上課前，請先做這些事...

- 安裝套件**tidytext, jiebaR, stopwords**
- 下載本週的code與資料檔
- 使用Windows作業系統的同學，請安裝Rtools (4.0) (https://cran.r-project.org/bin/windows/Rtools/)

R 語言
斷詞

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

課程主題重點

- 英文斷詞
- 中文斷詞

英文怎麼斷詞？

- 利用空白處（Whitespace）斷詞

中文怎麼斷詞？

- 沒有空白，所以要用別的方式。
- 使用辭典與演算法來決定。
- 目前較多人使用 jiebaR 來進行斷詞。中研院資科所前幾年已開放 CKIP 開源使用（但目前只能在Python執行）。

jiebaR

- 原本是python的package，被轉換到R語言上。
- 簡體語料，先將分析文件轉為簡體再進行斷詞，效果會更好。

- 兩步驟：先製作斷詞引擎，然後才進行斷詞。

Coding style: styler

- https://github.com/r-lib/styler
- 繳交作業前，請使用 styler 將 code 調整成比較易讀的 style。