巨量資料與統計分析

政治大學統計系余清祥

2024年12月10日

第十三週：文字資料分析案例

http://csyue.nccu.edu.tw

《紅樓夢》的寫作風格分析

https://i2.kknews.cc/SIG=37pq4el/n64000538o613p937n1.jpg

關於《紅樓夢》

- 《紅樓夢》的其他名稱：

→空空道人：《石頭記》、
《情僧錄》

→東魯孔梅溪：《風月寶鑑》

→曹雪芹：《金陵十二釵》

註：一般認為曹雪芹作1～80回，高鶚（程偉元）續作81～120回。

滿紙荒唐言，一把辛酸淚！
都云作者癡，誰解其中味？

關於紅樓夢

- 《三國演義》、《西遊記》、《水滸傳》、《紅樓夢》四部中國古典小說。

→四部著作都有很高的藝術水平，細緻的刻畫和所蘊含的思想都為歷代讀者所稱道，有「四大名著」的說法。 —— *維基百科*

- **紅學**就是研究中國古典文學名著《紅樓夢》的學問。紅樓夢由於傳世版本多，加以欣賞角度與動機的不同，因此學者們對於紅樓夢的作者與內容，有許多不同的看法。

文學與統計的相關研究

- 陳炳藻(1980)：【紅樓夢】的字彙統計，UW-Madison Thesis。
- 趙岡、陳鍾毅(1980),【紅樓夢研究新編】,聯經出版社。（胡適、俞平伯）
- Mosteller and Wallace (1984, Springer)：美國憲法The Federalist Papers的作者問題。
- Efron and Thisted (1976, Biometrika)：莎士比亞(Shakespeare)的字彙統計。

註：金庸的天龍八部（倪匡代寫，香港明報）

紅樓夢的名句

- **「假作真時真亦假，無為有處有還無」**，這是《紅樓夢》第五回賈寶玉夢遊太虛幻境時，在一個大石牌坊上看到的一副對聯。

→滿紙荒唐言，一把辛酸淚！都云作者痴，誰解其中味？

→任憑弱水三千，我只取一瓢飲

→女兒是水作的骨肉，男人是泥作的骨肉

→儂今葬花人笑痴，他年葬儂知是誰？

好了歌

作詞：曹雪芹　作曲：吳楚楚

世人都曉神仙好　唯有功名忘不了
古今將相在何方　荒塚一堆草沒了

世人都曉神仙好　唯有金銀忘不了
終朝只恨聚無多　聚到多時眼閉了

世人都曉神仙好　唯有嬌妻忘不了
君生日日說恩情　君死又隨人去了

世人都曉神仙好　唯有兒孫忘不了
痴心父母古來多　孝順子孫誰見了

非結構資料分析的注意事項

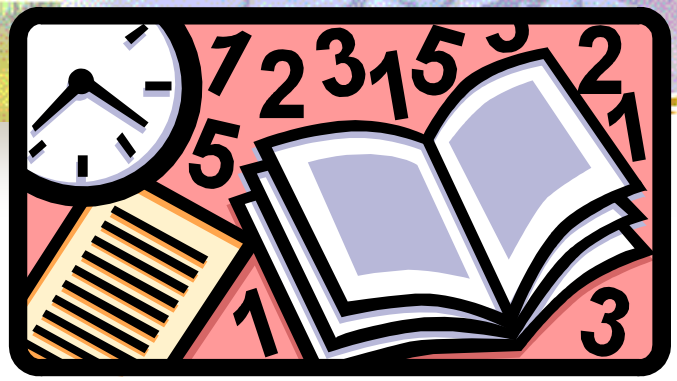
- 如何量化文字及音樂等資料？

→依賴應用領域知識，選取重要特性(Variable或Feature)以供分析。（結構資料！？）

- 資料品質（眉批？）

→由於印刷的昂貴與古代中國人缺乏著作權觀念的影響下，一部小說或因謄抄的錯誤；或因原著的散佚；或因後人的篡改增刪，可能產生與原著出入甚多的各種版本，令後來的讀者無所適從。（聖經密碼？）

紅樓夢的資料選取



- 較知名的版本

→「脂本」：甲戌本、己卯本、庚辰本、甲辰本、戚本

→經高鶚輯補：程甲本、程乙本

- 使用的資料庫：

→陳郁夫老師的資料庫（庚辰本 與 程甲本）

→中央大學的網路版（但有十二回散佚）

Efron & Thisted 的新物種估計(1987)

- 若泰勒詩作者是莎士比亞的假設為真，估計值與觀測值相近!

Observed value – Estimates

X	0	1	2	3	4	5	6	7	8	9
0-9	2.03	2.79	1.67	1.16	1.47	-0.43	1.84	-2.01	0.13	1.24
10-19	-0.62	-1.50	1.48	-1.36	-0.38	-0.38	-0.33	0.72	-0.25	-1.22
20-29	0.82	0.84	-0.13	3.91	1.94	-0.06	-1.04	0.98	1.00	2.02
30-39	3.04	0.06	0.07	0.10	1.12	0.12	-0.86	-0.85	2.17	2.18
40-49	0.20	1.21	-0.77	-0.75	1.26	0.26	0.27	1.28	0.30	0.31
50-59	-0.68	0.33	0.34	0.36	0.37	-0.63	-0.62	0.39	-0.60	1.41
60-69	-0.58	0.43	-0.56	-0.54	0.47	0.47	-0.52	-0.51	0.50	-0.50
70-79	-0.49	-0.48	0.52	-0.47	-0.46	0.54	-0.45	-0.45	0.56	0.56
80-89	-0.43	-0.42	0.58	0.59	-0.40	-0.40	-0.39	-0.39	-0.38	-0.38
90-99	-0.37	-0.36	-0.36	0.65	-0.34	0.66	0.66	-0.33	-0.32	-0.32

研究方法（1998年）

- 兩個樣本

→將前80回及後40回視為兩個樣本，可使用一般的統計方法(如t-test, Time Series Analysis)

註：考慮全書而非抽樣調查！

- 變動點問題(Change-point Problem)

→視全書120回為同一系列的產品，在於檢測是否有變動點產生，其發生的位置是否在第80回前後。

實證分析（1998年）

- 問題:如何量化資訊以供比較?

1.結構研究:

→每回的總字數、詩詞與對話比例

2.用字分析

→虛字「兒」、「在」、「了」、「的」、「著」

→趙岡與陳鍾毅建議的其他字詞：「嗎」和「麼」、「給」和「與」、「都」和「多」、「我們」和「咱們」

→每回結尾用語（是否使用「下回分解」）

五個常見的虛字

「兒」、「在」、「了」、「的」、「著」

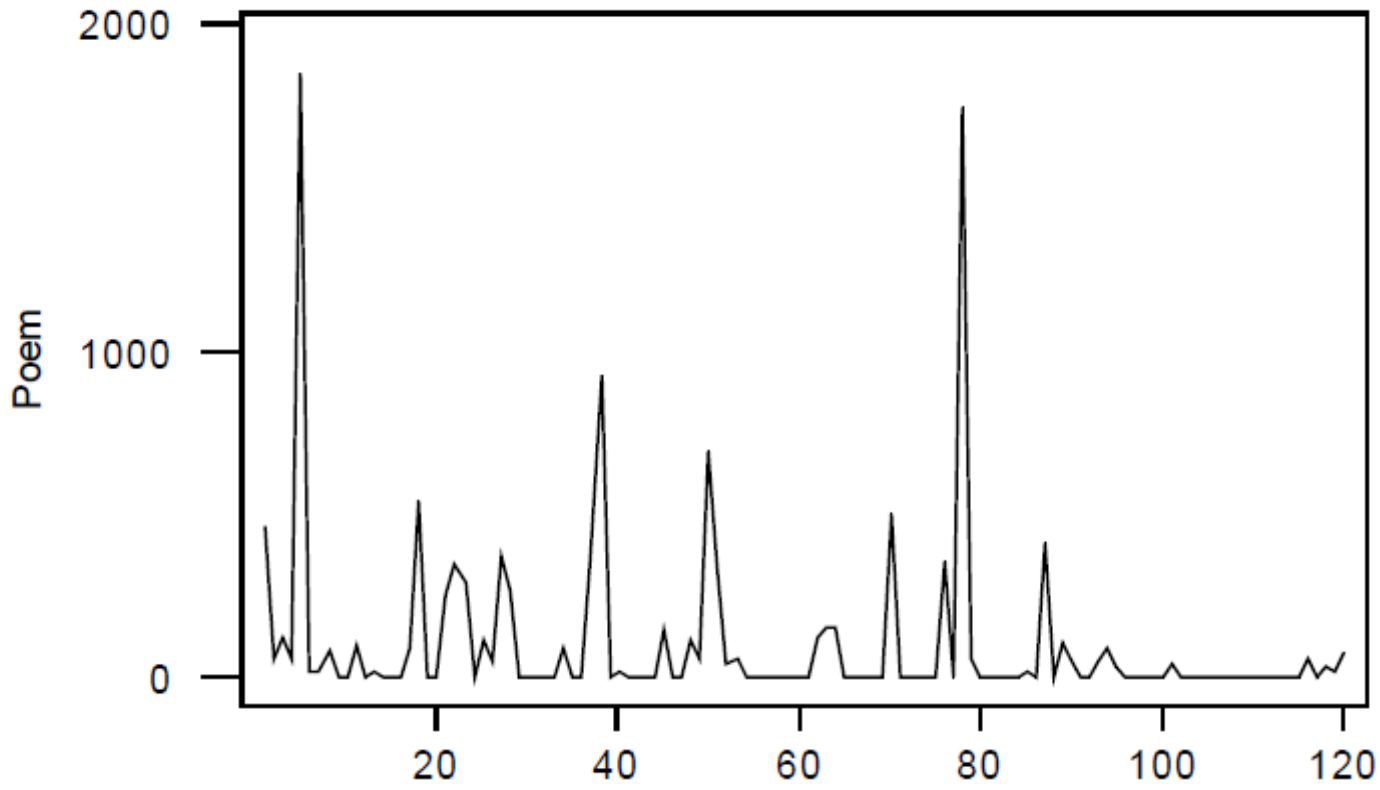
→「坐一坐兒」和「坐一坐」

→「便伏在枕上歇一會」和「便伏枕上歇一會」

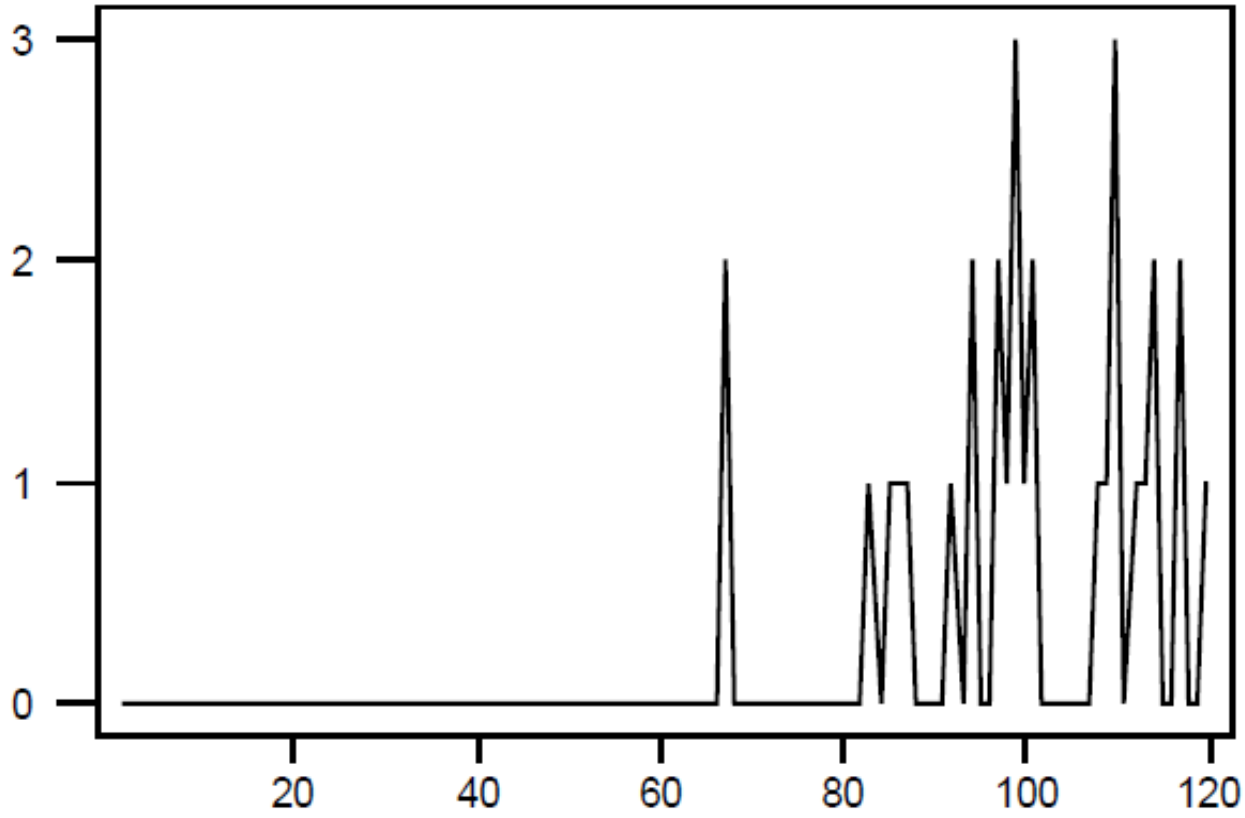
→「寶玉已醒了」和「寶玉已醒」

→「各房的丫頭」和「各房丫頭」

→「笑著說」和「笑說」

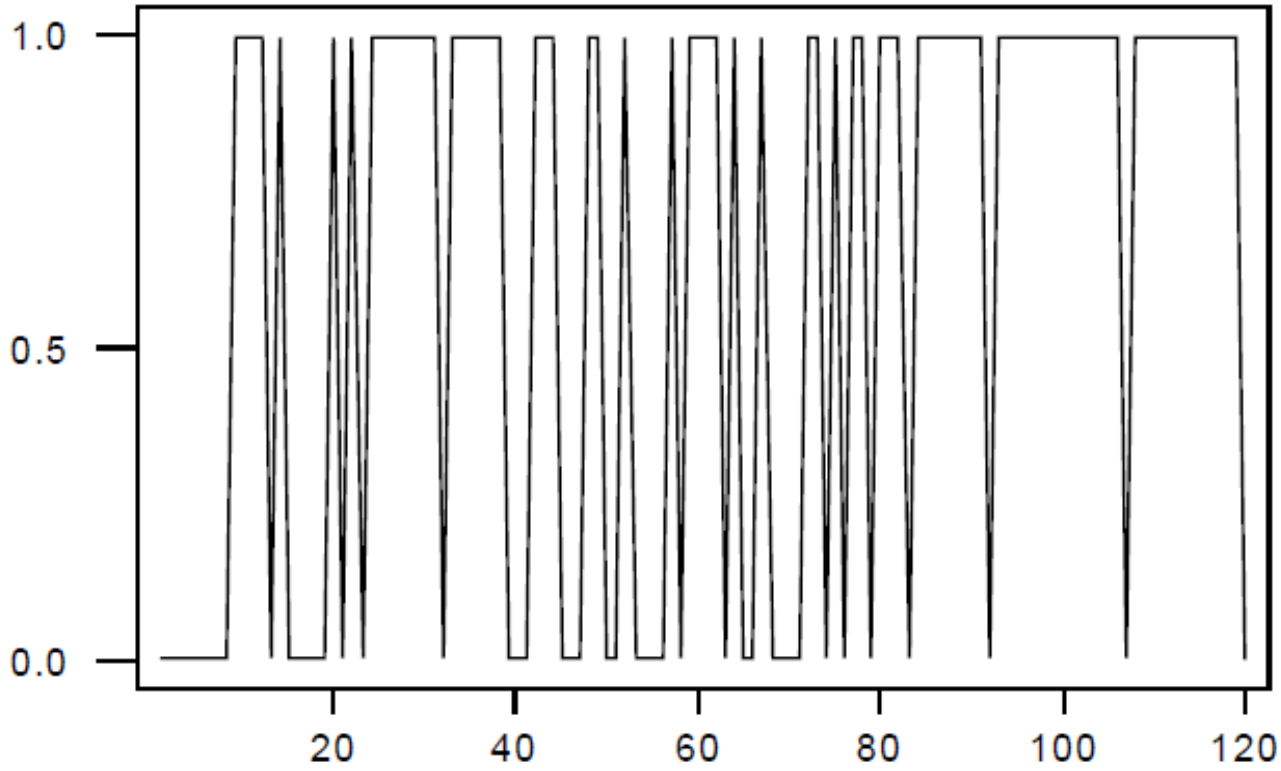


每回詩詞字數序列圖



每回問句以「嗎」結尾的字數

回末用詞	前80回	後40回
下回分解	3	29
要知端的，且聽下回分解	23	0
且聽下回分解	14	7
無任何結語	15	0
詩	6	1
要知端的	7	0
欲知後事且聽下回	1	3
其他（共八種）	11	0
總數	80	40



每回回末是否以「下回分解」結尾

註：前後半部各是40/80 vs. 36/40

- 使用上頁的17個變數，將《紅樓夢》區分為前八十回、後四十回，分類結果還算不錯。（準確率111/120≈92.5%）

	True Group	
Put into Group	1	2
1	75	4
2	5	36
total	80	40

	一般統計檢定	變動點分析 (變動點是否在第80回附近)
每回總字數	不顯著	——
每回詩詞字數	顯著	是
每回對話字數	不顯著	——
兒	顯著	不是
在	顯著	是
了	顯著	不是
的	顯著	不是
著	顯著	是
嗎	顯著	是
麼	顯著	是
給	不顯著	——
與	顯著	是
都	顯著	不是
多	顯著	不是
我們	不顯著	——
咱們	不顯著	——
每回結尾用語	顯著	是

註：一般統計檢定不顯著者，不再考慮變動點分析。

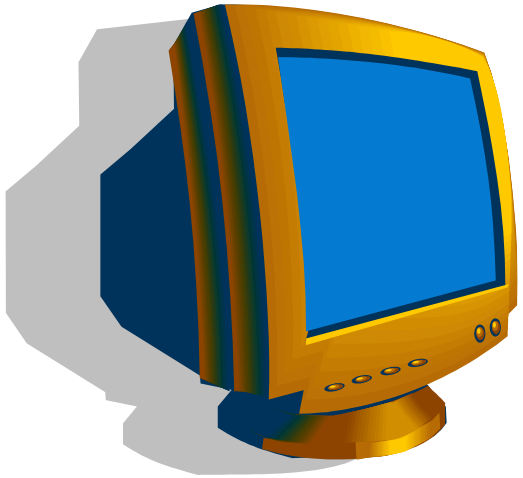
研究方法（2008年）

- 以物種的角度思考，估計全書一共有多少不同字彙（類似 Efron and Thisted, 1976）

→區塊抽樣(quadrat sampling)估計母體種類數

→使用方法包括：

- Bootstrap Method
- Jackknife Estimate
- Chao Estimate



紅樓夢代入E-T法–新字估計(1976)

- 以十回為一組作K-fold cross–validation檢視兩樣本之內部一致性。每次扣去一組作為 training data，被扣去的一組作為testing data。

前八十回視為同一母體，每十回為一組之8-fold CV

Testing Data	t	Est.	Variance of Est.	UC.I.	LC.I.	新字	是否在95%C.I.
1~10	1/7	98.42	65.40	226.60	0	144	Y
11~20	1/7	101.50	65.42	229.73	0	141	Y
21~30	1/7	98.75	65.83	227.78	0	87	Y
31~40	1/7	98.58	65.97	227.88	0	69	Y
41~50	1/7	96.14	65.71	224.94	0	103	Y
51~60	1/7	100.99	65.83	230.03	0	87	Y
61~70	1/7	101.27	65.89	230.41	0	80	Y
71~80	1/7	90.46	64.63	217.13	0	244	N

紅樓夢代入E-T法

後四十回視為同一母體，每十回為一組之4-fold CV

Testing Data	t	Est.	Variance of Est.	UC.I.	LC.I.	Real data	是否在95%C.I.
81~90	1/3	**172.47**	55.00	280.27	64.665	**261**	**Y**
91~100	1/3	**170.36**	55.38	278.91	61.817	**219**	**Y**
101~110	1/3	**175.28**	55.33	283.72	66.838	**225**	**Y**
111~120	1/3	**182.97**	55.68	292.10	73.843	**186**	**Y**

紅樓夢前八十回預測後四十回

Testing Data	t	Est.	Variance of Est.	UC.I.	LC.I.	Real data	是否在95%C.I.
81~120	1/2	294.87	66.49	425.19	164.547	**231**	**Y**

Jackknife在《紅樓夢》與金庸小說的覆蓋機率

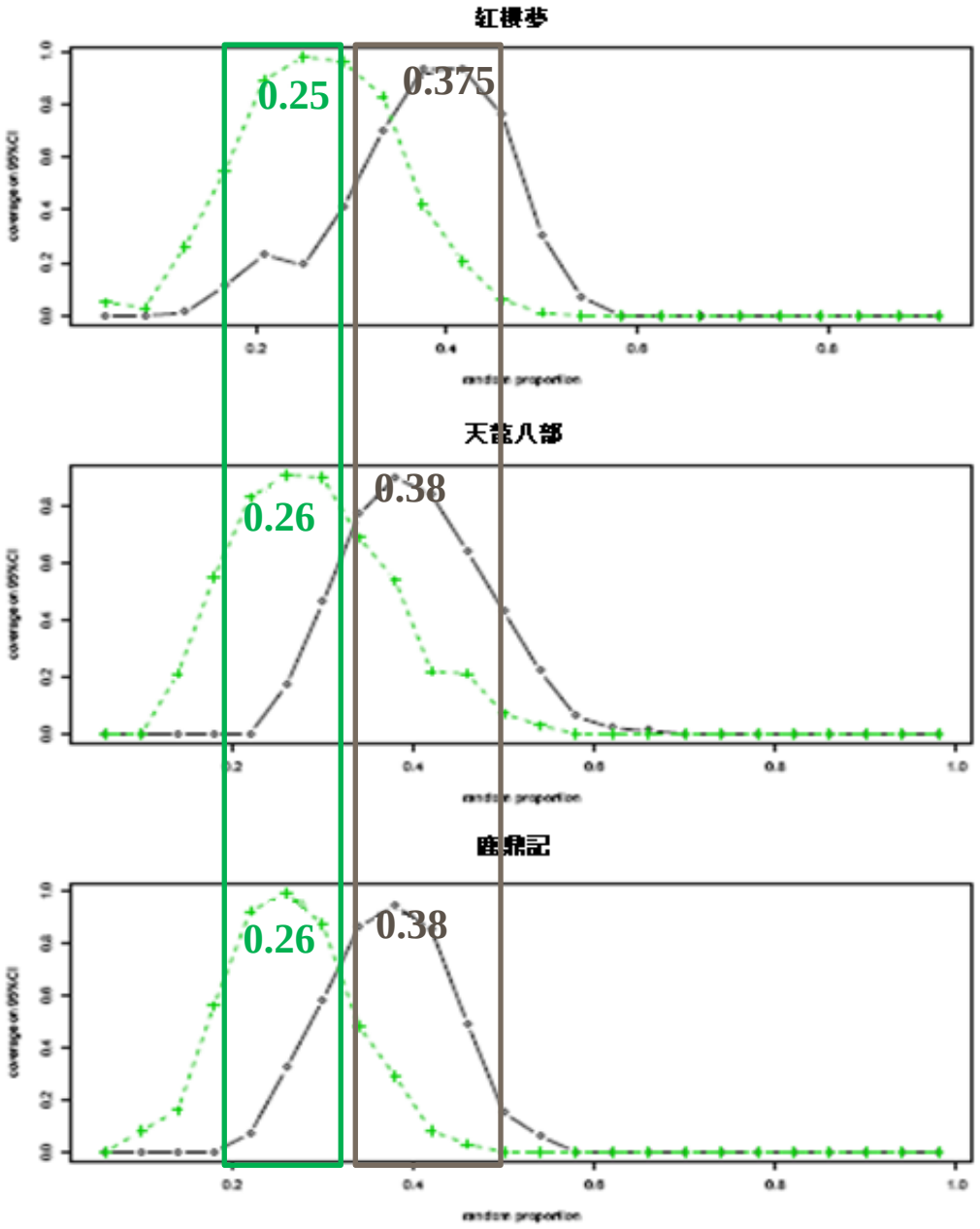
＊Magic Number!

只要抽出一部小說37.5%(25%)即可知道整部小說使用多少種不同的字彙！

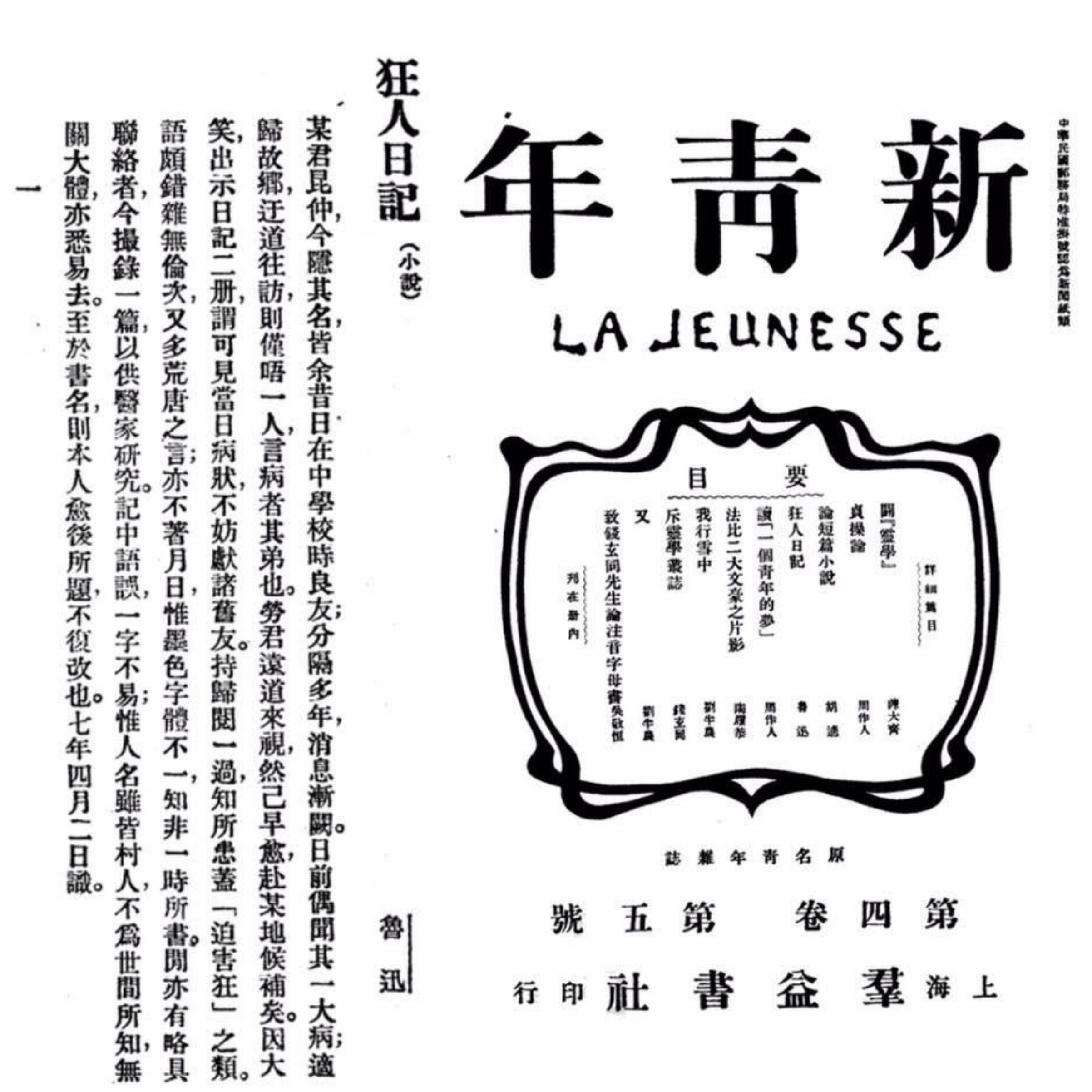
○ Jack1
＋ Jack2

X軸: random proportion

Y軸: coverage probability



《新青年》雜誌的分析



狂人日記（小說） 魯迅

某君昆仲，今隱其名，皆余昔日在中學校時良友；分隔多年，消息漸闕。日前偶聞其一大病；適歸故鄉，迂道往訪，則僅晤一人，言病者其弟也。勞君遠道來視，然已早愈，赴某地候補矣。因大笑，出示日記二冊，謂可見當日病狀，不妨獻諸舊友。持歸閱一過，知所患蓋「迫害狂」之類。語頗錯雜無倫次，又多荒唐之言；亦不著月日，惟墨色字體不一，知非一時所書。間亦有略具聯絡者，今撮錄一篇，以供醫家研究。記中語誤，一字不易；惟人名雖皆村人，不爲世間所知，無關大體，亦悉易去。至於書名，則本人愈後所題，不復改也。七年四月二日識。

一

中華民國郵務局特准掛號認爲新聞紙類

新青年

LA JEUNESSE

要目

原名青年雜誌

第四卷 第五號

上海羣益書社印行

研究目標：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。

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：

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

羅吉斯迴歸採用變數

變數型態	變數
字彙豐富度	總字數、不同字數
	每千字出現新字數
	累計字彙數
	Simpson Index
	Entropy
句子長短	平均每句字數
	每句字數的變異數
虛字	文言、白話各10個虛字
共用字、雙字詞	10 個常見字及10個常見雙字詞

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

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% ！

- 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%

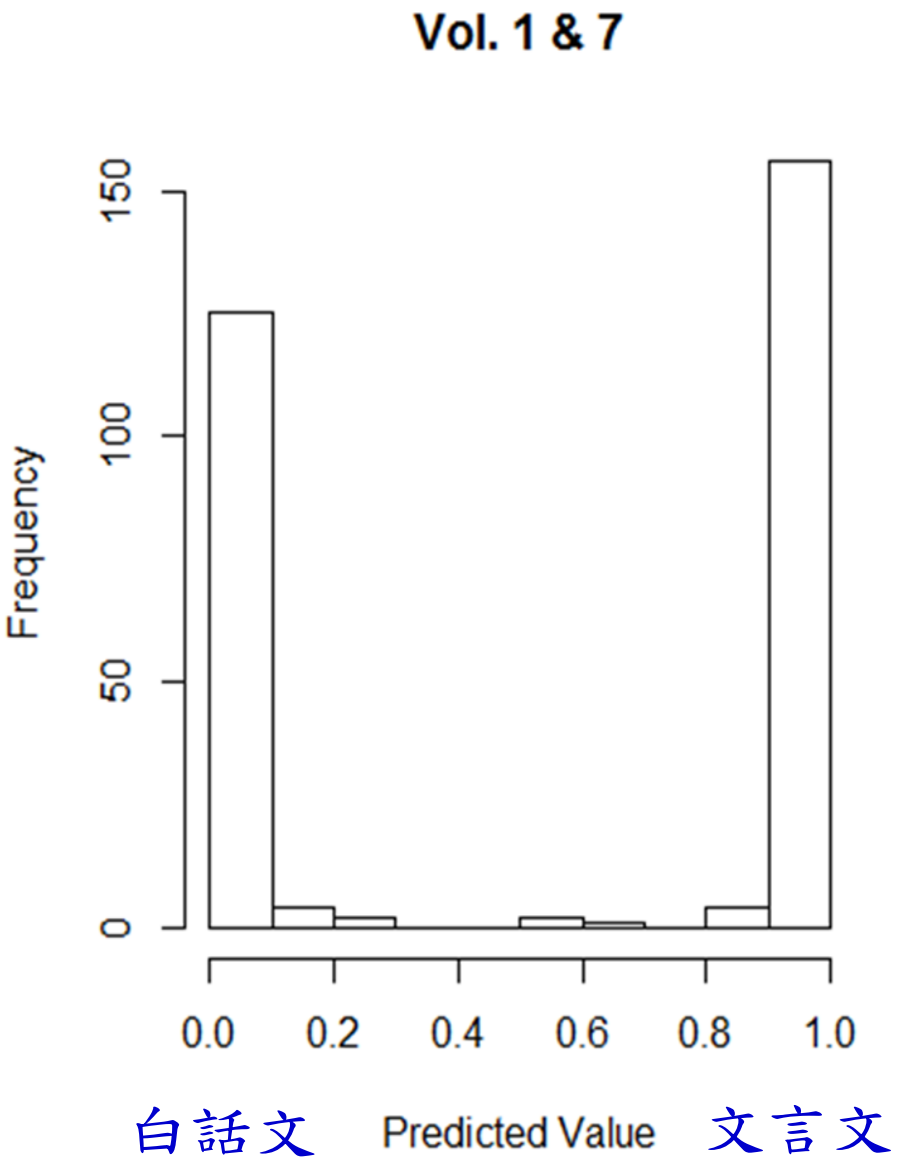
註：模型可視為穩健、可靠。

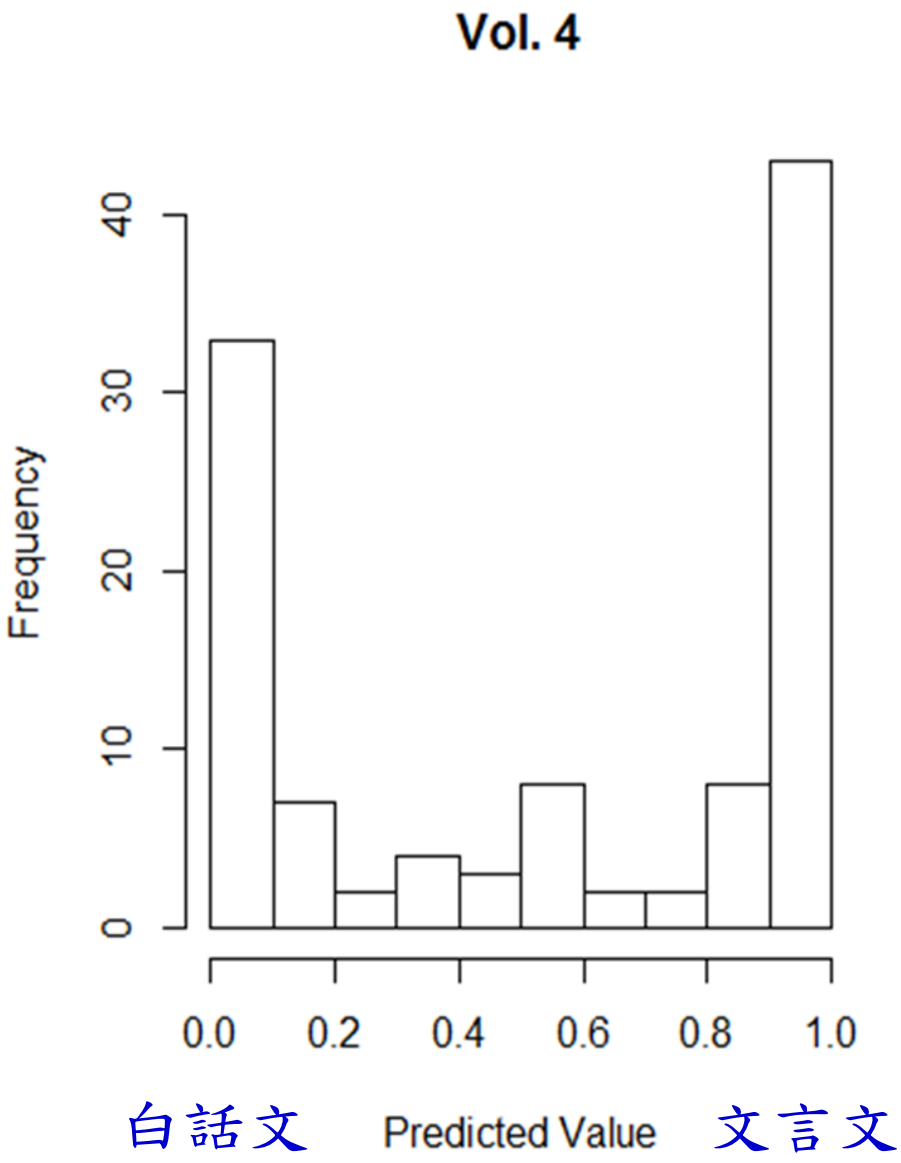
羅吉斯迴歸的預測結果（表列）

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

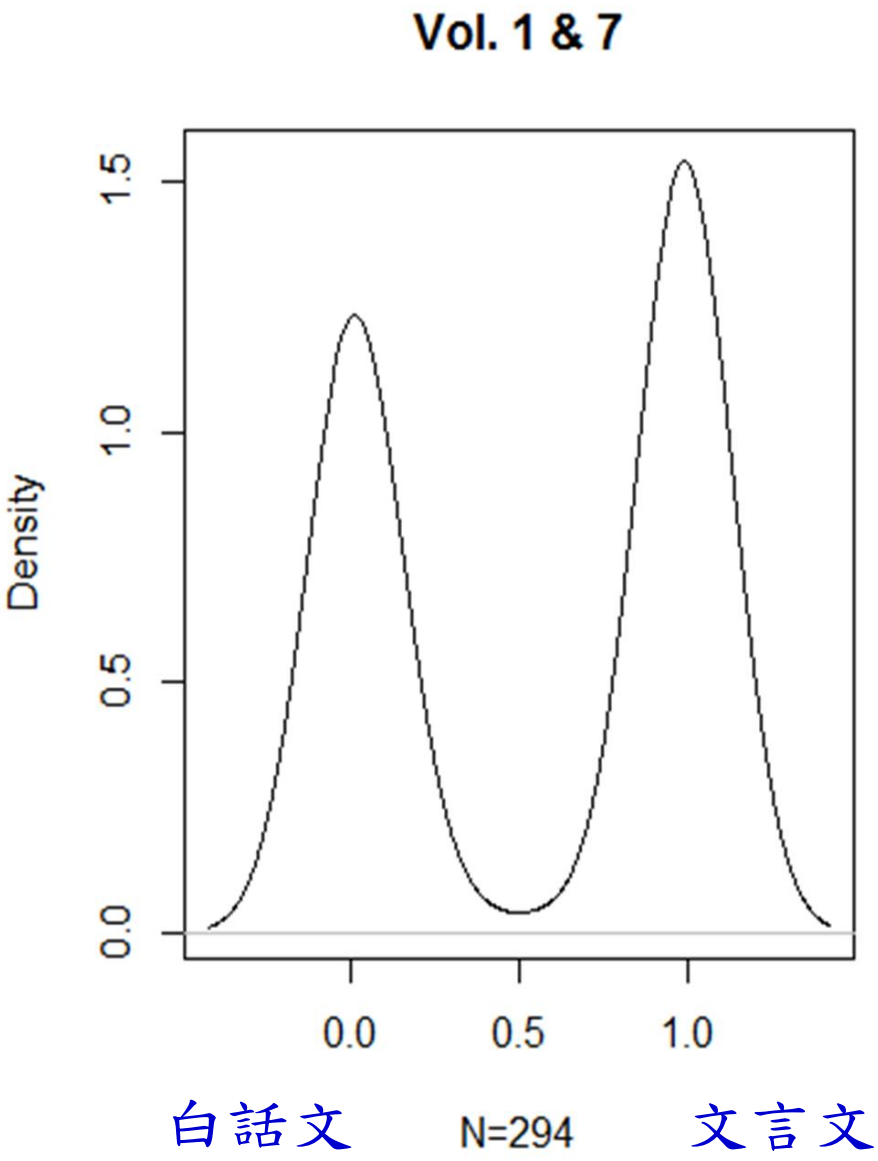
註：準確率83.54% ！

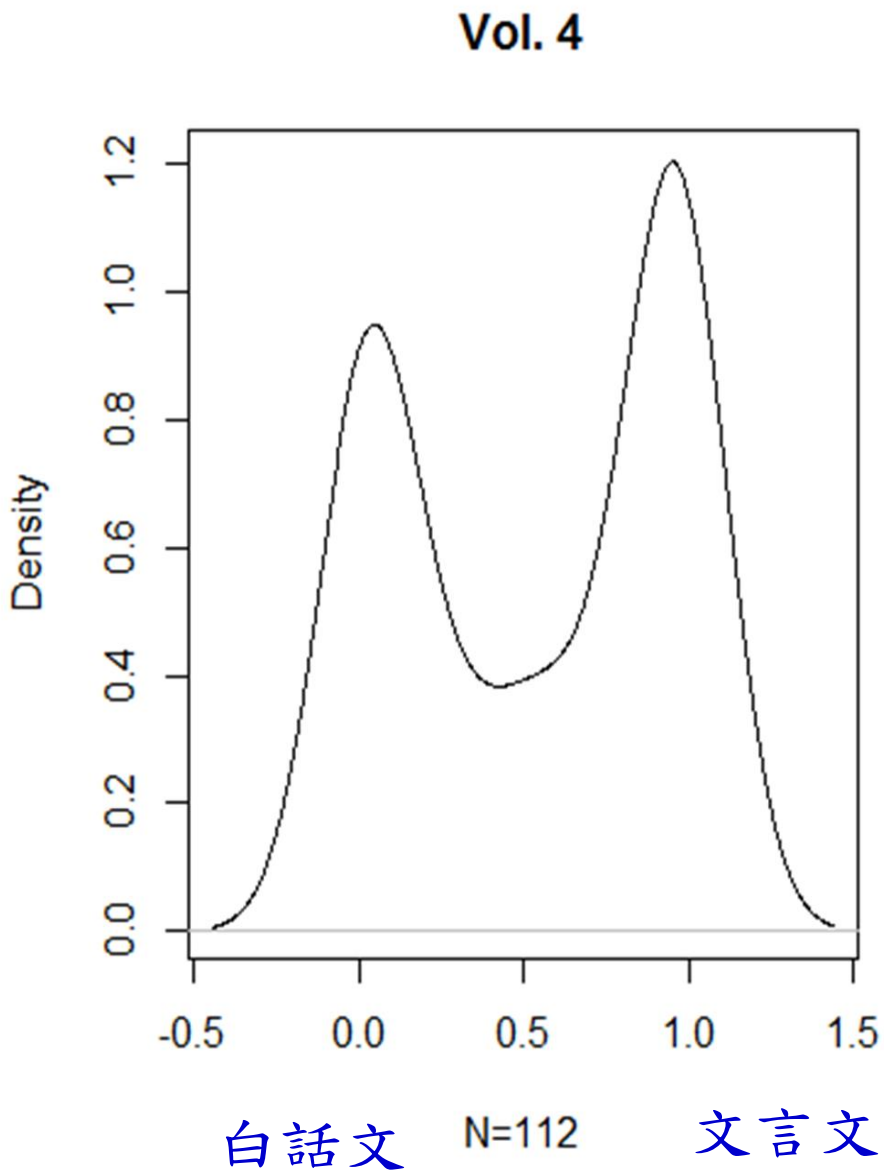
羅吉斯迴歸的分類結果


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





白話文分析範例

白話文EDA範例

- 在此也以《人民日報》為例，示範白話文的EDA。

→透過網路爬蟲等程式，下載《人民日報》1946～2003年所有文章報導。

- EDA整理項目包括：

→字詞多樣性及豐富度；

→多樣性指標、相似指標；

→生態多樣性（新生及滅絕物種）

《人民日報》歷年字彙豐富度

	總字數	不同字數	前500字 字數比例	前500雙字詞 字數比例
1950年	3,149,369	4442	82.8%	54.4%
1960年	2,541,397	3953	83.4%	59.6%
1970年	1,629,993	3674	**84.8%**	**63.2%**
1980年	2,133,879	4267	82.1%	54.0%
1990年	2,474,419	4366	82.1%	54.6%
2000年	3,027,230	4403	82.3%	55.4%
2010年	535,066	3016	84.4%	58.5%

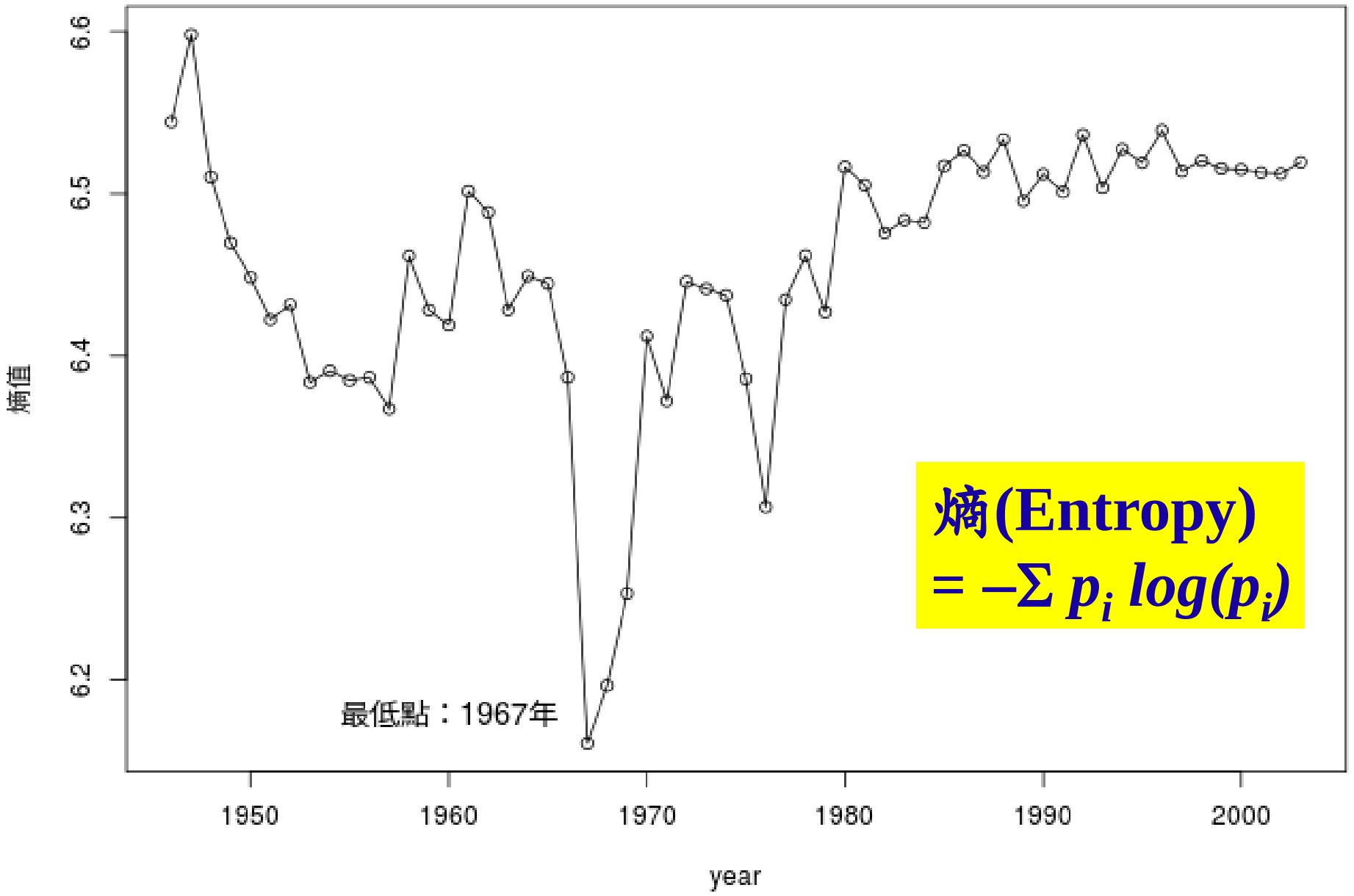
《人民日報》十大最常見字彙

Rank	1950	1960	1970	1980	1990	2000	2010
1	的	的	的	的	的	的	的
2	國	國	大	一	國	國	國
3	人	和	一	國	一	中	人
4	民	人	主	和	中	和	和
5	一	一	了	了	和	一	中
6	中	了	人	人	人	人	會
7	在	中	和	在	在	了	法
8	會	在	革	有	了	在	在
9	和	民	國	是	會	大	一
10	工	大	在	中	是	會	發

《人民日報》十大最常見雙字詞

Rank	1950	1960	1970	1980	1990	2000	2010
1	人民	人民	革命	我們	中國	發展	發展
2	我們	生產	人民	工作	我們	中國	中國
3	中國	中國	群眾	生產	發展	工作	社會
4	工作	我們	他們	發展	工作	建設	國家
5	代表	進行	我們	問題	建設	我們	問題
6	美國	發展	學習	同志	國家	經濟	美國
7	朝鮮	工作	思想	人民	人民	問題	國際
8	生產	總理	生產	他們	經濟	一個	工作
9	會議	鬥爭	鬥爭	國家	問題	企業	建設
10	進行	國家	自己	經濟	我國	國家	人權

《人民日報》單字熵值


熵值
6.6
6.5
6.4
6.3
6.2
1950
1960
1970
1980
1990
2000
year
最低點：1967年
熵(Entropy)
$= -\Sigma\, p_i \log(p_i)$

文言文、白話文的標點符號

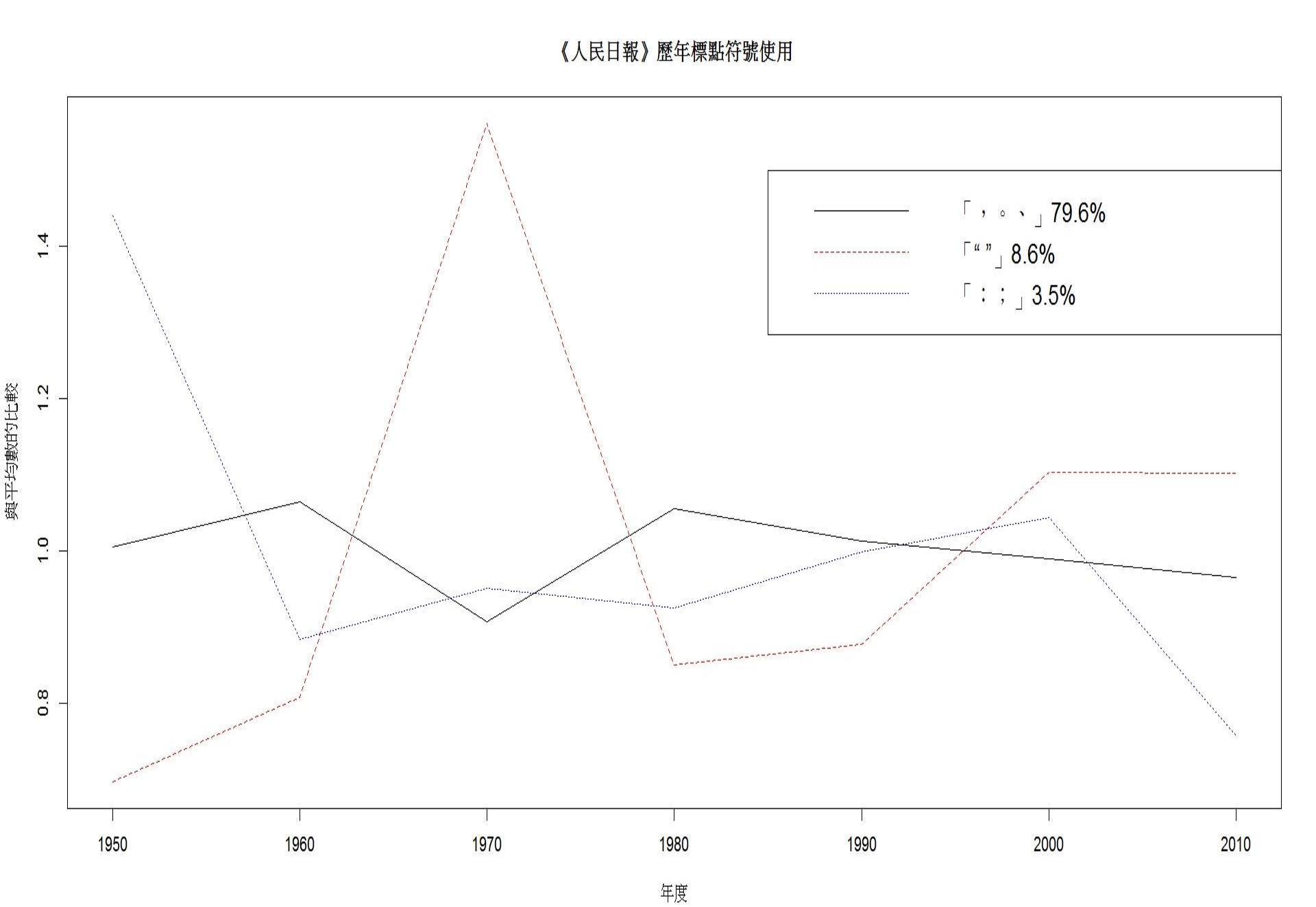
- 中文標點符號可追溯智商周時代，文言文通常不加標點符號，因而常有不同解釋。

→1920年教育部頒佈法令後逐漸一致。

- 標點符號可結合其他面向的分析，像是「斷句」的標點符號可反映句子長度。

→《人民日報》以「，。、」比例最高；

→引號「“”」及「『』」或可反映引述及對話等。



《人民日報》各年度標點符號使用比例趨勢

相似指標與其趨勢變化

- 兩兩樣本的相似指標(Similarity Index)可用於描述多樣性的變化，常用者包括：

Jaccard Index　　Yue Index

$$\theta_J = \frac{s_{12}}{s_1 + s_2 - s_{12}} = \frac{\text{相同物種數}}{\text{所有物種數}}; \quad \theta_Y = \frac{\sum_i p_i q_i}{\sum_i (p_i - q_i)^2 + \sum_i p_i q_i} = \frac{\text{相同性}}{\text{相異性+相同性}}$$

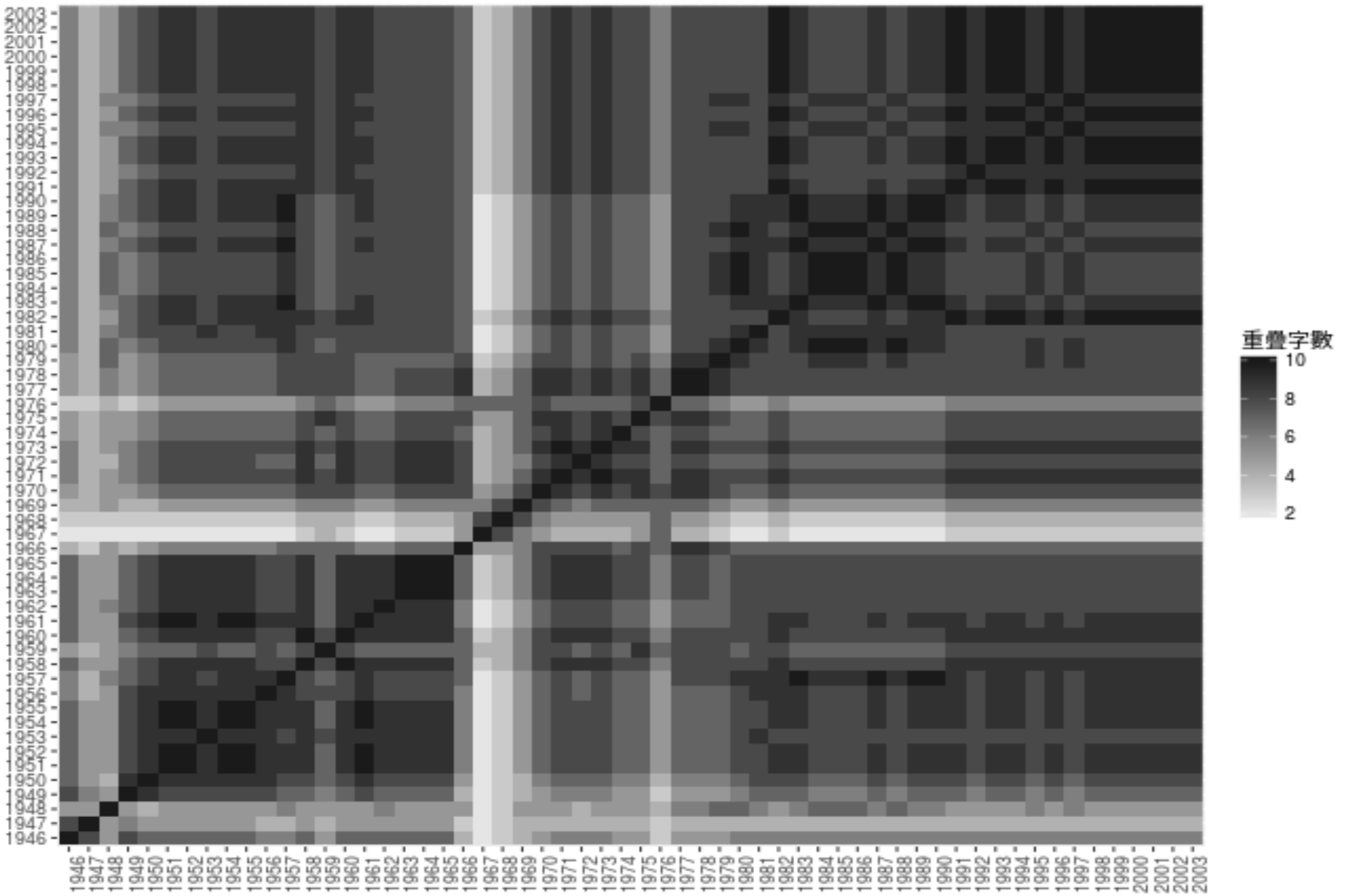
- 趨勢分析也可透過**重複程度**，像是常見字詞的出現個數，與滅絕及新生的詞彙。

→半衰期（或半生期）與新陳代謝速度。

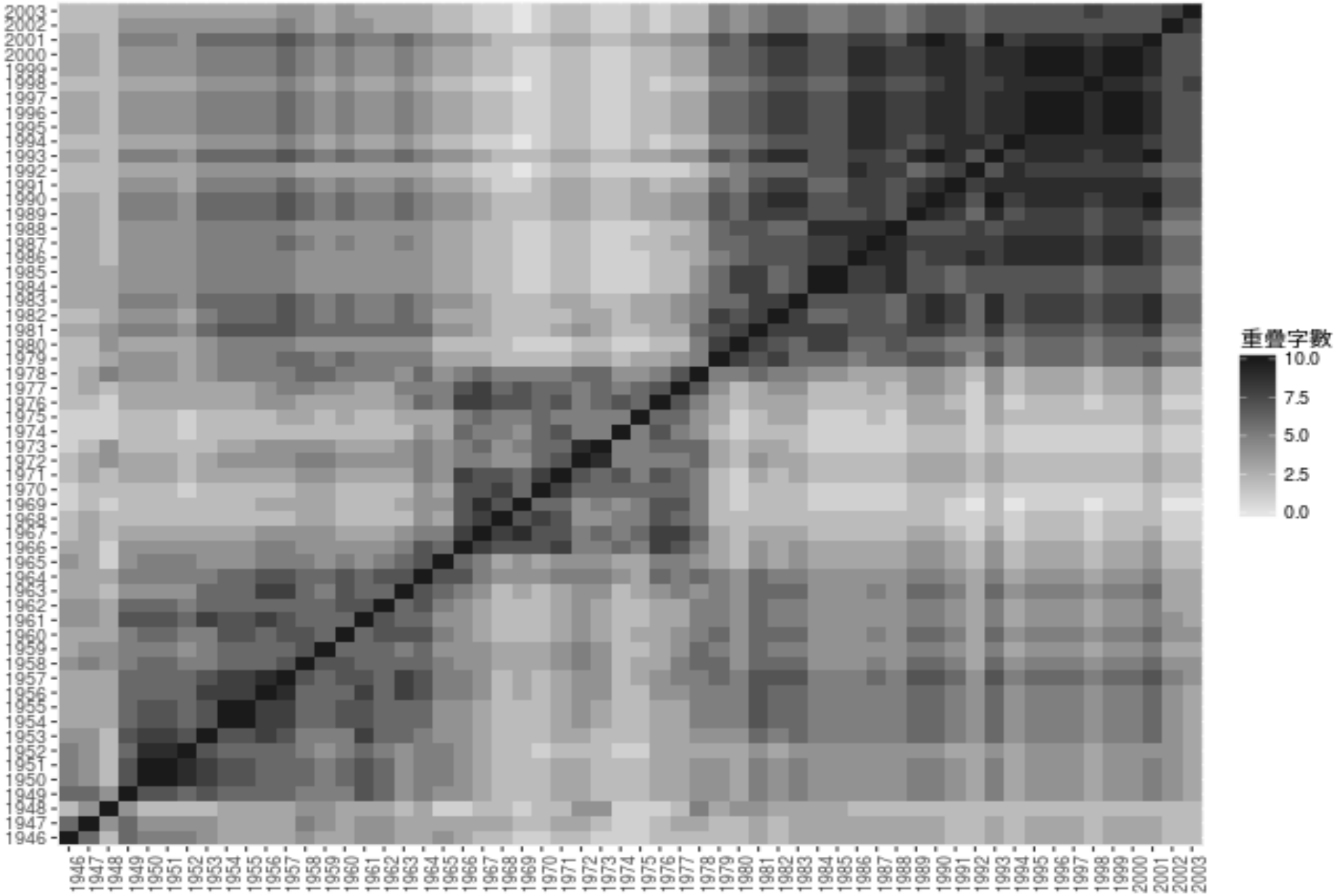
《人民日報》前二十單字重疊度

年度	1988-1990	1991-1993	1994-1996	1997-1999	2000-2002	2003-2005	2006-2008
1988-1990	20	16	16	18	17	16	15
1991-1993	16	20	17	18	17	17	16
1994-1996	16	17	20	18	18	17	17
1997-1999	18	18	18	20	18	18	16
2000-2002	17	17	18	18	20	18	18
2003-2005	16	17	17	18	18	20	18
2006-2008	15	16	17	16	18	18	20

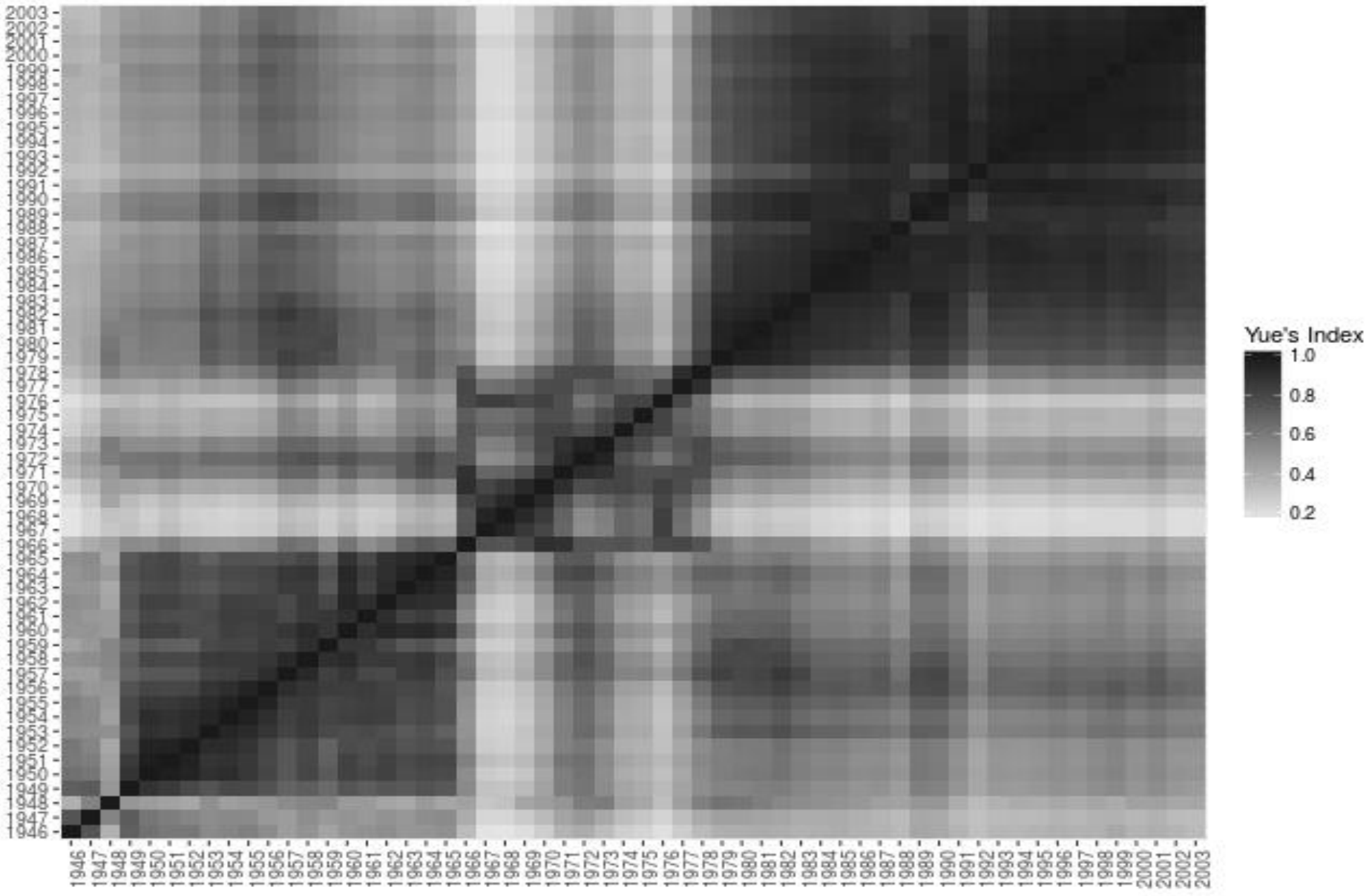
《人民日報》前十單字重疊度



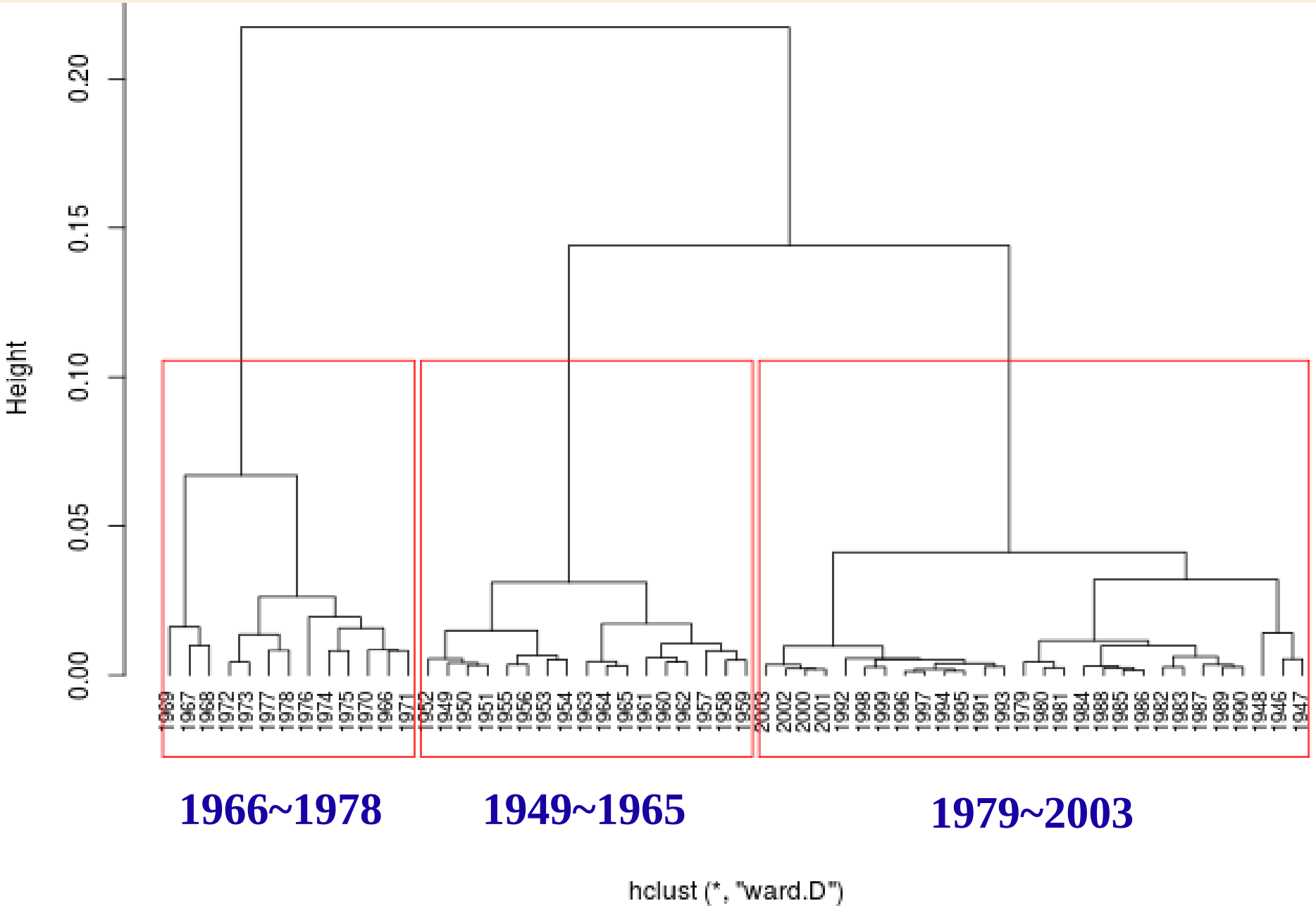
《人民日報》前十雙字詞重疊度



《人民日報》前十雙字詞Yue相似度

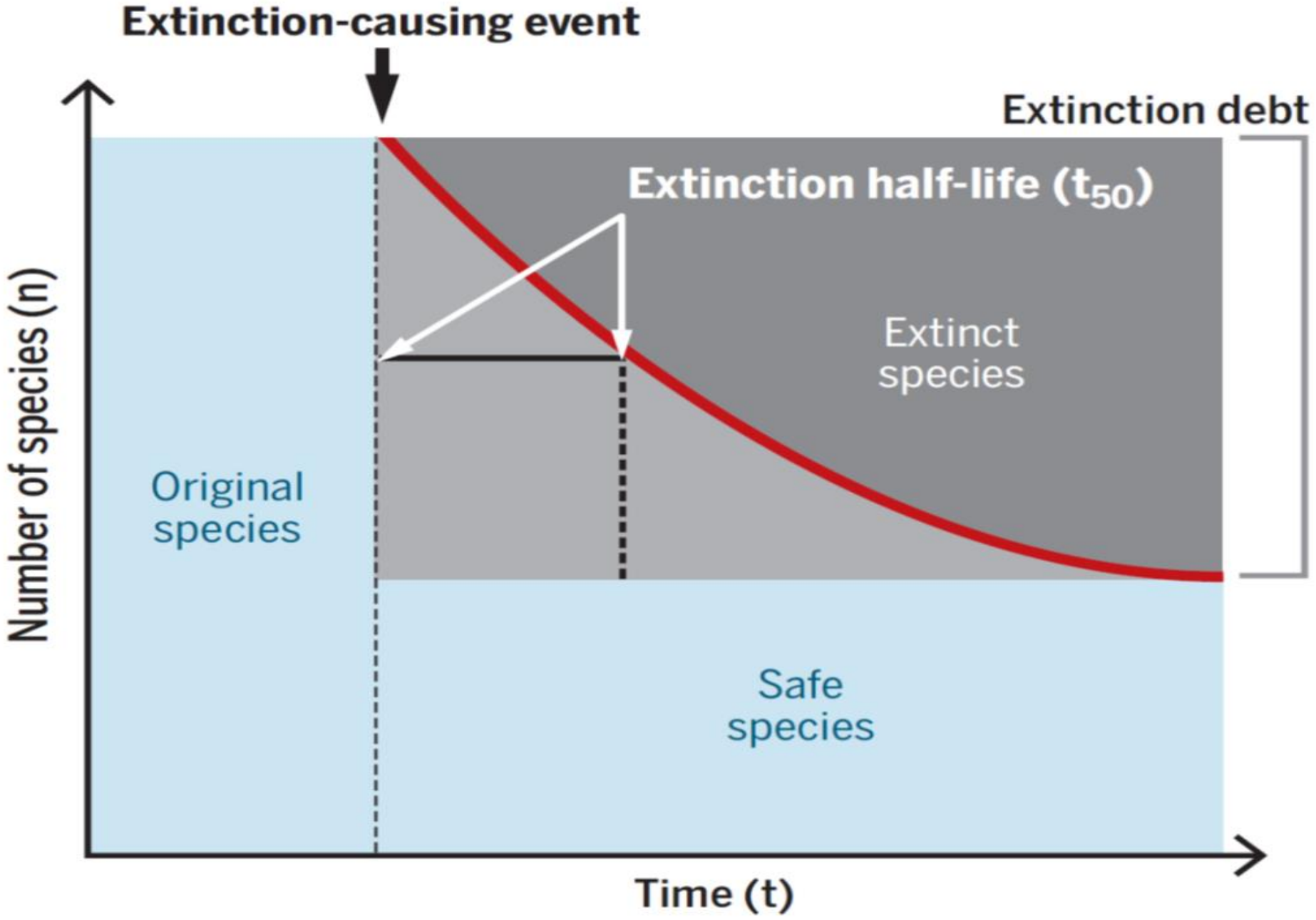


《人民日報》前三十雙字詞集群分析



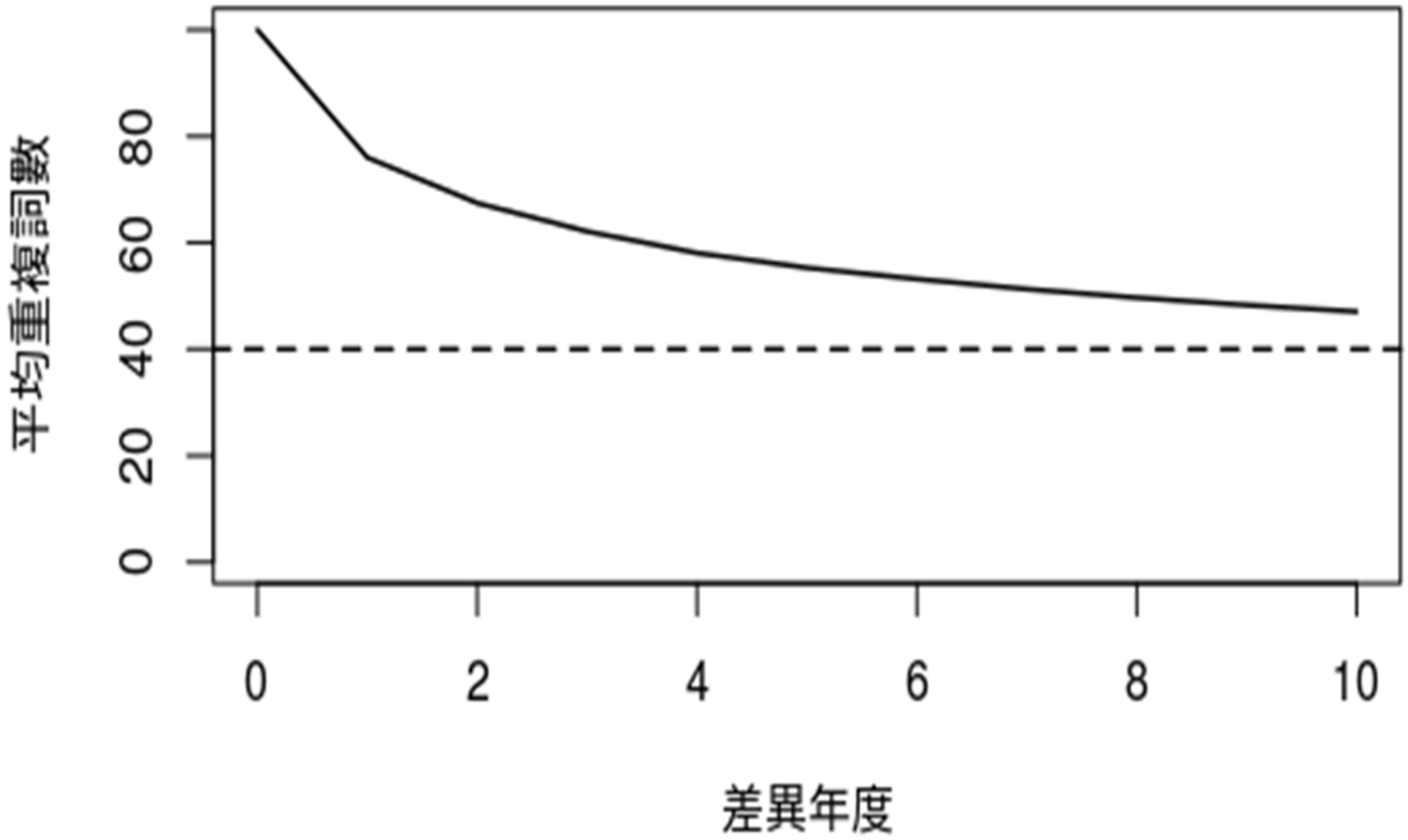
集群分析使用的前三十個雙字詞

中國	人民	政府	我們	代表
他們	一個	問題	會議	群眾
主義	委員	進行	全國	工作
主席	國家	經濟	中央	社會
鬥爭	同志	生產	領導	發展
干部	革命	思想	建設	階級



物種絶種速度(2016 Science)

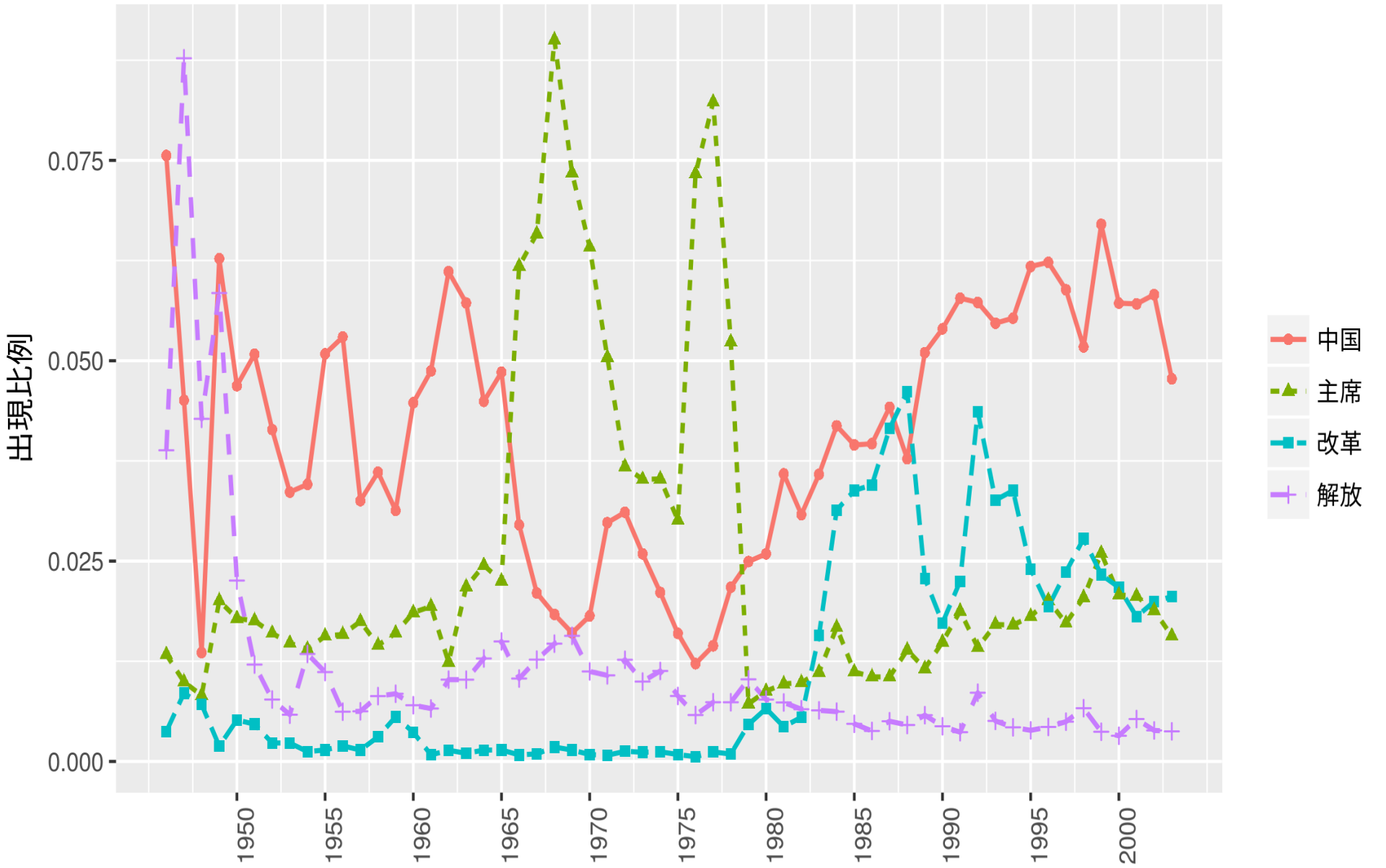
《人民日報》歷年前一百大雙字詞的半生期



《人民日報》前百大雙字詞的變化

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

《人民日報》雙字詞出現比例範例



結論與發現

- 以EDA分析《人民日報》，發現：

→可分為1946-1948、1949-1965、1966-1978、1979-2003四個時期，字詞使用變化快速；

→關聯性指標可用於描述關係變化；

→新生及滅絕字詞可進一步發展應用。

- 透過EDA更能發揮資料驅動(Data Driven)的精神，從資料挖掘找出更多附加價值。

結論與發現

- 透過統計方法分析《新青年》，發現第一卷至第七卷的文字風格有明顯變化。

→字彙豐富度降低、不均度卻升高；

→變化趨勢固定和緩；

→選定合適解釋變數，透過統計模型判別文言文、白話文，準確性相對穩定。

- 其他重要發現：

→用詞改變（代名詞、「今日」vs.「現在」）