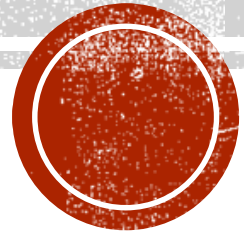
網路爬蟲實戰心法

國立臺灣大學共同教育中心

蔡芸琤



大綱

- 爬蟲基本概念介紹
- URL 的格式
- 網頁基本架構介紹
- 檢視原始碼
- 開發人員工具
- POSTMAN
- 使用 R 語言 **GET** Function
- 使用 R 語言 **POST** Function

爬蟲基本概念介紹

- 網頁的內容是由 HTML+CSS+JavaScript 組合而成。
- HTML+CSS+JavaScript 透過瀏覽器編譯後呈現給使用者。
- 爬蟲是一種自動抓取網頁內容 (HTML+CSS+JavaScript) 的程式。
- 爬蟲透過網址 (URL：Uniform Resource Locator) 提取網頁內容。

https://github.com/pecu/RCrawler101-201504/blob/master/Try/emap_pcsc_com_tw.R

URL 的格式

- URL 分成三部分：http://www.ntu.edu.tw/about/about.html
  - 第一部分是協議 (http、https)。
  - 第二部分是存有該資源的主機位址 (140.112.8.116、www.ntu.edu.tw)。
  - 第三部分是主機上的具體目錄或文件名等 (about/about.html)。
  - 第一部分和第二部分用「://」符號隔開。
  - 第二部分和第三部分用「/」符號隔開。
  - 第一部分和第二部分不可缺少，第三部分看實際狀況決定是否加上。

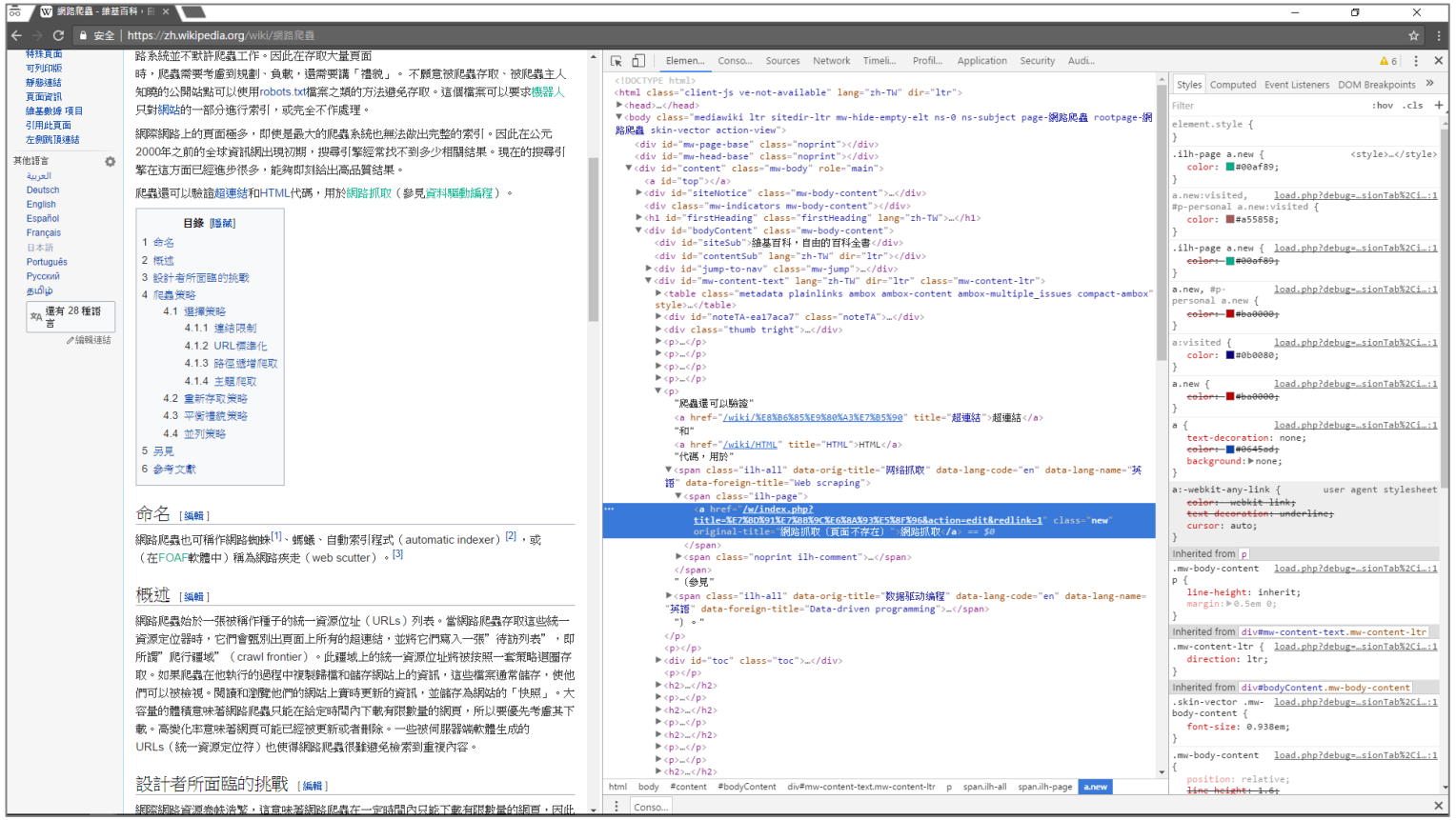
網頁基本架構介紹

Request
GET OR POST
HTTP/1.1

Response
200 OK
HTTP /1.1

http://httpbin.org/
https://www.w3.org/Protocols/rfc2616/rfc2616.html

檢視原始碼


網路爬蟲 - 維基百科
https://zh.wikipedia.org/wiki/網路爬蟲
目錄 [隱藏]
1 命名
2 概述
3 設計者所面臨的挑戰
4 爬蟲策略
命名 [編輯]
概述 [編輯]
設計者所面臨的挑戰 [編輯]

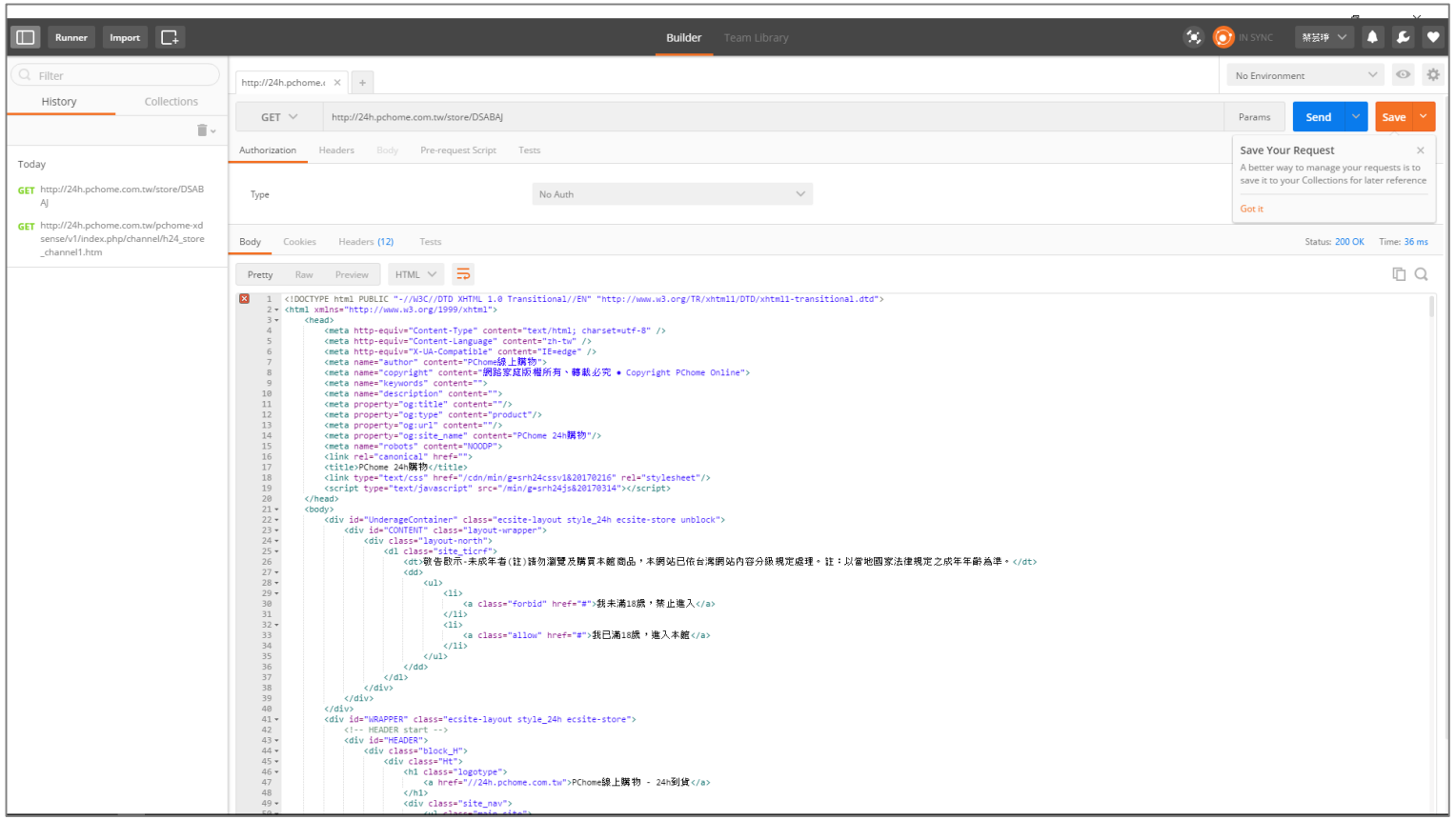
開發人員工具


Acer►Predator - PChome
24h.pchome.com.tw/store/DSABAJ
PChome 24h購物
Acer►Predator
LCD 螢幕
精選品牌
Predator Z1 眼球追蹤功能
•1800R曲面•144Hz•NVIDIA® G-SYNC™
對準！瞄準！掩護！
隨貨附的刺客教條遊戲序號卡
acer Predator Z271T 27型VA曲面電競螢
網路價$24880
PREDATOR Z35
35吋VA曲面
電競螢幕
買就送Tt eSPORTS電競LED藍光鍵盤滑鼠組
網路價$34900
買就送Tt eSPORTS電競LED藍光鍵盤滑鼠組
3840X2160 4K超高解析
網路價$23900
Elemen... Conso... Sources Network Timeli... Profil... Application
Request URL: http://24h.pchome.com.tw/store/DSABAJ
Request Method: GET
Status Code: 200 OK
Remote Address: 210.242.216.153:80
Response Headers
Request Headers

開發人員工具

- 在網頁任何位置按右鍵，選擇“檢查”，就可以看到相對應的原始碼。
- Network 頁面，可以看到網頁各項的執行細節。
- Console 頁面，可以檢查錯誤訊息。
- 直接對 CSS 樣式表更改參數，畫面就會直接可以預覽。
- 爬蟲觀察技巧：http://tech-marsw.logdown.com/blog/2016/01/10/crawler-tips-mining-chrome。

POSTMAN


Runner
Import
Builder
Team Library
Filter
History
Collections
Today
GET http://24h.pchome.com.tw/store/DSABAJ
GET http://24h.pchome.com.tw/pchome-xdsense/v1/index.php/channel/h24_store_channel1.htm
GET
http://24h.pchome.com.tw/store/DSABAJ
Params
Send
Save
Authorization
Headers
Body
Pre-request Script
Tests
Type
No Auth
Save Your Request
A better way to manage your requests is to save it to your Collections for later reference
Got it
Body
Cookies
Headers (12)
Tests
Status: 200 OK
Time: 36 ms
Pretty
Raw
Preview
HTML
No Environment

R 語言 GET FUNCTION

```
GET("http://google.com/")
GET("http://google.com/", path = "search")
GET("http://google.com/", path = "search", query = list(q = "ham"))

# See what GET is doing with httpbin.org
url <- "http://httpbin.org/get"
GET(url)
GET(url, add_headers(a = 1, b = 2))
GET(url, set_cookies(a = 1, b = 2))
GET(url, add_headers(a = 1, b = 2), set_cookies(a = 1, b = 2))
GET(url, authenticate("username", "password"))
GET(url, verbose())

# You might want to manually specify the handle so you can have multiple
# independent logins to the same website.
google <- handle("http://google.com")
GET(handle = google, path = "/")
GET(handle = google, path = "search")
```

R Package httr：https://cran.r-project.org/web/packages/httr/index.html
User Guide：https://cran.r-project.org/web/packages/httr/httr.pdf

R 語言 POST FUNCTION

```
b2 <- "http://httpbin.org/post"
POST(b2, body = "A simple text string")
POST(b2, body = list(x = "A simple text string"))
POST(b2, body = list(y = upload_file(system.file("CITATION"))))
POST(b2, body = list(x = "A simple text string"), encode = "json")

# Various types of empty body:
POST(b2, body = NULL, verbose())
POST(b2, body = FALSE, verbose())
POST(b2, body = "", verbose())
```

R Package httr：https://cran.r-project.org/web/packages/httr/index.html
User Guide：https://cran.r-project.org/web/packages/httr/httr.pdf

牛刀小試

- 利用 GET 或 POST 存下 10 頁以上的網頁原始碼
- http://www.cookbook-r.com/Data_input_and_output/Writing_data_to_a_file/
- 提示 FUNCTIONS：
  - FOR
  - DUMP
  - GET
  - POST
- 回家練習參考： https://github.com/pecu/RCrawler101-201504/tree/master/CaseStudies