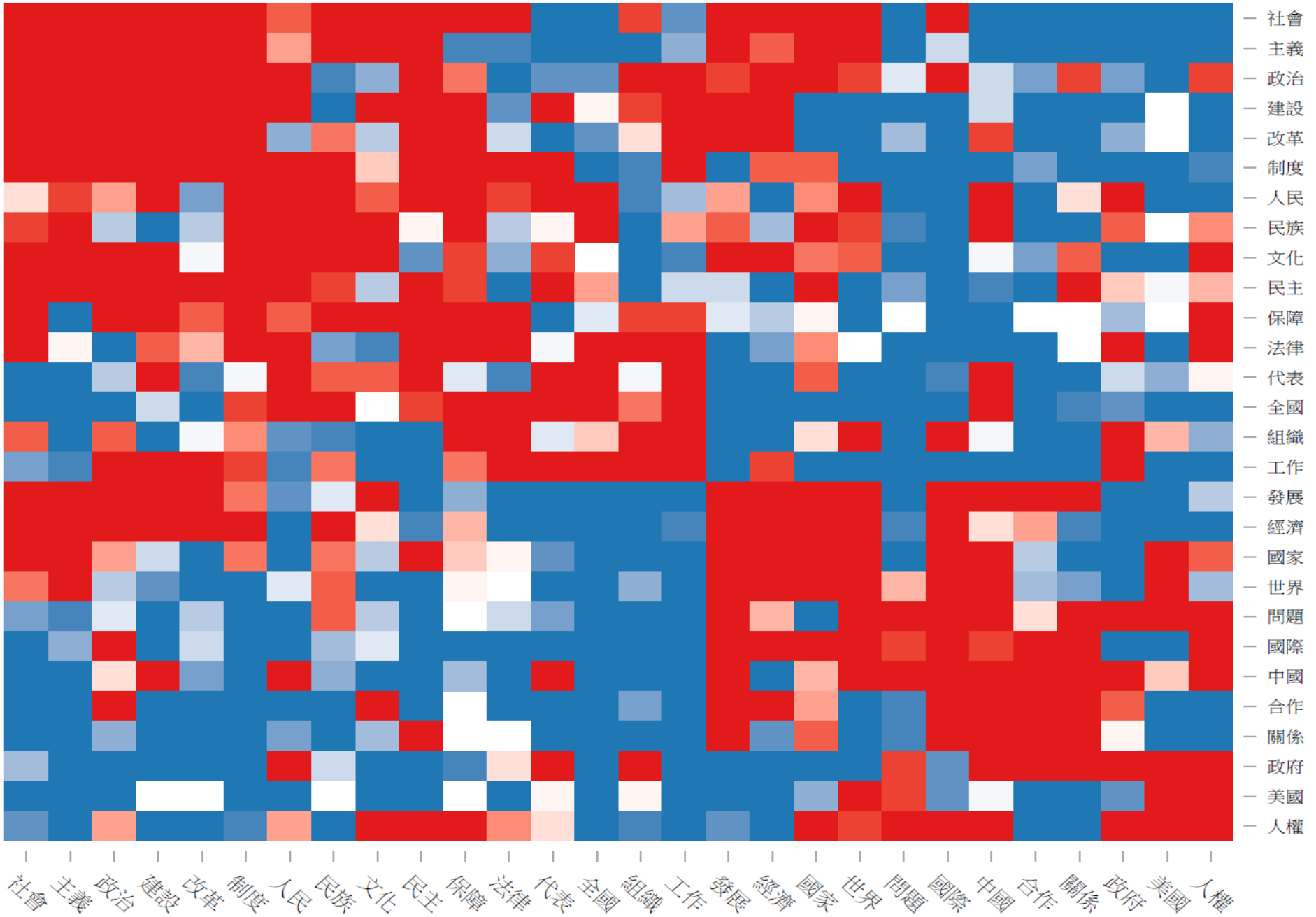


物種數的存活曲線（來源：Science, 2016)



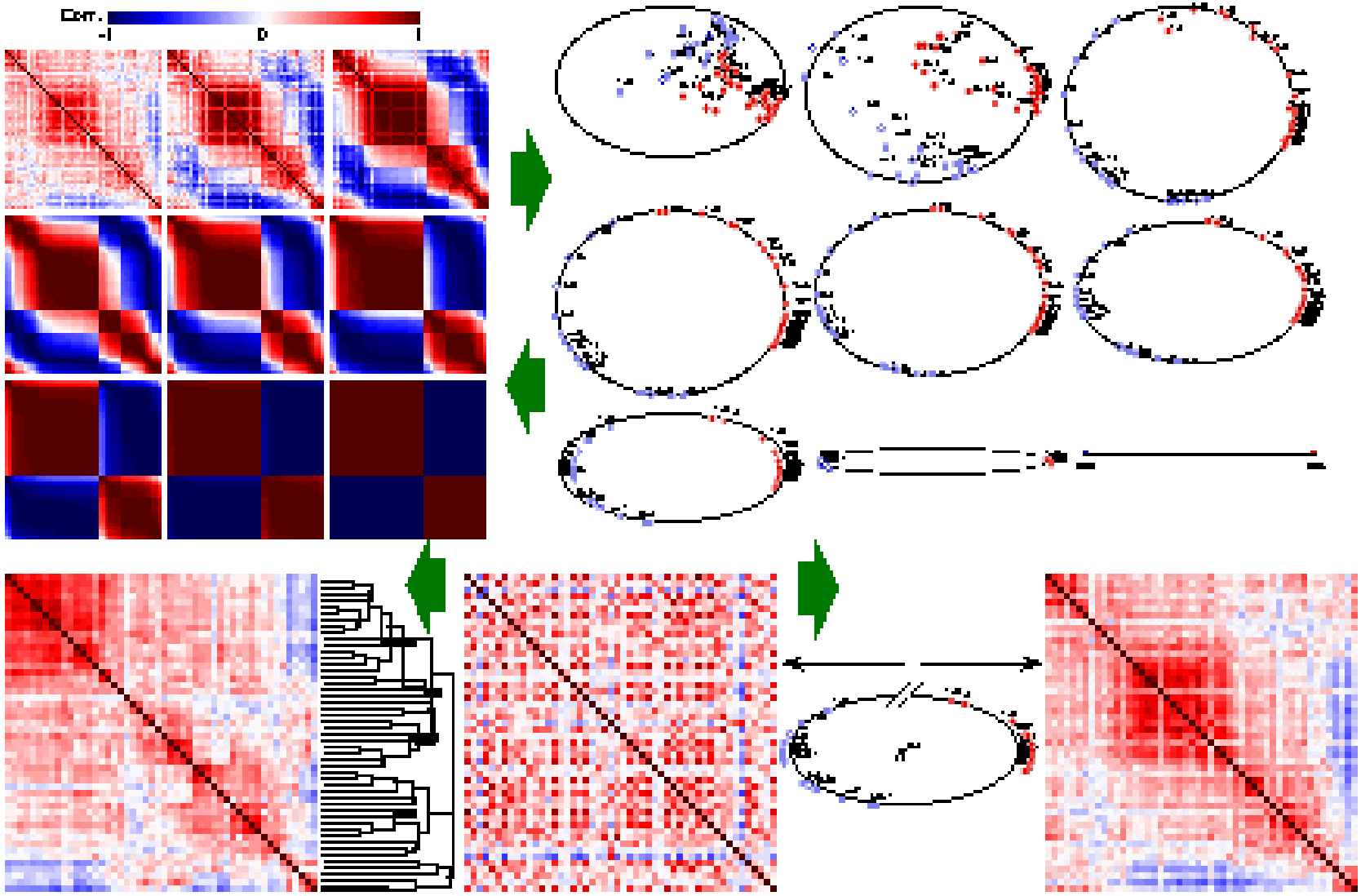
《人民日報》前百大雙字詞存活曲線(1988~2005年)

常見字彙	滅絕字彙	
	1988-1990	1991-1999
中國、國際、主義、人權、人民、問題、工作、改革、政治、文化、民主、發展、社會、經濟、美國、關係、領導	分子、同志、資產、革命	階級
	新生字彙	
	1991-1999	2000-2015
	貿易、平等、市場、友好、精神、提高、尊重	依法、實施、完善、推進、開展、法治、機制、服務、環境、執法、和諧

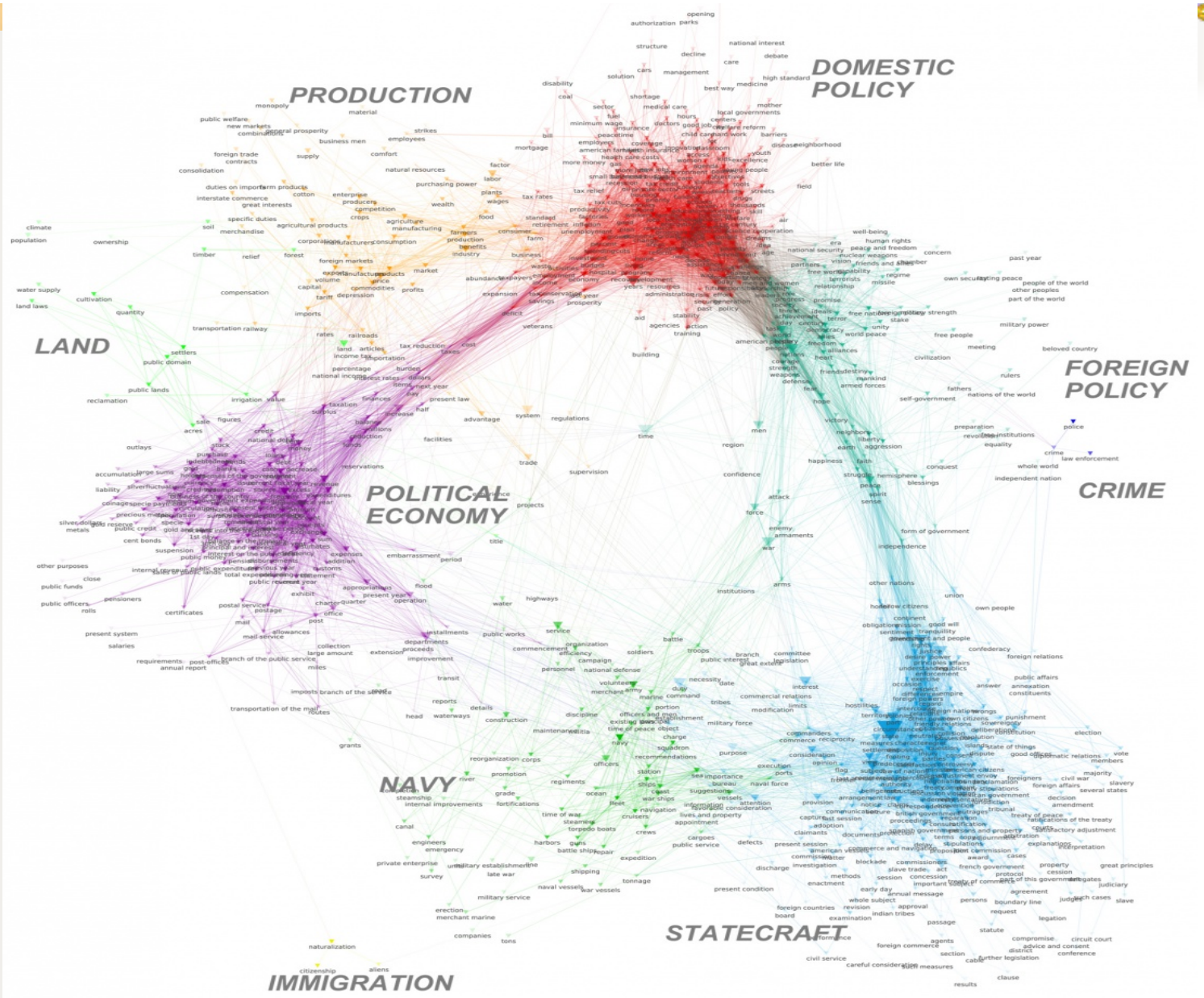


2000年《人民日報》前30大雙字詞關係圖

Basic Concepts for Generalized Association Plots

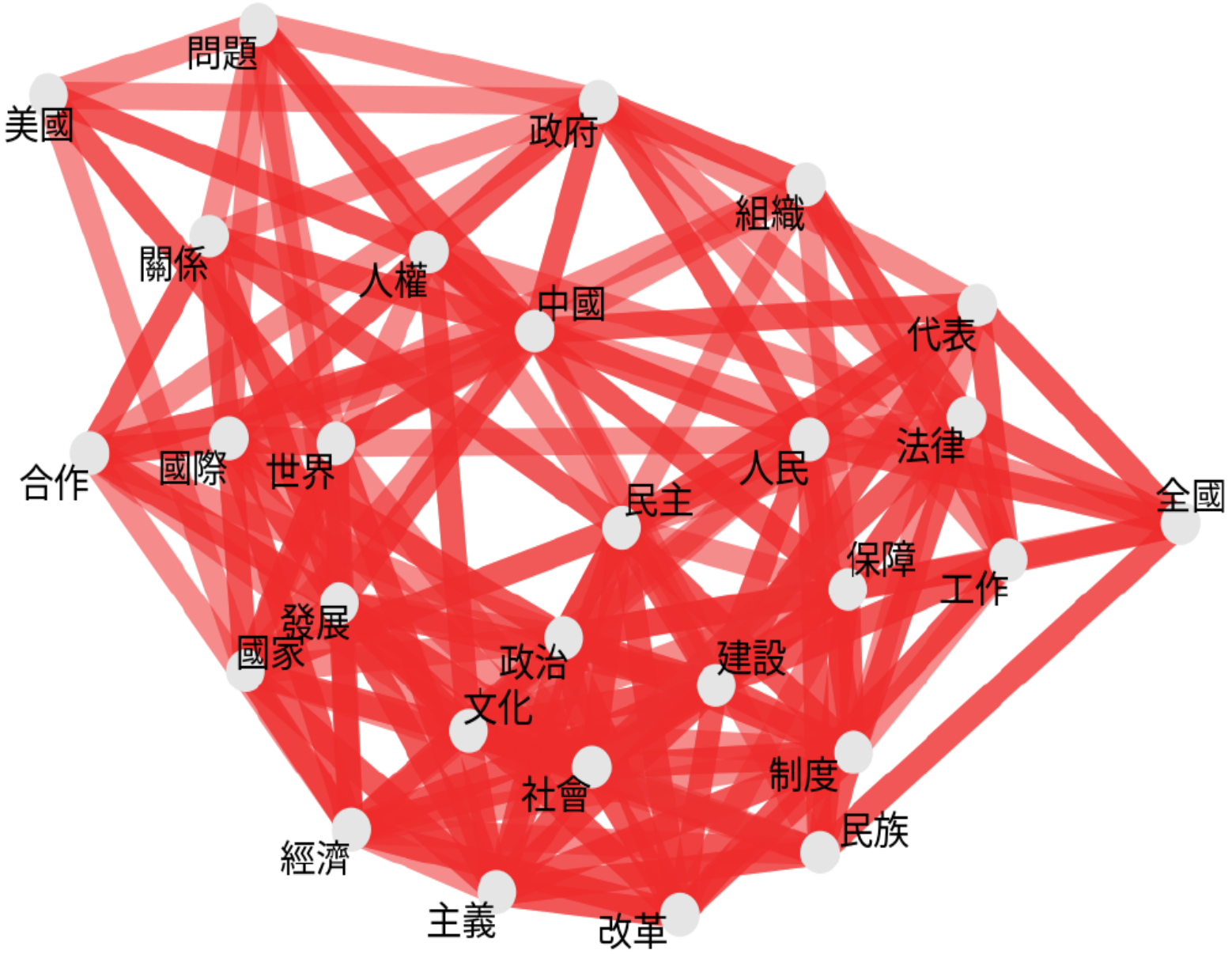


美國總統國情咨文(State of Union;1790年至今)

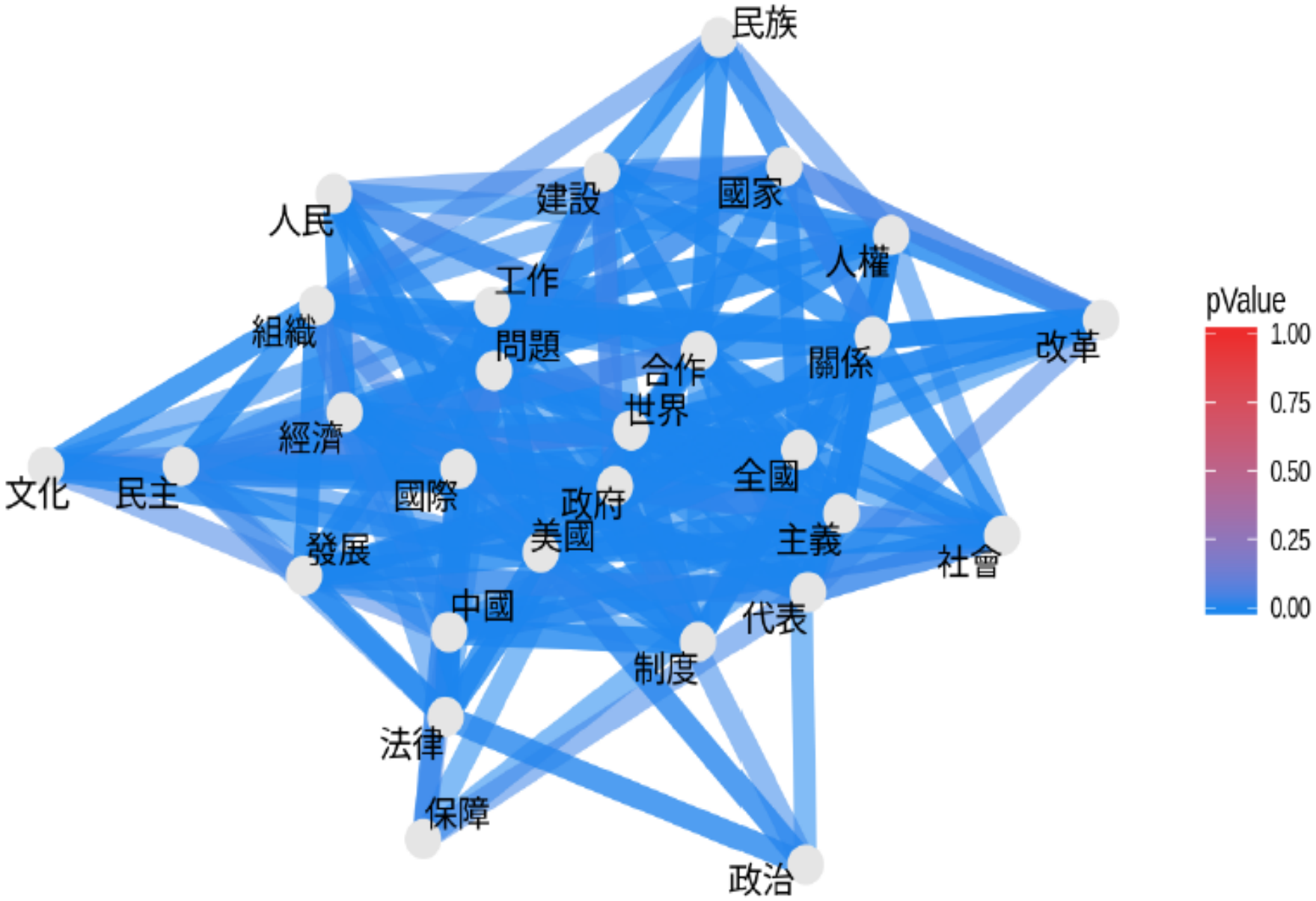


來源：https://sciencenode.org/feature/text-mining-strikes-gold-

2000年正相關詞彙


問題
美國
政府
組織
關係
人權
中國
代表
合作
國際
世界
人民
法律
全國
民主
保障
工作
發展
國家
政治
建設
文化
制度
社會
民族
經濟
主義
改革
pValue
1.00
0.75
0.50
0.25
0.00

2000年負相關詞彙


民族
建設
國家
人民
人權
工作
組織
問題
改革
合作
關係
世界
經濟
全國
文化
民主
國際
政府
美國
主義
發展
社會
中國
代表
制度
法律
保障
政治
pValue
1.00
0.75
0.50
0.25
0.00