blueanchor

Scraping avec Javascript

par Julien Moulin, Christophe Mendes et Thomas Paoli

github.com/materazu/meetup-javascript-scraping

Généralités

- Le scraping est une technique permettant d'extraire du contenu web pour le manipuler

- Il peut concerner une page web, un flux xml, une réponse en json ou toute données accessible depuis une url

- Avec un peu de technique, on peut aller jusqu'à automatiser des processus avec le scraping (et les bons outils)

Dans le navigateur

Jquery et manipulation

jQuery

- Bibliothèque javaScript libre, créée en 2006
- Permet la navigation dans le DOM, la gestion d'évènement, la manipulation des feuilles CSS ...
- Plugin disponible pour les navigateurs :

Navigation dans le DOM

- Récupérer le <body>...</body> de la page
- Chercher l'élément qui nous intéresse
- Traiter les données

Avantage

- Rapide
- Léger
- Facile

Inconvénient

- Pas de persistence
- Méthode manuelle

En console

Cheerio : du jQuery, mais dans node

Cheerio

- Implémentation de jQuery coté Node.js
- Permet de traverser le dom en le chargeant dans l'outil, et de le modifier, aussi
- Bien plus rapide que les concurrents en ce qui concerne le parsing de data
- Attention, ce n'est pas un navigateur, donc ni js, ni css dans les structures

Charger une structure

```
const cheerio = require('cheerio');
const $ = cheerio.load('<ul id="fruits">...</ul>', {
  normalizeWhitespace: true,
  xmlMode: true
});
```

https://github.com/fb55/htmlparser2/wiki/Parser-options

Selecteurs

```
$('.apple', '#fruits').text()
//=> Apple

$('ul .pear').attr('class')
//=> pear

$('li[class=orange]').html()
//=> Orange
```

Comme avec jQuery !

Attributs

```
$('.apple').attr('id', 'favorite').html()
//=> <li class="apple" id="favorite">Apple</li>

$('input[type="checkbox"]').prop('checked')
//=> false

$('<div data-apple-color="red"></div>').data()
//=> { appleColor: 'red' }

$('input[type="text"]').val()
//=> input_text

$('.pear').hasClass('pear')
//=> true

$('<form><input name="foo" value="bar" /></form>').serializeArray()
//=> [ { name: 'foo', value: 'bar' } ]
```

Formulaires

```
$(`<form>
   <input name="foo" value="bar" checked />
   <input name="foo" value="qux" checked />
  </form>`).serialize()
//=> foo=bar&foo=qux
```

```
$('<form><input name="foo" value="bar" /></form>').serializeArray()
//=> [ { name: 'foo', value: 'bar' } ]
```

Traverser le dom

```
$('#fruits').find('li').length
//=> 3

$('#fruits').find($('.apple')).length
//=> 1

$('.pear').parent().attr('id')
//=> fruits

$('.orange').parents().length
// => 2

$('.orange').parents('#fruits').length
// => 1
```

Exemple de script

On mange où ?

Manipuler le web

CasperJs, le petit fantôme qui automatise

CasperJS

Qu'est ce que c'est?

- Un utilitaire de navigation et de tests pour le navigateur
- Une librairie qui fournit des méthode avec une syntaxe simple et rapide
- Permet **SURTOUT** de récupérer des données

Pré-requis

- Python 2 ou 3
- PhantomJS

Comment ça marche?

Des questions ?

Merci à tous !

Et merci à la Wild Code School
d'avoir hébergé notre Meetup.

Mainteant, place à l'apéro !