

鍵盤也能蒐資料：資料爬析在社會學的運用

工作坊課程

- 時間：12月14-15、21-22日（共4天）
 - 地點：中研院社會所（802會議室）
 - 邀請「DSP 智庫驅動」之講師與助教群
-
- 謝宗震（Johnson）：國立清華大學統計所博士，致力推廣資料科學。結合統計學、網路爬蟲、文字探勘技術於2014台灣社會學年會發表「媒體報導關聯性分析—以太陽花學運為例」。
 - 劉佳欣（Rafe）：國立交通大學運管所博士候選人，從手搖飲經營到航空營收管理，相信資料分析是職場不變的道理。

※第一天(12/14)早上 9：30 開始辦理報到手續，請於 10：00 正式上課前，確認自己的電腦設備可以執行上機作業（包含軟體安裝與無線上網）。

課程內容與時間表

Date	Time	Topic
12/14	Day1	Main Topic: Quick R Introduction
	10:00 – 12:00	<ul style="list-style-type: none"> • What is R, Why R • Intro & Install packages • Numeric manipulations and description <ul style="list-style-type: none"> • vector
	12:00 – 13:00	Lunch and Break
	13:00 – 17:30	<ul style="list-style-type: none"> • String manipulation <ul style="list-style-type: none"> • gsub, grep, grepl, substr • Data manipulations - list and data.frame
12/15	Day2	Main Topic: R programming
	10:00 – 12:00	<ul style="list-style-type: none"> • File I/O <ul style="list-style-type: none"> • ReadLines • Read / Write table • Read data from web (csv and json) • Control flow <ul style="list-style-type: none"> • If...else... • for loop • apply
	12:00 – 13:00	Lunch and Break
	13:00 – 17:30	<ul style="list-style-type: none"> • Pipe coding style - magrittr • Data ETL - dplyr <ul style="list-style-type: none"> • filter, select, arrange • mutate • group_by, summarise • Basic web crawler – intro to ‘rvest’ • (Case 1: 台灣公益資訊中心 - 非營利組織名錄)

12/21	Day3	Main Topic: Basic R crawler
	10:00 – 12:00	From HTML to Crawler <ul style="list-style-type: none"> • Crawler strategy • How to parse a statistic webpage • Run parsing (control, error handling) • Exercise: (Case 2: 善耕 365 首頁)
	12:00 – 13:00	Lunch and Break
	13:00 – 17:30	HTTP Request (Case 3: SER API) <ul style="list-style-type: none"> • GET • POST • SESSION / Cookies • Exercise: (Case 4 台灣高鐵)
12/22	Day4	Main Topic: Team Project
	10:00 – 11:00	Get data from opendata APIs <ul style="list-style-type: none"> • Case 5: data.taipei • Case 6: Google Alerts & News
	11:00 – 12:00	Advance parsing showcases
	12:00 – 13:00	Lunch and Break
	13:00 – 16:00	Team Project
	16:00 – 17:30	Text mining application: media association based on sunflower movement data