International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 08 | Aug -2017 www.irjet.net p-ISSN: 2395-0072

Study of Various Tools for Data Science

Gowrishankar.S[1]

[1]*Department of Computer Science and Engineering,*
Dr. Ambedkar Institute of Technology, Bengaluru – 560056, Karnataka, India.

Abstract – *Increased generation of heterogeneous data from different applications in various domains like healthcare, e-marketing and cloud based and their fast growth has led to huge demand to churn out information from the data at the disposal of corporate houses to increase profitability. This paper focuses on identifying different software packages that are available across variety of platforms to perform data analysis. These software packages span across different domains of Data Science like Machine Learning, Big Data and Deep Learning.*

Key Words: Python, R Programming, Big Data, Machine Learning, Neural Networks

1. INTRODUCTION

Data Science is a term which encompasses various fields like Machine Learning, Neural Networks, Data Mining and Natural Language Processing. Data Science involves collection of data from various sources, inferring the data and predicting the outcome of the data. Huge volume of data is being generated in the current digital era consisting of internet and heterogeneous sensor devices and organizations are faced with huge challenges of processing these data and have to come up with novel approaches and technologies to process the data. Data science as a domain is useful for health research, future predictions for business community and support real time opportunities [1, 2, 3].

The large amount of data that is getting generated requires sophisticated analytical techniques to analyze which requires sophisticated data science tools. Data science involves interdisciplinary domains with multiple data sources. Different analytical technique are required to analyze and share the insights that we gain from analyzing these heterogeneous complex data [4, 5].

Automation and analysis of the data effectively is the need of the hour to impart computational intelligence. Data Science platform should support Accessibility, scalability, robust and be cost effective [6].

In section 2, we identify different Python tools for analysis of data, various R programming packages are analyzed in section 3, brief discussion about Tableau is carried out in section 4. Finally we conclude the discussion in section 5.

2. PYTHON TOOLS

Scikit-learn is a popular machine learning package in Python ecosystem. Scikit-learn supports algorithms under both Supervised Learning and Unsupervised Learning. In supervised learning different categories of algorithms like classification, and regression are supported while under unsupervised learning various algorithms under clustering are supported. Scikit-learn comes with its own standard datasets such as iris and digits. Scikit-learn makes use of Numpy and Scipy to develop many of its algorithms. Scikit-learn provides well defined API end points to extend the existing models. If any new algorithms has to be included in the package then the new algorithms should support the fit, predict and transform theme of this package. Under supervised learning, algorithms like ordinary least squares, ridge regression, least angle regression, linear discriminant analysis, support vector machines, nearest neighbor classes, Naïve Bayes algorithms are supported. Under unsupervised learning, popular algorithms like Gaussian mixture models, manifold learning, K-means clustering are supported [7].

Statsmodels is a Python Package to implement various statistical models. Statsmodels provides statistical methods to perform statistical data analysis. Statsmodels is released under modified BSD license. Statsmodels support linear regression, generalized linear models, robust linear models, ANOVA, time series analysis, nonparametric methods, contingency tables, multivariate statistics, empirical likelihood and other algorithms [8].

Numpy package is used to perform complex scientific computation. It is the core package for other packages like Pandas which are developed on top of Numpy. It supports linear algebra, Fourier transform and other numerical capabilities. Numpy is licensed under BSD license [9].

Matplotlib is the most popular plotting library in Python. Publishing quality images can be produced using Matplotlib [10].

Seaborn is another plotting library in Python ecosystem. Seaborn is basically a wrapper around Matplotlib. Its functionalities are well suited to plot statistical techniques [11].

Plotly is a popular charting package available for different programming languages including Python. Plotly is available as community edition as well as paid edition. Plotly is well integrated with Pandas to perform data analysis. Plotly supports building of chart dashboards. Plotly makes use of cufflinks to work display plots [12].

Pandas is a popular library to perform data analysis. Pandas includes well defined data structures to perform highly

critical and scalable data analysis. Pandas can be used in production ready codebase. Data object manipulation is carried out using DataFrame. Data can be read from and be written to many different formats like CSV, XML, JSON, Microsoft Excel, SQL databases and HDF5 format. Functionalities like Time series representation, slicing, handling of missing data, merging, reshaping of data is supported. With the support to data frames, Pandas allow us to represent data in rectangular format. Pandas easily handles missing data [13].

H2O is a machine learning with support to in-memory. It is highly scalable and is supported across different programming languages. H2O can also be integrated with Big data technologies like Hadoop and spark. H2O supports popular deep neural networks and other popular algorithms. H2O can be easily integrated into various IDE tools like Visual Studio, PyCharm and RStudio. H2O models are optimized to deliver scalable and high performance even on large datasets [14].

spaCy has been projected as Natural Language Processing library with industrial strength. spaCy can be used with deep learning frameworks like tensorflow and other machine learning packages like Scikit-learn. spaCy claims to be the fastest among natural language processing packages [15].

Natural Language Toolkit (NLTK) is a very popular Natural Language Processing library used by academicians. It supports variety of corpora and lexical resources. NLTK is an open source software and is released under Apache 2.0 license [16].

TensorFlow is a Deep Learning Library released as an Open source package by Google. TensorFlow supports multidimensional data arrays and dataflow graphs. TensorFlow is even hosted on Cloud with a support for 1,000 cloud TPUs. Tensorflow is also available on Android platform. Multiple GPUs are also supported [17].

Keras provides APIs which act as frontend for other neural network libraries like TensorFlow. Keras allows developers to develop prototypes a lot quicker as it abstracts a lot of functionalities of the backend neural network library. Keras APIs are well defined and designed for faster experimentation of the kind of application that the developer is developing. New modules can be developed and integrated or the existing models can be extended by adding other functionalities [18].

Arrow is a popular library for date manipulation in Python. This improves upon the existing core datetime library found in Python. It has support for various time zones and provides various APIs for manipulating common scenarios like data difference, time difference and others [19].

3. R TOOLS

Tidytext is a package based on tidy data principle which ensures that the data is easy to manipulate, model and visualize. Tidytext ensures the task of mining the text data is done efficiently. Tidytext manipulates the text to a format which is much easier to work with other mining packages [20].

Readr is a popular package to read data from files like csv, tsv and other formats. Data is read in rectangular format where each column is the appropriate part [21].

Haven is a package that is used to transfer the data from R to other statistical packages like SAS, Stata and others. It also handles the missing values found in SAS, Stata and other statistical packages [22].

Feather is a package that helps in improving the interoperability of data between Python and R communities. Feather stores the data frames in an easy to use binary file format. Feather is designed to read and write quickly with columnar format. Feather supports different numeric types, dates, time and null types [23].

Rvest package is used to perform web scraping of HTML pages in R programming language ecosystem. HTML tags can be converted to data frames using Rvest package [24].

Tidyr makes data munging and manipulation much easier and allows the data to be much tidier than by using the default data.frame [25].

Dplyr is called “grammar of data manipulation”. It provides the various verbs to manipulate common data problems. It provides methods like select(), filter(), summarize() and others to for manipulation of data. Dplyr works seamlessly with data frames and the syntax is consistent with the built in manipulation method of R programming language [26].

Lubridate allows working with date and time much easier in R programming language. Lubridate code is vectorized allowing it to process date related queries faster [27].

Stringr provides a comprehensive list of string functions that makes working with string much easier. Stringr makes use of ICU C library which provides faster and correct implementation of string functions [28].

Broom package takes the output of various builtin functions like lm that are not in data frame format and converts them to tidy data frame format which can be used for further manipulation [29].

ggplot2 is the plotting package for R programming based on grammar of graphics. It provides powerful models to display complex plots with high quality aesthetics [30].

Purrr provides additional set of consistent functions to enhance the functional programming aspect of R programming language [31].

Shiny helps in building web applications by using R programming language along with minimal use of JavaScript. Shiny web applications support live changes to the application indicating that any changes made to the code gets reflected instantly. It provides efficient bidirectional communication between R code and the browser. These shiny applications can be integrated by other developers into their own applications through widgets [32].

Rmarkdown is used to prepare reports, presentation and dashboards of high quality and to share them for further analysis of the code. Rmarkdown documents have the code embedded into them which allows the code to be produced on other developer systems and to share them. Rmarkdown

can produce dynamic outputs in various formats like HTML, pdf, doc and other popular formats [33].

4. TABLEAU

Tableau is Graphical User Interface based Analytical software. Tableau is available as desktop version as well as Cloud hosted version. Tableau supports unlimited data analysis with drag and drop facility. Interactive dashboards are also supported. Tableau supports various data connections like SQL, spreadsheets and also google analytics. For power users Tableau supports R programming for data manipulation and visualization [34].

3. CONCLUSIONS

Due to unprecedented scientific and technological progresses we are seeing complex heterogeneous data being generated from variety of sources. Interconnection of large number of smart devices connected to internet plays a major role in generating these data based for different applications. In this paper we have identified and reviewed different packages in Python and R Programming ecosystem for analysis of data. In future we plan to propose and implement a framework that can be used to automate the analysis of data.

REFERENCES

[1] Nigel Davies and Sarah Clinch, "Pervasive Data Science", Proceedings of IEEE Pervasive Computing, Vol: 16, Issue: 3, pp: 5- -58, July 2017.

[2] Nirmal Keshava, "Opportunities for Data Science in Pharmaceutical Industry", Proceedings of IEEE Pulse, Vol:8, Issue: 3, pp: 10: 14, May 2017.

[3] Fang Cherry Liu, Fu Shen et al., "Building a research data science platoform from industrial machines", In Proceedings of 2016 IEEE International Conference on Big Data, Washington DC, USA, 5-8 Dec, 2016.

[4] Frank S Haug, "Bad Big Data Science", In Proceedings of 2016 IEEE International Conference on Big Data, Washington DC, USA, 5 – 8 Dec, 2016.

[5] Zhou Zhao, Hanqing Lu et al., "User Preference Learning for Online Social Recommendation", Proceedings of IEEE Transactions on Knowledge and Data Engineering, Vol: 28, Issue:9, pp: 2522 – 2534, September 2016.

[6] Narada Wickramage, "Quality assurance for data science: Making Data Science more scientific through engaging scientific method", In proceedings of Future Technologies, San Francisco, USA, 6 – 7, December, 2016

[7] Pedregosa et al., "Scikit-learn: Machine Learning in Python", Proceedings of JMLR, pp: 2825 – 2830, 2011.

[8] Seabold, Skipper, and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python." Proceedings of the 9th Python in Science Conference. 2010.

[9] Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22-30 2011.

[10] John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95, 2007.

[11] Seaborn, https://github.com/mwaskom/seaborn, 2017.

[12] Plotly, https://plot.ly/feed/, 2017

[13] Wes McKinney, "Pandas: A Foundational Python Library for Data Analysis and Statistics", PyHPC, 2011.

[14] Arora, A., Candel, A., Lanford, J., LeDell, E., and Parmar, V. (Oct. 2016). Deep Learning with H2O.

[15] Matt, 'https://spacy.io", 2017

[16] Bird, Steven, Edward Loper and Ewan Klein, "Natural Language Processing with Python", O'Reilly Media Inc, 2009.

[17] Google Brain Team, " https://www.tensorflow.org/", 2017.

[18] Fran Chollet , https://github.com/fchollet/keras, 2015.

[19] Chris Smith, "https://github.com/crsmithdev/arrow', 2017

[20] Silge J and Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R

[21] Readr, "https://github.com/tidyverse/readr", 2017

[22] Hadley Wickham, "https://github.com/tidyverse/haven", 2017

[23] Wes McKinney, "https://github.com/wesm/feather", 2017

[24] Rvest, "https://github.com/hadley/rvest", 2016

[25] tidyr, "https://github.com/tidyverse/tidyr", 2017

[26] Hadley Wickham, "http://dplyr.tidyverse.org/", 2017

[27] Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25.

[28] stringr, "https://github.com/tidyverse/stringr", 2017

[29] broom, "https://github.com/tidyverse/broom", 2016

[30] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

[31] Purrr, "https://github.com/tidyverse/purrr", 2016

[32] Shiny, "http://shiny.rstudio.com/", 2017

[33] R Markdown, "http://rmarkdown.rstudio.com/", 2017

[34] Christian Chabot, Chris Stolte, Pat Hanrahan, "Tableau Software", Seattle, Washington 2003