Le Web, comment ça marche ?

Kossi Neroma

07/01/2020

Web scraping : qu'est-ce que c'est ?

Le Web scraping désigne les techniques d'extraction du contenu des sites internet. Des programmes informatiques se comportant comme des robots parcourent le Web et y extraient les données jugées intéressantes. Souvent, deux termes sont utilisés sans beaucoup de distinction pour désigner l'extraction de contenus : **scraping** et **crawling** . En fait, le **crawling** se fait en général à grandes échelles et par des robots des plateformes comme google, Yahoo et MSN. Par contre, le **scraping** se fait souvent à "petite" échelle et les scripts sont écrits pour analyser des pages web bien spécifiques (comparateur de prix, extracteur de données financières, automobiles ...). Dans ce cours, il n'est pas nécessaire de faire grande différence entre les deux concepts mais il faut tout de même garder à l'esprit que les deux ne sont pas tout à fait les mêmes.

Assez de parler de choses non concrètes ! J'imagine que vous bouillonnez d'envie de **scraper** un peu le Web. Et alors, comment allons-nous s'y prendre ? Comme d'habitude, il s'agira d'identifier les bonnes librairies et de savoir comment les utiliser à bon escient. Il y aura quelques bonnes habitudes à adopter et c'est surtout la pratique qui vous consacrera **scrapeur** de l'année.

Sinon, et le Web, qu'est-ce que c'est ?

Même si je ne vais pas vous faire un cours sur le développement Web (ce n'est ni mon travail ni le but de ce cours) il nous faut néanmoins certaines bases pour comprendre comment *Internet* et ses milliers de pages fonctionnent.

Un site Web est un ensemble de pages écrits en HTML. **HTML** est un langage de balisage qui permet de décrire à la fois le contenu et la mise en forme d'une page Web. Une page n'est ni plus ni moins qu'une chaîne de caractères structurées, un peu comme un document **Word**. Les éléments de structuration ou de mise en forme pour faire simple, sont appelés **éléments HTML** (**HTML elements** en anglais). Un élément **HTML** commence toujours par **une balise ouvrante** et se termine souvent par une **balise fermante**. Il est assez courant d'emploier l'équivalent anglais **tag HTML** au détriment de son ami français **balise HTML** (l'anglais est vraiment la lingua franca de l'informatique, et un peu de tout aujourd'hui). Revenons à nos moutons, ou plutôt à nos tags, ou balises si vous voulez. Comme dit-on souvent, *une image vaut 100 mots* :

- ** **voici un élément fourre-tout, dit élément **div**, qui sert à regrouper d'autres tags** **
- **<p>**Voici un élément, dit élément **p** qui sert à délimiter un paragraphe**<p>**
- **<a>**Voici un élément, dit élément **a** qui sert à définir un lien vers une ressource Web**<p>**
- **<body>**Voici un élément, dit élément **body** qui encadre toujours le corps de tout document **HTML<body>**
- **
** Voici un élément orphelin qui ne possède qu'une seule balise (ouvrante).

```
Le mot 'balise html' est souvent confondu avec 'élément html'. On verra donc sur le Web certains utilis
```

Du HTML, mais pas que !

Vous souvenez-vous ? Dans le paragraphe précédent je semblais dire que le Web est fait de pages HTML qui ne sont ni plus ni moins que des amas de texte. Ce n'est pas tout à fait vrai. Une "bonne" page Web contient généralement plus que du HTML. Le langage qui permet de donner vie au document HTML s'appelle CSS (**C**ascading **S**tyle **S**heet). Il permet de former la peau du site Web en lui donner sa couleur, sa disposition et son oragnisation d'ensemble. Si HTML définit des possibilités de mise ne page, c'est le CSS qui vient spécifier clairement comment cette opération de mise en forme sera effectuée.

HTML + CSS, est-ce tout ?

On est presqu'au bout de la liste. Vous vous doutez bien au vu de la question posée que ce n'est pas tout ! Il reste sûrement quelque(s) chose(s). Mais qu'est-ce que c'est ? Eh bien, c'est du JavaScript. En effet sans **JavaScript** les pages Web ressembleraient beaucoup à un document Word ou tout simplement à un pdf : un contenu statique qu'on visionne encore et encore ! Vous conviendrez avec moi que c'est pas le cas pour bon nombre de pages Web. Prenons l'exemple d'un site de paris sportifs par exemple, le contenu change presque à la seconde ! Le langage qui permet de réaliser cette prouesse s'appelle **JavaScript**. C'est le trio **HTML + CSS + JavaScript** qui permet forment l'architecture cachée de tous les sites Web présents sur Internet.

Le javascript vient ajouter une surcouche de complexité à la tâche du scrapeur en ce sens qu'il rend le contenu du site dynamique. Il ne s'agit donc plus suffisant d'aller télécharger le code HTML de la page Web et d'extraire son contenu une fois pour toute. Il va falloir répéter cette opération régulièrement et opérer des actions comme *remplir un formulaire, saisir ses identifiants pour s'authentifier, voire vérifier un* **CAPTCHA**.

Rassurez-vous, il n'existe de problème sans solutions, en tout cas sans éléments de solution : il existe bel et bien des outils python qui permettent d'activer les scripts javascript contenus dans le code source d'une page HTML afin de rendre disponible les données d'intérêt. Nous reviendrons sur tous ces points dans la suite de cours quand nous allons aborder les pages dynamiques.

Résumé

Dans cette deuxième partie de notre cours, nous avions rapidement aborder :

- Le fonctionnement du Web
- Le langage HTML, ses éléments et ses balises (*tags*)
- Le langage de feuille de style **CSS**
- Le langage **JavaScript** qui permet d'ajouter de l'interactivité aux pages Web