International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I1P109

Volume 68 Issue 1, 38-41, January 2020

Review Article

Programming Languages for Data Mining A Review

Neha Walia[1], Arvind Kalia[2]

[1]*Research Scholar, Department of Computer Science Himachal Pradesh University, Shimla, Himachal Pradesh, India*
[2]*Professor, Department of Computer Science Himachal Pradesh University, Shimla, Himachal Pradesh, India*

Received Date: 17 November 2019
Revised Date: 12 January 2020
Accepted Date: 19 January 2020

Abstract *- Explosive growth of data from zettabytes to yottabytes resulted in the flooding of data in this digital universe, but the availability of knowledge is very much limited. There are prevalent tools and techniques available nowadays to mine data and find useful knowledge from it. Data mining is one of the utmost commonly used approaches for exploring new relations /knowledge from huge databases. There are a number of programming languages available to accomplish this task. The paper focuses on the study of four prevailing programming languages, namely R, Python, Julia and Java, with their technical features. The study emphasizes the programming languages that are easily available, open-source, high level, platform-independent and object-oriented. The study can proliferate the choice of programming languages for extracting valuable information by the programmers and developers.*

Keywords *- Programming Languages, Data mining, R, Python, Julia, Java, Open Source, Object-Oriented.*

I. INTRODUCTION

Information needed in today's world must be very precise and crisp in order to make the right decisions. But the extraction of meaningful information is a burdensome task. The ease of availability of data is due to the existence of high-tech data collection tools, database systems, www and digitized culture. Since it is not possible of exploring data using traditional tools and techniques, there is a requirement of some method/technique that works on vast unprocessed data. Data mining is one of the dominant techniques nowadays used for information extraction. Data mining, in particular, is a data analysis technique that stresses statistical modelling and knowledge discovery for correct decision making. Discovering interesting, unexpected or valuable structures from huge datasets is the working description of data mining [1] [2].

Data mining is not a current discipline. Data dredging, fishing through data, to name a few, are earlier names used for this technique. Present-day data mining technique is a collection of statistics with tools and methods from computer science, machine learning, database technology and other standard data analytical technologies [3]. Data mining is the field of computer science where a huge amount of data is processed to extract some useful data. Data mining is a very important technique for identifying future trends and behaviours [4]. Many data mining techniques have been studied to process and analyse several types of data patterns, which includes Classification, Summarization, Association Rule, Regression, Visualization and Clustering [5].

The technique of data mining comprehensively depends on computer processing speed and the methods used for data collection. Data mining tools are used to accurately forecast forthcoming actions to make knowledgeable decisions. In the current scenario, there is a need to have excellent and powerful programming languages through which data mining can be implemented. There is a number of programming languages available, but the paper is restricted to four significant and dominant languages for data mining.

II. PROGRAMMING LANGUAGES FOR DATA MINING

A. *R LANGUAGE*

R, open-source, interpreted programming language, was developed by Ross Ihaka and Robert Gentleman in the year 1993 at the Department of Statistics of the University of Auckland, Auckland, New Zealand. Language is mainly used for statistical computing and graphics and is maintained by R Foundation for Statistical Computing [6]. The R language is extensively used by developers and programmers for the analysis of data and the development of statistical tools and software. It is a collection of a language, run-time environment with graphics, a debugger, accessibility to certain system functions and the capability to execute programs warehoused in the form of scripting files [7][8]. R language interface is available as a command-line as well as the graphical user. Its design has been comprehensively based on the features of S language and semantics from Scheme language. This is written chiefly in C, FORTRAN and R itself.

A number of statistical functions and procedures are available in the R language, such as linear and generalized linear models, nonlinear regression models, time series analysis, to name a few[7][8]. The r language is a powerful, modest and effective language for computing. It handles data and storage efficiently and effectively. R language contains a collection of operators for calculations on arrays, lists, vectors and matrices [9]. The language supports an enormous, intelligible and integrated group of transitional tools and graphical interfaces for data exploration. The r language is one of the dynamical languages used nowadays used for statistical computing, designing and data analytics. It consists of huge collections of packages supporting all the areas of computing. R also supports a number of various features for machine learning procedures such as classification, regression and developing artificial neural networks.

One of the important strong points of this language is the simplicity with which well- designed excellent quality plots can be created by using mathematical operations [10]. Technical specifications of the R language are prepared in a tabulated format that is depicted in Table 1.

Table 1. Technical specifications of R language [10] [11]

Developer	R Core Team
Operating System	UNIX platforms and similar systems (including FreeBSD and Linux), Windows and Mac OS.
Designed by	Ross Ihaka and Robert Gentleman
License	Free Software Foundation's GNU General Public License
Latest Released Version	3.6.2
Latest Released Date	December 12, 2019
Website	www.r-project.org
Released year	August 1993
Language	Interpreted
File extensions	.r, .rdata, .rds, .rda

Table 1 portrays various technical characteristics of the R language. Features such as license, official logo, operating system, released year, latest released version and so on are included.

B. *PYTHON*

Python is a widespread, powerful, open-source, interpreted and object-oriented programming language developed by Guido Van Rossum in the year 1990 at the National Research Institute for Mathematics and Computer Science in the Netherlands. According to a survey done in 2019 by KDnuggets, Python is the most popular and promising programming language [12]. Python is prevailing, developer-friendly, general-purpose, easy to use and learn language having community help to support the beginner to expert users and adds to the ever-increasing open-source knowledge base [13][14]. It is a portable, flexible, integrated, extensible and dynamic language that supports other languages such as C, C# and Java [15] and GUI applications. It consists of huge standard libraries that provide strong interfaces to databases.

A number of third-party modules for Python are hosted in Python Package Index. Because of the standard library of Python and the community-contributed modules, this language is used in various application areas such as Web & Internet Development, Database Access, Desktop GUIs, Scientific & Numeric applications, Education, Network Programming, Software & Game Development [16]. Features of Python language are prepared in a tabulated format and are depicted in Table 2.

Table 2. Technical specifications of python language [16] [17]

Developer	Python Software Foundation
Operating System	Windows, Linux/Unix, Mac OS X, Solaris, AIX, IOS
Designed By	Guido Van Rossum
License	Python Software Foundation
Latest Released Version	3.8.1
Latest Released Date	December 18, 2019
Website	www.python.org
Released Year	1990
Language	Interpreted
File Extensions	.py, .pyi, .pyc, .pyd, .pyo (prior to 3.5)[18],.pyw, .pyz (since 3.5)[19]

Table 2 exposes various technical characteristics of the Python language. Features such as license, file extensions, operating system, released year, latest released version and so on are incorporated.

C. *JULIA*

Julia is an interpreted, dynamic, high-level scripting language developed by Jeff Bezanson, Alan Edelman, Stefan Karpinski, Viral B. Shah in the year 2012 at the Massachusetts Institute of Technology (MIT). It provides powerful tools for artificial intelligence, machine learning and deep learning. Though Julia is a scientific purpose language, it is not limited and can be used for web and general-purpose programming. Julia is a modern, extensible, and expressive, high-performance programming language designed for scientific computations and data manipulations [20]. Julia is a rich language for descriptive data types and type declarations that can be used to clarify and solidify programs [21].

It consists of wide-ranging mathematical based library functions with great arithmetical precision. Fast, easy to learn and use, and open-source features of this language help in performing extremely rigorous computing tasks. It

is a language designed for data visualization and plotting, scientific and parallel computation in a distributed environment and data manipulation [22]. Julia provides efficient, specialized and automatic generation of code for different argument types. Lightweight green threading devectorised code, compact and fast user-defined types, powerful shell-like capabilities for managing other processes in the system are a few of the important features of Julia language. Julia also works with nearly all databases and also integrates with the Hadoop ecosystem [21]. Julia incorporates the convenience of dynamic languages like Python and the integration of open-source [23]**.** Features of Julia language are prepared in a tabulated format that is depicted in Table 3.

Table 3. Features of julia language [21] [22]

Developer	Jeff Bezanson, Stefan Karpinski, Viral B. Shah and other contributors
Operating System	Linux, Mac OS, Windows and FreeBSD
Designed by	Jeff Bezanson, Alan Edelman, Stefan Karpinski, Viral B. Shah
License	MIT (core) GPL v2
Latest Released Version	1.3.1
Latest Released Date	December 30, 2019
Website	www.JuliaLang.org
Released year	2012
Implementation Language	Julia, C,C++,Scheme, LLVM
Availability	Open source
File extensions	.jl

In Table 3, a number of technical characteristics of Julia language are depicted. Features such as license, official logo, operating system, released year, latest released version and so on are integrated.

D. JAVA

Java is a high level, interpreted, dynamic, strongly typed, flexible, object-oriented programming language and platform designed by James Gosling in the year 1995. Initially, it was designed only for writing code for handheld devices and set-top boxes. It is an open-source, simple, strong, secure and portable language. It has the features of high performance, multi-threaded concurrency and platform independence. Java is the best language for cloud development with Oracle cloud platform and infrastructure services [24].

Java assures to be write-once, run-anywhere language. On compilation, the Java program develops bytecode on a compilation that can run on any Java Virtual Machine (JVM) irrespective of the underlying internal computer design, is platform-independent and secure. Java language is very supportive in developing web, mobile, desktop GUI, enterprise and scientific applications. Deallocations of the objects are done automatically by means of a garbage collector program.

Java programming language syntax is mostly influenced by the C++ language. Contrasting to C++, which associates the syntax for structured, generic and object-oriented programming, Java is built almost exclusively as an object-oriented language [25]. Java does not support operator overloading or multiple inheritances for classes. Technical features of the Java language are prepared in a tabulated format and are illustrated in Table 4.

Table 4. Features of java language [24] [25]

Developer	Sun Microsystems
Operating System	Cross Platform
Designed By	James Gosling
License	Oracle License
Latest Released Version	Java SE 13
Latest Released Date	September 17, 2019
Official Website	www.oracle.com/java/
Release Year	May 23, 1995
Language	Interpreted and compiled
Availability	Open-source
File Extensions	.java, .class, .jar

In Table 4, a number of technical characteristics of the Java language are represented. Features such as license, operating system, released year, official website, availability, latest released version and so on are integrated.

III. CONCLUSION AND FUTURE SCOPE

Coding through programming languages is a significant aspect of computer systems. The paper focuses on the study of only four prevalent programming languages, namely R, Python, Julia and Java, with their importance and technical features. The study emphasized the programming languages that are easily available, open-source, high level, platform-independent and object-oriented. The analysis of these programing languages reveals the important characteristics which are required for solving today's era problems. All languages are popular languages, having their own significance. The r language is mainly used for statistical computing and graphics, and Python is often acclaimed for its general-purpose feature with simple and easy-to-understand syntax, Julia is a high-performance programming language designed for scientific computation and data manipulation, and Java is a language with write-once, run-anywhere feature, multi-threaded concurrency and platform independency. The study can be further explored by considering more prevalent programming languages for data mining.

REFERENCES

[1] D.J Hand, H. Manila and P Smyth, Principles of Data Mining, Cambridge (MA): The MIT Press., ISBN: 978-0-26-208290-7. (2001)

[2] S.Sreekanth and P.C. Rao., Anomaly Detection Using Data Mining Methods., International Journal of Computer Trends and Technology (IJCTT) 67(12) (2019) 1-4.

[3] David J. Hand., Principles of Data Mining, Drug Safety., 30(7) (2007) 621 – 622.

[4] M. Dunham, Data Mining Introductory and Advanced Topics, Prentice-Hall, ISBN 978-0-13-088892-1, (2003).

[5] Han Jiawei, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd Ed., Morgan Kaufmann Publishers, ISBN 978-0-12-381479-1, (2011).

[6] R Core Team., R: A language and environment for statistical computing., R Foundation for Statistical Computing, Vienna, Austria. [Online]., (2016) Available: http://www.R-project.org/.

[7] Kurt Hornik., R FAQ What is R (2018). [online]. Available:https://cran.r-project.org/doc/FAQ/RFAQ.html.,

[8] Kurt Hornik, R Foundation [online]. Available: https://cran.r-project.org/doc/FAQ/R-FAQ.html., (2018).

[9] R language (2019). [Online]. Available: https://www.tutorialspoint.com/r/r_overview.html.,

[10] R language [Online] Available: https://www.r-project.org/.,(2020).

[11] language [Online]. Available: https://en. wikipedia.org/wiki/R_(programming _language)., (2019).

[12] Data Science Software Poll (2020). [Online]. Available: https://www.kdnuggets.com/2019/05/new-poll-software-analytics-data-science-machine-learning.html.,

[13] M. Pastell., Teaching Instrumentation and Data Analysis Using Python., International Conference on Agricultural Engineering, (2016).

[14] E. Mészárosová., Python and Teaching Programming at Upper Secondary Schools., International Conference on Information and Communication Technologies in Education.(2015).

[15] Python (2019). [Online]. Available: https:// www. Edueeka.co/blog/python-features/

[16] Python (2020) [Online]. Available: https:/ /www .python.org/about/.,

[17] Python [Online]. Available: (2020). https://en.Wiki pedia.org/wiki/Python(programming language).,

[18] Holth Moore., PEP 0441—Improving Python ZIP Application Support., Retrieved on (2019).

[19] PEP 488--Elimination of PYO files 2015 Retrieved on (2019).

[20] Julia (2020). [Online]. Available: https://www. cseworldonline.com/articles/features-of-julia.php/.,

[21] Julia (2020). [Online]. Available: https://julialang.org/.,

[22] Julia (2020). [Online] Available: https:// en.wikipedia.org/wiki/Julia_(programming_language)/.,

[23] T.A, Cabutto, S.P. Heeeney, S.V. Ault, G. Mao and J. Wang., An Overview of the Julia Programming Language., Proceedings of 2018 International Conference on Computing and Big Data-ICCBD '18., (2018).

[24] Java,(2020). [Online]. Available: https://go.java/ index. html/.

[25] Java [Online]. Available: https://en.wikipedia.org/wiki/ Java_(programming_language)/.