

黃洧析
Bruce Huang

Python爬蟲

Very Good Experience

With Laravel Development

```
You, 7 months ago | 1 author (You)
Extends('Layout.app')
Section('content')
Section class="website-content">
    Div id="sub-header">
        Div class="height-20p"></div>
    Div class="container">
        Section id="feature-post" class="py-5">
            Div class="row align-items-center">
                Div class="col-lg-7">
                    Div class="text-lg font-weight-300 text-gray-500">
                        Having a good time with my new project.
                    <h1 class="text-lg font-weight-300 text-gray-500">
                        Very Good Experience
                    </h1>
                    Div class="height-20p"></div>
                    P class="text-regular font-weight-300 text-gray-500">
                        This repository provides you a development environment without
                    </p>
                    Div class="height-30p"></div>
                    Div class="d-flex">
                        A href="/" class="btn btn-primary font-weight-500 px-3 py-2">
                            Home
                        <a href="#" class="d-flex align-items-center justify-content-end px-3 py-2">
                            
                            <span class="ml-2">Go to GitHub repository
                        </a>
                    </div>
```

Python爬蟲-學習目標

.

1 : Python爬蟲簡介與基礎工具

2 : 進階爬蟲技術

3 : Flask Web框架介紹

4 : 結合Flask與爬蟲數據展示

5 : 專案實作與展示

1：Python爬蟲簡介與基礎工具

什麼是網路爬蟲？

- **定義：** 網路爬蟲 (Web Crawler) 是一種自動化腳本，用於系統性地瀏覽網頁並提取所需的資訊。
- **應用：**
 - 資料收集與分析
 - 搜尋引擎索引建立
 - 價格比較
 - 市場調查

1 : Python爬蟲簡介與基礎工具

HTTP 協議基礎

- **請求方法：**

- GET : 請求獲取資源。
- POST : 提交資料以處理請求。

- **狀態碼：**

- 200 OK : 請求成功。
- 404 Not Found : 資源未找到。

-

1 : Python爬蟲簡介與基礎工具

HTML 基礎結構

- 元素：標籤（Tags）、屬性（Attributes）、內容（Content）
- 常見標籤：
 - <a> : 超連結
 - <p> : 段落
 - <div> : 區塊
 - : 行內元素

1：Python爬蟲簡介與基礎工具

開發環境設置

- 安裝 Python：
 - 從 Python 官方網站下載並安裝最新版本。
- 安裝 VSCode：
 - 從 VSCode 官方網站下載並安裝。
- 安裝 Python 擴充套件：
 - 在 VSCode 中，前往擴充套件市集，搜尋並安裝「Python」擴充套件。

1 : Python爬蟲簡介與基礎工具

建立虛擬環境：Windows

- 在終端機中執行：`python -m venv env`
- 啟用虛擬環境：`.\env\Scripts\activate`

1 : Python爬蟲簡介與基礎工具

建立虛擬環境 : Windows

- 在終端機中執行 : `python -m venv env`
- 啟用虛擬環境 : `.\env\Scripts\activate`
- 安裝 `pip install requests`
- 安裝 `pip install beautifulsoup4`

1 : Python爬蟲簡介與基礎工具

介紹 requests 模組

- 功能：用於發送 HTTP 請求，簡化了網路請求的處理。

1：Python爬蟲簡介與基礎工具

req.py

```
python req.py M X  
python req.py > ...  
1 import requests  
2  
3 url = "https://www.example.com"  
4 # url = "https://news.google.com/home?hl=zh-TW&gl=TW&ceid=TW:zh-Hant"  
5 response = requests.get(url)  
6  
7 if response.status_code == 200:  
8     print("成功取得網頁內容")  
9     print(response.text)  
10 else:  
11     print(f"無法取得網頁內容，狀態碼：{response.status_code}")  
12     # print()  
13
```

1 : Python爬蟲簡介與基礎工具

介紹 BeautifulSoup 模組(簡稱bs4)

- 功能：用於解析 HTML 和 XML 文件，方便地提取所需的資料。

1：Python爬蟲簡介與基礎工具

req-bs4.py

```
python req-bs4.py > ...
1  from bs4 import BeautifulSoup
2
3  html_content = """
4      <html>
5          <head><title>範例網頁</title></head>
6          <body>
7              <h1>這是主標題</h1>
8              <p>這是一個段落。</p>
9              <a href="https://www.example.com">這是一個連結</a>
10         </body>
11     </html>
12     """
13
14  soup = BeautifulSoup(html_content, "html.parser")
15  title = soup.title.string
16  print(f"網頁標題 : {title}")
17
```

1：Python爬蟲簡介與基礎工具

實作練習：抓取指定網站的標題和基本資訊

- 目標：從指定的網站抓取網頁標題和所有連結。
- 步驟：
 1. 使用 requests 獲取網頁內容。
 2. 使用 BeautifulSoup 解析 HTML。
 3. 提取網頁標題和所有連結。

1：Python爬蟲簡介與基礎工具

req-bs4a.py

```
python req-bs4a.py M > ...  
python req-bs4a.py > ...  
4 # url = "https://news.google.com/home?hl=zh-TW&gl=TW&ceid=TW:zh-Hant"  
5 url = "https://books.toscrape.com/" # 替換為您要抓取的網站 URL  
6 response = requests.get(url)  
7  
8 if response.status_code == 200:  
9     soup = BeautifulSoup(response.text, "html.parser")  
10    title = soup.title.string  
11    print(f"網頁標題: {title}")  
12    print("所有連結:")  
13    for link in soup.find_all("a"):  
14        href = link.get("href")  
15        if href:  
16            print(href)  
17    else:  
18        print(f"無法取得網頁內容，狀態碼: {response.status_code}")  
19
```

1：Python爬蟲簡介與基礎工具

The end