Python: Empowering Data Science Applications and Research

Alunos: Gabriel Samarane, Gabriel Araújo, Arthur Matos, João Madeira, Vitória Símil

PUC Minas

Informações Do Artigo

Autores: Mritunjay Kr. Ranjan, Krishna Barot, Vaishnavi Khairnar, Vaishnavi Rawal, Anujaa Pimpalgaonkar, Shilpi Saxena, Arif Md. Sattar

Retirada: Journal of Operating Systems Development & Trends. Agosto de 2023

PUC Minas

Introdução

- O objetivo do estudo é discutir como todos os recursos fundamentais para desenvolver aplicações Python podem ser encontrados em um único lugar com a ajuda de um IDE.

- A pesquisa começa com uma visão geral das bibliotecas de Ciência de Dados em Python, e depois passa para uma discussão sobre as diversas aplicações práticas da ciência de dados.

PUC Minas

Introdução

- Diferencia-se de outras linguagens por prover todas as ferramentas necessárias para as aplicações de ciência de dados
- Python é extremamente usada na indústria de análise de dados
- Útil na coleta, análise, modelagem e visualização dos dados
- Também muito usado na área de IA e Machine Learning

PUC Minas

Características de Python

Table 1. Features description of python.

Citation	Features	Description
[1]	Easy to learn	Python's most appealing attribute is that it can be rapidly and easily learned by anybody, even complete novices; this is one of the reasons why learners favor python for data science. Another reason is that it is one of the most widely used programming languages.
[4]	It is an open source	Python is an excellent programming language in terms of its capacity to scale. It allows data scientists the ability to be flexible and to approach various challenges in many ways.
[5]	Scalability	Python is becoming the language of choice for data science tasks because of its growing popularity in the data science community, which provides users with access to a variety of cutting-edge processing tools and methods that are now on the market.
[6]	Vast ecosystem of libraries	Python's data science libraries include NumPy, Pandas, Matplotlib, SciPy, and Scikit-learn. These libraries include fundamental data processing, numerical calculation, visualization, statistical analysis, and machine learning techniques.
[7]	Data handling capabilities	Data handling: Pandas includes strong data structures like Data Frames and tools for data cleaning, transformation, merging, and filtering. Data manipulation and analysis are simplified.

Características de Python

[7]	Data handling capabilities	Data handling: Pandas includes strong data structures like Data Frames and tools for data cleaning, transformation, merging, and filtering. Data manipulation and analysis are simplified.
[8]	Support for numerical computing	Numerical computing: Python's NumPy library efficiently handles multi-dimensional arrays and matrices. Data science numerical computations require vectorized operations and a wide range of mathematical functions.
[9]	Visualization capabilities:	Matplotlib, Seaborn, and Plotly are Python data visualization libraries. These libraries let data scientists produce relevant and attractive plots, charts, and graphs to communicate insights and conclusions.
[10]	Machine learning and AI libraries	Scikit-learn, TensorFlow, Keras, and PyTorch are strong Python libraries for developing and deploying machine learning models. Python is ideal for AI application development because these libraries offer classification, regression, clustering, and deep learning methods.
[11]	Integration and interoperability	Python integrates easily with C, C++, and Java. Interoperability lets data scientists use existing code, low-level performance, and task-specific libraries.
[12]	Large and active community	Developers and data scientists use Python in large numbers. This community develops open-source libraries, supports newbies through forums and online resources, and offers many tutorials. Python is an accessible data science language due to community support and resources.

Características do Python

[13]	Data Structure	Sets, lists, dictionaries, queues, strings, and other data structures are all available in Python. When developing an algorithm, it is essential for a programmer to utilize the appropriate DS. This is especially necessary in the realm of coding that is focused towards study.
[14]	Balance of low-level programming	Python has the ability to balance high level programming with low level language. This is one of the characteristics of python. This makes it possible to work with entire array than a single number.
[15]	Portable and Flexible	python being flexible offers an option for developers to choose either to use OOPs approach and perform scripting, therefore, is right for developer for many purposes.

Bibliotecas do Python

Table 2. Main python libraries for data.

S. No.	Library	Purpose
1	Pandas	It is free and open source for working with and analyzing Python data. Panda is used for working with data sets. It has attributes for analyzing, cleaning, exploring and manipulating data. We can read and load any type of data using pandas [16].
2	NumPy	Python is not made for mathematical operation but due to interest of scientists and research community python developers were forced to develop packages of high level. For numerical operation and calculation python list data structure is not enough. So, NumPy is developed as core data structure called Ndaarry (Ndarry is usually fixed size multidimensional container of items) [16].
3	Scrapy	Scrapy is well known python library for data science. Python framework for web scraping is known as scrapy. It provides packages for developers without worrying about maintaining code. Applications use in makes spider bots that crawl the web and collect data. The interface is designed with “don’t repeat” mind set. Scrapy is also used to collect data from APIs [2].
4	Tensor Flow	The first in the list of python library for data science is TensorFlow which is used across various scientific fields. Features for better computation reduce errors by 60% in machine learning and provide quick updates of new features [17].

Bibliotecas do Python

Table 3. Additional library for data-science.

S. No.	Library	Purpose
1	Matplotlib	Python's Matplotlib generates static, animated, and interactive visualizations. Line, scatter, bar, histogram, heatmap, and others are plotted. Data scientists may create appealing graphs with Matplotlib [3]
2	Seaborn	Matplotlib-based Seaborn visualizes statistical data. It allows for more sophisticated statistical graphics. With a few lines of code, Seaborn makes distribution, regression, and categorical charts [18]
3	SciPy	This scientific computing package extends NumPy. Numerical optimization, integration, interpolation, linear algebra, signal processing, and more are available. Data science initiatives employ SciPy for advanced mathematical and scientific computations [18].
4	Scikit-learn	A popular Python machine learning library. It includes classification, regression, clustering, dimensionality reduction, and model selection tools and techniques. Scikit-learn contains data preprocessing, model evaluation, and cross-validation utilities [17].

Bibliotecas do Python

4	Keras	TensorFlow's high-level neural networks library. Its API makes deep learning model creation and training easy. Keras supports CNNs, RNNs, and transformers for quick prototyping [3].
6	PyTorch	Another prominent open-source deep learning package is PyTorch. It emphasizes simplicity and dynamic computation graphs. PyTorch supports GPU and CPU tensor calculations and has several tools for creating and training neural networks [3].
7	NLTK	NLTK is an NLP library. It offers tokenization, stemming, part-of-speech tagging, sentiment analysis, and topic modelling libraries and corpora. Data science initiatives employ NLTK for text mining and analysis [17].

Aplicações Diversificadas

- Bioinformática e Genômica
- Neurociência
- Física e Astronomia
- Ciência do Clima
- Análise de Redes Sociais

PUC Minas

Conclusões

- Versatilidade: Python é utilizado em diversas áreas.
- Python é essencial para a análise e visualização de dados
- Diversas bibliotecas poderosas para manipulação de dados
- Facilita a criação de aplicações e pesquisas avançadas

PUC Minas

Bibliografia

Ranjan, Mritunjay & Barot, Krishna & Khairnar, Vaishnavi & Rawal, Vaishnavi & Pimpalgaonkar, Anujaa & Saxena, Shilpi & Sattar, Arif. (2023). Python: Empowering Data Science Applications and Research. Journal of Operating Systems Development & Trends. 10. 27-33. 10.37591/joosdt.v10i1.576.

PUC Minas