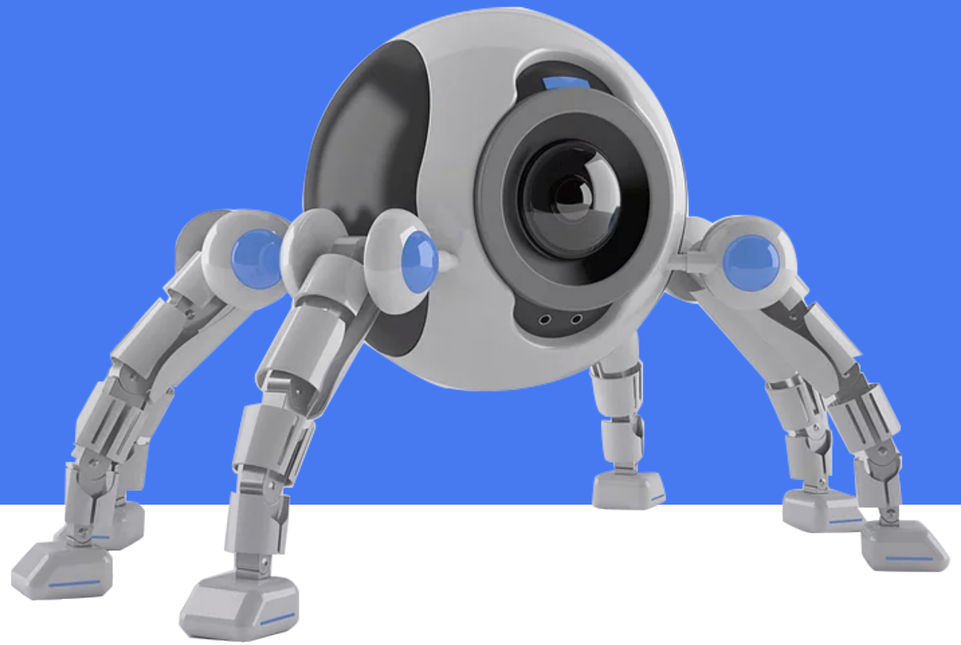
Day 19 動態網頁資料爬蟲

動態網頁爬蟲 - 使用 Selenium



出題教練：張維元

python

本日知識點目標

- 了解 Selenium 用於動態網頁爬蟲的原理
- 能夠使用 Selenium 撰寫動態網頁爬蟲

第一種動態網頁爬蟲策略

關於這種利用到 JavaScript 的非同步特性載入更多資料的網頁稱為動態網頁。而爬蟲程式也會因為沒有執行到 JavaScript 導致資料不完全的現象。

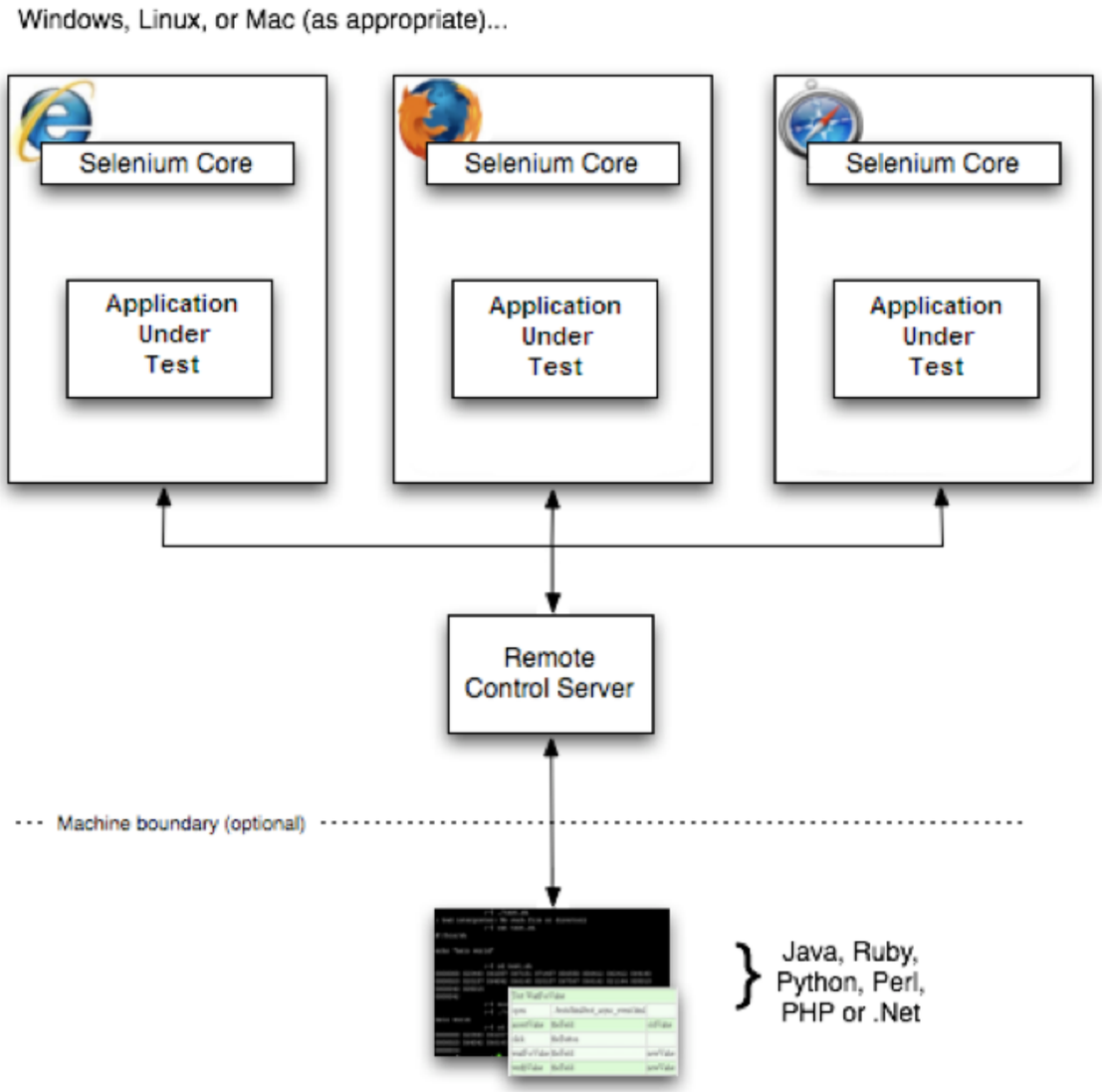
第一種解法會採用 selenium 這樣的瀏覽器模擬工作，從模擬使用者打開瀏覽器的行為，到模擬器執行JavaScript 動態載資料之後。

利用 Selenium 模擬操作瀏覽器

Selenium 是一個瀏覽器自動化（Browser Automation）工具，讓程式可以直接驅動瀏覽器進行各種網站操作。最早的目的是用來進行網頁測試使用，這邊我們藉由特性來運行 JavaScript 作為爬蟲用。

利用 Selenium 模擬操作瀏覽器

CUPOY



準備 Selenium 環境

CUPOY

1. 安裝 selenium 套件

```
1
2
3
$ pip install selenium
```

2. 下載 Chrome 驅動程式

上面第一步驟只是安裝 Selenium 模組而已，必須要下載對應的瀏覽器 Chrome 的驅動程式（建議放在程式相同目錄下）：http://chromedriver.chromium.org/downloads

範例：使用 Selenium 進行爬蟲

CUPOY

```
1  from selenium import webdriver
2
3  browser = webdriver.Chrome(executable_path='./chromedriver')
4  browser.get("http://www.google.com")
5  browser.close()
6
7  browser.page_source
```

這邊設定會去讀取我們放在相同目錄下的 driver 檔案

執行後會真的看到電腦打開一個新個瀏覽器，而且跳轉到設定的網址上！

透過 browser.page_source 可以取出，目前網頁上當下的 HTML，不過這是一個 HTML 格式的字串，此時就可以再利用 BeautifulSoup 進行解析。

重要知識點複習

CUPOY

- 了解 Selenium 用於動態網頁爬蟲的原理
- 能夠使用 Selenium 撰寫動態網頁爬蟲

解題時間
LET'S CRACK IT

請跳出 PDF 至官網 Sample Code & 作業

開始解題

START