IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

Python Libraries and Tools for Data Science: A Review

Siddhesh Chandrakant More
Student, Department of MCA
Lt. Bhausaheb Hiray S.S. Trust's Institute of Computer Application, Mumbai, Maharashtra, India

Abstract: *This document talks about features, characteristics and reasons behind python programming language and its modules to become popular among data science and its application developers compared to other programming languages like R. And also, their role and importance in implementing essential techniques to improve applications for solving real-world problems.*

Keywords: Python, Data Analytics, Data science, Artificial Intelligence, Machine learning, Numpy, Matplotlib, Pandas, Skit-learn

I. INTRODUCTION

In modern civilization of groups of people, technology has emerged and evolved as a robust tool to resolve modern-day problems and challenges. From the invention of computers, which are initially used as computing devices for mathematics, have extended their compatibility with other machines and improved their capability to supply big selection of operations from a distinct and diverse kinds of applications. This computing revolution forced every industry for an exponential growth by better performance and quick improvements by overcoming the challenges.

Computer science sub fields like data science which uses statistics, probability and their related methods to analyze and understand the insights of information, Machine learning for exploratory data analysis and building model by training data, AI which is employed to form intelligent systems, Deep learning which uses different layers during a network to predict etc. These technologies have evolved as an important need within the technology industry to seek out solutions for ever challenging problems. Last decade witnessed a considerable and extraordinary amount of stored data. Growth of knowledge in every industry including healthcare, automotive, manufacturing, finance, food processing etc., then came a desire to utilize this information for building and inventing best new products and to renovate the present ones, and also to enhance customer experience in their respective fields. To handle such amounts of information, there's a necessity for mathematical tools like statistics, the calculus, infinitesimal calculus, probability etc., they play a prominent role in understanding, interpreting and converting information to information.

Now comes a desire for an honest programming language which is powerful and versatile to implement the methods required to develop data science applications, which is simple to use and popular among the developers. Python could be a high-level general purpose programming language which had built-in data types like lists, arrays etc. python ASCII text file is compiled to be byte code without a necessity for separate compilation. In recent years, python with the assistance of mathematical libraries like Numpy, Pandas, Scipy and Scikit-learn made python really for machine learning and deep learning.

II. WHY PYTHON?

Scientists and developers use complied languages like lisp, C++, C for data analytics and for developing other scientific applications. A good but clean basic syntax, flexible but robust integration programming language is required due to the large number of integrated platforms and environments. Python satisfies all these qualities and it is also easy to learn. Let's discuss some important characteristics of python.

- **Integrity**: Python is a programming language that is well-known for its ability to integrate with other languages. It can be used with a variety of other programming languages, including C, C++, Java, CORBA, and TensorFlow, as well as a wide range of Computer Science and Machine Learning tools, including Google Cloud ML Engine, Amazon Machine Learning, and others. Python not only interacts with platforms and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

programming language interfaces, but it also has a library stack that demonstrates the strength of its integration capabilities.

- **Ease of Use:** Python is simple to use because it bases its operations on normal language rather than on complicated syntax rules. Python programming is as easy to learn as entering an English sentence into your computer. Installing and downloading Python is also simple.
- **OOPS:** In Python, object-oriented Programming (OOPs) is a programming paradigm that uses objects and classes in programming. It aims to implement real-world entities like inheritance, polymorphisms, encapsulation, etc. in the programming. The main concept of OOPs is to bind the data and the functions that work on that together as a single unit so that no other part of the code can access this data.
- **Python's Built-in Data Structures:** Python has a variety of mutable and immutable data structures, including arrays, Strings, and tuples for mutable data and list, set, and dictionary for immutable data. We can simply organize and perform operations on data using these data structures.
- **Compilation**: Python is generally called an interpreted language however; it combines compiling and interpreting. When we execute a source code. Python first compiles it into a bytecode. The bytecode is a low-level platform-independent representation of your source code, even so, it isn't the binary machine code and cannot be run by the target machine directly. Actually, the Python Virtual Machine is a set of instructions for a virtual machine(PVM). Byte code is a lower level, platform independent, effective, and intermediate.

III. OPERATIONS IN DATA SCIENCE

- **Data Extraction**: Data extraction is that the method of obtaining data from an information base or SaaS platform so as that it's replicated to a destination — sort of a data warehouse, designed to support online analytical processing (OLAP). Data Science operations starts with extracting information from planet, this data is in any format, shape, size. Python provides many libraries for extracting data from web and universal machines like requests, beautiful soup, scrapy, pypdf. you will be ready to extract data from SQL files and databases using the Pandas library. this will be by opening a database, or by running an SQL query. Two python libraries are going to be accustomed make the connection hoping on the type of database.
- **Data Processing**: This operation entails steps to transform raw data into usable information. Missing values, corrupted values, time zone differences, and date range issues are all crucial checks to make during this procedure. Numpy and Pandas libraries are provided by Python for data processing, which is also known as data cleaning. The conversion of information into something that a computer can understand, such as 0's and 1's, is known as raw data.
- **Data Visualization & Analytics**: Once the data has been cleaned and prepared for use, it is critical to understand the data's insights. Graphs are the best way to learn about data since they provide the data an overall meaning. The Python modules pandas and matplotlib are excellent graph visualization tools. Any firm or corporation relies heavily on data. To uncover information helpful for corporate decision-making, it is necessary to gather, handle, and evaluate data flow in a fast and accurate manner. Data analysis is a process for gathering, transforming, and organizing data in order to generate future predictions and data-driven decisions. It also aids in the discovery of potential solutions to a business problem.
- **Data Modelling**: After data analysis there are many machine learning algorithms to create a model based on the data. The design of models heavily relies on statistics and probability. Python provides a Skit-learn library which had inbuilt methods for machine learning models such as linear regression, logistic regression etc. for supervised, unsupervised and reinforcement learning
- **Scientific Computations:** For scientific computations for researchers, students and scientist's python provides a library called sci-py which have all the methods that are used for many mathematical and scientific operations.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

IV. PYTHON LIBRARIES FOR DATA SCIENCE

4.1 Pandas

Pandas is a fast, important, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

Pandas give fast and effective DataFrame object for data manipulation with integrated indexing.Pandas is used as a tool for reading and writing data between in-memory data structures and different formats CSV and text files, Microsoft Excel, SQL databases, and the fast HDF5format.Intelligent data alignment and integrated care of missing data gain automatic label-based alignment in performing calculations and easily transform disordered data into a structured format. pivoting and flexible reconfiguration of data collections.

example: import numpy as np

```
import pandas as pd
s = pd.Series([1, 3, 5, np.nan, 6, 8])
s = pd.Series([1, 3, 5, np.nan, 6, 8])
s
0  1.0
1  3.0
2  5.0
3  NaN
4  6.0
5  8.0
dtype: float64
```

4.2 Matplotlib

A tool for visualising data, Matplotlib is a low-level graph charting framework written in Python. John D. Hunter is the author of Matplotlib. We are free to utilize Matplotlib because it is open source. For platform portability, it is primarily written in Python, with a few pieces also written in C, Objective-C, and Javascript.

Example:

```
import matplotlib.pyplot as plt
import numpy as np
x = np.linspace(0, 2 * np.pi, 200)
y = np.sin(x)
fig, ax = plt.subplots()
ax.plot(x, y)
plt.show()
```

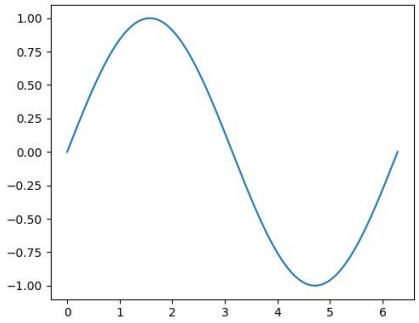

Fig. 2 Matplotlib Plot Result

Python's Matplotlib toolkit provides a complete tool for building static, animated, and interactive visualizations. Easy effects are made feasible by Matplotlib, as are challenging effects

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

4.3 Numpy

Python isn't developed to perform numerical operations. But the raising interests for python from all the engineering, scientific and exploration communities forced the inventors of python to create a package with high position array perpetration. Jim Fulton and Jim Hugunin created Numeric, a matrix perpetration for numerical operations, with Guido van Rosuum. They also changed the name toNumpy.Python has a list data structure but that is n't enough for numerical transformations and computations. So Numpy is principally developed on a core data structure calledndarray.ndarray is a type of matrix array, which has rudiments of same type. Numpy array by dereliction has a fixed size and shape m * n with equal m rows and n columns when a new element is added to the matrix exceeding its size, also it clones all the rudiments from also existing array and creates a new array with equal size also deletes the originalarray.Numpy can be used in any python integrated development terrain. For illustration Initialization

import numpy as np

This command significances numpy library with a object np.

This object can be used whenever we've to use the numpy functionalities, stylesetc.

To create an array

a = np.array((1, 2, 3))

This law creates a ndarray object and inserts 1, 2, 3 figures to variable a.

It gives complete control on vectorized array computations and operations perfecting the results and performance of the CPU. Scipy and Pandas libraries are developed on the ndarray structure bed of the numpy.

4.4 Skit-Learn

This library is developed to function with scipy and Numpy libraries to design and implement models. It has a range of algorithms from the combination of statistics and probability. By using this library, the original data set can be divided into desired ratio of train and test datasets. Developers train the model from the training dataset which is previously divided from the original dataset and then evaluate it from the test dataset to measure the accuracy, quality of the model.Skit-learn library also provides a variety of methods for supervised, unsupervised and reinforcement learning. Classification and clustering of data items can be done using this library.

4.5 Nltk Library

Natural language processing tool kit is the acronym for Nltklibrary. As the name suggests this library is used to develop NLP modules. It is the designed for improving the English language in machine learning models.Tokenization, stemming and lemmatization are the key operations which are to be performed on the dataset before training an NLP model, to perform these operations Nltk library provides a wide range of inbuilt methods and functions that can be directly used by just using the values in the dataset.

4.6 Scipy

Scientific python called as Scipy is used for N-dimensional array manipulation. This library runs on the core of Numpy. This library provides a numerous method for scientific computations such as optimization, linear programming, calculating distances etc

Table 1: Information about Python libraries from GitHub

Library	Stars	Forked	Contributors
NumPy	20.8K	7K	1330
SciPy	9.8K	4.3K	1158
Cython	7.1	1.3	405
Pandas	34.3K	14.6K	2612
PyTables	1.1K	230	72
h5py	1.7k	463	175
Matplotlib	15.7K	6.4K	1172
seaborn	5722	905	87
Plotly	11.7k	2.2K	181

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

Bokeh	16.4K	4K	525
ggplot	5.5K	1.9K	262
scikit-learn	50.5K	23.2K	2399
mlpy	5	2	1
Shogun	2.9K	1.1K	175
mlxtend	4K	771	81
TensorFlow	166K	86.9K	3129
Keras	55.5K	19.1K	1029
PyTorch	56.9K	15.8K	2326
Caffe	32.7K	19K	269
Caffe2	8.4K	2K	200
Dumbo	1.1K	153	6
Spark (PySpark)	33.2K	25.8	1808

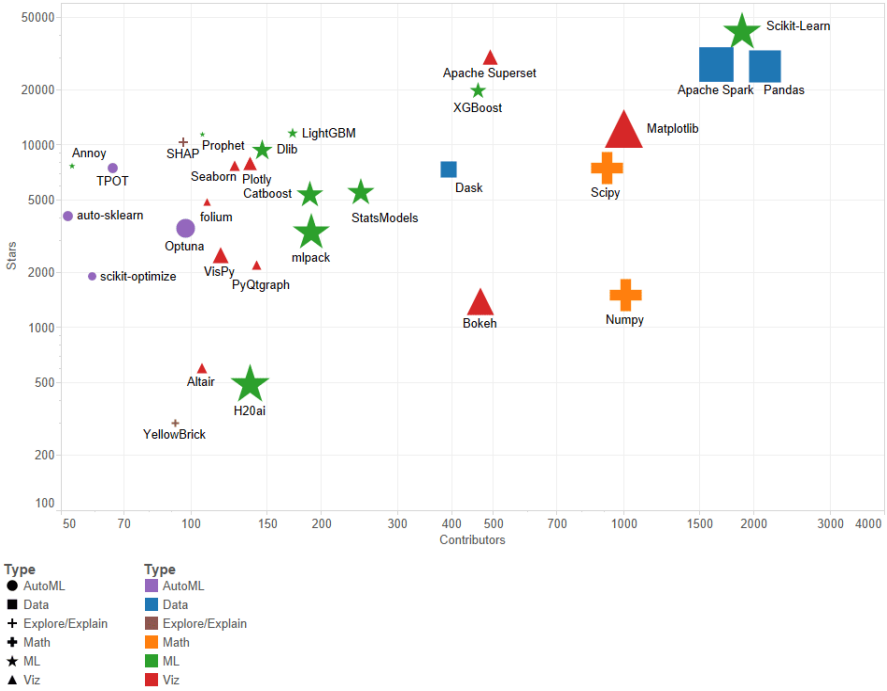

Figure 1: Top Python Libraries Data Science Data Visualization Machine Learning

V. DEEP LEARNING

Deep learning uses a variety of characteristics and representations and is a type of unsupervised learning. The Keras framework is one of the most important extensions that Python provides for deep learning. Numerous modules, including initializers, regularizes, restrictions, activations, losses, metrics, and optimizers, are supported by Keras.

We can create a wide variety of cutting-edge applications with Keras, including robotics, picture recognition, audio/video recognition, and more.

VI. ARTIFICIAL NEURAL NETWORKS

Many Python packages, modules, and libraries are available for artificial intelligence. One such library with a potent neural network is neurolab. Single layer neural networks and multi-layer neural networks are among its primary functionalities. Numpy, Scipy, and Matplotlib libraries are extensions.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

VII. TOOLS FOR PYTHON DATA SCIENCE

7.1 Jupyter Notebook

An interactive web-based tool called Jupyter Notebook is utilised in data science initiatives. Jupyter notebooks include more useful capabilities in addition to offering kernels for programming languages like Python, Scala, and R.

It blends materials written in natural language with code. The second justification is that Jupyter Notebooks are interactive. It is perfect for data scientists and researchers since it allows them to experiment with data and see how the code responds to each command they input.

7.2 Spyder

Spyder is a free and open source scientific environment created by and for scientists, engineers, and data analysts in Python, for Python. The functionality of a comprehensive development tool's powerful editing, analysis, debugging, and profiling skills are combined with a scientific package's data exploration, interactive execution, deep inspection, and beautiful visualization features to create a singular product.

VIII. LITERATURE REVIEW

Most DM and ML operations in Python depend on vectorized and rapid numerical computing with the NumPy and SciPy modules. Many of these libraries' features are essentially wrappers for Netlib's secure, reliable scientific algorithm implementations. The primary benefit of SciPy and NumPy is their capacity for broadcasting over n-dimensional arrays and conducting effective vectorized computation. Another benefit of adopting Python in this area is how simple it is to integrate outside code into the Python interpreter. Cython is likely the library that is used the most frequently in DM for that reason. A language based on Python called Cython additionally allows for the invocation of C functions and the use of C-type variables and classes. Some people may object to the use of Cython. parts of code several times faster.

Another benefit of adopting Python in this area is how simple it is to integrate outside code into the Python interpreter. Cython is likely the library that is used the most frequently in DM for that reason. A language based on Python called Cython additionally allows for the invocation of C functions and the use of C-type variables and classes. Some people may object to the use of Cython.

IX. CONCLUSION

In this paper we have discussed about characteristics of python programming language and the reasons behind python to become the most popular language. We also discussed about various python libraries and there functionalities on developing data science applications and analysis. We discussed about the disadvantages of using python in data science projects and improvements required to meet future needs of the industry. we also discussed about deep learning and artificial neural networks and python libraries which support their functionality.

Machine learning is rapidly growing area and its sub branches such as deep learning and neural networks are headed towards new innovations and advancements. There is a need for every technology to evolve to meet machine learning needs in the future, this evolution process can be either by advancing the existing systems or by knowing its limitations and improving them. There are many other technologies which are in their respective developing stages are getting ready for more powerful computational speed, flexibility and being robust systems. But today python libraries are more popular in the data science industry for their dynamic usage and functionalities.

ACKNOWLEDGMENT

I would like to acknowledge the University of Mumbai, India to give me the opportunity to do the research work under the title “Python Libraries and Tools for Data Science: A review”. I would like toacknowledge Prof. Sadhana Pandeyfor providing guidance and the college L.B.H.S.S Trust’s Institute of Computer Application, Mumbai India to support during the research process.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

REFERENCES

[1]. R. Tohid, Bibek Wagle, Shahrzad Shirzad, Patrick Diehl, Adrian Serio, Alireza Kheirkhahan, Parsa Amini, Katy Williamst , Kate Isaacst , Kevin Huck, Steven Brandt and Hartmut Kaiser Asynchronous Execution of Python Code on Task-Based Runtime Systems. Louisiana State University, University of Arizona, University of Oregon E-mail: {mraste2,bwagle3,sshirzl,patrickdiehl,akheirl}@lsu.edu, {hkaiser,aserio,sbrandt,parsa} @cct.lsu.edu,

[2]. Abhinav Nagpal, Goldie Gabrani, Python for Data Analytics, Scientific and Technical Applications.

[3]. Sanzu: A Data Science Benchmark R. Nicole, Alex Watson, Deepigha Shree Vittal Babu, and Suprio Ray

[4]. Stančin and A.Jović A nover view and comparison of free Python libraries for data mining and big data analysis

[5]. Matthew Mayo, KDnuggets on November 2, 2020 in Automated Machine Learning, AutoML, Data Exploration, Data Processing, Data Science, Data Visualization, Explainability, Machine Learning, Pythonhttps://www.researchgate.net/publication/347444225_Python_And_Its_libraries_in_Data_Science_and _Related_fields