Herramientas para ciencia de datos en lenguaje Python

Los científicos de datos necesitan lidiar con cuatro pasos principales: recolección y limpieza de datos, exploración de datos, modelado de datos y visualización de datos.

Recopilación de datos y limpieza

Se puede trabajar con casi todo tipo de datos en diferentes formatos, como CSV, TSV o JSON. Se puede importar tablas SQL directamente a su código, Python lo ayuda a realizar estas tareas fácilmente con sus bibliotecas dedicadas como **PyMySQL** y **BeautifulSoup**, respectivamente.

Exploración de datos

Se exploran los datos para identificar sus propiedades y segregarlos en diferentes tipos, como numéricos , ordinal, nominal, categórica, etc. Luego se emplea **NumPy** y **Pandas**. Estas bibliotecas ayudan a liberar información de los datos al permitirle manipularlos de manera fácil y eficiente.

Modelado de datos

Esta es crucial en el proceso de ciencia de datos, donde se minimiza la dimensionalidad del conjunto de datos. Para un análisis de modelado numérico de sus datos se puede usar **Numpy**. Con **SciPy** puede realizar cálculos. Una vez finalizado el modelado de datos, necesitaría visualizar e interpretar datos para obtener información procesable

Visualización e interpretación de datos

Python tiene muchos paquetes de visualización de datos. **Matplotlib** es la biblioteca más utilizada entre ellas para generar gráficos y cuadros básicos. Para gráficos avanzados se puede probar otra biblioteca de Python, Plotly .

Otra biblioteca de Python, IPython lo ayuda con la visualización interactiva de datos y admite el uso de un kit de herramientas GUI.