Collège de Bois de Boulogne

420-A53-BB (collecte de données)

Tp1: collecte de données

Les objectifs du TP sont les suivants:

1. Introduction

Le Web Scrape sert pour collecter des données dans la Web.
Après l'apparition de l'Internet et sa évolution, elle s'est devenu une source inépuisable de données. Le Web Scrape est une façon de collecter ces données.
Pour le faire, on utilise le code Python avec quelques bibliothèques :

import requests
Requests est une bibliothèque HTTP Python, publiée sous la licence Apache2. Le but du projet est de rendre les requêtes HTTP plus simples et plus conviviales. Et aussi permet la connexion au website désiré.

from bs4 import BeautifulSoup
Beautiful Soup est la bibliothèque Python permettant d'extraire des données de fichiers HTML et XML. Il travaille avec votre analyseur préféré pour vous fournir des moyens idiomatiques de naviguer, de rechercher et de modifier l'arbre d'analyse. Cela permet généralement aux programmeurs d'économiser des heures ou des jours de travail.

Après la collecte, on stocke de données dans un fichier.

2. Caractéristiques de votre processus

Le Web Scrape, est la récupération Web ou l'extraction des données de sites Web. Un logiciel de récupération Web peut accéder directement au Web via le protocole de transfert hypertexte ou via un navigateur Web en utilisant la fonction request (après l'importation de la bibliothèque requests) . Bien que le scrape Web puisse être effectué manuellement par un utilisateur de logiciel, le terme désigne généralement les processus automatisés mis en œuvre à l'aide d'un robot ou d'un robot d'indexation Web. Il s'agit d'une forme de copie, dans laquelle des données spécifiques sont rassemblées et copiées à partir du Web, généralement dans une base de données locale centralisée ou un tableau, pour une récupération ou une analyse ultérieure.

3. Étapes principales de mise en place du processus

J'ai utilisé comme langue le Python avec l'IDE PyCharm.

On visite une page Web, prend son Uniform Resource Locator (URL) avec la bibliothèque Requests. Et ensuite, on utilise la méthode get() pour accéder cette URL.

Comme la page est habituellement en HTML, il faut qu'on import la bibliothèque Beautiful Soup (d'autres éléments tels que CSS et JavaScript peuvent styler, transformer et ajouter des couches

d'interaction à une page). BeautifulSoup peut nous aider à entrer dans les couches et à extraire le contenu avec find () et find_all().

Maintenant nous sommes capables d'accéder, d'analyser et extrait les données désirées.

On fait notre code avec une boucle **for** et par la suite, on extrait les données et les garde dans un fichier.

Le code est auto-explicatif :

```
import csv
import requests
from bs4 import BeautifulSoup
import datetime
import pandas as pd


url = 'https://www.mitsubishi-motors.ca/en/vehicle/showroom/outlander/2019/'
page = requests.get(url)
print(page.status_code)

soup = BeautifulSoup(page.text, 'html.parser')

# Create a file Price, give headers: Price and Name of the shoe
f = csv.writer(open('car.csv', 'a'))
f.writerow(['Car Model', 'Price', 'Date' ])

# from datetime import datetime
dt_object = datetime.datetime.now()

# all Vehicles
specifications = soup.find('section', class_="specifications l-row background-white")
ul = specifications.find_all('div', class_='vcontainer')
clas = ul

# To get all specification for each Outlander model
name = []
price = []
for i in clas:
    name.append(i.h2.text)
    price.append(i.h3.text)
    f.writerow([i.h2.text, i.h3.text, dt_object])

# Display the DataFrame
```

4. Comment tester l'application

L'application, à mon avis, est testé avec son résultat en si même quand on la visulaize dans un fichier CSV:

Car Model,Price,Date
ES AWC,"Starting at $29,198 MSRPΔ",2019-04-24 13:24:58.010199
ES AWC TOURING,"Starting at $31,698 MSRPΔ",2019-04-24 13:24:58.010199
ES AWC PREMIUM,"Starting at $34,198 MSRPΔ",2019-04-24 13:24:58.010199
SE AWC,"Starting at $32,498 MSRPΔ",2019-04-24 13:24:58.010199
SE AWC BLACK EDITION,"Starting at $35,998 MSRPΔ",2019-04-24 13:24:58.010199
SE AWC TOURING,"Starting at $34,998 MSRPΔ",2019-04-24 13:24:58.010199
GT S-AWC,"Starting at $38,398 MSRPΔ",2019-04-24 13:24:58.010199

5. Conclusion

Le web est vraiment riche de données et avec le Web Scraping, on est capable de trouver pas mal de ces données-là.
On était capable d'accéder la page, extrait le contenu désiré et à la fin garder les données dans un fichier. Particulièrement, je trouvé la bibliothèque Beautiful Soupe très puissante et facile pour travailler avec. Dans ce TP j'ai de difficultés au côté HTML et son arborescence. Peu en peu, j'ai devenu plus à l'aise avec ce Framework et à la fin j'ai réussi et arrivé à mon but.