Scrapy Tutorial #11: How To Extract Data From Native Javascript Statement

Last updated on Feb 26 2019 by Michael Yin

CSS expression and XPath expression are not silver bullet

When scraping some web pages, the data is included in some native javascript statement (js object), we need to find out a way to extract the data without importing heavy browser such as phantomjs. `css` expression and `xpath` expression can not get this job done well, we need other options to solve this problem.

If you search `Scrapy tutorial` on google, most would tell you how to use `css` or `xpath` expression to extract the data as you like. However, `regex` is another powerful tool helping you extract the data in some cases, and it seems it has been ignored by most people.

In this Scrapy tutorial, I will show you how to extract data from native javascript statement using Regex and Json. But you can also use the method when developing web scraper using other frameworks such as BeautifulSoup.

Inspect HTML source

Let's assume the target web page we are crawling is `http://quotes.toscrape.com/js/`, If you check the html source code of that page, you will find out the data is in javascript object called `data`.

```
var data = [
{
    "tags": [
        "change",
        "deep-thoughts",
        "thinking",
        "world"
    ],
    "author": {
        "name": "Albert Einstein",
        "goodreads_link": "/author/show/9810.Albert_Einstein",
        "slug": "Albert-Einstein"
    },
    "text": "\u201cThe world as we have created it is a process of our thinking. It canr
},
......
```

The javascript in the web page would iterate the `data` object and create DOM when the page is rendered in web browser, considering it is not included in raw html, so you can not use `xpath` expression or `css` expression to extract the data in Scrapy or BeautifulSoup.

Before we get started, you should have a basic understanding about what Json is.

> JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language

Here, we need extract the valid json text from the HTML source code, and then use `json` library in Python to load the data, after that we can easily access the data as we like.

Regex to extract the data, JSON to load the data

> A regular expression is a special text string for describing a search pattern. You can think of regular expressions as wildcards on steroids.

First, we can enter Scrapy shell and import `re` library from Python.

```
$ scrapy shell 'http://quotes.toscrape.com/js/'

In [1]: import re
```

Now we need to extract the javascript list from the code block

```
In [2: data = re.findall("var data =(.+?);\n", response.body.decode("utf-8"), re.S)
```

Let me explain this statement. We use `re.findall` method helping us to find matched text from `response.body`. `var data =(.+?);` here is the regex expression we created, The meaning of the expression is to find all content between `var data` and the first `;\n`. The `re.S` is regex mode passed with regex expression, here `re.S` told regex engine that the dot sign in expression match linebreak.

If the `re.findall` can find something mateched, it will retuen it back, if not, it will return `None`. After extracting the json data text, we can use json library in Python to load the data so we can easily access the data.

```
import json

ls = []
if data:
    ls = json.loads(data[0])

if ls:
    print(ls)
```

Now we can see the data can be accessed just like Python native dict and list.

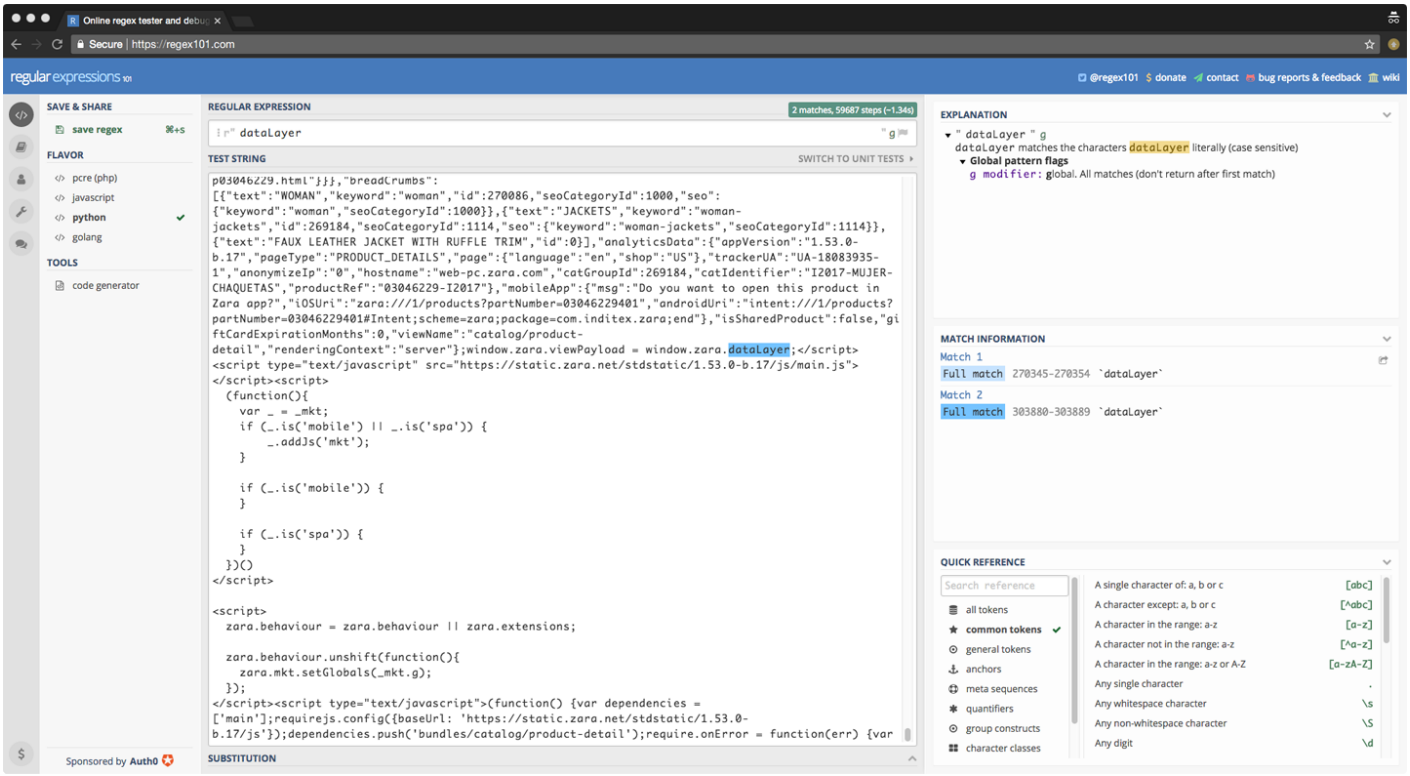
```
[{'author': {'goodreads_link': '/author/show/9810.Albert_Einstein',
   'name': 'Albert Einstein',
   'slug': 'Albert-Einstein'},
  'tags': ['change', 'deep-thoughts', 'thinking', 'world'],
  'text': '“The world as we have created it is a process of our thinking. It cannot be c
 {'author': {'goodreads_link': '/author/show/1077326.J_K_Rowling',
   'name': 'J.K. Rowling',
   'slug': 'J-K-Rowling'},
  'tags': ['abilities', 'choices'],
  'text': '“It is our choices, Harry, that show what we truly are, far more than our abi
 {'author': {'goodreads_link': '/author/show/9810.Albert_Einstein',
   'name': 'Albert Einstein',
   'slug': 'Albert-Einstein'},
  'tags': ['inspirational', 'life', 'live', 'miracle', 'miracles'],
  'text': '“There are only two ways to live your life. One is as though nothing is a mir
  ......
]
```

Are there some good resources or tools for me to learn regex?

If you want to learn regex expression, you can check doc of `re` module.

- python2 re module
- python3 re module

If you want to debug your regex expression, you can check regex101, which is great online free tool for you to debug your regex expression.



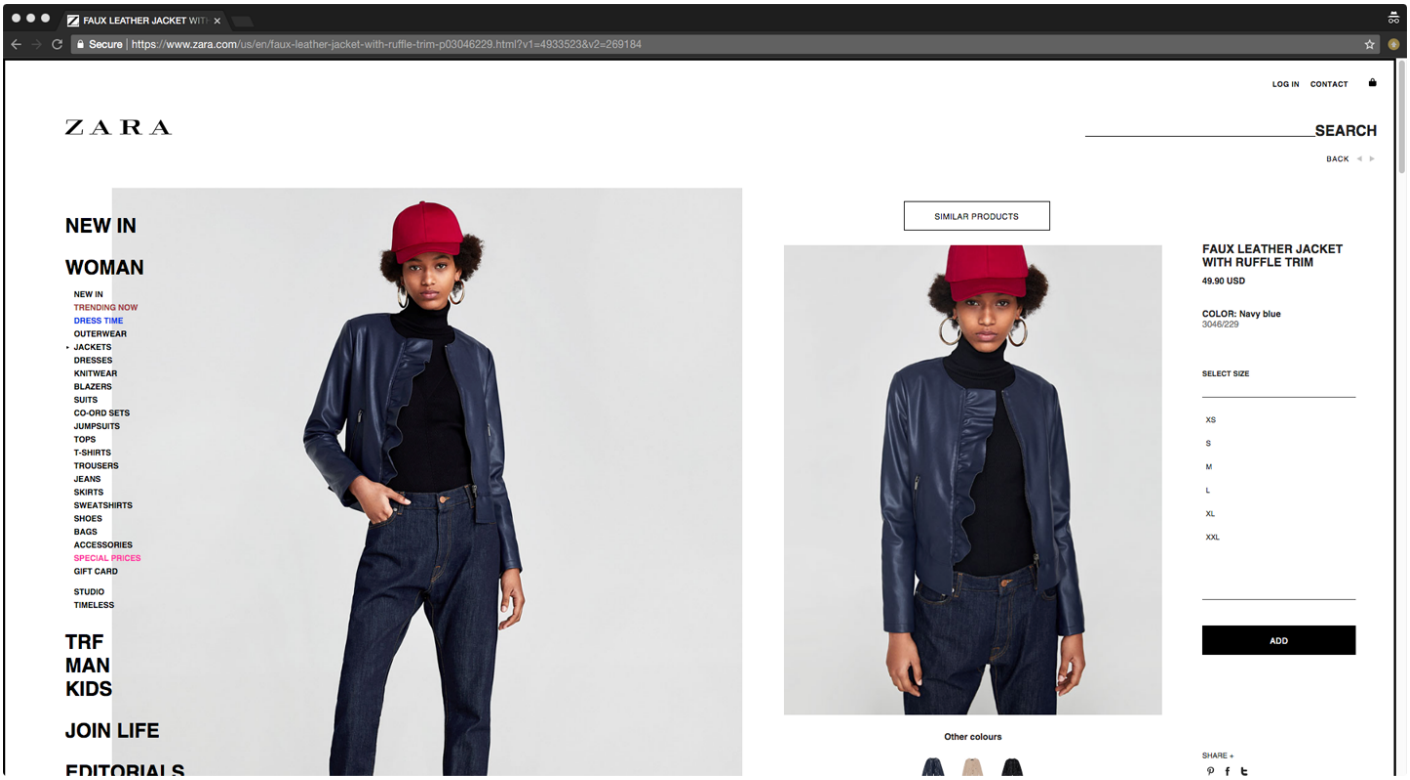
Are there any other ways to crawl data from javascript

The data of javascript would be used to create the html element of web page, so if you find it hard to write regex in scrapy spider, you can use Selenium to get the job done.

Real world example

Now I can give you one real example for you to better understand the key point.

For example, your boss need you to collect some data about fashion industry, then you might want to write spider to get the data of zara.com, when scraping the data of zara products, you figure out that the detail of product is included in javascript code block, so you can use the methods I talked above to quickly extract the data.



You can find the detail of the product in the code block like

```
<script data-compress="true" type="text/javascript">window.zara.appConfig =
......
......
window.zara.viewPayload = window.zara.dataLayer;</script>
```

Now you can enjoy your coffee now.