資料科學導論與程式設計期末成果報告

台灣升大學考試國寫情意題佳作詞頻相關研究

第三組 想不到組名

朱修平 盧德原 楊舒晴 陳宛瑩

分工

	工作內容
朱修平	- 文章相似性分析
盧德原	- 詞彙之詞頻、詞性與詞意量分析 - 呈現網頁建置
楊舒晴	- 基礎結構分析 - 引用資料分析 - 呈現網頁建置
陳宛瑩	- 基礎結構分析 - 引用資料分析 - 分析結果敘述

研究方法

資料來源

- 大考中心網站每年提供之升大學考試國文作文佳作範本
  1. 民國95年至110年學測佳作
  2. 98年至110年指考佳作
- 共計208篇

資料前處理

- 手動繕打轉換資料為txt.檔
- 統一資料格式
  （分段換行、無空格、完整標點符號、刪除底線、標準化檔名）
- 匯入R、Python進行後續資料分析

研究成果

PART 01 基礎結構分析

PART 02 引用資料分析

PART 03 文章相似性分析

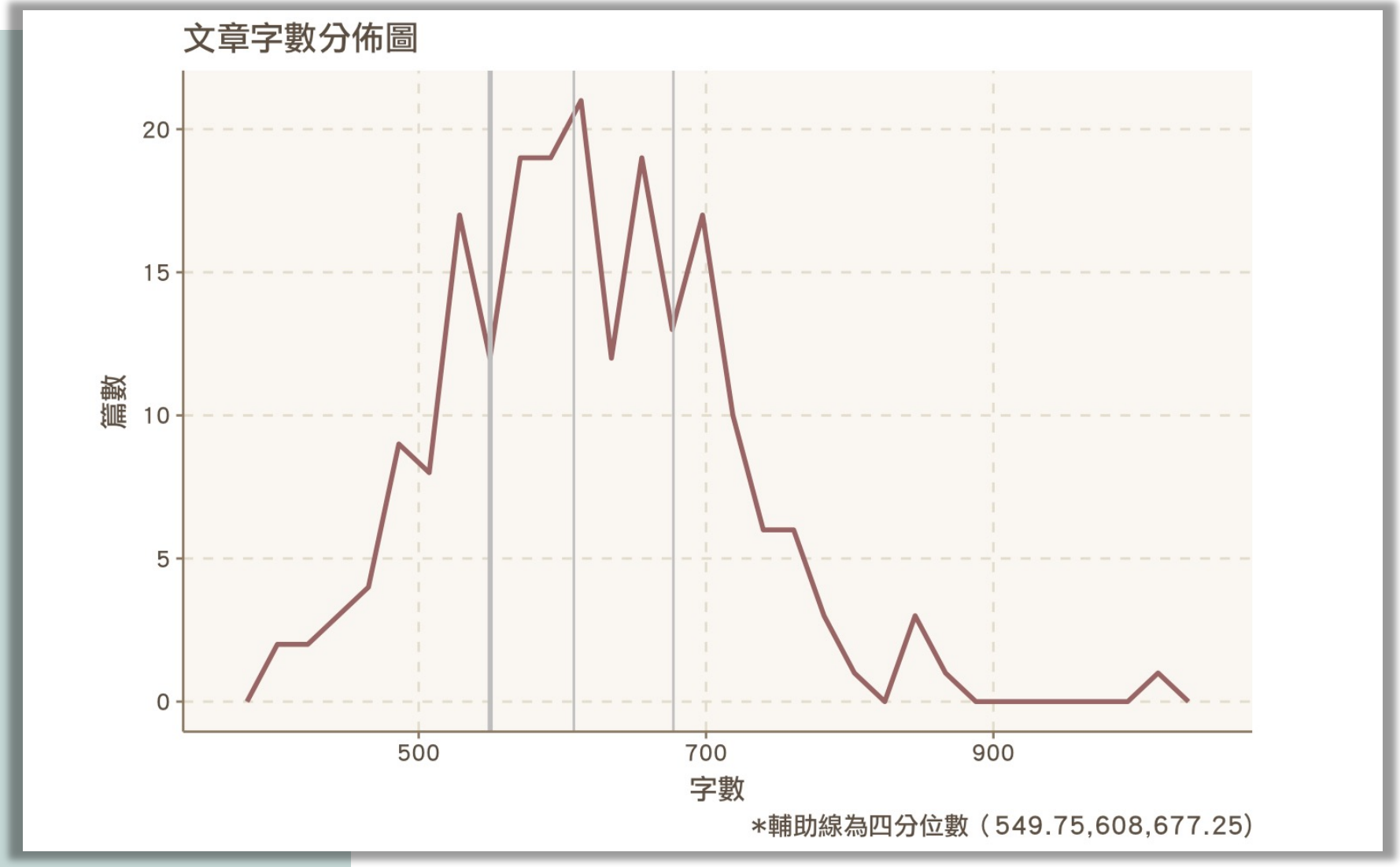
PART 04 詞彙之詞頻、詞性與詞意量分析

基礎結構分析

PART 01

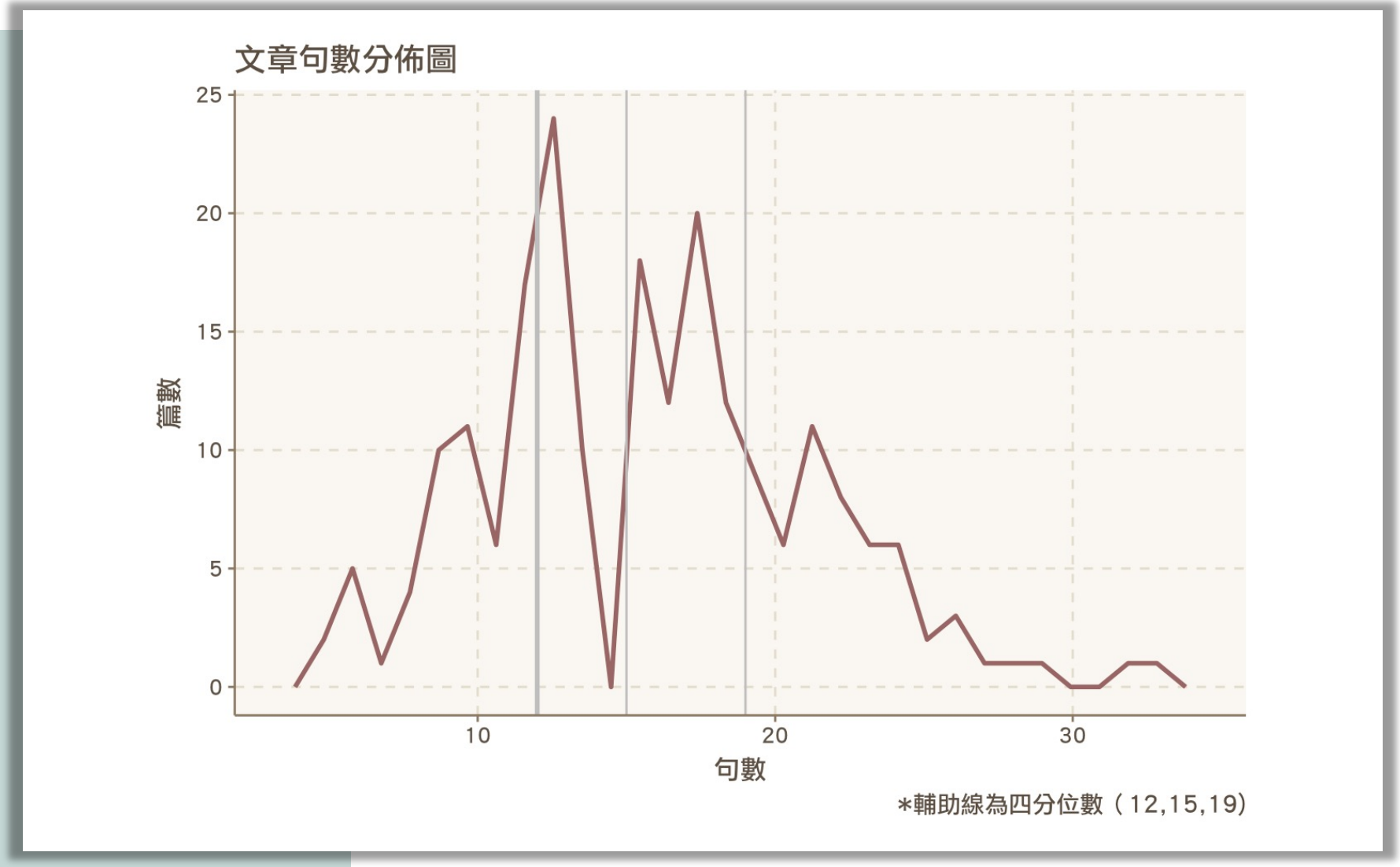
佳作文章長度

- 平均：615.21
- 標準差：94.16
- 中位數：607
- 眾數：650



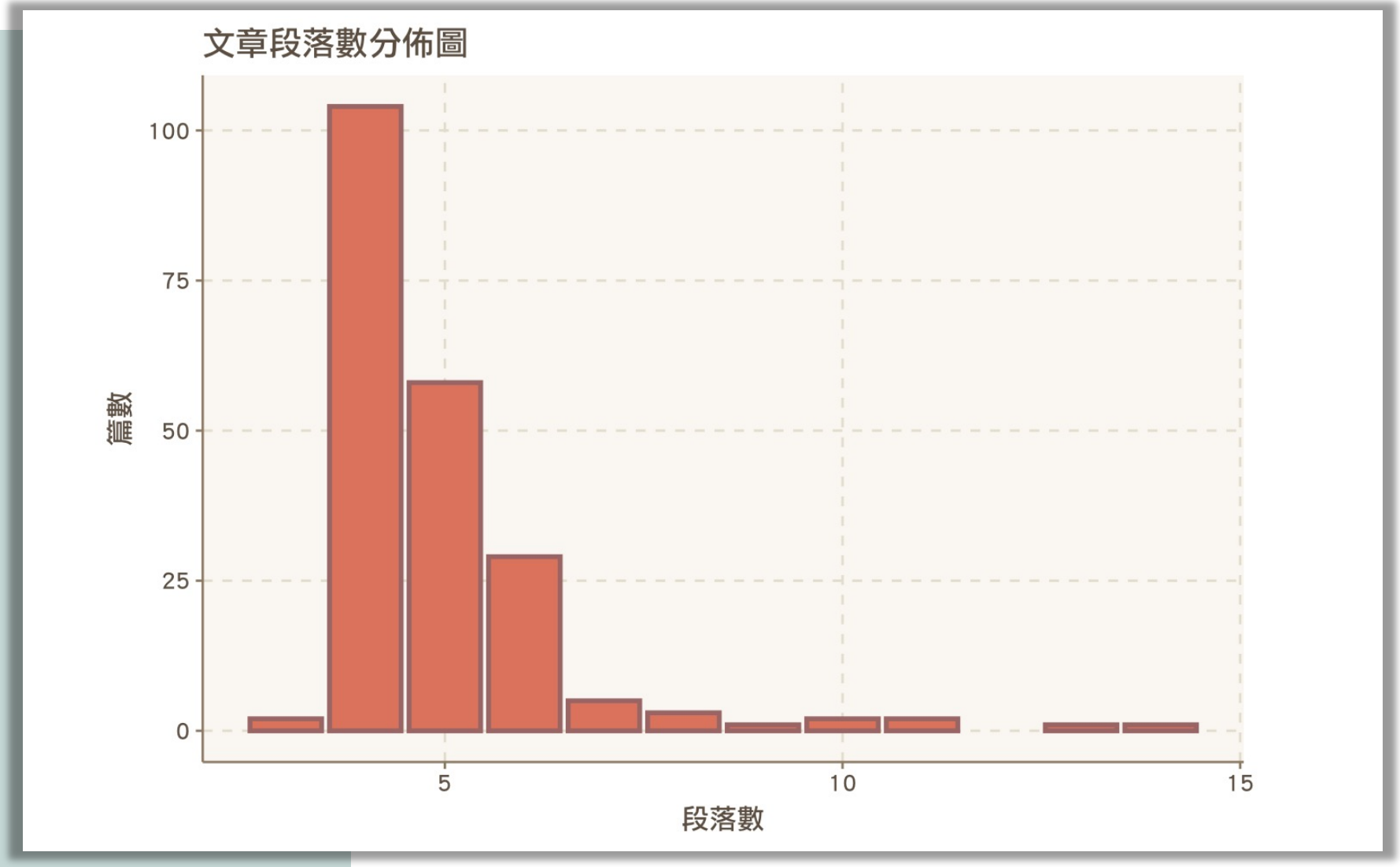
佳作文章句數

- 平均：15.75
- 標準差：5.17
- 中位數：15
- 眾數：13



段落分佈狀況

- 平均：4.91
- 標準差：1.5
- 中位數：4
- 眾數：4



引用資料分析

PART 02

名人偉人

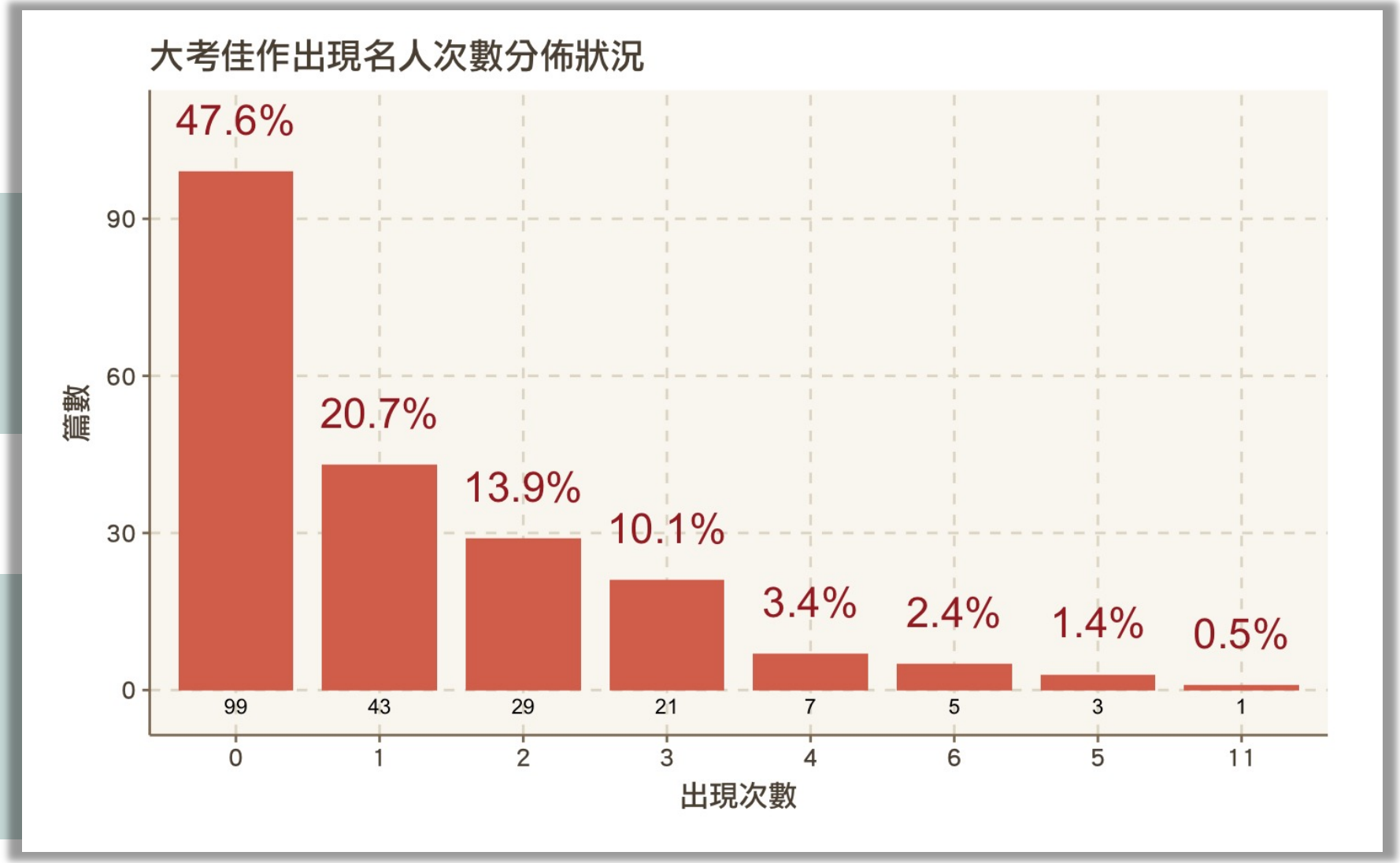
作法

建立含有各種名人偉人名字之.txt檔

發現

並非所有佳作皆會引用名人偉人之事跡或名言

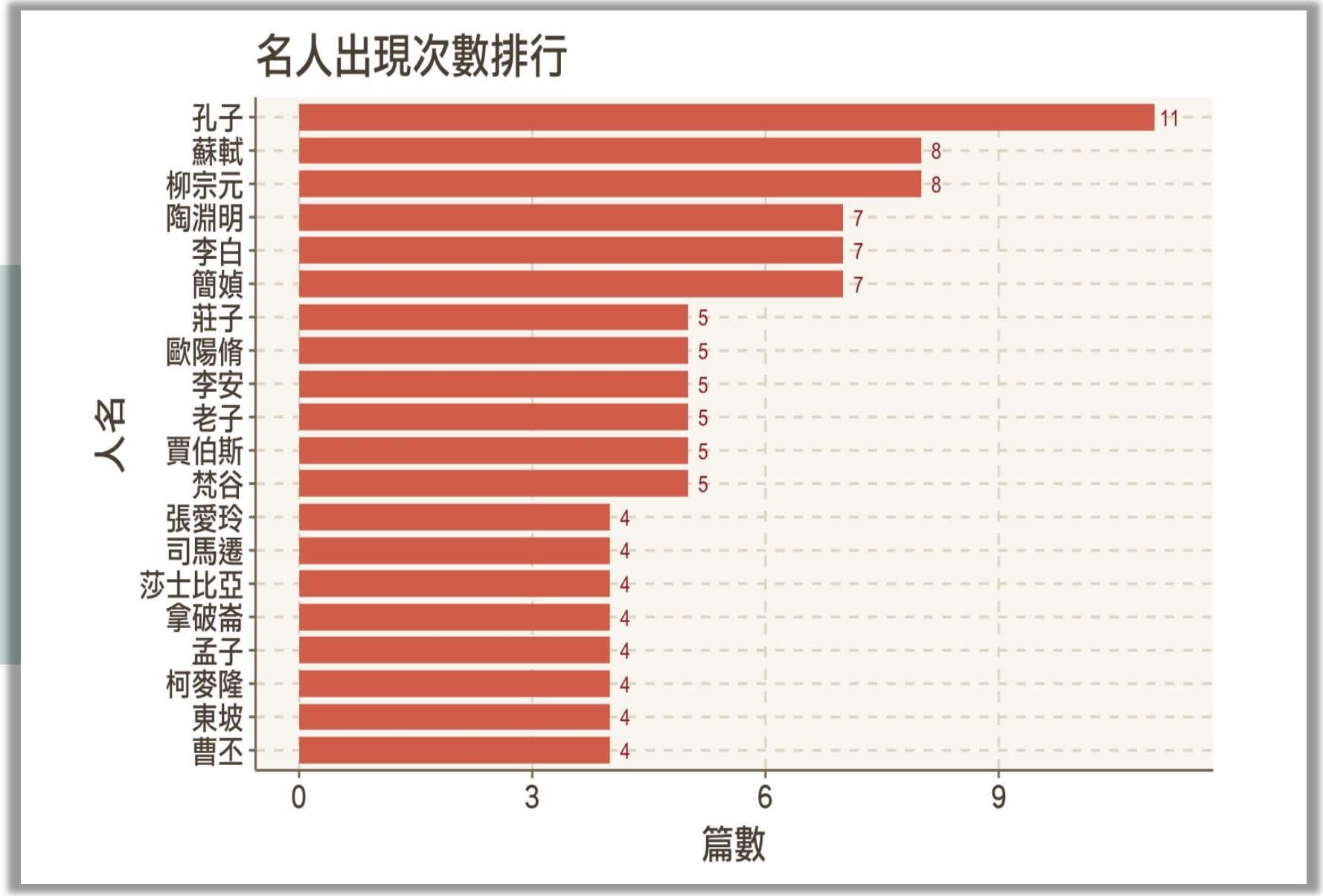
平均出現次數



名人偉人 出現次數排行

發現

引用之名人類型以中國古代文人為主



找相似佳句

作法

ＳＴＥＰ１　設定regular expression

篩選出格式為在冒號後方或在引號內，且長度大於4之句子
→：(\\w|，)**{5,}**(。|？|！)及「(\\w|，|。|！|；|、)**{5,}**?」

ＳＴＥＰ２　運用jieba中的simhash相似度分析

兩兩佳句比較，找出距離小於12之佳句組合

找相似佳句

結果

	類型	年份	文章編號	佳句
第一組	指考	101	12	有自信心的人可以化平凡為偉大，化腐朽為神奇。
	學測	101	1	有自信的人可以化渺小為偉大、化平凡為神奇。
第二組	指考	103	6	人因夢想而偉大
	指考	103	7	人類因夢想而偉大
第三組	指考	101	6	回首向來蕭瑟處，歸去，也無風雨也無晴
	學測	102	10	回首向來蕭瑟處，也無風雨也無晴
第四組	學測	101	1	勝利是屬於堅忍卓絕的人。
	學測	101	14	勝利屬於堅忍卓絕的人。
第五組	學測	109	1	人間是我的根本用情處。
	學測	109	2	人間是我的根本用情處。
⋮				

文章相似性分析

PART 03

相似度分析

作法

找出分析文章中
出現頻率前30個高的
有效字(CKIP)

	0	1	2	3	4	5	6	7	8	9	...	198	199	200
0	人生	才	就	個	中	能	深	中	也	所	...	著	著	著
1	和	深	能	中	所	中	寬	深	不	寬	...	來	陽光	人類
2	各	能	人生	與	與	與	而	寬	能	深	...	山坡	根	人們
3	深	個	深	問題	生命	而	能	要	更	有	...	不再	只	中
4	中	要	越	有	最	深	上	會	那	作曲家	...	怎麼	那	知道
5	與	中	有	更	成為	寬	中	與	但	練習曲	...	但	便	不
6	寬	寬	更	去	使	不	文學	海洋	會	而	...	今天	沒有	同伴
7	生命	未知	活	學習	不同	沒有	面對	只	著	與	...	人類	大地	們
8	便	人	寬度	因為	醫生	就	來自	學問	充實	個	...	內	道	忘
9	牛頓	深度	個	便	能	使	作家	峽谷	人們	每	...	家園	邊	這
10	能	天地	花草	因此	唯有	又	瓦歷斯	心	知識	更	...	不	仍	而
11	廣	永遠	木板	議題	但是	卻	傅山	人	學習	廣泛	...	刀	而	生命
12	心力	西方	生命	可	更	讓	才	可以	各	寬度	...	如此	也	也
13	實踐	強權	中	其實	名	人生	比較	知識	寬大	學習	...	過	下	來自
14	使	無限	樹木	是否	醫師	次	藝術	草原	挑戰	基本功	...	而	片	劈啪
15	但	思想	寬	知識	寬度	累積	與	然而	就	讓	...	那	網	所

相似度分析　文本向量表徵

```
Document-feature matrix of: 208 documents, 29 features (91.71% sparse) and 0 docv
ars.
                features
docs                    寬       人生       深       生命     世人     深度
  AST_100_1.txt  5.438072 5.3141326 6.102200 1.3983215 1.414973 4.244920
  AST_100_10.txt 4.350458 0.7085510 5.085167 0.6991608 2.829947 4.244920
  AST_100_11.txt 4.350458 2.8342041 2.034067 1.0487412 0        2.829947
  AST_100_12.txt 2.175229 0         1.017033 0         0        4.244920
  AST_100_13.txt 1.087614 0.3542755 2.034067 1.7479019 0        4.244920
  AST_100_14.txt 5.438072 1.0628265 6.102200 0         2.829947 0
                features
docs                    廣      無限       越     天地
  AST_100_1.txt  1.716003 0         0        0
  AST_100_10.txt 1.716003 2.926922 1.414973 3.61236
  AST_100_11.txt 0        0         2.829947 0
  AST_100_12.txt 0        0         0        0
  AST_100_13.txt 1.716003 0         0        0
  AST_100_14.txt 0        0         0        0
[ reached max_ndoc ... 202 more documents, reached max_nfeat ... 19 more features
 ]
```

相似度分析

量化比較文本之間的相似度

$$d(\vec{p},\vec{q}) = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_n-q_n)^2}$$

$$cos(\theta) = \frac{\vec{p}\cdot\vec{q}}{\|p\|\|q\|}$$

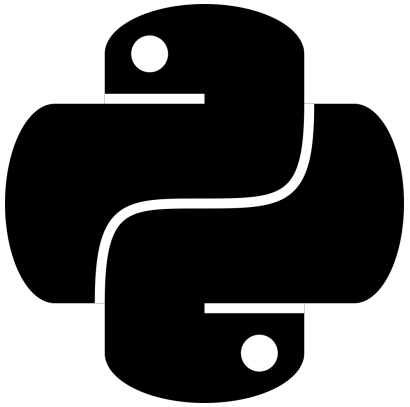
AST_100_1.txt	AST_100_5.txt	AST_100_11.txt	AST_100_14.txt	AST_100_2.txt
1.0000000	0.9649792	0.9228791	0.8584520	0.8566117
AST_100_9.txt	AST_100_8.txt	AST_100_3.txt	AST_100_6.txt	AST_100_13.txt
0.8483397	0.8353900	0.8117013	0.8043561	0.7923995
AST_100_12.txt	AST_100_7.txt	GSAT_99_5.txt	AST_100_4.txt	AST_100_10.txt
0.7863128	0.7645895	0.5575813	0.5367095	0.5320222
GSAT_104_10.txt	GSAT_101_4.txt	GSAT_103_3.txt	GSAT_104_6.txt	AST_101_6.txt
0.5244957	0.4695750	0.4048583	0.3981934	0.3897797

詞彙之詞頻
詞性與詞意量分析

PART 04

詞彙含義分析

- **斷詞與詞性** 使用CKIP LAB開發的ckiptagger套件
- **詞彙含義搜尋** 使用Chinese WordNet的cwn套件



```
from CwnGraph import CwnBase
from ckiptagger import data_utils, construct_dictionary, WS, POS, NER
```

使用ckiptagger斷詞與分析詞性

佛祖見迦葉而拈花微笑，

兩人之間的會心無法為外人共享。

佛祖(Na) 見(VE) 迦葉(Nb) 而(Cbb) 拈花(VA) 微笑(VA) ，(COMMACATEGORY)

兩(Neu) 人(Na) 之間(Ng) 的(DE) 會心(Na) 無法(D) 為(P) 外人(Na) 共享(VJ) 。(PERIODCATEGORY)

中研院平衡語料庫詞類標記集

104學測作文<獨享> 佳作

ADV	Dfb	Dfb	/*動詞後程度副詞*/
ASP	Di	Di	/*時態標記*/
ADV	Dk	Dk	/*句副詞*/
ADV	D	*Dab*, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
N	Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
N	Nb	Nba, Nbc	/*專有名稱*/
N	Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
N	Ncd	Ncda, Ncdb	/*位置詞*/
N	Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
DET	Neu	*Neu*	/*數詞定詞*/.

使用cwn套件計算詞意量

- 果　詞意量＝7
- 渺遠　詞意量＝0（冷僻詞）

分析步驟與分析議題

步驟

1. 斷詞
2. 計算每篇文章每個詞彙的含義數量
3. 整合統計
   - 佳作文章詞彙使用頻率與比較
   - 佳作文章使用詞彙的平均含義量：詞意量分析
   - 佳作文章的冷僻詞彙使用率

佳作詞彙頻率排行

普通名詞 (Na)		
排名	詞彙	頻率
1	人	481
2	心	247
3	生命	234
4	人生	206
5	夢想	107

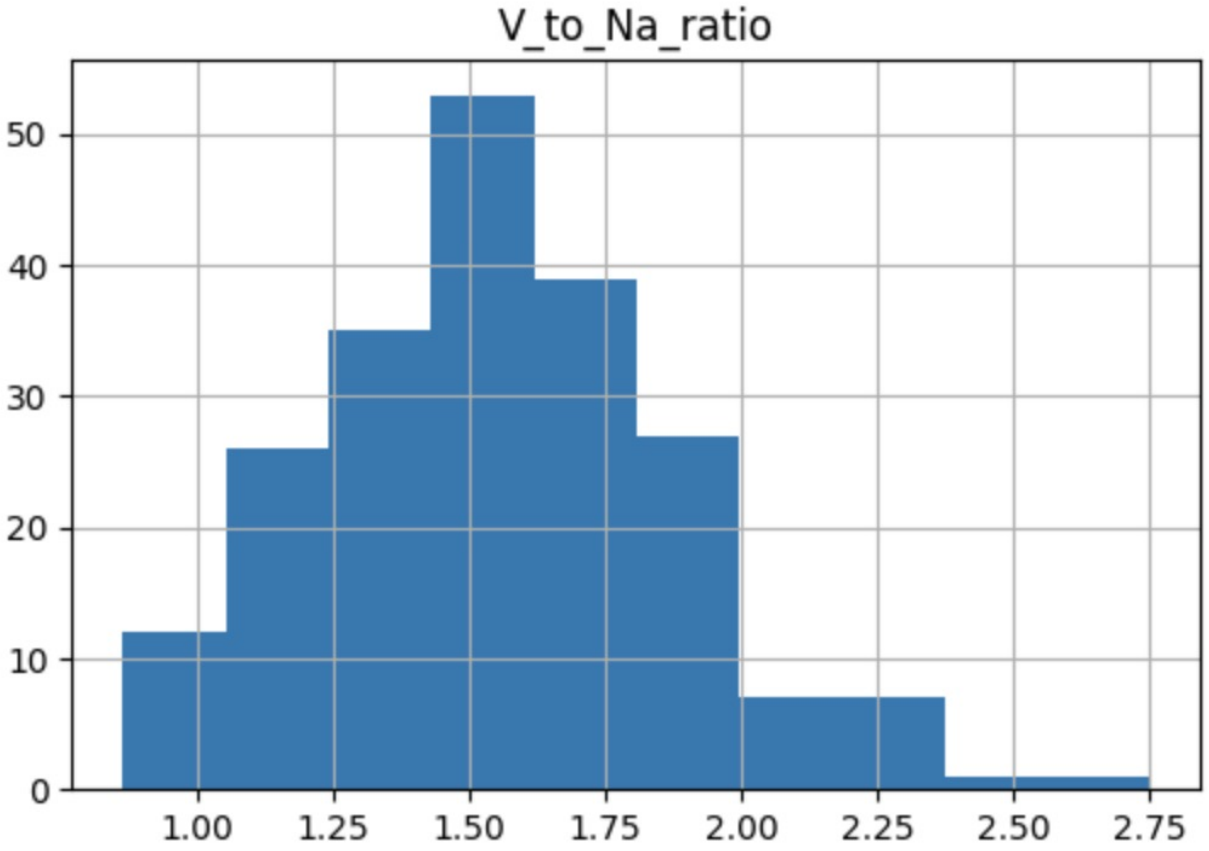
動詞			
排名	詞彙	頻率	詞性
1	讓	183	VL 狀態不及物動詞
2	使	181	VL 狀態不及物動詞
3	説	132	VE 動作句賓動詞
4	大	120	VH 狀態不及物動詞
5	面對	119	VC 動作及物動詞

(使用中研院平衡語料庫詞類標記集)

名詞：Na(普通名詞)

動詞：除 V_2(有) 外所有動詞

各佳作動詞/名詞使用量比例直方圖



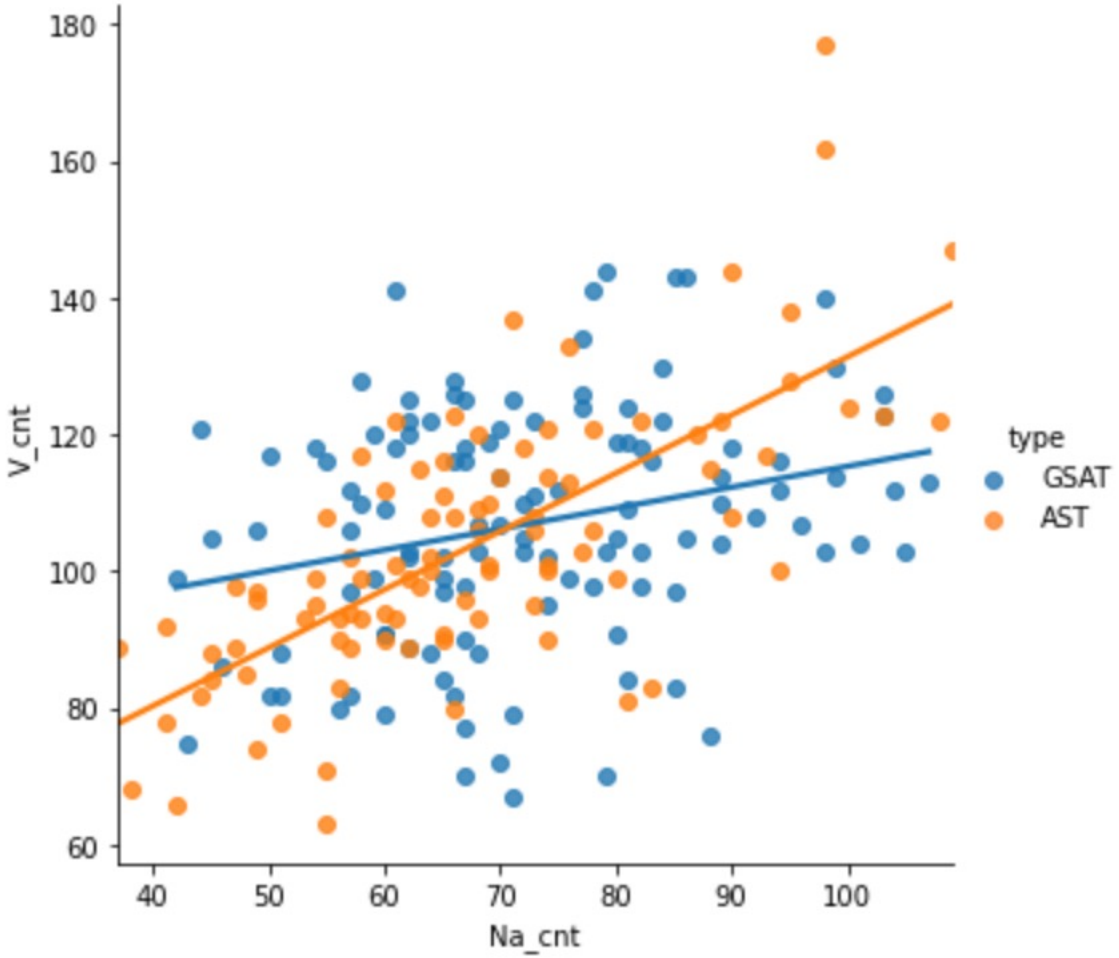
結果

名詞： Nα(普通名詞)

動詞： 除 V_2(有) 外所有動詞

平均動詞/名詞比例 = 1.551

各佳作動詞/名詞使用量分佈圖

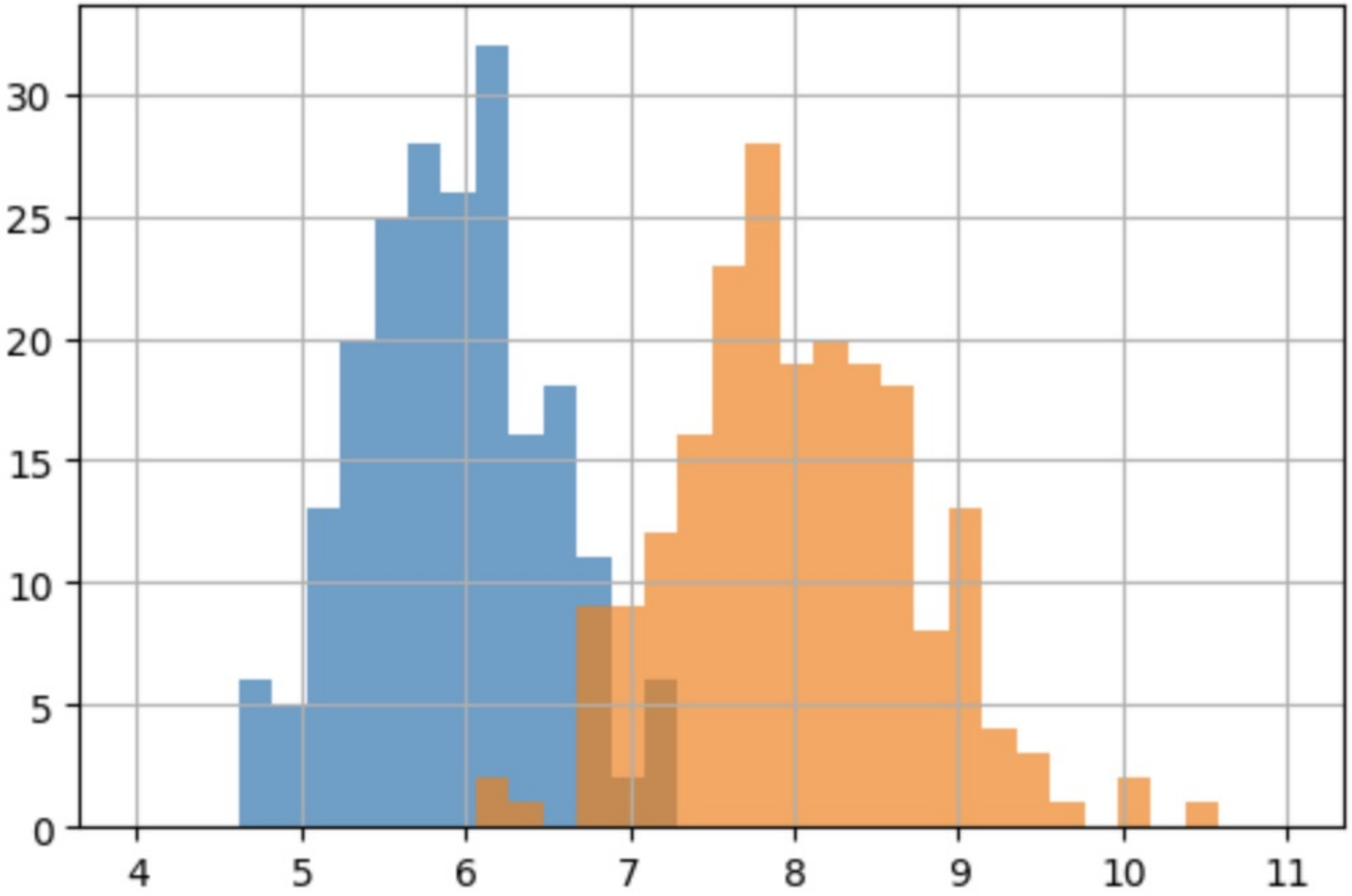


結果

名詞： Na(普通名詞)

動詞： 除 V_2(有) 外所有動詞

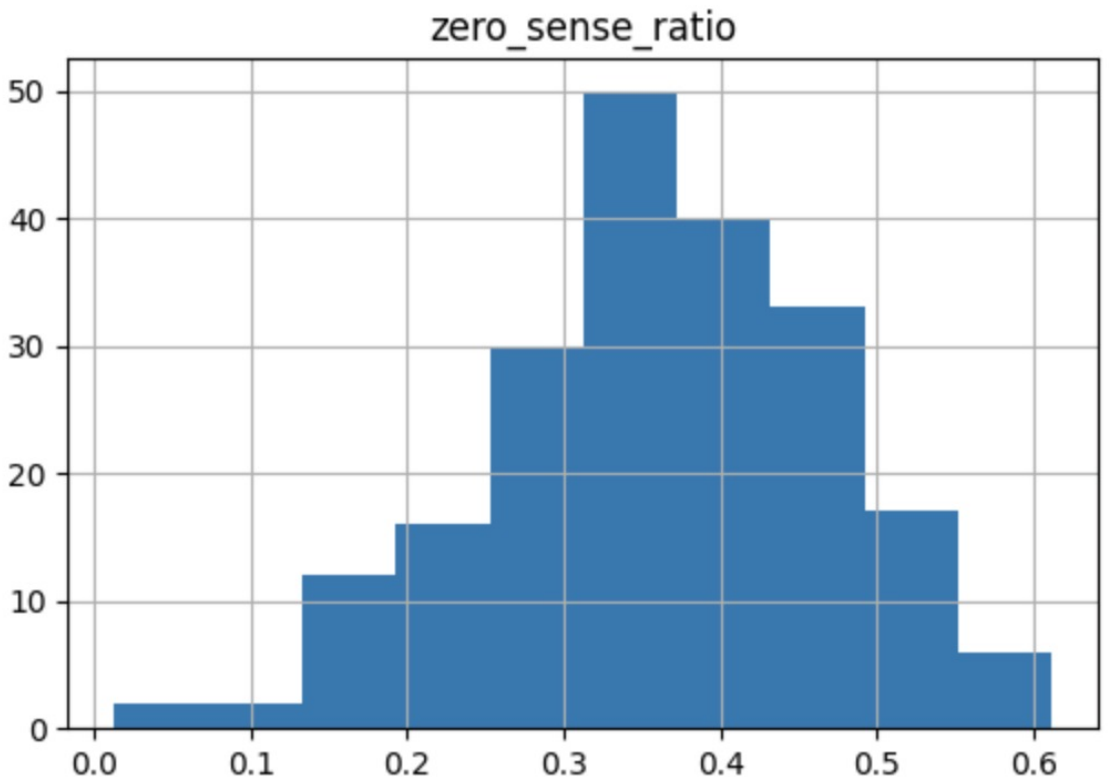
各佳作平均詞意量直方圖



平均詞意量 = 5.911

詞意量非零詞彙平均詞意量 = 8.013

各佳作冷僻詞使用量比例直方圖



視零詞意量者為冷僻詞

冷僻詞使用量比例
＝冷僻詞彙量／總詞彙量

平均冷僻詞使用量＝36.14%

結論

小結

文章結構

字數介於6百上下、總句數平均約15句、
以四段式作文為主

引用資料

無普遍引用名人偉人之現象
引用之佳句類型各異，相似性低

小結

文章相似性

只要同屆的作文篇幅超過5篇，能在相似度大於50%下找到同屆作文的比率超過70%

詞頻分佈狀況

平均詞意量約 6
動詞名詞比例約為 1.5
佳作冷僻詞使用量約 36%