Memoria EDA: Tiempo en pantalla de los personajes del MCU

Una vez definido el tema de estudio, el primer paso consistió en obtener los datos necesarios. Después de explorar varios datasets en Kaggle y GitHub, se identificó que todos ellos obtenían la información de una lista de IMDb creada por ninewheels0.

Con la información localizada, la siguiente tarea fue determinar la mejor forma de adquirirla. Inicialmente, se consideró el uso de una API de IMDb; sin embargo, al no encontrar una manera de obtener datos de listas creadas por usuarios, se descartó esta opción. La alternativa más viable resultó ser el web scraping.

Tras revisar la estructura de la página web, se determinó que era factible obtener la información deseada a partir del HTML de la página. Se optó por utilizar BeautifulSoup para este propósito. Una vez obtenido el HTML, fue posible extraer la información sobre el tiempo en pantalla de cada personaje, la cual estaba contenida en las etiquetas <p>.

La información estaba organizada en párrafos separados por cambios de línea, con el formato "personaje <tiempo>". Durante este proceso, se registró también la información de los nombres de las películas y series, así como la duración de cada una. Se implementó un bucle para limpiar los datos, y se excluyeron algunos títulos no relevantes, como documentales y la serie animada "What If...", ya que no es de acción real.

Tras revisar los datos, se identificaron casos en los que a un mismo personaje se le refería con diferentes nombres según la película/serie. Para abordar esto, se creó un CSV con los diferentes nombres asociados a un único personaje, asignándoles una ID única. Luego, se combinaron las filas correspondientes a un mismo personaje en una única fila, consolidando así los datos.

Con los datos limpios, se desarrollaron funciones para representar diversas gráficas. Estas funciones fueron adaptadas para permitir filtros flexibles. Las gráficas generadas incluyen:

Una gráfica con los x personajes con más horas en pantalla en total.
Una gráfica con los x personajes con más apariciones en títulos diferentes.
Una gráfica con los títulos diferentes y las horas en pantalla del personaje seleccionado.
Un gráfico en forma de tarta que representa el tiempo en pantalla de cada personaje en el título seleccionado.
Todas las gráficas pueden personalizarse para incluir solo películas o películas y series, así como para seleccionar combinaciones específicas de las cuatro fases del MCU.

Estas gráficas, acompañadas de imágenes y formato de texto en Markdown, se integran en un código que utiliza Streamlit para facilitar la presentación.