

Proyecto de Scraping de Películas desde FilmAffinity

Extracción, Almacenamiento y Análisis de Datos con Java

Contenido

Introducción: Objetivos, alcance, justificación y utilidad del proyecto 2

Metodología 3

 Extracción de Datos 3

 Almacenamiento en Base de Datos 4

Análisis de Datos 5

Conclusiones..... 5

Mejoras posibles 10

Código Fuente..... 11

Introducción: Objetivos, alcance, justificación y utilidad del proyecto

Este proyecto tiene como objetivo aplicar los conocimientos adquiridos sobre automatización web utilizando Selenium para capturar datos dinámicos de una página web (en este caso, FilmAffinity). Posteriormente, los datos extraídos serán almacenados en una base de datos MySQL y se realizará un análisis básico utilizando Excel. El propósito es practicar la extracción de datos (scraping), la gestión de bases de datos y la creación de informes analíticos.

Se ha elegido FilmAffinity como el sitio web para este proyecto debido a que es una de las plataformas más populares para obtener información sobre películas. La extracción de datos como el título, año de lanzamiento, puntuación, votos, duración y género de las películas permitirá realizar análisis interesantes sobre tendencias y preferencias de los usuarios. Además, FilmAffinity tiene datos actualizados que permiten realizar comparativas entre películas en tiempo real.

Este proyecto es útil para aprender cómo interactuar con sitios web dinámicos y cómo almacenar y manipular grandes volúmenes de datos utilizando Java. Es especialmente relevante para desarrollar habilidades en la creación de sistemas automatizados de recopilación de datos y en la realización de análisis a partir de esos datos.

The screenshot shows the FilmAffinity website interface. At the top, there is a search bar and navigation links for 'Iniciar sesión' and 'Registrarse'. Below the search bar, there are buttons for 'Limpiar Filtro' and 'Aplicar Filtro'. The main content area displays a list of movies and TV series ranked by votes. The list includes the following items:

- 1. El padrino** (1972) by Francis Ford Coppola. Cast: Marlon Brando, Al Pacino, James Caan, Robert Duvall, Diane Keaton. Rating: 9.0, 177,062 votes.
- 2. Planeta Tierra II** (2016) by Elizabeth White, Justin Anderson, Ed Charles. Genre: Documental. Rating: 8.9, 4,577 votes.
- 3. El padrino. Parte II** (1974) by Francis Ford Coppola. Cast: Al Pacino, Robert De Niro, Diane Keaton, Robert Duvall, John Cazale. Rating: 8.9, 141,461 votes.
- 4. The Wire (Bajo escucha)** (2002) by David Simon (Creador), Joe Chappelle. Rating: 8.8.

The left sidebar contains navigation links for various categories, including 'Netflix (próximo)', 'Amazon Prime', 'Movistar Plus+', 'HBO Max', 'Disney+', 'Filmin', 'Apple TV+', 'SkyShowtime', 'Rankings FA', 'Secciones', 'Series de TV', and 'TOPs'.

Metodología

Extracción de Datos

Para la automatización de la extracción de datos, se utilizó la librería Selenium WebDriver en Java. El proceso de extracción implicó la creación de un código en java que accediera a la página de películas más populares de FilmAffinity ([link](#)), hacer clic en el botón "Ver más resultados" hasta cargar más de 3600 películas, y luego entrar en la página principal de cada una de ellas para extraer la información relevante.

Los datos extraídos de cada película incluyen:

- **Título original**
- **Año de estreno**
- **Puntuación**
- **Número de votos**
- **País de origen**
- **Duración**
- **Género**

El flujo general consistió en abrir cada película en una nueva pestaña, extraer los datos y luego almacenarlos en una base de datos MySQL.

El padrino

Ficha | Créditos | Críticas [688] | Tráilers [9] | Imágenes [78] | Blu-ray [6] | *Films con Oscar™ a ...*

Título original The Godfather
Año 1972
Duración 175 min.
País Estados Unidos
Dirección Francis Ford Coppola
Guion Francis Ford Coppola, Mario Puzo. Novela: Mario Puzo
Reparto
Marlon Brando | Al Pacino | James Caan | Robert Duvall | Diane Keaton | John C.
Música Nino Rota
Fotografía Gordon Willis
Compañías Paramount Pictures, Alfran Productions. Productor: Albert S. Ruddy
Género Drama | Mafia. Crimen. Años 40. Años 50. Familia. Película de culto
Grupos Trilogía El Padrino | Adaptaciones de Mario Puzo
Sinopsis América, años 40. Don Vito Corleone (Marlon Brando) es el respetado y temido jefe de una de las cinco familias de la mafia de Nueva York. Tiene cuatro hijos: Corleone...

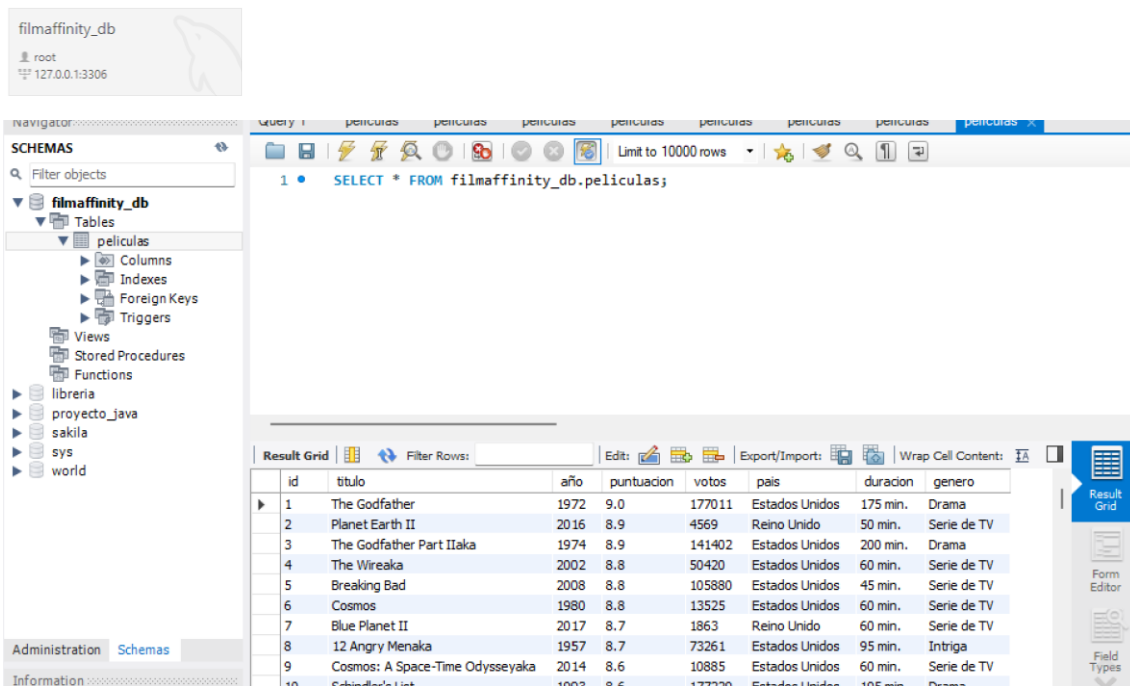
9,0 177.062 votos
688 críticas - títulos

Almacenamiento en Base de Datos

Se creó una base de datos en MySQL llamada `filmaffinity_db` con una tabla `peliculas`, que tiene las siguientes columnas:

- `titulo` (VARCHAR)
- `año` (INT)
- `puntuacion` (DECIMAL)
- `votos` (INT)
- `pais` (VARCHAR)
- `duracion` (VARCHAR)
- `genero` (VARCHAR)

Los datos se insertaron utilizando sentencias SQL `INSERT INTO`, y antes de insertarlos, se verificó si la película ya existía en la base de datos para evitar duplicados.



The screenshot shows a MySQL database interface. On the left, the 'Schemas' panel displays the 'filmaffinity_db' database. The 'Tables' section under 'filmaffinity_db' shows the 'peliculas' table. The main window displays the 'Result Grid' for the query `SELECT * FROM filmaffinity_db.peliculas;`. The grid shows 10 rows of movie data.

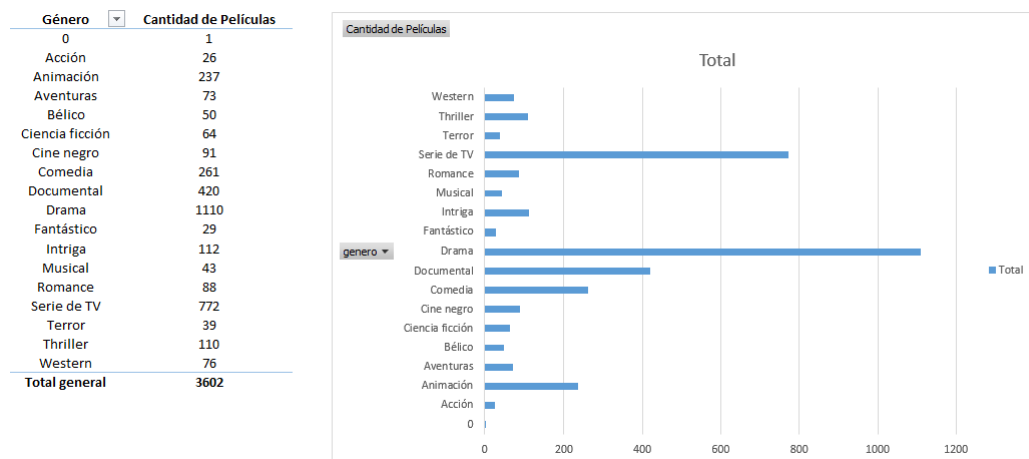
	id	titulo	año	puntuacion	votos	pais	duracion	genero
▶	1	The Godfather	1972	9.0	177011	Estados Unidos	175 min.	Drama
	2	Planet Earth II	2016	8.9	4569	Reino Unido	50 min.	Serie de TV
	3	The Godfather Part II	1974	8.9	141402	Estados Unidos	200 min.	Drama
	4	The Wire	2002	8.8	50420	Estados Unidos	60 min.	Serie de TV
	5	Breaking Bad	2008	8.8	105880	Estados Unidos	45 min.	Serie de TV
	6	Cosmos	1980	8.8	13525	Estados Unidos	60 min.	Serie de TV
	7	Blue Planet II	2017	8.7	1863	Reino Unido	60 min.	Serie de TV
	8	12 Angry Men	1957	8.7	73261	Estados Unidos	95 min.	Intriga
	9	Cosmos: A Space-Time Odyssey	2014	8.6	10885	Estados Unidos	60 min.	Serie de TV
	10	Schindler's List	1993	8.6	177779	Estados Unidos	195 min.	Drama

Análisis de Datos

Una vez que los datos fueron almacenados en la base de datos, se exportó la base de datos en un formato compatible con Excel para realizar análisis sobre los datos.

De ahí se obtuvieron los siguientes resultados:

Películas por género



Los datos muestran que los géneros más representados en la colección son el drama (1110 películas), los documentales (420) y la animación (237), mientras que géneros como el terror (39) y el musical (43) tienen una menor presencia. El drama es, con diferencia, el género más predominante, representando casi un tercio del total de películas (3602). Además, el gráfico refleja visualmente esta distribución, destacando el dominio del drama y la serie de TV en comparación con otros géneros.

Puntuación promedio por país

Tras realizar una tabla con las respectivas puntuaciones promedio por país, además de considerar la cantidad de votos.

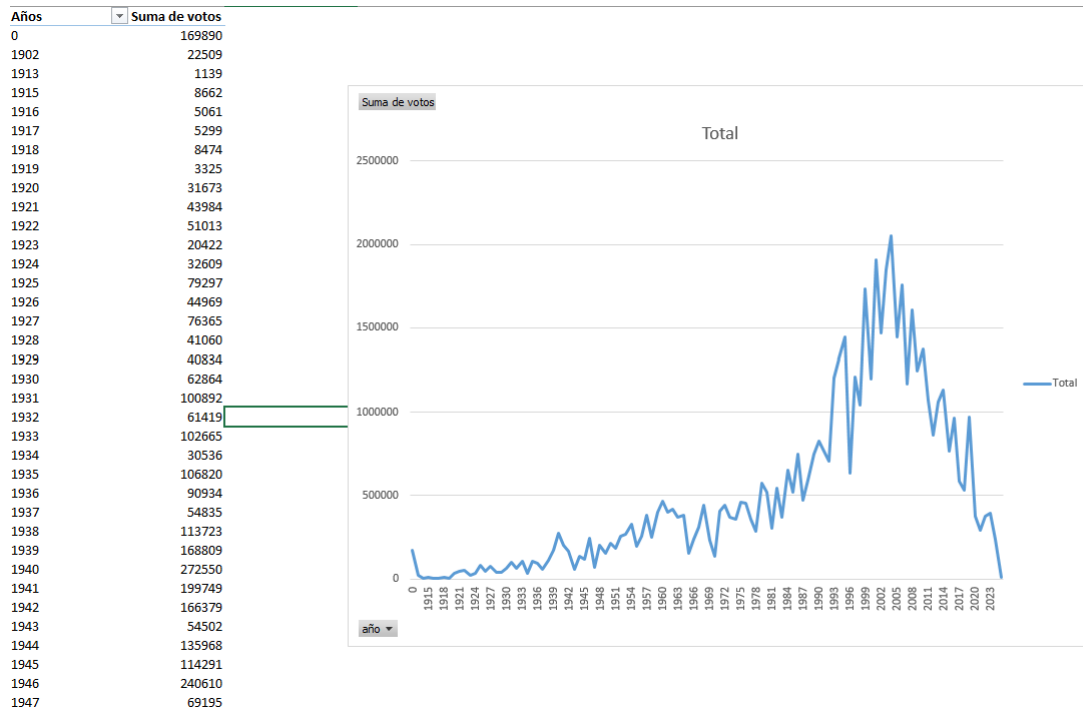
Al analizar estos datos, se observa que, si bien Islandia tiene el promedio de puntuación más alto (8,00), su número total de votos (2104) es bajo en comparación con países como Estados Unidos (más de 36 millones de votos), Reino Unido (5,4 millones) y Japón (1,9 millones). Esto sugiere que Islandia tiene una base de votantes más reducida pero altamente positiva.

Por otro lado, países con gran cantidad de producciones y reconocimiento, como Estados Unidos, Italia y Francia, tienen millones de votos, lo que refleja una mayor presencia en la industria cinematográfica y una audiencia más amplia. Sin embargo, sus puntuaciones promedio se mantienen en el rango de 7,4 a 7,5, lo que indica una distribución más diversa de valoraciones.

Finalmente, algunos países con puntuaciones relativamente altas, como Ucrania o Macedonia, tienen muy pocos votos, lo que podría indicar que su cine es apreciado por una audiencia de nicho en lugar de una popularidad masiva. En general, los datos reflejan que

una alta puntuación no siempre se correlaciona con un alto volumen de votos, lo que puede estar influenciado por factores como la cantidad de producciones, la accesibilidad y la distribución internacional.

Votos a lo largo de los años



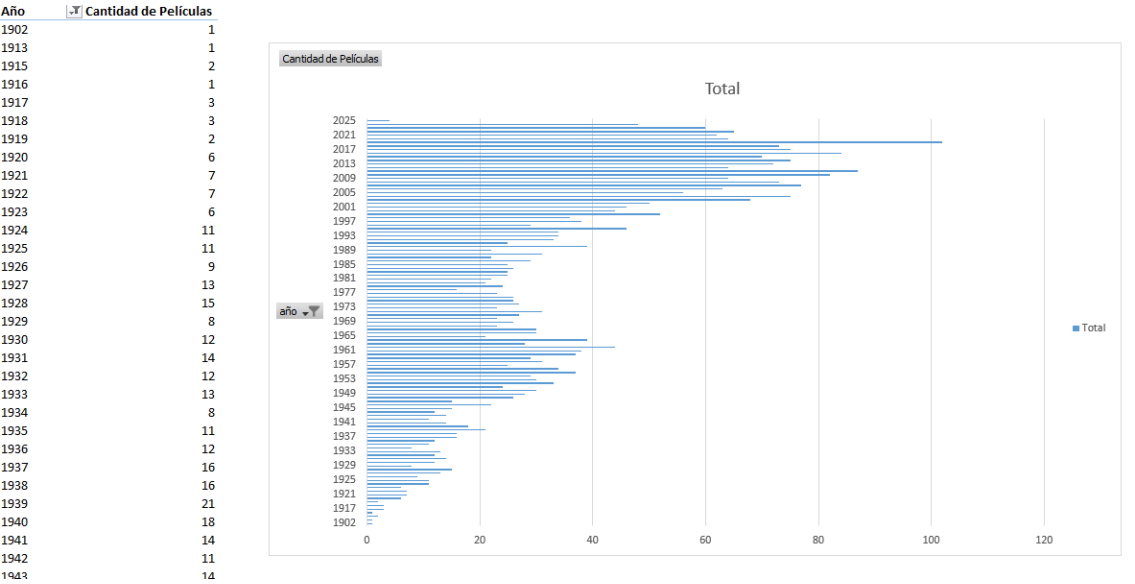
Estos datos muestran la evolución de la cantidad de votos recibidos por películas a lo largo de los años, reflejando tendencias en la popularidad y el acceso al cine. Se pueden destacar varios puntos clave:

1. Aumento progresivo de votos en el siglo XX: A partir de 1930, se observa un crecimiento sostenido en la cantidad de votos, con picos en décadas como los años 50 y 60, lo que sugiere una mayor producción y reconocimiento del cine clásico.
2. Explosión de popularidad en los años 90 y 2000: A partir de 1990, hay un notable incremento en los votos, alcanzando un pico en 1999 con más de 1,7 millones y en 2001 con más de 1,9 millones. Esto puede atribuirse a la digitalización del cine, la expansión del internet y el auge de plataformas de votación en línea.
3. Máxima popularidad en los años 2000 y 2010: Los años 2004-2010 muestran los mayores volúmenes de votos, coincidiendo con el auge de grandes producciones de Hollywood, el crecimiento del cine internacional y el acceso masivo a bases de datos de películas en línea.

4. Descenso en la cantidad de votos recientes (2017-2024): A partir de 2017, se observa una disminución en los votos, posiblemente debido a la fragmentación del consumo de entretenimiento, con el auge de las plataformas de streaming, el impacto de la pandemia en la industria cinematográfica y un menor interés en la votación de películas recientes.

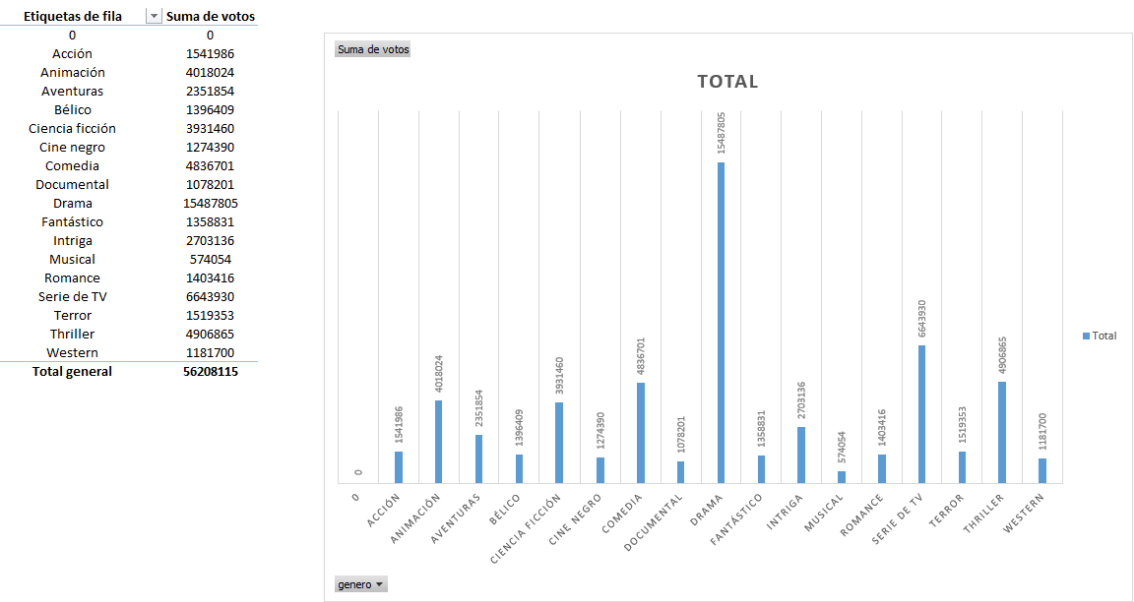
En general, los datos reflejan la evolución del cine y los cambios en los hábitos de consumo del público a lo largo de más de un siglo.

Películas por año



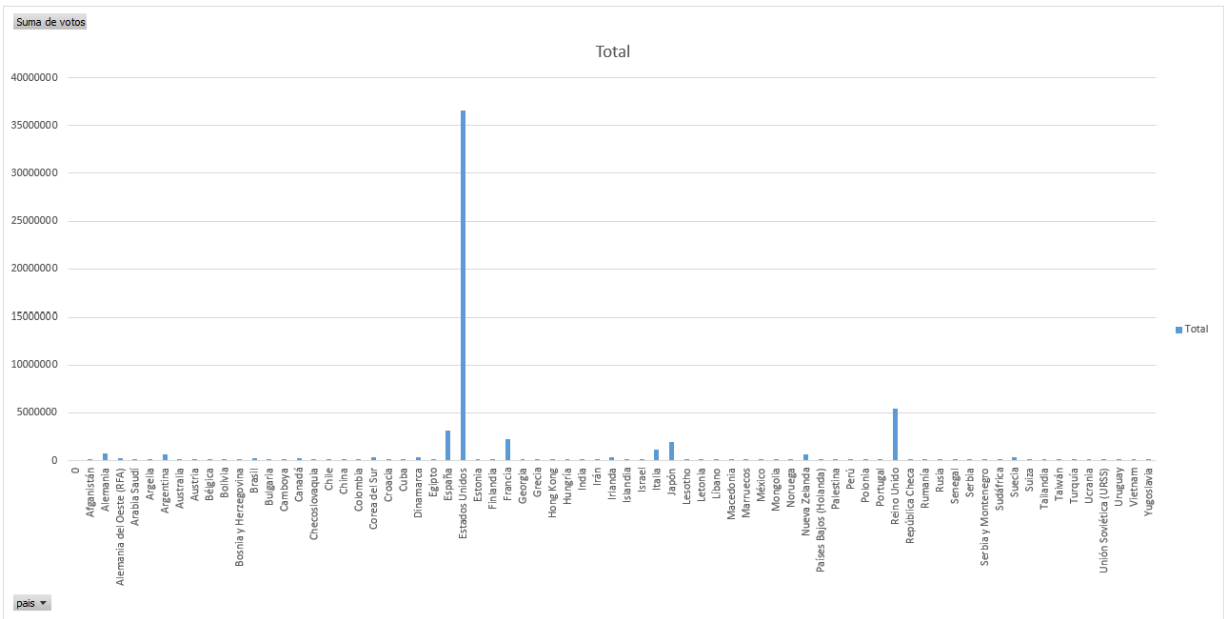
Los datos reflejan un crecimiento sostenido en la producción cinematográfica desde principios del siglo XX, con un aumento notable en los años 90 y 2000 debido a la digitalización y la globalización del cine. La cantidad de películas alcanzó su punto máximo en 2019 con 102 producciones, coincidiendo con el auge del streaming y la diversificación del mercado. Sin embargo, a partir de 2020, la producción disminuyó, posiblemente debido a la pandemia y los cambios en los modelos de distribución, priorizando las plataformas digitales sobre los estrenos en salas.

Votos por género



El drama es el género con mayor cantidad de votos (más de 15 millones), seguido por la serie de TV, el thriller y la comedia, lo que indica una fuerte preferencia por historias emocionales y narrativas extensas. La animación y la ciencia ficción también destacan con más de 4 millones de votos cada una, reflejando su impacto en la industria. En contraste, géneros como el musical, el western y el documental tienen menos votos, lo que sugiere un público más de nicho. Estos datos muestran cómo ciertos géneros capturan más la atención del público y generan mayor participación en votaciones.

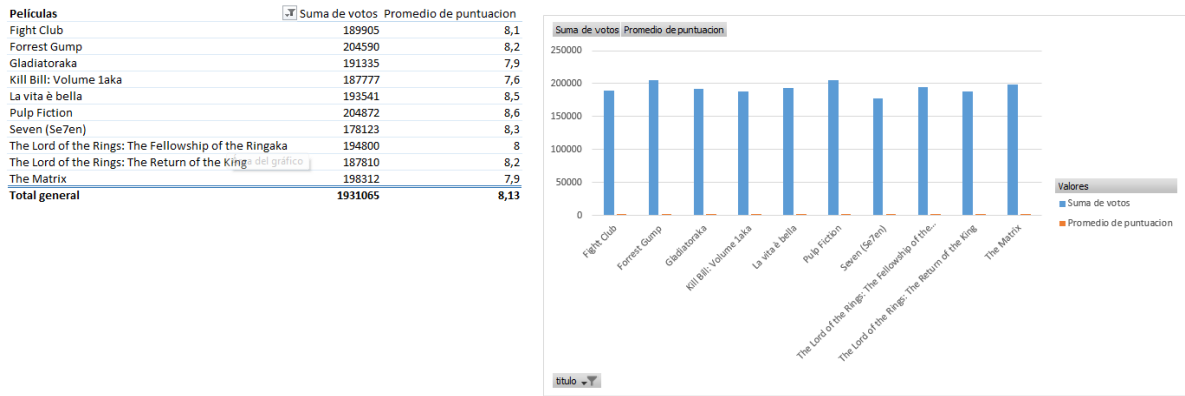
Votos por país



Estados Unidos domina con más de 36 millones de votos, muy por encima de cualquier otro país, lo que refleja su liderazgo en la industria cinematográfica global. Le siguen Reino Unido, Francia, Japón, Italia y España, con varios millones de votos, lo que indica una fuerte

presencia en la producción y distribución de películas. Otros países como Alemania, Argentina, Brasil y Corea del Sur también tienen una participación significativa. En contraste, países como Bolivia, Lesotho y Mongolia tienen cifras muy bajas, lo que sugiere un menor impacto en la producción o distribución internacional de cine.

Mejores 10 películas



Las 10 mejores películas analizadas presentan una puntuación promedio bastante alta, con una media de 8,13, lo que refleja una gran aprobación por parte de los usuarios. Entre ellas, *Pulp Fiction* destaca como la película con la mayor puntuación (8,6), seguida por *La vita è bella* (8,5), lo que indica que los filmes con historias complejas y profundas, como los mencionados, tienden a ser más apreciados. Además, el número de votos muestra que estas películas no solo tienen una buena recepción, sino que también cuentan con una amplia base de espectadores. Es interesante observar que la mayoría de las películas son de géneros como drama y acción, y que la saga de *El Señor de los Anillos* mantiene una valoración alta, pese a ser una franquicia extensa.

Conclusiones

El proyecto demostró con éxito cómo se puede automatizar la extracción de datos de una página web dinámica como FilmAffinity, utilizando herramientas como Selenium, que permiten interactuar con sitios web y extraer información de manera eficiente. Este enfoque automatizado no solo facilita la recolección de grandes volúmenes de datos, sino que también permite realizar un análisis más detallado y actualizado de las preferencias de los usuarios en la plataforma. Al integrar estos datos en una base de datos MySQL, se logra almacenar y organizar la información de manera estructurada, lo que facilita su posterior consulta y análisis.

Los análisis realizados sobre los datos extraídos revelaron tendencias interesantes que ofrecen una visión más profunda de las preferencias del público. Por ejemplo, se observó una clara inclinación hacia las películas de drama, que no solo obtenían altas puntuaciones, sino también un número considerable de votos. Además, se identificó que las películas con más votos tienden a recibir puntuaciones más altas, lo que podría indicar que las producciones populares y ampliamente vistas son también aquellas que logran captar el interés y la satisfacción de la audiencia. En general, el proyecto no solo mostró la viabilidad de la automatización de la recolección de datos, sino también cómo estos pueden ser utilizados para descubrir patrones valiosos que podrían orientar tanto a cineastas como a plataformas de streaming sobre las preferencias del público.

Mejoras posibles

- Ampliar la cantidad de datos extraídos (más de 3600 películas).
- Realizar análisis más complejos utilizando herramientas como Python y Pandas.
- Implementar un sistema de actualización en tiempo real para obtener datos nuevos automáticamente.

Código Fuente

El código fuente se encuentra en el repositorio de GitHub:

[GitHub - Proyecto de Scraping FilmAffinity](#)