#### 上課前,你可以先...

- •下載課程檔案(一份Rmd與兩份資料檔 [udn\_token\_clean.rds, news\_df.rds])。
- •如果決定題目後,請盡早開始進行。還沒決定的組別,也可以再找我或是助教討論。 ( https://reurl.cc/V5OGkA )
- •檢討作業四(簡報檔、solution已上傳)
- 最後一次的課後回饋調查 (<u>https://forms.gle/3aP9MPpEvsKGmbuh6</u>)。

# R語言 詞袋模型

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

- 文件-詞彙矩陣 ( Document-Term Matrix, **DTM** ) 與詞彙-文 件矩陣 ( Term-Document Matrix, TDM )
- 詞頻-反文件頻率 ( term frequency—inverse document frequency, TF-IDF )
- •中文 bigram 斷詞
- 額外補充:用詞袋模型+羅吉斯迴歸(logistic regression) 預測報導來源

### 文字探勘在生活中的應用

• 怎麼判斷是垃圾郵件?

• 實際案例分析: 兒少保護案件之精準派案

### 詞袋模型

- 詞袋模型就是當斷詞完成後,不考慮用 詞的先後順序,計算每個詞出現的次數 (或是詞頻),並將其轉換成矩陣的方 式儲存。
- 一般最常見的是文件-詞彙矩陣 (Document-Term Matrix, DTM),將不 同的文件放在每一列(row),而每一個 特有的詞彙則是儲存在不同欄位 (column)。

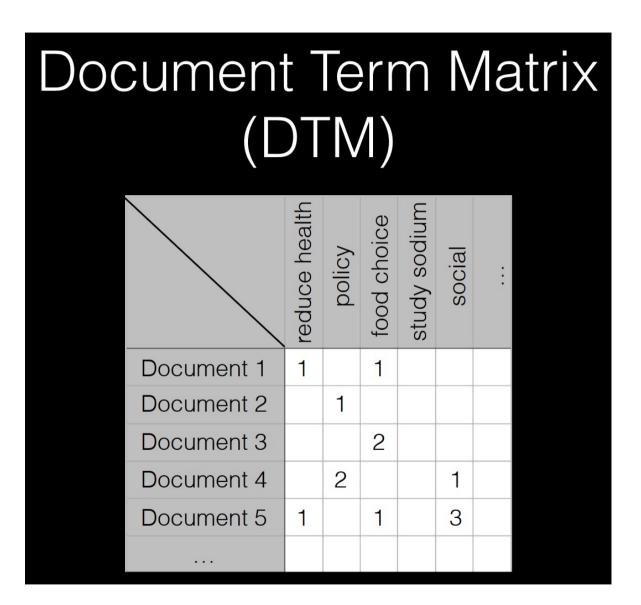


#### 詞袋模型

# the dog is on the table



#### 詞袋模型



#### 詞頻-反文件頻率

(term frequency-inverse document frequency, TF-IDF)

$$w_{x,y} = tf_{x,y} \times log(\frac{N}{df_x})$$

**TF-IDF**Term x within document y

 $tf_{x,y}$  = frequency of x in y  $df_x$  = number of documents containing x N = total number of documents

#### 詞頻-反文件頻率

(term frequency-inverse document frequency, TF-IDF)

• 詞頻:某個詞彙出現在該文件中的頻率

$$tf(term) = n_{tokens} / n_{tokens in that document}$$

• 反文件頻率: 文件頻率的倒數, 再取自然對數

$$idf(term) = ln(n_{documents} / n_{documents \, containing \, term})$$

• 詞頻 × 反文件頻率 = TF-IDF