[sfeir] Institute

SFEIR Institute | PYDATA

Python pour la science des données

Apprenez les connaissances fondamentales pour manipuler, explorer et analyser vos données avec Python.

3jours / 21h

Objectifs pédagogiques

- d'utiliser Python en standalone et à travers JupyterLab.
- d'acquérir des données à partir des formats de fichiers standards (ex: CSV) ou par web scraping.
- de mettre en oeuvre des analyses de données exploratoires quantitatives avec Numpy et Pandas.
- de mettre en oeuvre des analyses de données exploratoires visuelles avec Matplotlib, Pandas, Seaborn, Bokeh,...
- de résoudre des problèmes réels d'analyse de données à grande échelle.

Public cible

- Consultants, développeurs, chefs de projet, data scientists, data engineers

Prérequis

- Connaissances de base en programmation (logique, structures de données)
- Connaissances de base en mathématiques (fonctions, vecteurs, matrices)

A noter: des bases en Python seront un plus mais ne sont pas requises, les concepts fondamentaux du langage étant présentés la première journée

Programme

Jour 1

Python au sein d'un environnement data science type Anaconda.

Module 1: Introduction à l'analyse de données avec Python

- Pourquoi Python pour l'analyse de données ?
- Introduction au langage Python: bref historique, versions, outils.
- Les librairies Python essentielles pour l'analyse de données.
- Notebooks interactifs collaboratifs: JupyterLab, Kaggle Kernels, Google Colab.

Ateliers:

- Installer un environnement Python complet pour l'analyse des données, type Anaconda.
- Prendre en main de l'environnement de développement JupyterLab.

Module 2: Les bases de Python pour la manipulation de données

- Premier programme Python.
- Organiser et écrire son code, PEPs.
- Expressions, variables et types.
- Chaînes de caractère et texte.
- Interagir avec les utilisateurs: inputs et affichage formaté.
- Structures de données: listes, ensembles, dictionnaires, tuples.
- Logique et boucles.
- Fonctions.
- Travailler avec des fichiers.
- Itérateurs et Générateurs.

Atelier:

- Explorer des jeux de données dans des fichiers tenant ou non en mémoire.

Module 3: Au delà du prototypage dans les notebooks

- Bases de POO en Python (classes, objets).
- Modulariser son code.

- IDE et débuggers.

Ateliers:

- Refactoring du code d'exploration de données avec une approche modulaire et POO.
- Développer une application web basique avec Flask permettant de partager les résultats de l'exploration des données.

Jour 2

Les librairies Python pour la data science.

Module 4: Les bases de Numpy: tableaux et calculs vectoriels

- Représentation des tableaux avec Numpy.
- Fonction universelles et accès rapide aux éléments.
- Programmation orientée tableaux.
- Tableaux et fichiers.
- Algèbre linéaire.
- Génération de nombres pseudo aléatoire.
- SciPy.

Ateliers:

- Implémenter des fonctions avec Numpy.
- Traiter des données avec Numpy et SciPy.

Module 5: Les bases de Pandas

- Les structures de données Pandas.
- Fonctionnalités essentielles.
- Statistiques descriptives.

Atelier:

- Mettre en oeuvre des analyses de données exploratoires quantitatives.

Module 6: Gestion des fichiers de données sous Pandas

- Les différents formats de données.
- Bonnes pratiques en terme de manipulation de gros fichiers.
- Stratégies pour analyser les données à l'échelle.

Atelier:

- Explorer des jeux de données tenant ou non en mémoire.

Module 7: Visualisation des données en Python

- Visualisation avec MatplotLib.
- Visualisation avec Pandas et Seaborn.
- Autres outils de visualisation: bokeh, plotly,...

Atelier:

- Mettre en oeuvre des analyses de données exploratoires visuelles.

Jour 3

Utilisation avancée des librairies Python pour la data science.

Module 8: Pandas avancé

- Nettoyer et préparer les données.
- Joindre, combiner et reshaper des données.
- Agréger et grouper les données.
- Séries de temps.

Atelier:

- Manipulation avancée des fichiers .CSV avec Pandas.

Module 9: Web scraping

- Rappels sur le fonctionnement du web et des sites web.
- Qu'est-ce que le web crawling et le web scraping ? Est-ce légal, faut-il se conformer ? Quels sont les outils communément utilisés ?
- Télécharger des pages web et effectuer des requêtes HTTP.
- Web scraping avec BeautifulSoup.

Atelier:

- Scraping web avec Request et BeautifulSoup.

Module 10: Introduction à Scikit-learn

- Pré-traitement des données: chargement et transformations.
- Analyse et prédiction sur les données avec les algorithmes de machine learning: régression, classification, clustering, réduction de dimensionnalité.

Atelier:

- Prise en main de Scikit-learn: pré-processing et premier modèle ML.