

Installation de l’environnement de travail

Dans ce cours nous allons utiliser le langage Python pour programmer les algorithmes vus en cours, ainsi que pour développer le projet final. Pour ceux qui ne connaissent pas du tout le langage, il est fortement conseillé d’installer Python sur sa propre machine et de suivre le tutoriel suivant :

<https://docs.python.org/3/tutorial/>

Tous les autres peuvent aller directement à la section suivante.

1 NumPy

NumPy est une bibliothèque qui introduit les `ndarray` qui va être une structure de données fondamentale dans nos projets. Elle permet aussi de manipulations numériques sur ces objets avec une efficacité remarquable. Pour installer cette bibliothèque consultez la page :

<https://numpy.org/install/>

Il est aussi conseillé de jeter un coup d’œil à ce tutoriel :

<https://numpy.org/devdocs/user/quickstart.html>

2 Pandas

Cette bibliothèque introduit deux autres structures de données fondamentales pour nous : les `Series` et les `DataFrame`. Les `Series` peuvent être vues comme une sorte de liste uni-dimensionnelle apte à contenir n’importe quelle type de données. Un `DataFrame` peut être vu comme une version 2D d’une `Series` mais il est en réalité bien plus riche car il permet, par exemple, d’avoir des types différents sur des colonnes différentes et aussi de faire des opérations arithmétiques sur les lignes ou sur les colonnes. Vous pouvez trouver un tutoriel assez complet [ici](#).

3 Scikit-learn

Cette bibliothèque va être notre outil de travail pendant la première partie du cours. Il s’agit d’une bibliothèque spécialement dédiée à l’apprentissage automatique qui interagi de manière assez naturelle avec les deux bibliothèques vue précédemment. Vous pouvez trouver la page d’installation à cet adresse :

<https://scikit-learn.org/stable/install.html>

Un tutoriel (trop) rapide se trouve [ici](#), mais avant d'y aller commençons par faire quelques exercices de préparation.

4 Charger une BD

Nous allons utiliser des données qui se trouvent sur un site web spécialisé dans l'hébergement gratuit de BD qui servent (ou ont servi) comme *benchmark* ou point de départ pour le développement d'algorithmes d'apprentissage ou plus simplement comme exemples pédagogiques.

Exercice 1

1. Allez sur le site web [Kaggle.com](https://www.kaggle.com/zynicide/wine-reviews) et téléchargez la base 'Wine reviews' (<https://www.kaggle.com/zynicide/wine-reviews>);
2. Utilisez la méthode `read_csv` de `Pandas` pour charger les données dans un dataframe de nom `mydata`.
3. Vérifiez avec la commande `head()` le contenu de votre base.
4. Utilisez l'attribut `columns` d'un dataframe pour visualiser les noms des colonnes de votre BD. Cherchez sur la page de téléchargement de la base plus de détails sur la signification des données contenus dans ces colonnes.