Practical No. 01

Aim: To setup Python Environment for Data Science and study how python is useful for data science.

S/W Required: Linux OS/Win OS

Theory:

Python is a popular high-level programming language used mainly for data science, automation, web development, and Artificial Intelligence. It is a general-purpose programming language supporting functional programming, object-oriented programming, and procedural programming. Over the years, Python is known to be the best programming language for data science, and it is commonly used by big tech companies for data science tasks.

Python has many uses in Data Science. It comes with tons of free libraries that directly can be used for Data Science and fields associated with it. NumPy, Pandas, Matplotlib, Seaborn, Pytorch, Keras etc. are some of the most popular libraries in Python for Data Science.

Numerical Python – NumPy:

NumPy is one of the most commonly used data science libraries. It allows you to work with numeric and scientific tasks in Python. Data is represented using arrays or what you may refer to as lists, which can be in any dimension: 1-dimensional (1D) array, 2-dimensional (2D) array, 3-dimensional (3D) array, and so on.

Pandas:

Pandas is also a popular data science library used in data preparation, data processing, data visualization. With Pandas, you can import data in different formats such as CSV (comma-separated values) or TSV (Tab-separated values). Pandas works like Matplotlib because it allows you to make different types of plots. Another cool feature Pandas offers is that it allows you to read SQL queries. So, if you have connected to your database, and you want to write and run SQL queries in Python, Pandas is a great choice.

Matplotlib and Seaborn:

Matplotlib is another awesome library Python offers. It has been developed on top of MatLab - a programming language used mainly for scientific and visualization purposes. Matplotlib allows you to plot different kinds of graphs with just a few lines of code.

You can plot graphs to visualize any data, helping you to gain insights from your data, or giving you a better representation of the data. Other libraries like Pandas, Seaborn, and OpenCV also use Matplotlib for plotting sophisticated graphs.

Seaborn (not Seaborne) is just like Matplotlib, just that you have more options - to give different parts of your graphs different colors, or hues. You can plot nice graphs and customize the look to make the data representation better.

Python Offers Many Data Science Tools:

Though Python is simple because of its syntax; there are tools that have been specifically designed with data science in mind. Jupyter notebook is the first tool, it is a development environment built by Anaconda, to write Python code for data science tasks. You can write and instantly run codes in cells, group them, or even include documentation, as provided by its markdown capability.

A popular alternative is Google Colaboratory, also known as Google Colab. They are similar and used for the same purpose but Google Colab has more advantages because of its cloud support. You have access to more space, not having to worry about your computer storage getting full. You can also share your notebooks, log in on any device and access it, or even save your notebook to GitHub.

Jupyter Notebook/Lab:

Jupyter Notebooks are a tool for easily integrating text, code, and code output into a single document. This not only makes them incredibly useful for instructional materials (this entire site is actually built with Jupyter Notebooks), but it also makes them useful as a method of sharing analyses. Using Jupyter Notebooks, you can not only share the conclusions of your analysis with colleagues, but also the code that generated those analyses, making it easy for others to see how you reached your conclusions, and, crucially, play with that code to see what happens if the analysis is changed slightly. Overall it’s an amazing free and open source tool that is available for Data Science.

Installation of Jupyter:

The easiest way to install jupyter is by using Anaconda, it is a tool that provides all the things needed for Data Science in one place.

In our lab environment, another way to setup jupyter is by using python package manager “pip”. Any linux system comes with Python pre-installed on them. To install “jupyter” on Debian based linux execute the following series commands in the terminal.

```
sudo apt update && sudo apt upgrade
sudo apt install python3-pip
pip install jupyter
```

After installation use the following command to start the jupyter notebook server. Make sure to always keep the server running. The following command opens default web browser with the fairly simple jupyter notebook interface.

```
jupyter notebook
```

Conclusion:

Thus, we have studied how python is used for data science with wide range of libraries and learned how to setup a python environment for data science.