R 語言
作業五重點回顧

王貿

國立臺灣大學行為與資料科學研究中心助理研究員

國立臺灣大學政治學系博士、兼任講師

maowang01@gmail.com

沒用到的package就不要library

- 不是吹毛求疵，因為套件的namespaces（所有functions的名稱）可能會有同樣的function名稱，後library的套件就會蓋過先library的套件。
- 常見案例：dplyr::filter

- 當然重複library也是沒意義的。
  library(tidyverse)
  library(ggplot2)
  library(dplyr)

轉換成html的時候，有很多warnings...

- 可以在code chunk最上方，寫上**{r , warning=FALSE}**，警告訊息就不會產生。

為什麼第一題有中文？Encoding！

- trump <- read.csv("data/trump_twitter_archive.csv", stringsAsFactors = FALSE, **encoding = "unknown"**)
- Encoding(trump$text) %>% head(10)

- **readr::read_csv() 預設是UTF-8**
- trump <- readr::read_csv("data/trump_twitter_archive.csv ", **locale = default_locale()**)
- Encoding(trump$text) %>% head(10)

處理錯誤日期

- 有9筆轉換時產生錯誤，先轉成NA，然後再用tidyr::fill來填上

```
trump <- readr::read_csv("data/trump_twitter_archive.csv") %>%
    # 如果日期為5個字元，轉為NA，若不是則維持一樣。
    mutate(created_at = dplyr::if_else(
     str_length(created_at) == 5, # 邏輯判斷式
       true = NA_character_,
       false = created_at)) %>%
    mutate(date = lubridate::mdy_hms(created_at),
           date = as.Date(date)) %>%
    select(text, date) %>%
    # 用NA的前一筆觀察值補上
    tidyr::fill(date, .direction = "down")
```

辭典很重要！

- 使用不同的停用字辭典就會產生不同的結果。
- 這次作業我用的是：

  tidytext::stop_words (728 unique words)

  stopwords::stopwords() (175 unique words)

geom_col為什麼變成漸層色？

- aes(fill = month)
- 如果**month是數值（numeric）**的資料，上色時ggplot2會自動選擇漸層色。

- aes(fill = factor(month))
- 如果**month是類別（factor）**的資料，上色時ggplot2會選擇不同色系的顏色。

資料清理要更注意！

- 因為這是作業，主要是讓同學練習一些重要的概念與技能，資料清理我不會太要求。但**對期末報告來說，如果要提高分析的品質，資料清理一定要多加注意。**
- 例如：哪些要建入斷詞辭典，哪些要建入停用詞辭典，哪些token應該也要在斷詞後移除。