Chapter 11
Data Mining for Junior Data Scientists:
Basic Python Programming

ABSTRACT

The availability of off-the-shelf tools for data mining has made it easier to process data. However, in many cases, such software packages are not flexible enough to allow for algorithmic improvements. Therefore, data scientists need to write computer programs to customize the processing methods in tandem with the software packages. This chapter introduces Python, a programming language that provides libraries that support data science work. The content includes Python programming syntax, such as variables, structured programming, decision-making programming, recursive programming, data structure handling, and file handling. Additionally, the chapter introduces Google Colab as a tool for programming experiments. It provides a crucial foundation for data science students who are going to process data using data mining techniques in the next chapter using Python programming. Although data scientists don't need a deep understanding of computer programming, learning computer languages is essential, and this chapter caters to beginners.

INTRODUCTION

A number of industries, nowadays, focus on data science as they recognize the value of data to support decision-making (Raschka & Mirajalili, 2018). Whether it is the medical industry, automotive industry or the scientific community, data scientists have become a sought-after position for many organizations. A data scientist can either be developed from the knowledge expert within the organization itself or being adopted from outsiders. Anyhow, every organization analyzes data on the principle of questioning, seeking resources, exploring data and pre-Processing, analyzing data using data mining techniques, and presenting the results to the questioners. At present, there are many tools for exploring and preparing data, as well as machines for using data mining techniques to analyze data. One of these tools is Python, which provides a library of commands for developing visual data, exploring and preparing data, as well

DOI: 10.4018/978-1-6684-4730-7.ch011

as analyzing data with data mining techniques. In addition, data scientists can process data with Python's algorithms without limiting the size of the data. Python itself is considered an open source that allows organizations to analyze data with data mining techniques, which reduces the cost of digital resources (Sneeringer, 2016: Reges, Obourn & Stepp, 2019). Therefore, Python is a useful tool for beginners in data science (Bowles, 2015: Joshi, 2017).

INTRODUCTION TO PYTHON

Python, developed by Buido van Rossum in 1980, is a high-level language which is so close to human language, making it ideal for beginners in computer programming languages (Sunkpho & Ramjan, 2020: Tanantong & Ramjan, 2022). It also has an easy-to-remember syntax, so software developers can use Python for programming in both the Structural Programming and the Object-Oriented programming styles (Hill, 2015: Bader, 2018: Mueller, 2018: Lubanovic, 2019). Python is an open software source, so software developers worldwide can work together to develop Python without the cost of licensing (Stewart, 2017: Zelle, 2010).

Software developers can use software editors such as Google Colab to write Python and use the interpreter to translate Python in order to run the digital device with the process as shown in the following:

1. Software developers write a set of commands on a software editor.
2. The interpreter translates Python into the language that the digital devices can operate.
3. The digital devices accept the commands for processing.
4. The digital devices operate as commanded.
5. Software developers check the operation accuracy.

In the figure, software developers can write a set of commands on the Software Editor and then uses Interpreter to translate Python into the format that can be run by digital devices. The digital devices then receive the command and process it to perform the tasks as designed by the software developers. In the final step, software developers can verify whether the digital devices work as designed so that their command sets can be improved to make digital devices work more accurately.

Python is Case Sensitive, meaning lowercase and uppercase English letters are different. For example, dX and DX, when interpreted, the letters will totally be different. Although developers have different duties from data scientists, when data scientists intend to use programming for data analysis, they can begin their Python study with the following basic commands.

Variables

Data scientists can analyze data by packing it into variables within Python. The variables act as a container containing data for data transfer, processing, and display. The process goes as presented in the figure below.