

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/main/tw>



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



```
library(jiebaR)
```

```
library(tmcn)
```

```
library(dplyr)
```

```
library(lubridate)
```

```
library(stringr)
```

```
library(wordcloud)
```

Add a variable to add the context

variable

context

```
content <- "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，維持正常上班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。台灣也對世界各國提供各項協助與經驗分享，期望幫助各國盡快從疫情的陰霾中走出。"
```

content

show context

```
cutter <- worker(bylines = F)
cutter[content]
```

#斷詞

define cutter

Word Segment

Word Segmentation

```
0 沙  
▶ cutter <- worker(bylines = F) ← define cutter  
  cutter[content] ← segment  
  #斷詞
```

☞ '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' · '上班' · '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' · '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' · '對' · '世界' · '各國' · '提供' · '各項' · '協助' · '與' · '經驗' · '分享' · '期望' · '幫助' · '各國' · '盡快' · '從' · '疫情' · '的' · '陰霾' · '中' · '走出'

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]+?")
```


```
cutter[content]
```

← delete numbers and English words

#去掉數字、英文(如果有需要)

new words and stop words

```
writeLines(new_words, "new_words.txt")
```



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 <- cutter[content]  
seg_words
```



Word Segment

Word Segmentation

```
cutter <- worker(user = "new_words.txt", stop_word = "stop_words.txt", bylines = FALSE)
seg_words <- cutter[content]
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' · '台灣' · '防疫' · '政策' · '正確' · '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

```
txt_freq <- freq(seg_words)
```

```
# 計算詞彙的頻率
```

calculate frequency

```
txt_freq <- arrange(txt_freq, desc(freq))
```

```
# 從大到小排列
```

sort by number

```
head(txt_freq)
```

```
# 檢查前5名
```

check the top 5

A data.frame: 6 × 2

	char	freq
--	------	------

	<chr>	<dbl>
--	-------	-------

1	各國	4
---	----	---

2	疫情	4
---	----	---

3	台灣	4
---	----	---

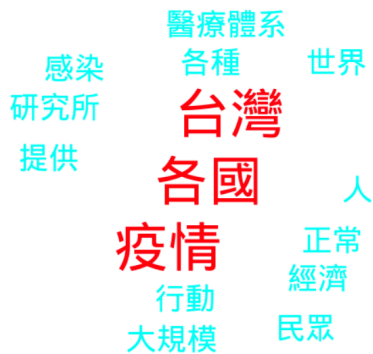
4	行動	2
---	----	---

5	各種	2
---	----	---

6	經濟	2
---	----	---

WordCloud

```
wordcloud(txt_freq$char, txt_freq$freq, min.freq = 2, random.order = F, ordered.colors = F, colors = rainbow(nrow(txt_freq)))  
# 印出文字雲
```



WordCloud packages

Q&A

