CS-E4002
Course materials
Your points
Code Vault

This course has already ended.

« Data and Input Reading Course materials Importing Package Functions »

CS-E4002 / Using Data in Python / Libraries and Packages

Libraries and Packages

Once the type of the file is known, we need to be able to analyse the dataset and work on it. One way to do this is to manually create an instance of an object (Python variables) from the file. To do this, we would need to choose the right dataset type to use and worry about the RAM available, and we would need to implement functions manually to work on the dataset... Doable, but there is a simpler/faster way to do this: we can use Python types and functions that have already been developed in order to analyse datasets. This implies using Python packages, libraries and modules.

1. Definitions:

- Module: A file containing Python definitions and statements.
- Package: A collection of modules following a precise naming convention (for example, calling any function can look like module.submodule1.function).

```
scipy.
numpy.random.randint
pandas.dataframe
```

- Library: A set of functions and types (it's the generic term). Packages and Modules are libraries. In Python, we often talk about the Python Standard Library, which contains data types and functions that make up the core of the Python language.

2. Installing packages in Python:

Packages need to be installed on the computer, and can be found on the Python Package Index. All of the detailed information on how to install a package is available here; make sure to pick the method that actually suits your Python installation.

3. Most-used Python libraries for data science:

- SciPy: Scipy is a huge library that includes, in particular, Numpy and Pandas. It also includes Matplotlib, which we will use later on for plotting results of analysis on the dataset. Its name comes from being a scientific computation library.
- Numpy: numerical Python package focusing on the use of n-dimensional arrays. It is the most useful Python tool for working with vectors and matrices.
- Pandas (panel data): its goal is to facilitate data analysis with Python. It provides high-level types and functions to work on data.

When using these libraries, even the best programmers always refer to the documentation. It is an essential tool when using Python packages. In the documentations you can find everything you could need: how to install the library, which functions and types does it provide, the parameters and outputs of different functions, examples of uses of all the functions provided, etc.

4. Using modules during this course:

IMPORTANT: The automatic Grader on A+ supports only a handful of external libraries and Python Standard Library modules. For the A+ coding exercises of this course, you only need to use the external libraries Numpy, Pandas, and Scikit-Learn (which will be discussed later in the course) and the Python Standard Library modules datetime, random, math, and time. Other modules (including sys, os, and enum) are forbidden by the Grader, as is the exit() function.

« Data and Input Reading Course materials Importing Package Functions »