

Python avancé pour l'analyse de données scientifiques

Les 23, 24 et 25 novembre 2020

| | |
|-------------------------------|---|
| Objectifs de formation | Approfondir les connaissances de python Utiliser les principales bibliothèques de calcul numérique dont Numpy, SciPy, Pandas et Matplotlib Créer des interfaces homme machine avec le module tkinter Utiliser les modules pour la parallélisation des applications |
| Public Pré-requis | Biologistes (au sens large), chercheurs, ingénieurs ou techniciens. Avoir fait la formation : initiation au langage python pour les biologistes ou avoir une connaissance des bases de python Remplir consciencieusement le questionnaire lié à la demande d'inscription. |

Partie I : Présentation générale

Vue d'ensemble des contributions scientifiques qui se sont développées autour du langage Python.

Les distributions Python pour les DataScientists : (WinPython, , Enthought Canopy, Anaconda, Pycharm)

Partie II : Python avancé : modules scientifiques

Dans ce chapitre nous étudions les modules scientifiques de Python. Nous découvrons les bibliothèques Numpy, Pandas et Matplotlib, sur lesquelles s'appuient pratiquement toutes les autres bibliothèques.

La Scipy Stack

Rapide présentation de l'écosystème Scipy : Scipy Stack, Numpy, scikit-learn, scikit-image...

Numpy

Numpy est une bibliothèque Python de calcul numérique (algèbre linéaire) permettant de manipuler des vecteurs et matrices à plusieurs dimensions. Elle offre des types de données avancés pouvant être gérés très simplement. Implémentée en langage C elle mixe le meilleur des 2 mondes: la souplesse du langage Python avec les performances du C.

Les tableaux Numpy et leurs particularités

Création de tableaux - Sélection de données – Calculs - Filtres

Pandas

Pandas est une bibliothèque permettant d'analyser avec efficacité des structures de données complexes. Pandas permet de manipuler des matrices dont chaque colonne peut être d'un type différent et offrant de nombreuses et puissantes fonctionnalités de manipulation, réduction, recherche et statistiques.

Les Séries

Les DataFrames

Création de Série et de DataFrame

Sélection des données

Dépouillement de fichiers .xlsx, .csv

Analyses statistiques

Matplotlib

Matplotlib est une bibliothèque de tracé et de visualisation de données. Elle permet de dessiner tout type de représentations, 2D comme 3D, lesquelles peuvent être interactives avec l'utilisateur.

Affichage de courbes 2D - Affichage de nuages de points - Affichage d'histogramme

Figure avec plusieurs graphiques

Annotations des courbes

Partie III : Parallélisme avec Numba et Dask

Numba et Dask

Just in Time Compilation

Partie IV : Machine learning avec Scikit-learn

Ce chapitre vous initiera aux bases du Machine Learning à travers plusieurs exercices concrets. Nous utiliserons les implémentations mises à disposition par la bibliothèque Scikit-Learn

L'apprentissage automatique

Principe

Enjeux

Applications

Apprentissage supervisé - Apprentissage non-supervisé - Apprentissage par renforcement

Classification

Régression

Traitement des données

Quelques algorithmes par apprentissage automatique

Régression linéaire - Régression logistique

Algorithme de Ridge - Algorithme de Lasso

SVM

Validations croisées

Random Forest

Partie VI : Création d'IHM avec tkinter

Introduction

Concepts de base

Les widgets disponibles dans Tkinter

Le Label

Les Boutons (Button)

Les champs de saisie (Entry)

Les widgets checkbutton - Les widgets Radiobutton

La Spinbox

Le widget Listbox - Le widget Scale

Les conteneurs

Les fenêtres (Toplevel)

le widget Frame

LabelFrame

le widget PanedWindow

Dates : 23, 24 et 25 novembre 2020

Date limite d'inscription : Le 6 octobre 2020 sur <https://www.sirene.inserm.fr/>

Lieu : Paris 13^{ième} (48 /50 rue Albert)

Nathalie SUZANNE
Chargée de développement RH

formation.paris11@inserm.fr

Tél : 01 85 55 38 18 / 44

INSERM DR Paris 11
Formation Continue
48-50 rue Albert
75013 PARIS