政見與「民意」之風向研究

—— 試析台大學生共識議題之轉變

第九組 匿名花栗鼠

組員 陳品睿 陳亦菱 許庭瑜 賴迎曦

2021-06-17

目錄

1.簡介

上個月，臉書的臺大交流版充斥對於一年一度學生會會長的討論，從宿舍、社團、到各種學生權益的爭取，候選人無不絞盡腦汁想出最能博得大家喜愛的政見。這樣的學生會長選舉已經行之有年，面對每學年度都會大量出現的政見，我們想藉此了解近十三年來學生們的想法，我們主要想探討的有：

1. 臺大學生在意的議題有哪些？
2. 是否一直存在著還沒有被解決的難題？
3. 這些難題經過了長時間的討論後，大家關注的面向有什麼改變？

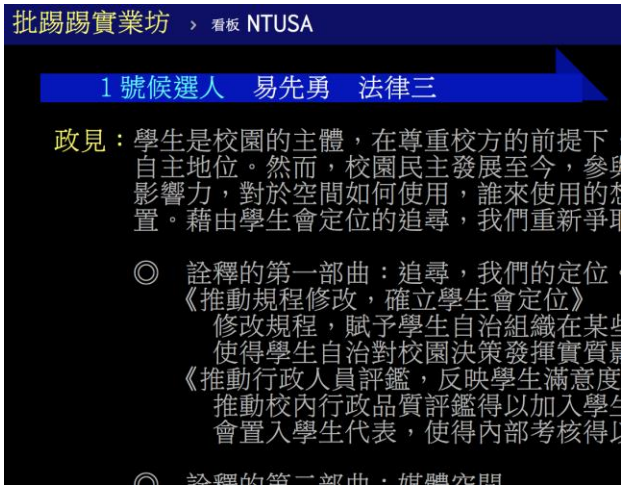
我們資料是從粉專爬文和 PTT 所查找的，總共有 33 筆資料，總共 67827 字。查找到資料之後，我們嘗試了三種分析方法，包括 jiebaR、CKIPtagger 和 Textrank，其中各自有各自的優點，也可以用不同的方式對我們所爬找的資料進行分析。

2. 方法

a. 資料取得

我們與臺大學生會選委會聯繫過後，對方表示未曾有人整理歷年來之選舉政見，因此我們的資料出自二處：

1. **粉專爬文**(104~109 學年度)：自臺大學生會選委會的粉專，找出每年度「選舉公報」中之學生會會長候選人的政見，將其儲存成 txt 文字檔。
2. **查找 PTT**(97~103 學年度)：這七年的政見皆有人整理放在 PTT 網站中，同樣儲存成 txt 文字檔。

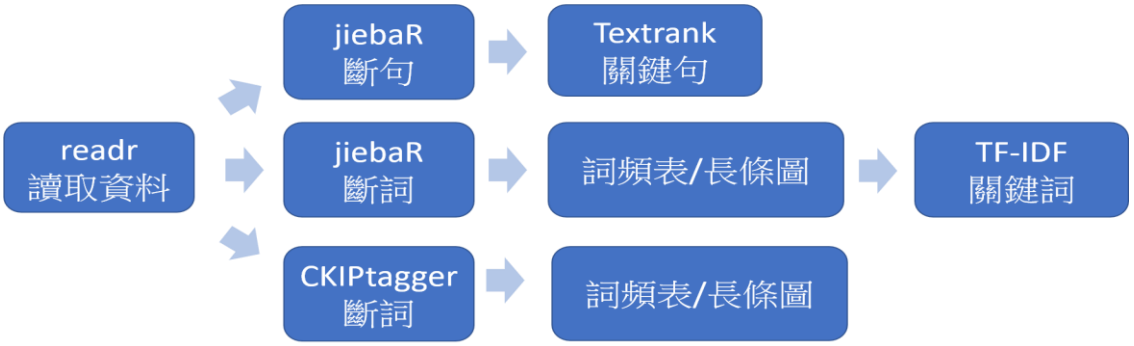


我們將蒐集的資料共同儲存在「science_data」資料夾中，以便分析。

由於 96 學年度以前的學生會會長候選人政見不曾有人整理，因此我們只蒐集 97 至 109 學年度共 13 年、33 筆資料，總字數為 67827 字。

b.原始碼運作說明

我們使用的套件包含 dplyr, jiebaR, tidytext, tidyverse, tibble, quanteda, highcharter, readr, stringr, reticulate, textrank 和 pacman。以 readr 讀取資料後，將資料整理成多個向量儲存在變數中。首先使用 tibble 和 quanteda 進行斷詞，加入自訂詞典及停用詞典，並且同時以 dplyr, tedytext, tidyverse, tibble, stringr 整理成表格。其中，在 CKIPtagger 部分使用了 reticulate 作進一步的斷詞。由於每次以 CKIPtagger 斷詞時，都要重新跑一次他們的 NLP 模型檔案、相當費時，所以我們以 jiebaR 的斷詞結果，透過 tf-idf 和 textrank 跑出關鍵詞和關鍵句。



以下步驟將分為四個部分：

1. jiebaR 斷詞

   我們透過函數 candidate，直接將原始文字檔進行資料清洗、增加/刪除詞彙和斷詞，輸出 data frame 詞頻表。並且以 highchart 繪製詞彙分佈圖。
   此一詞頻表將用於第三部分。

2. CKIPtagger 斷詞

   第二部分相對來說較獨立，由於我們發現以 jiebaR 處理後的詞頻表不夠精確，因此希望能夠透過呈現 CKIPtagger 斷詞後的結果，來比較兩者的差異。讀檔後進行初步斷詞，並取得詞性，接著篩選我們需要的詞性製作詞頻表與詞彙分佈圖。

3. TF-IDF 技巧

   將資料經過第一部分的處理之後，使用 jiebaR 的 worker()和 vector_keywords()讀取關鍵詞並製作詞頻表。
   原理：其會將經常出現的字給予較小的權重，不常出現的詞給予較大的權重。如此一來，如果不常出現的詞卻在這份資料中多次出現，則它很可能是這篇文章的重點，於是將它標記為關鍵詞。

4. Textrank 技巧

   透過函數 get_sentence_table 及 get_word_table，將原始文件檔斷句，其中會使用到 stringr 和 dplyr。接著透過 textrank 分析每個句子之間的相關性，並進行降冪排列，輸出一個 dataframe。

原理：其主要是將句子排名，排名的流程如下：

缺點：像是 “我今天不會回家” 跟 “我今天會回家” 這兩個句子的語意上是相反的，但用上面的公式算相似度會很高。一篇文章的摘要基本上就是重點，那重點應該會散佈再這篇文章中，所以藉由利用計算相似度將是重點的句子挑出來，但他是利用詞袋(bags of word)的方式，沒有考慮前後文的關係。

- CKIPtagger 進行流程

1. 使用 r-reticulate 將 ckiptagger 從 python 移植到 R

```
use_condaenv("r-reticulate")
ckip <- reticulate::import(module = "ckiptagger")
ws <- ckip$WS("C:/Users/yu/Documents/data")
pos <- ckip$POS("C:/Users/yu/Documents/data")
```

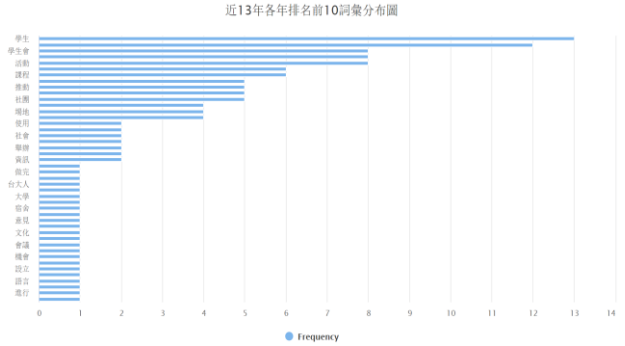
2. 寫了四個函數，只要輸入年度作為函數的 Input

t()	word()	pos_word()	result()
讀取與整理 txt 檔	斷詞	標註詞性	化成表格最後輸出詞彙分布長條圖

3. 撰寫 top10()函數，取出每年出現次數前 10 多的詞, 並繪製成表格：

y97	y98	y99	y100	y101	y102	y103	y104	y105	y106	y107	y108	y109
學生	學生	學生	社團	學生	學生	校園	校園	校園	學生	學生	學生會	學生
校園	空間	校園	學生	台大	同學	學生	活動	學生	臺大	空間	學生	同學
議題	校園	議題	校園	校園	學生會	校	學生	性別	校	推動	校園	社團
學生會	宿舍	空間	活動	做完	意見	場地	推動	制度	學生會	校	同學	校
台大	性別	校	空間	學生會	校園	社會	社團	校	校園	活動	語言	校園
台大人	詮釋	建立	性別	建立	臺大	空間	舉辦	國際	舉辦	同學	校	學生會
同學	推動	會議	學生會	會長	領域	使用	課程	推動	課程	爭取	活動	資訊
媒體	曲	社會	資源	社團	場地	性別	台大	爭取	爭取	課程	社團	課程
政見	爭取	設立	場地	資訊	機會	活動	場地	課程	推動	進行	空間	大學
校	使用	爭取	想像	人	活動	深度	學生會	台大	活動	文化	課程	活動

4. 將 97~109 年各年前 10 詞彙出現的年數進行統計，最後繪製成長條圖：



扣除學校、學生、學生會等相似詞彙，這 13 年間各年十大高頻詞彙出現 4 次以上的名詞可分為四類，分別為活動與社團、空間與場地、課程、性別；動詞則包含推動、爭取。

- TF-IDF 進行流程
  1. 使用 jiebaR 進行初步斷詞
  2. 以 tf_idf()函數找出前 20 關鍵詞:

```
tf_idf <- function(x){
  key = worker("keywords", topn = 20)
  vector_keywords(seg[x], key)
}
```

  3. 結果

```
#tf_idf
tf_idf(title[8])

#> 93.9136  58.696 46.9568 35.2176 35.2176 35.2176 23.4784 23.4784 23.4784 23.4784
#>  "校園"  "學生"  "校內"  "Open"  "開放"  "社會"  "議題"  "經費"  "空間"  "系統"
#> 23.4784 23.4784 23.4784 23.4784 23.4784 23.4784 23.4784 14.7838 14.1789 13.9151
#>  "資訊"  "監督"  "場地"  "活動"  "關心"  "性別"    "從"  "深化"  "透明"  "深度"
```

TF-IDF 將詞的重要程度量化後，相較我們一開始的方法，更能找到精確的關鍵詞。

- Textrank 進行流程
  1. 以 jiebaR 斷句
  2. 以 sentences()函數找出關鍵詞

```
textrank_sentences(data = st,terminology = wt) %>%
   summary(n = 3)   #n 代表要 top 多少的關鍵句子
```

在 summary(n=3)的地方，可以決定關鍵句的句數。我們先以 n=3 進行排名，然而，開始出現與一開始詞頻表相同的問題。有一些符號及非議題的句子出現，因此，操作上我們會多取幾年，且由於此方法是按照句子的重要程度排名，因此我們可以略過那些不精確的地方，按順位取得比較重要的三句話。後面再從由取得的每年三句話，用人工的方式，定義出句子中的議題，並將議題分成八大類，取這幾年間有重覆出現過的議題進行分析。詳細方法，在分析結果中說明。

```
#Textrank
sentences(11)

#> Warning in stri_split_regex(string, pattern, n = n, simplify = simplify, :
#> argument is not an atomic vector; coercing

#> [1] "3.進行成立各課程討論版,讓願意幫助有需要之後生課業之前輩能有平台給予協助"
#> [2] "2.舉辦校系系茶會,增進不同校系領域學子間學術交流"
#> [3] "4.鼓勵全校師生提案,選出具迫切性及可行性之提案,聽取全校師生需求進行可能協助"
```

上圖為理想結果：sentences(11)

```
#Textrank
sentences(13)

#> Warning in stri_split_regex(string, pattern, n = n, simplify = simplify, :
#> argument is not an atomic vector; coercing

#> [1] "2."                                    "請問各位,要向下沉淪,還是向上提升?"
#> [3] "經濟學系四年級"
```

上圖為不理想結果：sentences(13)

3. 結果

a. 試析台大學生共識議題之轉變

1. 臺大學生在意的議題有哪些？

我們根據 Textrank 的關鍵句，從每年政見中找出被大家關注的三個議題，再將這些議題分類歸納。進而發現，以下是十年中有被重複提過的議題，而我們認為這些最可能是台大學生在意的議題--「**參選預算、文化與藝術、性別友善、社會回饋、選舉制度改善、技術交流(跨領域、跨國際)、空間分配、學生自治**」

參選預算	文化與藝術	性別友善	社會回饋
-提案討論預算(109) -預算(106)	-文史資產(109) -文化季(100) -台大裝置藝術節(98)	-性別平權(108) -性別友善校園(104) -性別議題(103) -性別空間(98)	-社會議題調查(106) -校園社區沒距離、民主牆(100) -學生服務隊、提出議題給其他學校(99)
選舉制度改善	**技術交流(跨領域、跨國際)**	**空間分配**	**學生自治**
-成立友善台大 NTU's (105) -學生會官方通訊軟體帳號(105) -電子投票公投(102) -議題懶人包(97)	-跟世界大師接觸的活動(106) -跨領域交流、國際徵才展(102)	-煮食空間、閒置空間(109) -監督校內工程(103) -空間分配(99) -性別空間(98)	-自治組織(101) -公共媒體(98) -學生自治(97)

2. 是否一直存在著還沒有被解決的難題？

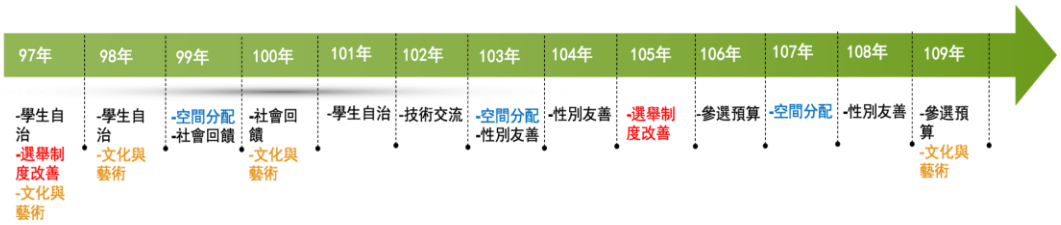
我們根據以上議題，我們在將他們的年份列出，發現其中「文化與藝術、選舉制度改善、空間分配」，相隔五年以上又被提出。由於我們這次所研究的資料為 97~109 年(約有 13 年)，因此五年將近一半，而認為這些可能是沒有被解決

的難題。然而，不排除此議題本身並非一個需被解決的難題(如:文化與藝術)，而是每個年度大家都會關注的議題，所以一再被提出。

文化與藝術：提出年度 109, 100, 98, 97，相隔 9 年再度被提出

選舉制度改善：提出年度 105, 97，相隔 8 年再度被提出

空間分配：提出年度 107, 103, 99，每隔 4~5 年會被提出

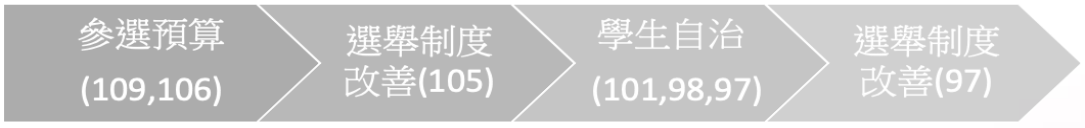


值得關注的是，在 104 年時，其實學生會長有進行補選，而 105 年中**選舉制度改善**為分析該年的重點議題，我們懷疑兩者之間應該存在什麼關係。

3. **這些難題經過了長時間的討論後，大家關注的面向有沒有改變？**

我們再將上面的議題，重新按時間順序排列後，會發現從學生會長提出的政見中，議題有這樣的演變關係。

我們想針對「**參選預算、學生自治、選舉制度改善**」的這三個議題進行討論，其實這些議題都跟學生會有關，雖然分類上將他分為三類，但或許我們可以把它們共同視作「學校自治活動」議題來看。並推測這些議題的演進，可能跟學生會從一開始成立、到越來越完備，同學們對於爭取學生自治權利的關注議題已經產生改變有關。



109	議題	Textrank
1	提案討論預算	2. 舉辦參與式預算：參與式預算透過公民提案、在會議上進行討論...
2	關注體育學生	2. 對體育專長學生、體育學程的關注：從許多國外大學到政大今年的例子可證明，一個學校運動人才與組織的培育，能帶動整個學校運動的風氣...
3	文史資產	文史資產」曾提過的以住代護政見也是一個可嘗試的方法。
108	議題	Textrank
1	性別平權	肆、捍衛性別平權:看見差異與不平等,真正落實友善於校園生活
2	環境議題	學生會將與環境議題相關的社團進行合作,推動工作坊、新生書院生態導覽等
3	本土語言	二、捍衛本土語言在公眾場合的使用
107	議題	Textrank
1	煮食空間	針對現有空間進行檢討報告，並推動校園自煮運動...
2	閒置空間	重新檢視校內閒置空間，針對同學練習表演之時段與地點進行調查...
3	換宿資訊平台	（三）學生換宿資訊平台的常態化
106	議題	Textrank
1	社會議題調查	11. 針對社會議題進行校內民意調查，提供台灣社會進步的新方向。
2	跟世界大師接觸的活動	我們將透過學生會的力量，舉辦更代表性、傳承性的活動…
3	預算	參與式預算，學生政策提案

105	議題	Textrank
1	成立友善台大NTU’s	是去年會長補選時我提出的政見,為了實踐提出的理念,我們成立了「友善台大NTU’s Heart」粉絲專頁,作為獨立的校園政策智庫與倡議團體
2	學生會官方通訊軟體帳號	(二)、臺大1999:設立「學生會官方通訊軟體帳號」,雙向接收、推播校內重大議題,並即時回應。
3	飲食選擇	五、飲食選擇友善
104	議題	Textrank
1	學校事務透明	➤學生會事務透明化,加強校務與會務的資訊轉譯與揭露...
2	教務改革	➤推動教務改革:教學意見調查結果公開、停修改為報備制...
3	性別友善	三、性別友善校園:女廁及性別友善廁所設置衛生棉/棉條販賣機...
103	議題	Textrank
1	監督校內工程	◎監督校內工程，參與建築空間設計
2	場地借用系統	◎改善場地借用系統，學生活動更順利
3	性別議題	◎深化性別議題推動，從軟體硬體強化性別友善校園
102	議題	Textrank
1	跨領域交流	跨領域交流活動
2	電子投票公投	→透過使用電子投票讓公投廣泛使用，讓每個議題大家都有參與的機會...
3	國際徵才展	國際徵才展

101	議題	Textrank
1	自治組織	◎轉角遇到學生會，自治組織在你身邊
2	廁所有衛生紙	3. 學校的廁所裡面要有衛生紙
3	社團程序友善化	◎檢討社團相關制度，推動程序友善化
100	議題	Textrank
1	校園社區沒距離	4. 校園社區零距離
2	民主牆	1. 設置臺大民主牆
3	文化季	3. 多元文化季
99	議題	Textrank
1	學生服務隊	(12)設立學生會服務專線與服務隊，解決校內需求並提供社區服務。
2	提出議題給其他學校	(v)由學生主動提出其他議題給學校
3	空間分配	(i)系所空間分配
98	議題	Textrank
1	公共媒體	《建置公共媒體，學生自主審查》
2	台大裝置藝術節	《台大裝置藝術節》
3	性別空間	借由討論性別空間，一方面引入審議式民主論壇...
97	議題	Textrank
1	藝術風氣	8. 邀請藝術家、作家駐校，打造校園藝術風氣
2	議題懶人包	一、校園議題懶人包，降低關心議題成本...
3	學生自治	《學生自主，學生治校》

b. 斷詞系統與關鍵詞技術之比較

除了使用 Textrank 技術找出關鍵句來幫助分析臺大學生關注議題的轉變，我們也希望透過呈現以 jiebaR, CKIPtagger, tf-idf 三種方式跑出來的詞頻表，對於這幾種方法做一個比較。

jiebaR 的概念最簡單，使用者透過自訂詞典和停用字典，將文章斷詞，詞頻表呈現出來的就會是篩選過後最直接的統計結果。即便我們盡量增補了兩個字典，透過一再地修正讓 jiebaR 跑出來的斷詞結果越來越精確，仍然有許多漏網之魚，這也是我們選擇額外用 CKIPtagger 斷詞的原因。如 109 和 108 年的結果可以看到，由於我們未將「在」這個介系詞加入停用字典中，因此它便出現在詞頻表上累積數量最多的地方，容易干擾分析結果。

109	jiebaR	CKIPtagger	tf-idf	108	jiebaR	CKIPtagger	tf-idf
1	在	學生	與	1	學生會	學生會	與
2	學生	同學	學生	2	在	學生	學生會
3	我們	社團	同學	3	學生	校園	社團

因此，我們決定使用第二種方法，以 CKIPtagger 重新將資料斷詞。CKIPtagger 除了上述提到 jiebaR 的功能外，也可以標記詞性和辨識各領域的專有名詞，並且透過每次讀取 NLP 模型檔案，使得斷詞更精確。我們利用了 CKIPtagger 詞性標記的技術，只篩選出 Na(普通名詞), Nb(專有名詞), Nc(地方詞), 和 VC(動作及物動詞)來製作詞頻表和詞彙分佈圖，因此 CKIPtagger 的詞頻表和其他兩種方法相比，基本上都是最精確的。不過如同前面曾提及的，做每一個斷詞步驟時，都要跑一次 CKIPtagger 的 NLP 模型，十分耗時，因此我們實在無法將 CKIPtagger 跑出來的詞頻表拿去以 tf-idf 或 textrank 做進一步的分析。如 107～103 年的結果可以看到，CKIPtagger 的斷詞結果和其他兩種相比非常精確，由於進行了詞性標記，我們得到的幾乎都是能夠直接拿來分析的資料。

107	jiebaR	CKIPtagger	tf-idf	106	jiebaR	CKIPtagger	tf-idf
1	校	學生	接軌	1	學生	學生	臺大
2	學生	空間	與	2	臺大	臺大	與
3	空間	推動	推動	3	內	校	舉辦

105	jiebaR	CKIPtagger	tf-idf	104	jiebaR	CKIPtagger	tf-idf
1	校園	校園	學生	1	活動	校園	活動
2	學生	學生	校園	2	校園	活動	課程
3	友善	性別	與	3	學生	學生	與

103	jiebaR	CKIPtagger	tf-idf
1	校園	校園	校園
2	學生	學生	學生
3	內	校	校內

CKIPtagger 再怎麼神通廣大，畢竟它就是個好用的斷詞系統，能夠將一份文件有效地斷詞後，和 jiebaR 一樣將詞彙出現的次數進行降冪排列。然而有許多文件中的重要詞彙，或許不是那麼常出現，又或者一些主詞、常用詞彙雖不是該文件重點，卻常常出現在詞頻表的首幾欄。所以我們決定使用第三種方法 tf-idf。同樣屬於 jiebaR，tf-idf 顧名思義是以詞頻加上逆向文件頻率，將斷詞後的資料透過相關性做進一步的分析。雖然說因為與學校有關的文件本身具有的詞彙和其他文件相比之下就比較少見，它仍無法完全看出政見中真正的關鍵詞，但 tf-idf 的結果仍然好過於 jiebaR。由於我們是以 jiebaR 的斷詞結果進行 tf-idf 的分析，可以從 102～99 年的結果看到還是有一些副詞、介系詞穿插其中，干擾分析。我們認為比較理想的方式是克服 CKIPtagger 需要耗費大量時間跑的問題後，以 CKIPtagger 的斷詞結果進行 tf-idf 分析，應該會最準確。

102	jiebaR	CKIPtagger	tf-idf	101	jiebaR	CKIPtagger	tf-idf
1	讓	學生	學生	1	更	學生	更
2	有	同學	讓	2	多	臺大	臺大
3	學生	學生會	我們	3	學生	校園	學生
100	jiebaR	CKIPtagger	tf-idf	99	jiebaR	CKIPtagger	tf-idf
1	社團	社團	與	1	之	學生	與
2	我們	學生	社團	2	學生	校園	學生
3	活動	校園	學生	3	校園	議題	校園
98	jiebaR	CKIPtagger	tf-idf	97	jiebaR	CKIPtagger	tf-idf
1	學生	學生	學生	1	學生	學生	學生
2	空間	空間	空間	2	校園	校園	校園
3	部	校園	校園	3	議題	議題	議題

討論與貢獻

a. 研究限制與未來展望

1. 儘管 ckiptagger 可以將大部分無意義的字詞刪去，然而詞頻表整理出的高頻率詞大多是學生會、校園、學生、臺大等指涉範圍較廣的詞，無法有效率的看出學生實際關心的議題
2. 關鍵句可以看出學生關注的某些議題，但由於政見是比較偏重點式、條列式的文本，不會大篇幅闡述某項主題，使得關鍵句無法發揮最佳效果
3. 以我們目前寫的程式碼，還無法得出詞彙與詞彙之間的關聯，譬如無法將學生、校園、學生會，統整成一個主題造成分析上的困難與沒有效率，因為只能用人工的方式分類，未來也許能夠利用詞向量的方式歸納出各個主題。
4. 由於關鍵句是按照重要程度排名的，因此我們應該用加權的方式，給每個議題不同的重要程度，再進行統計，結果會更精確
5. 由於我們在議題的分類上是用「人工」的方式，可能比較主觀，或許我們可以參考之前「文本與詞彙的向量表徵」所學的，用向量重新分類，會比較客觀

b. 組員分工表

姓名	工作內容
陳品睿	PPT 排版、資料整理、code、書面資料排版、書面資料總整理
陳亦菱	Textrank 斷詞系統、結論分析、Jieba 斷詞系統、書面報告結論一、網頁架設
許庭瑜	CKIPtagger 斷詞系統詞頻歷年分析比較、研究限制、書面資料校稿與排版、Github readme.md 與資料上傳
賴迎曦	Jieba 斷詞系統詞頻歷年分析比較、各分析方法總整理及比較、網頁架設