為何需要用 Scrapy

- 穩健
- 功能眾多
- 暫時作為 B 計畫，但未來可能會成為 A 計畫
  - Scrapy 是專門設計來做網路爬蟲的框架
    - 相較之下， requests + beautifulsoup 顯得陽春

Scrapy [編輯]

維基百科，自由的百科全書

Scrapy（/ˈ skreɪpi/ *SKRAY-pee*[2]是一個用Python編寫的自由且開源的網絡爬蟲框架。它在設計上的初衷是用於爬取網絡數據，但也可用作使用API來提取數據，或作為生成目的的網絡爬蟲[3]。該框架目前由網絡抓取的開發與服務公司Scrapinghub公司維護。

Scrapy項目圍繞「蜘蛛」（spiders）建構，蜘蛛是提供一套指令的自包含的爬網程序（crawlers）。遵循其他如Django框架的一次且僅一次精神[4]，允許開發者重用代碼將便於構建和拓展大型的爬網項目。Scrapy也提供一個爬網shell，開發者可用它測試對網站的效果。[5]

使用Scrapy的知名公司和產品有：Lyst[6][7]、Parse.ly[8]、Sayone Technologies[9]、Sciences Po Medialab[10]、Data.gov.uk的世界政府數據網站[11]等。

Scrapy

開發者	Scrapinghub, Ltd.
初始版本	2008年6月26日
穩定版本	2.4.1（2020年11月17日，3個月前[1]）
原始碼庫	github.com/scrapy/scrapy
程式語言	Python
作業系統	Windows、macOS、Linux
類型	網絡爬蟲
許可協議	BSD許可證
網站	scrapy.org

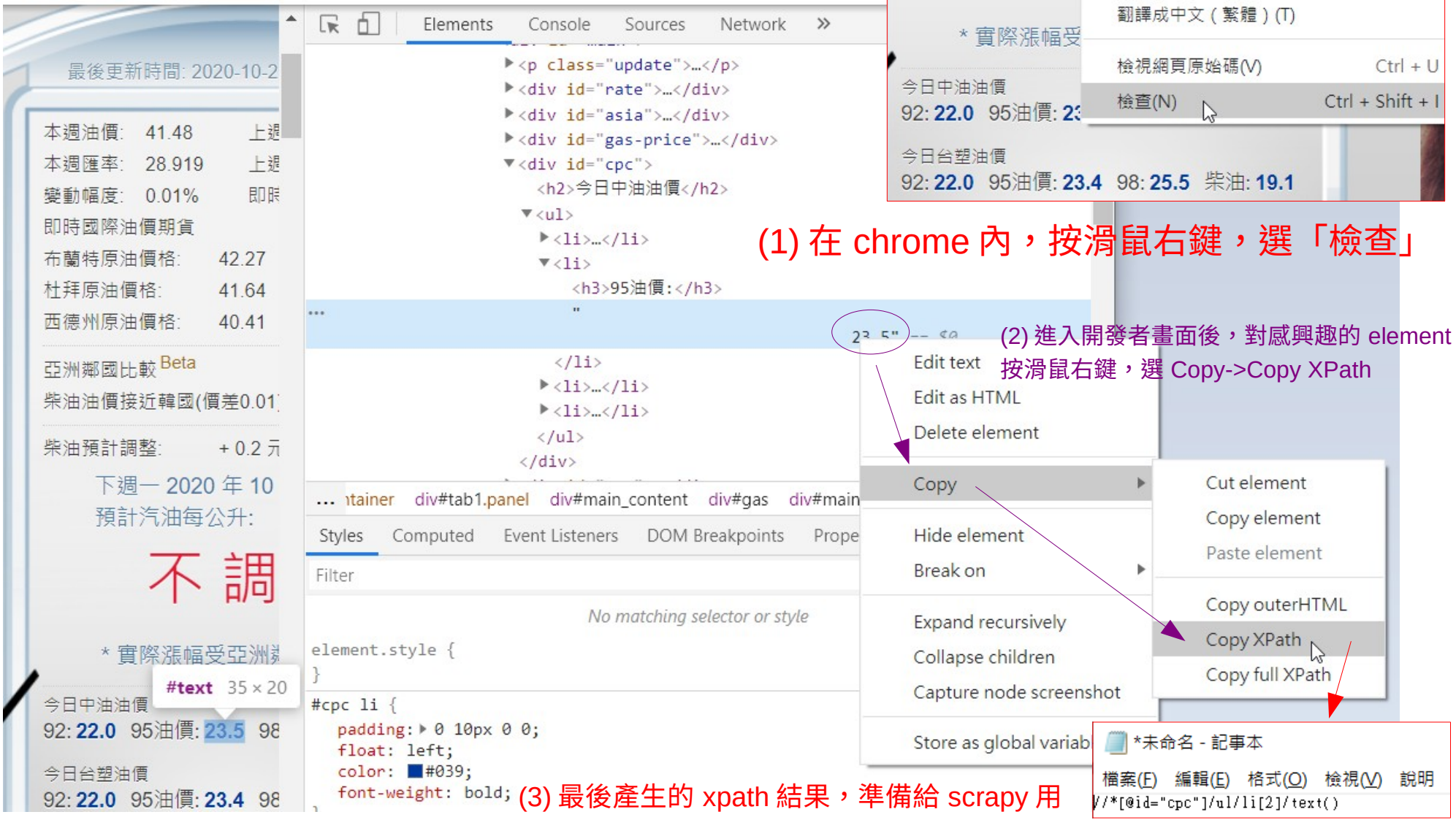
歷史 [編輯]

Scrapy誕生於網絡聚合和電子商務公司Mydeco，它由Mydeco和Insophia公司的員工開發和維護。2008年8月首次以BSD許可證公開發布，2015年6月發布有里程碑意義的1.0版本[12]。2011年，Scrapinghub成為新的官方維護者[13][14]。

操作 Scrapy (概觀)

- 以 gas.goodlife.tw 為例，試著找出油價資訊
  - 打開該網頁，觀察資訊，並建立新 Scrapy 專案
  - 進入 google chrome 的開發者畫面
  - 複製感興趣的 element 其對應的 xpath


浮動油價漲跌監測 - 中油油價查詢及國際油價走勢 - GoodLife
浮動油價漲跌監測 - 中油油
gas.goodlife.tw
油價公告
全新改版
功能升級
信用卡優惠
加油站導航
推播通知
App Store
Google play
最後更新時間: 2021-03-23 10:10 (6分鐘 前)
本週油價: 63.35 上週油價: 65.99
本週匯率: 28.492 上週匯率: 28.315
變動幅度: -2.72% 即時匯率: 28.45
即時國際油價期貨
布蘭特原油價格: 64.02 09:36 更新
杜拜原油價格: 63.32 03/22 更新
西德州原油價格: 61.47 02:30 更新
亞洲鄰國比較 Beta
92油價接近韓國(價差0.05)，可能受影響
柴油油價接近韓國(價差0.01)，可能受影響
柴油預計調整: + 0.8 元
中油累計吸 0.9 元, 若不列入計算, 應調整: +1.2 元
下週一 2021 年 03 月 29 日 起,
預計汽油每公升:
漲 0.5 元
＊實際漲幅受亞洲鄰國油價限制
今日中油油價
92: 26.5 95油價: 28.0 98: 30.0 柴油: 23.8
今日台塑油價
92: 26.5 95油價: 27.9 98: 30.0 柴油: 23.6


最後更新時間: 2020-10-2
本週油價: 41.48
本週匯率: 28.919
變動幅度: 0.01%
即時國際油價期貨
布蘭特原油價格: 42.27
杜拜原油價格: 41.64
西德州原油價格: 40.41
亞洲鄰國比較 Beta
柴油油價接近韓國(價差0.01
柴油預計調整: + 0.2 元
下週一 2020 年 10
預計汽油每公升:
不調
* 實際漲幅受亞洲
今日中油油價
92: 22.0 95油價: 23.5 98
今日台塑油價
92: 22.0 95油價: 23.4 98
#text 35 × 20
Elements Console Sources Network
…
…
…
…

<h2>今日中油油價</h2>

…

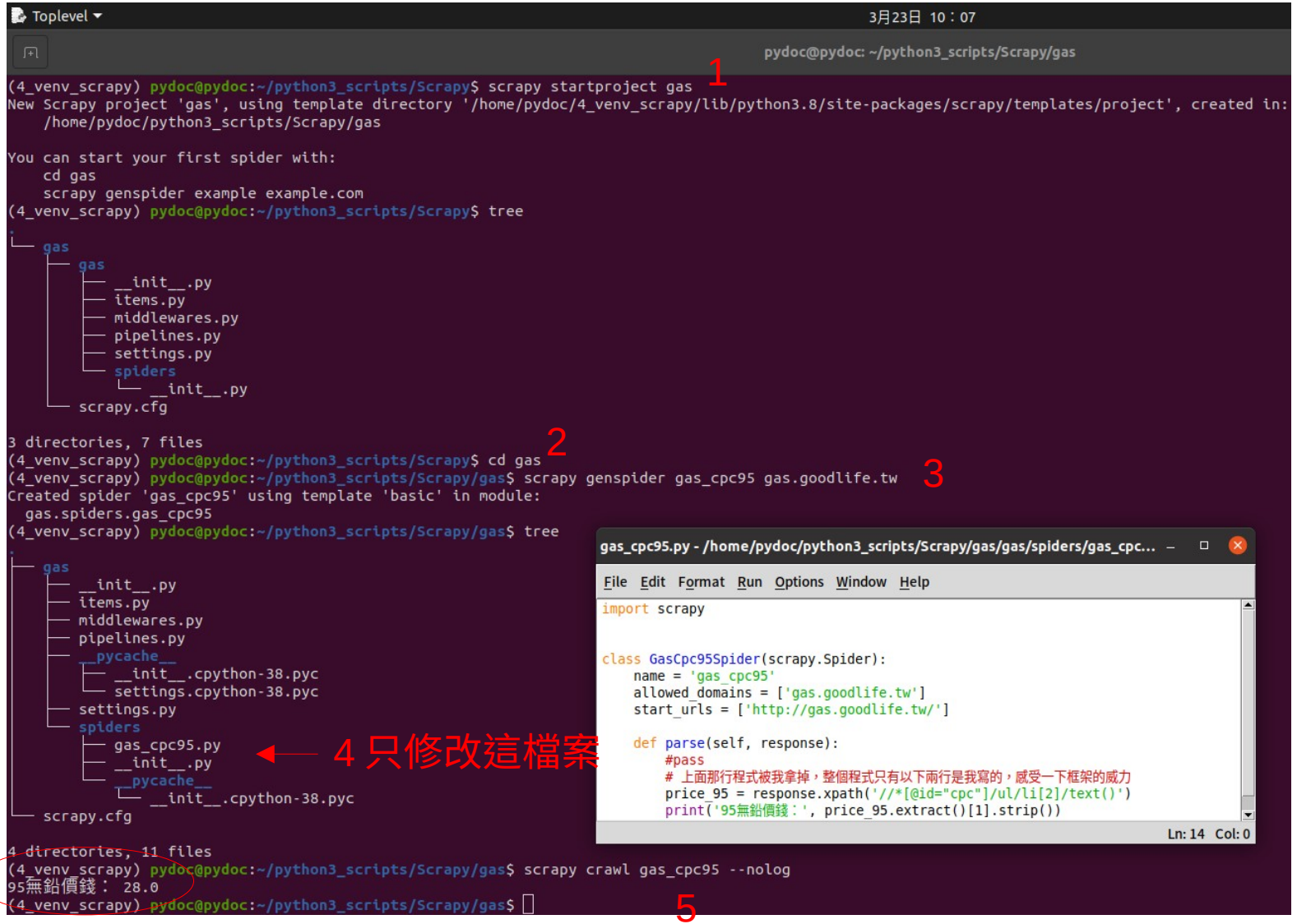
<h3>95油價:</h3>
"
23.5" == $0

…
…

div#tab1.panel div#main_content div#gas div#main
Styles Computed Event Listeners DOM Breakpoints
Filter
No matching selector or style
element.style {
}
#cpc li {
padding: 0 10px 0 0;
float: left;
color: #039;
font-weight: bold;
翻譯成中文（繁體）(T)
檢視網頁原始碼(V) Ctrl + U
檢查(N) Ctrl + Shift + I
今日中油油價
92: 22.0 95油價: 23
今日台塑油價
92: 22.0 95油價: 23.4 98: 25.5 柴油: 19.1
(1) 在 chrome 內，按滑鼠右鍵，選「檢查」
(2) 進入開發者畫面後，對感興趣的 element
按滑鼠右鍵，選 Copy->Copy XPath
Edit text
Edit as HTML
Delete element
Copy
Hide element
Break on
Expand recursively
Collapse children
Capture node screenshot
Store as global variab
Cut element
Copy element
Paste element
Copy outerHTML
Copy XPath
Copy full XPath
*未命名 - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明
//*[@id="cpc"]/ul/li[2]/text()
(3) 最後產生的 xpath 結果，準備給 scrapy 用

使用 Scrapy 來實現網路爬蟲 (細節)

- pip install scrapy (安裝 Scrapy 模組，官方建議在虛擬環境下安裝)
- **scrapy startproject** gas (新建一個 project)
- **cd** gas (進入新建 project 的目錄下)
- **scrapy genspider** gas_cpc95 gas.goodlife.tw
  - 產生 spider 時得指定一個起始網址，即使以後用不到
- **scrapy crawl** gas_cpc95
  - 不一定得用 crawl 指令，也可以用其他方式跑程式 (例如 runspider)


Toplevel
3月23日 10：07
pydoc@pydoc: ~/python3_scripts/Scrapy/gas
1
(4_venv_scrapy) pydoc@pydoc:~/python3_scripts/Scrapy$ scrapy startproject gas
New Scrapy project 'gas', using template directory '/home/pydoc/4_venv_scrapy/lib/python3.8/site-packages/scrapy/templates/project', created in:
/home/pydoc/python3_scripts/Scrapy/gas
You can start your first spider with:
cd gas
scrapy genspider example example.com
(4_venv_scrapy) pydoc@pydoc:~/python3_scripts/Scrapy$ tree
.
└── gas
├── gas
│ ├── __init__.py
│ ├── items.py
│ ├── middlewares.py
│ ├── pipelines.py
│ ├── settings.py
│ └── spiders
│ └── __init__.py
└── scrapy.cfg
3 directories, 7 files
2
(4_venv_scrapy) pydoc@pydoc:~/python3_scripts/Scrapy$ cd gas
(4_venv_scrapy) pydoc@pydoc:~/python3_scripts/Scrapy/gas$ scrapy genspider gas_cpc95 gas.goodlife.tw
3
Created spider 'gas_cpc95' using template 'basic' in module:
gas.spiders.gas_cpc95
(4_venv_scrapy) pydoc@pydoc:~/python3_scripts/Scrapy/gas$ tree
.
├── gas
│ ├── __init__.py
│ ├── items.py
│ ├── middlewares.py
│ ├── pipelines.py
│ ├── __pycache__
│ │ ├── __init__.cpython-38.pyc
│ │ └── settings.cpython-38.pyc
│ ├── settings.py
│ └── spiders
│ ├── gas_cpc95.py
│ ├── __init__.py
│ └── __pycache__
│ └── __init__.cpython-38.pyc
└── scrapy.cfg
4 只修改這檔案
gas_cpc95.py - /home/pydoc/python3_scripts/Scrapy/gas/gas/spiders/gas_cpc...
File Edit Format Run Options Window Help
import scrapy
class GasCpc95Spider(scrapy.Spider):
name = 'gas_cpc95'
allowed_domains = ['gas.goodlife.tw']
start_urls = ['http://gas.goodlife.tw/']
def parse(self, response):
#pass
# 上面那行程式被我拿掉，整個程式只有以下兩行是我寫的，感受一下框架的威力
price_95 = response.xpath('//*[@id="cpc"]/ul/li[2]/text()')
print('95無鉛價錢：', price_95.extract()[1].strip())
Ln: 14 Col: 0
4 directories, 11 files
(4_venv_scrapy) pydoc@pydoc:~/python3_scripts/Scrapy/gas$ scrapy crawl gas_cpc95 --nolog
95無鉛價錢： 28.0
(4_venv_scrapy) pydoc@pydoc:~/python3_scripts/Scrapy/gas$
5
最後結果