

R語言推廣 探索性資料分析與文字探勘初探

游為翔

中央研究院資訊科學研究所

Data
Insights
Research
Lab



資料洞察實驗室



Lecturers



- 成大心理 101 級
- 台大心理所畢
- 中研院資訊所 資料洞察實驗室
資料科學家 / 研究助理
- 研究專長
 - 網路數據探勘
 - 消費及使用者行為分析
 - 機器學習
 - 推薦系統



最新消息

- 我們將舉辦 ACM Multimedia Systems 2017，日期在 6/20 – 6/23，敬請期待。:)
- 科學人專訪：〈讓資料對你說真心話——陳昇璋〉
- 歡迎十位來自四面八方的優秀暑期實習生加入本實驗室一起做研究 :)
- 2016 台灣資料科學年會花絮
- 2016 台灣資料科學年會圓滿落幕，點此看精采議程
- 徵才啟事：研究助理 / 博士後研究人員 / 軟體工程師
- 徵才啟事：博士後研究人員
- IEEE Spectrum: Reducing World of Warcraft Power Consumption
- 雜誌專訪：IT人甘苦談—從程式人跨到學術人的深度歷險

關於我們





Outline

- A. 取得資料 & 清理 -- 資料收集與清整
- B. 檢視資料與建立特徵 -- 探索性資料分析 + 模型
- C. 以不同面相建立更多特徵 -- 文字探勘
 - ** 如果我們時間來得及的話 ...
 - ** 同時若尚未安裝今天需要的 packages, 請同步安裝

今天課程結束後你(妳)應該會什麼？

□ 軟實力

- 知道分析資料的流程與常碰到的問題
- 基礎/常用的模型有哪些
- 文字可以怎麼處理

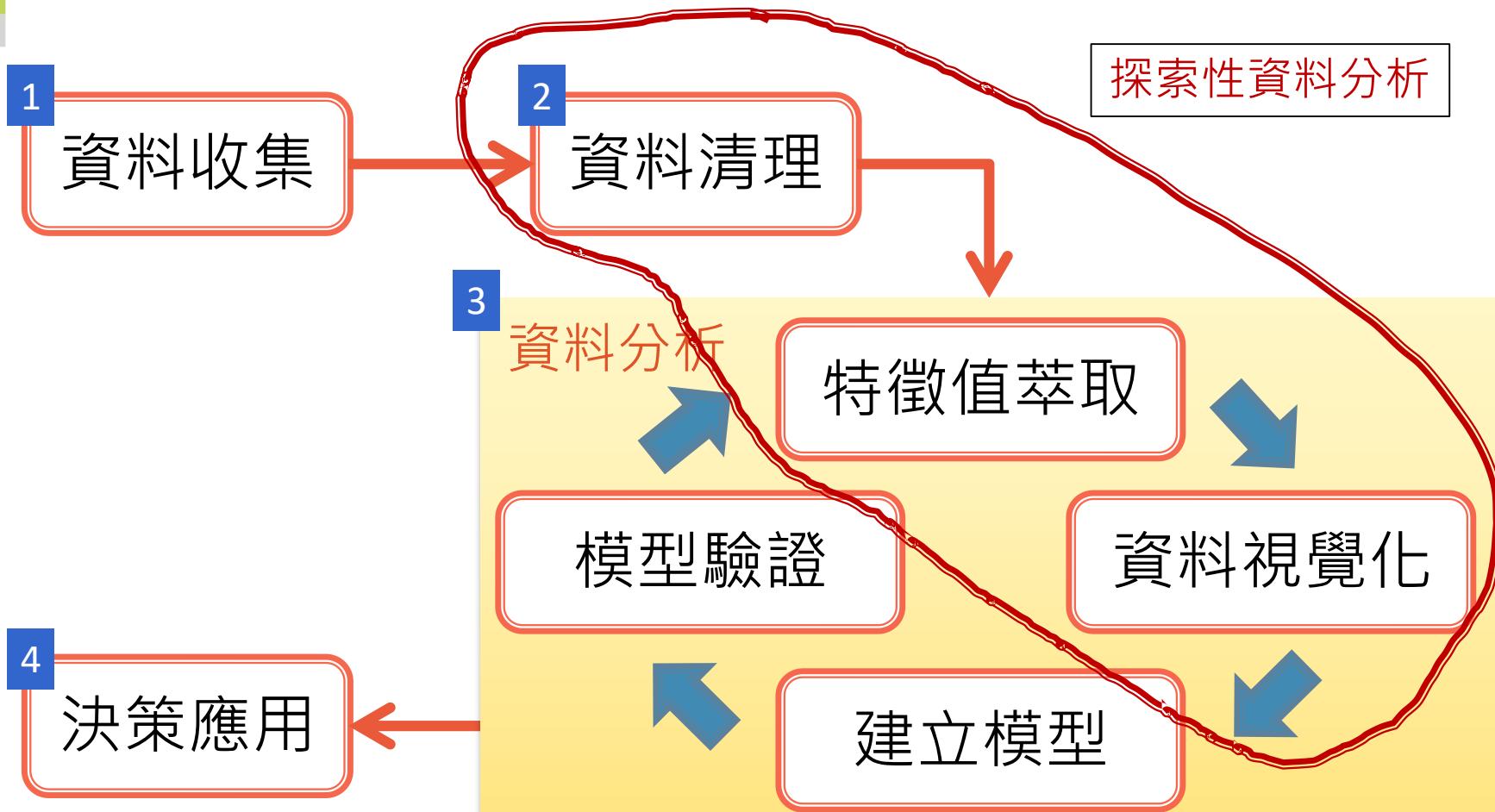
□ 硬底子

- 小小爬蟲：從一個野生的網站 download 數據下來
- 用“data.table”這個 package 來做資料的處理
- 用“ggplot”來畫精美的圖
- 學會如何在 R 實現 SVM/RandomForest
- 常用的文字分析



先來談談資料分析的流程

資料分析流程





什麼是探索性資料分析 (EDA)?

EDA (Exploratory Data Analysis)

- 初步透過視覺化方法進行分析，達到三個主要的目的
 - 了解資料
 - 發現 outliers 或異常數值
 - 找出重要的變數

- 不做過度假設地從原始數據看出隱含意義

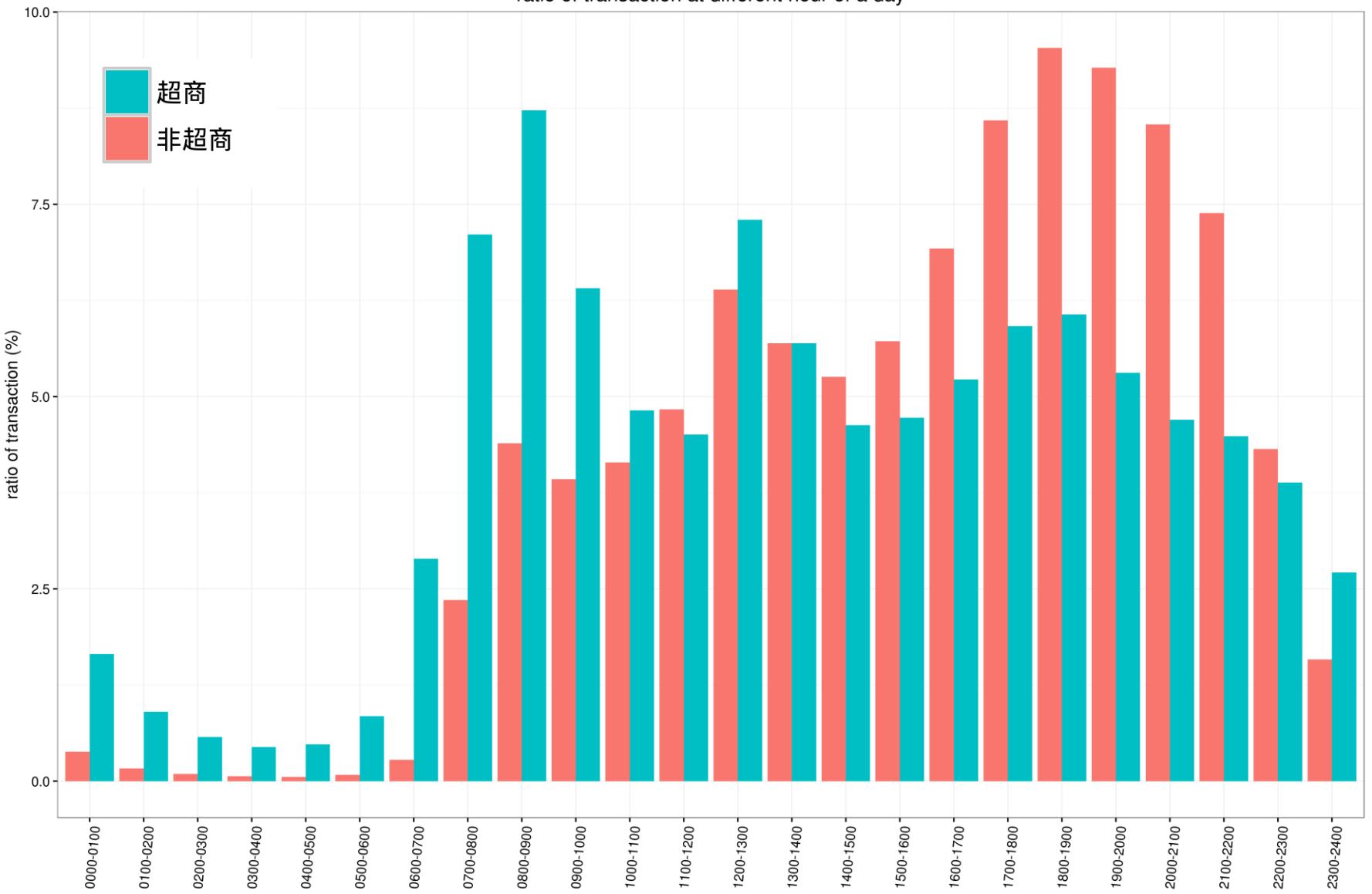
簡單來說，就是各種畫圖啊！

從畫圖的過程中，了解資料與觀察現象

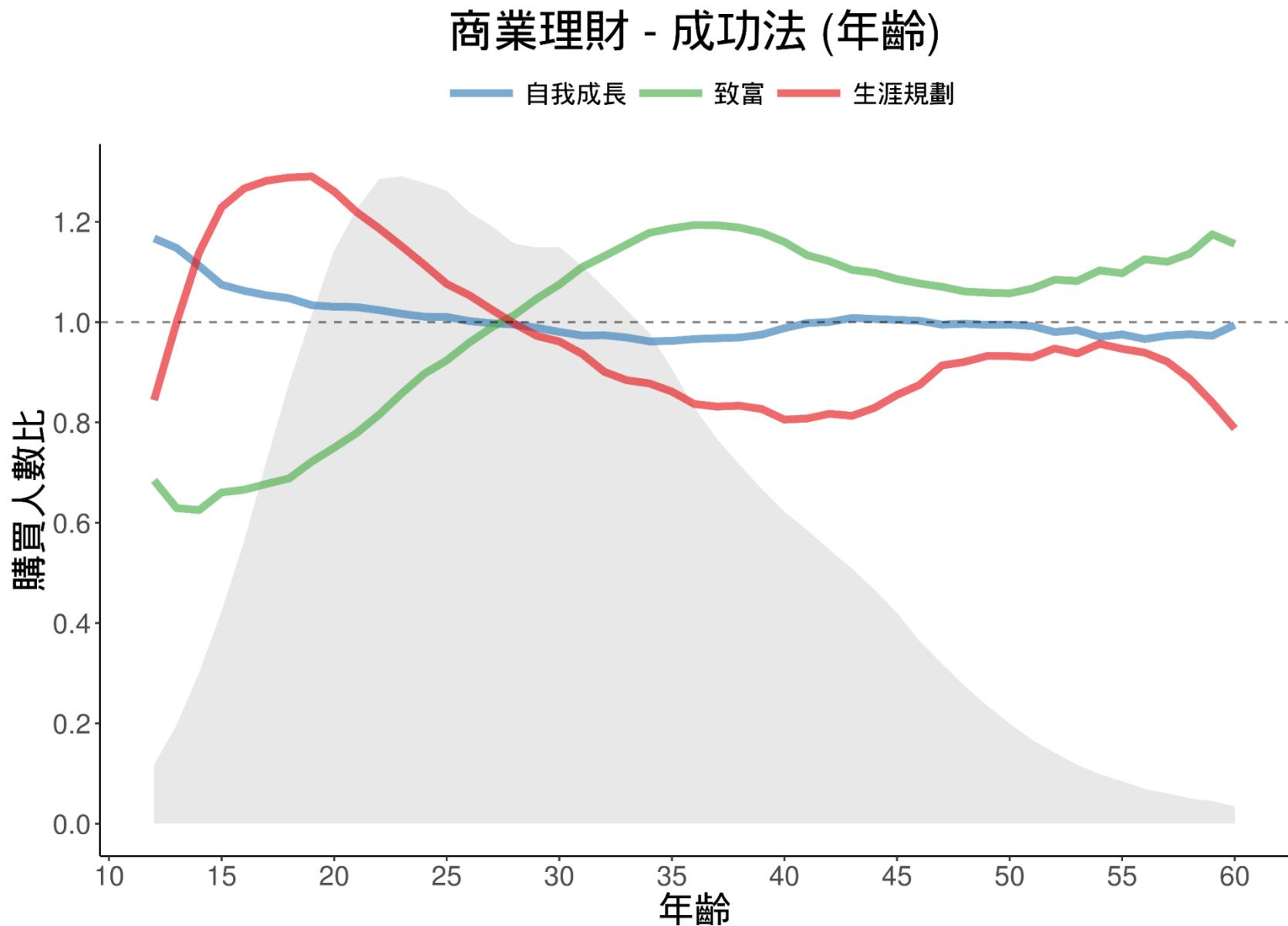
比方說...

Case 電子票證於各時段在超商與非超商之交易分布

ratio of transaction at different hour of a day

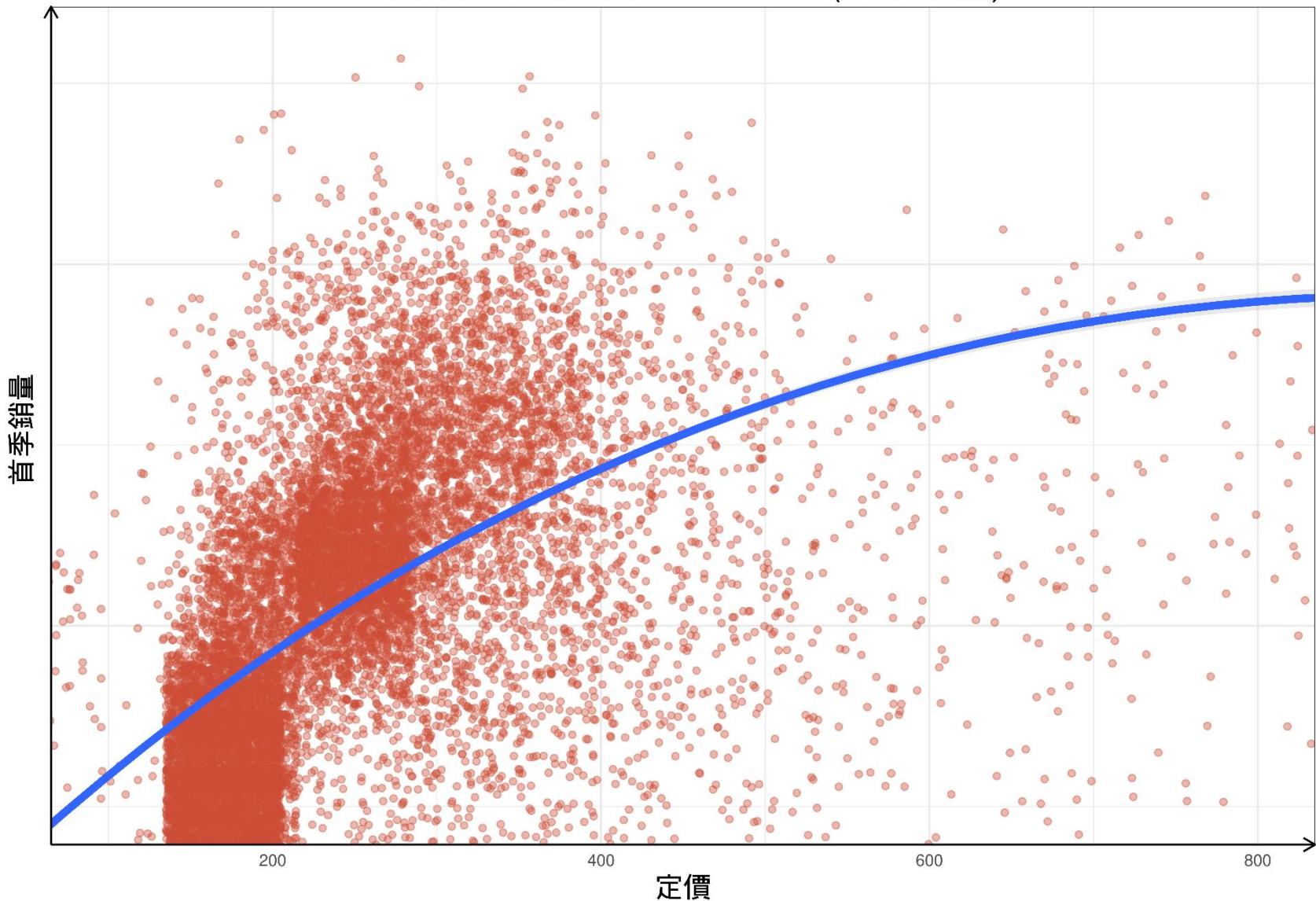


Case 網路購書在年齡與購買類型的改變



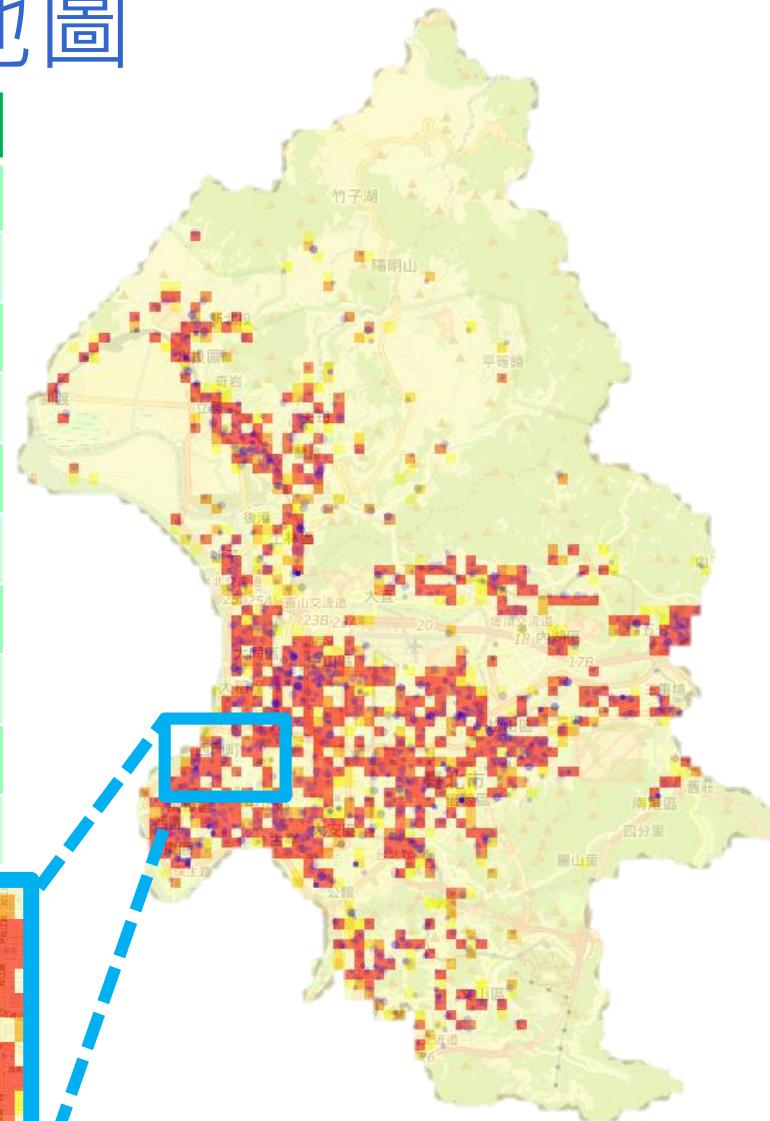
Case 網路書商在書籍定價與首季銷量之關係

首季銷量 ~ 定價, cor = 0.48 (文學小說)



Case 北市開放犯罪地圖

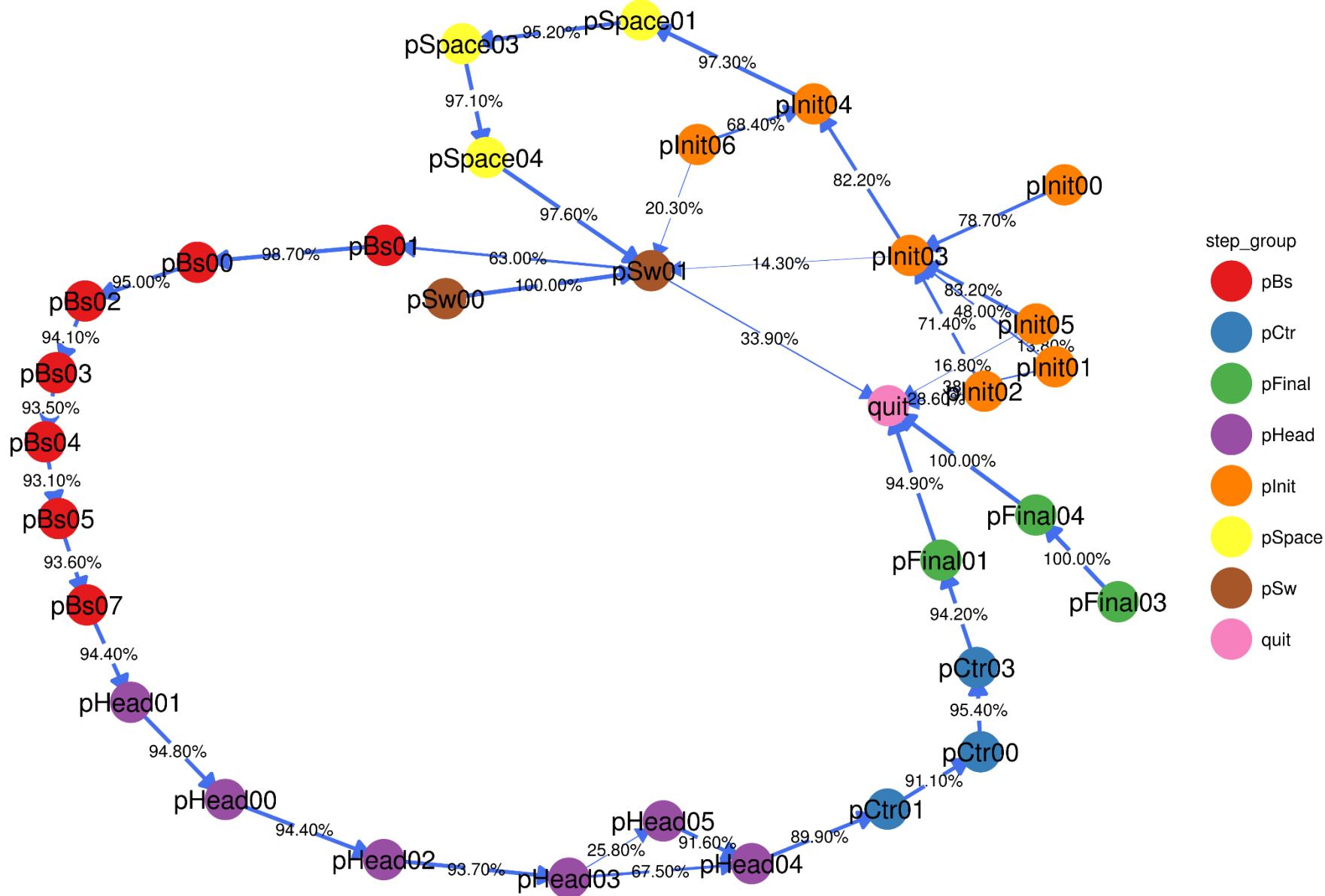
變數	高犯罪熱點	低犯罪熱點
教育程度	0.529	0.632
離婚比例	0.0733	0.0495
性別比例	89.5	93.4
公車站數	0	0
監視器數量	30	3
路燈數量	204	31
監視器距離	21.4	77.7
超商距離	94.8	206.3
捷運站距離	203.1	426.9
警察局距離	102.1	352.8



中山區長安西路19巷10-1號

中正區徐州路15號

Case: Setup Flow



常會用到的圖

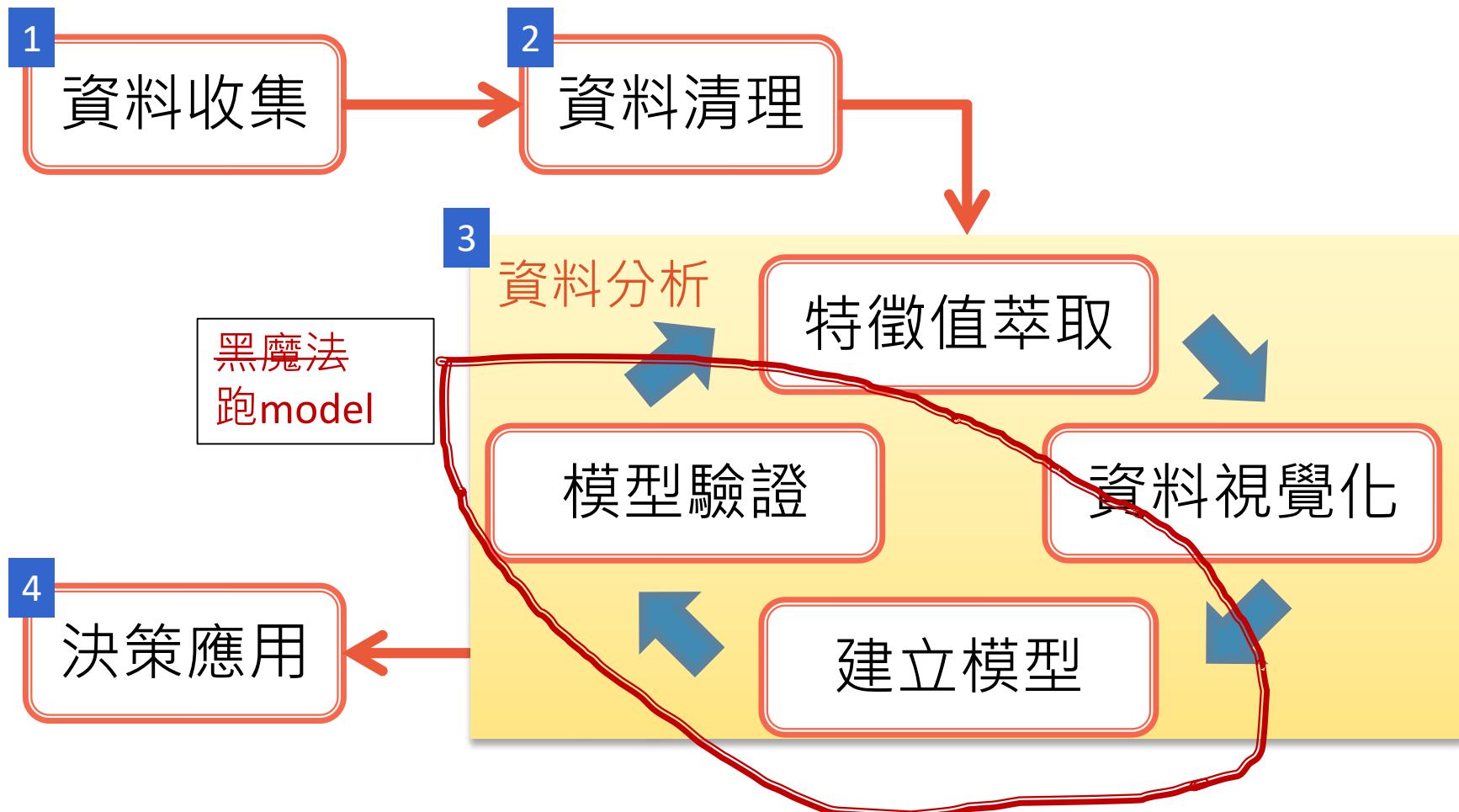
- 長條圖 (Barplot)
- 散布圖 (Scatter)
- 線圖
- ECDF (cumulative density function)
- 熱圖
- ...



檢視資料過程常碰到的狀況

- 資料型態跑掉, 數字的 leading 0 消失
 - 001, 002 → 1, 2
- 資料中有空白
- 要偵測特定的字串 pattern
 - 從地址中取出縣市鄉鎮區
- 值要做對應轉換
 - 比方說把 非常差 ~ 非常好 轉成 5 點量表
- 運算速度太慢

資料分析流程



模型要幹嘛，能吃嗎？



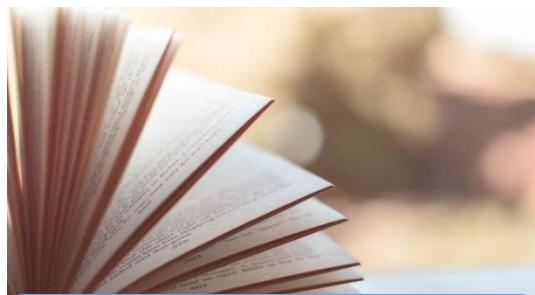
建立模型

□ 目標：

找到一個方法 $f(x)$ 來盡可能地解釋/預測我們有興趣的東西 y

$$y \approx f(\text{WORLD})$$

Case: 預測書籍暢銷機率與其首季銷量之關係



書籍與商品呈現特徵



書名關鍵字



上市前的市場狀況

建
模

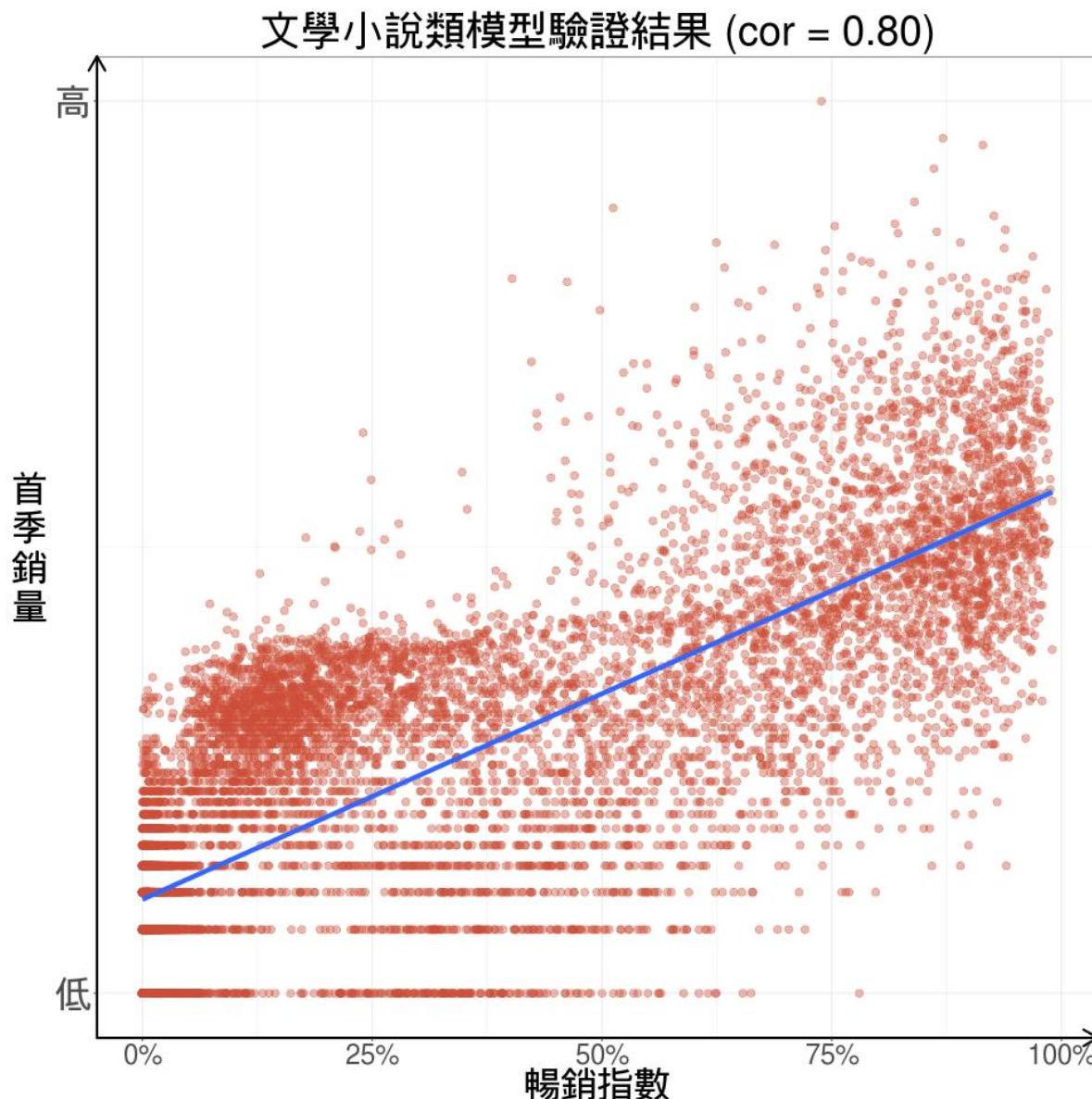


預測模型

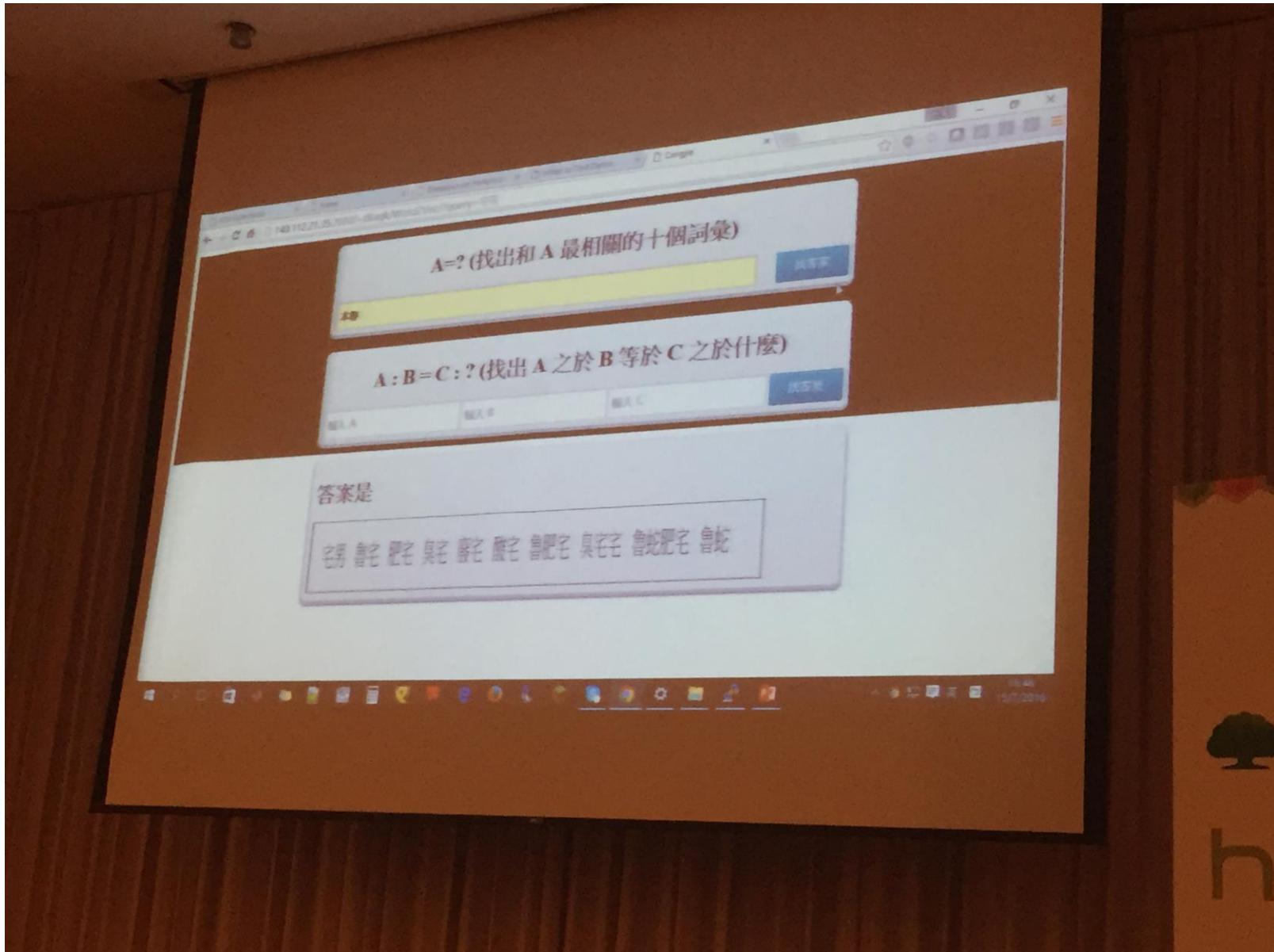
預
測



Case: 預測書籍暢銷機率與其首季銷量之關係



Case: 用 PTT 文章訓練填空漏字機器人



簡單介紹完資料分析的流程與一些案例後
大家的 package 應該也差不多裝完了 ...

來動手做看看吧!

資料收集
Data Collection

|
Session A



資料：愛評網 (ipeen)
早午餐快搜

<http://www.ipeen.com.tw/>

要抓什麼資料? - iPeen 早午餐搜尋榜

1

1. Bon tree 好樹

電話 : 02-2737-3082
地址 : 台北市大安區樂業街8號 [MAP](#)

★★★★★ (15評)
早餐 / 西式早餐
本店均消 91 元

店家資訊

Bon tree 好樹是一個有早餐、西式早餐、好樹的西式早餐，網友認為值得推薦的有：薯餅塔、心有所薯米勒千層蛋餅、地瓜金沙蛋餅、總匯蛋餅

閃 250元美食折抵券，現金特價 175 元
在店家使用，結帳享有250元消費折抵

立即閃購

1. 阜杭豆漿店

電話 : 02-2392-2175
地址 : 台北市中正區忠孝東路一段108號2樓 [MAP](#)

★★★★★ (365評)
早餐 / 中式早餐
本店均消 79 元

2. 好初早餐

電話 : 02-2253-2087
地址 : 新北市板橋區文化路二段125巷70號 [MAP](#)

★★★★★ (160評)
早餐 / 其他類型早餐

目標網頁架構:

1. 搜尋頁
2. 商店頁
3. 評論文

選擇頻道

« 全部頻道

« 美食

« 美食店家

早餐

其他類型早餐(536)

中式早餐(919)

西式早餐(2011)

早午餐(210)

選擇地區

« 全部地區

台灣(3608)

台灣以外(74)

平均價位

99元以下

100元~299元

300元~499元

500元~699元

700元~999元



阜杭豆漿店

中式早餐 \$ 本店均消 79 元 ☎ 02-2392-2175 ◉ 台北市中正區忠孝東路一段108號2樓

綜合評分 ★★★★★ (共 365 人評分) | 我的評分 ★★★★★ | 共 1,209,925 人瀏覽 | 2,501 人收藏

營業時間 今日 05:30~12:30 顯示全部 | 補充說明 暫無提供

首頁

照片

分享文 (232)

媒體報導

優惠/KoKo回饋金 ▾

發表分享文

評分

收藏



1

G+1

分

推文

綜合評分詳細資訊

美味度	★★★★★
服務品質	★★★★★
環境氣氛	★★★★★

我的評分詳細資訊

美味度	★★★★★
服務品質	★★★★★
環境氣氛	★★★★★

店家資訊報錯/更新

加入行程

我是老闆，上傳店家照片



今年重新裝潢開幕的阜杭豆漿店面明亮又大氣，盡皆有人...
by 米絲肌

› 滑覽照片集

我是老闆，編輯簡介

2號出口

02-2312-2066

台北市中正區中華路一段59-23號

刊登廣告

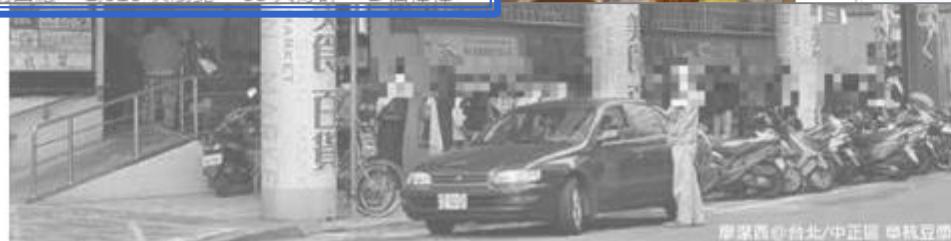


會員分享文 (232)



潔西麻的美食日記—【台北/中正區】阜杭豆漿

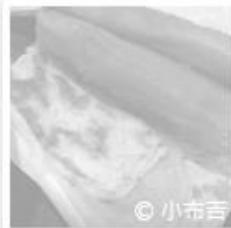
發表於 2016/12/23 4 則回應 1,616 次瀏覽 35 人好評 2 個檯棒



潔西麻 © 台北/中正區 阜杭豆漿

阜杭豆漿是位於台北華山市場二樓的傳統早餐店
遠近馳名
很多觀光客都會慕名而來

FELISSIMO



阜杭豆漿店

中式早餐 本店均價 79 元 | 02-2392-2175 | 台北市中正區忠孝東路一段108號2樓

綜合評分 ★★★★★ (共 365 人評分) | 我的評分 ★★★★★ | 共 1,209,925 人瀏覽 | 2,501 人收藏

營業時間 今日 05:30~12:30 | 顯示全部 | 補充說明 | 預約提供

首頁

照片

分享文 (232)

媒體報導

優惠/KoKo回饋金

發表分享文

評分

收藏

3

綜合評分詳細資訊

美味度 ★★★★★

服務品質 ★★★★★

環境氣氛 ★★★★★

我的評分詳細資訊

美味度 ★★★★★

服務品質 ★★★★★

環境氣氛 ★★★★★

店家資訊報錯/更新

加入行程

會員分享文 (232)



潔西麻的美食日
發表於 2016/12/23

我是



潔西麻的美食日記-【台北/中正區】阜杭豆漿 實用

商家名稱：阜杭豆漿店

發表日期：2016-12-23

09:09:57

評分：★★★★★

同步發表：<http://juicybaby0068.pixnet.net/blog/post/3381...>

美味度：很好 服務品質：很滿意 環境氣氛：好

潔西貝比的生活小
事

貢士三級 | 文 (258)

0

G+1

分

推文



阜杭豆漿是位於台北華山市場二樓的傳統早餐店
遠近馳名

很多觀光客都會慕名而來

你以為資料都會乖乖的整理好放在
你面前?!!



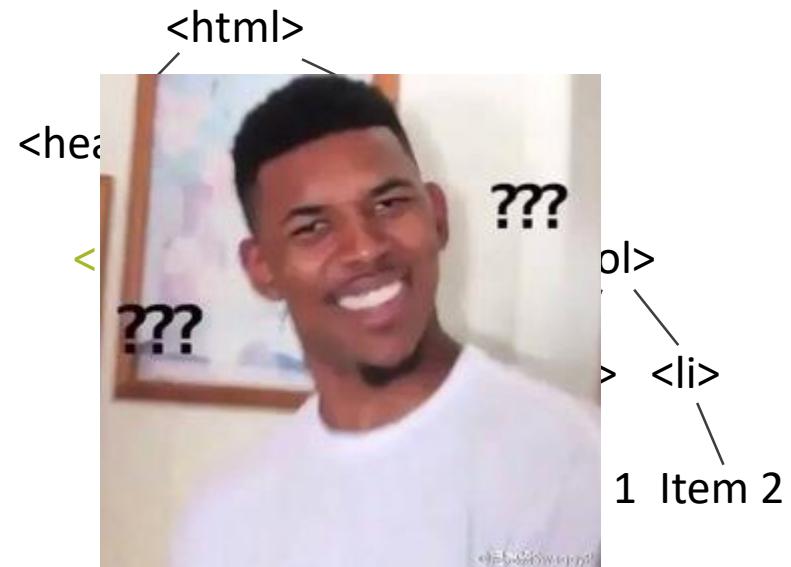
要做爬蟲前，
先來簡單瞭解一下網頁架構吧！



網頁架構及語法

<標籤 屬性> 內容 </標籤>

```
<html>
  <head>
  </head>
  <body>
    <h1 id="title"> Title </h1>
    <p> Paragraph </p>
    <ol>
      <li> Item 1 </li>
      <li> Item 2 </li>
    </ol>
  </body>
</html>
```



打開網頁, 按下你的F12 - 利用瀏覽器檢查原始碼

依： 預設 | 綜合評價 | 分享文數 | 平均價位 排序

蒂兒廚房(左營店)

電話：07-345-1292
地址：高雄市左營區文天路123號 [MAP](#)

 [店家資訊](#)

只提供健康新鮮美味的餐點,是我們的堅持 百分百手做牛排堡+打拋豬 蛋牛奶可是必點唷

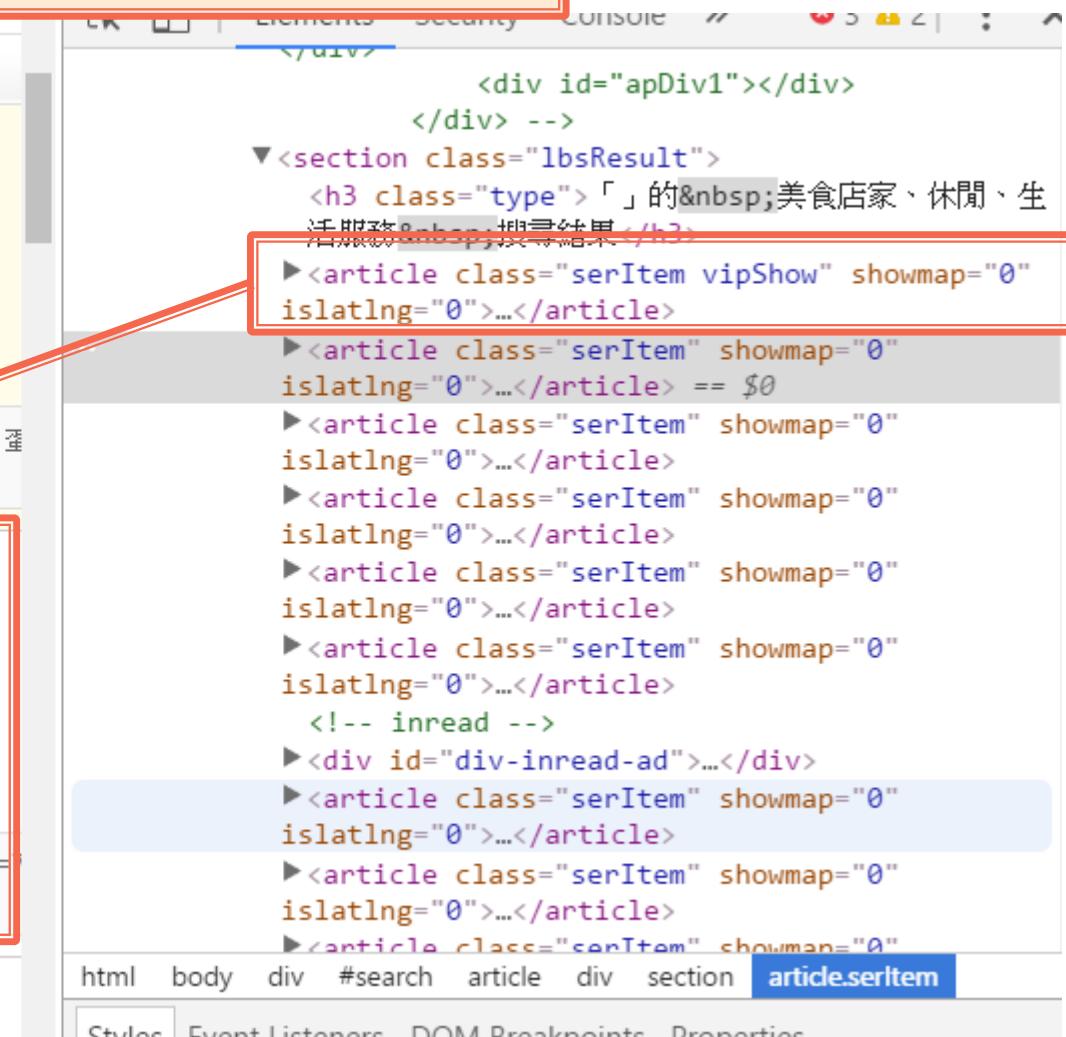
1. 豐盛號

電話：02-2880-1388
地址：台北市士林區中正路223巷4號1樓 [MAP](#)

 [店家資訊](#)

【媒體推薦-台北十大必吃早餐】 <https://www.youtube.com/watch?v=1>
豐盛號開始..... [\(繼續閱讀\)](#)

2. 豐滿咖啡早午餐



The screenshot shows the browser's developer tools open to the 'Elements' tab. A red box highlights a specific article element in the DOM tree, which corresponds to the listing for '豐盛號' (Fong Sheng Hao). An arrow points from the highlighted text in the main content area to this element in the DOM tree. The DOM tree also includes other sections like 'lbsResult' and various 'article' elements for other food items.

```
<div id="apDiv1"></div>
</div> -->
▼<section class="lbsResult">
  <h3 class="type">「」的美食店家、休閒、生活敗家&nbsp;地圖結果</h3>
  ▶<article class="serItem vipShow" showmap="0" islatlng="0">...</article>
  ▶<article class="serItem" showmap="0" islatlng="0">...</article> == $0
  ▶<article class="serItem" showmap="0" islatlng="0">...</article>
  ▶<article class="serItem" showmap="0" islatlng="0">...</article>
  ▶<article class="serItem" showmap="0" islatlng="0">...</article>
  ▶<article class="serItem" showmap="0" islatlng="0">...</article>
  <!-- inread -->
  ▶<div id="div-inread-ad">...</div>
  ▶<article class="serItem" showmap="0" islatlng="0">...</article>
  ▶<article class="serItem" showmap="0" islatlng="0">...</article>
  ▶<article class="serItem" showmap="0" islatlng="0">...</article>
  ▶<article class="conItem" showmap="0" islatlng="0">...</article>
```

html body div #search article div section **article.serItem**

Styles Event listeners DOM Breakpoints Properties

邊滑邊抓網頁資訊 (R.pkg xml2)

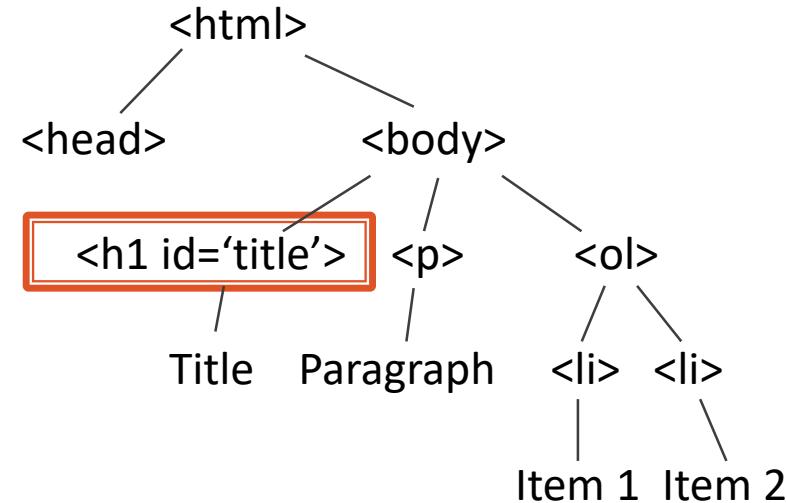
□ xml2

- 讀取網頁 : `read_html`; `read_xml`
- 選擇節點 : `xml_find_all`
- 撿取資訊 : `xml_text`; `xml_attr`

<標籤 屬性> 內容 </標籤>

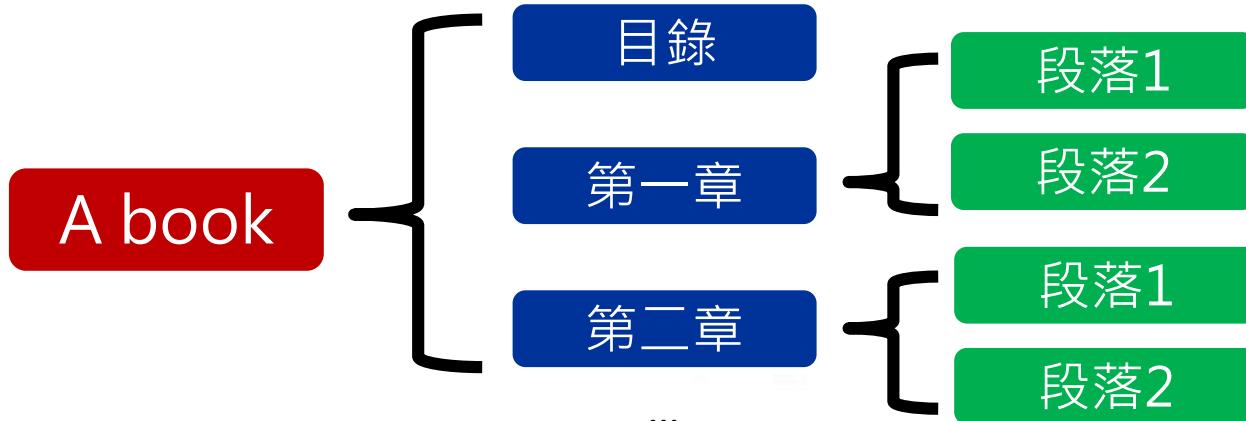
如何找到需要的節點 ?

Xpath



XPath: a path of XML Tree

標記	意義
/	選取某標籤的根節點 (children)
//	選取某標籤所有節點 (descendant)
[@]	選取屬性 (attribute)
*	任何標籤的節點
	OR



Xpath = “//div[@class="serShop"]/h3/a”

1. 豐盛號



電話 : 02-2880-1388
地址 : 台北市士林區中正路223巷4號1樓 [MAP](#)

[店家資訊](#)

【媒體推薦-台北十大必吃早餐】 <https://www.youtube.com/watch?v=...>
豐盛號開始.....[\(繼續閱讀\)](#)

2. 豐滿咖啡早午餐



電話 : 0952-258-700
地址 : 新北市板橋區雨農路18巷1號 [MAP](#)

3. Ona.Rina Icing 歐娜。瑞娜 早午餐



電話 : 02-2256-7314
地址 : 新北市板橋區文化路二段125巷11號 [MAP](#)

[店家資訊](#)

▼<article class="serItem" showmap="0" islatlng="0">
▼<div class="serShop" id="shop_row_613262">
▼<h3 id="shop_h3_1" class="name">
" 1.

" " " "

豐盛號 == \$0
</h3>
►<div class="serPic">...</div>
►<div class="serData">...</div>
</div>
►<div class="special">...</div>

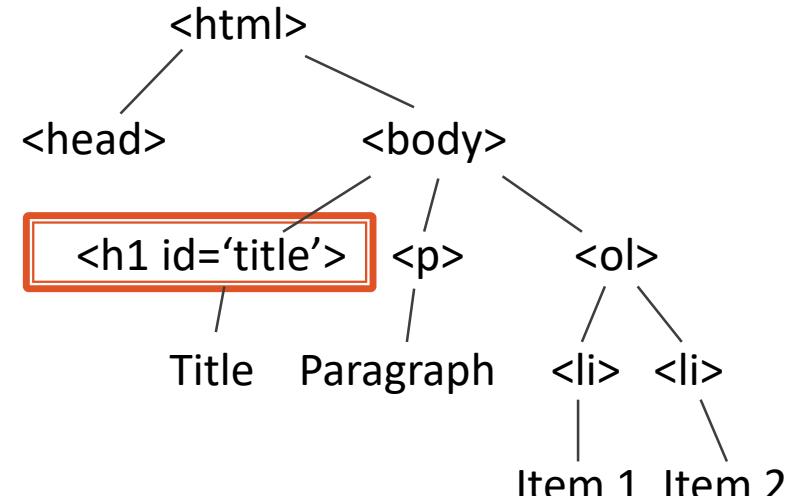
本店均消 200 元



擷取資訊

- `xml_text(doc): <標籤 屬性> 內容 </標籤>`
- `xml_attr(doc, attr): <標籤 屬性> 內容 </標籤>`

```
<html>
  <head>
  </head>
  <body>
    <h1 id="title"> Title </h1>
    <p> Paragraph </p>
    <ol>
      <li> Item 1 </li>
      <li> Item 2 </li>
    </ol>
  </body>
</html>
```



邊看邊做時間



- 開啟 00_getData.R

- 幾個小重點
 - 換頁怎麼換
 - 是不是 unique tag
 - 你要的東西是 text 還是 attribute

還有更多資料呢？



- 每頁的細節資料？
- 每家的評論內容？

- 有興趣的話，我把程式碼都放在那了！想要就去研究吧！
 - `00_getShop_details.R`



練習時間 (optional)

- 電影票房資訊
- <http://taipeibo.com/weekly/>

- 把每天每部電影的票房訊息抓下來做成 data frame 吧



小結

- 學會怎麼看網頁架構
- 了解 Xpath 運作方式, 透過 F12 鎖定目標與抓取資料
- 這是靜態網頁 ..., 動態網頁需要其他小技巧與技能
 - 建議不用執著用 R 做爬蟲拉 ..., Python 很好用的

我們現在手上有何資料

- df_all_shop.csv

- 每間店的基本訊息 (我們撈了前 20 頁)

- df_shopDetails.rds

- 每間店的細節資訊 (這 320 間店的對應資料)

- df_review.rds

- 每間店的評論者資訊

- ipeen_txt/...txt

- 每間店的被評論文章

探索式資料分析
Explanatory Data Analysis



Session B



開啟你的 RMD, 邊做邊看!

□ 重點

- 會用 `data.table` 這個 package (WHY?)
 - 快
 - 很快
 - 非常快
 - 還滿好懂 (跟 SQL 語法相似)
- 會用 `ggplot` 來畫圖 (WHY?)
 - 架構清晰
 - 美美 der
- 能從原來的資料中, 創造更多變數(特徵)

Summary Functions in R

Function	Description	白話文
names()	Functions to get or set the names of an object	看欄位名稱
head(), tail()	Returns the first or last parts of a vector, matrix, table, data frame or function	看前/(後) 幾筆資料
str()	Compactly display the internal structure of an R object	物件屬性
summary()	Produce result summaries	物件的基本數值狀態
dim()	Retrieve or set the dimension of an object	矩陣大小
length()	Get or set the length of vectors	向量長度
complete.cases()	Return a logical vector indicating which cases are complete, i.e., have no missing values	回傳各元素 NA 邏輯值
as.Date()	Convert between character representations and objects of class "Date" representing calendar dates	轉成日期型態

Function name and parameter 的縮寫解釋：

<http://jeromyanglim.blogspot.tw/2010/05/abbreviations-of-r-commands-explained.html>

Handy function in data.table

□ 讀檔

□ fread

- 別懷疑, 它就是比 read.csv / read.table 快

□ 基本操作概念

□ data.table(i, j, by)

- i: 條件 (拿來篩選要保留/丟掉哪些 row)
- j: 操作 (對欄位操作)
- k: grouping (根據哪個 index grouping 後, 進行 j 的操作)

Concept of ggplot

- 用 + 做層次疊加
- aes (aesthetics): 動態選擇變數
- geom_xxx: 圖樣
 - geom_point
 - geom_line
 - geom_bar
 - geom_tile
 - ...
- stat_xxx: function
 - stat_smooth
 - stat_ecdf
 - ...

大致的架構

```
ggplot +  
geom +  
stat +  
scale +  
labs(x = ..., y = ...) +  
theme +  
ggtitle +  
coord
```



我們也來簡單的來跑個模型

□ Model

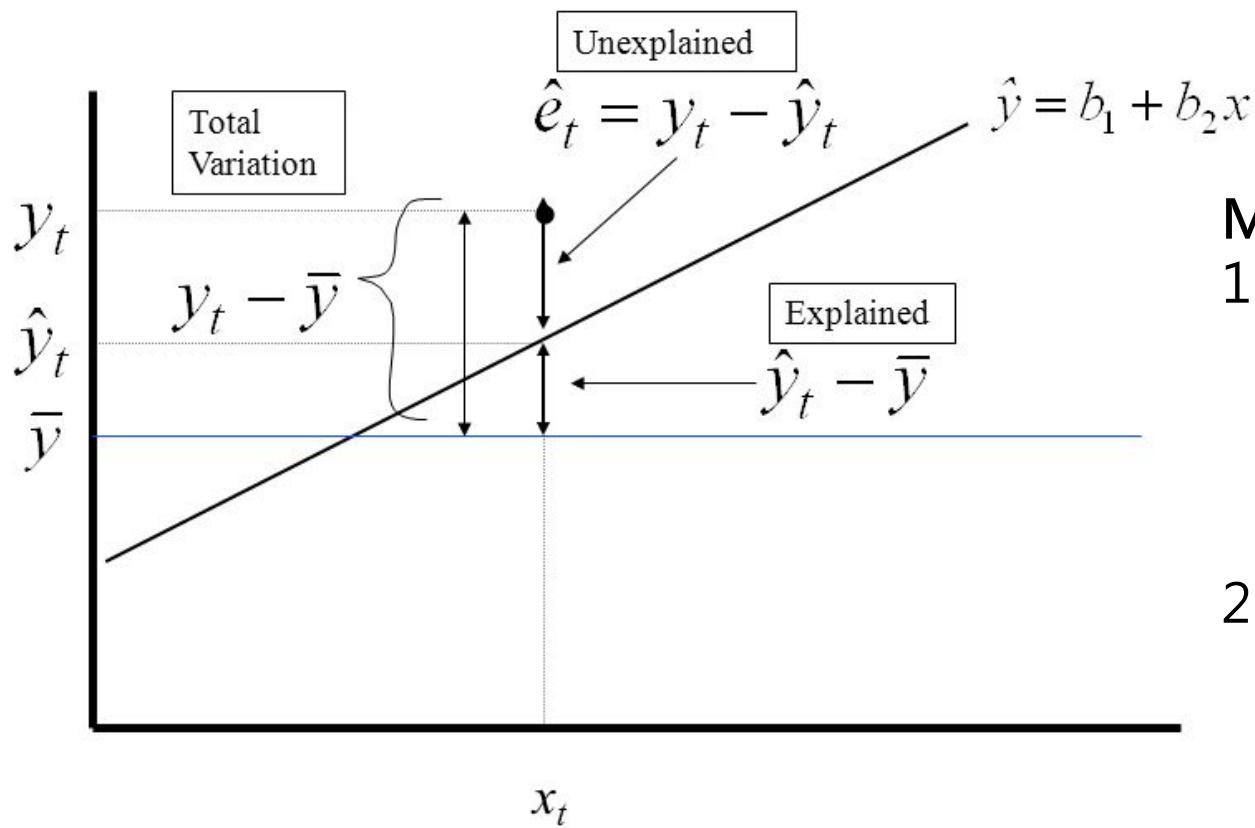
- Linear
- SVM (Support Vector Machine)
- RandomForest

□ How to do?

線性迴歸模型 (Multiple regression model)

目標:

在空間中找到一條線, 讓誤差項總和最小化



Model evaluation:

1. R-square (解釋量)
 $= SS_{\text{reg}}/SS_{\text{tot}}$

$$2. \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n SSE_i^2}$$

支持向量機 (support vector machine)

目標:

在空間中找到一條向量, 把海劈開, 讓兩群樣本分開



		預測值	
		1	0
實際值	1	TP	FN
	0	FP	TN

如何驗證

□ Continuous Variable

□ Pearson correlation coefficient = $\frac{cov(X,Y)}{\sigma_x \sigma_y}$

□ Coefficient of determination (R^2) = $\frac{SSreg}{SStot}$

□ RMSE (Root mean square error) = $\sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}}$

□ Categorical Variable

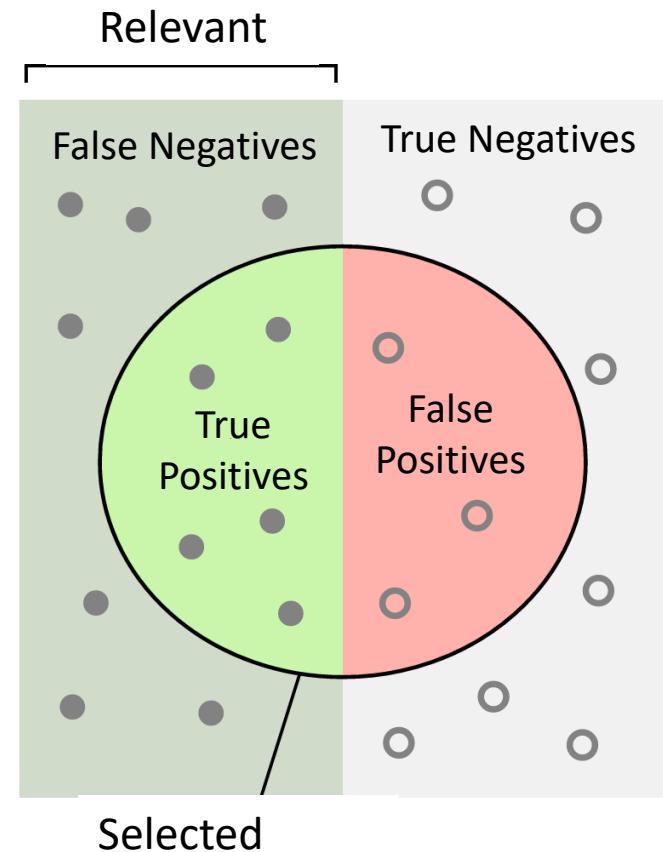
□ Accuracy = $\frac{TP + TN}{P + N}$

□ F1-score = $2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$

Diagnostic Testing

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

		預測值	
		1	0
實際值	1	TP	FN
	0	FP	TN



$$F1\text{-score} = \frac{2TP}{2TP+FP+FN}$$

		預測值	
		1	0
實際值	1	TP	FN
	0	FP	TN

Precision



Recall





跑模型的 take home message

- 定好公式
- 要切 Training 和 Testing data
 - 8:2, 9:1 都是常見的切法
 - leave-1-out 也是一招
 - k-fold cross validation (這個我們剛才沒講)

複習一下到此為止的重點

- 會用 `data.table` 與 `ggplot`
- 能檢視資料, 移除重複值, 合併不同的 `table`
- 創造新變數 (特徵)
- 視覺化
- 跑模型與驗證結果好壞

基礎文字探勘

Session C



非結構化資料 – 文字

□ 文字可以用做什麼分析？

- 創造新的變數
- 做探索性分析，比方說
 - 關鍵字文字雲
 - 詞與特並變數的關聯性
 - ...
- 語意分析、情緒分析
- 聊天機器人
- ...

周末兩隻懶惰鬼睡到了十點半，賴床摸魚摸一摸就十一點多，就決定早午餐一起吃啦~

對豐滿澎湃的擺盤念念不忘的我，就決定蒂阿亞來二訪，
上回是10點的早餐時間，這回是12點的午飯時間。
然而不變的是，這間來自板橋的連鎖早午餐店無時無刻都在排隊啊! >A<

就連咱們酒足飯飽走人之際外面還是坐著不少排隊的人們。



來了來了，擺盤超澎湃的皇后拼盤!!

五彩繽紛又豐富的拼盤端到桌上的那一刻，真是忍不住哇出來耶>//<然後嘻擦嘻擦開始拍照XD

最左邊的就是單點菜單中的"綜合梅果脆片法式吐司 \$75"，
法式吐司香軟鬆甜很好吃，脆片的話因人而異，因為我不愛吃硬的所以我上次沒吃完，而且整份吃完好飽。

但梅果....我們都一樣把它留了下來QQ 因為梅果妝點擺盤很美，
可是真的太酸了....
有德式香腸，起司火腿，兩種吃起來都很夠味~
薯條+蜂蜜芥末醬，還有飯後甜點柳丁。

雖然餐點有些是視覺藝術營造的效果，例如都切成對半讓東西看起來加倍，不過份量是男生也會有飽足感的唷!

不愧是連鎖店一家一家開的知名店家，CP值有高! OuO~
要再來的話，大約是衝著美麗的擺盤、大份量的誘惑再訪的機率較高，
因美味度算中上，不過和一般早午餐的口味相去不遠，
不過三五好友一道嘗鮮是滿推薦的唷~^^

P.S 新莊另有中信店及後港店



第一次來的時候點了王子拼盤，因為真的太想太想吃德式香腸了!!
但因為又想吃法式吐司，所以又把麵包換成法式，但其實這樣長得跟皇后拼盤就差不多啦~

仔細看的話會發現每種都只有一些小差異，但都能依照喜好再做調整，
例如也有朋友將嫩蛋換做太陽蛋，這點很貼心!

然後那次也對薄餅超有興趣，掙扎著選了拼盤，不過這回有阿亞的可以搶劫分享，
我就無後顧之憂的選了墨西哥薄餅囉~~(轉圈圈



歡迎光臨，優醬的部落格.....好吃不藏私♥

需要的材料？

□ 文本: data\ipeen_txt\ooo.txt



IM Home除了是Ivan & Mike一起創辦的餐廳外，亦都

來自香港的我們會選擇台北成為起步點，當中沒有什麼偉大的夢想：
命更成長。

不斷重新學習及適應，令人生更豐實，或許這就是我們，亦都

但我們更相信，當我們奉獻了青春換取更多
有時倦了、累了、想分享、想獨樂樂…，總要有個家作為
故此……歡迎來到我們的家。

From

IM Home

以上故事取自於IM HOME粉絲團

【IM home】
交通便利，
口味上比較
若是你還在
現在還有F
安排個時間



如果只是路過我會覺得它是一間小酒吧

東區的義式料理很多有名的也吃過好幾家
Mei姐的朋友推薦去年12月開始試營運的店

上網看了一下他們粉絲團發現是香港人開的
原本以為是港式飲茶結果居然是義式料理耶!!



平常都在東區走跳的瑞塔

發現了一家超質感的Brunch餐廳要跟大家

這間位於忠孝復興2號出口的『IM HOME』

就在BR4 sogo復興館旁邊

去年12月才開始營業

目前還在試營運階段

以後也會再陸續新增菜單呀調整營業時間

一大面簡約白牆配上大紅色的大門十分搶眼

再看到地面上的貓頭鷹咕咕

就表示你已經抵達餐廳啦~XDDD

『IM HOME』表示著兩位香港創辦人Ivan & Mike

兩人脫離舒適圈來到台灣築夢

也代表著 I'm Home的意思

相當溫暖溫馨的小天地 ♥

我們一開店10:00就跑來用餐你看Mei姊有多累XD





好問題 - 這麼多文章怎麼讀進R

□ 讀檔的步驟有 ...

1. 列出有哪些檔案
2. 讀檔

□ 寫迴圈一個一個讀 ...?

□ 還是有更好的做法

□ 打開 ipeen_review_text.Rmd



關於創造變數

□ 暴力法

□ 智慧解



暴力法 - grep

- 窮舉你想到的詞吧!
 - 比如
 - `grep(pattern = "難吃", x = article_txt[i])`
- 玩不完的 ...

認真分析文字的第一步：斷字

□ 中文好難



□ 中文需要斷詞

- 每個英文單字 (word / term) 都用空格分開
- 下雨天留客，天留我不留

文章斷詞

□ jiebaR <https://qinwenfeng.com/jiebaR/>

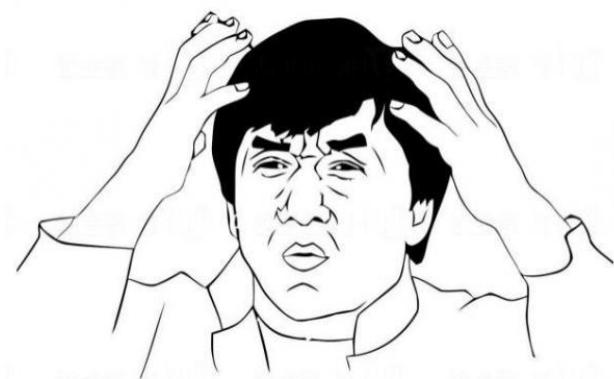
- 號稱最好的 Python 中文斷詞組件的 R 語言版本
- 支持四種斷詞引擎
 - 最大概率法、隱式馬爾科夫模型、混和模型、索引模型
- 可以標註詞性

□ 中研院斷詞系統 <http://ckipsvr.iis.sinica.edu.tw/>

- 號稱地表最強中文斷詞系統 (96% 精準度)
- 自動標註詞性
- 需要申請.....

斷完詞，之後呢

- 斷完詞，電腦還是看不懂
- 建立詞庫：記得所有看過的詞
- 飽讀詩書，瞭解詞義



詞庫、詞義

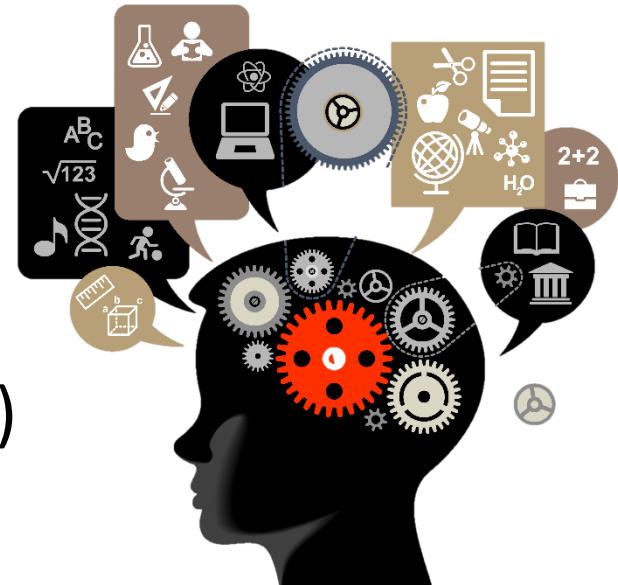
□ _____是最炎熱的季節

- 季節：春季、夏季、秋季、冬季
- 哪個比較適合？

□ 詞與詞之間的距離

- 同義字：溫拿 \leftrightarrow 人生勝利組
- 反義字：溫拿 \leftrightarrow 魯蛇

□ 詞的特徵化 (word embeddings)



建立詞庫

- 詞庫：所有看過的字詞的集合，通常用 V 表示
 - V : vocabulary
- 詞庫內的詞可以用 $1 \times |V|$ 的向量表示
 - 稱為 one-hot vector 可視為索引
 - 彼此獨立，沒有詞意
- 利用 jiebaR 斷詞和 text2vec 建立詞庫吧！

1	0
2	0
:	0
:	0
k	1
k+1	0
:	0
:	0
V	0

$1 \times |V|$



教電腦從文章中解讀詞意

□ word2vec

- Tomas Mikolov et.al 2013 年於 Google 開發
- Prediction-based method [reference](#)

□ Glove

- Jeffrey Pennington et al. 2015 年開發 (Stanford)
- Count-based method [reference](#)
- Dmitriy Selivanov 開發其 R 套件 text2vec
- <https://cran.r-project.org/web/packages/text2vec/index.html>

淺談 word2vec

- 語意相近的字較常出現在一起

- Local window

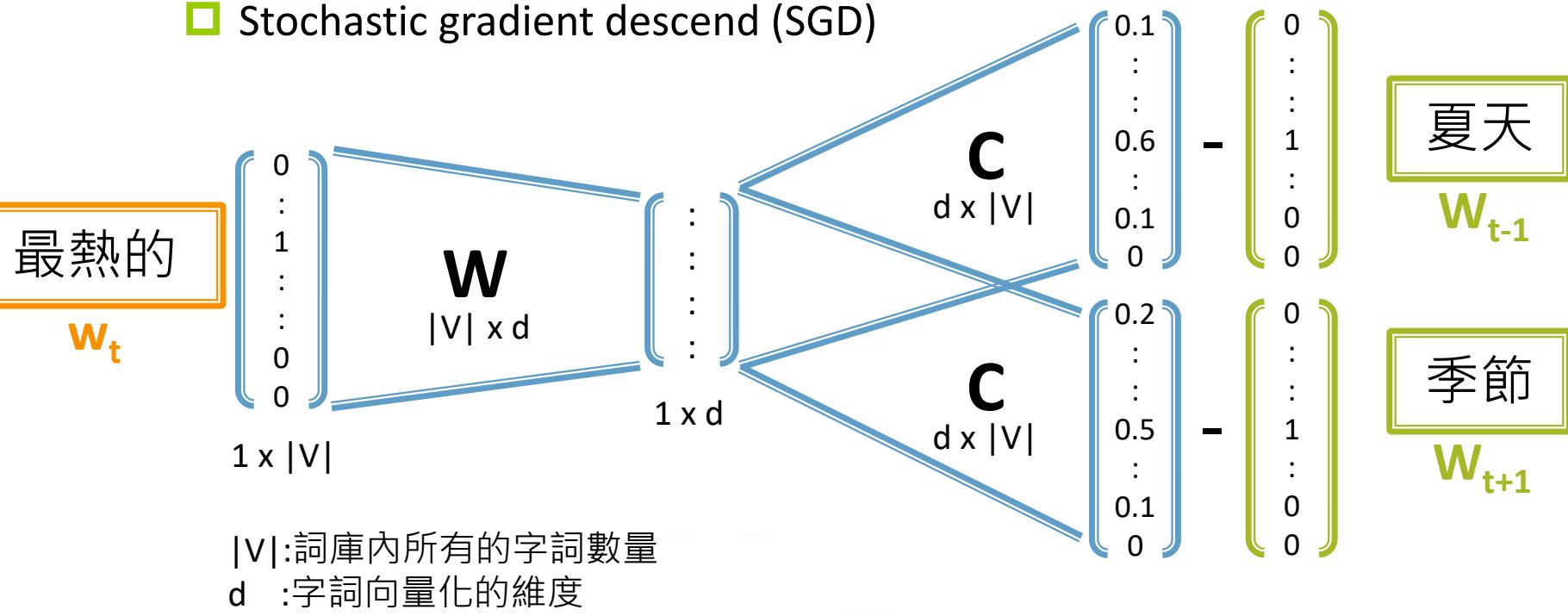


- 讓電腦玩克漏字填空學習 word embedding

- Skip-gram
 - Continuous bag of words

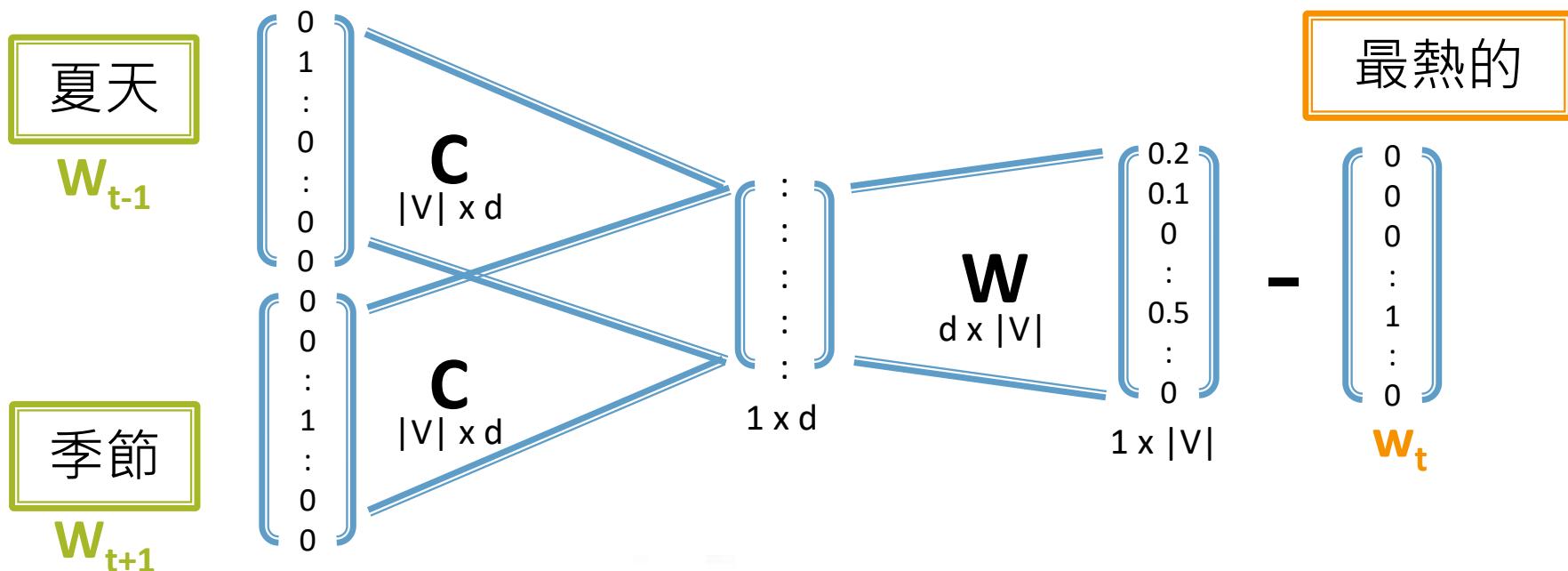
Skip-gram 模型

- 藉由 current word 推測 context words
- Neural network model
 - Stochastic gradient descend (SGD)



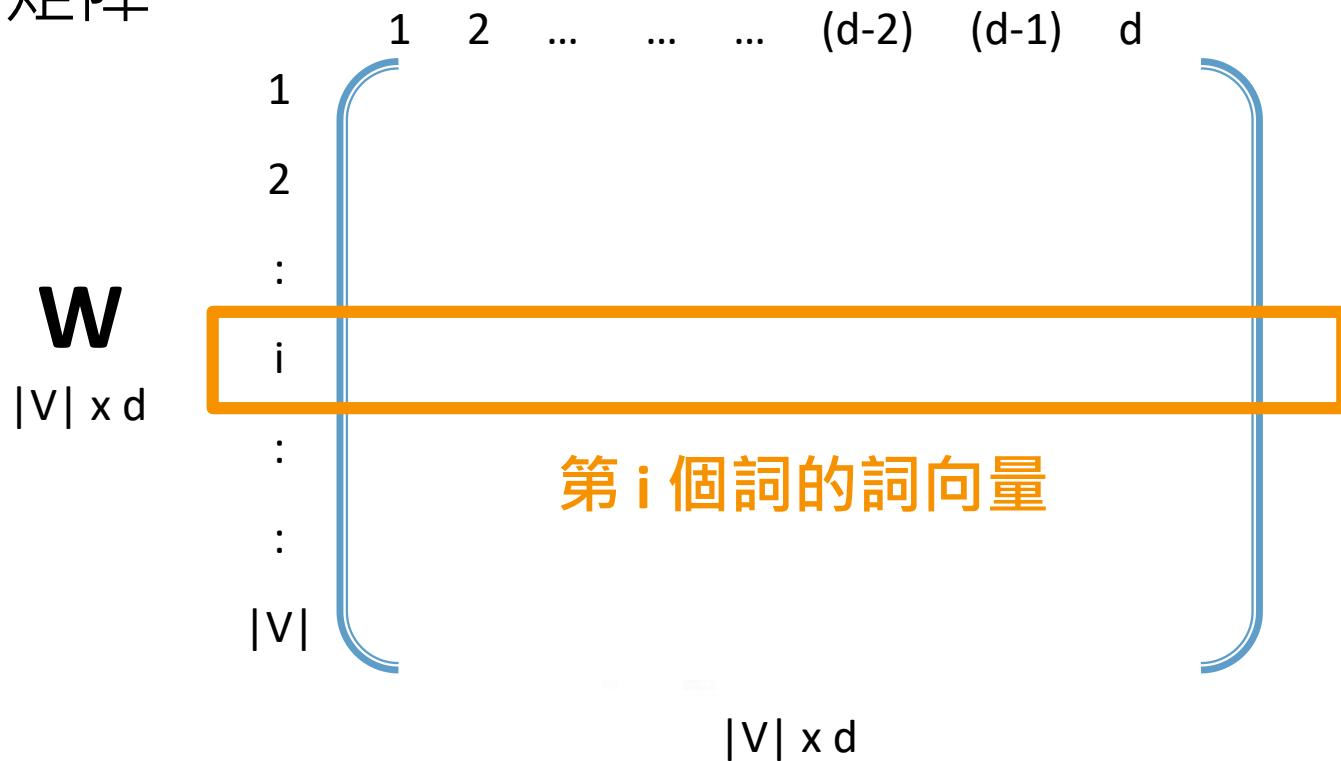
Continuous Bag of Words 模型

- 由 context words 推測 current word



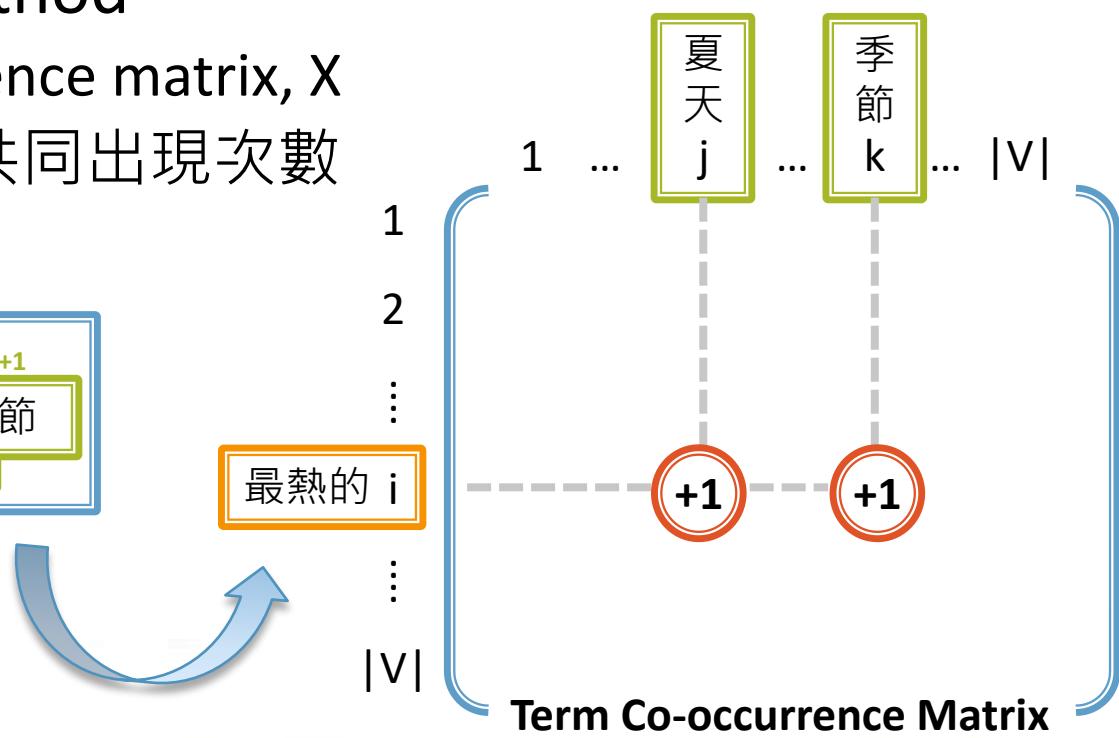
詞向量

- 訓練結束後，將 $1 \times |V|$ 字詞轉換成 d 維向量的矩陣



淺談 text2vec

- Global corpus statistics + local window
- Count-based method
 - Term co-occurrence matrix, X
 - X_{ij} : 詞-i 和 詞-j 共同出現次數



GloVe 模型

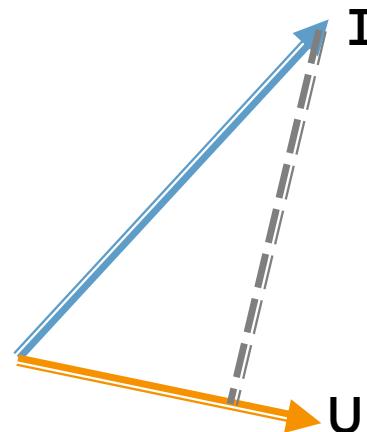
- 目標：訓練出一個詞矩陣 \mathbf{W}
最能表示 term co-occurrence statistics
- 使用 AdaGrad, a variant of stochastic gradient descend

$$\left[\begin{array}{c} \mathbf{W} \\ |V| \times d \end{array} \right]$$

1. 對於一組詞向量 $\vec{w_i}$ 和 $\vec{w_j}$ ，我們希望： $\vec{w_i}^T \vec{w_j} + b_i + b_j = \log X_{ij}$
2. 目標函數：最小化 $J = \sum_{i=1}^V \sum_{j=1}^V f(X_{ij})(\vec{w_i}^T \vec{w_j} + b_i + b_j - \log X_{ij})^2$
3. $f(X_{ij})$ 為加權函數，減少常見詞：
$$f(X_{ij}) = \begin{cases} \left(\frac{X_{ij}}{x_{max}}\right)^\alpha, & X_{ij} < x_{max} \\ 1, & otherwise \end{cases}$$

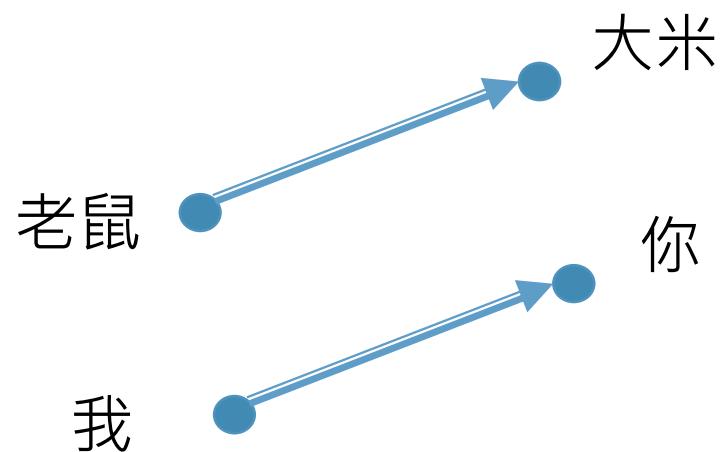
神奇的詞向量

- 計算詞之間的距離
 - 同義、反義
 - Cosine-similarity



在你(U)身上看見部份的自己(I)

- 計算字詞間的關係
 - 老鼠跟大米



找出你我之間的關係



文詞向量化到底可以幹嘛！



文詞向量化到底可以幹嘛!

- 創造新變數!
 - 找出相近的詞意的字, 然後把它們撈出來
 - 把這些向量分群當成新特徵

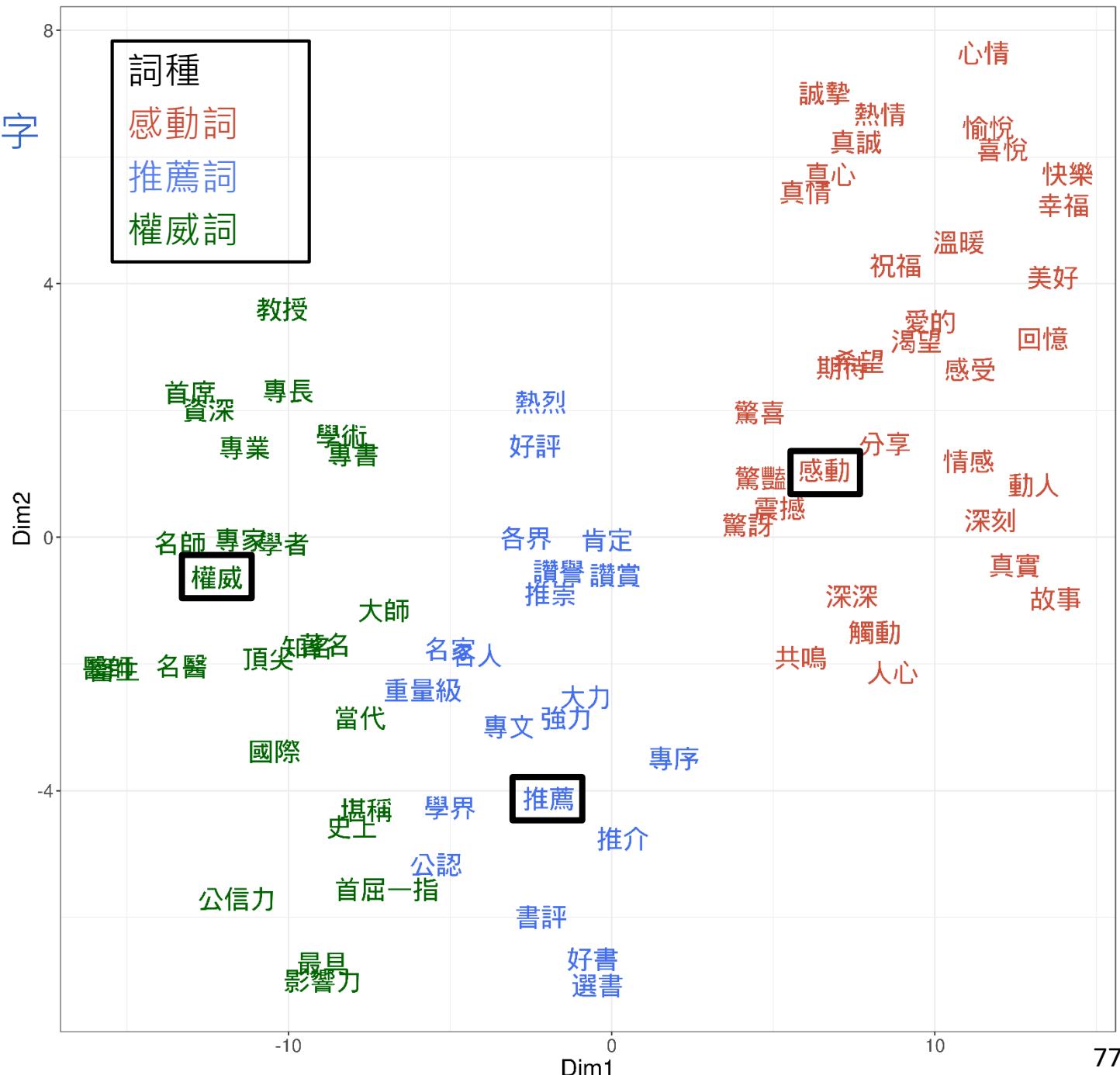


文詞向量化到底可以幹嘛!

- 創造新變數!
 - 找出相近的詞意的字, 然後把它們撈出來
 - 把這些向量分群當成新特徵

Case 找出相近詞意的字

t-SNE on Word2Vec



Case 詞向量化後的 kmean 分群 (蘋果日報公益)

donor.de vs k133 (死亡用詞), cor = 0.203

```
> w[w$k == 133,]$v1
[1] "走(vi)" "病逝(vi)" "逝(vi)" "留下(vt)" "死(vi)" "去世(vi)" "生前(N)" "喪葬(N)" "喪葬費(N)" "往生(vi)"
[11] "猝(ADV)" "交代(vt)" "後事(N)" "留(vt)" "驟(ADV)" "心肌(N)" "梗塞(vi)" "來不及(ADV)" "身亡(vi)" "悲傷(vi)"
[21] "喪夫(vi)" "喪事(N)" "遺照(N)" "送走(vt)" "突然(vi)" "出殯(vi)" "辦妥(vt)" "靈堂(N)" "猝死(vi)" "喪子(vi)"
[31] "程(N)" "國民(N)" "不敵(vt)" "白髮人(N)" "黑髮人(N)" "臨終(vi)" "病故(vi)" "含淚(vi)" "溺斃(vi)" "辦好(vt)"
[41] "撒手(vi)" "遺體(N)" "辭世(vi)" "老幼(N)" "人世(N)" "辦完(vt)" "斷氣(vi)" "殯儀館(N)" "冰櫃(N)" "悲慟(vi)"
[51] "桂圓(N)" "奠儀(N)" "睡夢(N)" "繼(P)" "葬儀社(N)" "離世(vi)" "紙錢(N)" "再見(vi)" "黑髮(N)" "平復(vt)"
[61] "小晟(N)" "喪禮(N)" "林美琴(N)" "遺言(N)" "棺木(N)" "葬(vt)" "積勞成疾(vi)" "遽逝(vi)" "享年(vt)" "陳芷涵(N)"
[71] "孤(vi)" "安葬(vt)" "永隔(N)" "杜氏(N)" "火化(vt)" "薛(N)" "葉榮進(N)" "含悲(vi)" "悲痛(vi)" "遺書(N)"
[81] "處理完(vt)" "周麗珠(N)" "江鎧安(N)" "莫惠萍(N)" "自焚(vi)" "莊台(N)" "阿治(N)" "李瑪美(N)" "忠正(N)" "悲(vi)"
```

donor.de vs k180 (食物用詞), cor = 0.1008

```
> w[w$k == 180,]$v1
[1] "吃(vt)" "餐(M)" "飯(N)" "晚餐(N)" "煮(vt)" "便當(N)" "青菜(N)" "餓(vi)" "省錢(vi)" "配(vt)"
[11] "罐頭(N)" "飯菜(N)" "泡麵(N)" "早餐(N)" "稀飯(N)" "炒(vt)" "熱(vt)" "白飯(N)" "麵(N)" "肉(N)"
[21] "餓肚子(vi)" "飽(vi)" "包(M)" "鍋(M)" "碗(M)" "拌(vt)" "果腹(vi)" "粥(N)" "碗(N)" "吃飽(vi)"
[31] "瓶(M)" "頓(M)" "好吃(vi)" "吃完(vt)" "一口(ADV)" "蛋(N)" "醬油(N)" "麵包(N)" "湯(N)" "加菜(vi)"
[41] "餐桌(N)" "道(M)" "鍋(N)" "樣(M)" "麵線(N)" "煎(vt)" "剩菜(N)" "饅頭(N)" "弄(vt)" "麵條(N)"
[51] "開水(N)" "澆(vt)" "鹹(vi)" "填飽(vt)" "煮好(vt)" "鹽(N)" "肉鬆(N)" "炒飯(N)" "盤(M)" "蛋炒飯(N)"
[61] "加熱(vi)" "荷包蛋(N)" "苦瓜(N)" "吃到(vt)" "水餃(N)" "營養(vi)" "滷肉(N)" "豆漿(N)" "菜色(N)" "年夜飯(N)"
[71] "放進(vt)" "自助餐(N)" "省下來(vi)" "扒(vt)" "舀(vt)" "吃剩(vi)" "醬瓜(N)" "餓死(vt)" "打發(vt)" "蒸(vt)"
[81] "充飢(vi)" "燙(vt)" "滷(vt)" "菜肉(N)" "匙(M)" "攪拌(vt)" "絲瓜(N)" "會兒(N)" "豆腐(N)" "配上(vt)"
[91] "罐(N)" "飯桌(N)" "包子(N)" "白開水(N)" "津津有味(vi)" "食量(N)" "吐司(N)" "熱騰騰(vi)" "醬菜(N)" "用餐(Nv)"
[101] "填(vt)" "湯汁(N)" "配飯(N)" "吃光(vt)" "大鍋(N)" "菜葉(N)" "湯麵(N)" "午飯(N)" "食欲(N)" "菜湯(N)"
```



開啟你的 RMD, 邊做邊看!

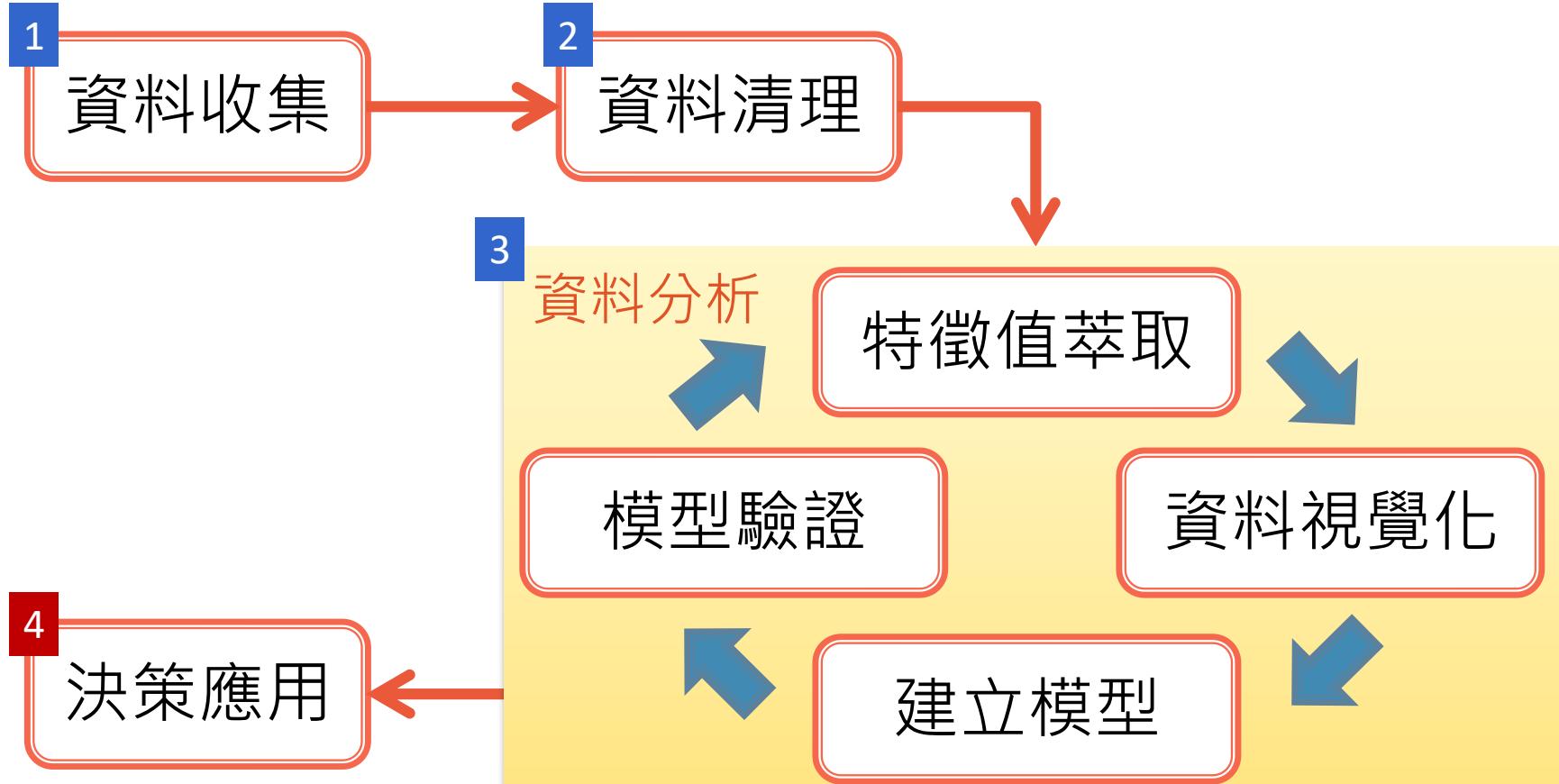
□ 重點

- 忍住你心中小宇宙的迴圈慾望, 快速讀檔法
- 斷字怎麼斷 (注意缺陷)
- 詞的向量化
- 找到相近詞義
- wordcloud



今天學到了什麼？

資料分析流程



還有那些好用的 package 和小技巧啊！



THANK YOU ALL!!

Hard work pays off.