今日上課內容

- 作業三檢討與重點提醒（solution檔案已上傳）。
- 爬蟲需要的小工具：請先至 SelectorGadget（https://selectorgadget.com/）安裝瀏覽器套件（限Chrome），或是依照網頁說明加入書籤。
- 作業四的內容已經公布。

- DataCamp與Piazza都有手機App，可以下載使用。

R 語言
作業三重點回顧

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

str_extract_all 與 str_extract

- str_extract_all的結果會回傳成list，如果結合mutate就會變成list column，要解開的話，要在使用unnest()。
- str_extract的結果直接就是character的vector形式，不需要再轉換。

```
> udn_corpus
# A tibble: 221 × 3
      id date      text
   <int> <list>    <chr>
 1     1 <chr [1]> 監院糾正政
 2     4 <chr [1]> 洪仲丘案起
 3     5 <chr [1]> 管教不當 國
 4     6 <chr [1]> 酒駕撞人兄
 5     7 <chr [1]> 鞏固SOGO經
 6     8 <chr [1]> 悠活審照 4
 7     9 <chr [1]> 「沒洗錢」
 8    10 <chr [1]> 「沒貪汙」
 9    11 <chr [1]> 換3394億紓
10    12 <chr [1]> 公懲法首例
# … with 211 more rows
```

```
> udn_corpus
# A tibble: 221 × 3
      id date       text
   <int> <chr>      <chr>
 1     1 2013-08-21 監院糾正政
 2     4 2013-08-01 洪仲丘案起
 3     5 2013-07-16 管教不當
 4     6 2013-06-16 酒駕撞人兄
 5     7 2013-05-10 鞏固SOGO經
 6     8 2013-05-06 悠活審照
 7     9 2013-05-01 「沒洗錢」
 8    10 2013-05-01 「沒貪汙」
 9    11 2013-04-30 換3394億紓
10    12 2013-04-02 公懲法首例
# … with 211 more rows
```

以data frame為主要的資料處理格式

- 部分同學在處理資料時，習慣將df$column取出來處理，處理完後再 df$column <- ，比較不建議這種做法。
- 因為這樣一次只能處理一個column的資料，處理完就必須assign回去那個column，**無法藉由 %>% 的方式將所有的資料清理步驟串連起來**。
- 比較建議的做法是：

```
data_clean <- data %>%
      mutate(text = str_replace_all(...)) %>%
      filter(str_length(word) >= 3)
```

dplyr::distinct在做什麼？

- Retain only unique/distinct rows from an input tbl. This is similar to unique.data.frame(), but considerably faster.
- data %>% **distinct**(date, word, .keep_all = TRUE)

group_by() 與 ungroup()

- group_by(column1, column2, …)
- ungroup()

使用完group_by後，請養成好習慣**使用ungroup**。

- count(column1) == group_by(column1) %>%
  summarize(n = n()) %>%
  ungroup()

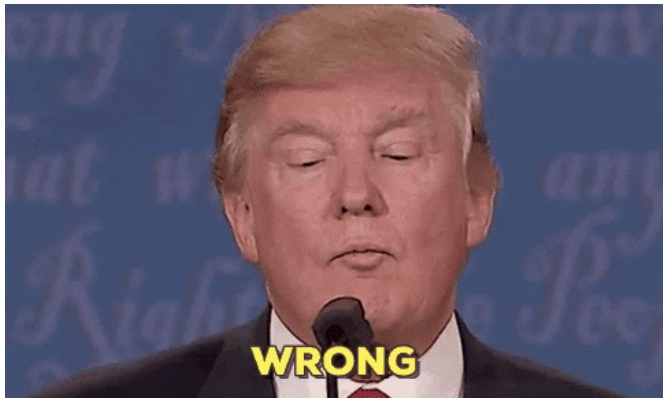
當join兩個data frames的欄位名稱不同

- ****_join 兩個data frames的key column名稱不同。
- 不能直接使用：
  df1 %>% anti_join(df2, by = “key”)

  要改用：
  df1 %>%
  anti_join(df2, **by = c(“df1.key1” = “df2.key2”)**)

為什麼沒辦法轉成html檔？

- 我們使用的Rmd檔案，會幫我們在轉換成html檔時，類似先重開一次RStudio，從頭到尾將你的每一個 code chunk 執行一次，**如果其中有任何 error message 產生，則會無法成功轉換**。
- 另外的情況則可能是讀檔的路徑有誤，如果無法成功排除，建議可以另外建立Rproj檔案再執行，比較不會有問題。



一些提醒

善用註解

- Help me help you!
- #

記得看一下跑出來的結果

- Seeing is believing

程式碼代表你的想法

- 寫程式沒有「正確」的寫法，重點在你怎麼思考問題！