JavaScript 程式設計新手村

單元29 NodeJS 網路爬蟲入門實戰

@kdchang

Outline

1. 網路爬蟲概論
2. NodeJS 簡介
3. 網路爬蟲實戰

網路爬蟲概論

網路蜘蛛（Web spider）也叫網路爬蟲（Web crawler），可以使用自動化程式採集所有其能夠存取到的頁面內容。網路爬蟲廣泛運用於搜尋引擎檢索當中

NodeJS 簡介

Node.js 是一個開放原始碼、跨平台的、可用於伺服器端和網路應用的執行環境。Node.js 採用 Google 的 V8 引擎來執行代碼。Node.js 的大部分基本模組都是用 JavaScript 寫成的

使用 Node 環境下我們可以擺脫瀏覽器束縛，在使用 JavaScript 撰寫爬蟲上更為簡易方便

網路爬蟲實戰

準備開始動工

1. 分析目標網頁
2. 撰寫爬蟲程式
3. 儲存爬取資料
4. 分析爬取資料內容

我們的目標：PTT Soft_Job 版

使用工具

1. NodeJS：JavaScript 執行環境
2. Request：方便執行 HTTP 呼叫
3. Cheerio：方便在 server 端使用類似 jQuery DOM 操作
4. fs：Node 內建的檔案處理工具

安裝 NodeJS 和套件

確認 NodeJS 已經安裝完成：

```
// v6.9.1
$ node -v
```

安裝所使用的套件：

```
$ npm install request cheerio
```

來寫程式吧！

進階議題

1. throttling：有些網站碰到大量的存取時，會逐漸降低同來源的優先序，使得爬蟲需要花非常久的執行時間來執行
2. cookie verification：有些網站需認證後才能進入，這時候我們必須額外對不同網址呼叫 request ，並紀錄回傳的 cookie 值
3. agent verification：`request` 這個套件可以帶入自訂的 HTTP Header，偽裝成一般使用者
4. reCAPTCHA：常見的網站認證碼欄或判斷是否為機器人，比較難處理這類問題
5. JavaScript 動態載入的頁面： 使用：phantomjs、CasperJS

參考文件

1. 輕輕鬆鬆用 Nodejs 寫網路爬蟲
2. 資料爬蟲實戰 – 使用 node.js
3. Node.js 網路爬蟲
4. How to make a web crawler in JavaScript / Node.js
5. Scraping the Web With Node.js

總結

在這個單元中我們學會了：

1. 網路爬蟲概論
2. NodeJS 簡介
3. 網路爬蟲實戰