"El saber de mis hijos
hará mi grandeza"

IMDB, ANALIZANDO LA BASE DE DATOS

COMPARACIÓN DE PERSPECTIVAS
MAESTRÍA EN CIENCIA DE DATOS

1 Introducción

Se realizó un análisis de las películas mejor y peor calificadas en promedio haciendo uso de la base de datos IMDb, utilizando dos enfoques distintos: línea de comandos en Linux y Python en Windows. A continuación se da un análisis comparativo de las ventajas y desventajas de cada perspectiva.

2 Usando la Línea de Comandos (Linux)

- **Descripción de los pasos:** Para resolver el problema mediante la línea de comandos, considerando que ya hemos descargado la base de datos previamente, se hace uso de los siguientes comandos en el orden que sigue:
  - Uso de **for** para extraer el identificador y la calificación de los archivos.
  - Uso de **sort** para ordenar las calificaciones.
  - Uso de **awk** para combinar las calificaciones con las URLs y calcular los promedios.
  - Uso de **sed** para limpiar las URLs.
  - Uso de **head** para extraer las 10 mejores o peores películas.
- **Ventajas:**
  - El script es mucho más sencillo y directo.
  - No se necesitan instalar librerías adicionales, lo que hace que sea menos tedioso.
  - A mi parecer, se manejan de mejor manera los archivos de texto.
- **Desventajas:**
  - Es más complicado si los datos cambian o si necesitas usarlos para otros fines.
  - Puede ser difícil de entender para quienes no estén acostumbrados a trabajar con la línea de comandos.

3 Usando Python (Windows)

- **Descripción de los pasos:** Para resolver el problema mediante python se creó el proyecto con peotry y se instalaron las paqueterías correspondientes, siguiendo los pasos:
  - Uso de módulos personalizados como **IMDBAnalysis**, **downloader** y el script **test_script**
  - Automatización de la descarga y procesamiento de los archivos.
- **Ventajas:**
  - Permite hacer un análisis más detallado y reutilizar los datos más fácilmente.
  - El código es un poco más comprensible, especialmente para quienes están familiarizados con Python.
  - Gracias a beautilsoup4 podemos obtener el nombre de las películas, cosa que no podíamos lograr con la línea de comandos.
  - Se puede correr en colab, mediante el scrip, jupyter.
- **Desventajas:**

- Necesitamos instalar algunas dependencias o librerías adicionales..
- Puede tardar más tiempo en desarrollarse en comparación con usar comandos simples en Linux.
- El código es mucho más largo.

4 Conclusiones

Ambos enfoques, el uso de la línea de comandos en Linux y la implementación con Python en Windows, tienen sus propias ventajas y desventajas que los hacen útiles en diferentes contextos.

Desde mi perspectiva, trabajar con la línea de comandos resulta más cómodo y eficiente para tareas específicas como este análisis de películas. Las principales ventajas radican en su simplicidad y el hecho de que no es necesario gestionar múltiples dependencias ni preocuparse por la configuración del entorno. La línea de comandos ofrece un enfoque directo y rápido. Además, en términos de configuración inicial, el proceso es menos propenso a fallos relacionados con la instalación de librerías o problemas de compatibilidad.

En cuanto al tiempo de ejecución, ambos métodos toman un tiempo similar para procesar los datos, ya que en ambos casos estamos trabajando con archivos relativamente grandes y realizando cálculos similares. Sin embargo, se invierte menos tiempo en la preparación del entorno y ejecución de los comandos en Linux, lo que lo hace más eficiente desde una perspectiva de desarrollo inmediato. No obstante, es importante mencionar que la línea de comandos puede volverse limitada cuando se necesita realizar tareas más complejas, personalizaciones (como por ejemplo, obtener el resultado de manera más bonita) o reutilizar el código en otros proyectos o plataformas.

Por otro lado, Python ofrece una gran ventaja en cuanto a la portabilidad y la posibilidad de reproducir el análisis en diferentes plataformas, como Windows, Linux o incluso en servidores remotos. Al usar Python, el proceso es más estructurado y escalable, lo que facilita la automatización y la reutilización del código en otros contextos. Además, Python cuenta con una gran cantidad de librerías especializadas para el análisis de datos que podrían facilitar la implementación de tareas más complejas o generar informes más detallados en el futuro. Este enfoque es más adecuado para proyectos que requieren mantenimiento a largo plazo, colaboración entre equipos o la integración con otros sistemas de análisis.