International Journal of Modern Electronics and Communication Engineering (IJMECE) ISSN: 2321-2152
Volume No.-7, Issue No.-2, March, 2019

Multidimensional View of Python Packages

G. Kavitha1st
Assistant Professor in Information Technology
Dr. N.G.P. Arts and Science College
Coimbatore, India.

D. Sowmyadevi 2nd
Assistant Professor in Computer Science
Sri Ramakrishna College of Arts and Science for Women
Coimbatore, India

Abstract: Python is a powerful dynamic programming language which is used in wide variety of application development. It has built-in and third party packages to speed up the process of programs implementation. Python implementation is under open source licenses that make it freely usable and distributable. That is why Python is used to create in thousands of real-world business applications like, Product developments, websites creations and web app development, creating models in Machine learning and useful for data analyses. This paper is mainly focused on, why to learn python language and also discussed about Python's basics and its features. There are plenty of packages available to deal with data analysis process. Here, represented a few important packages, which are used to create and analyze the data for scientific computing.

Keywords: Python's packages, Big Data, Data Analysis, Machine Learning, MySQL.

I. INTRODUCTION

Python is based on C, it is a General-purpose high-level programming language which is deep, huge and intuitive. It is an object-oriented programming language, which means it integrates data and code into objects which increase the performance of the system activities like Java, C++ and Scala. It acts as a tool to deploy and implement machine learning model at a large-scale by performing data wrangling, engineering, feature selection and web scrapping. It has cutting-edge API for machine learning or Artificial Intelligence. It is also called as English statement language. It doesn't require a programming paradigm to do the coding. It is mostly used in Data Science field by the engineers and mathematicians. At the same time the doing coding in python are easy and more robust than R.

II. WHY PYTHON?

The python language is used by many developers since it is easier to learn, implementation time is very less by using Predefined packages .The important one is, it is an Up growing language in this era. The other main reasons are follows,

1. Rich Data Structures

1. Python has rich built-in Data Structures like lists, tuples, sets, dictionaries, strings, thread safe queues, and many other types to hold the data [3].
2. Lists hold arbitrary data objects and it can be sliced, indexed, joined, split, and used as stacks.
3. Heaps are available as operations on top of lists.
4. Sets hold unordered and unique items.
5. Dictionary in Python is an unordered collection of data values, used to store data values like a map, it holds key and value pair which is separated by a colon (:).Keys of a Dictionary must be unique and of immutable data type. It is created by the built-in function dict().
6. NumPy is an n-dimensional array structure that supports optimized and flexible broadcasting and matrix operations.

2. Readability and Indentation

Python's syntax is very simple. In fact, tuning pseudo code into correct Python code is a matter of correct indentation.

The most distinctive features of Python is its use of indentation to mark blocks of code. To indicate a block of code in Python, need to do indent each line of the block by the same level. A missing or extra space in a Python block could cause an error or unexpected behavior. The IDLE is designed to automatically handle indentation. If indentation is missed this "Indentation Error: expected an indented block" error will be thrown by the compiler, The below is the simple code correct indentation in python.

```
fav_Language = input("What's your favorite Language ? ")
if (fav_Language == 'Python'):
print("Your favourite Language is", fav_Language)
else:
print("Not a python Language")
```

3. Cross-Platform Python Framework

Python scripts can be used on different operating systems such as Windows, Linux, UNIX, Amigo, Mac OS, etc[2]. The migration is easy from one platform to another, and run it without any changes. The development and portability rate in Python are very high, which allows the same application to operate across platforms. Python consists of rich libraries and many other packages to tackle migration. The cross-platform Python framework works for Android, Windows 7, Linux, and Mac.

4. Great Community

Python's development is conducted largely through the Python Enhancement Proposal (PEP) process. The PEP process is the primary mechanism for proposing major new features, for collecting feedback on an issue, and for documenting the design decisions that have gone into Python. Outstanding PEPs are commented and reviewed by the Python Community [3]. The Stack Overflow is a programming Q&A site which is useful for beginners [1].

5. Python is beating with R

Till 2015-2016, R has been more popular language. But, in the last 2 – 3 years, Python gained tremendous popularity and Rich libraries for numerical and scientific analysis. KDnuggets also did survey to figure out the top platforms among data scientists and

analytics professionals [6]. Have a look at the results below in infographics format Fig.1 [6].

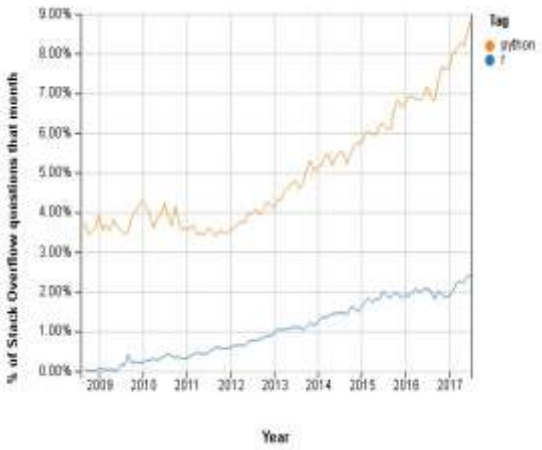

Fig.1: Popularity of Python and R

III. POPULAR PACKAGES IN PYTHON

1. Big Data

Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. The term big data has become one of the hottest technology buzzwords in the past two years. By utilizing the list of available packages to handle these kinds of Big data for doing data science and analytics job in short time.

2. Data Analysis

Data analysis is the act of getting useful insights or decisions out of data. It involves asking questions about what happened, what is happening, and what will happen (Descriptive, Perspective and Predictive Analytics) to get optimized solution for a particular business problem.

3. Packages

A package is collection of related modules in python, where as module is a collection of related classes and functions. These modules are used for mathematical calculations, string manipulations, web programming, data science and many more. One can either import or develop a package own to perform the desired task. The Popular Libraries are numpy, SciPy, Matplotlib and so on. Overall, it is a powerful environment for scientific computing. Some of the popular python's packages are discussed below.

4. Popular Python's Package:

NumPy

Numpy is the core library for scientific computing in Python. The main object of the N-dimensional array type is called ndarrays which describes the collection of items of the same data type [4]. Items in the collection can be accessed using a zero-based index, its dimensions are called axes and mutable. It is used to perform different operations on ndarrays by using its built-in functions. There are several ways to index an array in numpy are, slicing, integer array indexing, and Boolean array indexing. Intrinsic array will be created by built-in functions and some of the functions are ones(), zeros(), arrange() and so on. It has some functions for Linear Algebra, Fourier Transforms and Random Number Generation.

SciPy

The basic data structure used by SciPy is a multidimensional array provided by the NumPy module. SciPy is a collection of mathematical algorithms and convenient functions due to this many scientists are switched from ruby to python. It is organized into subpackages covering different scientific computing domains and these sub-packages need to be imported separately [4]. Basically, it is mainly used in scientific computing and to do this so many constants are required.

Pandas [Data Structures and Analysis]

The panda is a high-performance open source library for data analysis. It is a Python package providing fast, flexible, and expressive data structures designed to make working with structured and time series data. It built upon the NumPy libraries. It is a perfect tool for data wrangling which means that the process of cleaning data and unifying messy and complex data sets for easy access and analysis. It designed for quick and easy data manipulation, aggregation, and visualization. There are three main data structures are present ,The first one is "Series Data" Structure which is one-dimensional, second is "Data Frame" which is two-dimensional and third one is "Panel" which is three-dimensional and all these are size-mutable array. It facilitates loading/importing data from various resources such as CSV and DB. According to python context Pandas is nothing but streamlining the data and then performing analysis on it. The pandas are a core library of the Python toolkit for data analysis. The pandas data structures are much easier to use and more user-friendly than Numpy's ndarrays, since it has row indexes and column indexes in Data Frame and Panel data structure .It has many features and capabilities for data analysis than many other popular languages such as Java, C, C++, and Ruby.

Scikit-learn

Scikit-learn is a simple and efficient tools for data mining and data analysis. It is the most powerful library for machine learning which is creating models from data. It has range of supervised and unsupervised machine learning algorithms. This library contains a lot of efficient tools for machine learning as well as statistical modeling including classification, Regression, Clustering, Model Selection, Preprocessing and Dimensionality Reduction[11]. It was developed in Python itself with the help of NumPy, SciPy and Cython [4]. This package is mainly for analyzing the data and not be used for reading, manipulating and summarizing the data. There are better libraries like NumPy, Pandas to perform reading, manipulating and summarizing the data.

Another terminology is used in data mining is that, Web scraping which is used for extracting data from websites. It is a form of copying data into a central local database or spreadsheet, which can be used for analysis. For example Scrapy is a Python framework for doing web scraping in large scale [2]. It gives all the tools to efficiently extract data from websites, process those data, and store them in preferred structure and format for analyses.

Matplotlib.pyplot

The most popular graphing and data visualization module is Matplotlib. It is a plotting library used for 2D graphics. It can be used in python scripts, shell, web application servers and other graphical user interface toolkits along with NumPy packages. The matplotlib.pyplot is a collection of command style functions that make matplotlib work like MATLAB [7] language. It can also be used with graphics toolkits like PyQt and wxPython [4]. There are various plots which can be created using python Matplotlib. It includes Line Plot, Bar graph, Histogram, Scatter Plot, Area Plot, Stem plots, Pie plot and multiple subplots in one figure(Many plot types can be combined in one figure to create powerful and flexible representations of data).

The pcolormesh () function is useful for colored representation of a two-dimensional array. The mplot3d toolkit has support to plot simple 3d graphs which includs surface and wireframe. The polar() function generates polar plots. The legend() function automatically generates legends of a plot. It has support for visualizing information with a wide array of colors and colormaps. The mathtext module provides TeX style mathematical expressions. So that creating labels, grids, legends, and many other formatting entities with Matplotlib is easy and here everything is customizable.

There are many file formats available to save the plots in Matplotlib and the supported file formats are below

eps: Encapsulated Postscript	pdf: Portable Document Format, ps: Postscript	pgf: PGF code for LaTeX
raw: Raw RGBA bitmap	rgba: Raw RGBA bitmap	svg: Scalable Vector Graphics
jpg: Joint Photographic Experts Group	jpeg: Joint Photographic Experts Group	tif: Tagged Image File Format
png:Portable Network Graphics	svgz: Scalable Vector Graphics	Tiff: Tagged Image File Format

Seaborn

Seaborn is a library for creating statistical graphs in Python to explore and understand data. It is built on matplotlib and tightly integrated with the pandas data structures. In the OOPS paradigm, Seaborn extends the matplotlib library. It is mainly focused on the visualization of statistical models, such as thermal maps. the load () function in the seaborn package IS to load the data for Data exploring. The Pair plots graph functions are useful for exploring correlations between multidimensional data and sns.kdeplot() function is useful for estimating kernel density . Other plotting functions in Seaborn package are Factor plot, violin plot, PairGrid, distplot and joint plot. There is a principle related to seaborn and matplotlib, "If matplotlib "tries to make easy things easy and hard things possible", seaborn tries to make a well-defined set of hard things easy too." [5]. Finally, the main idea of Seaborn is that, it provides high-level commands to create a variety of plot types useful for statistical data exploration, and even some statistical model fitting.

Bokeh

Bokeh is an interactive visualization library that uses modern browsers to display data on a Web page via zoom, resize, reset, and wheel zoom. It works like "D3.js" which is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. It helps to easily create interactive plots, dashboards and data applications. Bokeh has interfaces in Python, Scala, Julia, and R. It provides output in different formats such as html, notebook and server. It can be integrated with flask and djangoweb applications. The Plots created by matplotlib, seaborn and ggplot can also be converted to Bokeh [8].

Pydot

Pydot is a free, open source package and has an interface to connect Graphviz Tool (Graph Visualization Software which uses DOT languages to describe the graphs) written in Python[12]. With its help, it's easy to create structure of graphs, which are very often requited when building neural networks and decision trees-based algorithms.

Here there is a small discussion about DOT languages which is a graph description language to draw directed and undirected graphs, various attributes can be applied to graphs, nodes and edges in DOT files. These attributes can control aspects such as color, shape, and line styles. For nodes and edges, one or more attribute-value pairs are placed in square brackets. It can be used to generate graphs in different formats such as .ps and .pdf. Thus, the usage of Pydot library is for generating complex oriented and non-oriented graphs.

MySQL.Connector

MySQL Python connector package enables Python programs to access MySQL relational data base systems. Python connector runs on any platform where Python is installed. It is designed specifically to connect MySQL. To use this package in python, first import the MySQL.connector. Invoke connect () function by passing parameters like host, database, user and password to connect with a database. Once the connection is established from python to MySQL database, it returns a MySQL Connection object. This object can execute DDL and DML queries. Even stored procedures can be created by using callproc() method and returns a result set, which can be capture stored results() method.

Natural Language Toolkit (NLTK)

Natural Language Processing (NLP) is a category of AI that helps computers to understand, interpret and manipulate human language such as English, Spanish, Hindi and so on and it requires Python 2.7, 3.4, 3.5, 3.6, or 3.7 version. The five essential components of Natural

Language processing are 1) Morphological and Lexical Analysis 2)Syntactic Analysis 3) Semantic Analysis 4) Discourse Integration 5) Pragmatic Analysis. Also Three types of the Natural process writing system they are Logographic, Syllabic and Alphabetic. Machine learning and Statistical inference are two methods to implement NLP. The manipulations of NLP are Tokenization, Stemming, Lemmatization, Punctuation, Character count and word count. Tokenization is the process of replacing

Library	Type	Commits	Contributors	Releases	Watch	Star	Fork	Commits / Contributors	Commits / Releases	Star / Contributors
NumPy	Data wrangling	15980	522	125	280	4286	2012	31	128	8
SciPy	Data wrangling	17213	489	91	244	3643	1775	35	189	6
Pandas	Data wrangling	15089	762	78	826	9394	3709	20	193	12
Matplotlib	Visualization	21754	588	60	413	5190	2817	37	363	9
Seaborn	Visualization	1699	71	11	176	3878	568	24	154	55
Bokeh	Visualization	15724	223	49	322	5720	1401	71	393	26
Plotly	Visualization	2486	33	7	149	2044	512	75	355	62
SciKit-Learn	Machine learning	21793	842	88	1680	18245	9897	26	272	22
Keras	Machine learning	3519	428	28	1025	15043	5227	8	126	35
TensorFlow	Machine learning	16785	756	29	5002	55486	29433	21	579	76
Theano	Machine learning	25870	300	23	520	6171	2116	86	1125	21
Scrapy	Data scraping	6325	243	78	1427	20124	5353	26	81	83
NLTK	NLP	12449	196	28	376	4649	1358	64	622	24
Gensim	NLP	2878	179	43	300	4182	1595	16	67	23
Statsmodels	Statistics	9980	119	19	194	2219	977	75	472	17

sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security. Stemming and Lemmatization are Text Normalization (Word Normalization) techniques in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing.

There are various Natural Language Processing packages are present apart from this , they are SpaCy, Stanford CoreNLP Python, Text Blob, Gensim, Pattern and Polyglot, The important applications of NLP are Information retrieval and Web Search, Grammar Correction, Question Answering, Text Summarization, Machine Translation and Sentiment analysis.

StatsModels

Statsmodels is a complement to SciPy package for statistical computations including descriptive statistics , estimation and inference for statistical models. It is built on top of NumPy, SciPy, and matplotlib. The StatsModels are used in econometrics, generalized-linear-models, time series-analysis, and regression-models. There is a small difference between Scikit-learn and StatsModels [9].Scikit-learn are machine-learning and data-science. StatsModels is for complex statistics. And at the same time Scikit-learn has a variety of tools which helps to pick the correct models and variables. Whereas StatsModels doesn't have this variety of options, it offers statistics and econometric tools that are validated againt other statistics software like Stata and R. Both Scikit-learn and Statsmodels useful to data scientists to run models and get results fastly. But good engineering skills and a solid background in the fundamentals of statistics are required. TheFig.2 depicts about the development activity between Scikit-learn and StatsModels by Github and Scikit-learn development began in 2007 and was first released in 2010 [10]. The current version, 0.19, came out in July 2017. StatsModels started in 2009, with the latest version, 0.8.0, released in February 2017.

Fig.2 and Fig.3 is the detailed status of GitHub activities for a few libraries in python [13].

Fig.2: Development Activities Development Activity of Scikit-learn and statsmodels.

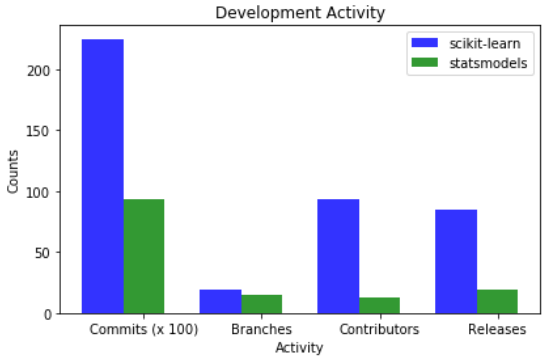

Fig.3: Development activities of different libraries by GitHub.

IV. CONCLUSION

This paper focused on why the Python programming language has a appropriate choice for learning and developing real world programming. It has gained wide popularity as the syntax is crystal clear to understand. Python allows one to easily leverage object oriented and functional design patterns mainly used to provide resolution for difficult problem in tiny time period by using least lines of code than other language. Python makes reproducibility and accessibility easier than R. This paper also discussed about packages which are used for analytics, web page creations, and real world applications development and so on. Most of the data science jobs can be done with five Python libraries they are Numpy, Pandas, Scipy, Scikit-learn, and Seaborn. Python provides a scalable, well-supported, and complete programming solution for Research, Analysis and scientific coding.

REFERENCES

[1] K. R. Srinath, "Python – The Fastest Growing Programming Language", International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 12 | Dec-2017.

[2] FankarArmashAslam, HawaNabeel Mohammed, Prof. P. S. Lokhande, "Efficient Way Of Web Development Using Python And Flask", Prof. P. S. Lokhande Volume 6, No. 2, March-April 2015.

[3] KalyaniAdawadkar, "Python Programming-Applications and Future", International Journal of Advance Engineering and Research Development,Special Issue SIEICON-2017, April -2017.

[4] Hoyt Koepke, "Why Python Rocks for Research" - Packages

[5]https://jakevdp.github.io/PythonDataScienceHandbook/04.14-visualization-with-seaborn.html

[6]https://www.kdnuggets.com/2017/06/top-15-python-libraries-data-science.html

[7]https://opensourceforu.com/2016/01/dot-a-language-that-helps-you-to-draw-graphs/
[8]https://www.analyticsvidhya.com/blog/2015/08/interactive-data-visualization-library-python-bokeh/
[9] https://pypi.org/project/statsmodels/
[10]https://blog.thedataincubator.com/2017/11/scikit-learn-vs-statsmodels/
[11] https://scikit-learn.org/stable/
[12]https://en.wikipedia.org/wiki/DOT_(graph_description_language)
[13] https://github.com