# Web scraping - MP ETIC BOT

Python and beautiful soup

*Linda BELKESSA*

November 12, 2022

## 1 C'est quoi le Web scraping

Certains sites web offrenet des API (Application Programming Interface); càd des commandes ou librairies qu'on pourrait intégrer dans notre code, pour intégragir avec ce site ou son contenu. **Exemple : l'API Facebook**

Par contre la majorités des sites n'intègre pas cette possibilité, d'où le recours ou web scraping pour récupérer le contenu (fetching data).

**Notion 1 : Web scraping :** le fait d'automatiser des scripts (codes) qui récupérent le contenu (texte, images, URLs, ...) d'un site web.

### 1.1 Les outils pour créer votre propre Scraper

Ceci dépend du langage dans lequel vous voudriez l'implémenter.Par exemple dans le cas de **Python** y a :

- **Beautiful Soup**
- Scrapy
- Selenium
- MechanicalSoup

### 1.2 Comment marche le web scraping ?

- Le code envoie une requete (a request) au site cible avec l'URL spécifique, ce dernier répond par retrouner une page HTML (c'est exactement ce que fait un web browser !)
- La seule différence c'est qu'il interprète pas le code HTML visuellement (HTML Rendering), donc faudrait écrire du code pour filtrer les éléments HTML et extraire ce qu'on cherche.

### 1.3 Les étapes générales d'un scraper

1. **Envoyer requete :** pour récupérer le code HTML de la page cible.

2. **Identifier les éléments HTML** qui nous intéressent après avoir inspecté le contenu HTML

3. **Extraire et reformatter** les éléments filtrés

## 2 Les composants d'une page Web

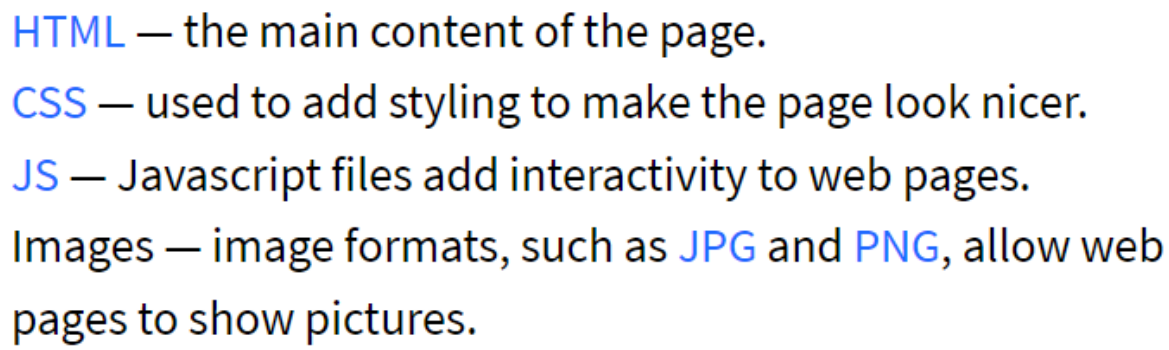


**Figure 1:** Les fichiers composants une page web

**Exemple d'un code HTML :**

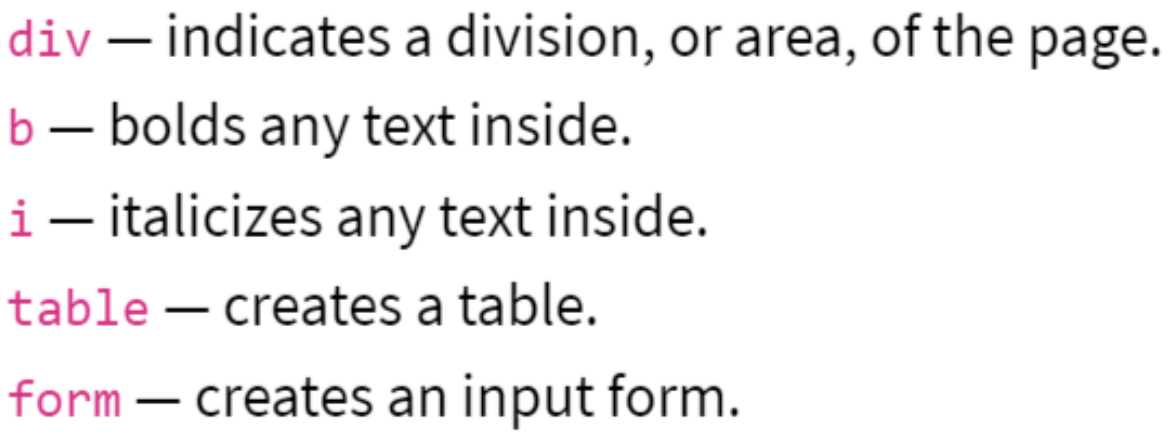


**Figure 2:** Les balises HTML usuelles

```
<html>
<head>
</head>
<body>
<p class="bold-paragraph">
Here's a paragraph of text!
<a href="https://www.dataquest.io" id="learn-link">Learn Data Sci
</p>
<p class="bold-paragraph extra-large">
Here's a second paragraph of text!
<a href="https://www.python.org" class="extra-large">Python</a>
</p>
</body>
</html>
```

# 3 Scrapping the HTML with Python and BeautifulSoup

## 3.1 Outils et installation

- **IDE :** VScode -PyCharm ou autre
- **Interpreteur Python3** tuto ici
- **Installer les librairies :** ***Requests, Html5lib, BeautifulSoup4, Pandas***
  - pip Install requests
  - pip install html5lib
  - pip install bs4
  - pip install pandas

## 3.2 Comment faire son propre web scraper

*A présenter durant la réunion.*

## 3.3 Exercice :

Scrappez un site web de votre choix dont le contenu est cohérent avec les valeurs et objectifs d'ETIC (tech, business, startups, entrepreneurship, student life, ...).