Web Scraping, récolter des données sur le web avec Python

Cours Pratique de 4 jours - 28h
Réf : WPY - Prix 2024 : 2 390€ HT

Vous devez extraire des données du web, les manipuler, les vérifier ou les archiver ? Pour être plus performant, automatisez vos récoltes, élargissez le champ de ces opérations. Optez pour le web scraping avec Python, ses bibliothèques de scraping et sa simplicité permettent rapidement d'industrialiser les processus.

OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Maîtriser les bases du langage Python

Connaître des éléments de programmation avancée en Python

Posséder une vue d'ensemble des principales librairies Python disponibles pour gérer tous types de données de sites

Sélectionner la bonne librairie Python pour votre projet de web scraping et être capable de la mettre en œuvre

Savoir automatiser des récoltes d'envergure (large-scale web scraping) avec des scripts

MÉTHODES PÉDAGOGIQUES

Pédagogie active, des retours d'expérience, des démonstrations sont mises en œuvre par le formateur pour une mise en pratique rapide par les participants.

TRAVAUX PRATIQUES

De nombreux exercices sont réalisés pour illustrer et assimiler les sujets.

PARTICIPANTS

Développeurs, consultants, analystes, chefs de projet et toute personne souhaitant automatiser la récolte de données sur le web.

PRÉREQUIS

Maitriser les bases de l'algorithmique ou savoir programmer. Avoir des connaissances en HTML et CSS est recommandé.

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

LE PROGRAMME

dernière mise à jour : 08/2023

1) Les bases du langage Python

- Les principaux types de variables.
- Effectuer des opérations et travailler sur les chaînes de caractères.
- Les structures de données en Python.
- Comprendre les types mutable et immutable.
- Les structures de contrôle.
- Créer et utiliser des fonctions
- Lire et écrire des fichiers textes ou binaires.

Travaux pratiques : Se familiariser avec le langage, les structures de contrôle et la manipulation de données. Utiliser des fonctions. Créer un petit jeu.

2) Eléments de programmation avancée en Python

- Les fonctions anonymes lambda.
- Comprendre l'utilité des générateurs et savoir en créer.
- Traiter les erreurs avec la gestion des exceptions.
- Créer de nouveaux types de données avec la programmation objet.
- Notions d'héritage en programmation objet.
- Utiliser une librairie.
- Sélectionner et évaluer les librairies développées en open source

Travaux pratiques : Écrire un générateur. Créer et manipuler un objet avec ses attributs et des méthodes.

3) Extraire des données via des API Restful

- Se connecter à une API web.
- Effectuer différents types de requêtes HTTP.
- Le format JSON, analyser les données extraites.
- Rechercher des données spécifiques dans du texte avec des expressions régulières.
- Gérer les erreurs de connexion.

Travaux pratiques : Extraire les bonnes données dans un texte. Récolter et exploiter des informations de géolocalisation via une API RESTful. Évaluer les résultats d'une connexion API réelle.

4) Capter des tableaux de données sur internet et les traiter avec Pandas

- Rappel des bases HTML et CSS.
- Les fondamentaux de Pandas.
- Importer et exporter des données dans différents formats.
- Manipuler des données avec Pandas.
- Scraper des tableaux de données sur le web.

Travaux pratiques : Extraire des données numériques à jour sur Internet. Traiter et archiver les données récoltées.

5) Scraper des sites web avec Beautiful Soup

- Scraping facile : Beautiful Soup.
- Mettre en œuvre le parser.
- Rechercher dans l'arborescence du parser.

Travaux pratiques : Scraper des sites web avec Beautiful Soup. Trouver rapidement les données utiles, les sauvegarder avec les informations correspondantes.

6) Automatiser des récoltes d'envergure avec Scrapy

- Le fonctionnement de base du framework Scrapy.
- Identifier du contenu à scraper.
- Structurer une spider.
- Automatiser une récolte Scrapy et enregistrer les résultats.
- Évaluer la performance d'une campagne.

Travaux pratiques : Crawler des articles web et récolter les données pertinentes avec Scrapy.

LES DATES

CLASSE À DISTANCE	PARIS	LYON
2025 : 18 mars, 03 juin, 09 sept., 30 déc.	2025 : 11 mars, 20 mai, 02 sept., 16 déc.	2025 : 18 mars, 03 juin, 09 sept., 30 déc.