Iteración eficiente sobre filas en Pandas

Andrea Manuel Simón y Ana López Palomo

November 2023

1 Reflexión y Contextualización

Tras reflexionar hemos llegado a la conclusión de que después de 3 años analizando bases de datos y datasets, con distintas perspectivas, el manejo de grandes datos resulta un gran problema a la hora de su ejecución. Además, trabajar con muchos datos resulta tedioso tanto de entender, como de limpiar y de realizar.

Estos problemas pueden verse reflejados en la parte computacional, por ejemplo: el ordenador puede recalentarse, bloquearse, y hasta puede reiniciarse tras un pantallazo azul.

2 Investigación y comparación de técnicas

Hay varias maneras de iterar filas en pandas:

1. Iterar usando iterrows(): es un método fácil de entender y usar, aunque puede ser lento en dataframes grandes debido a la sobrecarga de crear objetos de tipo serie para cada fila. Este método itera sobre el dataframe como pares (indice, fila) y permite acceder a los valores de cada celda.

2. Vectorización: mucho más rápido en dataframes grandes debido a la optimización interna de Pandas. Sin embargo, no siempre es posible aplicarlas, especialmente para operaciones personalizadas complejas. Esta técnica no itera, sino que aprovecha al máximo las operaciones vectorizadas. Una manera de aplicar la vectorización en pandas es con la función apply().

3. Cython o Numba: es una técnica que se usa si necesitas bucles más rápidos ya que puede mejorar significativamente la velocidad del procesamiento.Sin embargo, requiere un conocimiento más avanzado y es más complicada de implementar.

4. Iteración paralela: esta técnica puede acelerar en gran medida el procesamiento. Aprovecha al máximo la potencia de la CPU en maquinas multiproceso, aunque es más complejo y requiere un hardware adecuado.

A la hora de implementar estas técnicas, hay que tener en cuenta las necesidades de nuestro proyecto, así como las limitaciones del hardware. De cualquier modo, siempre hay que intentar hacer el menor número de iteraciones posibles.

Con nuestro conocimiento, lo único que usabamos era el iterrows(). Sin embargo no tuvimos en cuenta el rendimiento del programa, utilizando lo que nos fuera más sencillo de aplicar.

De todas las técnicas mencionadas anteriormente, la más eficiente es la vectorización. Es un enfoque que se basa en realizar operaciones en bloques de datos en lugar de iterar sobre elementos individuales. Utiliza operaciones vectorizadas proporcionadas por Pandas y NumPy, para procesar datos de manera eficiente sin necesidad de bucles explícitos. Es decir, manipula columnas completas en lugar de datos individuales.

Por otra parte, aprovecha el broadcasting. Es un concepto que se usa en NumPy que permite realizar operaciones entre elementos de diferentes formas (dimensiones). Por ejemplo, podemos sumar un número a una columna sin necesidad de un bucle.

3 Aplicación de técnicas y reflexión

Tras programar el método apply y el iterrowrs, hemos podido ver que es más eficiente el primero. Esto cambia totalmente la manera en la que podemos programar en el futuro, ya que en vez de hacer cambios uno por uno a base de bucles, los realizamos todos a la vez, siendo una manera más eficiente.

Aunque en el ejemplo es una operación sencilla y no se nota mucho la diferencia, en códigos más extensos y complejos esto puede cambiar totalmente el rumbo de la investigación.

En resumen, es muy importante entender y saber las funciones y los métodos que nos proporcionan. Así, podremos evaluar cuál es mejor en cada caso y poder elegir el mejor procedimiento entre todas las opciones, en vez de hacer la única manera que se sabe.