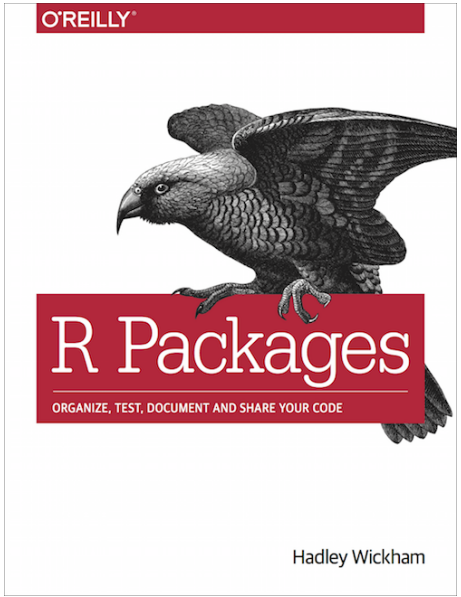
pttR

廖永賦

September 28, 2018

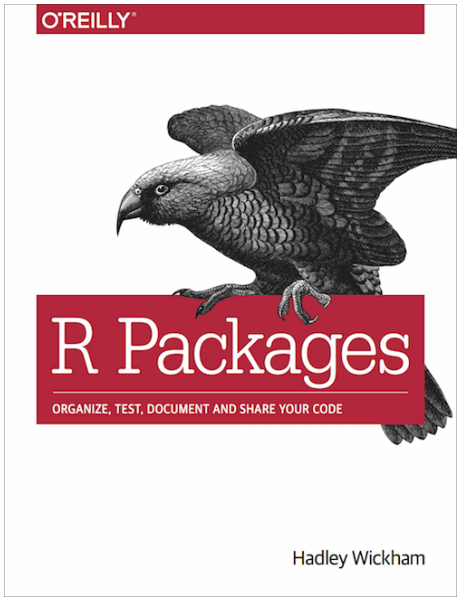
bit.ly/0928pttR

巨人的肩膀



Read Online

巨人的肩膀



Read Online

- Analyses as Packages

Outline

- pttR 簡介
  - 目標
  - 資料抓取(操作)
  - 斷詞(操作)
- 批踢踢詞(?)庫
  - 鄉民百科、用語擷取規則
  - 自動更新

pttR 簡介

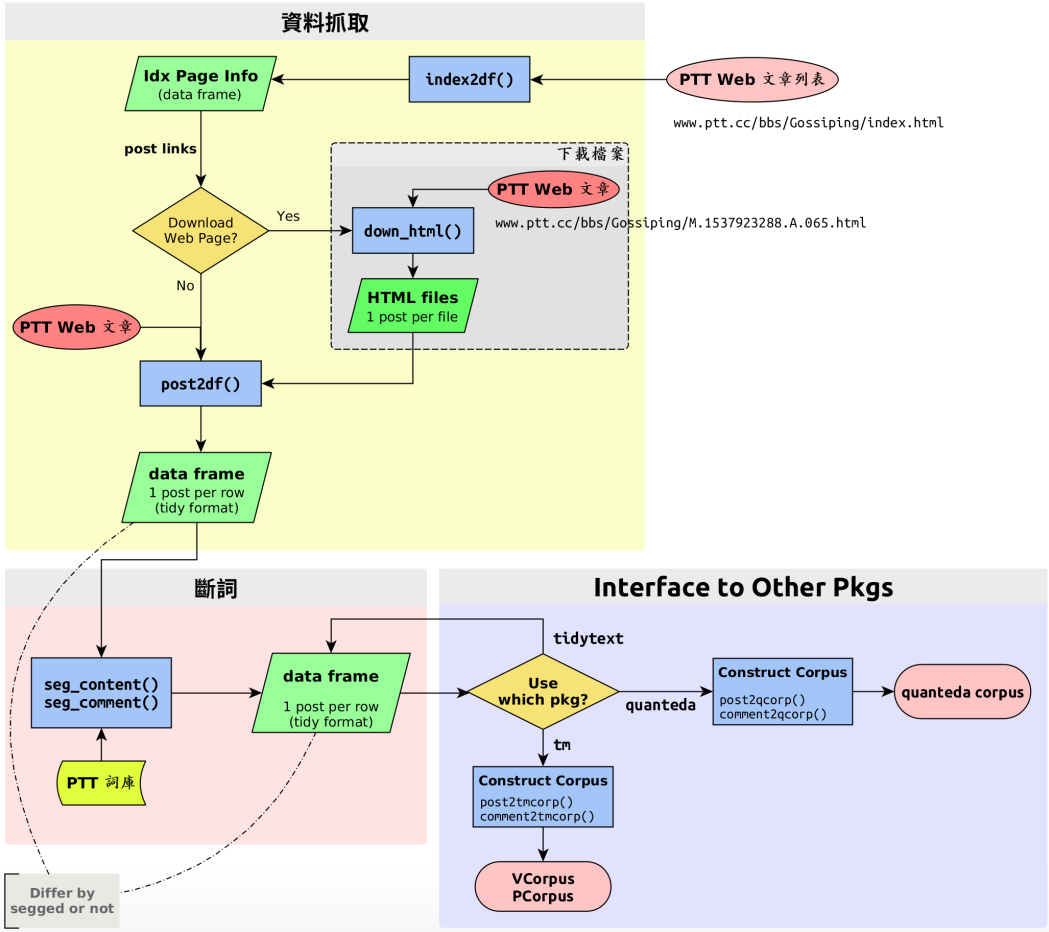
pttR 目標

1. 減輕抓取 PTT 資料的負擔
2. 符合 PTT 的斷詞處理
3. 與 R Text Mining 套件銜接

pttR 目標

1. 減輕抓取 PTT 資料的負擔
2. 符合 PTT 的斷詞處理
3. 與 R Text Mining 套件銜接

How pttR Works



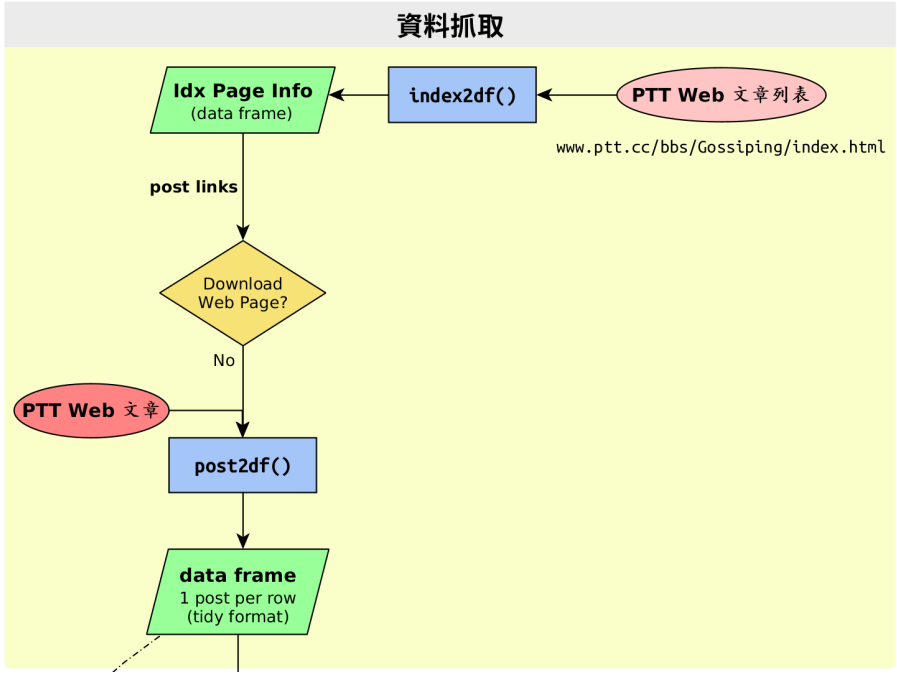
Get Started

```
devtools::install_github("liao961120/pttR", ref = "build")
```

範例程式碼

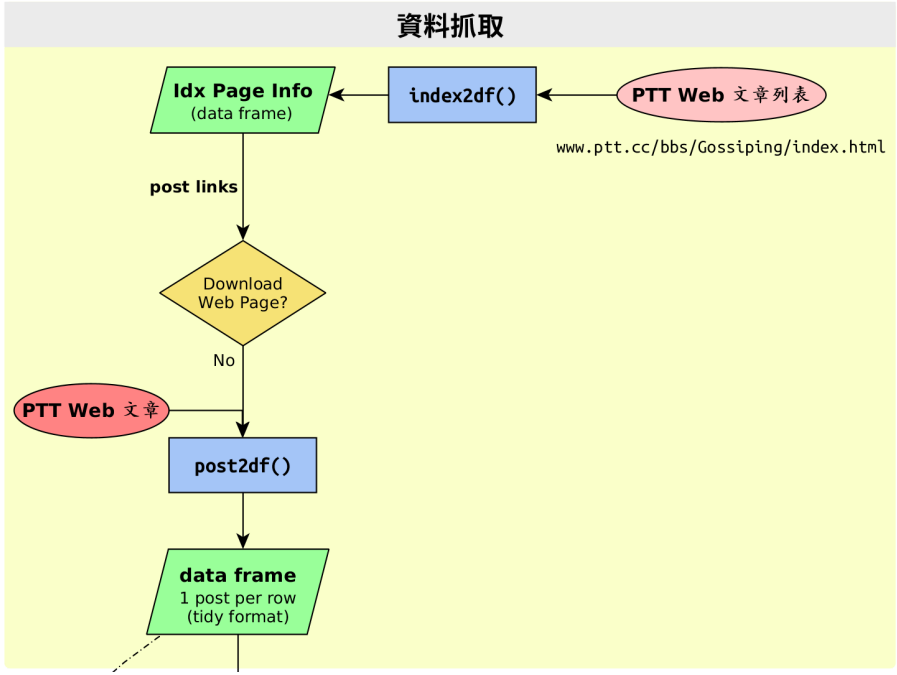
```
library(dplyr)

# 資料抓取
idx_df <- pttR::index2df("gossiping", newest = 1)
pst_df <- idx_df$link[1:5] %>%
  pttR::as_url() %>%
  pttR::post2df()

# 斷詞
pst_df_segged <- pst_df %>%
  mutate(content = pttR::seg_content(content),
         comment = pttR::seg_comment(comment))

# 第一篇文章的留言
pst_df_segged$comment[[1]]

# Construct Corpus Object
post_qcorp <- pttR::post2qcorp(pst_df_segged) # Corpus object
cmt_qcorp <- pttR::comment2qcorp(pst_df_segged) # Corpus list-col in df
```

資料抓取


資料抓取
Idx Page Info
(data frame)
index2df()
PTT Web 文章列表
www.ptt.cc/bbs/Gossiping/index.html
post links
Download
Web Page?
No
PTT Web 文章
post2df()
data frame
1 post per row
(tidy format)

- PTT 網頁

資料抓取

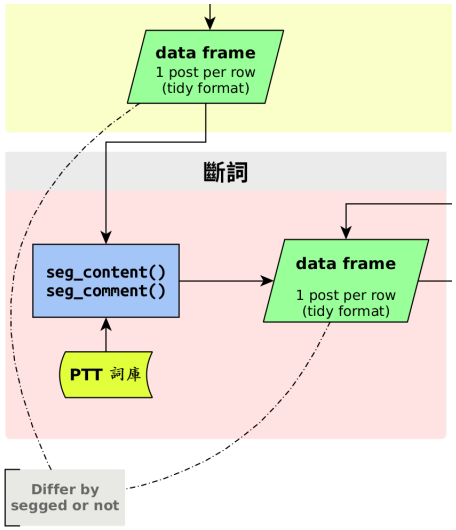


- PTT 網頁
- `index2df()` ≡ www.ptt.cc/bbs/*看板名稱*/index.html

  `post2df()` ≡ www.ptt.cc/bbs/*看板名稱*/xx..xx.html

Demo

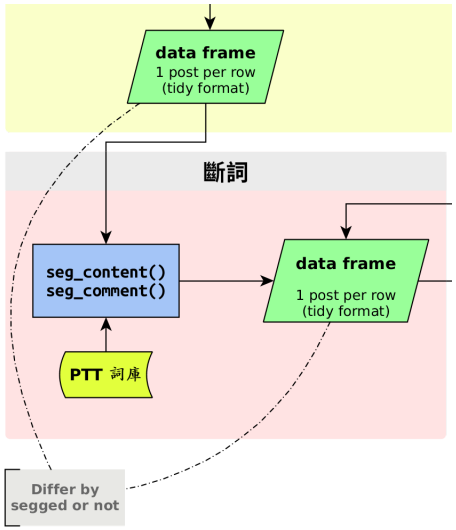
斷詞



- jiebaR
- user dictionary (批踢踢用語)

(Demo)

斷詞



- jiebaR
- user dictionary (批踢踢用語)
- `dplyr::mutate()` + `pttR::seg_content()` + `pttR::seg_comment()`

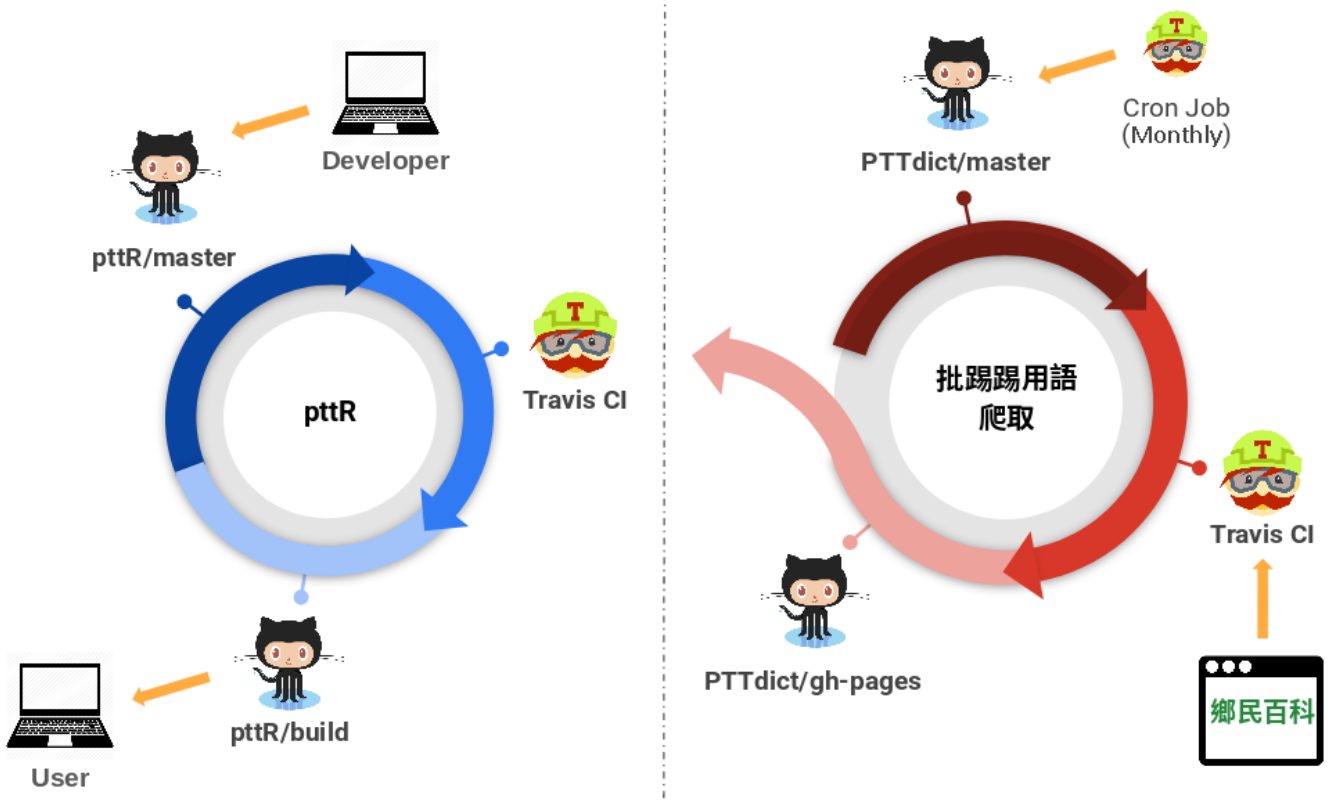
(Demo)

批踢踢詞庫

鄉民百科資料擷取

- PTT鄉民百科
- Scrapy
  - Start URLs:
    流行用語、鄉民文化、流行符號、基本用語、PTT名人、看板、事件、相關事物
  - PTT用語來源
    - `<h1 class="page-header__title">` (標題)
    - `<b>` (粗體)
    - `<a class="new">` (紅字)
    - 「、」 (上、下引號)

自動更新


Developer
pttR/master
pttR
Travis CI
pttR/build
User
PTTdict/master
Cron Job
(Monthly)
批踢踢用語
爬取
Travis CI
PTTdict/gh-pages
鄉民百科

相關連結

- pttR (Vignette, GitHub)
- 批踢踢用語 (GitHub)
- PTT鄉民百科