

Introduction à la programmation Python

Master 2 Ingénierie économique et analyse de données, CY Cergy Paris Université

2020-2021

20h CM/TP

Enseignant : William Kengne (william.kengne@cyu.fr)

Nous vivons aujourd'hui dans un monde inondé des données. On estime aujourd'hui à 2.5 trillions (i.e. 2.5×10^{18}) d'octets la quantités de données générées chaque jour dans le monde et 90 % de ces données auraient été créées au cours des deux dernières années. Malheureusement, beaucoup de ces informations (données) ne peuvent être utilisées par les humains ; en général parce qu'elles sont trop volumineuses pour être traitées par les méthodes d'analyse standard et les outils et technologies disponibles. La quête permanente des solutions pour l'exploitation de ces données volumineuses (données massives ou big data) a permis le développement de nouvelles technologies/outils, prenant en compte les problématiques big data.

Il s'agit dans ce cours d'étudier le langage de programmation Python ainsi que les outils qu'il offre pour la gestion, le traitement, le trafic (data munging/wrangling) des données volumineuses et le big data. Des cas pratiques sont étudiés sur des données réelles.

Plan du cours

Chapitre 1 : Introduction générale : Présentation de Python, Python pour la data science, Installation et environnement de travail, Premier pas avec Python.

Chapitre 2 : Principaux types de données Python : Types numériques et booléens, Chaînes de caractères, Listes, Tuples, Dictionnaires.

Chapitre 3 : Syntaxe du langage Python : Structures de contrôle, Bloc et séquence d'instructions, Structure conditionnelle, Structures répétitives, Fonctions, Fonction *input()*, Définir une fonction.

Chapitre 4 : Les fichiers sous Python : Ecriture dans un fichier texte, Lecture d'un fichier texte.

Chapitre 5 : Tableaux et matrices avec NumPy : Tableaux et matrices, Manipulation des matrices, Eléments de calcul matriciel.

Chapitre 6 : Trafic de données avec pandas : Structures de données Pandas, Lecture et écriture des tables de données, Gérer des tables de données, Statistiques descriptives élémentaires, Statistiques descriptives univariée et bivariée, Traitement de données manquantes.

Chapitre 7 : Analyse exploratoire et Machine learning avec Python : Analyse en composantes principales, reconnaissance de caractères, données "German credit data", régression logistique, arbre de décision, random forests, réseaux de neurones : reconnaissance de caractères.

Chapitre 8 : Web scraping avec Python : Expressions régulières, BeautifulSoup, utilisation de l'API de Twitter, exemple d'analyse de sentiments.

Chapitre 9 : Introduction au Big data : Spark, PySpark, MapReduce, MLlib : Configuration de Spark, Paradigme MapReduce, utilisation de MLlib avec PySpark, exemple de reconnaissance de caractères.

Références

- [1] BIERNAT, E. , LUTZ M. Data Science : fondamentaux et études de cas. *Eyrolles*, (2015).
- [2] LAYTON, R. Learning Data Mining with Python. *Packt Publishing*, (2015).
- [3] NELLI, F. Python Data Analytics. *Apress*, (2015).
- [4] MÜLLER, A. C. AND GUIDO, S. Introduction to Machine Learning with Python, A Guide for Data Scientists. *O'Reilly*, (2016).
- [5] RASCHKA, S. Python machine learning. *Packt Publishing*, (2015).
- [6] RICHERT, W. Building machine learning systems with python. *Packt Publishing*, (2013).
- [7] SWINNEN, G. Apprendre à programmer avec Python. *Editions Eyrolles* , (2009).
- [8] WES, M. Python for data analysis. *O'Reilly*, (2012).