R語言與文字探勘
學期課程回顧

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

為什麼要學程式語言？

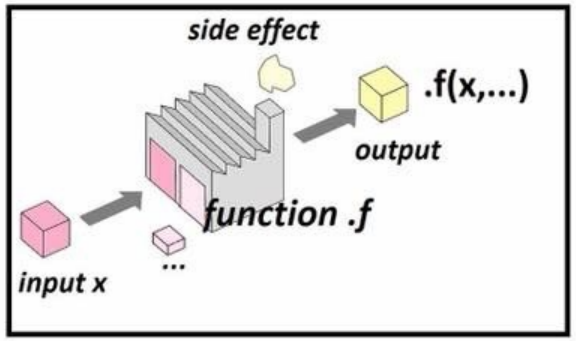
- 簡化繁瑣重複的工作
- make your life easier

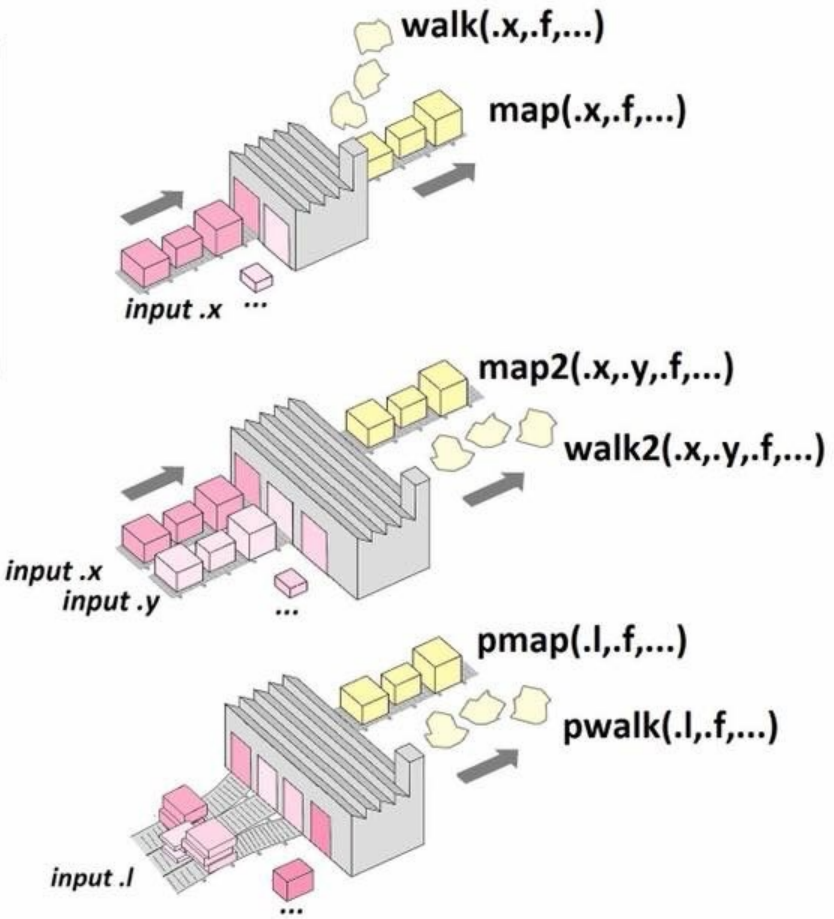
為什麼程式語言要學 R？

- 免費！
- 功能強大（各種分析都可以辦到，繪圖尤其強大）。
- 易學（對人文社會科學背景者較容易）。
- 友善的學習社群。

R 語言的基本要素

- 物件（Object）
- 函式（Function）





About computation in R

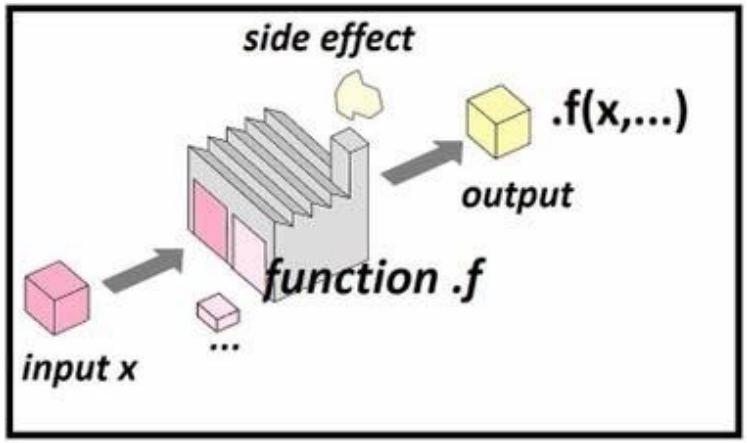
"To understand computations in R, two slogans are helpful:

- Everything that exists is an object.

- Everything that happens is a function call."

— John Chambers

R 語言範例



- print(x,

  digits = getOption("digits"), ...)

- 參數（Arguments）

object an object for which a summary is desired.

digits minimal number of significant digits

- R語言的程式碼是**有區分大小寫**（case sensitive）

R 語言參數說明

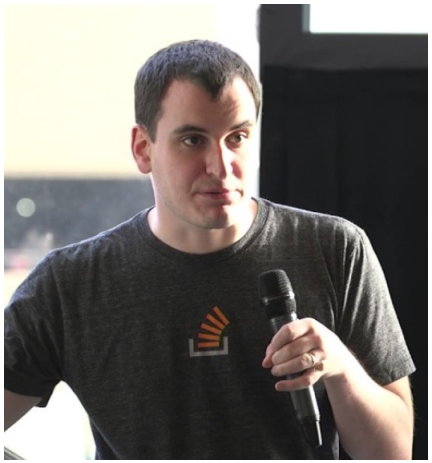
- 必要參數
- x　a numeric vector, matrix or data frame
- 預設參數
- y　NULL (default) or a vector, matrix or data frame with compatible dimensions to x. The default is equivalent to y = x (but more efficient).

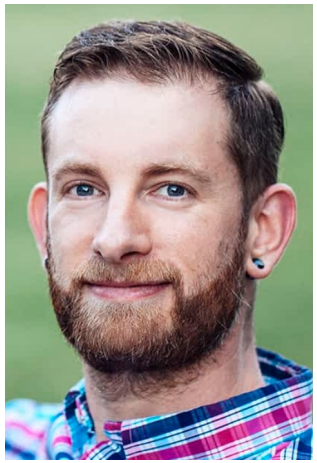
- 依參數位置（position）
- 依參數名稱

```
cor(  x,
      y = NULL,
      use = "everything",
      method = c("pearson", "kendall", "spearman"))
```

R 語言開發生態

- R Development Core Team (Base R)
- 其他套件（package）製作者





為什麼有人覺得 R 不好學？

- **難處1：要記各種函式（function）**
- 解答：沒有人真的能全部記住，重點是該函式的說明文件（documentation）清楚嗎？

- **難處2：入門的門檻不低，要學的套件（package）太多！**
- 解答：理解基本的資料結構，可以降低後續學習的門檻；學習具有同樣設計邏輯的套件（如tidyverse）。

怎麼問問題？

- **Repr**oducible **ex**ample (**reprex**)
  https://github.com/tidyverse/reprex

- 讓別人能最小化的**重現**你的問題，才能夠幫你處理問題。
  - A **minimal dataset**, necessary to reproduce the error
  - The **minimal runnable code** necessary to reproduce the error, which can be run on the given dataset.

https://stackoverflow.com/questions/5963269/how-to-make-a-great-r-reproducible-example

Coding style

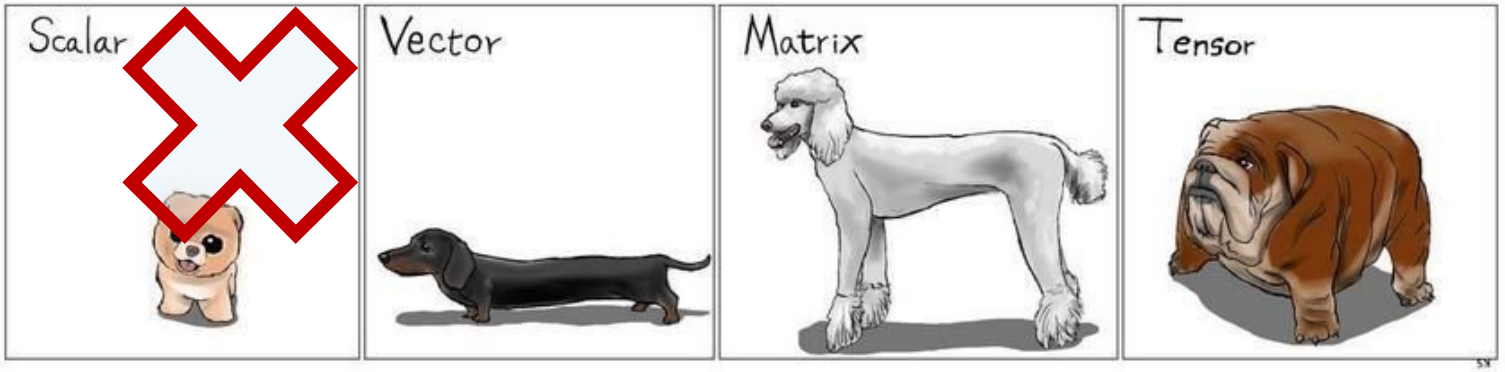
- https://style.tidyverse.org/
- 讓你自己及合作者更容易讀你的code。

資料類型

- 數值（numeric）
- 字串（character）
- 邏輯判斷（logical）

儲存格式

	Homogeneous	Heterogeneous
1d	Atomic vector	List
2d	Matrix	Data frame
nd	Array	



什麼是文字探勘？

- Github repository
- https://github.com/aleszu/textanalysis-shiny

- Shiny App實做
  https://storybench.shinyapps.io/textanalysis/

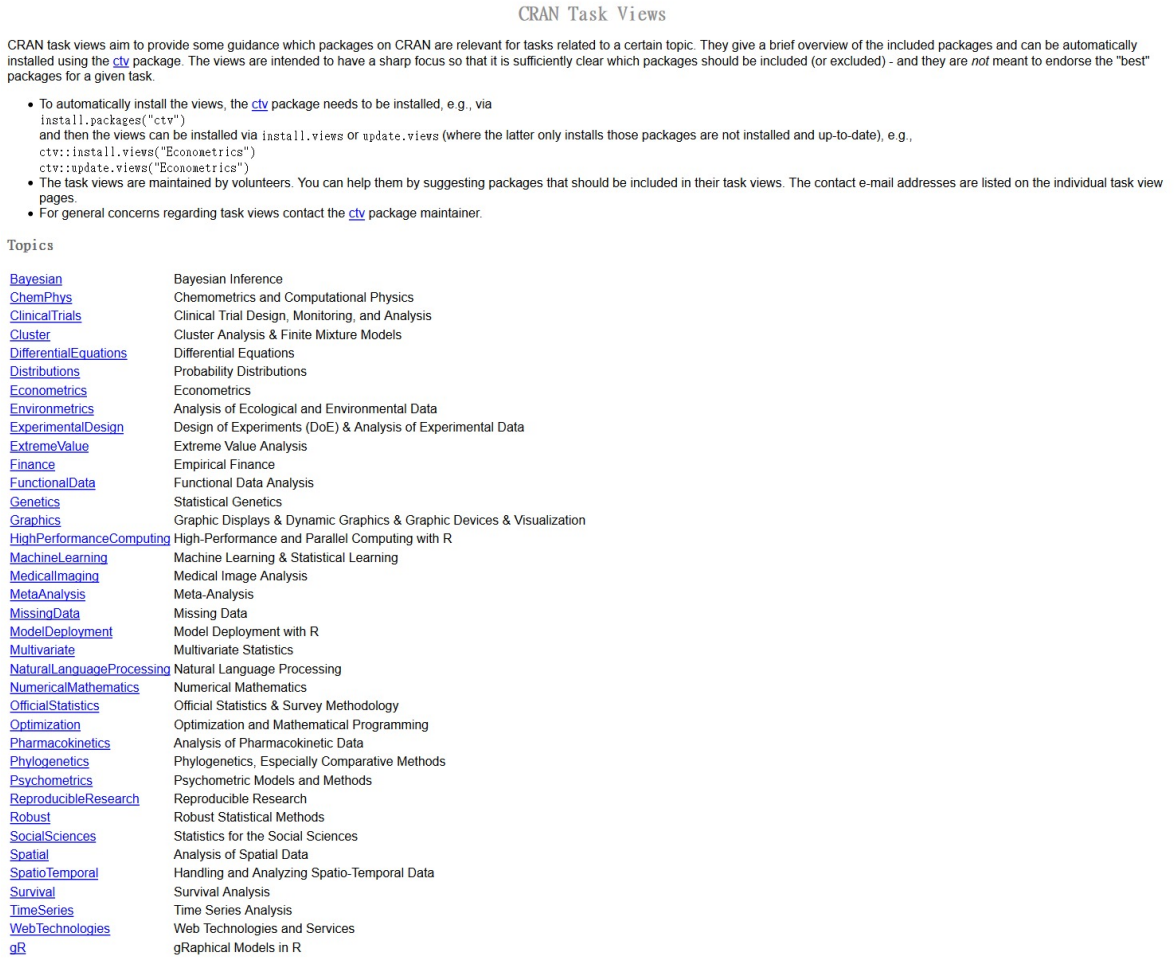
你學了哪些文字探勘的工具？

- 詞袋模型（bag of words, BOW）：DTM, TF-IDF
- 情緒分析
- 主題模型
- 自然語言處理
- 詞嵌入

R 能做什麼事？

- 統計分析
- 文字探勘
- 社會網絡分析
- 空間分析
- 網路爬蟲
- App
- ……

https://cran.r-project.org/web/views/



CRAN Task Views

CRAN task views aim to provide some guidance which packages on CRAN are relevant for tasks related to a certain topic. They give a brief overview of the included packages and can be automatically installed using the ctv package. The views are intended to have a sharp focus so that it is sufficiently clear which packages should be included (or excluded) - and they are *not* meant to endorse the "best" packages for a given task.

- To automatically install the views, the ctv package needs to be installed, e.g., via
  `install.packages("ctv")`
  and then the views can be installed via `install.views` or `update.views` (where the latter only installs those packages are not installed and up-to-date), e.g.,
  `ctv::install.views("Econometrics")`
  `ctv::update.views("Econometrics")`
- The task views are maintained by volunteers. You can help them by suggesting packages that should be included in their task views. The contact e-mail addresses are listed on the individual task view pages.
- For general concerns regarding task views contact the ctv package maintainer.

Topics

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
ExtremeValue	Extreme Value Analysis
Finance	Empirical Finance
FunctionalData	Functional Data Analysis
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
MissingData	Missing Data
ModelDeployment	Model Deployment with R
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods
ReproducibleResearch	Reproducible Research
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
SpatioTemporal	Handling and Analyzing Spatio-Temporal Data
Survival	Survival Analysis
TimeSeries	Time Series Analysis
WebTechnologies	Web Technologies and Services
gR	gRaphical Models in R

其他學習資源

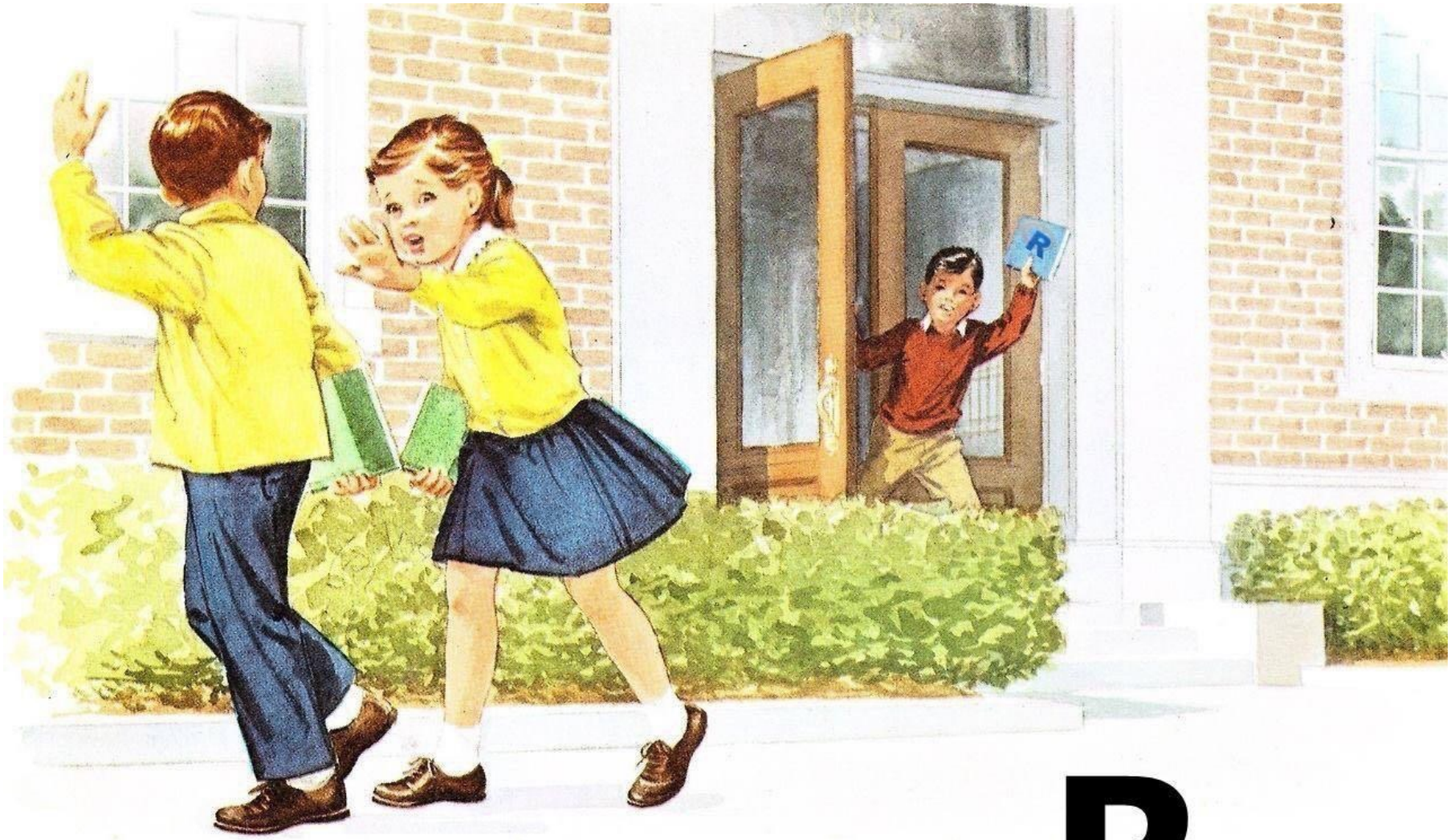
- DataCamp 

- Coursera 

- edX 

- Cheat sheet 

https://www.rstudio.com/resources/cheatsheets/



Run, or he's going to tell us about **R** again!