The Big Picture : Scrape any Website in 4 Steps

In this chapter, we overview on common web scraping steps. Actually this is really simple process.

Step 1 : Web scraping always start with a web page and data items which we want to scrape out from the page. At first step, we need to understand web page then find out HTML tags contain our wanted data. Result from this step will be used in final step to actually scrape data.

Step 2 : Wanted Data is located inside HTML page. We download HTML to local. We use Selenium for download.

Step 3 : After have HTML page, we use Beautiful Soup to parse HTML content in to object. Then we search for HTML tags which contain our data.

Step 4 : Final step is scraping data and store it to file or database. In this book we will make it simple by store data to file.

Selenium

We use Selenium to control browser and download HTML content.

Why we not just use simple library like requests to download HTML content ?.

Have 2 reasons:

- Many modern web page use a lot of JavaScript for dynamic HTML render, requests package could not render HTML from JavaScript.
- In some web pages, in order access wanted data, we need to do actions like : login, click link to navigate. Selenium can do that perfectly.

Beautiful Soup

We need Beautiful Soup to parse HTML in to object. Beautiful Soup provide functions help us to search HTML tags inside HTML object.

After have HTML tags, final step just about access wanted data and save.