

Contents

JOUR 29M	2
📖 Objectifs	2
Les requêtes HTTP (hyper text transfer protocol)	3
Comprendre le code HTML	4
Automatiser la récupération de données	6
Conclusion	7
Requêtes et parsing – Live-coding	7
Recherche de tag – Live-coding	7

JOUR 29M

📖 Objectifs

A partir de ce module, nous quittons l'analyse de données "conventionnelle", c'est-à-dire l'analyse d'arrays numpy et de dataframes pandas, pour nous diriger vers de **nouveaux types d'analyse**. Cette semaine, nous allons nous concentrer sur l'**extraction de données depuis le web**, ce qu'on appelle le **scraping**.

Après ce module, vous saurez :

- **Requêter** une page web avec Python afin d'obtenir son contenu (sous forme de code **HTML**)
- **Rechercher et extraire** les données voulues dans ce code HTML, grâce au package **BeautifulSoup**
- **Mettre en pratique** ces méthodes de scraping sur des cas concrets

Web scraping

Qu'est ce que c'est ?

- Extraire des informations du web grâce à des requêtes automatisées

Pour quoi faire ?

- Growth hacking | analyse de marché | automatisation

Comment ça marche ?

- Transformer le code des pages web en données structures (images, csv)
- Automatiser l'analyse de ces données



Exemples d'utilisation

- Analyse des commentaires sur TripAdvisor
- Extraction des informations d'un site e-commerce



Les requêtes HTTP (hyper text transfer protocol)

La 1ère étape du scraping est la **requête HTTP**, qui permet d'obtenir le contenu du site web voulu, sous forme de code HTML.

Key takeaways :

- Le **protocole HTTP** est le principal système de transfert d'informations sur internet ; c'est donc ce que nous allons utiliser pour obtenir le contenu d'un site web
- Le package **requests** permet d'effectuer une requête HTTP
- La réponse à notre requête contiendra le contenu demandé (format HTML) et un **code de statut** (par ex: 404 = contenu inexistant / requête incorrecte)

Les étapes du web scraping (1/3)

1. **Requête** : obtenir le code de la page souhaitée (format HTML).

✓ **Package requests**

Permet d'obtenir le code HTML brut sous forme de longue chaîne de caractères.



Codes de réponses HTTP

Lorsqu'on requête le contenu d'une page avec `requests`, on obtient un code de réponse, les plus communs étant :

- **2xx : succès**. Ex: 200 = successful request.
- **3xx : redirection**. Ex: 302 = redirection temporaire.
- **4xx : erreurs client**, i.e. erreurs de votre part. Ex: 404 = page non trouvée.
- **5xx : server errors**, i.e. problème avec l'hébergeur du site. Ex: 503 = serveur indisponible (surchargé) ou en maintenance, 504 = time-out (pas de réponse).



→ https://fr.wikipedia.org/wiki/Liste_des_codes_HTTP

Comprendre le code HTML

La 2ème étape du scraping est le **parsing**, c'est-à-dire le **découpage** du code HTML. Le but est de découper le code reçu afin de pouvoir isoler les parties contenant les données qui nous intéressent.

Key takeaways :

- Le code HTML correspond à la **structure** des sites web, il est donc constitué de **balises** (ou **tags**) qui définissent les différentes zones de la page web
- Le parsing consiste à **découper le code HTML** pour pouvoir isoler les tags / zones de contenu qui nous intéressent
- Le parsing et la recherche de tags sont effectués à l'aide du package **BeautifulSoup**

Les étapes du web scraping (2/3)

2. **Parsing** : séquencer (= morceler) le code web brut pour le rendre manipulable avec Python.

✓ Package Beautiful Soup

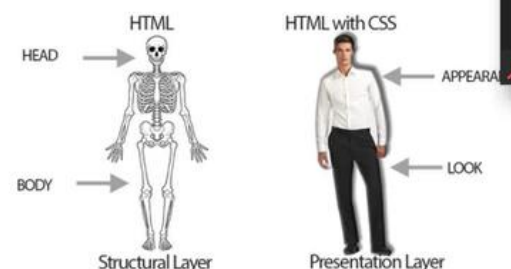
Transforme le code HTML en "arbre",
exploitable par Python pour rechercher
et extraire du contenu



BeautifulSoup

La structure d'une page web

- **HTML** : le squelette de la page, organisé par blocs de contenu, caractérisés par leurs attributs.
- **CSS** : le code contenant l'esthétique de la page (mise en forme, couleurs...).
- **JavaScript** : utilisé pour coder les parties interactives du site et processus scriptés (envoi de formulaires...).



```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8" />
    <title>Titre</title>
  </head>
  <body>
    <div>
      <h1>
        <h2>
          <h3>
            <h4>
              <h5>
                <h6>
              </h6>
            </h5>
          </h4>
        </h3>
      </h2>
    </h1>
  </div>
</body>
</html>
```



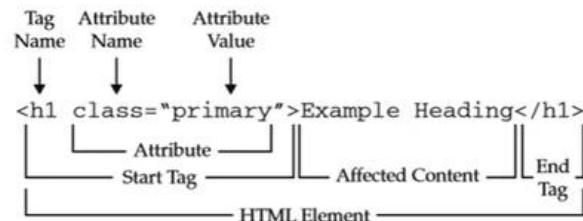
Comment fonctionne le code HTML ?

```
<!DOCTYPE html>
<html>
<body>

<h1>My First Heading</h1>
<p>My first paragraph.</p>

</body>
</html>
```

HTML Tags



```
<a href="https://www.w3schools.com">This is a link</a>
```

Automatiser la récupération de données

La 3ème étape du scraping est optionnelle. Elle consiste à automatiser l'extraction de données sur **plusieurs pages** d'un site web. C'est donc ce qu'on appelle le **crawling**, en anglais "ramper".

Key takeaways :

- Il existe **plusieurs méthodes** pour automatiser l'extraction de données sur plusieurs pages
- Le plus simple est d'utiliser une **boucle** avec Python, pour effectuer plusieurs requêtes HTTP et utiliser BeautifulSoup plusieurs fois de suite
- Une technique plus avancée consiste à utiliser le package **Selenium** pour simuler entièrement un utilisateur (créer un bot), capable de scroller, cliquer, etc.

Conclusion



A retenir

- Le contenu des page web est stocké sous forme de **code HTML**
- On peut **obtenir** et **découper** ce code avec des **packages Python**
- On peut ensuite **parcourir** ce code pour en **extraire** les données requises.

[Requêtes et parsing – Live-coding](#)

[PDF](#)

[Recherche de tag – Live-coding](#)

[PDF](#)

[Exemple concret – Live-coding](#)

[PDF](#)

Quizz récap

QUESTION 1 SUR 6

Quelles sont les utilisations principales du webscraping?

Choisissez TOUTES les réponses applicables.

A

Growth Hacking



B

Analyse de marché



C

Création de site Web

D

Faire de la dataviz



Cette réponse est incorrecte. La réponse correcte est 'A' , & 'B' .

QUESTION 2 SUR 6

Quelle est le package qui permet d'obtenir le code HTML brut sous forme de longue chaîne de caractères.

Choisissez la meilleure réponse.

A

BeautifulSoup

B

bs4

C

Pandas

D



Request



QUESTION 3 SUR 6

Par quel chiffre commence un code de réponse http qui stipule que notre requête a fonctionné?


Choisissez la meilleure réponse.

- A** 0 
- B** 1
- C** 2 
- D** 4

QUESTION 4 SUR 6

A quoi correspond le code Html?

Choisissez la meilleure réponse.

- A** Le squelette de la page, organisé par blocs de contenu, caractérisés par leurs attributs. 
- B** Le code contenant l'esthétique de la page (mise en forme, couleurs...).

QUESTION 5 SUR 6

Qu'est ce que le crawling?

Choisissez la meilleure réponse.

A une technique de nage

B Un style de graphique

C Un langage de programmation

D Le fait de parcourir plusieurs pages web pour récupérer les données brutes



QUESTION 6 SUR 6

soup.find('p')

Choisissez la meilleure réponse.

A Trouve le contenu de la premiere itération d'une balise p



B Trouve le contenu de toutes les itérations des balises p