Scrapy Tutorial #7: How to use XPath with Scrapy

Last updated on Feb 25 2019 by Michael Yin

Introduction:

This is the #7 post of my Scrapy Tutorial Series, in this Scrapy tutorial, I will talk about how to use XPath in scrapy to extract info and how to use tools help you quickly write XPath expressions.

Basic points of Xpath

First, we can did some tests on the homepage of Quotes to Scrape to understand the basic points of Xpath.

```
$ scrapy shell

In [1]: fetch("http://quotes.toscrape.com/")
```

In the code above, first we enter Scrapy shell by using `scrapy shell` commands, after that, we can use some built-in commands in scrapy shell to help us. For example, we can use `fetch` to help us to send http request and get the response for us. You can get the detail of the HTTP response by accessing property of the response object. There are many useful methods in response object, in the code below, we use the `xpath` method to extract info for us.

```
#If we want to get html node
response.xpath("/html").extract()

#If we want to get body node, which is the child of html node
response.xpath("/html/body").extract()

#If you want to get all div descendant of this html
response.xpath("/html//div").extract()

#we can also drill down without having to start with /html, this expression would extrac
response.xpath("//div").extract()
```

From the code above, you should know how to use `/` and `//` to select the node. If you want to filter all div elements which have `class=quote`

```
response.xpath("//div[@class='quote']").extract()

# you can use this syntax to filter nodes
response.xpath("//div[@class='quote']/span[@class='text']").extract()

# use text() to extract all text inside nodes
response.xpath("//div[@class='quote']/span[@class='text']/text()").extract()
```

You should copy the code to your terminal, check the output, to make sure you really understand how it works.

Advanced Xpath

Many people like to learn XPath by reading the Cheatsheet or online doc, however, I do not think this is a good way because many patterns would not be used in many cases. It is better to search the answer when you have encountered specific problem of XPath, if you still have problem after these steps, you can leave me message here.

Here are some good resources for you to check when you have problem.

1. Xpath cheatsheet

2. xml path language

How to get XPath in Chrome

To make you quickly get the XPath in Chrome, it is recommended to install Chrome Extension called XPath Helper, I would show you how to use this great extension.

1. Press `Command+Shift+x` or `Ctrl+Shift+x` to activate it in web page, you will console in page.
2. Press `Shift`, then move your mouse, then the console will show the XPath expression and the right side will show the result.
3. In most cases, the XPath expression generated in the console is very long, so you can edit if you like. You can edit the XPath query directly in the console. The results box will immediately reflect your changes, which is the most powerful feature of this Plugin.

What you should notice is that, sometimes the HTML elements and property can be modified by Javascript, which means the XPath expression which works in your browser might not work in XPath shell, so you should test all XPath expressions in your scrapy shell before writing it in your code

How to get XPath in Firefox

FirePath is a FIrebug Extension which can generate XPath for you, it is very easy.

1. Install FireBug, which is a prerequisite to install FirePath.
2. Install FirePath. Remember to restart firefox after installation.
3. Right-click on the element you want to extract and select "Inspect in FirePath".
4. You can see the XPath generated in the box

What you should notice is that sometimes the HTML elements and property can be modified by Javascript, which means the XPath expression which works in your browser might not work in XPath shell, so you should test all XPath expressions in your scrapy shell before writing it in your code

Conclusion

In this scrapy tutorial, we learned how to how to use XPath in scrapy to extract info, if you have any questions about your project, just left a message here and I will respond ASAP. What is more, you really should install the plugin mentioned above to increase your productivity, it can help you a lot if you have not much experience with Xpath.