Python Saturday

Crawling - The Basic

Content

1. What & Why
2. Equipment
3. Selenium
4. Demo
5. Q & A

WHAT & WHY

- Collecting information
- Collecting information
- Collecting information

EQUIPMENT

equipment

Danh sách công cụ cần thiết

1. Selenium
2. Scrapy
3. Beautiful Soup (BS4)

SELENIUM

Selenium WebDriver là một thư viện cho phép chúng ta lập trình (scripting) test script trên các ngôn ngữ lập trình khác nhau như Python, Java, C#, Ruby.

SCRAPY

Scrapy là một framework được viết bằng Python, nó cấp sẵn 1 cấu trúc tương đối hoàn chỉnh để thực hiện việc crawl và extract data từ website một cách nhanh chóng và dễ dàng.

BEAUTIFUL SOUP

Beautifulsoup là một module hỗ trợ crawling với những site thông thường, ít dữ liệu, nhằm đưa ra một cấu trúc tương tự như khi debug trực tiếp trên trình duyệt, dễ dàng xử lý và can thiệp.

SELENIUM WEBDRIVER

REQUIREMENT

Python 2.7.x +
Download: http://python.org

Pip (nếu không đi kèm)
- Download:
https://bootstrap.pypa.io/get-pip.py
- Thực thi:
python get-pip.py

WEBDRIVER

Các dạng webdriver thông dụng

PhantomJS
(khuyên dùng khi triển khai trên server)

Firefox
(khuyên dùng khi testing)

ChromeDriver

HTMLUnit

WEBDRIVER

PhantomJS

Link download:

- **Windows**: https://goo.gl/hX5dYl
- **MacOS**: https://goo.gl/eKNVSe
- **Linux**: https://goo.gl/CUZasw

WEBDRIVER

PhantomJS

Tips:

- Không sinh log khi chạy:
  `service_log_path = os.path.devnull`
- Không load ảnh khi chạy:
  `service_args = ['--load-images=no']`

WEBDRIVER

```
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By
from selenium.common.exceptions import *
import os, time, sys
```

DEMO

QUESTION?

THANK YOU!