R 語言與文字探勘
課程介紹

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

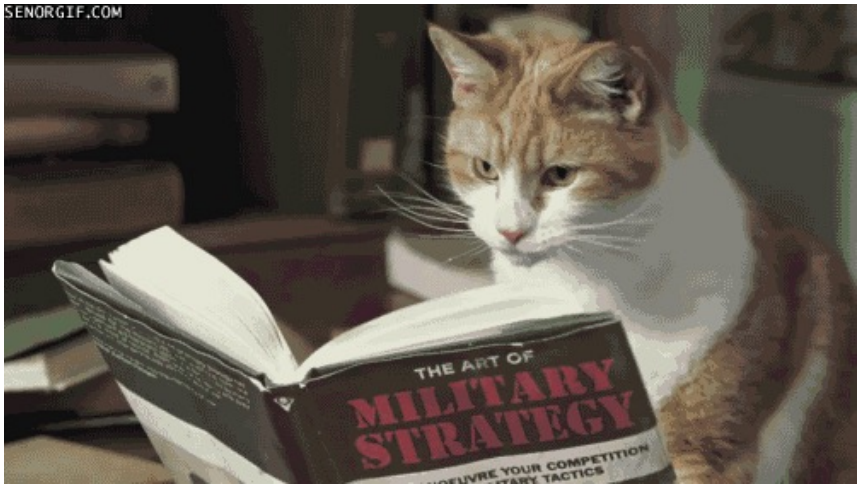
maowang01@gmail.com

課程適合對象

- 尚無 R 基礎或是有使用過但不熟，但對學習 R 有強烈興趣；
- 對文字探勘有興趣；
- 不排斥大量的英文說明文件及練習。
- 有初等統計的基礎較佳（平均數、變異數、基礎矩陣運算）。
- 對公共議題有興趣。

你主要會學到什麼？

- 基礎與中階的 R 語言
- 文字資料清理
- 資料視覺化
- 基礎爬蟲
- 文字探勘技術



各週內容簡介

週次	日期	主題	備註
1	9/28	課程概覽	
2	10/5	R 語言基礎：簡介	
3	10/12	R 語言基礎：資料篩選與整理	# 作業一
4	10/19	R 語言基礎：資料探索與分析	
5	10/26	R 語言基礎：程式設計基礎	# 作業二
6	11/2	文字資料前處理（regex, tm, stringr）	
7	11/9	文字資料前處理：斷詞（tidytext, jiebaR）	# 作業三
8	11/16	資料視覺化（ggplot2）	
9	11/23	爬蟲（web scraping）	# 作業四
10	11/30	期末報告主題說明	
11	12/7	詞袋（bag of words）模型	
12	12/14	情緒分析（sentiment analysis）	# 作業五
13	12/21	主題模型（topic models）	
14	12/28	自然語言處理（Natural Language Processing, NLP）	# 作業六
15	1/4	詞嵌入（word embedding）模型	
16	1/11	資料溝通（shiny）	
17	1/18	期末口頭報告	
18	1/25	期末口頭報告	

這學期你需要做什麼？

- 練習、練習、練習！
- 理解文字分析模型的原理。
- 實作出一個小專案。

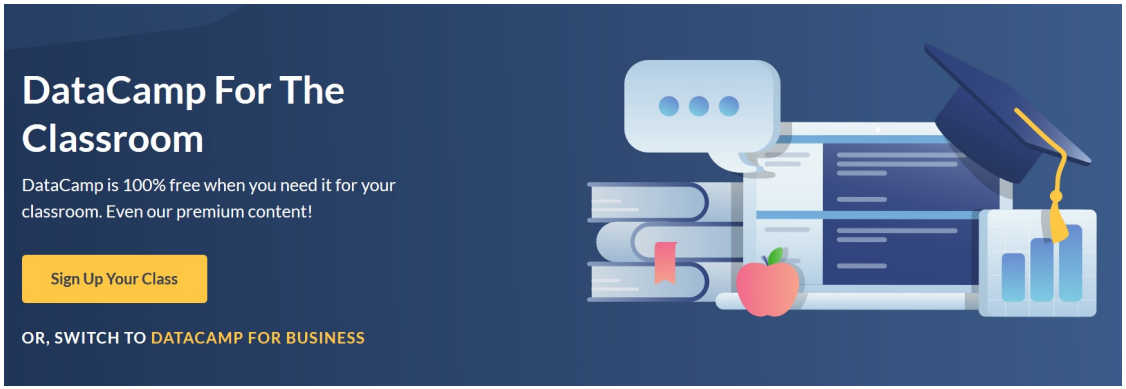
評分標準

- **課堂作業（60%）**：作業共6次，取5次最高分作業計算，每一作業各12分。
- **期末口頭報告（10%）**：各組人數1-3人。報告需有明確的研究問題、文字資料蒐集方式、分析方法、研究發現、研究結論與建議等部分。
- **期末書面報告（25%）**：字數至少4,000字以上，以組為單位，遲交者不予計分。
- **線上討論（5%）**：於課程之線上討論平臺具名提問或是回答他人提問，每次1分，5次以上則滿分。

本課程使用的線上平臺

- Piazza

- Datacamp classroom

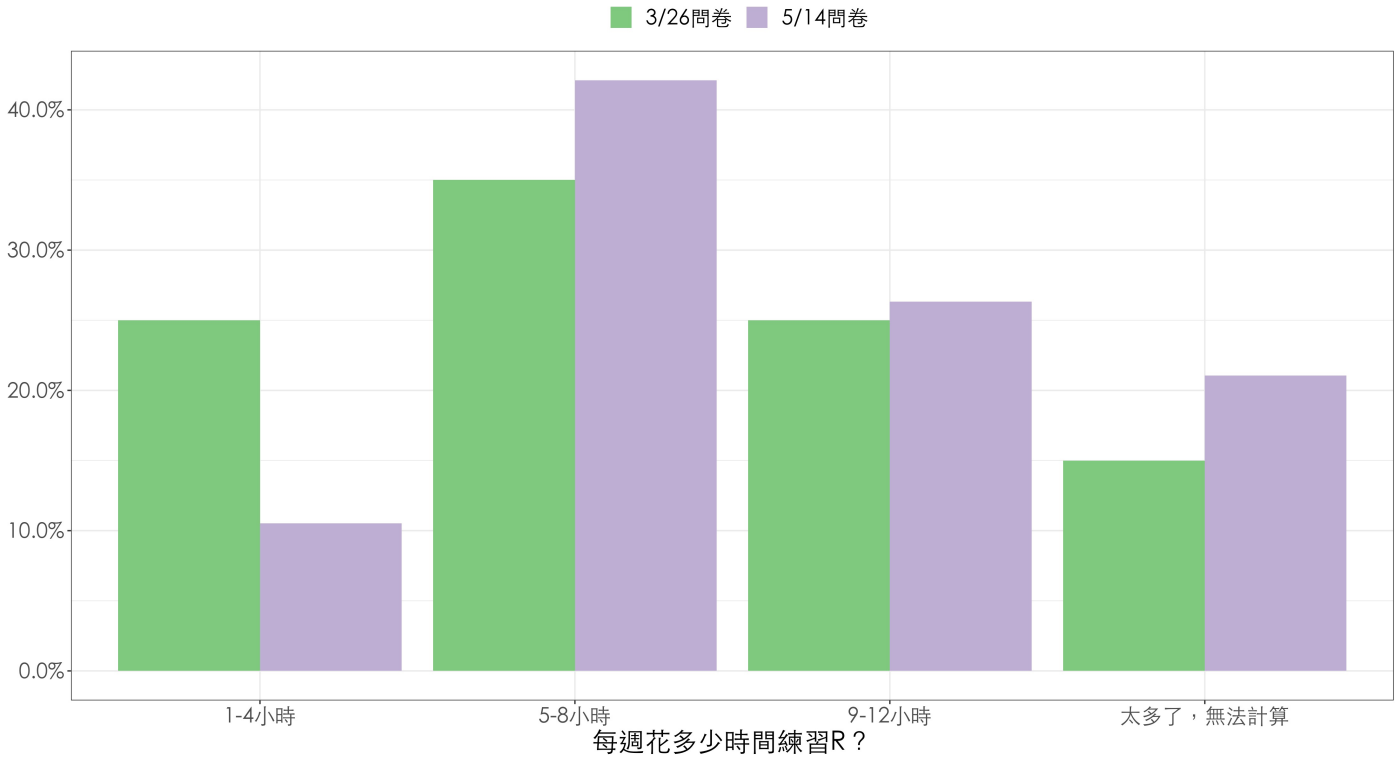


參考文獻簡介

- Thieme, N. (2018). R generation. Significance, 15(4), 14-19. **[R]**
- Grolemund, G. (2014). *Hands-On Programming with R*. **[manual]**
- Grolemund, G. & Wickham, H. (2017). *R for Data Science*. **[manual]**
- Silge, J. & Robinson, D. (2018). *Text mining with R: A tidy approach*. **[manual]**
- Wilkerson, J., & Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science*, *20*(1), 529-544. doi:10.1146/annurev-polisci-052615-025542 **[overview]**
- Jurafsky, D. & Martin, J. H. (2018). *Speech and Language Processing* (3rd ed. draft) **[NLP]**

修課同學需要先知道的一些事...

- 每週花費時間
  （期末課程意見）
  - 1小時-2小時 1 人 (6.3%)
  - 2小時-3小時 0 人 (0.0%)
  - 3小時-4小時 3 人 (18.8%)
  - 超過4小時 12 人 (75.0%)



共20人填答，7人修課前未使用過R，13人略懂。

下週上課前要完成的事

- 安裝最新版的 R（4.1.1）
- 安裝最新版的 RStudio（1.4以上）
- 下載好上課教材。

- 如果安裝軟體或是下載上課教材有任何問題，請務必上Piazza發問，或是第二週提前半小時到上課教室排除。
- 已選上課程的同學，請確認可以進入臺大COOL系統，以後所有的課程公告，都會透過該系統通知。