

Web scraping avec Python

Code : DEV-PYTHON-310

Durée : 3 jours (21 heures)

Public visé

Développeurs, Data Scientists, Data Analysts, Data Engineers, chefs de projets.

Prérequis

Expérience professionnelle en développement. Connaissances Python souhaitées.

Objectifs pédagogiques

À l'issue de cette formation, vous serez capable de :

- Décrire le fonctionnement des Web scrapers
- Développer un Web scraper avec fonctionnalités avancées
- Activer des architectures multi-thread pour le scraping

Programme

Jour 1

Bases du Web scraping et premiers outils

- Bases du Web scraping (modèle Internet, HTML/CSS/JS, aspects légaux)
- BeautifulSoup (navigation, regex, sélecteurs CSS)
- Scrapy (scrapers, règles, items, pipelines)

Jour 2

Traitement de données et scraping avancé

- Traitement de texte (CSV, PDF, Word, Pandas, NLTK)
- Formulaires et logins (Requests, cookies, authentification)
- Scraping JavaScript (Ajax, Selenium WebDriver)

Jour 3

API, CAPTCHAs et parallélisation

- Scraping d'API (JSON, API non documentées)
- CAPTCHAs (Pillow, Tesseract, NumPy)

- Parallélisation (multi-thread, multi-process)