Web Scraping

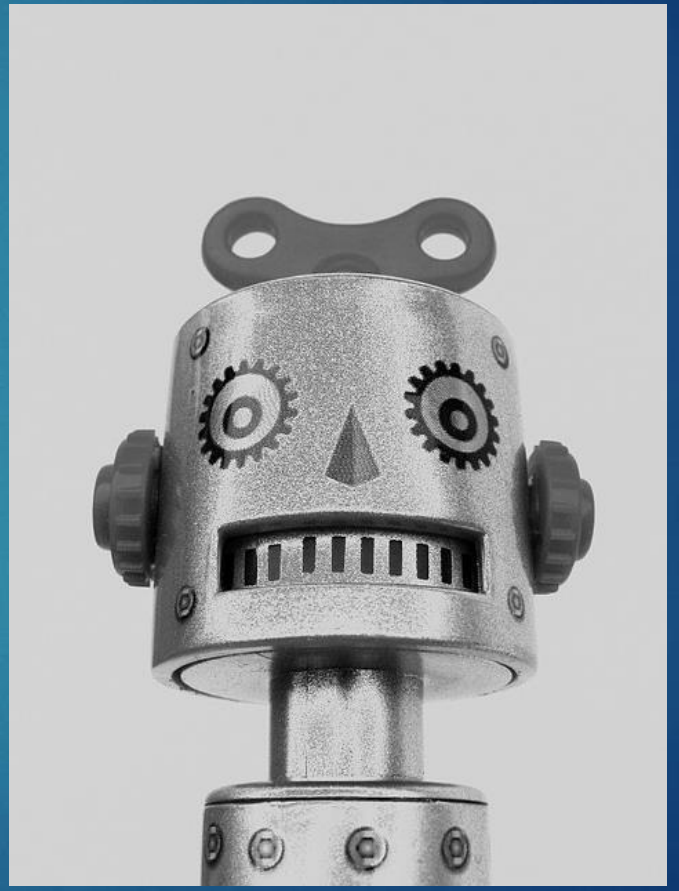
AVOIR UNE API LÀ OÙ ELLE N'EXISTE PAS

Web scraping, Was ist das?

Ce que nous dit Wikipedia :

Le web scraping (parfois appelé Harvesting) est une technique d'extraction du contenu de sites Web, via un script ou un programme, dans le but de le transformer pour permettre son utilisation dans un autre contexte, par exemple le référencement.

Pour faire simple, c'est un robot qui se balade sur votre site !



Ce que dit la loi

Article L341-1 du Code de la propriété intellectuelle :

" Le producteur d'une base de données, entendu comme la personne qui prend l'initiative et le risque des investissements correspondants, bénéficie d'une protection du contenu de la base lorsque la constitution, la vérification ou la présentation de celui-ci atteste d'un investissement financier, matériel ou humain substantiel.

Cette protection est indépendante et s'exerce sans préjudice de celles résultant du droit d'auteur ou d'un autre droit sur la base de données ou un de ses éléments constitutifs."

Dans quels cas utiliser du scraping ?

- Référencement
- Extraction de données pour statistiques
- Pas d'API existante
- Test d'intégration
- …

Quel langage choisir ?

- Votre langage de programmation peut faire une requête HTTP vers l'extérieur ?

IL EST PARFAIT !

- Votre langage gère l'asynchrone ?

C'EST ENCORE MIEUX !

Coder un scraper, simple ?

Malgré l'idée, faire son scraper est :

- Simple
- Rapide

Je vous le prouve ?

Coder un scraper, simple ?


```
npm install scrap
```

Voilà, vous avez fait le plus dur …

request + cheerio = 

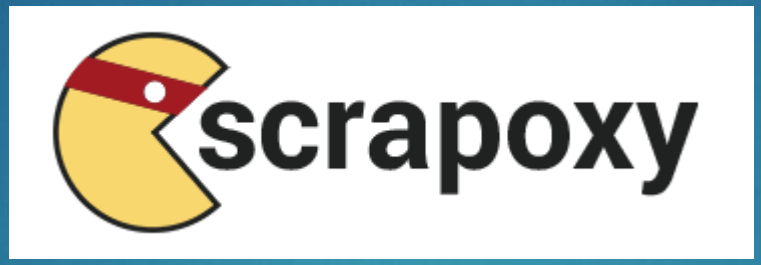
La notion de vitesse

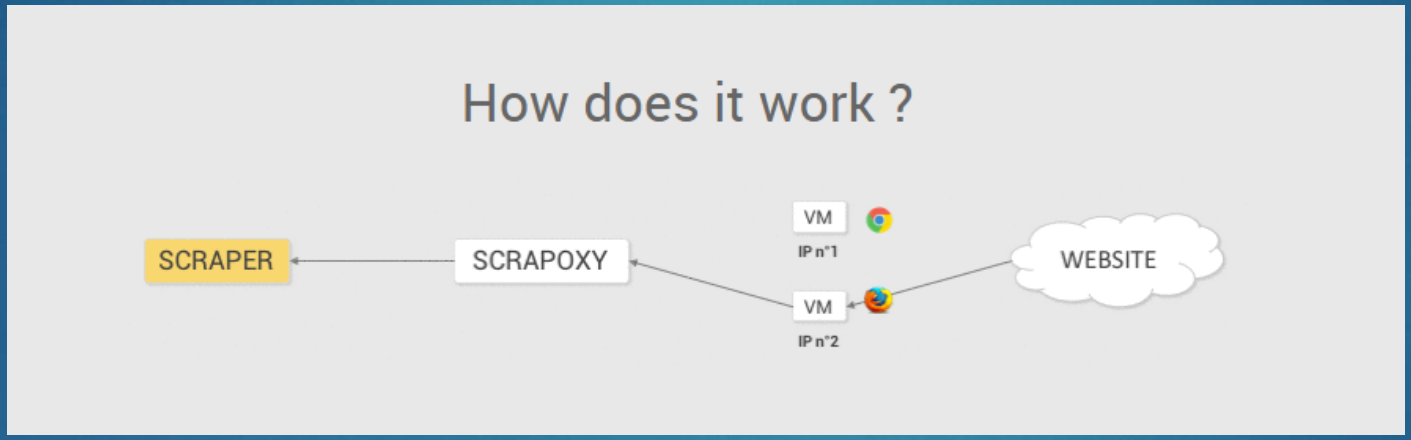
Il est toujours important de faire du scrap *éthique*.

Il faut toujours paraitre comme un humain et non un robot.

Un humain ne peut pas cliquer sur 2500 liens à la fois.

Passer un proxy





Scrapoxy permet de passer par des serveurs à part

Cliquer et faire des screenshots


```
npm install -g phantomjs
npm install -g casperjs
```

Le cas des sites en JS

Alternative au code

- YQL (Yahoo Query Language)
- import.io

Quelques cas de scraping

À votre avis ?

À VOUS DE JOUER