**VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**



PROGRESS REPORT

PROJECT: RECOMMENDATION SYSTEM

Lecturer: Nguyễn Trác Thức

Student: Nguyễn Ngọc Tường

*Tp. Hồ Chí Minh, ngày 1 tháng 1 năm 2019*

# **CHAPTER 1: TÌM HIỂU CÁCH CRAWLER DATA**

## **Tìm hiểu Scrapy Framework**

### Scrapy là gì ?

### Scrapy Framework Model

#### Model

Figure 1 - Scrapy Framework Model

#### Element

|  |  |  |
| --- | --- | --- |
| STT | Thành phần | Nhiệm vụ |
| 1 | Scrapy Engine | Kiểm soát luồng dữ liệu giữa tất cả các thành phần của hệ thống và kích hoạt các sự kiện khi một số hành động xảy ra |
| 2 | Scheduler | Giống như một hàng đợi (queue), Scheduler sắp xếp thứ tự các URL cần download |
| 3 | Downloader | Thực hiện dowload trang web và cung cấp cho Engine |
| 4 | Spiders | Spiders là class được viết bởi người dùng, chúng có trách nhiệm bóc tách dữ liệu cần thiết và tạo các url mới để nạp lại cho Scheduler qua Engine |
| 5 | Item Pipeline | Những dữ liệu được bóc tách từ Spiders sẽ đưa tới đây, Item Pipeline có nhiệm vụ xử lý chúng và lưu vào cơ sở dữ liệu |
| 6 | Các Middlewares | - Là các thành phần nằm giữa Engine với các thành phần khác, với mục đích là giúp người dùng có thể tùy biến, mở rộng khả năng xử lý cho các thành phần.  - Bao gồm:  + Spider Middlewares  + Dowloader Middlewares  + Scheduler Middlewares |

#### Data Flow (Xem Figure 1)

① Khi bắt đầu crawl một website, Engine sẽ xác định tên miền và tìm vị trí của spider đó và yêu cầu spider đó tìm các urls đầu tiên để crawl

② Engine nhận danh sách các urls đầu tiên từ spider, gửi cho Scheduler để sắp xếp

③ Engine yêu cầu danh sách cách urls tiếp theo từ Scheduler

④ Engine nhận danh sách các url tiếp theo từ Scheduler vào gửi đến Dowloader (requests)

⑤ Downloader nhận request và thực hiện việc tải trang, sau khi tải xong sẽ tạo một response và gửi lại Engine

⑥ Respone từ Dowloader sẽ được Engine đẩy qua Spiders để xử lý

⑦ Tại Spiders, khi nhận được response, chúng bóc tách thông tin từ response (tilte, content, author, date publish...) và những url có khả năng để crawl và đẩy lại cho Engine (requests)

⑧ Ở bước này, Engine nhận được kết quả từ Spiders sẽ thực hiện 2 công việc: đẩy những dữ liệu đã được bóc tách tới Item Pipeline để xử lý và lưu vào Databases, đẩy những url mới (requests) mới về Scheduler và quay về bước 3

### Cài đặt và Demo

#### Cài đặt

- Cài đặt bằng lệnh:

pip install scrapy

- Yêu cầu:

* Python 2.7 hoặc Python 3.4 +
* Làm việc trên môi trường Linux, Windows, Mac OS

#### Structure

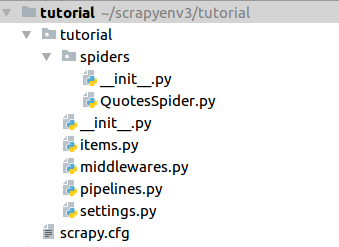
Cấu trúc thư mục Scrapy Project

Figure 2 - Structure of Scrapy Project

#### Demo

\* Bài toán đặt ra: thu thập thông tin 50 question và địa chỉ của các question đó trên trang *https://stackoverflow.com*

\* Giải quyết bài toán:

- Tạo project bằng lệnh:

scrapy startproject demo

- Trong thư mục spiders của project, ta tạo một spider con có tên demo\_spider.py

- Khai báo các giá trị cần crawler về trong file

## Giải pháp Crawler các trang web Javascrip – Splash Tool

### Aaa

### Aaa

### Aaa

## **Phân loại và lưu trữ dữ liệu**

### Aaa

### Aaa

### aaa

# **CHAPTER N: REFERENCES**

<https://viblo.asia/p/gioi-thieuhuong-dan-ve-crawler-voi-scrapy-framework-ByEZkWoEZQ0>