政見與「民意」之風向研究

—由歷年台大學生會長候選人之政見，試析學生共識議題之轉變

ㄨ匿名花栗鼠ㄨ
陳品睿、陳亦菱、許庭瑜、賴迎曦

目錄

01 研究動機(探討議題)

02 資料來源

03 分析方法

04 結論(比較三種方法)

研究動機

臺大學生在意的議題

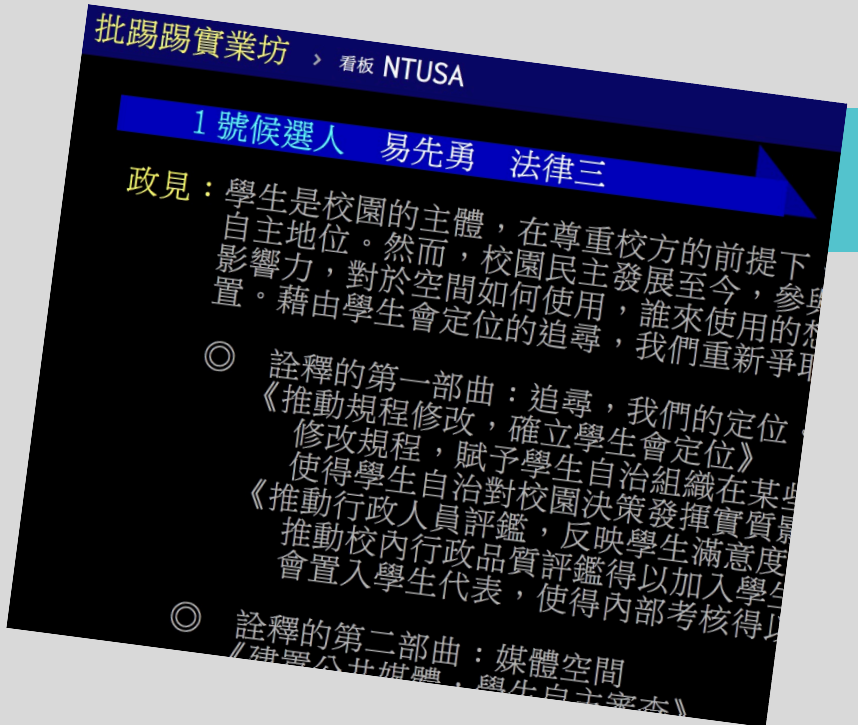
+

=

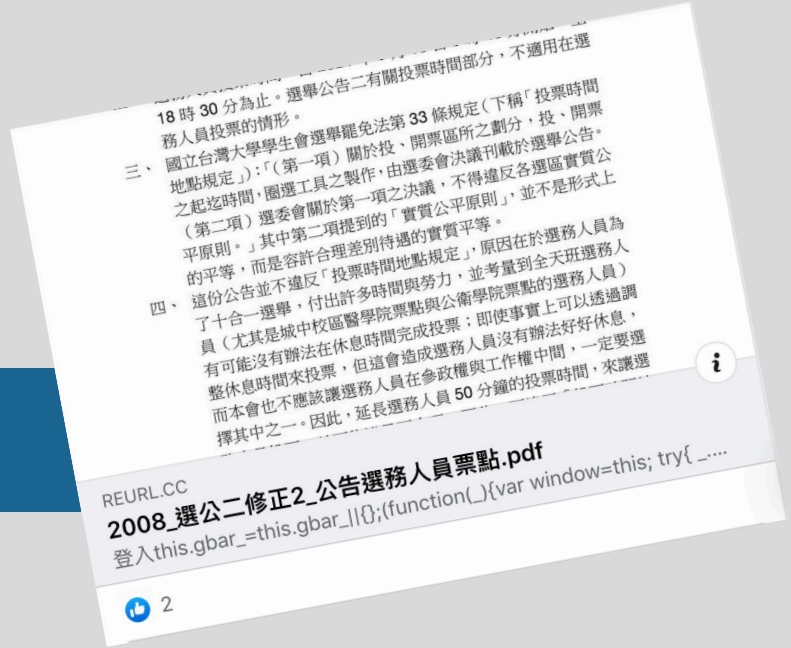
還沒有被解決的難題？
經過長時間討論，關注的面向有無改變？

資料來源

PTT



FB粉絲專頁



33筆政見資料

97～109共13年

總字數67827字

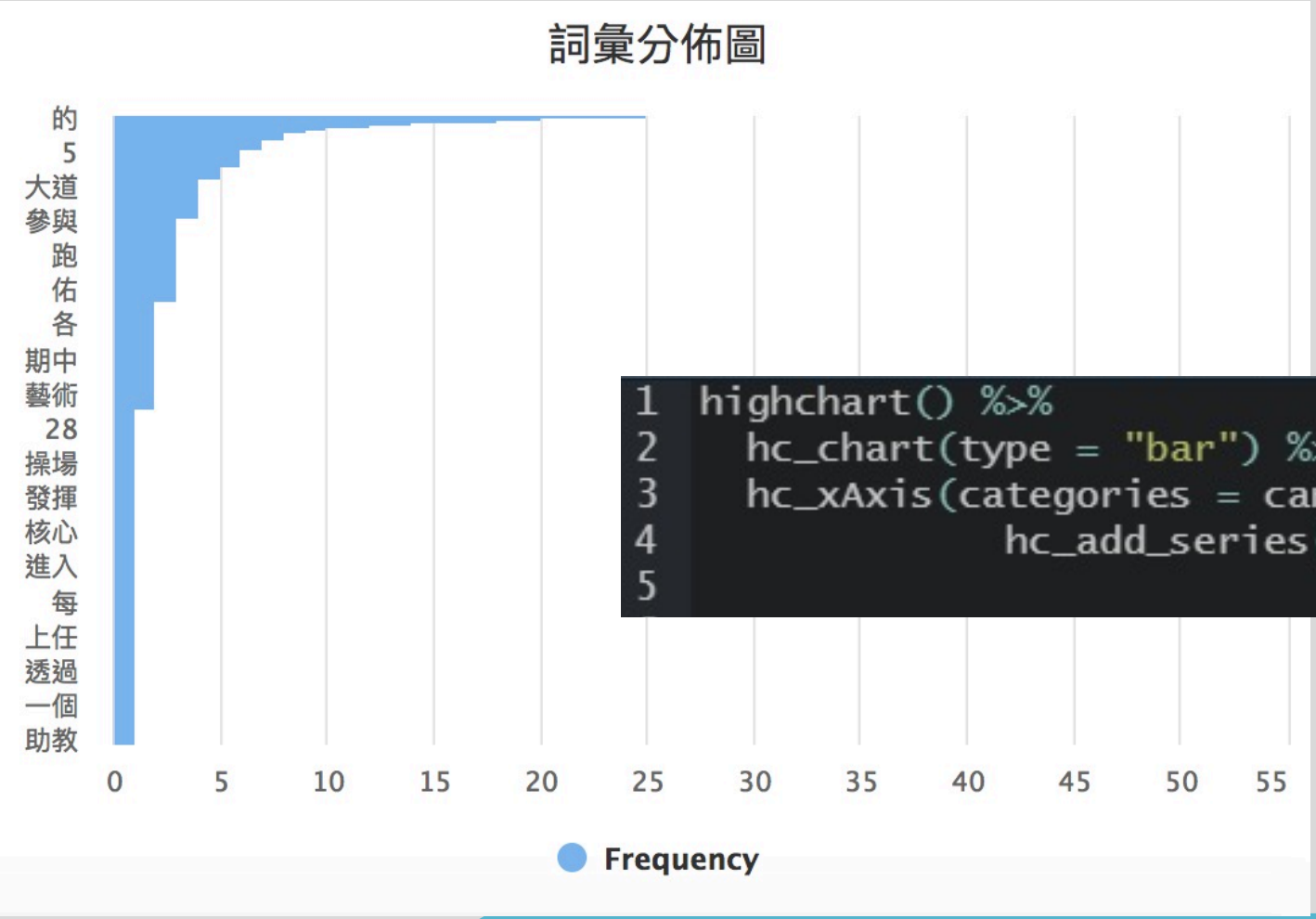
分析方法

斷詞、詞頻表

	word	n
1	臺大	12
2	的	10
3	將	8
4	內	8
5	校	8
6	校園	8

```r
candidate <- function(x) {
  docs_segged <- vector("character", length = length(x))
  segged <- paste0(segment(x, seg), collapse = "\u3000")
  docs_df <- tibble::tibble(
    doc_id = seq_along(docs_segged),
    content = segged
  )
  knitr::kable(docs_df, align = "c")
  frequency_df <- docs_df %>%
    unnest_tokens(output = "word", input = "content")%>%
    group_by(word) %>%
    summarise(n = n()) %>%
    top_n(3,n) %>%
    arrange(desc(n))
  View(frequency_df)
}
```

斷詞、詞頻表




```r
highchart() %>%
  hc_chart(type = "bar") %>%
  hc_xAxis(categories = candidate(x) %>%
             hc_add_series(name = "Frequency", data = candidate(x) %>%
                           hc_title(text = "詞彙分佈圖")
```

斷詞、詞頻表

CKIPtagger
原理

CKIPtagger 原理

- 中文處理開源套件
- 辨識11類專有名詞
- 繁體中文表現佳
- 缺點

unlist_words	unlist_pos
1	Neu
.	PERIODCATEGORY
訂定	VC
完善	VH
補考	VC
機制	Na
部分	Neqa
學生	Na
可能	D
因	Cbb

```r
# 製作初步斷詞及其對應詞性的dataframe

df <- tibble(unlist_words = unlist(words),
             unlist_pos = unlist(pos_words))
```

CKIPtagger

unlist_words	unlist_pos
訂定	VC
補考	VC
機制	Na
學生	Na
狀況	Na
成績	Na
學期	Na
成績	Na
科目	Na
擋修	VC

```r
# 抽取詞性為普通名詞(Na)、專有名詞(Nb)、地方詞(Nc)、
# 動作及物動詞(VC)的詞彙

df2 <- filter(df, (unlist_pos == "Na" )|
                  (unlist_pos == "Nb" )|
                  (unlist_pos ==  "Nc")|
                  (unlist_pos ==  "VC"))
```

CKIPtagger

unlist_words	n
增設	8
空間	8
宿舍	6
停車場	4
學生	4
教學館	4
校	4
自行車	4
人	3
台大	3
同學	3
大便	3
娛樂	3

```r
# 製作詞頻表

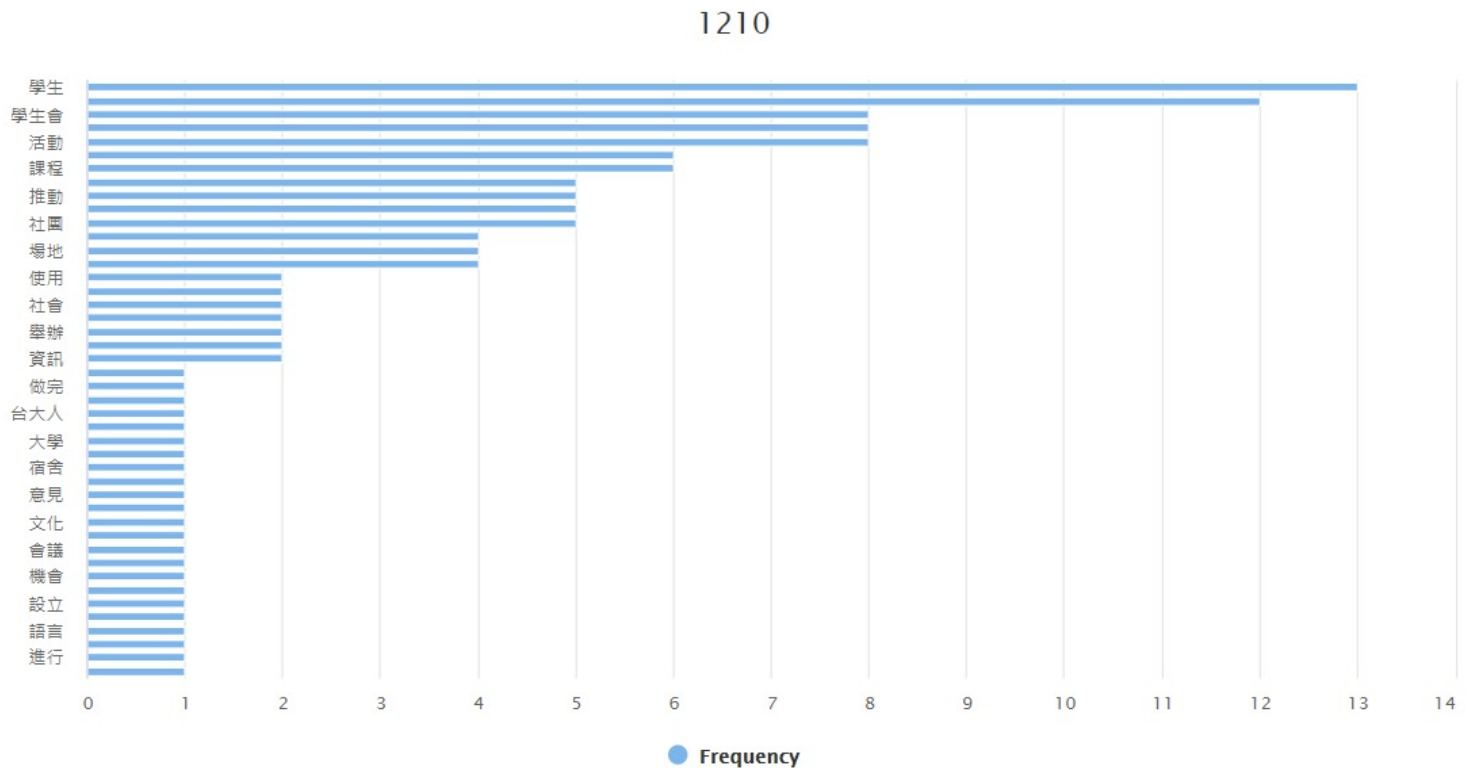
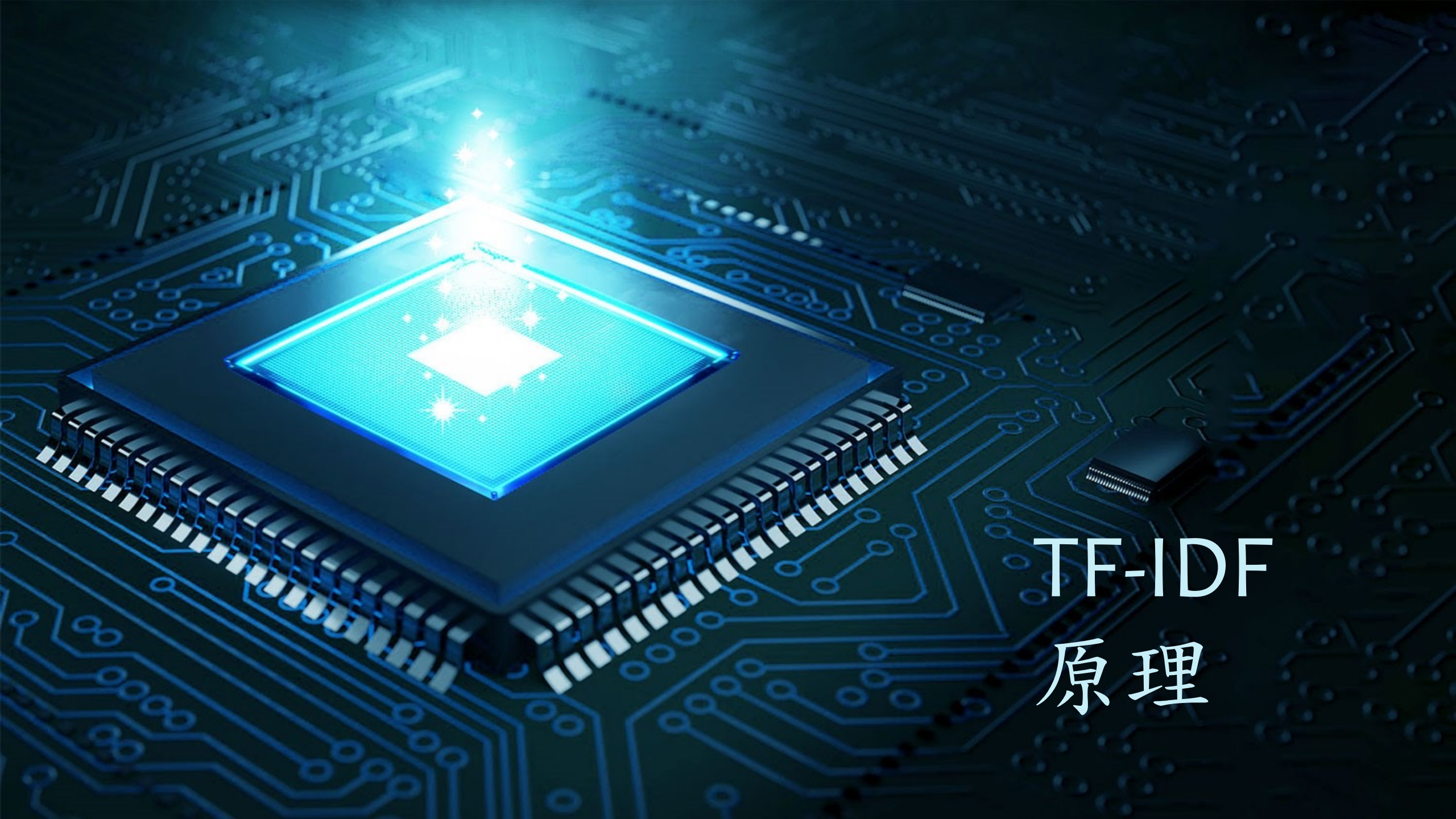
df3 <- df2 %>%
  group_by(unlist_words) %>%
  summarise(n = n()) %>%
  arrange(desc(n))
```

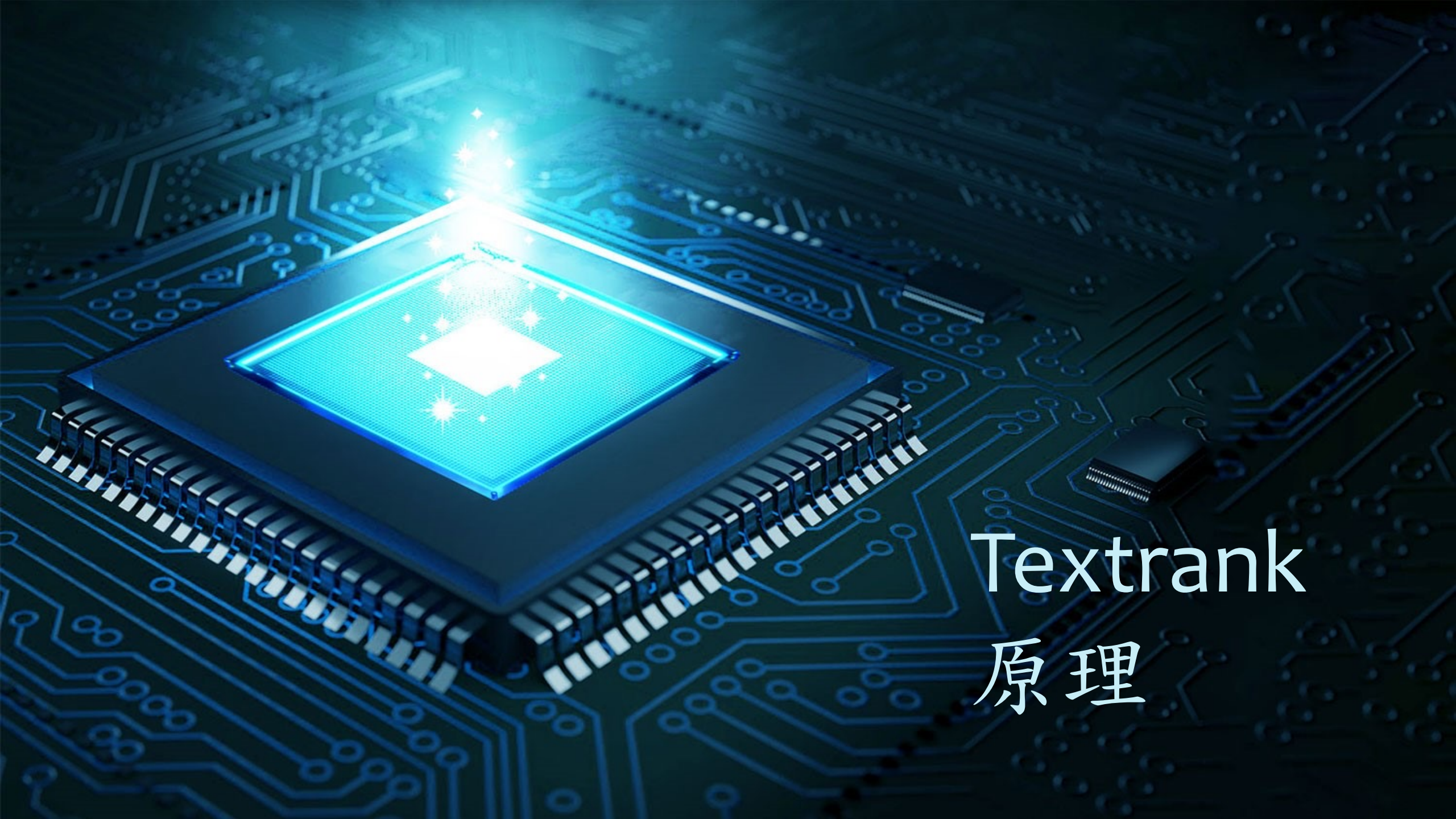
CKIPtagger

df	df2	df3
1	訂定	增設
.	補考	空間
訂定	機制	宿舍
完善	學生	停車場
補考	狀況	學生
機制	成績	教學館
部分	學期	校
學生	成績	自行車
可能	科目	人
因	擋修	台大

```
# 比較表1、表2與表3前10列的詞:
```{r}
df4 <- tibble(df = head(df$unlist_words, 10L),
              df2 = head(df2$unlist_words, 10L),
              df3 = head(df3$unlist_words, 10L))
```

CKIPtagger
```

對比各年出現次數最多的前10詞彙

```
top10 <- function(x){
    df <- tibble(unlist_words = unlist(word(x)), unlist_pos = unlist(pos_word(x)))
    df2 <- filter(df, (unlist_pos == "Na" )| (unlist_pos == "Nb" )| (unlist_pos ==  "Nc")| (unlist_pos ==  "VC"))
    df3 <- df2 %>%
        group_by(unlist_words) %>%
        summarise(n = n()) %>%
        arrange(desc(n))
    head(df3$unlist_words, 10L)
}
y97 <- top10(97)
y98 <- top10(98)
y99 <- top10(99)
y100 <- top10(100)
y101 <- top10(101)
y102 <- top10(102)
y103 <- top10(103)
y104 <- top10(104)
y105 <- top10(105)
y106 <- top10(106)
y107 <- top10(107)
y108 <- top10(108)
y109 <- top10(109)
compare_by_year_df <- tibble(y97 = y97, y98 = y98, y99 = y99, y100 = y100, y101 = y101, y102 = y102, y103 = y103, y104 = y10
4, y105 = y105, y106 = y106, y107 = y107, y108 = y108, y109 = y109)

knitr::include_graphics("C:/Users/yu/OneDrive/Desktop/R Lab Note/png/top10_compare_by_year_df.png")
```

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

CKIPtagger

97~109年各年前10詞彙出現年數統計

表格

```
all_top10 <- c(y97, y98, y99, y100, y101, y102, y103, y104, y105, y106, y107, y108, y109)
top10_freq_by_year_df <- tibble(word = all_top10) %>%
    group_by(word) %>%
    summarise(n = n()) %>%
    arrange(desc(n))
knitr::include_graphics("C:/Users/yu/OneDrive/Desktop/R Lab Note/png/top10_freq_by_year_df.png")
```

	word	n
1	學生	13
2	校園	12
3	學生會	8
4	校	8
5	活動	8
6	空間	6
7	課程	6
8	同學	5
9	推動	5
10	爭取	5
11	社團	5
12	台大	4
13	場地	4
14	性別	4

CKIPtagger

長條圖

```
top10_freq_by_year_graph <- highchart() %>%
    hc_chart(type = "bar") %>%
    hc_xAxis(categories = top10_freq_by_year_df$word) %>%
    hc_add_series(name = "Frequency", data = top10_freq_by_year_df$n) %>%
    hc_title(text = "近12年各年排名前10詞彙分布圖")
top10_freq_by_year_graph
```



CKIPtagger



TF-IDF
原理

TF-IDF 原理

- tf=term frequency
  idf=inverse document frequency
- 計算詞頻，給予權重
- 缺點

TF-IDF

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



Textrank 原理

Textrank 原理

- 重要的詞或句子
- 句子排名
- 缺點

```r
get_sentence_table = function(string){
  string %>% str_split(pattern = "[:space:]+") %>% unlist %>% as_tibble() %>%
    transmute(sentence_id = 1:n(),sentence = value)
}
seg = worker(user = "user_dict.txt",stop_word ="/content/science_data/stop_word.txt") #在外部構造一個jieba分詞器
get_word_table = function(string){
  string %>% str_split(pattern = "[:space:]+") %>% unlist %>% as_tibble() %>%
    transmute(sentence_id = 1:n(),sentence = value) %>%
    mutate(words = map(sentence,segment,jieba = seg)) %>% select(-sentence) %>% unnest()
}
test_text = title[x]
st = test_text %>% get_sentence_table
wt = st%>% get_word_table

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

Textrank

分析方法整理

Jieba

使用者自訂字典

CKIPtagger

使用者自訂字典+
辨識專有名詞+
詞性標記

TF-IDF

詞頻+
逆向文件頻率

Textrank

網絡呈現+
交集字詞數量+
相關性

Limitations

結論

臺大學生
在意的議題
有哪些？

是否一直存在著
還沒有被解決
的難題？

難題經長時間討論，
關注面向
有無改變？

臺大學生在意的議題有哪些

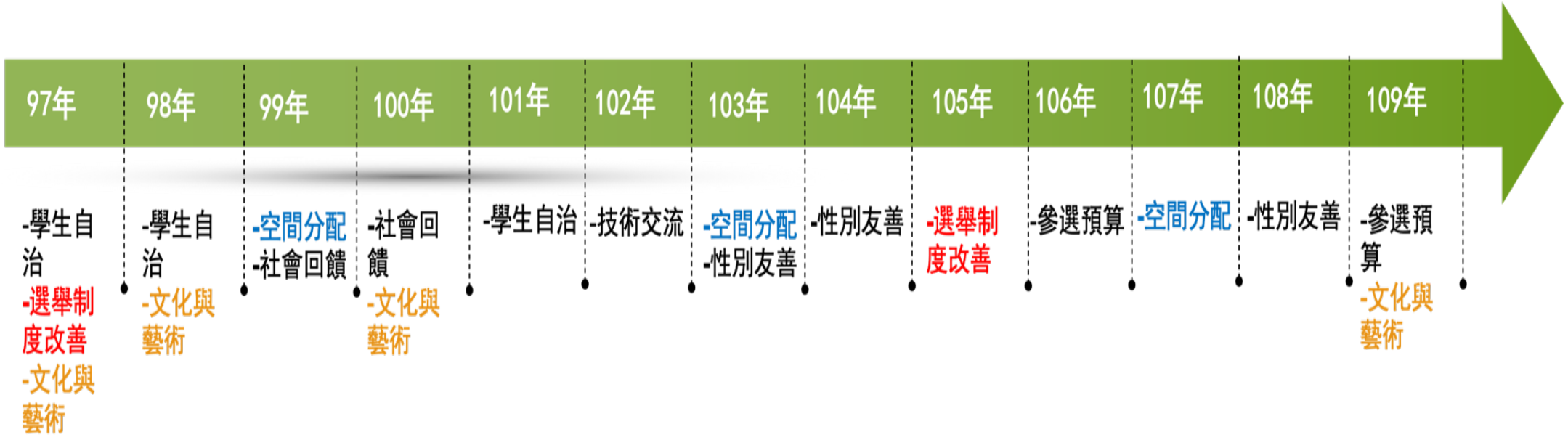
參選預算	文化與藝術	性別友善	社會回饋
-提案討論預算(109) -預算(106)	-文史資產(109) -文化季(100) -台大裝置藝術節(98)	-性別平權(108) -性別友善校園(104) -性別議題(103) -性別空間(98)	-社會議題調查(106) -校園社區沒距離、民主牆(100) -學生服務隊、提出議題給其他學校(99)
選舉制度改善	技術交流(跨領域、跨國際)	空間分配	學生自治
-成立友善台大NTU’s(105) -學生會官方通訊軟體帳號(105) -電子投票公投(102) -議題懶人包(97)	-跟世界大師接觸的活動(106) -跨領域交流、國際徵才展(102)	-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

是否一直存在
還沒有被解決的難題

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



難題經長時間討論，關注面向有無改變

Thanks for listening