Le scrap : accessible à tous

@Julien_Gadanho

HACK THE PLANET

Les robots vont supprimer 10 000 postes chez Foxconn cette année

Les salariés d'Innolux trop chers

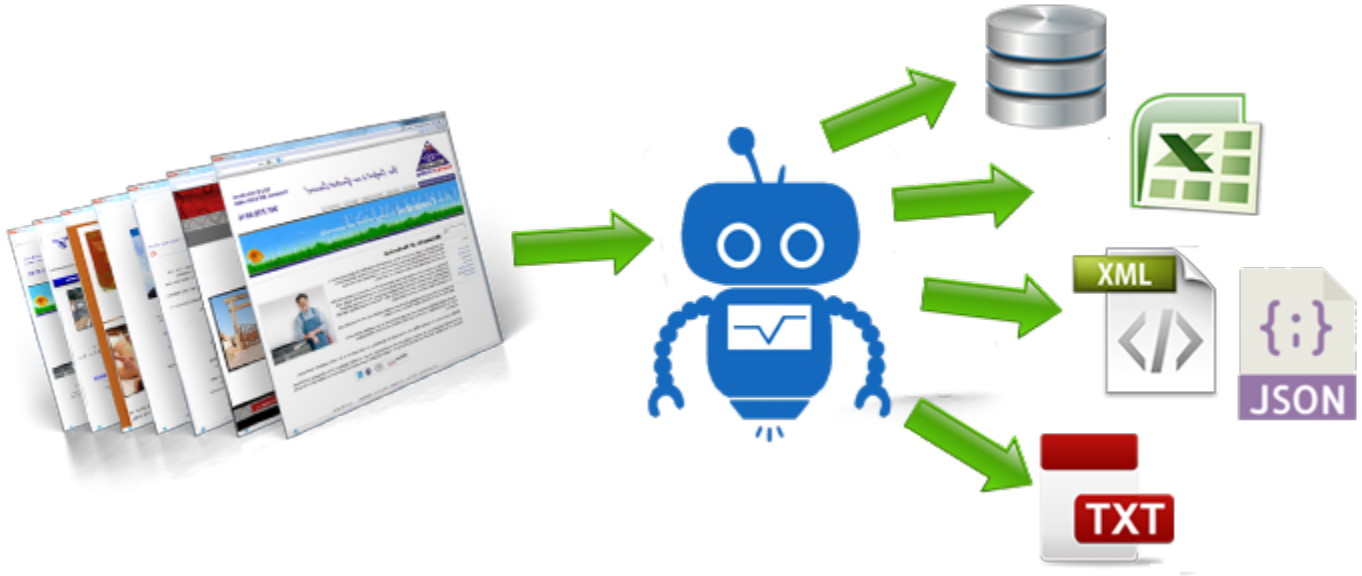
Publié le 13/02/18 à 16h49

Foxconn a régulièrement été dans le viseur d'associations dénonçant les conditions de travail déplorables et les salaires trop faibles de ses employés. Des plaintes qui ne cibleront plus les robots, de plus en plus nombreux au sein de ses usines.

“Aucun bateau n'est **ni bon ni mauvais**. Tu es libre de construire n'importe quel navire dont tu as envie ! Mais retiens bien ceci: qu'il prenne des vies ou qu'il précipite le monde dans le chaos, **un créateur ne doit jamais renier un navire qu'il a conçu** ! Il doit toujours l'aimer ! Le navire n'est jamais responsable ! SOIS FIER DE CE QUE TU AS CONSTRUIT!”

Le web scraping : définition

Extraction du contenu d'une page web à l'aide d'un **script**. Le but est de réutiliser ces données.



Cas concrets

- Veille tarifaire
- Veille des stocks
- Centralisation de vos datas
- Récupération de données sur des sites n’ayant pas d’API
- Intégration d’images / vidéos

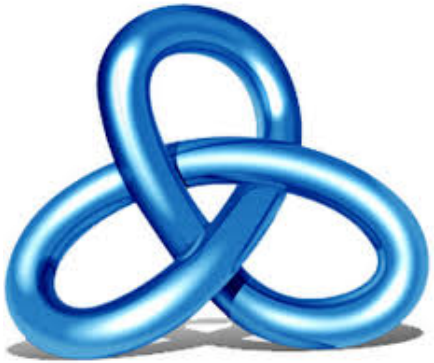
Le web scrapping : juridique

Les technologies

Scrapy

Laquelle choisir ?

			Casper.js
Temps de dev	Rapide	Lent	Moyen -
Stabilité	Faible	Forte	Moyen
Performance	Faible	Forte	Moyen
Flexibilité	Forte	Faible	Moyen

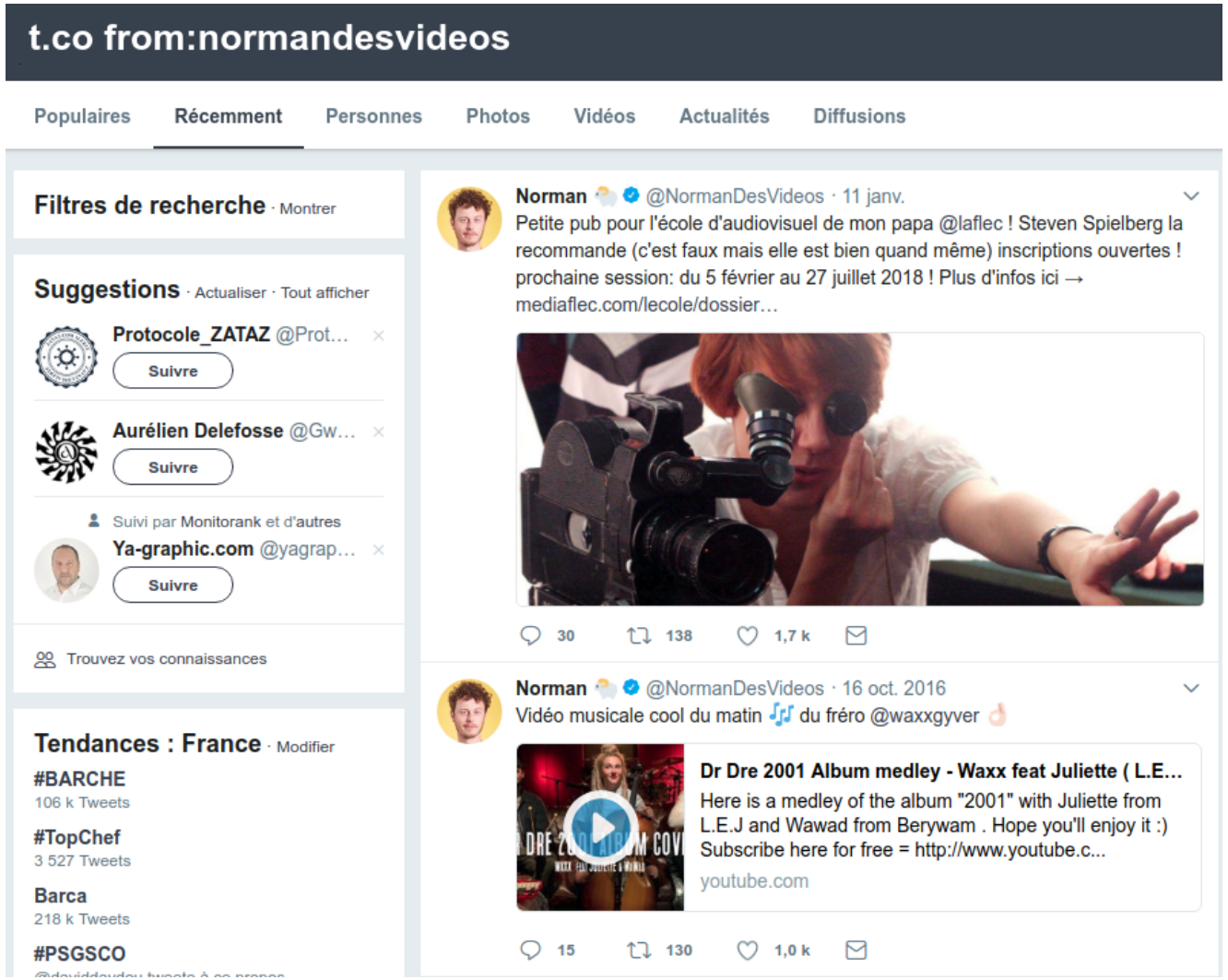


Zennoposter

Hijacking Twitter

Hijacking Twitter

Footprint Twitter



Identifier la cible

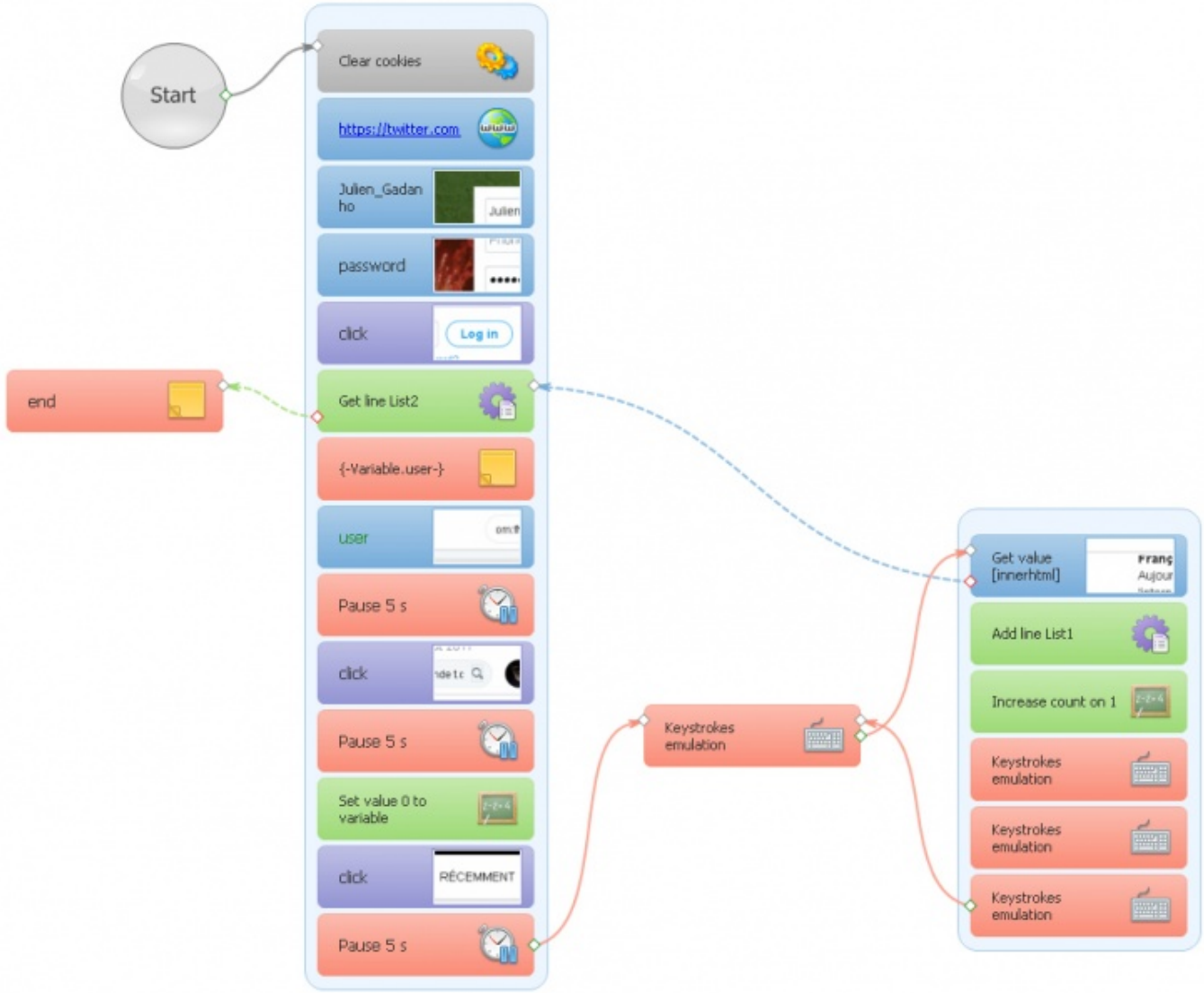
Twitter HTML

```
▼<p class="TweetTextSize  js-tweet-text tweet-text" lang="fr" data-aria-label-part="0">
  "Petite pub pour l'école d'audiovisuel de mon papa "
 ▶<a href="/laflec" class="twitter-atreply pretty-link js-nav" dir="ltr" data-mentioned-user-id="460241992">…</a>
  " ! Steven Spielberg la recommande (c'est faux mais elle est bien quand même) inscriptions ouvertes ! prochaine session: du 5 février au 27 juillet 2018 ! Plus d'infos i
 ▶<a href="https://t.co/WeKMrEFHeG" rel="nofollow noopener" dir="ltr" data-expanded-url="http://www.mediaflec.com/lecole/dossier-de-candidature/" class="twitter-timeline-l
www.mediaflec.com/lecole/dossier-de-candidature/">…</a>
  <a href="https://t.co/bUthfeunsW" class="twitter-timeline-link u-hidden" data-pre-embedded="true" dir="ltr">pic.twitter.com/bUthfeunsW</a>
 </p>
```

Element HTML

p.TweetTextSize js-tweet-text tweet-text

Template


Start
Clear cookies
https://twitter.com
Julien_Gadanho
password
click
Log in
end
Get line List2
{-Variable.user-}
user
Pause 5 s
click
Pause 5 s
Set value 0 to variable
click
RÉCEMMENT
Pause 5 s
Keystrokes emulation
Get value [innerhtml]
Add line List1
Increase count on 1
Keystrokes emulation
Keystrokes emulation
Keystrokes emulation

Template

Get value

What to take:

innerhtml

Classic item search:

Which tab: Active
First
By name:
By number:

Document: 0 Form: -1 Tag: p

Search criterias:

	Group	Att...	Search t...	Value	Match #
	0	class	regexp	TweetTextSize\ \ js-tweet-text\ tweet-text	{-Variable.count-}

Put to variable

Variable3

CasperJs – Headless Browser

Projet (photo non contractuelle)

Initialisation du script

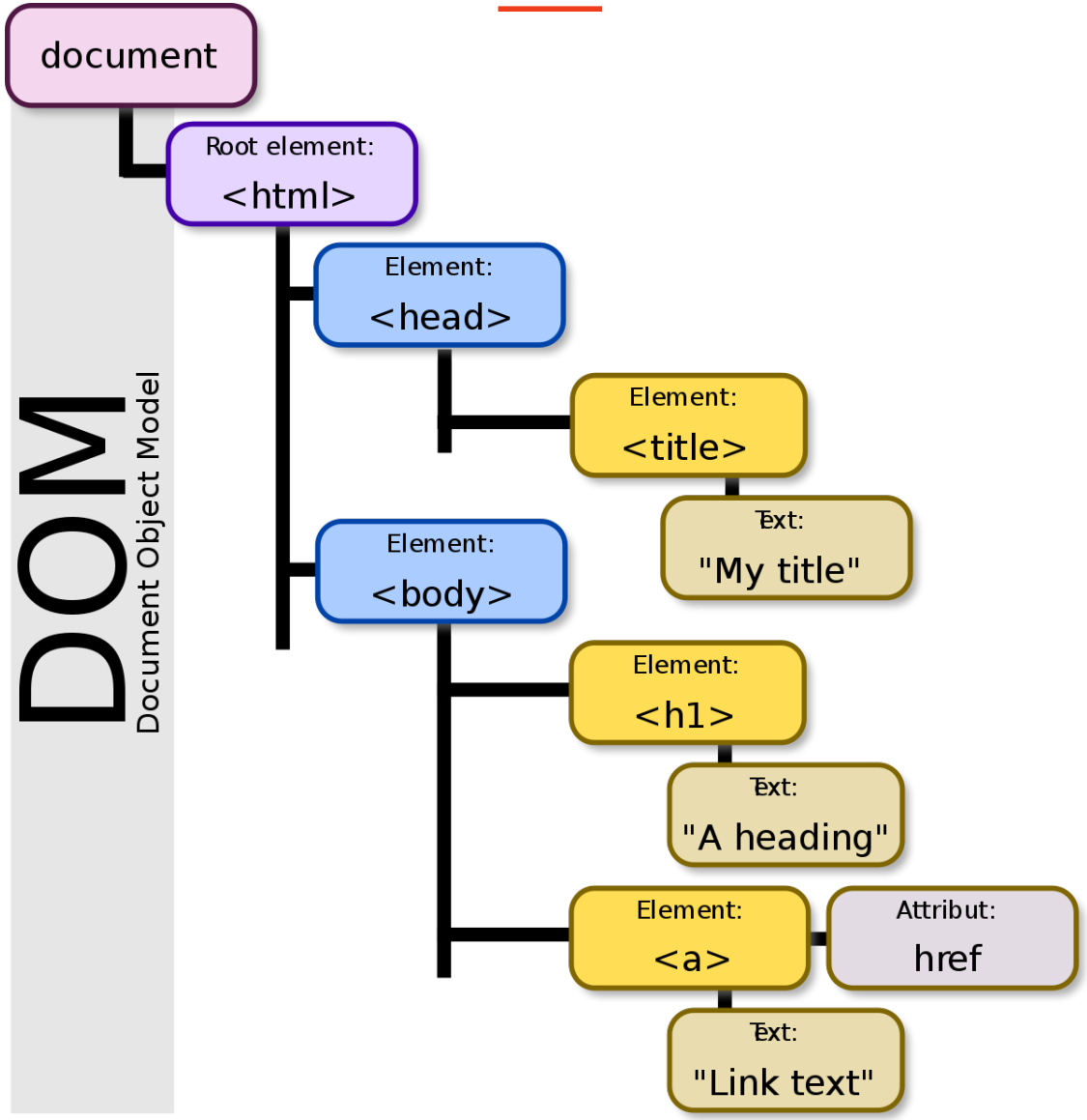
```
//ini casper
var casper = require('casper').create();

//include xpath module
var x = require('casper').selectXPath;


casper.userAgent('Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)');


//visit homepage
casper.start('https://partenaires.amazon.fr/');
```

DOM


document
Root element:
<html>
Element:
<head>
Element:
<title>
Text:
"My title"
Element:
<body>
Element:
<h1>
Text:
"A heading"
Element:
<a>
Attribut:
href
Text:
"Link text"
DOM
Document Object Model

Connexion

```
//click on login
casper.thenClick(x('//*[@id="a-autoid-0-announce"]'), function() {});

//fill form
casper.then(function() {
    this.evaluate(function () {
        document.getElementById("ap_email").value = "youremail@yourdomain.com";
        document.getElementById("ap_password").value = "yourpassword";
        document.getElementsByTagName('form')[0].submit();
    });
});
```

Récupération du solde

```
//get money
casper.wait(1000, function () {
        var text = casper.fetchText(x("(//div[@class='ac-card-data-content'])[1]"));
        casper.echo(text.trim().replace("€","").replace(",",."));
});


//run application
casper.run();
```

GetMoney.js

```
//ini casper
var casper = require('casper').create();

//include xpath module
var x = require('casper').selectXPath;


casper.userAgent('Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)');


//visit homepage
casper.start('https://partenaires.amazon.fr/');

//click on login
casper.thenClick(x('//*[@id="a-autoid-0-announce"]'), function() {});

//fill form
casper.then(function() {
    this.evaluate(function () {
        document.getElementById("ap_email").value = "youremail@yourdomain.com";
        document.getElementById("ap_password").value = "yourpassword";
        document.getElementsByTagName('form')[0].submit();
    });
});

//get money
casper.wait(1000, function () {
        var text = casper.fetchText(x("(//div[@class='ac-card-data-content'])[1]"));
        casper.echo(text.trim().replace("€","").replace(",",."));
});


//run application
casper.run();
```

Bonus : création d'un webservice

```php
<?php

echo exec('casperjs getMoney.js');
```

Simple HTML Dom

Demo

Conclusion

- Une infinité de technologies pour le web scrapping
- Elles sont plus ou moins complexe
- Simplifiez vous le travail

Sponsors

Oncrawl

netlinking.fr

linkjuice.io

TRUSTPILOT

AUTHORITAS

BOTIFY

MONITORANK

textbroker
Chaque Mot Compte

SEOQUANTUM
OPTIMISEZ VOTRE SEO GRÂCE À LA SÉMANTIQUE

DEEPCRAWL

MAJESTIC

youdot
FIND YOUR VISIBILITY

SF
SEARCH · FORESIGHT

Myposeo®

ROCKETLINKS