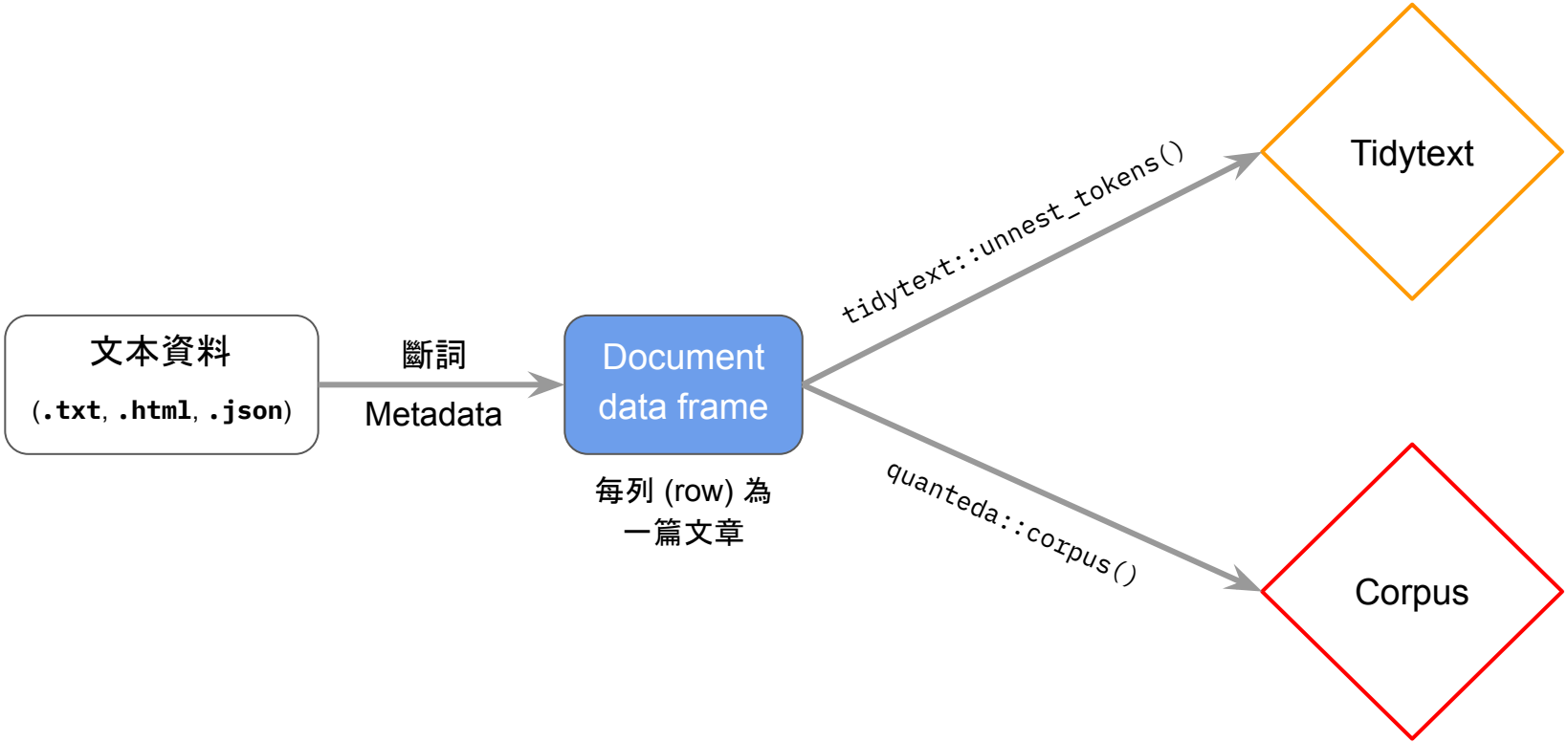
Lab 08

中文文本資料處理

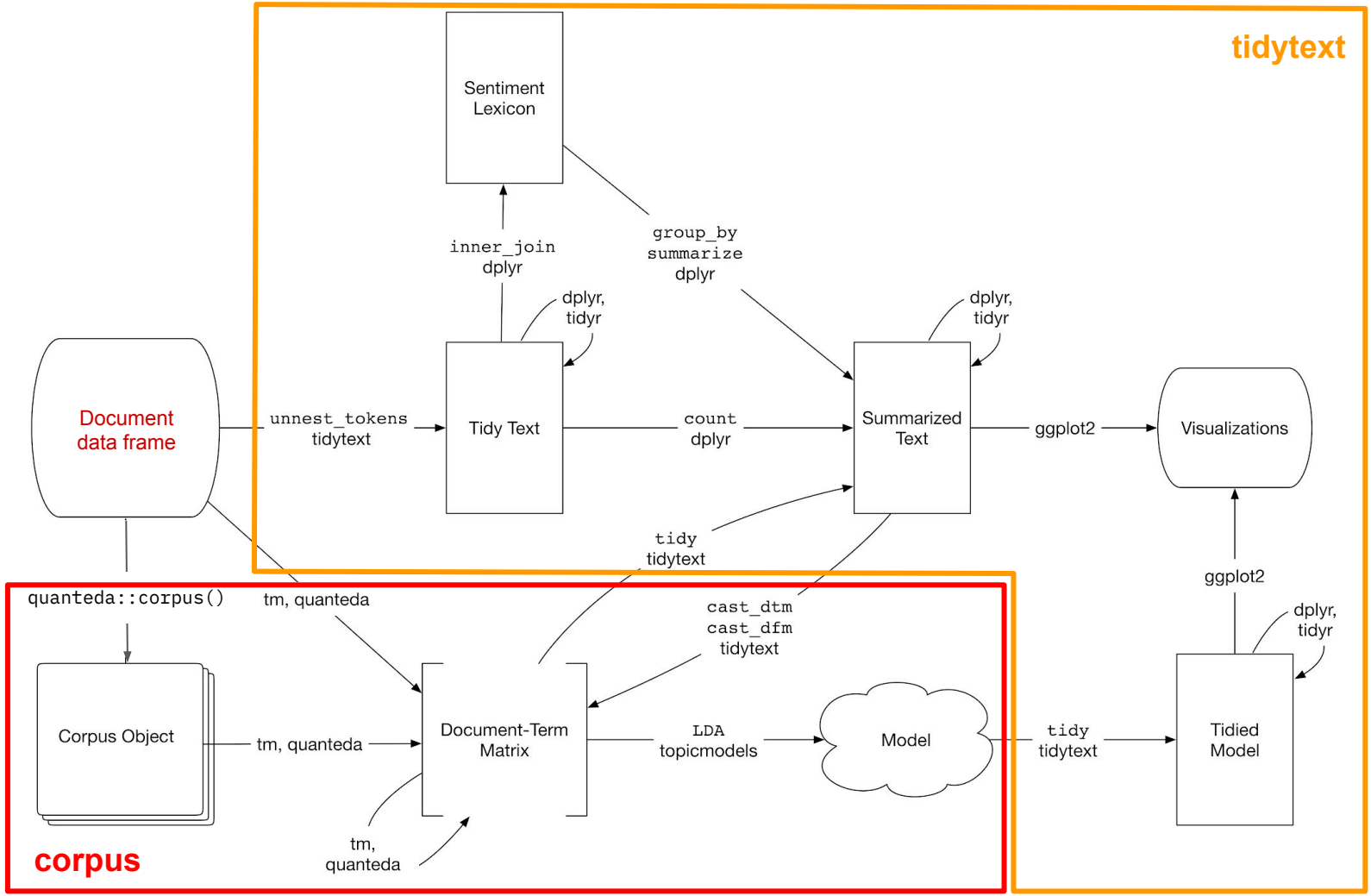
斷詞、tidytext、quanteda

R Text Mining Frameworks


文本資料
(.txt, .html, .json)
斷詞
Metadata
Document
data frame
每列 (row) 為
一篇文章
tidytext::unnest_tokens()
quanteda::corpus()
Tidytext
Corpus

Document data frame

doc_id	author	title	content
1	楊喚	夏夜	蝴蝶 和 蜜蜂 們 帶著 花朵 的 蜜糖 回來 了 羊隊和 牛群 告別 了 田野 回家 了 火紅 的 太陽 也 滾著...
2	宋晶宜	雅量	朋友 買 了 一件 衣料 綠色 的 底子 帶 白色方格 當 她拿給 我們 看 時 一位 對 圍棋 十分 感 與趣 的 同學 說 啊 好像 棋盤 似 的 我 看 倒 有點 像 稿紙 ...
3	胡適	母親的教誨	每天 天剛亮 時 我 母親 便 把 我 喊醒 叫 我 披衣 坐起 我 從 不 知道 她 醒來 坐 了 多久 了 她 看 我 清醒 了 便 對 我 說 昨天 我 做錯 了 什麼 事 說錯 了 什麼 話 要 我 認錯 要 我 用功讀書 ...
...	...	...	...



原圖: https://www.tidytextmining.com/images/tidyflow-ch-6.png

jiebaR

```
library(jiebaR)
seg <- worker(user = "path/to/user_dict")
segment("失業的熊讚陪柯文哲看銀翼殺手", seg)
```

Document data frame

doc_id	author	title	content
1	楊喚	夏夜	蝴蝶 和 蜜蜂 們 帶著 花朵 的 蜜糖 回來 了 羊隊和 牛群 告別 了 田野 回家 了 火紅 的 太陽 也 滾著...
2	宋晶宜	雅量	朋友 買 了 一件 衣料 綠色 的 底子 帶 白色方格 當 她拿給 我們 看 時 一位 對 圍棋 十分 感 與趣 的 同學 說 啊 好像 棋盤 似 的 我 看 倒 有點 像 稿紙 ...
3	胡適	母親的教誨	每天 天剛亮 時 我 母親 便 把 我 喊醒 叫 我 披衣 坐起 我 從 不 知道 她 醒來 坐 了 多久 了 她 看 我 清醒 了 便 對 我 說 昨天 我 做錯 了 什麼 事 說錯 了 什麼 話 要 我 認錯 要 我 用功讀書 ...

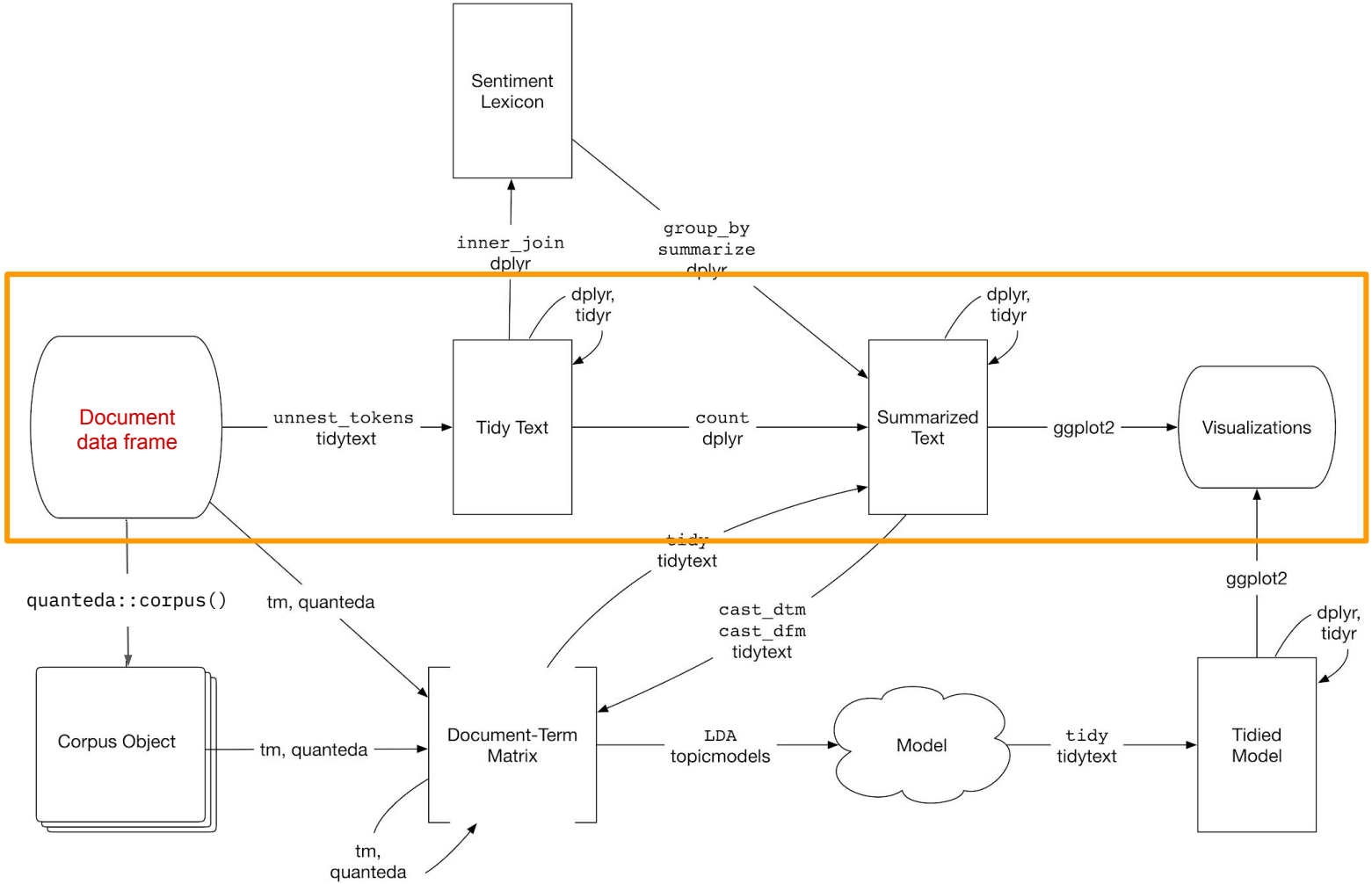
語料

```
docs_df <- readRDS("samesex_marriage.rds")
docs_df
```

id	topic	title	content
pro_292.txt	pro	國防大學生染愛滋被「逼」退學　醫:國軍應公開道歉	東森 新聞 記者 嚴云岑 台北 報導 2016-8-19 愛滋病 感染者 權益 促進會 舉辦 傳染病 病人 權益 案例 座談會...
anti_16.txt	anti	性別平等課本 一點都不平等	不 平等 的 公民 課本 公民課 綱中 寫明 要以 平等 尊重 原則 來 撰寫 課本 內容 經查 後 發現 撰寫 ...
pro_175.txt	pro	國際跨性別現身日 性別友善刻不容緩	近年來 跨 性別 學生 希望 依照 自身 性別 認同 使用 空間 的 聲音 與 訴求 已 浮現 在 台灣 校園 中 但 在 社會 未 正視 跨性別者 處境 ...
...	...	...	...

台灣同性婚姻論述分析 (洪漢唐, 2019)

Tidytext framework



原圖: https://www.tidytextmining.com/images/tidyflow-ch-6.png

例子:詞頻表

`unnest_tokens()` + `dplyr` + `ggplot2`

Text Mining with R

quanteda framework

例子：Key Word in Context

```
corpus()

tokenize_regex()

tokens()

kwic()
```

quanteda tutorials