Webscraping avec Python

5 j (35 heures) Ref : IAD009

Public

Développeurs, administrateurs et architectes cherchant à automatiser la collecte d'informations

Pré-requis

Maitrise de Python et programmation orientée objet
Maitrise en data science et en traitement de données
Connaissances de base en HTML et CSS
Notions de protocole HTTP et d'architecture web

Moyens pédagogiques

Formation réalisée en présentiel ou à distance selon la formule retenue
Nombreux exercices pratiques et mises en situation, échanges basés sur la pratique professionnelle des participants et du formateur, formation progressive en mode participatif. Vidéoprojecteur, support de cours fourni à chaque stagiaire

Modalités de suivi et d'évaluation

Feuille de présence émargée par demi-journée par les stagiaires et le formateur
Exercices de mise en pratique ou quiz de connaissances tout au long de la formation permettant de mesurer la progression des stagiaires
Questionnaire d'évaluation de la satisfaction en fin de stage
Auto-évaluation des acquis de la formation par les stagiaires
Attestation de fin de formation

Cette formation de 5 jours vous apprendra à extraire des données de sites web avec Python. Vous découvrirez comment réaliser des requêtes HTTP, analyser des pages HTML et naviguer dans des sites web. Vous apprendrez à utiliser les principales bibliothèques Python de webscraping comme BeautifulSoup, Selenium et Scrapy. De nombreux exercices vous permettront de mettre en pratique ces techniques sur des sites web réels. Vous verrez également comment automatiser la navigation, gérer l'authentification et respecter les bonnes pratiques et l'éthique du scraping.

En fin de formation, vous saurez concevoir et déployer des robots d'extraction pour collecter des données à partir du web de façon efficace et responsable.

Objectifs

Comprendre le fonctionnement du protocole HTTP et du langage HTML

Savoir envoyer des requêtes HTTP GET et POST avec Python

Être capable d'analyser le contenu HTML d'une page web

Extraire des données structurées à partir de pages web avec BeautifulSoup

Savoir parcourir plusieurs pages d'un site en suivant les liens

- Utiliser les sélecteurs CSS et XPath pour cibler finement des éléments
- Automatiser la navigation dans un site avec Selenium
- Savoir gérer les formulaires, les cookies et l'authentification
- Découvrir le framework Scrapy pour développer des robots d'extraction
- Comprendre les enjeux légaux et éthiques du webscraping
- Mettre en place des stratégies pour éviter le blocage des robots
- Savoir déployer un robot d'extraction sur un serveur ou dans le cloud

Programme détaillé

BASES DU SCRAPING AVEC PYTHON ET BEAUTIFULSOUP

- Introduction au webscraping
- Envoi de requêtes HTTP en Python
- Analyse de pages HTML avec BeautifulSoup
- Suivi de liens et pagination
- Exercices pratiques

TECHNIQUES AVANCEES AVEC BEAUTIFULSOUP ET SELENIUM

- Sélecteurs CSS et expressions XPath
- Gestion des formulaires et de l'authentification
- Webscraping dynamique avec Selenium
- Exercices pratiques

SCRAPY ET BONNES PRATIQUES

- Introduction à Scrapy
- Création d'un projet Scrapy
- Optimisation des performances
- Bonnes pratiques et éthique
- Exercice pratique

SCRAPING AVANCE ET ASTUCES

- Gestion du javascript et des sites single-page
- Contournement des protections anti-robots
- Extraction de données à partir d'APIs
- Cas pratiques

DEPLOIEMENT ET PROJET FINAL

- Déploiement d'un robot de scraping
- Sauvegarde et exploitation des données
- Projet fil rouge
- Conclusion et perspectives