FORMATION - FORMATION - DATA SCRAPING : COLLECTER DES DONNÉES À GRANDE ÉCHELLE SUR LE WEB

Public et pré-requis
Tout public : professionnel·le de l'information en rédaction, ONG, organisme public ou privé
Aucun, mais si vous avez déjà un projet en tête, c'est mieux !

Durée et modalités pédagogiques
Formation Présentiel de 14h

Moyens pédagogiques
Le support numérique de formation est transmis aux apprenant·e·s pendant ou après la session.

Accessibilité
Nous mettons tout en œuvre pour vous accueillir toutes et tous dans les meilleures conditions. Notre charte handicap détaille nos engagements pour l'accessibilité de nos formations. Contactez notre référent handicap si vous avez des besoins spécifiques d'accompagnement avant, pendant et après la formation: handicap@samsa.fr

Evaluation
1- Positionnement
Un questionnaire en ligne est adressé aux participant·e·s avec la convocation de manière à évaluer leur niveau initial et leurs attentes.
2-Evaluation des acquis
Au cours de la formation, les exercices réalisés permettent au formateur ou à la formatrice d'évaluer l'acquisition des compétences
3- Attestation
Une attestation des acquis de fin de formation est adressée par mail aux apprenants suite à la session

Dernière modification contenu formation - 25/03/2026

Objectifs

- Comprendre le fonctionnement et les limites du scraping
- Connaître les bases de programmation en Python
- Programmer un robot simple en Python
- Collecter automatiquement des données en ligne

Programme

Jour 1

- Présentation du formateur et tour de table
- Définition du scraping et de ses différents niveaux de difficulté sur plusieurs supports (depuis le web, depuis du papier, depuis des PDF)
- Passage en revue d'utilisation du scraping dans des projets pour bien comprendre l'intérêt d'une telle pratique ainsi que ses limites
- L'environnement légal : découverte de ce qu'il est autorisé de scrapper et de ce qui ne l'est pas
- L'architecture d'internet : pour scraper des sites, il faut avant tout savoir comment Internet fonctionne
- Qu'est ce qu'un « client » ? Qu'est qu'un « serveur » ? Pourquoi est-ce important ?
- Utiliser des données déjà structurées au travers d'une API
- Sur Internet, les échanges utilisent HTTP et HTML : comment impactent-t-ils nos scrapers ?
- Comprendre comment fonctionne le HTML pour mieux extraire des informations. Qu'est qu'une balise HTML ? Un attribut ? Comme identifier certains éléments avec une CLASS ou un ID ?
- Initiation au Python, le langage le plus répandu pour coder un scraper : les fichiers, les variables, l'affichage d'information, les conditions et les boucles
- Utiliser des fonctions et manipuler des CSVs pour lire et enregistrer des données
- Prise en main d'outils spécifiquement dédiés au scraping : Beautiful Soup – pour automatiser les opérations les plus courantes – et CSS Select – pour mieux cibler les éléments à extraire

Jour 2

- Mise en pratique sur un scraper simple (requêtes GET, pages séquencées)
- Identifier la stratégie à adopter pour naviguer sur le site
- Travail individuel pour coder du scraper
- Analyse des résultats obtenus et des difficultés rencontrées
- Mise en pratique sur un scraper complexe : envoyer des données à un site Internet pour obtenir des résultats plus complexes
- Identifier la stratégie à adopter en fonction du site

- Travail individuel pour coder le scraper
- Analyse des résultats obtenus et des difficultés rencontrées
- Débrief de la session et évaluation à chaud de la formation par les participant·es