## 大數據程式實作

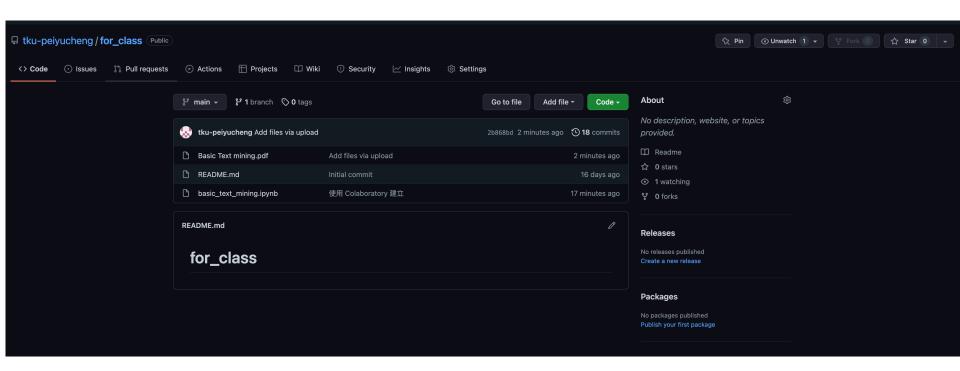
Big Data Programming



Week 5
Basic Text mining

Pei-Yu Cheng 鄭培宇 Assistant Professor Dept. of Information Management Tamkang University

#### Slide and Code



https://reurl.cc/MbY1VK

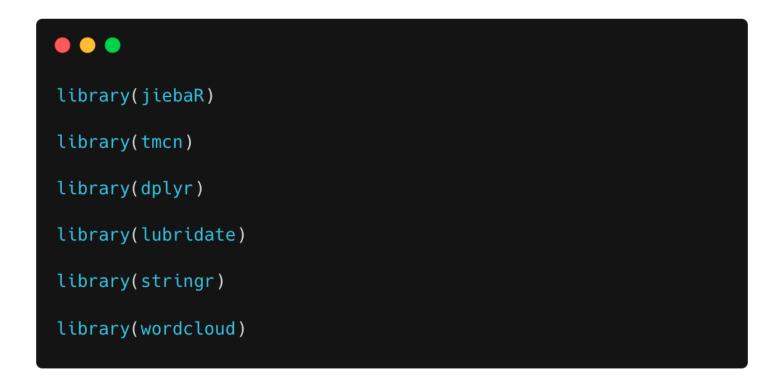
https://reurl.cc/X4ODg3



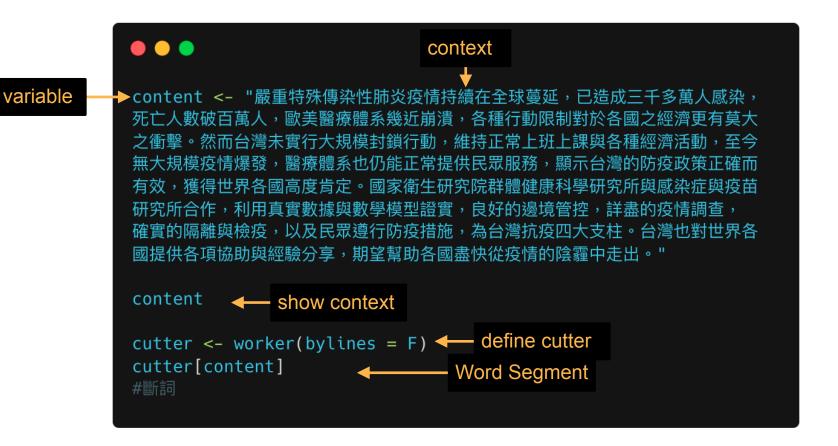
### install packages

```
install.packages("rJava")
install.packages("jiebaR")
install.packages("tmcn")
install.packages("dplyr")
install.packages("lubridate")
install.packages("stringr")
install.packages("wordcloud")
install.packages("RColorBrewer")
```

## import packages



#### Add a variable to add the context



#### **Word Segmentation**



□→ '嚴重'・'特殊'・'傳染性'・'肺炎'・'疫情'・'持續'・'在'・'全球'・'蔓延'・'已'・'造成'・'三千多萬'・'人'・'感染'・'死亡'・'人數'・'破'・'百萬'・'人'・'歐美'・'醫療'・'體系'・'幾近'・'崩潰'・'各種'・'行動'・'限制'・'對於'・'各國'・'之'・'經濟'・'更'・'有'・'莫大'・'之'・'衝擊'・'然而'・'台灣'・'未'・'實行'・'大規模'・'封鎖'・'行動'・'維持'・'正常'・'上課'・'與'・'各種'・'經濟'・'活動'・'至今'・'無'・'大規模'・'疫情'・'爆發'・'醫療'・'體系'・'也'・'仍'・'能'・'正常'・'提供'・'民眾'・'經濟'・'活動'・'至今'・'無'・'大規模'・'疫情'・'爆發'・'醫療'・'體系'・'也'・'仍'・'能'・'正常'・'提供'・'民眾'・'服務'・'顯示'・'台灣'・'的'・'防疫'・'政策'・'正確'・'而'・'有效'・'獲得'・'世界'・'各國'・'高度肯定'・'國家衛生研究院'・'群體'・'健康'・'科學'・'研究所'・'與'・'感染'・'症'・'與'・'疫苗'・'研究所'・'合作'・'利用'・'真實'・'數據'・'與'・'數學模型'・'證實'・'良好'・'的'・'邊境'・'管控'・'詳盡'・'的'・'疫情'・'調查'・'確實'・'的'・'隔離'・'與'・'檢疫'・'以及'・'民眾'・'遵行'・'防疫'・'措施'・'為'・'台灣'・'抗疫'・'四大'・'支柱'・'台灣'・'也'・'世界'・'各國'・'提供'・'各項'・'協助'・'與'・'經驗'・'分享'・'期望'・'幫助'・'各國'・'基出'

#### proper nouns - disconnected

'醫療'、'體系'→'醫療體系'

### add new words to corpus

```
new_words <- c("傳染性肺炎","醫療體系","防疫措施") 		─ Add proper nouns
for (i in 1:length(new_words)) {
 new_user_word(cutter, new_words[i])
# 以迴圈方式,一次加入一個詞
content <- str remove all(content, "[0-9a-zA-Z]+?")
#去掉數字、英文(如果有需要)
```

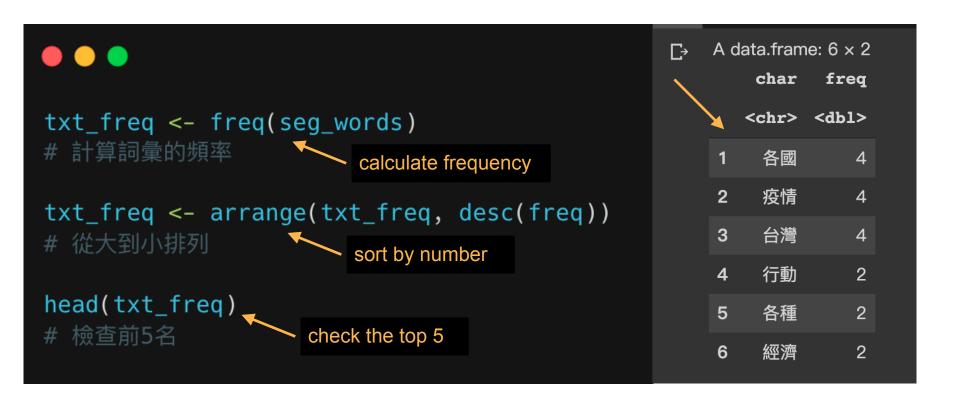
#### new words and stop words

```
writeLines(new_words, "new_words.txt") \rightarrow Add the new words
stop_words <- c("在","之","更","個","未","而","也","與","的")
writeLines(stop_words, "stop_words.txt")
# 設定停止詞
                      Add the stop words
cutter <- worker(user = "new words.txt", stop word =
"stop words.txt", bylines = FALSE)
seg_words
```

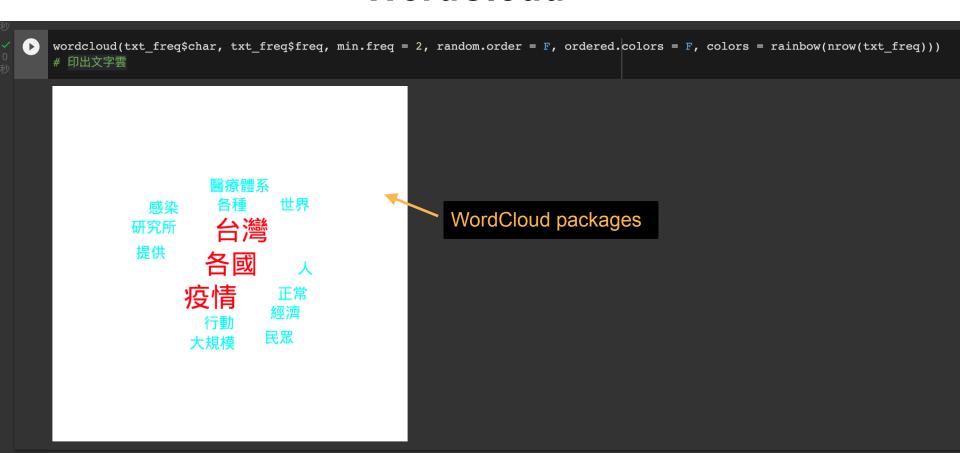
## **Word Segmentation**

```
\Theta
cutter <- worker(user = "new words.txt", stop word = "stop words.txt", bylines = FALSE)</pre>
seg words <- cutter[content]</pre>
seg words
'嚴重'・'特殊'・'傳染性肺炎'・'疫情'・'持續'・'全球'・'蔓延'・'已'・'造成'・'三千多萬'・'人'・'感染'・'死亡'・'人數'・'破'・
'百萬'・'人'・'歐美'・'醫療體系'・'幾近'・'崩潰'・'各種'・'行動'・'限制'・'對於'・'各國'・'經濟'・'有'・'莫大'・'衝擊'・
'然而'・'台灣'・'實行'・'大規模' ♪∕封鎖'・'行動'・'維持'・'正常'・'上班'・'上課'・'各種'・'經濟'・'活動'・'至今'・'無'・
'大規模'・'疫情'・'爆發'・'醫療體系'・'仍'・'能'・'正常'・'提供'・'民眾'・'服務'・'顯示'
'有效'・'獲得'・'世界'・'各國'・'高度肯定'・'國家衛生研究院'・'群體'・'健康'・'科學'・'研究所'・'感染'・'症'・'疫苗'・
'研究所'・'合作'・'利用'・'真實'・'數據'・'數學模型'・'證實'・'良好'・'邊境'・'管控'・'詳盡'・'疫情'・'調查'・'確實'・
'隔離'・'檢疫'・'以及'・'民眾'・'遵行'・'防疫措施'・'為'・'台灣'・'抗疫'・'四大'・'支柱'・'台灣'・'對'・'世界'・'各國'・
'提供'・'各項'・'協助'・'經驗'・'分享'・'期望'・'幫助'・'各國'・'盡快'・'從'・'疫情'・'陰霾'・'中'・'走出'
```

### Calculate frequency



#### WordCloud



# Q&A

