Python pour la data science

Durée

4 jours.

Participants

Ingénieur, développeur, chercheur, data scientist, data-analyst et toute personne désireuse de se former à l'univers scientifique de Python.

Prérequis

Bonne connaissance de base en programmation

Objectifs pédagogiques

- Posséder une vue d'ensemble de l'écosystème scientifique de Python
- Connaître les librairies scientifiques incontournables pour la science des données
- Être capable de manipuler des données volumineuses avec Python
- Comprendre l'intérêt de la datavisualisation
- Savoir visualiser des données avec Python

Méthodes pédagogiques

Ce séminaire se base sur des présentations, des échanges et des études de cas. Des outils comme Lasagne ou Keras seront présentés.

Description

La science des données est un domaine multidisciplinaire en constante expansion. Elle s'appuie sur des méthodes scientifiques, des algorithmes et des processus que Python a su maîtriser grâce à un écosystème particulièrement riche. Il est devenu aujourd'hui le langage de référence pour l'analyse de données, quels qu'en soient les formats. Notre formation vous permet la prise en main des outils, bibliothèques et modules Python pour obtenir de rapides compétences en data science avec ce langage.

Programme

Introduction à Python

- Panorama de l'univers Python : Domaine d'utilisation, PEP8, librairies, gestion des environnements virtuels, Jupyter Notebooks
- Presentation des concepts fondamentaux de python : variables, structures de contrôle et de boucle, fonctions…
- Programmation objet en Python : Attributs et méthodes de classe, manipulation, héritage et encapsulation

- Librairies standard de python et fonction communes : os, sys, pathlib, re, compréhension de liste

Travaux pratiques Mise en pratique sur les différentes fonctionnalités

La SciPy Stack

- Le socle de librairies scientifiques incontournables sur lequel sont basées toutes les autres : la SciPy Stack.
- Numpy : calcul numérique et algèbre linéaire (les vecteurs, matrices, images).
- SciPy, basée sur Numpy pour : les statistiques, les analyses fonctionnelles, géo spatiales, le traitement du signal…
- Pandas : l'analyse de données tabulaires (csv, excel…), statistiques, pivots, filtres, recherche…
- Matplotlib : la librairie de visualisation de données incontournable.
- PySpark : L'interface Python pour Apache Spark

Travaux pratiques Analyses et manipulations de fichiers en utilisant les outils de la scipy stack

Les librairies de visualisation

- Panorama des librairies de visualisation de Python : 2D/3D, desktop/web, statistiques, cartographie, big data…
- Les librairies orientées desktop : Matplotlib, Seaborn.
- Les librairies orientées web : Bokeh, altair, Plotly…
- Les librairies pour la 3D : Plotly, pythreejs, ipyvolume…
- Les librairies cartographiques : Cartopy, folium, ipyleaflet, Bokeh, cesiumpy…

Travaux pratiques Réalisation de multiples exercices avec différentes librairies. Comprendre dans quelle situation utiliser une librairie plutôt qu'une autre. Visualisation de données cartographiques (DVF).

La datavisualisation

- L'intérêt de la datavisualisation
- Utiliser les écosystèmes PyViz et HoloViz.
- Utiliser les outils SuperSet, Mayavi, Paraview et VisIt.

Travaux pratiques Poursuivre l'utilisation des librairies de visualisation et manipulations des outils.

Les formats de fichiers scientifiques et la manipulation de données volumineuses

- Panorama des principaux formats de fichiers scientifiques : NetCDF, HDF5, GRIB, JSON, PARQUET, MATLAB, CGNS…

- Manipuler des données volumineuses avec Pysparks.

Travaux pratiques Manipulation de données dépassant les Go, lecture et écriture de fichiers NetCDF/HDF5.