Une image contenant Police, logo, texte, Graphique

Description générée automatiquement

**Département D’INFORMATIQUE**

**Veille Technologique**

**420-1SH-SW**

**Le Web Scraping**

**Présenté à**

**M. Nicolas Bourré**

**09 juin 2023 Par : Olivier Lefebvre**

Table des matières

[Introduction 1](#_Toc137029352)

[Développement 1](#_Toc137029353)

[Son histoire 1](#_Toc137029354)

[Qu’est-ce que c’est? 2](#_Toc137029355)

[Les outils et langages 2](#_Toc137029356)

[Les techniques 3](#_Toc137029357)

[HTTP programming 3](#_Toc137029358)

[Text pattern matching 3](#_Toc137029359)

[HTML parsing 3](#_Toc137029360)

[Computer vision web-page analysis 3](#_Toc137029361)

[Crawler VS Scraper 3](#_Toc137029362)

[Les avantages 4](#_Toc137029363)

[Les défis 4](#_Toc137029364)

[Domaine où le web scraping est utilisé 4](#_Toc137029365)

[Les cas d’utilisations possibles 5](#_Toc137029366)

[Est-ce légal ? 5](#_Toc137029367)

[Utiliser le web scraping de façon éthique 6](#_Toc137029368)

[Exemples de web scraper 6](#_Toc137029369)

[Premier exemple 6](#_Toc137029370)

[Deuxième exemple 7](#_Toc137029371)

[Conclusion 10](#_Toc137029372)

[Rappel 10](#_Toc137029373)

[Ouverture 10](#_Toc137029374)

[Médiagraphie I](#_Toc137029375)

# Introduction

De nos jours, les technologies jouent un rôle crucial dans nos vies et elles sont devenues essentielles au bon fonctionnement du monde moderne. Elles offrent de nombreux avantages, notamment en facilitant l'accès à l'information. Aujourd'hui, il est possible pour tout le monde d'obtenir une quantité incroyable d'informations sur presque tous les sujets en quelques clics.

Cependant, la collecte de données sur le web peut être un défi. De nombreuses personnes utilisent la méthode du copier-coller, mais cela peut prendre beaucoup de temps. C'est là que le *web scraping* intervient. Le *web scraping* est un outil puissant qui permet d'automatiser la collecte de données sur le web et de les transformer dans le format souhaité.

Cependant, le *web scraping* soulève des questions : est-ce légal ou éthique dans différents contextes? Comment fonctionne-t-il ? Dans quels domaines le web scraping est-il utilisé ? Ces questions sont essentielles pour bien comprendre cet outil qu'est le *web scraping*.

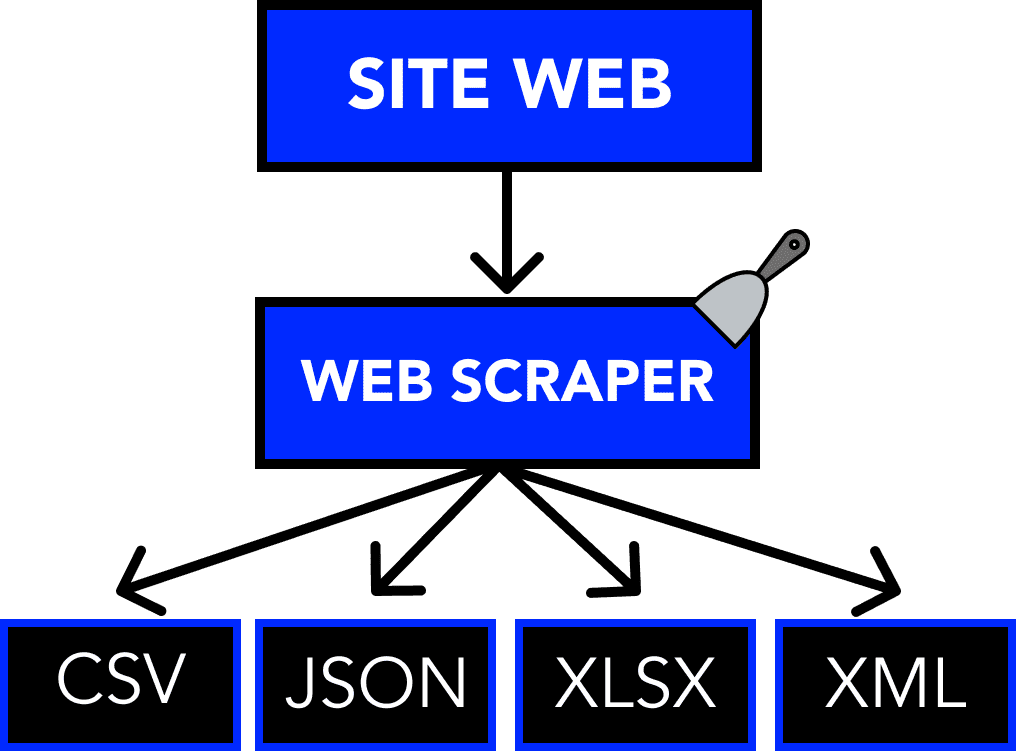
# Développement

## Son histoire

Le *web scraping* ne date pas d’hier, il est présent depuis le tout début du *World Wide Web*. Il a cependant pris plusieurs formes. En effet, en 1993, *World Wide Web Wanderer* fut le premier robot sur le web. Il a cependant été conçu que pour mesurer l’étendue du web. Ensuite, toujours en 1993, le premier moteur de recherche basé sur un robot est né. Il s’agit de *JumpStation*. En 2000, deux entreprises ont pris les devants dans le monde de la collecte de données sur le web. En effet, *Salesforce* et *eBay* ont introduit leur propre *API*. Ceci a permis aux développeurs d’avoir la possibilité d'accéder à une sélection de données disponibles au grand public et de les récupérer. Par la suite, plusieurs entreprises ont emboité le pas. Cependant, ce ne sont pas tous les sites web qui ont une *API*. C’est pour cela qu’en 2004, une nouvelle bibliothèque Python a été créée. En effet *Beautiful Soup* a vu le jour. Cette bibliothèque facilite le *web scraping* et donne un accès à énormément d’information. Aujourd’hui, il existe une énorme quantité d’outils pour faire du *web scraping*, même les gens qui ne connaissent rien en programmation peuvent en profiter.

## Qu’est-ce que c’est?

Le *web scraping* est un procédé d’extraction de données d’un site web pour par la suite les récupérer pour que l’utilisateur puisse s’en servir plus facilement. L’utilisateur peut donc ensuite les utiliser à des fins d’analyse par exemple, car elles seront dans un format prévu à cet effet. En effet, les formats finaux peuvent être : un fichier *JSON*, un fichier *CSV* ou bien un fichier Excel par exemple. En utilisant le web scraping, on automatise l’extraction de données qui serait normalement effectué par une personne en utilisant le copier et coller.



## Les outils et langages

Il y a différentes façons de faire du *web scraping*. En effet, il y a plusieurs langages de programmation qui permettent d’en faire et ceux-ci ont à leur disposition plusieurs outils qui permettent de faciliter cette extraction de données. Il y a par exemple Python qui est généralement considéré comme le langage le plus conseillé pour le *web scraping* avec *Scrapy* ou *Beautifulsoup* qui sont deux outils liés à ce langage. Par la suite, le *web scraping* peut être fait à l’aide du langage PHP avec les librairies *Guzzle* ou *Goutte*. Ensuite, un autre langage utilisé est le C++ qui utilise *libcurl*, *libtidy* et *libxml* comme outils pour récupérer les données d’un site. Comme dernier exemple, il est possible de faire du web scraping avec le langage Ruby, celui-ci utilise *HTTParty*, *NokoGiri* et *Pry* comme outil.

## Les techniques

Il existe plusieurs techniques de *web scraping*. Cette panoplie de technique apporte plusieurs approches pour collecter des données à partir d’un site web. Voici quelques exemples de technique de *web scraping*.

### HTTP programming

Avec cette technique, il est possible de récupérer des données en envoyant des requêtes HTTP et en utilisant de la programmation par socket sur deux types de sites. En effet, avec ceci, et en envoyant les requêtes appropriées, il sera possible de récupérer les pages web désirées du serveur que le site soit fait de façon statique ou dynamique.

### Text pattern matching

Cette technique permet de rechercher quelque chose de spécifique sur la page web ou l’on tente de récupérer les données. Cette recherche peut être faite de deux différentes façons. En effet, la commande grep en UNIX et les expressions régulières dans les langages de programmation sont utilisées afin de trouver les données à extraire.

### HTML parsing

Cette technique est divisée en 4 étapes et elle permet de récupérer les données en analysant le code HTML. La première consiste à prendre connaissance du code HTML de la page à l’aide d’un inspecteur afin de trouver ce que l’on veut extraire comme données. Par la suite il faut choisir un outil et un langage afin de construire notre web scraper. Après avoir inspecté le code HTML et choisi son langage ainsi que son outil, il faut bien analyser le code de la page pour être en mesure de récupérer les éléments voulus en les isolant du reste du contenu de la page. Finalement, il ne reste qu’à extraire les données voulues avec l’outil.

### Computer vision web-page analysis

Cette technique veut que l’ordinateur regarde une page web comme un être humain le ferait. En effet, grâce à l'apprentissage automatique et la vision par ordinateur, ceci permet de reconnaitre et d’extraire les données en analysant la page web visuellement.

## Crawler VS Scraper

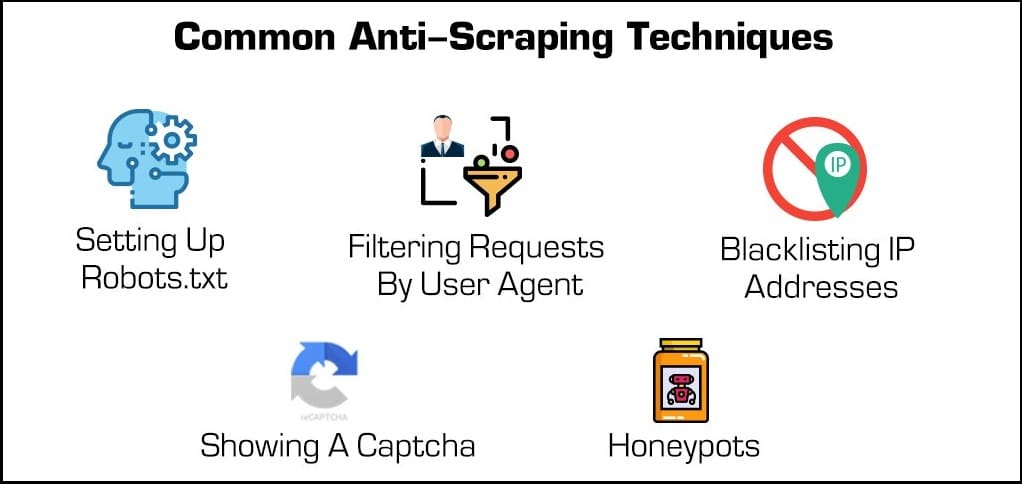
Il faut bien distinguer la différence entre un *web crawler* et un *web scraper*. En effet, un *web crawler* est un robot qui permet de recenser et de lister toutes les URL présentes sur une page web. Tandis qu’un *web scraper* est un programme ou un *script* qui permet d’extraire des données de sites web, et ce, de manière automatique.

## Les avantages

Le *web scraping* comporte plusieurs avantages qui permettent de faciliter l’extraction de données et son analyse par ses utilisateurs. En effet, le *web scraping* extrait les données de façon rapide, automatique, et en grande quantité.

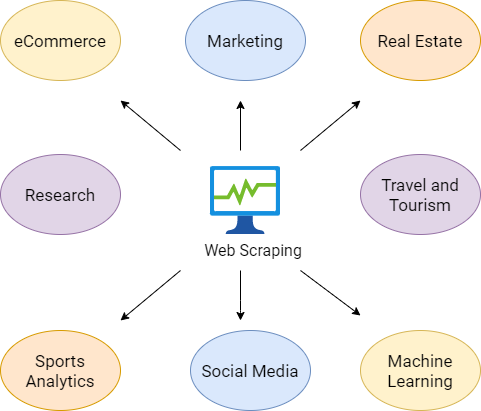
## Les défis

Certes, le *web scraping* apporte son lot davantage, mais il comporte aussi quelques défis lors de son utilisation. En effet, dans un premier lieu, il faut rester à l'affût du moindre petit changement sur la page web ou l’on veut obtenir de l’information, car la page peut changer à n’importe quel moment la façon dont sa structure est conçue. De plus, un autre défi pour le web scraping consiste à éviter les pièges, aussi appeler les *HoneyPots*. Ceux-ci consistent à mettre des liens qui sont dissimulés dans la page. Ensuite, les *web scraper* sont également aux prises avec des technologies *anti-scraping*. Il n’est pas rare que des gros sites web vous bloque, à l’aide de votre adresse IP, à la suite de tentative de *scraping*. Ces sites web sont prêts à tout pour ne pas permettre aux scrapers d’extraire leurs données afin de protéger leur propriété intellectuelle. Ils arrivent à faire ceci, entre autres, avec du code dynamique fait au travers de leur algorithme. Il faut également être certain que les données soient exactes et utiles pour votre usage. L’utilisation de *captchat* est aussi très utilisée pour contrer l’utilisation de *web scraper*.



## Domaine où le web scraping est utilisé

Le *web scraping* n’est pas seulement utilisé pour la collecte de données personnelles, il est utilisé dans une multitude de domaines afin d’aider les utilisateurs dans toutes sortes de tâches. En effet, le *web scraping* peut être utilisé dans les domaines suivants : le *marketing*, l’analyse sportive, le voyage et le tourisme, le *machin learning* ou même le commerce électronique.



## Les cas d’utilisations possibles

Tous les domaines où le *web scraping* est utilisé, l’utilisent de différentes façons, toujours selon leur besoin. En effet, le *web scraping* est généralement utilisé de ces façons : pour surveiller le montant que coute un article, pour étudier le marché visé par une campagne, pour suivre l’actualité ou même pour surveiller les données présentes sur le web des entreprises concurrentes à la sienne.

## Est-ce légal ?

Plusieurs personnes se posent la question à savoir si le *web scraping* est légal, et il est tout à fait pertinent de se la poser. Le point le plus pertinent qui entoure cette question est de savoir si les données que nous voulons récupérer sont destinées au public ou non. Si c’est le cas, il n’y a alors aucun problème à vouloir utiliser un *web scraper* afin de récupérer ces données. En revanche, il faut s’assurer qu’il n’y a aucune donnée personnelle lors de la récupération, car c’est à ce moment que le *web scraping* peut devenir source de problème. On parle de donnée personnelle quand il s’agit par exemple de : nom, numéro de téléphone, courriel, adresse personnelle, information de connexion, adresse IP ou même un numéro de sécurité sociale. De plus, pour que le *web scraping* reste légal, il faut aussi s’assurer de bien suivre les conditions d’utilisation du site web que nous voulons scraper.

## Utiliser le web scraping de façon éthique

Il est bien beau d’utiliser le *web scraping* de façon légale, il faut également l’utiliser de manière éthique. Pour ce faire, il faut par exemple, utiliser l’API du site web s’il en offre un, il ne faut pas scraper le site web trop souvent pour ne pas surcharger le serveur, comprendre ce qu’il est inscrit dans le fichier robot.txt du site web ou être en mesure de séparer les données dont on a réellement besoin et celle que nous n’avons pas besoin.

## Exemples de web scraper

Dans le cadre de ce cours, j’ai fait 2 exemples de *web scraper* en python. Le premier exemple reprend une liste des meilleurs joueurs de la saison en cours et le deuxième exemple reprend le classement des équipes pour l’année voulu. Par la suite, pour ces deux exemples, les données sont récupérées et inscrites dans un fichier Excel afin de faciliter leur analyse.

### Premier exemple

Une image contenant texte, capture d’écran, logiciel

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, motif

Description générée automatiquement

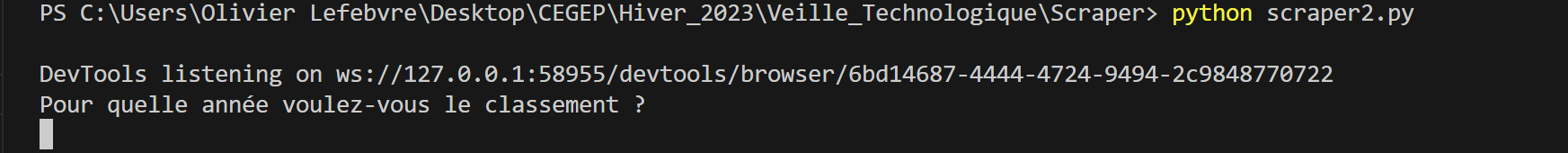
### Deuxième exemple

Une image contenant texte, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement



Une image contenant texte, Police, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran, nombre

Description générée automatiquement

# Conclusion

## Rappel

En somme, *le web scraping* est utilisé pour extraire des données automatiquement sur un site web, de les convertir dans le format voulu et ensuite de les analyser. Il est souvent utilisé en marketing ou dans le commerce électronique par exemple pour faire des suivis de prix ou une étude de marché. Il est également utilisable de plusieurs façons et avec plusieurs outils.

Certes le *web scraper* est rapide et efficace dans sa collecte de données, mais il ne faut pas oublier qu’il n’est pas parfait et qu’il comporte son lot de défis.

Cependant, le *web scraping* doit être utilisé de manière responsable et légale. Certaines personnes peuvent être préoccupées par l'utilisation de leurs données personnelles, ce qui soulève des questions d'éthique et de confidentialité. Il est essentiel de respecter les politiques de confidentialité des sites web ciblés et de s'assurer que le scraping est effectué seulement pour extraire des données publiques.

## Ouverture

En explorant les différentes méthodes d'acquisition de données en ligne, il est important de se demander : existe-t-il d'autres approches qui évitent le recours au web scraping ? Comment pouvons-nous obtenir des données sans violer les politiques de confidentialité et les droits d'auteur des sites web ?

Une possibilité intéressante à considérer est l'utilisation de données *open data*. Les gouvernements, les organisations et les communautés du monde entier ont adopté l'idée de partager des ensembles de données publiques dans un format ouvert et accessible.

En exploitant les *open data*, nous pouvons accéder à des informations précieuses sans avoir recours au web scraping. Ces données sont souvent fournies dans des formats standardisés, facilitant ainsi leur utilisation et leur intégration dans nos analyses.

# Médiagraphie

Martel Arnaud. (2021, 16 septembre). Qu’est-ce que le Web scraping et A quoi sert-il ? *Octoparse*. <https://www.octoparse.fr/blog/quest-ce-que-le-web-scraping>

Mesnildrey, S. (2023, février 14). Qu'est-ce que le Web Scraping et à quoi ça sert ? *sales-hacking*. <https://www.sales-hacking.com/post/web-scraping>

Wikipedia contributors. (2023). Web scraping. Wikipedia. <https://en.wikipedia.org/wiki/Web_scraping>

Karatas, G. (2023). A Comprehensive Guide to Web Scraping Techniques in 2023. *AIMultiple*. <https://research.aimultiple.com/scraping-techniques/>

Comment et pourquoi faire du scraping de données avec Python ? (2023, May 30). *Data Bird*. <https://www.data-bird.co/blog/web-scraping-python#:~:text=Il%20s'agit%20d'une,Rapidement>

What is Web Scraping used for? (2021). *webharvy*. <https://www.webharvy.com/articles/web-scraper-use-cases.html>

Mesnildrey, S. (2022, December 5). Le Web Scraping est-il légal? Les Bests Practices? Éthique? *sales-hacking*. <https://www.sales-hacking.com/post/legal-web-scraping>

Naik, C., & Naik, C. (2023). Web Scraping Done Right: Best Practices to ensure Ethical Data Collection & Web Scraping. *Merit Data Tech*. <https://www.meritdata-tech.com/resources/blog/data/web-scraping-best-practices-ethical-data-collection/#:~:text=Ethics%20of%20Web%20Scraping&text=All%20your%20data%20scraping%20efforts%20must%20be%20ethical.,is%20available%20through%20the%20API>

Talend. (n.d.). Open Data : définition et enjeux professionnels. Talend - a Leader in Data Integration & Data Integrity. <https://www.talend.com/fr/resources/open-data-definition/>