An introduction to Web scraping & crawling with Scrapy Framework

Quách Huy Tùng

HUST

12-11-2019

Web scraping & Web crawling - An Introduction

Web scraping & Web crawling in a nutshell

Scrapy Framework

Web scraping & Web crawling - Why we need it?

- Data is the "new gold" in the 21st century
- Cheap way to collect data without buiding real Information Systems with numerous users
- Data mining
- Information Retrieval
- Machine Learning, Deep Learning

Web scraping & Web crawling - What it is?

- Just for easy, fetch many web pages, extract the data we need and then fetch the next page we want!
- Store the data scraped "somewhere"

Next step: Extract hidden gems in the data with "some smart algorithms"

Web scraping vs Web crawling: Use cases & examples

Web crawling

1. *Purposes*: Find the links between nodes(web pages) in the graph(the Internet) for Ranking and Searching.
2. *Use cases*: Information Retrieval, Web pages ranking

Web scraping

1. *Purposes*: Collect data for specific motivations: User Behaviour prediction, Recommendation System, ...
2. *Use cases*: Data mining, Machine learning, Deep learning, ...

Web scraping vs Web crawling: Use cases & examples

Web crawling

3. *Examples*: GoogleBot
4. *Implementation*: Fetch webpage, save webpage(HTML) and store it to a search engine(Nutch, Lucence, Elastic search), follow **ALL** links.

Web scraping

3. *Examples*: Baomoi, Sosanhgia, ...
4. *Implementaion*: Fetch specific webpages, extract the information needed, then follow the next resources **WE WANT**

The big picture

- Fetch web page(HTTP, XHR, Websocket)
- Save the data — information we need
- Decide what is the next resources to follow

Not so short introduction to HTTP

- Stateless protocol, connect and forget! (sessions, cookies, JWT)
- Method: GET, POST, PUT, PATCH, DELETE
- Status code: 1xx, 2xx, 3xx, 4xx, 5xx.
- RESTful webservices

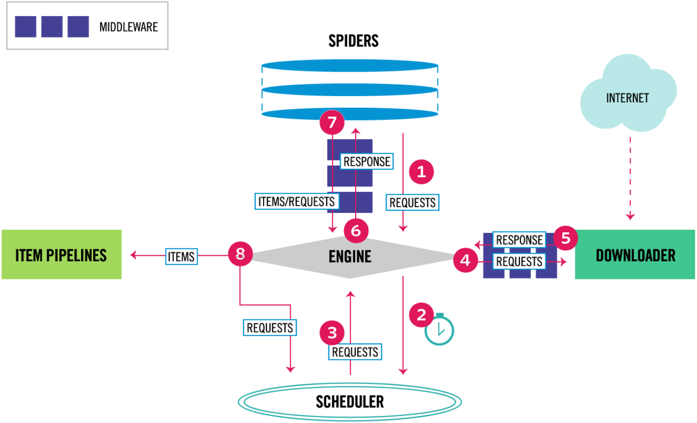
How HTTP impacts the scraping & crawling

- Authentication(Sessions, Cookies, JWT)
- Proxy
- User Agent
- POST request with Form data

Thinks as a Web developer

- Multi page app vs Single page app
- Which links to be followed next
- Basic of web app security(XSS, CSRF, ...)
- AJAX is your best friend!
- Mimick the HTTP headers(Man in the Middle - attack yourself)
- **Be flexible is all you need!**

Scrapy framework - 10k feet overview



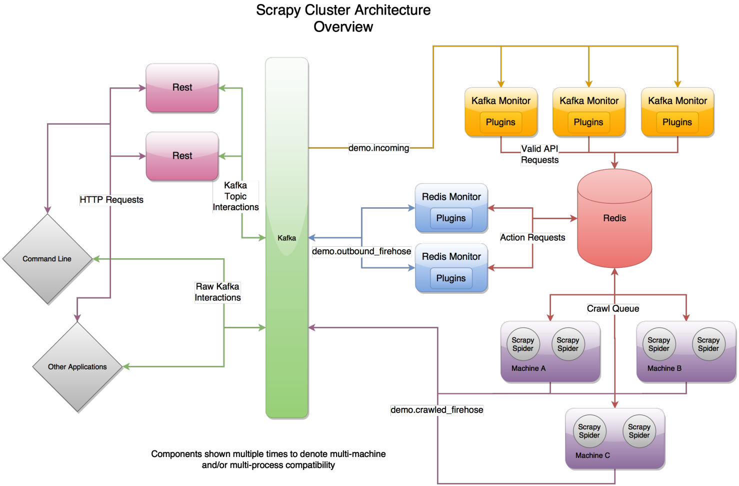
Javascript rendering sites: What we will do

- Selenium
- Scrapy + Splash
- Scrapy with AJAX

Scrapyd: Control your spiders over HTTP

- Efficent way to control spiders instead of having enter commands to your terminals.
- Can control, schedule multiple spiders, jobs at once.
- Rich supporting extensions: Monitoring(scrapyd-monitor), Clustering.

Scrapy Cluster: control thousands of spiders



Examples

- Scrape all articles from some categories from dantri.com.vn(MPA)
- Scrape all products, ratings from shopee(SPA + AJAX)
- Scrape all films, description, meta-data from tamnhinso.vn(JS rendering, Splash)