Original Research Paper

Volume-9 | Issue-6 | June-2019 | PRINT ISSN No. 2249 - 555X

Computer Science

A STUDY ON DATA SCIENCE WITH PYTHON IN COMPARISON TO OTHER LANGUAGES

Amita Dahiya PGT, Computer Science, Amity International School, Power Grid Complex, Sector 43, Gurugram – 122022.

ABSTRACT In this paper we will be discussing about Data Science, Data Science Flow Process, Python and Comparison of Python with other languages so as to analyze the difference and preference of Python as Data Science tool to extract the desired knowledge from the data available or gathered to improve the business strategies and also advances the scalable computing and data management. In this paper we will be focusing on the important of data science and how python is supporting data science.

KEYWORDS : Comparison, Data Science, Data Science Flow Process, Data Science Tool, Libraries, Python, Scalable Computing.

INTRODUCTION

Data Science is the field that deals with the progressive research in the study of processes and systems developed and being developed to extract the desired knowledge from the data available or gathered to improve the business strategies and also advances the scalable computing and data management.

Data Science is a vast research area which includes Data Mining, Algorithmic Innovations, Machine Learning, Pedagogy and Training.

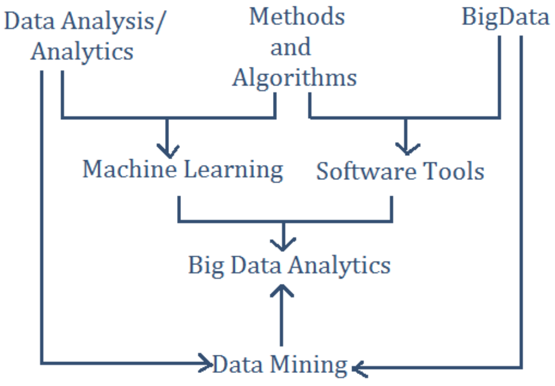

Figure1: Various fields of Data Science and their relation

Data Science Work Flow involves:

1. Machine learning field deals with building and implementing the top machine learning algorithms, creating workflows, and in general helping to solve machine learning problems. They provide the primary toolkit for different classification, regression, and other problems.
2. Data manipulation and analysis field represents libraries that carry out data scraping, ingestion, cleaning, pre-processing and other operations that allow you to process data and as a result to perform the analysis itself.
3. Visualization allows displaying the data visually which is necessary for better understanding and interpreting the data. These packages contain numerous visualization charts as well as different options for representation.
4. Libraries for mathematics and engineering provide the abilities to store numerical data in a convenient form and perform

DATA LIFE CYCLE

Data Life Cycle includes six stages:

1. Inquire: Inquire includes the possible queries for the data collection such as:

a. Which communities are more popular?
b. Is the user engagement increasing?
c. What is the distribution of publishing time?
d. What is the distribution of user interactions?
e. Is there a relationship between publishing hour and number of interactions?

2. Obtain: Obtain will find and use the resources for the collection of required data.

a. Download data from another location (e.g., a webpage or server)
b. Query data from a database (e.g., MySQL or Oracle)
c. Extract data from an API (e.g., Twitter, Facebook)
d. Extract data from another file (e.g., an HTML file or spreadsheet)
e. Generate data yourself (e.g., reading sensors or taking surveys)

3. Scrub: Scrub will identify the data that is not required or useful and refine the data accordingly.
4. Explore: Explore the refined data according to the survey/research done on specific domain of users/communities. Such as:

a. Which communities are more popular?
b. Is user engagement increasing?
c. What is the distribution of publishing time?
d. How is distribution of user interactions?
e. Relationship between distribution time and interactions

5. Model: Model actually develops the operational process for the survey made to benefit the business or community. Example:

a. Finding relevant word in the post
b. Similar posts

6. iNterpret: Interpret evaluation of the the model made on basis or data available and the requirement of user. Evaluation may be done by finding such as:

a. Drawing conclusions from your data
b. Evaluating what your result means
c. Communicating your results

Python is more of a general-purpose language with a rich set of libraries for a wide range of purposes. It's as good for mathematics, engineering, and deep learning problems as for data manipulation and visualizations. This language is an excellent choice for both beginners and advanced specialists which makes it extremely popular among data scientists.

Python is preferred over other data science tools. The reason is Python's Features which makes it to be a generous choice for data science. Python Features are: Easy to learn, Scalable, Variety of Libraries, it's Ecosystem and Visualization.

POWER OF PYTHON - LIBRARIES

Python's biggest power for data science lies in the great Python libraries and modules available for data science.

NumPy is the most important Python library for scientific computing with Python. NumPy provides a lot of convenient functions and data structures to work with numbers, linear algebras, random number generation, Fourier transformation, etc. Python is inherently slow for its dynamic behavior. So, a library written in Python could be slower than a library in C/C++. But, scientific computation and working with large arrays and multidimensional arrays requires the system to be faster. Python is great for working without numbers without thinking about limitations or the size of the number.

Matplotlib is a Python library for data visualization. It can visualize structured data in multiple forms and with a lot of customization. It can also output the visualization in multiple file types like, JGP, PNG, PMP, GIF, SVG, PDF, etc. It is fast and easy to make line graphs, pie charts, scatter plots, histograms, and other types of figures. Matplotlib was primarily created for 2D graphics, but it is also possible to create

3D graphics and effects.

Panda is a Python library that provides various high level convenient data structures, functions, and classes for easy and fast data analysis and manipulation operations. It is built on top of NumPy and thus it provides interchangeability between many other popular Python libraries that uses NumPy. Use of NumPy data structures and functions makes it relatively faster. It also works with a wide range of data format.

SciPy is a collection of modules that provide optimization, linear algebra, integration, interpolation, signal processing, image processing, etc. SciPy is also built on top of the versatile library NumPy and thus reuses all of its convenient features.

Scikit-learn is a machine learning library for the Python programming language. It provides very commonly used machine learning algorithms with a consistent API. It can be used to implement commonly used algorithms on various data sets. Some features include: classification, regression, clustering, etc.

Theano is similar to NumPy. It provides various numerical computation for Python. It can run both on CPU and GPU seamlessly. It helps define, optimize and evaluate mathematical expressions. Theano expresses computations using NumPy-esque syntax.

Scrapy is not a data analysis Python library, instead it is a data aggregation script development library built on top of Twisted. In data science we cannot work with data unless we get it for various sources. The web for example is a good resource for data. But data is not available through some clean API through these websites. So, we need to scrape them and extract the data out to put in a certain format. It is a hard task to create a scraper from scratch. With the help of Scrapy we can use it as a framework to scrape the web.

On the web data is not formatted in most cases and with machines it is a not possible to analyze the data unless it is not in suitable format for the machine. Beautiful Soup is the best library to parse a malformed web pages.

COMPARISON OF PYTHON WITH OTHER LANGUAGES IN REGARD OF USE IN DATA SCIENCE

According to the KDnuggets poll, the share of python in comparison to other languages and platforms usage for Analytics, Data Science and Machine Learning increases in two aspects: Loyalty and Switching.

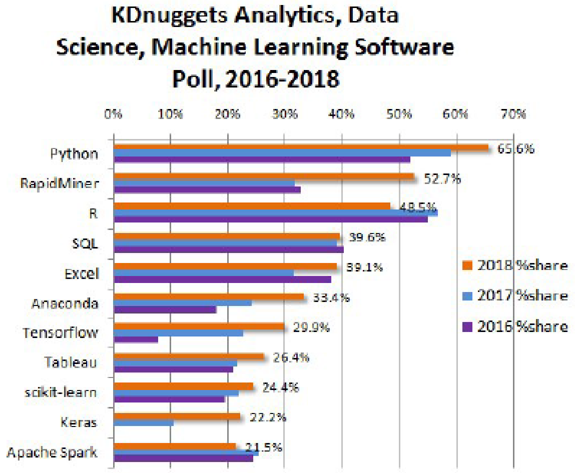

Figure2: KDnuggets 2018 Software Poll: Share of Python and other languages

Python seems to swallow not only R, but also most other languages, except for SQL, Java, C/C++ which remained at about the same level. R has declined for the first time since we have run this survey. Other languages have also declined.

CONCLUSIONS

In this paper, we discussed the features of Python which enriches the field of Data Science to extract the desired knowledge from the data available or gathered to improve the business strategies and also advances the scalable computing and data management. This is possible because of Python's vast library which leads this field to get heights as it makes it easy and fast to collect and analyze data. KDnuggets poll results of year 2018 also justified the survey done in the paper.

REFERENCES:

[1] Francine Berman, Rob Rutenbar, Brent Hailpern, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Margaret Martonosi, Padma Raghavan, Victoria Stodden, Alexander S. Szalay: "Realizing the Potential of Data Science", Communications of the ACM, , April 2018, Vol. 61 No. 4, Pages 67-72, 10.1145/3188721.

[2] Gabriel Moreira: "Python for Data Science", The Developers Conference 2015.

[3] Maruti Techlabs: "Is Python the most popular language for data science?" 2018.

[4] Igor Bobriakov: "Comparison of top data science libraries for Python, R and Scala", June 2018.

[5] Gregory Pietatsky: "Python eats away at R: Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis", KDnuggets, 2018.

[6] Md. Sabuj Sarker: "Top Python Data Science Libraries", DiscoverSDK Blog, 2018

[7] T. Giri Babu, Dr. G. Anjan Babu "A Survey on Data Science Technologies & Big Data Analytics", International Journal of Advanced Research in Computer Science and Software Engineering, Feb. 2016.