

Web Scraping : analyse et extraction de données avec Python

Date et durée
Code formation : WSAEP Durée : 3 jours Nombre d'heures : 21 heures
Description
<p>Aujourd'hui dans le monde numérique, les données (Data) constituent une source d'information de très grande valeur. La technique du scraping consiste à extraire ces données depuis des sites web. Plusieurs méthodes existent pour y parvenir et l'utilisation de Python associé à des outils tels que la bibliothèque BeautifulSoup 4 est l'un des meilleurs moyens.</p> <p>Dédiée principalement aux développeurs, data analyst, intégrateurs ou encore aux consultants en business intelligence, cette formation de 3 jours vous apportera des connaissances et des compétences approfondies sur le scraping de données avec python. Vous apprendrez à collecter des données pertinentes et exploitables en appliquant des techniques d'analyse modernes. Grâce aux 8 cours de scraping python proposés, vous découvrirez en détail ces processus et vous vous familiariserez avec les outils et les stratégies de scraping les plus prisés.</p> <p>Avec une approche participative et ludique, cette formation est destinée aux débutants qui souhaitent se former aux outils d'analyse de données de première nécessité dans un laps de temps rapide. Les nombreux travaux pratiques reposent sur des scénarios d'entreprise concrets afin que vous puissiez exercer et appliquer directement vos nouvelles compétences dans un environnement adapté.</p>
Objectifs
<p>Suivre la formation web scraping avec Python, vous permet d'acquérir les compétences suivantes :</p> <ul style="list-style-type: none">• scraper, isoler, modifier et supprimer des données avec une méthode appropriée et rationnelle ;• connaître et appliquer des techniques modernes du langage Python permettant de convertir des données en jeux de données exploitables ;• mettre en œuvre une bonne stratégie de scraping avec différentes sources afin de collecter des données pertinentes ;• coder un script avec une boucle pour scraper efficacement ;• exécuter des opérations de collecte afin de fournir des volumes structurés pour un lac de données (data lake).
Points forts
<p>Un mixte de cours théoriques et de travaux pratiques ; une auto-évaluation de vos connaissances et de vos compétences en programmation python ; des temps d'échanges et un accompagnement pédagogique individualisé.</p>
Modalités d'évaluation
Quiz / QCM Travaux Pratiques

Pré-requis

Suivre la **formation web scraping avec Python**, nécessite les prérequis suivants :

- avoir des connaissances et des compétences de base en matière de programmation et d'algorithmique ;
- avoir une bonne connaissance des fondamentaux de Python.

Public

Cette formation s'adresse aux publics suivants :

- tout professionnels ou toute entreprise issus de tout type de secteurs intéressés par le monde du Big Data et en particulier les Data scientists, les Data Analysts, les Business Analysts, les développeurs web ou encore les marketeurs.

Cette formation s'adresse aux profils suivants

Développeur

Chef de projet / Responsable de projet

Responsable Commercial / Marketing

Chef d'entreprise / Dirigeant

Programme

Module 1 : comprendre les structures de données en Python

- Le concept de récupération de données avec Python.
- Les composants de base d'une structure de données (les listes, les tuples, les séquences, les ensembles et les dictionnaires).

Module 2 : appliquer des techniques avancées pour les structures de données intégrées

- Les spécificités des structures de données intégrées en Python.
- Les manipulations courantes sur les fichiers de données.

Module 3 : utiliser les bibliothèques NumPy et Pandas

- Les étapes de création de tableaux avec NumPy.
- Les étapes de création de DataFrames avec Pandas.
- La visualisation de données et les statistiques avec NumPy et Pandas
- Le calcul de statistiques générales sur des DataFrames avec les modules NumPy et Pandas.

Module 4 : préparer les données avec Python (Data Wrangling)

- Qu'est-ce que le data wrangling ?
- Quels sont les processus du data wrangling ?
- Les sous-ensembles de données, le filtrage et le partitionnement de données.
- La recherche de données aberrantes et la gestion des valeurs de données erronées.
- La concaténation, la jonction et la fusion de données.
- Les techniques de data wrangling avec Pandas.
- L'utilisation avancée des listes et de la commande zip.

- Les techniques de formatage des données.

Module 5 : scraper en mode simple avec Python

- Qu'est-ce que le scraping ?
- Les niveaux de complexité en fonction du support choisi.
- La reconnaissance de données extraites de plusieurs pages web de type textuel et non-textuel.
- Les différents outils disponibles pour le scraping.
- Présentation de la bibliothèque BeautifulSoup.
- L'utilisation de la fonction CSS Select.

Module 6 : scraper en mode expert avec Python

- Les principes fondamentaux du scraping Web.
- L'importance de travailler avec la bibliothèque BeautifulSoup.
- L'utilisation de Python comme outil d'extraction, de transformation et chargement des données (ETL).
- Le scraping avec les fichiers structurés de type .csv, .xml et .json.
- Les processus de lecture et d'écriture de fichiers.
- L'analyse de données de fichiers issus de plusieurs sources.
- Les fonctionnalités d'accès et de traitement des données en blocs linéaires.

Module 7 : implémenter un scraper en Python

- Les techniques de scraping les plus courantes telles que les requêtes GET et les pages séquencées.
- La mise en place d'une stratégie de recherche sur la page pour identifier les données rapidement.
- La création d'un algorithme pour scraper automatiquement.
- L'envoi de données vers une page web.
- Les méthodes utilisées pour récupérer des données spécifiques.
- L'utilisation des requêtes POST et GET.
- Les techniques de navigation pour identifier les données d'un site web.
- La mise en place d'une stratégie de navigation.
- Le développement du scraper.

Module 8 : utiliser le scraping au quotidien

- Les connaissances transversales du data wrangling et du data scraping applicables à la réalité.

*Contenu de formation proposé en partenariat avec [Softeam Institute](#)
Python est une marque déposée de la [Python Software Foundation](#)*