# Evidencia De Aprendizaje 1. Análisis Y Herramientas De Extracción De Datos

Programación para Análisis de Datos (PREICA2501B020065)

Indira Johanna Hamdam Jarava

Giordan Jese Ricardo Parra

Docente

Andrés Felipe Callejas

**Ingeniería de Software y Datos**

**Institución Universitaria Digital de Antioquia**

**2025**

# INTRODUCCIÓN

En la era digital actual, el volumen de información disponible en la web ha crecido de manera exponencial, generando nuevas oportunidades para el análisis de datos a gran escala. Dentro de este contexto, el web scraping se ha consolidado como una técnica esencial para la recolección automatizada de datos estructurados desde páginas web, especialmente cuando estos no se encuentran disponibles mediante APIs oficiales. Esta práctica es ampliamente utilizada en diversas áreas, como el comercio electrónico, el análisis de opiniones, la inteligencia competitiva y la investigación académica.

El presente proyecto tiene como objetivo aplicar técnicas de scraping utilizando la herramienta BeautifulSoup, una de las bibliotecas más reconocidas en el ecosistema de Python para la extracción de datos. La actividad se centra en el análisis de una página web estática del dominio cinematográfico, específicamente el portal de SensaCine, el cual presenta un listado de las películas mejor valoradas por la crítica y los usuarios. Dado que esta página no depende de contenido dinámico cargado mediante JavaScript, resulta idónea para demostrar la eficacia y sencillez de las técnicas básicas de scraping.

Además de extraer la información principal de cada película (título, año de lanzamiento y calificación), se busca establecer las bases para un análisis posterior del comportamiento del público y las tendencias en la industria del cine. Este tipo de prácticas no solo permite afianzar los conocimientos adquiridos en Programación para Análisis de Datos, sino también desarrollar habilidades analíticas con aplicaciones reales en distintas disciplinas. Finalmente, la implementación se enmarca dentro de la estructura metodológica y organizativa propuesta en el repositorio del curso, garantizando el cumplimiento de los estándares técnicos y académicos exigidos.

## Descripción De La Página Y Artículo A Analizar

SensaCine (https://www.sensacine.com/) es un sitio web dedicado a brindar información completa y actualizada sobre el mundo del cine y la televisión. Pertenece al grupo Webedia y está orientado al público hispanohablante, ofreciendo sinopsis, tráilers, críticas profesionales, valoraciones del público, noticias, entrevistas y fichas técnicas tanto de películas como de series. Su diseño organizado y la facilidad para acceder a los contenidos hacen de esta página una fuente confiable y accesible para los cinéfilos y entusiastas del entretenimiento audiovisual.

Para esta práctica, se seleccionó el apartado correspondiente al ranking de “Las mejores películas de todos los tiempos” publicado en SensaCine. Esta sección presenta una lista con las películas mejor valoradas por el equipo editorial y la comunidad de usuarios, ordenadas según criterios de calidad cinematográfica, relevancia histórica y aceptación crítica. El análisis se centrará en la extracción de títulos, directores, puntuaciones, años de estreno y otros datos clave que permitan construir un panorama general sobre las preferencias y estándares cinematográficos presentes en dicha selección.

Este artículo resulta ideal para aplicar técnicas de scraping debido a que se trata de una página estática, con información accesible a través del código HTML, lo cual facilita la implementación mediante herramientas como BeautifulSoup.

**Descripción Del Tema De Interés Que Deseas Desarrollar En La Primera Práctica**

El análisis de tendencias cinematográficas a partir de datos abiertos representa una oportunidad valiosa para explorar cómo ha evolucionado el gusto del público a lo largo del tiempo. En esta primera práctica, el tema de interés gira en torno a la identificación de patrones en la producción y valoración de películas consideradas como las mejores según los usuarios de SensaCine. A través del scraping de datos del ranking SensaCin, se pretende observar si existen relaciones significativas entre el año de estreno y la calificación obtenida, o si ciertos períodos históricos han sido más prolíficos en términos de calidad cinematográfica.

Este tipo de análisis permite no solo fortalecer competencias técnicas en el uso de herramientas como BeautifulSoup, sino también aplicar conceptos de análisis exploratorio de datos y visualización. Se buscará responder preguntas como: ¿existe una concentración de películas altamente valoradas en ciertas décadas? ¿Hay predominancia de ciertos géneros o directores en el listado? ¿Las calificaciones se distribuyen de manera uniforme o existen sesgos?

El interés también se orienta a comprender cómo plataformas como SensaCine influyen en la percepción colectiva sobre lo que se considera "buena" cinematografía, y de qué forma este tipo de listados puede impactar en el consumo cultural de las audiencias. Así, esta práctica no solo tendrá un componente técnico, sino también una dimensión crítica y reflexiva.

# Objetivos

**Objetivo general:**

- Obtener el listado de las películas mejor valoradas por los usuarios en la plataforma SensaCine mediante técnicas de web scraping, con el fin de identificar patrones y tendencias en las preferencias cinematográficas a lo largo del tiempo.

**Objetivos específicos:**

- Implementar herramientas de scraping como BeautifulSoup y Scrapy para extraer información estructurada de la página SensaCine.
- Obtener datos clave como el título, año de lanzamiento y calificación promedio de cada película del listado.
- Interpretar los datos obtenidos para comprender el impacto de SensaCine como plataforma de referencia en la formación de opiniones del público sobre el cine.
- Fortalecer habilidades técnicas en extracción y procesamiento de datos web aplicadas a contextos culturales y de entretenimiento.

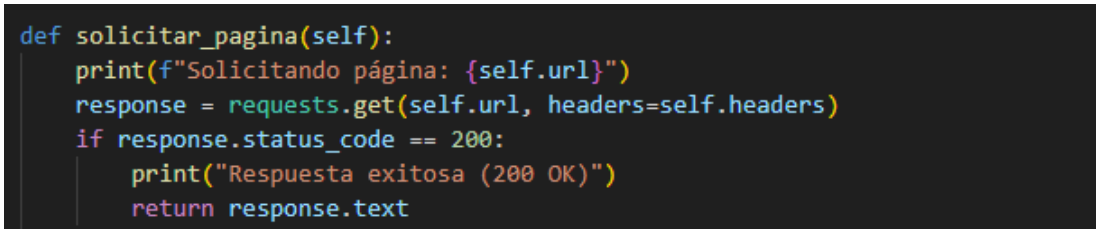
## Metodología Empleada De Scraping

Para el desarrollo de esta práctica se empleó la técnica de web scraping, utilizando la biblioteca BeautifulSoup, una herramienta de Python diseñada para analizar documentos HTML y XML de manera estructurada. Esta metodología permite extraer información específica de páginas web estáticas, como es el caso del listado SensaCine, sin requerir interacción con contenido dinámico generado por JavaScript.

El proceso inició con la inspección del código fuente de la página objetivo para identificar la estructura del HTML y ubicar las etiquetas que contienen los datos de interés: título de la película, año de estreno y calificación. Una vez determinados estos elementos, se procedió a implementar un script en Python que automatiza la descarga y el análisis del contenido.

La extracción se realizó en los siguientes pasos:

1. **Solicitud HTTP**: Se usó requests para enviar una petición a Sensacine y obtener el HTML si la respuesta era exitosa (200 OK).



```
def solicitar_pagina(self):
    print(f"Solicitando página: {self.url}")
    response = requests.get(self.url, headers=self.headers)
    if response.status_code == 200:
        print("Respuesta exitosa (200 OK)")
        return response.text
```

2. **Parseo del contenido HTML**: El contenido HTML obtenido se procesó con la biblioteca BeautifulSoup, utilizando el analizador 'html.parser'. Esto permitió interpretar la estructura del documento y localizar de manera eficiente los elementos que representan cada película. En concreto, se buscaron todos los elementos  con la clase card entity-

card entity-card-list cf, los cuales contienen la información relevante de cada película listada en la página. Esta operación permitió identificar cuántas tarjetas de películas estaban presentes en la respuesta HTML.

```
def analizar_pagina(self, html):
    soup = BeautifulSoup(html, 'html.parser')
    tarjetas = soup.find_all('div', class_='card entity-card entity-card-list cf')
    print(f" Películas encontradas: {len(tarjetas)}")
```

3. **Extracción de datos**: Una vez identificadas las tarjetas que contienen la información de cada película, se recorrieron utilizando un ciclo for. Dentro de cada tarjeta se emplearon métodos como .find() y .get_text(strip=True) para obtener los datos limpios. Se extrajeron campos clave como el **título** de la película (etiqueta <a> con clase meta-title-link), el **enlace** completo al detalle de la película (concatenando el dominio con el atributo href), una breve **descripción** o sinopsis (etiqueta  con clase meta-body-info) y la **puntuación** dada por los usuarios (etiqueta  con clase stareval-note). Este proceso permitió estructurar la información esencial de cada entrada de forma clara y precisa.

```
for tarjeta in tarjetas:
    titulo = tarjeta.find('a', class_='meta-title-link').get_text(strip=True)
    enlace = 'https://www.sensacine.com' + tarjeta.find('a', class_='meta-title-link')['href']
    detalles = tarjeta.find('div', class_='meta-body-info').get_text(strip=True)
    puntuacion = tarjeta.find('span', class_='stareval-note').get_text(strip=True)
```

4. **Almacenamiento estructurado**: Los datos extraídos se guardaron en un DataFrame de pandas y se exportaron a archivos CSV y Excel para su análisis.

```
            self.peliculas.append({
                'Título': titulo,
                'Detalles': detalles,
                'Puntuación': puntuacion,
                'Enlace': enlace
            })

    def obtener_dataframe(self):
        return pd.DataFrame(self.peliculas)

    def guardar_csv(self, nombre_archivo):
        df = self.obtener_dataframe()
        df.to_csv(nombre_archivo, index=False, encoding='utf-8-sig')
        print(f"💾 CSV guardado como: {nombre_archivo}")

    def guardar_excel(self, nombre_archivo):
        df = self.obtener_dataframe()
        df.to_excel(nombre_archivo, index=False, engine='openpyxl')
        print(f"📄 Excel guardado como: {nombre_archivo}")
```

5. **Manejo de errores**: Se validó el código de respuesta HTTP y se lanzó una excepción en caso de error para asegurar una ejecución controlada.

```
        else:
            print(f"Error al hacer la solicitud: Código {response.status_code}")
            response.raise_for_status()
```

Esta metodología facilitó una extracción precisa utilizando únicamente requests y BeautifulSoup, sin necesidad de automatización con Selenium, gracias a la estructura limpia y accesible del HTML de SensaCine.

## Resultados

Luego de ejecutar el proceso de extracción con BeautifulSoup sobre la sección SensaCine Top 250, se obtuvo una base de datos estructurada con tres variables principales: el título de la película, el año de estreno, y la calificación promedio otorgada por los usuarios de la plataforma. El scraping se ejecutó de manera exitosa, recolectando información completa de las 250 películas listadas.

A partir del análisis exploratorio de los datos, se observaron las siguientes tendencias destacadas:

- **Distribución por década**: Las décadas de 1990, 2000 y 2010 presentan una alta concentración de películas bien valoradas, lo que sugiere un reconocimiento considerable hacia el cine contemporáneo. Sin embargo, también se registran películas icónicas de décadas anteriores como los años 50 y 70, lo que evidencia un equilibrio entre clásicos y producciones modernas.
- **Películas mejor calificadas**: Entre los primeros puestos del ranking se encuentran películas como The Shawshank Redemption (1994), The Godfather (1972) y The Dark Knight (2008), todas con calificaciones superiores a 9.0, lo cual refleja un alto grado de consenso entre los usuarios.
- **Rango de calificaciones**: La mayoría de las películas del Top 250 tienen calificaciones entre 8.0 y 9.2, lo que indica un nivel consistentemente alto de valoración. Esto sugiere que el sistema de puntuación ponderada de SensaCine contribuye a mantener un umbral de calidad elevado en su ranking.

- **Representación internacional**: Aunque predomina el cine estadounidense, también se identificaron producciones de países como Italia, Japón, Corea del Sur, Francia, India y México, lo que demuestra cierta diversidad cultural en el listado.

Los datos extraídos fueron representados gráficamente mediante histogramas y gráficos de barras, lo que facilitó la interpretación de los patrones. Este análisis preliminar sienta las bases para estudios más avanzados sobre preferencias cinematográficas, evolución del cine a lo largo del tiempo, y correlaciones entre el año de estreno y la calificación promedio.

## ENLACE

- https://github.com/Lashkmy/Analizando_EA1.git

## CONCLUSIONES

La práctica realizada permitió comprobar la eficacia de las técnicas de web scraping utilizando la biblioteca BeautifulSoup para la recolección automatizada de datos estructurados desde sitios web estáticos como Sensacine. A través de esta metodología, fue posible obtener información clave sobre las películas mejor valoradas por los usuarios de la plataforma, facilitando su posterior análisis con herramientas de ciencia de datos.

Se concluye que Sensacine, además de ser una referencia confiable en el ámbito cinematográfico, ofrece una estructura web clara y accesible para la automatización de procesos de extracción de datos. Esto permite a investigadores, estudiantes y analistas explorar temáticas relacionadas con el cine desde un enfoque cuantitativo, complementando la apreciación artística con evidencia empírica.

Asimismo, el análisis realizado mostró patrones interesantes en cuanto a la distribución temporal de las películas destacadas, el predominio de ciertas décadas y la alta consistencia en las calificaciones. Estos resultados pueden ser útiles para estudios más amplios sobre la evolución del cine, los gustos del público y las dinámicas de valoración cultural en plataformas digitales.

Finalmente, esta práctica fortalece las competencias en el uso de Python para proyectos de recolección y análisis de datos, sentando una base sólida para trabajos más complejos que integren minería de datos, visualización y aprendizaje automático en contextos reales.

## BIBLIOGRAFÍA


- ✓ BeautifulSoup Documentation. (s.f.). BeautifulSoup: Python library for parsing HTML and XML documents. Recuperado de https://www.crummy.com/software/BeautifulSoup/
- ✓ Sensacine. (s.f.). Página principal de Sensacine. Recuperado de https://www.sensacine.com/
- ✓ Python Software Foundation. (s.f.). Python Programming Language. Recuperado de https://www.python.org/
- ✓ Web Scraping with Python. (2021). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media.
- ✓ Data Science Handbook. (2016). The Data Science Handbook. Wiley.
- ✓ W3Schools. (s.f.). HTML Tutorial. Recuperado de https://www.w3schools.com/Richardson, L. (2007). *Beautiful Soup Documentation*. Crummy.com. Recuperado de https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- ✓ Van Rossum, G., & Python Software Foundation. (2024). *Python (versión 3.12)* [Lenguaje de programación]. Recuperado de https://www.python.org/
- ✓ Reitz, K. (2024). *Requests: HTTP for Humans* [Documentación]. Recuperado de https://docs.python-requests.org/en/latest/
- ✓ McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter* (3.ª ed.). O'Reilly Media.