台灣升大學考試國寫情意題佳作研究

rlads2021 Final Project (LING 5505)

G03 想不到組名
朱修平、盧德原、楊舒晴、陳宛瑩

壹、簡介

國寫在升大學考試之國文科目中，與選擇題各佔一半分數之比重，惟我國之國文教學，向來重視古文閱讀、國學常識等，在教學大綱中以選擇題的答寫作為主要教學方向，許多學生對於國寫部分之掌握能力相對不足。因此，本組希望透過分析國寫情意題佳作詞頻，進一步洞察國寫在我國命題方向與佳作取材等寫作方式。

貳、資料取得

本組參酌大考中心每年提供之升大學考試國文作文佳作範本，收錄範圍為民國95年至110年學測與98年至110年指考，共計208篇。另外，由於大考中心係提供範文手寫影本，資料前處理較為耗時，為免壓縮後續資料處理時間，106至110年之學測與指考皆各收錄兩筆資料，並以完整結構之作文為主（新式國寫有部分簡答題）。

參、資料整理

本組以手動繕打將資料輸入電腦轉換成txt.檔，統一資料格式（例如：分段換行、無空格、完整標點符號、刪除底線、標準化檔名）後，再匯入R、python進行後續資料分析。

處理後之資料格式範例顯示如下:

file type	.txt
file name	GSAT_107_1

file content	四季遞嬗，各有各的紛繁絢麗，也各有各的落寞寂寥。然而，一個季節的組成絕不只是天氣上的變化，還要有繚繞的氣味，刻骨的畫面，以及活在季節的人們的故事，最好還要有一支如韋瓦第四季的曲子，使季節如電影般，有開演、有落幕。\n 閉上眼睛，常常想起的畫面通常背景是。

肆、程式碼運作說明

本組將成果報告分成以下4個部分，依序為**基礎結構分析**、**引用資料分析**、**相似性分析**以及**語彙之詞頻詞性與詞意量分析**，詳細說明如下:

一、 基礎結構分析

該部分主要是針對所有佳作的一般性特徵進行調查，從每篇佳作平均的字數量、段落數、句數量著手進行分析，製作的方式使用了老師上課所教授的R語言技巧，將佳作資料匯入r以後，進行初步資料分析，得出結果後，再以ggplot2將該結果以視覺化的方式呈現，並在重要部分加上輔助線，以供辨識。

二、 引用資料分析

此部分可再細分為兩個小段落，分別為名人引用分析及名言佳句引用分析。目的為觀察佳句是否皆會引用名人偉人之事跡，或名言佳句等來豐富文章內容，而當中又有哪些名人偉人以及哪些名言是較常被大家引用的。

（一） 名人引用分析

首先我們上網蒐集了各類型、古今中外名人的名字，以建立names.txt這個含有所有名人名字的文字檔，這些名人包含中國古代歷史人物、文人、小說家、散文家、慈善家、企業家、畫家、導演等等，雖然無法百分之百保證一定沒有缺漏，但已盡可能涵括了所有類型中較知名的名人，而後統計時也會只取出現較多次的名人來看，只出現少次的名人並不會對整體分析有太重大的影響。

建立好names.txt後，運用stringr套件中之str_detect函式逐一跑過所有佳作，去看某名人是否出現在某篇佳作中，再分別去統計「某篇佳作中引用了幾個名人」以及「某個名人出現在幾篇佳作中」，而在做這部分統計時，由於str_detect只會回傳TRUE或FALSE，它並不會去計算某名人出現在某篇佳作中的次數，所以若是同一篇佳作重複提到同個名人多次，統計結果也只會算一次而已，這也是我們選用str_detect而非其他像str_count等計數之函式的原因，

因為作者若想描述一個偉人之事跡，可能會須重複提到他的名字多次，若因此認為「這篇佳作引用了多個名人」，並不適當。

（二） 名言佳句引用分析

這部分我們運用Regular Expression去找出在冒號後方或在引號內的句子，定義它們為佳句，而由於在標示想特別指稱或是強調的詞語時也會使用引號，因此為了初步排除這些不相關的結果，我們設定字數需大於或等於5個字，綜上，Regular expression之設定如以下：`：(\\w|，){5,}(。|？|！)`及`「(\\w|，|。|！|；|、){5,}?」`。篩選出來後，由於當中仍會包含一些日常對話等非引用名言佳句之內容，因此需再透過手動刪除，排除不適當之句子。

接著我們想去看是否有哪些名言佳句是有重複被引用的，這部分我們運用了Jieba中的 SimHash算法來進行相似度分析。它首先會先對文本進行斷詞並賦予每個詞彙一個向量及權重以表達各個詞彙的含意及重要性，接著將每個詞彙的向量和其權重相乘以表示此詞彙，然後計算出每組詞彙向量的總和以表達該文本的含意，最後為了加速運算效率將文本向量的各維度進行化簡，若該維度大於零則更改其值為一，反之則為零，以此化簡後的向量與其它文本的化簡向量進行相比，即可知道文本之間的相似性。程式碼如以下所示，在此我們取距離12做為分界，取出相似性較高之佳句。

```
# 計算相似度
simhasher = worker("simhash", topn = 10)
for(i in 1:(length(good_sent)-1)){
  for(j in((i+1):length(good_sent))){
    sim <- distance(good_sent_clean[i], good_sent_clean[j], simhasher)$distance
    if (sim < 12){
      if (!(good_sent_id[i] == good_sent_id[j])){
        cat(good_sent_id[i], good_sent[i], "\n")
        cat(good_sent_id[j], good_sent[j], "\n")
        cat("\n")
      }
    }
  }
}
```

三、 相似性分析

（一） 首段尾段相似性

因為一個段落的詞彙量就非常少，所以沒有必要再用CKIPtagger去做過濾，因此直接使用jiebrR進行段詞後，使用上課所教的相似度分析函式，得出我們要的結果。

```
seg <- worker()
first_break <- vector("character", length(post))
last_break <- vector("character", length(post))
for(i in seq_along(first)){
  segged_F <- segment(first[i], seg)
  segged_L <- segment(last[i], seg)
  first_break[i] <- paste(segged_F, collapse = "\u3000")
  last_break[i] <- paste(segged_L, collapse = "\u3000")
}

for(i in seq_along(post)){
  a <- c(first_break[i], last_break[i])
  # print(a)
  quanteda_corpus_a <- corpus(a) %>%
    tokenizers::tokenize_regex(pattern =  "\u3000") %>%
    tokens()
  q_dfm <- dfm(quanteda_corpus_a) %>%
    # dfm_remove(pattern =  readLines("../source/stopwords.txt"), valuetype
    dfm_select(pattern = "[\u4E00-\u9FFF]", valuetype = "regex") %>%
    dfm_tfidf()
  similar[i] <- cossim(q_dfm[1, ], q_dfm[2, ])
}
print(similar)
```

（二） 文章相似度

首先使用CKIPtagger與CWN的套件進行斷詞，並且去除Stopwords之後，取出每個文章前30個出現頻率最多的詞，之後同樣使用上課所教的相似度分析去得出結果。

```
df <- read_excel("../source/frequency_30_word.xlsx")
test <- unname(unlist(as.list(df["1"])))

quanteda_corpus <- corpus(df_break,
                          docid_field = "id",
                          text_field = "content") %>%
  tokenizers::tokenize_regex(pattern =  "\u3000") %>%
  tokens()

q_dfm <- dfm(quanteda_corpus) %>%
  dfm_select(test) %>%
  # dfm_remove(pattern =  readLines("../stopwords.txt"), valuetype = "fixe
  dfm_tfidf()
q_dfm

doc_sim <- textstat_simil(q_dfm, method = "cosine") %>% as.matrix()
sort(doc_sim["AST 100 1.txt", ], decreasing = T)[1:20]
```

四、 語彙之詞頻詞性與詞意量分析

這個部分我們使用Jupyter Notebook以Python進行分析，部分資料處理的說明已附在此頁面中，此處僅對幾個重要的函示進行詳述。

我們利用CKIPtagger與CWN的套件進行斷詞與詞意搜尋，說明文件顯示CKIPtagger的斷詞正確率比Jeiba好，且其詞性標記使用中研院平衡語料庫詞類標記集，提供更豐富的詞性分析。

Tool	(WS) prec	(WS) rec	(WS) f1	(POS) acc
CkipTagger	97.49%	97.17%	97.33%	94.59%
CKIPWS (classic)	95.85%	95.96%	95.91%	90.62%
Jieba-zh_TW	90.51%	89.10%	89.80%	--

斷詞正確率比較

此部分的分析步驟為，先以CKIPtagger對每篇文章進行斷詞，並將斷詞結果儲存在該文章的一個欄位中，以保留對每個斷詞所屬文章、年份與考試類型的資料屬性，然後我們會依序提取或拆分資料以進行各項分析。

利用CWN套件的辭意搜尋函數，可以寫出一個搜尋並列出輸入單詞所有詞意的函式，如下所示。

```
cwn = CwnBase()

def all_sense_tree (word, verbal = False):
    cnt = 0
    for i in range(len(word)):
        snese_tree = word[i].senses
        cnt += len(word[i].senses)
        if(verbal == True): print(snese_tree)

    if(verbal == True): print("total senses = ", cnt)
    return cnt
```

列出所有詞意樹函式

有了上述函示後，我們依序每篇文章的每個詞彙進行詞意量計算，計算函式如下所示，在208篇文章中我們得到493,482條詞意。

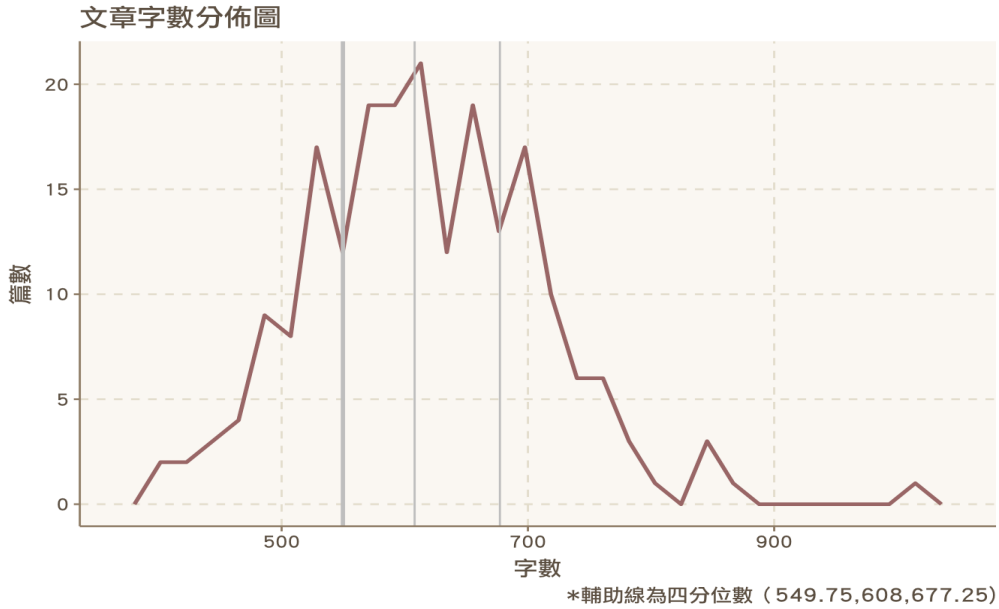
```
for i in range(all_list.shape[0]):
    senses_list = list()
    arr = pd.Series(word_sentence_list[i])
    ttl = 0
    for j in range(len(arr)):
        if(pos_sentence_list[i][j] not in pun_set):
            _word = arr[j]
            word = cwn.find_lemma("^" + _word + "$")
            sense_cnt = all_sense_tree(word)
            senses_list.append((arr[j], pos_sentence_list[i][j], sense_cnt))
```

對所有單詞計算詞意量函式

伍、結果

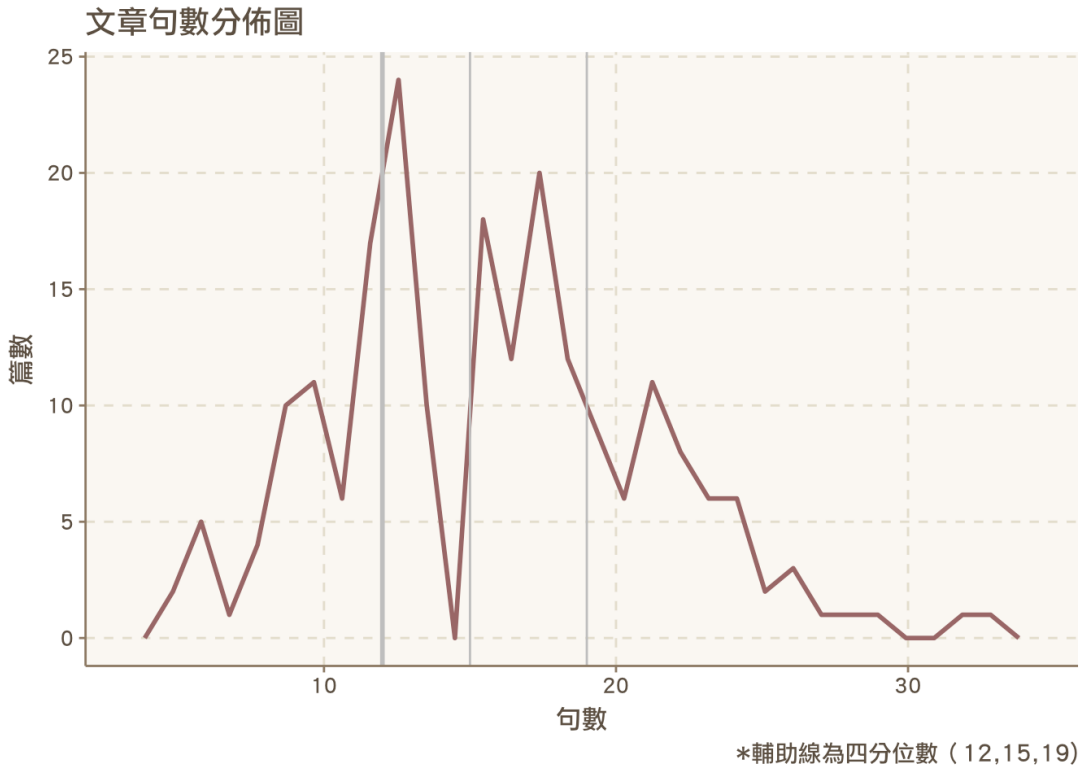
一、基礎結構分析

（一）佳作文章長度



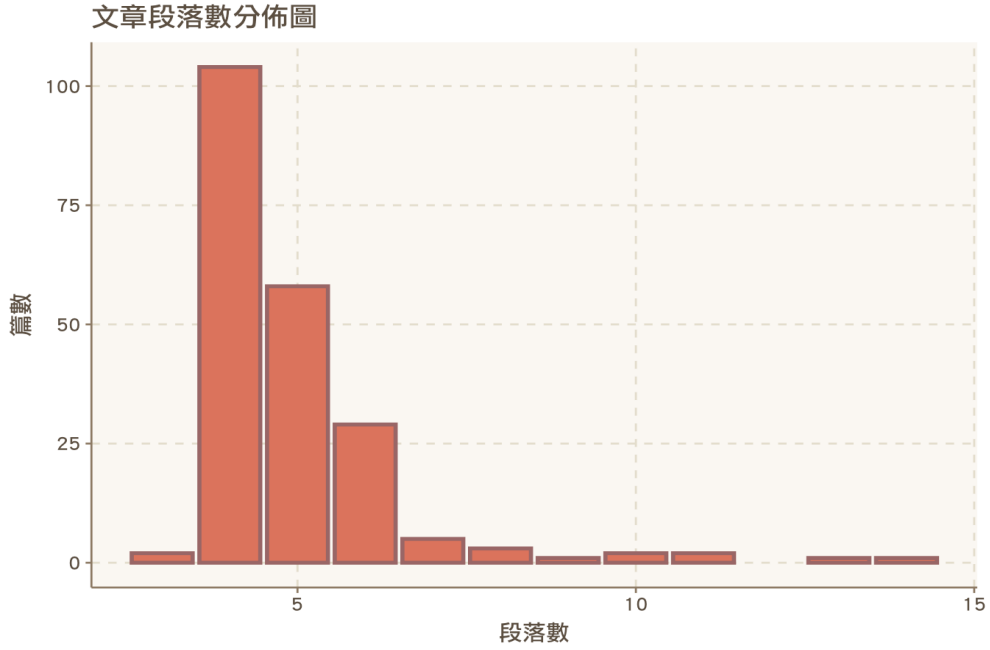
從該圖可以發現，文章字數多為600字上下，只有少部分的作文偏離該分佈到達800字或小於500字。因此，我們可以發現，作文與學校課綱所強調的600字作文吻合。

（二）文章句數



從文章句數來看，可以發現每篇作文介於15句上下。

（三）段落分佈狀況



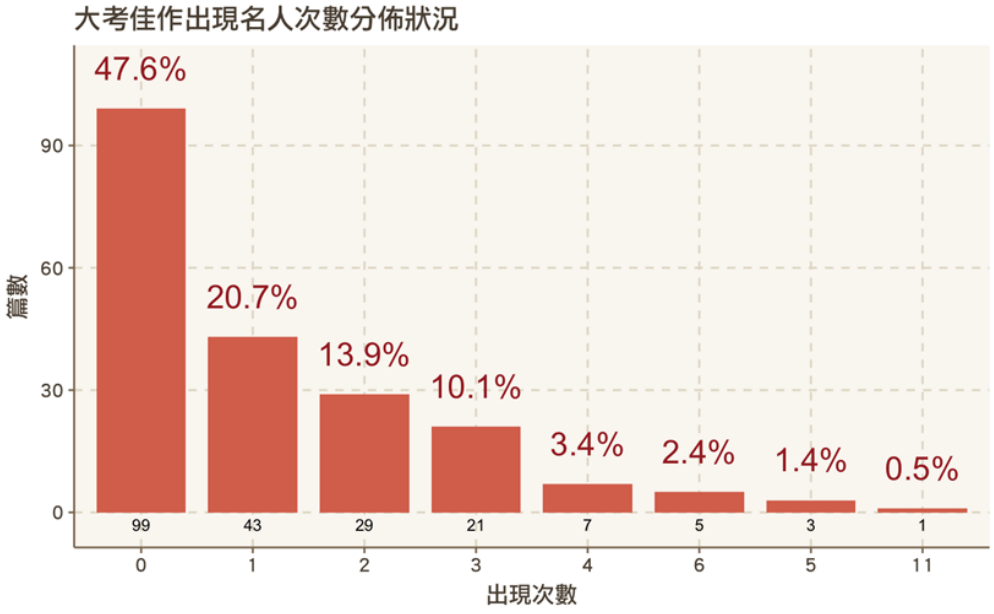
從段落數來看，佳作文章整體以傳統的四段式作文為主。

二、引用資料分析

（一）　名人引用分析

1.　一文中引用名人偉人數目

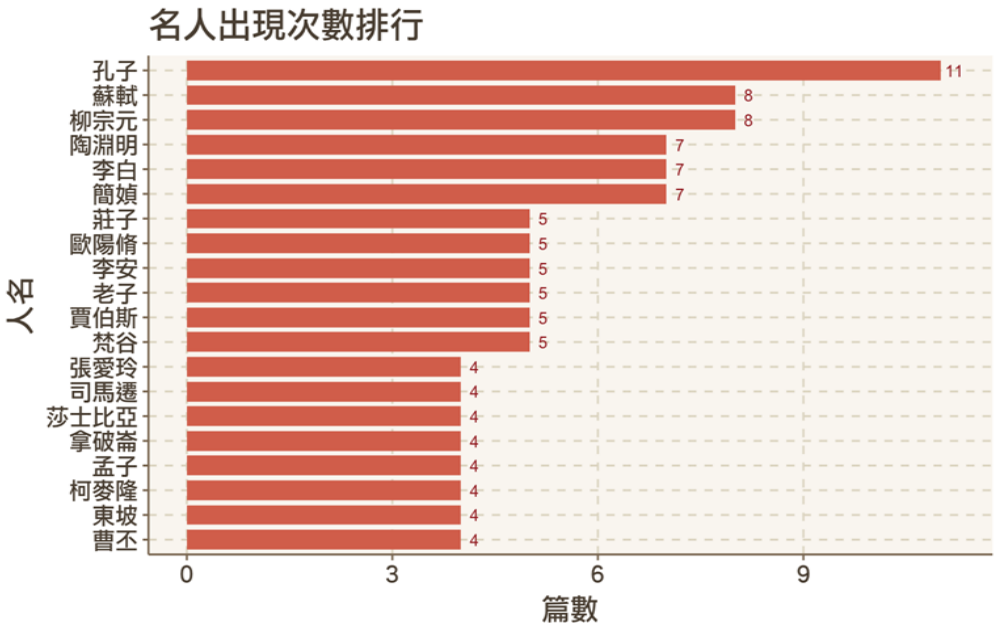
由下圖結果可看出並非所有佳作皆會引用名人偉人之事跡，沒有引用的就佔了47.6%，將近半數，而只引用一個與兩個名人的佳作也分別佔了20.7%與13.9%，有趣的是，其中有一篇佳作引用了11個名人之多。



圖一：一文中引用名人偉人數目

2. 名人偉人出現頻率

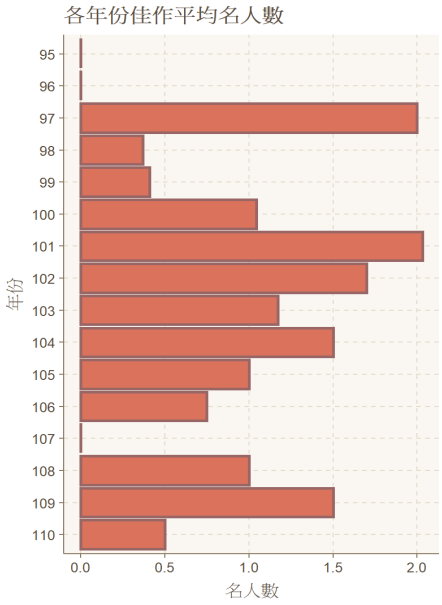
此部分我們只取出現次數在四次以上之名人偉人來做圖呈現，共計有20個，從下圖呈現之結果可以看出，佳作文章中的名人出現頻率以中國古人為主，其中又以孔子出現次數最多，又上述出現之中國古人物，皆屬課綱30古文之作者。



圖二：名人偉人頻率

3. 年份比較

由下圖觀察，各年之平均引用名人數起伏不定，並無一定趨勢，但若將各年之題目納入一起對照觀察，可以發現若該年度作文主題非典型之抒情、敘事文，名人引用比例有較高的趨勢。



歷年作文題目

年份	學測題目	指考題目
95年	雨季的故事	
96年	走過	
97年	如果當時…	
98年	逆境	惑
99年	漂流木的獨白	應變
100年	學校和學生的關係	寬與深
101年	自勝者強	我可以終身奉行的一個字
102年	人間愉快	遠方
103年	通關密語	圓一個夢
104年	獨享	審己以度人
105年	我看歪腰郵筒	舉重若輕
106年	關於經驗的N種思考	在人際互動中找到自己
107年	季節的感思	
108年	溫暖的心	
109年	靜夜情懷	
110年	如果我有一座新冰箱	

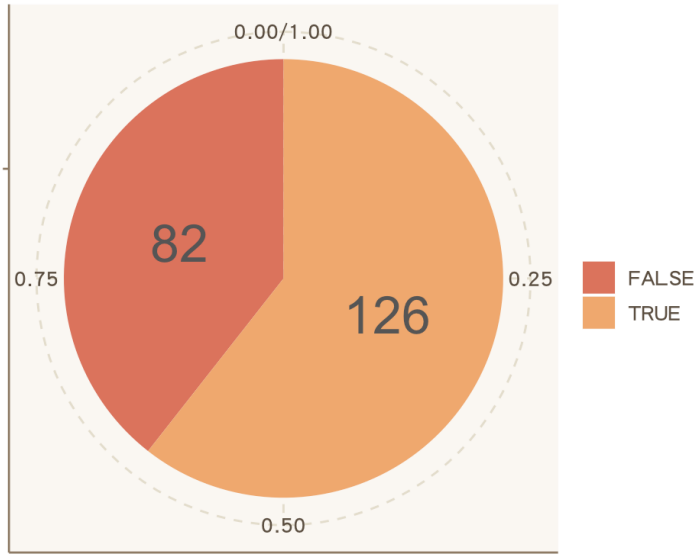
圖三：各年份比較

（二） 名言佳句引用分析

1. 是否引用佳句

從下圖可以看出，有引用佳句之文章約有6成，超過半數。

引用佳句之文章比例



2. 佳句集錦

在經過Regular Expression及手動清理後，208篇佳作中共計有引用245句之名言佳句，以下截取部分結果，完整之佳句集錦可在網頁中取得。

```
[232] "人間是我的根本用情處。"
[233] "彼可取而代之"
[234] "大丈夫當如是也"
[235] "時不利兮騅不逝，騅不逝兮可奈何"
[236] "項籍有志有恆，終成大業。"
[237] "一注血脈，四支分流。"
[238] "人生不如意，十之八九。"
[239] "最美麗的詩歌是最絕望的詩歌，有些不朽的篇章是純粹的眼淚。"
[240] "失敗的人，沒有悲觀的權利。"
[241] "比海洋寬廣的是天空，比天空寬廣的是你我的心。"
[242] "衡量付出的準則就是永無止境的奉獻"
[243] "只有歷經了無數高溫的試煉與考驗，才能製造出世上無堅不摧的寶劍，只有歷經碎石
攻擊，才會有蚌殼中那一顆圓潤飽滿，晶瑩剔透的珍珠。"
[244] "逝者如斯夫，不舍晝夜。"
[245] "天生我材必有用，千金散盡終復來。"
```

3. 重複出現之佳句

篩選出來後，總共有八組相似之佳句，都是兩兩相似，並沒有任何一組為三句以上相似之佳句，其實佳句之引用蠻多樣化的，相似句之比例並不高。而值得注意的是，這些重複出現之佳句中，其中有三組(第二組、第四組及第五組)皆為來自同類型、同年份之文章者，也就是針對同一題目去引用相同的佳句。

	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。
第六組	指考	101	4	採菊東籬下，悠然見南山。
	學測	103	5	採菊東籬下，悠然見南山
第七組	指考	101	6	榮華或清苦，都像第一遍茶，切記倒掉，飲到路斷夢斷，自然回甘。
	指考	103	10	榮華或清苦，都像第一遍茶，切記倒掉，飲到路斷夢斷，自然回甘。
第八組	指考	102	11	青年，多麼美麗，是書的第一章，是永無終結的故事。
	學測	102	13	青年，多麼美麗！是書的第一章，是永無終結的故事。

三、相似性分析

（一）　首段尾段相似度

```
> print(similar)
  [1]   0   0   0   0   0   0   0   0   0   0   0   0 NaN   0   0   0   0   0
 [19]   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 [37]   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 [55]   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 [73]   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 [91] NaN   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
[109]   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
[127]   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
[145]   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
[163]   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
[181]   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
[199]   0   0   0   0   0   0   0   0   0   0
```

得出的結果不盡人意，首段尾段的相似度不高。

（二）　文章相似分析

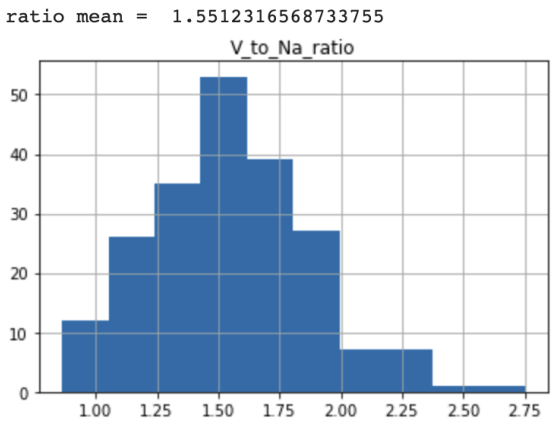
```
> sort(doc_sim["AST_100_1.txt", ], decreasing = T)[1:20]
  AST_100_1.txt   AST_100_5.txt  AST_100_11.txt  AST_100_14.txt
      1.0000000       0.9649792       0.9228791       0.8584520
  AST_100_2.txt   AST_100_9.txt   AST_100_8.txt   AST_100_3.txt
      0.8566117       0.8483397       0.8353900       0.8117013
  AST_100_6.txt  AST_100_13.txt  AST_100_12.txt   AST_100_7.txt
      0.8043561       0.7923995       0.7863128       0.7645895
  GSAT_99_5.txt   AST_100_4.txt  AST_100_10.txt GSAT_104_10.txt
      0.5575813       0.5367095       0.5320222       0.5244957
 GSAT_101_4.txt  GSAT_103_3.txt  GSAT_104_6.txt   AST_101_6.txt
      0.4695750       0.4048583       0.3981934       0.3897797
```

可以看到和100年指考作文相似度大於0.5的作文當中，13個文章來自於同屆作文。

四、語彙之詞頻詞性與詞意量分析

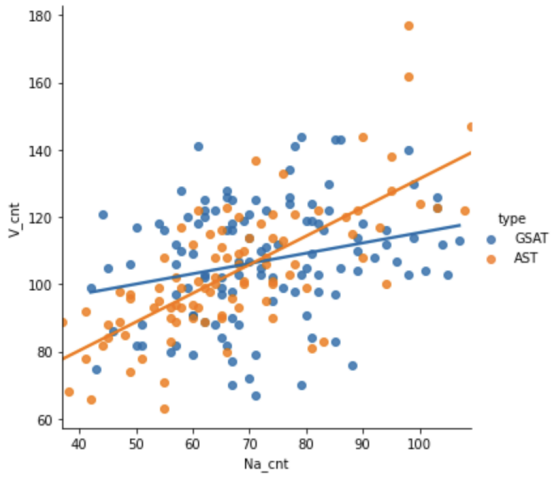
（一）　動詞名詞比例

我們想知道佳作文章的動詞對名詞使用比例為何，以中研院平衡語料庫詞類標記集中除V_2(有)以外的各種動詞為動詞計算，以N_a(普通名詞)為名詞計算，得到平均的動詞名詞比例為1.551，整體呈現鐘型分佈，如下所示。



動詞對名詞比例直方圖

同時，我們也將資料拆分成學測與指考來看，顯示指考的比例比學測多，如下所示。



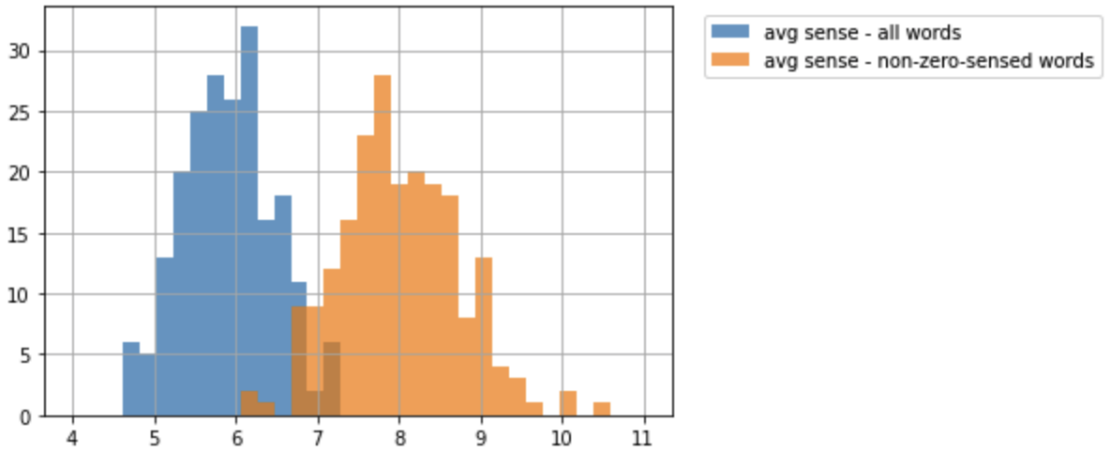
動詞對名詞數量散佈圖

我們亦分別對兩組資料進行回歸分析，分析結果放置於呈現頁面。

（二） 詞意量分析

我們以上述程式碼進行詞意量的計算與分析，初步結果顯示平均詞意量為5.9917，顯示在佳作中平均每個用詞含有約6條詞意。若扣除詞意為零的單詞後，平均詞意量為8.0131，顯示在佳作非零詞意中平均每個用詞含有約8條詞意。直方圖如下所示。

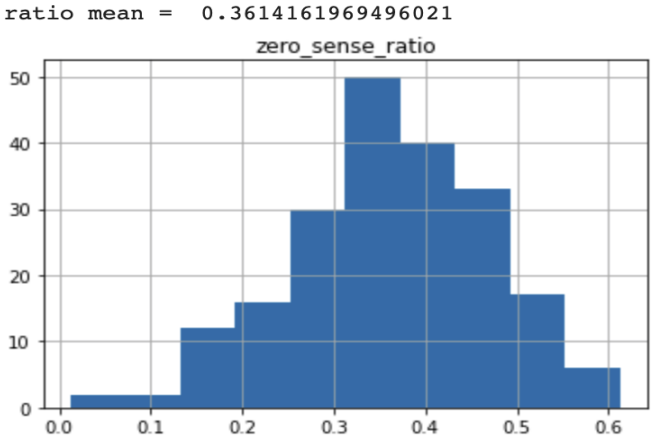
```
average sense from all words =  5.9117085939980205
average sense from non-zero-sensed words =  8.013096983316954
```



平均詞意直方圖

（三）　冷僻詞分析

普遍認為詞藻華麗、用艱澀冷僻的詞彙堆砌出來的文章可能比較吸睛，於是我們想知道這些佳作中的冷僻詞彙使用狀況為何。我們定義詞意量為零的單詞為冷僻詞，將一文章之詞意為零單詞數量除以所有單詞數量後即可得到該文章的冷僻詞使用比例，計算出所有文章的冷僻詞使用比例平均為36.1416%，直方圖如下所示。



冷僻詞使用比例直方圖

（四）　詞性詞頻分析

最後，我們利用中研院平衡語料庫詞類標記集篩選出來的詞性分析特定詞性的詞彙出現頻率，以下為所有詞性、普通名詞與除V_2(有)以外之動詞的結果。

排名	所有詞性	CKIP詞性	頻率	普通名詞(N_a)	頻率	所有動詞(除V_2)	CKIP詞性	頻率
1	的	DE	7052	人	481	讓	VL	183
2	我	Nh	2705	心	247	使	VL	181
3	一	Neu	1454	生命	234	說	VE	132
4	是	SHI	1362	人生	206	大	VH	120
5	在	P	1229	夢想	107	面對	VC	119
6	了	Di	836	學生	86	沒有	VJ	107
7	不	D	699	逆境	83	想	VE	106
8	著	Di	652	夢	81	深	VH	88
9	自己	Nh	601	郵筒	78	看	VC	88
10	而	Cbb	569	生活	78	愉快	VH	87

部分詞性之詞彙頻率表

陸、討論與貢獻

一、 基礎結構分析

從基礎結構分析中，我們可以看出佳作在結構上與學生在報告前所設想的樣態類似，在句數、字數、段落數中並沒有超乎預期，學生認為這可能還是跟從小接受的作文訓練有關，也可以說明這樣的結構是大考中學生比較能夠負荷的，以四段式作文為主，撰寫約600字的作文。做這個主題的時候，本組的作文並沒有進一步討論關於作文的主題分類，在未來如果有機會盡一步分析，學生希望可以從類型、年份去將作文進行更細緻的分類，仔細去觀察佳作文章在基礎結構中有無差異，也許能夠更近一步了解國寫作文在我國考生心中的意義與大考中心選材選題上的變化。

二、 引用資料分析

經由分析過後可以發現，引用名人偉人與名言佳句的比例並沒有預期的高，大概分別占了52%與60%而已，而引用的資料也十分多元化，有被引用四次以上的名人偉人只有20個，相似之名言佳句只有八組。而由於我們只能取得佳作，並沒有非佳作之文章能做分析，因此沒有辦法做出「佳作普遍較常引用名人偉人事跡與名言佳句」等等之結論，若是之後有機會，希望能分別取得得分高與得分低之作文來做對照比較，也許能有其他有趣之發現。

三、 相似性分析

在做文章相似度分析遇到最大的阻礙就是我們蒐集資料的年份樣本數不均，有些較久以前的作文可以到10幾篇，而最新的作文只有2篇(因為現有資料只找到掃描檔)，所以做出來的最終結果為「只要該年的樣本數大於5，得到該年且相似度大於0.5的文章機率超過70%」。最後也應該會嘗試使用TF-IDF模型(因為每年都有相對應的題目)來試試看能否有更好的結果。

四、 語彙之詞頻詞性與詞意量分析

分析資料集的詞彙使用的量化數據後，可以讓我們推定了解近幾年被選為佳作文章的量化特性，像是詞性、詞頻、詞彙的詞意量與冷僻詞使用比例等。然而，可惜的是我們無法獲得足夠且可信的大考非佳作文章，否則可以進行兩母數的差異檢定分析，進一步得知被選為佳作是否有其他更顯著的關係。

柒、總結

綜上所述，我們可以從基礎結構中發現佳作普遍以600字、15句、4段式作文為主。而引用資料的部分，有五成左右的文章有引用名人、六成左右的文章有引用佳句，名人以中國傳統文人（例如：孔子）為多，而佳句上則是非常多元，只有少部分作文引用相同佳句。在文章相似性的部分，可以發現多數文章首尾關連度不高，而綜合所有文章進行相似性分析後，有一年的作文佳作相似性特別高。在詞意上，指考作文的動名詞比例高於學測，學生在動詞使用上有較頻繁的傾向；而非零詞意中平均每個用詞含有約8條詞意。最後，在詞頻上，佳作使用冷僻詞的比例約為三成六，一般詞語頻率以說話常見的動詞「讓、使、說」、名詞「人、心、生命」為多。從本次報告中有幸一窺我國升大學國寫作文的全貌，實屬難能可貴的經驗，也理解了非常多相關知識，學會使用課程外延伸的套件，也期待未來有機會能夠更近一步進行相關分析。

捌、附錄

一、相關內容

（一） 投影片

（二） 專案網站

二、組員分工

組員	工作內容
朱修平	- 基礎結構分析 - 相似度分析
盧德原	- 詞彙之詞頻、詞性與詞意量分析 - 呈現、入口網頁建置
楊舒晴	- 引用資料分析 - R網頁建置
陳宛瑩	- 圖表呈現 - 協助引用資料、基礎結構分析 - 簡報彙整製作

三、參考資料

https://github.com/ckiplab/ckiptagger