Pandas 1.x en español

Kevin Farinango (Asuskf)

2020-09-08

Contents

Chapter 1

Introducción

Pandas es una librería usanda en la creación y manipulación de datos con Python. Es empleada para manipular información tabular (filas y columnas) cómo encontrarías en una base de datos o en una hoja de cálculo, es aprovechado por **Científicos de datos, analistas, programadores, ingenieros** que lo usan en el modelado de datos.

Pandas se limita a **"datos pequeños"** (datos que pueden caber en la memoria de una sola máquina). Sin embargo, la sintaxis y las operaciones se han adoptado o inspirado a otros proyectos: PySpark, Dask, Modin, cuDF, Baloo, Dexplo, Tabel, StaticFrame, etc.

1.1 Para quién es esta guia

Esta guía fue pensada en principiantes de habla hispana que estén empezando en el mundo de la ciencia de datos. El presente libro tiene la intención de dotar al principiante de los conocimientos necesarios para manipular y analizar datos con una herramienta potente y de fácil uso como es Pandas.

La presente guía tiene más de 50 ejercicios que van de básico a intermedio con los cuales el lector podrá poner en práctica e interiorizar lo aprendido. Para una mejor comprensión cada ejercicio se detalla pasos a paso.

1.2 Qué cubre esta guía

- **Capítulo 1. Introducción:** Descripción de lo que tiene el libro y para quien va dirigido.
- **Capítulo 2. Fundamentos de Pandas:** Presenta la anatomía y el vocabulario utilizado para identificar las dos estructuras de datos principales de Pandas, la Serie y el DataFrame con sus respectivos métodos.

- **Capítulo 3. Funciones esenciales en un Dataframe:** Se centra en las operaciones comunes que se realizan durante el análisis de datos.
- **Capítulo 4. Importación y creación de Dataframes:** Analiza las diversas formas de importar datos y crea DataFrames.
- **Capítulo 5. Empezando con la ciencia de datos:** Presenta técnicas de análisis básicas que sirven para comparar números y datos categóricos.
- **Capítulo 6. Trabajando con subconjuntos:** Cubre las muchas formas variadas y potencialmente confusas de seleccionar diferentes subconjuntos de datos.
- **Capítulo 7. Filtrando los subconjuntos:** Explica varias formas de cómo seleccionar diferentes subconjuntos de datos.
- **Capítulo 8. Consultas booleanas**
- **Capítulo 9. el uso de index en Pandas:** El uso correcto de index para evitar resultados erróneos.
- **Capítulo 10. Agrupación:** Presenta métodos de agrupación y la construcción de funciones personalizadas para aplicar a diferentes grupos
- **Capítulo 11. Ordenando el Dataframe:** Explica qué son los datos ordenados, por qué son tan importantes y que métodos pueden ser utilizados
- **Capítulo 12. Combinando objetos Pandas:** Presenta varios métodos para combinar DataFrames y Series vertical u horizontalmente.
- **Capítulo 13. Trabajando con las series de tiempo:** Presenta distintos métodos para trabajar con las series temporales en cualquier dimensión de tiempo posible
- **Capítulo 14. Visualización de datos:** Tips para crear visualizaciones estéticamente agradables usando: Pandas, Seaborn y Matplotlib
- **Capítulo 15. Test del Dataframe:**Explora los mecanismos de test de los DataFrames y los resultados que presenta Pandas.

Chapter 2

Fundamentos de Pandas

El objetivo de este capítulo es dar una introducción detallada de los fundamentos de Pandas y la diferencia entre **Series y Dataframes**. Esta librería es útil para trabajar con datos estructurados.

¿Pero qué son los datos estructurados?

Los datos estructurados son datos que se almacenan en tablas (filas y columnas) como pueden ser archivos CSV, XLSX (hojas de cálculos de Excel) o la tabla de una base de datos.

En este capítulo aprenderás a seleccionar una sola columna de datos de un Dataframe (conjunto de datos bidimensionales) que se devuelve como un conjunto de datos unidimensional también conocidos como series. Trabajar con este objeto unidimensional facilita cómo funcionan los diferentes métodos y operaciones. Varios métodos de Serie retornan una serie como salida. Esto conduce a la posibilidad de llamar a más métodos lo que se conoce como **encadenamiento de métodos** (ejecutar un método tras otro).

El index del Dataframe y de las series es lo que separa a Pandas de otras librerías de análisis de datos. El index es un poderoso objeto que se emplea como una etiqueta para los valores de serie. Entender este objeto es clave para comprender cómo funcionan los métodos.

Antes de empezar la presenta guía presentaremos a Colab un entorno de desarrollo que no necesita instalación de ningún tipo. En este caso usaremos Colab

2.1 Qué es Colab

Colaboratory, o Colab es un entorno de desarrollo que google a puesto para el uso del público en general. Este entorno es basado en jupyter notebook el cual

permite escribir y ejecutar código de Python en un navegador: Sin configuración requerida, con acceso gratuito a GPU/TPU y con gran facilidad para compartir

2.1.1 Qué es la GPU

La GPU es una unidad de procesamiento de Gráficos que tienen la capacidad de hacer varias tareas de forma simultánea, ya que las imágenes en HD no dejan de ser la unión de matrices de datos (Álvaro Gonzalo, 2019). La CPU realiza cálculos de forma secuencial, es decir, uno a la vez. Es posible incrementar núcleos adicionales para que pueda llevar a cabo dos, cuatro, ocho, dieciséis o más operaciones a la vez. Pero estos núcleos adicionales aumentan el precio de forma exponencial.

2.1.2 Qué es la TPU

Una unidad de procesamiento **tensorial** o TPU (del inglés tensor processing unit) es un circuito integrado de aplicación específica y acelerador de IA desarrollado por Google para el aprendizaje automático con redes neuronales artificiales y más específicamente optimizado para usar TensorFlow (Silva Selva, 2020). La TPU es entre 25 y 80 veces más eficiente para la realización de cálculos que las unidades de procesamiento normal (CPU y GPU).

2.1.3 Video presentación Colab

¿Qué es Colab? parte 1/2 (Uso de recursos)

¿Qué es Colab? parte 2/2 (Presentación de Colab)

Bibliography

Silva Selva (2020). Unidad de procesamiento tensorial.

Álvaro Gonzalo (2019). Comparando cpus y gpus para inteligencia artificial.