1. **Extracción de URLs**

El proceso comienza definiendo un conjunto de páginas específicas del sitio web de **Metacritic** que se quieren visitar. Estas páginas contienen listados de películas. Para cada una de ellas, se establece un número máximo de películas que se desean extraer, en este caso **24 por página**. A fin de evitar que el sitio bloquee el acceso automático, se configura una cabecera HTTP personalizada que simula ser un navegador real —usando un **User-Agent común como el de Google Chrome**.

Una vez configurada esta base, el script accede a cada una de las páginas listadas. Para cada una, **realiza una solicitud HTTP** con los parámetros apropiados.

Cuando se localizan estos enlaces, se extraen y se convierten en URLs completas, es decir, listas para usarse en accesos posteriores. **Solo se conservan los enlaces correspondientes a las primeras 24 películas de cada página, y todos estos enlaces se acumulan en una lista maestra**. A partir de esta, se construyen nuevas URLs que apuntan directamente a las secciones de reseñas de cada película. Estas URLs se generan de manera que incluyan filtros específicos:

Uno para reseñas mixtas ("Mixed Reviews") y otro para reseñas negativas ("Negative Reviews"). Esto se hace con el objetivo de focalizar el **scraping** en opiniones que no son totalmente positivas.

Con estas URLs ya generadas, se procede a la etapa más compleja: **la extracción de reseñas utilizando Selenium**, una herramienta que permite automatizar navegadores reales. Selenium es necesario aquí porque muchas de las reseñas no están disponibles directamente en el HTML original, sino que se cargan dinámicamente mediante JavaScript.

El navegador automatizado accede a cada una de estas URLs y espera unos segundos a que el contenido se cargue completamente. Una vez que la página está lista, se analiza el HTML renderizado en busca de los elementos que contienen las reseñas. Para cada reseña encontrada, se extrae información clave: el nombre del usuario que la escribió, la fecha de publicación, la puntuación que le dio a la película y el texto completo de la reseña.

Estas reseñas se acumulan temporalmente en una lista, y el proceso se repite para cada página de reseñas (aunque se impone un límite de 50 reseñas por película para no saturar el sistema). Cuando se completa la recolección, todas las reseñas obtenidas se combinan en una estructura tabular (como un DataFrame de Pandas), donde cada fila representa una reseña distinta con todos sus datos asociados.

Finalmente, esta tabla se puede visualizar para verificar que el scraping se ha ejecutado correctamente y que los datos están listos para ser utilizados en análisis posteriores, como clasificación de sentimiento, evaluación de polaridad, detección de incoherencias o estudios sobre lenguaje de reseñas.

4º.

1. **Preprocesamiento**

**Carga y consolidación de datos**

El procedimiento inicia con la lectura de cuatro archivos CSV que contienen reseñas de usuarios. Estos archivos son importados como DataFrames individuales (df1, df2, df3 y df4). Todos estos conjuntos de datos se concatenan en un único DataFrame unificado.

Para verificar la limpieza inicial, se detectan filas duplicadas y se imprimen para revisión.

**Extracción y transformación de información de URLs**

Se extrae el nombre de la película desde el enlace usando expresiones regulares, transformando el formato URL en un nombre legible y capitalizado.

**Filtrado y detección de idioma**

Se eliminan las filas en las que la reseña está vacía o contiene la frase "SPOILER ALERT", ya que estas no aportan valor al análisis de sentimiento. También se aplica una función personalizada que detecta el idioma de cada reseña utilizando un detector de lenguas. Este paso considera excepciones, como textos vacíos o no identificables. Tras identificar los idiomas, se filtra el DataFrame para conservar únicamente aquellas reseñas escritas en inglés.

**Asignación de etiquetas de sentimiento**

Se introduce una nueva columna llamada sentiment, derivada de la puntuación (score) otorgada por el usuario. Se utiliza un criterio simple: reseñas con puntuaciones de 0 a 3 se clasifican como negativas, de 4 a 6 como neutrales y de 7 a 10 como positivas.

Este el mismo criterio que sigue la página web, por lo que nos servirá para evaluar modelos más adelante.

**Análisis de "stopwords" y visualización**

Se importa un conjunto de palabras vacías en inglés (stopwords) desde la librería spaCy. Luego se define una función que calcula el porcentaje de palabras vacías dentro de cada reseña.

A continuación, se define otra función que genera un gráfico de barras que compara el porcentaje de palabras vacías y palabras con contenido real según la categoría de sentimiento. Esto permite observar, por ejemplo, si las reseñas negativas tienden a usar más palabras vacías que las positivas, o viceversa.

**Normalización y lematización de texto**

Se establece una función para limpiar a fondo cada reseña. Este proceso incluye:

* Eliminación de números.
* Tokenización y lematización usando spaCy.
* Eliminación de puntuación, espacios innecesarios y stopwords.
* Conversión a minúsculas.
* Reconstrucción del texto como una única cadena con las palabras procesadas.

Esta función se aplica directamente sobre la columna review, reemplazando el texto original con su versión limpia y estructurada. Este paso es fundamental antes de aplicar modelos de análisis de texto, ya que estandariza el contenido y reduce ruido.

1. **Exploratorio**

**Tokenización con spaCy**

Esta función aplica nlp (el modelo de spaCy) sobre cada reseña, creando una nueva columna llamada tokens que contiene los objetos Token generados por spaCy. Estos tokens son los elementos básicos para realizar análisis posteriores.

**Análisis de longitud de textos**

Se implementa una función para analizar la cantidad de palabras por reseña:

A cada texto se le calcula su longitud, es decir, cuántos tokens tiene. Luego, se generan histogramas para visualizar la distribución de longitudes por cada clase de sentimiento (NEG, NEU, POS).

Se muestra también una distribución global que agrupa todas las clases juntas.

En cada gráfico se incluye la media de longitud de reseñas para facilitar la interpretación.

**Nubes de palabras (WordClouds)**

Otra técnica de visualización textual utilizada es la nube de palabras. Para hacerla más informativa, se eliminan primero las palabras más frecuentes del conjunto general (las top\_n) para resaltar términos más distintivos por clase.

Se construye un listado de todas las palabras en el corpus completo y se cuentan las frecuencias y se eliminan las más comunes.

Para cada clase de sentimiento, se agrupan las palabras restantes y se genera una nube de palabras individual que representa visualmente la importancia relativa de los términos (palabras más frecuentes aparecen más grandes).

**Generación y visualización de n-gramas (bigramas y trigramas)**

Se analizan secuencias frecuentes de palabras consecutivas:

Bigramas: Secuencias de 2 palabras (como “very good” o “not bad”).

Trigramas: Secuencias de 3 palabras (como “i really liked” o “not worth watching”).

El procedimiento incluye:

Extraer todos los tokens por clase de sentimiento.

Calcular la frecuencia de aparición de cada bigrama y trigramas.

Seleccionar los 10 más comunes por clase.

Visualizarlos en gráficos de barras horizontales, uno por clase, para facilitar la comparación.

1. **Extracción de características**