TOP PROGRAMMING LANGUAGES for a DATA SCIENTIST

simplilearn

TABLE OF CONTENTS

Everything around us is related to Big Data – all the digital bits of information that help businesses grow. Programming languages have become a big part of Big Data. The programming languages make the programming process faster and easier, and in turn, helps find its growing importance in the current bombarding field of Big Data and Analytics.

This eBook lists out the top 10 programming languages for Data Scientists, in the order of their importance and use.

1 R PROGRAMMING LANGUAGE

As a programming language and software environment for graphics and statistical computing, R is recognized as one of the most popular languages amid data scientists. It is a vital tool for analytics and finance driven businesses like Facebook, Google, and LinkedIn.

Combined with the dialect of the S language and lexical scoping semantics, the R system is on a rapid rise. Every few months the language adds new features and abilities which are driving it as the most important tool for visualization, computational statistics, and data science.

There is no cost to any of the versions provided by this programming language. The 32-bit versions of Microsoft Windows are available for Linux; and OS X for UNIX and Macintosh. It is also available through the Comprehensive R Archive Network (CRAN).

The R Language was developed at AT & T Bell Laboratories by Rick Becker, John Chambers, and Allan Wilks and released in 1995.

As an interpreted language, R's users frequently access it through a command-line interpreter. When a user types in 2+2 at the R command prompt and enters the data, the computer replies with 4.

The data structure used by R includes vectors, matrices, arrays, data frames (similar to tables in a relational database) and lists.

★ FEATURES

- Supports matrix arithmetic
- Freely available under the GNU – general public license
- Contains pre-compiled binary versions that are available for every operating system
- Command line interface is used in the language
- Implements a wide variety of statistical and graphical techniques including linear and non-linear modeling, time-series analysis, classical statistical tests, classification, clustering, and others
- Because of functions and extensions it is easily extensible
- Most of R’s standard functions are written in the language itself which makes it simple for the users to follow the algorithm changes that are made
- Unlike most statistical computing languages, R has stronger object-oriented programming facilities
- The extending of R is simplified by its lexical scoping rules
- It contains static Graphics which produces publication-quality graphs, including mathematical symbols
- Through additional packages, interactive and dynamic graphics are available.
- Unlike other programming languages, R has its own LaTeX-like documentation format. This format is useful in supplying documentation that is comprehensive, both on-line in a number of formats and in hard copy.
- For some functions, R supports the procedural programming while for others it supports the object-oriented programming.

Data scientists and statisticians around the world use this programming language to solve some of their most challenging problems in fields that range from computational biology to quantitative marketing.

Since complex data is represented through charts and graphs, the language has become an essential part of the data analysis process.

Thought leaders in data visualization like Bill Cleveland and Edward Tufte, have influenced the language making it easy to draw meaning from multidimensional data with multi-panel charts, 3D surfaces and more.

Essentially every data manipulation, chart and statistical model that a modern data scientist needs ID included in the language.

You can easily find, download and use this cutting edge community reviewed method in statistics and predictive modeling from leading researchers in data science.

Download here for free.

2 PYTHON PROGRAMMING LANGUAGE

Python is a general purpose, high-level programming language. The design philosophy of Python highlights code readability. Programmers are allowed to express the concepts they use in fewer line codes than that which would be needed for languages such as Java, or C++, because of the syntax in Python.

The idea of a programming language like Python was established in the late 1980s, and it went on to being implemented in December 1989 by Guido van Rossum at CWI in the Netherlands. It was released in 1991.

It developed as the successor to the ABC language to be capable of exceptional handling and interfacing with the Amoeba Operating System. The language was also inspired by C, and Modula 3.

★ FEATURES

- The language equips constructs with the intention of enabling clear programs on both large and small scale.
- The language supports various programming paradigms, including imperative, object-oriented, and functional programming or procedural styles.
- It contains a dynamic type of system, a large comprehensive standard library and an automatic memory management system.
- The major advantage it holds is its breadth. An example to explain this: On a preprocessed dataset, R can run Machine Learning Algorithms. Python, however, is much better at processing data. Python uses Panda, which is an immensely useful library that can do everything that SQL or R does, plus more.

contd..

★ FEATURES contd..

- The program also transcends as a glue language for applications that are written in C, C++ and Forton.
- In the programming world, Python is known for its simplicity.
- The strength of it is its core libraries: NumPy, SciPy, Pandas, Matplotlib, IPython.
- The language offers high productivity in terms of prototyping and building reusable and small systems
- It is a general purpose programming language.

Data scientists are often involved in wiring together network applications, scripting and automating data processing jobs, programming for the web and other processes including data munging. They find it desirable to do all of this, in addition to the actual analysis and modeling in one single language.

To be skilled in an all-in-one language, professionals begin learning the Python language. Python is used by the larger companies for important purposes to evaluate vast data sets, which puts Python programmers on high demand.

To download Python, please click here

3 MATLAB PROGRAMMING LANGUAGE

Matrix Laboratory, commonly known as MATLAB, a fourth gen programming language, is a multi-paradigm numerical computing environment.

Developed by MathWorks, it was initially released in 1984.

Matlab crossed over one million users across the industry and academia in 2004.

Users of Matlab emerge from various backgrounds like engineering, economics, and science. The programming language is used in a number of research institutions and industrial enterprises.

★ FEATURES

- The programming language allows platting of functions and data, matrix manipulations, implementation of algorithms, interfacing with programs written in other languages, including C, C++, Forton, Java, and Python, and creation of user interfaces.
- Though the language was initially intended for numerical computing, the language now offers an optional toolbox that uses the MuPAD symbolic engine that allows access to symbolic computing capabilities.
- Simulink, an additional package that is offered, adds multi-domain graphical simulation and model-based design for dynamic and embedded systems.
- The language has an interactive environment for design, iterative exploration and problem solving.
- It offers mathematical functions for statistics, linear algebra, filtering, Fourier analysis, optimizations, solving ordinary differential equations, and numerical integrations.
- It has built-in graphics for visualizing tools and data to create custom plots.
- Development tools are present to improve code quality and maintainability and to maximize performance.

Though it seems very limiting, a wide range of the data analysis and scientific community use this language to solve problems that are represented as matrix problems. With this programming language, data scientists can perform analysis to gain insight into the data quicker than those like C, C++, or Visual Basic. Data Analysts use MATLAB to access data from spreadsheets, files, databases, data acquisition hardware, and other software, explore data for the identification of trends, estimate uncertainty, and test a hypothesis. It helps the data analyst in creating customized algorithms, models, visualizations and publishes customized reports.

Download MATLAB, here

4 HADOOP PROGRAMMING LANGUAGE

An open source framework, Hadoop that is used for distributed processing and distributed storage of large data sets. In simple terms, it is a framework that allows professionals to process and store big data across clusters of computers in a distributed environment using simple programming models. The language is designed to scale up from single servers to millions of machines, each of which offers local storage and computation.

Hadoop is written in Java; all the modules are devised with the central assumption that hardware failures are ordinary and common and should be handled automatically in a software.

Hadoop was created by two Yahoo employees Doug Cutting and Mike Cafarella in 2005. Released in 2011, Hadoop uses a cross-platform operating system.

The base Apache Hadoop framework is composed of the following modules:

- Hadoop Common – this module consists of utilities and libraries that are essential to other Hadoop modules.
- Hadoop Distributed File System (HDFS) – The HDFS module is a distributed file-system that is involved in storing data in commodity machines that provide high aggregate bandwidth across a cluster.
- Hadoop YARN – YARN is a resource management platform that handles the management of computation of resources in clusters and uses them for the scheduling of users' applications.
- Hadoop MapReduce – is a programming model that is used for large-scale data processing.

Hadoop opened new roads for data scientists to store and process data. Instead of depending on proprietary hardware and other systems to process and store data, Hadoop allows parallel distributed processing of massive amounts of data across industry standard servers that will process and store data. With Hadoop, there is no data that is too big.

Download Hadoop, here

5 SQL PROGRAMMING LANGUAGE

The Structured Query Language (SQL) was developed at IBM by Donald D Chamberlin and Raymond F Boyce in the 1970s. Originally the language was based on relational algebra and Tuple relational calculus.

It is a special-purpose programming language devised to manage data that is held in a Relational Database Management System (RDMS) or for the purpose of stream processing in a Relational Data Stream Management System (RDMS).

One of the immensely popular languages today, SQL as a language is endorsed for all the new data types and data cases. The need for SQL in Big Data is not just a default choice, but a realization that SQL is one of the best-suited languages for basic analysis.

The concept of this language is constructed from the idea of relational algebra that is a framework that is used for the organization and manipulation of data sets. The SQL syntax briefly uses this mathematical system.

★ FEATURES

- The language consists of data definition language, data control language and data manipulation language.
- The language is an ISO and ANSI standard computer language to create and manipulate databases.
- Allows users the creation, updating, deletion and retrieving data from a database.
- The language itself is easy and simple to learn.
- The language works with programs like Oracle, DB2, Sybase, MS Access, MS SQL Server, etc.
- SQL is vital in the data scrubbing stage.
- It allows querying and extraction of meaningful data from large and complex databases.

6 SAS PROGRAMMING LANGUAGE

A computer programming language, SAS is used for statistical analysis. It originated as a project that was run by the North Carolina State University. The language reads in data from the common databases and spreadsheets. It then sends out the results of the analysis in the form of tables and graphs.

The language runs under compilers that are used on Linux, Microsoft Windows, and various other UNIX and mainframe computers.

The language has been the undisputed market leader in the commercial analytics space.

SAS was developed from 1966 until 1976 at the North Carolina State University.

FEATURES

- The language offers an array of statistical functions.
- It has a good GUI for people to learn and provides even better technical support.
- Holds the highest market share in the Private Organizations.
- The language is easy to learn and provides an easy option for professionals who already have an established knowledge in SQL.

Download SAS, here

7 JAVA PROGRAMMING LANGUAGE

Green Team, a team of engineers, initiated the language in 1991. Originally, Java was known as Oak, being designed for handheld devices and set up boxes. Being unsuccessful as OAK, in 1995, the name was changed to JAVA and modified to take advantage of the World Wide Web.

Java, a general purpose, object oriented programming language, is considered to be among the top for developers and programmers. Currently, it stands at the top for the best programming language and has bagged the highest position with Android OS.

The language is used in various platforms like mobile-based applications, the creation of desktop applications, enterprise level purpose, establishing Android apps on smart phones and tablets.

★ FEATURES

- The intention of the application is to 'Write Once Run Anywhere' (WORA) – writing the code once and the code being able to run on all the platforms that support the language.
- The syntax for the language is derived from C++.
- The language is simplified to eliminate language features like those that cause common programming errors.
- The importance of Java comes from its vast array of libraries that are offered as a solution to most of the common problems that a professional encounters while developing enterprise applications.
- Simple: Easy to read and write – most concepts are drawn from C++.
- Secure: The program cannot harm any other system program, provides a secure way of building internet applications.
- Portable: The programs in Java can run in any environment where there is a Java run-time system.
- Multithreaded: it provides support for multithreaded programming.
- Dynamic: programs of Java carry a substantial amount of run-time information which is used to verify and resolve access to objects at run time.

Download Java, here

8 C++ PROGRAMMING LANGUAGE

A general purpose language, C++ is an important language in high-volume, high-frequency trading.

C++ has objective oriented and generic programming features that provide the facilities for low memory manipulation.

The language was developed by Bjarne Stroustrup and released in 1983.

FEATURES

- It is based on the earlier C language.
- The language uses features of Simula.
- C++ stands as the only language that supports RAII – it provides control over the lifetime of objects, unlike other management methods.
- Adding to its performance is its modular programming and closeness to hardware.
- Top tech-based companies like Facebook, Apple; use the language because of the brevity of its codes.

9 JULIA PROGRAMMING LANGUAGE

The language first appeared in 2012, and was developed by Jeff Bezanson.

As a high-level dynamic programming language, Julia was created to address the needs of high performance scientific and numerical computing in addition to being useful to general purpose programming.

★ FEATURES

- It has a multiple dispatch that provides the ability to define functional behavior across a combination of argument types.
- A dynamic type system for documentation, dispatch, and optimization.
- It has a built-in package manager.
- It contains Metaprogramming facilities: lisp-like macros.
- It uses the PyCall package.
- To manage processes, it has a powerful shell-like capability.
- It is designed for parallelism and distributed computation.
- User-defined types are as fast and compact as built-ins.
- Automatic generation of efficient, specialized code for different argument types.
- Elegant and extensible conversions and promotions for numeric and other types.

Download Julia, here

<GOOD LUCK>