Introduction au Web scraping

Web Scraping Le **web scraping** est une technique d'extraction du contenu d'un site web à l'aide d'un programme (écrit en python, par exemple) dans le but de stocker ou de traiter le contenu.

Par exemple, le moteur de recherche *Google* extrait des données des sites web, comme le titre de la page d'accueil, les articles des blogs, etc, dans le but de référencer les données extraites.

Le programme qui effectue le web scraping fonctionne comme un **web bot** qui navigue les sites web, et les stocke en local, puis traite les données recupérées. Il existe deux manières d'effectuer le web scraping en utilisant le langage du python:

1. à l'aide des modules **Requests** et **BeautifulSoup**.
2. en utilisant le module **Selenium** (avancé).

On va opter pour la première approche. Le module **Requests** permet d'envoyer des requêtes HTTP GET pour récupérer le contenu HTML du site web. Ce contenu sera traité par **BeautifulSoup** pour le parser (analyser la syntaxe du contenu).

Installation des modules L'utilisation des deux modules nécessite leurs installation à l'aide de l'outil de python **pip** avec les commandes suivantes:

```
pip install requests
pip install beautifulsoup4
```

Code source: Bot web-scraper basique Le code source suivant va procéder à la récupération du contenu du site web https://www.example.com, et automatiquement afficher à la console: le titre et le premier paragraphe.

```
# importer les modules nécessaires
from bs4 import BeautifulSoup
import requests

# définir le lien à web scraper
url = 'https://www.example.com/'

# récupérer le contenu du site web
reponse = requests.get(url)
contenu = reponse.text

# parser le contenu HTML à l'aide de BeautifulSoup
soup = BeautifulSoup(contenu, features="html.parser")

# extraire le titre du site web
```

```
title = soup.title.string
print(title)

# extraire le premier paragraphe du contenu HTML
paragraph = soup.findAll('p')[0].text
print(paragraph)
```

Références :

- Web Scraping Basics: https://towardsdatascience.com/web-scraping-basics-82f8b5acd45c
- Documentation de BeautifulSoup4: https://www.crummy.com/software/BeautifulSoup/
- Documentation de Requests: https://docs.python-requests.org/en/latest/

Exercice :

1. Le site web https://books.toscrape.com contient un catalogue de bouquins. Vous allez récupérer la page web situé à l'endroit : `Home > Books > Classics` et vous allez afficher le titre et le prix des bouquins (Classiques) du moins cher au plus cher.
2. Vous allez modifier votre bot de telle manière qu'il cherche un bouquin par titre. Puis, affiche les informations associées avec bouquin, incluant le prix. Par exemple, si l'utilisateur saisit "The Little Prince" , le prix "£45.42" sera affiché.

ALJI Mohamed.