Introducing Data Science with R
R 語言與資料科學導論
Thur., 14:20-17:10, 2016

謝舒凱 Shu-Kai Hsieh
shukaihsieh@ntu.edu.tw

課程概述與目標

大數據 (Big Data) 的時代, 數位資料累積與增長的速度已經遠遠倍增於人類史上的任何階段。這樣一種鉅量資料風潮, 不僅改變了人文社會與自然科學研究的面貌, 在各項產業也產生了分析資料輔佐決策的迫切需求。在此背景下, 數位素養 (digital literacy) 已經成爲現代公民必須具備的基本素養之一, 近年來新興的**資料科學家** (data scientist) 更成爲當前最爲熱門的行業之一。然而由於數據資料的發展, 隨著社交媒體與社會網路的發展, 非結構性的文本資料所佔比例已經遠超過結構性的表格性資料, 使得文本的語言分析在資料科學發展中的角色顯得愈來愈重要, 特別對於人文與社會科學的學生而言, 更是開啓了一個結合數位科技與人文關懷的新的發展方向。本課程的設計, 就是在這個動機之下, 透過介紹語言學與文本分析知識結合統計計算, 並運用新的教學與實作平台, 希望能夠刺激人文、社會、傳播與其他財務管理、醫學等各領域學生之間的互動協作與學習, 培養跨領域的興趣與分析能力。

資料科學家的工作, 可以視爲是一個探索、預測與解讀資料意義的互動歷程。而語言分析的工作, 在了解文本資料的語意與情緒表現上是重要的關鍵。本課程結合了目前統計程式設計與自然語言處理技術, 以較爲簡潔容易入門的設計與實際操作導引, 希望可以讓毫無相關程式學習基礎的學生在本課程的帶領下, 達到以下的學習目標:

- 了解結構與非結構性資料的特性與預處理工作, 特別是針對中文文本中呈現的語言特性的處理方法。
- 了解中文的語言特性與文本解析 (text analytics) 的基本概念。
- 選擇適當的變數與特徵並加以合理調製, 對之進行描述統計與視覺探勘, 針對不同的問題點與數據類型, 找出適當的圖形表達與統計分析。
- 學習簡易的自然語言處理與機器學習預測模式, 並應用在自己關心的領域。

Introducing Data Science with R
R 語言與資料科學導論
Thur., 14:20-17:10, 2016

謝舒凱 Shu-Kai Hsieh
shukaihsieh@ntu.edu.tw

Syllabus (tentative)

Week	Date	Topic
1	09/15	中秋節放假
2	09/22	Introduction to Data Science and Text Analytics `Installing R and Rstudio`
3	09/29	Introduction to Data Science and Text Analytics `data types and structures (vector)`
4	10/06	Preparing / Obtaining Data `data structures (matrix, data frame, factor)`
5	10/13	Scrubbing Data `list, Basic I/O`
6	10/20	Exploratory data analysis and Graphics `control flow; writing functions`
7	10/27	Exploratory data analysis and Graphics `data wrangling with dplyr and tidyr`
8	11/03	Corpus and Natural Language Processing `data manipulation (with regex)`
9	11/10	Corpus and Statistics `statistics with R`
10	11/17	Advanced Graphics
11	11/24	Machine learning Basics
12	12/01	Text classification and clustering
13	12/08	**mini-Hackathon**
14	12/15	Discussion
15	12/22	Sentiment analysis
16	12/29	Reporting and presenting Data `Web application with Shiny`
17	01/04	Term project competition/presentation
18	01/11	Term project / report due

Introducing Data Science with R
R 語言與資料科學導論
Thur., 14:20-17:10, 2016

謝舒凱 Shu-Kai Hsieh
shukaihsieh@ntu.edu.tw

課程概述與目標

- 上課方式 分爲兩部分：針對每週主題有 2 個小時的課堂講解，與 1 小時的實作教學與練習。每次上課都會分派作業，以階段性的題組一步一步的累積相關知識。每次作業預估的工作量約每週 5 小時。所有學習活動的進度、歷程與成果都在網路平台上進行, 一方面老師與助教容易掌握, 同學之間也可彼此觀摩學習。

- 成績評量項目與方式：課後作業與課堂表現 (30%)；期中評量 (20%)；期末計畫展演、組織與報告 (50%)。

Capstone Project: 期中進行第一次評量。將以黑客松 (hackathon) 方式進行公開之跨領域小組協作競賽。期末則將邀請語言科技與資料科學新創公司團隊來擔任評審，也增進同學對於如何進行知識應用的了解。計畫展演的方式將以開放的短會 unconference 方式進行，由各組同學彼此協調與自行組織，藉此練習學習、組織與團隊協作的翻轉學習精神。

由於此課程內容較新, 目前並未有完整的中文教材, 此外, 考量到本課程是以**培養基本能力**爲目的, 在閱讀教材的設計上, 將以授課教師自己編撰的教材講義爲主, 輔以列在延伸閱讀中的英語延伸線上教材。由於本課程著重在「觀念激盪」與「創新實作」。有別於傳統上以紙本書籍閱讀爲主的課程設計, 本課程將以預先設計的「主題」爲核心, 透過翻轉式的「網路閱讀」爲學習方式, 鼓勵同學在此過程中, 學習到協力合作並累積自學經驗, 結合自身專業而能轉化成未來可能的應用。

Bibliography

[1] Datacamp. https://www.datacamp.com/home

[2] Big Data University. https://bigdatauniversity.com.cn/learn/.

[3] Garrett Grolemund and Hadley Wickham. **R for Data Science** http://r4ds.had.co.nz/.

[4] Dan Toomey. 2014. *R for Data Science*. Packt Publishing.

[5] Gergely Daróczi. 2015. *Mastering Data Analysis with R*.Packt Publishing.