

# Actividad extracurricular 12 - Web Scraping

**Nombre:** Alexis Bautista

**Fecha de entrega:** 04 de enero de 2025

**Paralelo:** GR1CC

**Enlace de GitHub:** <https://github.com/alexis-bautista/Actividad12-MN>

## Web Scraping

El web scraping, también conocido como extracción de datos web, es el proceso automatizado de recolectar y estructurar información desde sitios web. Este procedimiento es ampliamente utilizado en diversas áreas como la monitorización de precios, investigación de mercados, recopilación de noticias y también en proyectos académicos o de aprendizaje automático (machine learning). En esencia, el web scraping emula la acción de copiar y pegar información manualmente, pero de forma sistemática y a gran escala, permitiendo extraer datos de una o varias páginas web en un tiempo considerablemente menor.

## Funcionamiento del Web Scraping

En el web scraping intervienen principalmente dos componentes:

1. **Crawler (rastreadores web):** Un programa diseñado para navegar por la web de manera automática. Utiliza los enlaces de las páginas web para explorar diferentes sitios y descubrir contenidos de interés. Este rastreador simula la navegación de un usuario humano y realiza solicitudes a servidores web para acceder a los contenidos.
2. **Scraper (extractor de datos):** Este componente se encarga de analizar las páginas obtenidas por el crawler, identificar las secciones relevantes y extraer los datos específicos deseados. Por ejemplo, un scraper puede buscar en una tabla de precios o extraer artículos de noticias. Para funcionar correctamente, se debe proporcionar al scraper la URL del sitio de interés y definir las características de los datos que se quieren recolectar.

## Herramientas Comunes en Web Scraping

El proceso de web scraping puede realizarse utilizando herramientas y frameworks que simplifican tareas como el acceso a sitios web, el análisis del código HTML y la extracción de datos. En Python, existen varias librerías populares que permiten implementar soluciones eficientes de scraping:

- **Requests:** Permite realizar solicitudes HTTP para acceder al contenido de las páginas web. Es sencillo de usar y ampliamente utilizado para interactuar con

servidores web.

- **BeautifulSoup:** Una biblioteca diseñada para analizar código HTML y XML. Facilita la navegación y búsqueda en estructuras jerárquicas, ayudando a localizar y extraer datos específicos.
- **Scrapy:** Un framework completo para web scraping que incluye funcionalidades avanzadas como la gestión de crawlers, manejo de datos estructurados y optimización del rendimiento.
- **Selenium:** Ideal para interactuar con páginas web dinámicas que utilizan JavaScript. Permite automatizar navegadores web para realizar tareas complejas como completar formularios o hacer clic en botones.

## Proceso del Web Scraping

El proceso de web scraping puede dividirse en las siguientes etapas:

1. **Acceso al sitio web:** Se envía una solicitud HTTP al servidor del sitio web deseado. En respuesta, el servidor proporciona el código HTML de la página.
2. **Análisis del contenido:** El contenido recibido es procesado para identificar las secciones relevantes. Esto se realiza mediante el análisis del DOM (Modelo de Objetos del Documento) de la página web.
3. **Extracción de datos:** Los elementos de interés (como textos, imágenes o enlaces) son localizados y extraídos según las especificaciones definidas previamente.
4. **Almacenamiento de datos:** La información recolectada se estructura y guarda en un formato adecuado, como CSV, JSON o bases de datos.

## Aplicaciones del Web Scraping

El web scraping es una herramienta poderosa y versátil. Algunas de sus aplicaciones más comunes incluyen: monitorización de precios, análisis de mercado, recopilación de noticias, Machine Learning, proyectos académicos (extracción de datos específicos para trabajos de investigación.)

## Consideraciones Éticas y Legales

Aunque el web scraping ofrece múltiples beneficios, es esencial tener en cuenta los aspectos legales y éticos relacionados con su uso. Algunos puntos importantes son:

- **Respeto a los Términos de Servicio:** Muchos sitios web especifican en sus términos y condiciones las restricciones sobre el uso de crawlers y scrapers.
- **Evitar sobrecargar los servidores:** Es fundamental diseñar soluciones que no generen un volumen excesivo de solicitudes al servidor, lo que podría causar problemas de rendimiento.

- **Protección de datos:** Asegurarse de no recopilar información que infrinja las leyes de privacidad o derechos de autor.

## Ejemplo en python para Web Scraping

Se usaron las librerías BeautifulSoup requests y Scrapy para realizar un ejemplo de scraping al sitio web de BBC News.

### BeautifulSoup

```
In [16]: import requests
from bs4 import BeautifulSoup
import csv

page = requests.get("https://www.bbc.com/news") # para obtener el html de la pa
soup = BeautifulSoup(page.text, "html.parser") # para parsear el html

# Extraer los titulares de las noticias
headlines = soup.find_all('h2')

# Imprimir los titulares
for headline in headlines:
    print(headline.get_text())

# Abrir un archivo CSV para escribir
with open('headlines.csv', mode='w', newline='', encoding='utf-8') as file:
    writer = csv.DictWriter(file, fieldnames=['headline'])
    writer.writeheader()

# Escribir los titulares en el archivo CSV
for headline in headlines:
    writer.writerow({'headline': headline.get_text()})
```

South Korea plane crash kills 179 with investigation into cause under way  
Video captures moments before South Korea plane crash  
Billionaire HBO creator Charles Dolan dies aged 98  
As Putin reaches 25 years in power, has he 'taken care of Russia'?  
Nigerians take to the streets for Calabar Carnival  
Video captures moments before South Korea plane crash  
Did bird strike contribute to South Korea plane crash? What we know so far  
New elections could take up to four years, Syria rebel leader says  
Chlamydia could make koalas extinct. Can a vaccine save them in time?  
Billionaire HBO creator Charles Dolan dies aged 98  
Lost city found by accident and a fly's brain mapped: 2024's scientific wins  
Quiz of the Year, Part 4: Why did 100 couples all say 'I do' together?  
Tourist killed in shark attack off Egyptian coast  
Trump sides with tech bosses in MAGA fight over immigrant visas  
Azerbaijan urges Russia to accept blame for plane crash  
More to explore  
Maggie Smith, Liam Payne and the other famous people who died in 2024  
Inside a Syrian 'reconciliation centre' where Assad's soldiers give up their weapons  
UK and EU look to 2025 for reset, but with little room for trade-offs  
Growth of women in power grinds to near-halt in a mega-election year  
From Squid Game to Blackpink, how South Korea became a culture powerhouse  
Inside a Syrian 'reconciliation centre' where Assad's soldiers give up their weapons  
A year of extreme weather that challenged billions  
How feminism, not Bollywood, drew global audiences to Indian cinema in 2024  
UK and EU look to 2025 for reset, but with little room for trade-offs  
Most watched  
Video captures moments before South Korea plane crash  
Watch: Huge waves strike Peruvian coastline  
Azerbaijan plane survivors recall moments before crash  
Watch: Police officer dressed as the Grinch leads drug raid  
Footage shows survivors walking from crashed Azerbaijani plane  
Also in news  
Toddler nearly runs off cliff at Hawaii volcano  
School chaplain killed in shark attack on Australia's Great Barrier Reef  
Trump sides with tech bosses in MAGA fight over immigrant visas  
Romeo and Juliet actress Olivia Hussey dies aged 73  
Watch: Huge waves strike Peruvian coastline  
School chaplain killed in shark attack on Australia's Great Barrier Reef  
Kiefer Sutherland grew up unaware of dad Donald's success  
Three migrants die attempting to cross Channel  
Trump sides with tech bosses in MAGA fight over immigrant visas  
Most read  
South Korea plane crash kills 179 with investigation into cause under way  
Did bird strike contribute to South Korea plane crash? What we know so far  
New elections could take up to four years, Syria rebel leader says  
As Putin reaches 25 years in power, has he 'taken care of Russia'?  
Billionaire HBO creator Charles Dolan dies aged 98  
Notable deaths 2024  
Chlamydia could make koalas extinct. Can a vaccine save them in time?  
Rebel Wilson marries Ramona Agruma in Sydney ceremony  
Lost city found by accident and a fly's brain mapped: 2024's scientific wins  
Kiefer Sutherland grew up unaware of dad Donald's success  
Sport  
PDC World Darts Championship: Owen beats Evans in last third-round tie  
Ruthless Liverpool thrash West Ham to go eight points clear  
Thuram scores twice but Juventus held by Fiorentina  
Rabada sends SA into Test final with thrilling win  
Australia rally after Bumrah genius on riveting day

Ruthless Liverpool thrash West Ham to go eight points clear  
Forest lend retro feel to table - here's why they can stay the course  
'We need help' - Guardiola targets January & says 'no chance' of title  
Thuram scores twice but Juventus held by Fiorentina  
Follow BBC on:

## Scrapy

Para usar Scrapy primero se creo un proyecto por terminal de la siguiente manera:

```
scrapy startproject noticias
```

 . Lo cual genera un directorio.

Luego se creo un spider en la ubicacion `noticias/noticias/spiders` el spider es un archivo de Python que se lo creo usando: `touch titulares_spider.py`

El archivo titulares\_spider.py contiene el siguiente código:

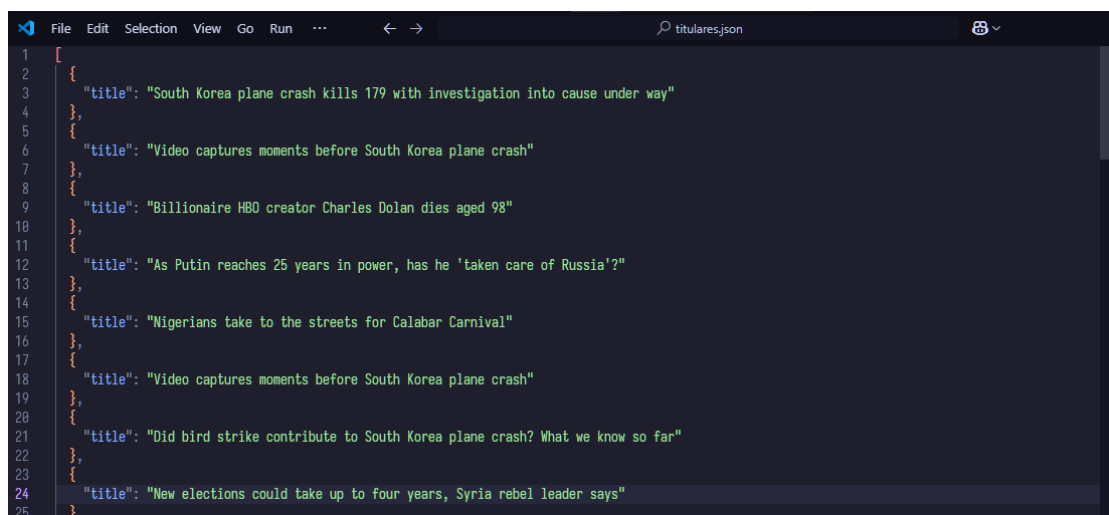
```
import scrapy

class TitularesSpider(scrapy.Spider):
    name = "titulares"
    start_urls = ["https://www.bbc.com/news"]

    def parse(self, response):
        # Selecciona los elementos que contienen los titulares
        for article in response.css("div.gs-c-promo"):
            yield {
                "title": article.css(
                    "h2::text"
                ).get(), # Extrae el texto del titular.
            }
```

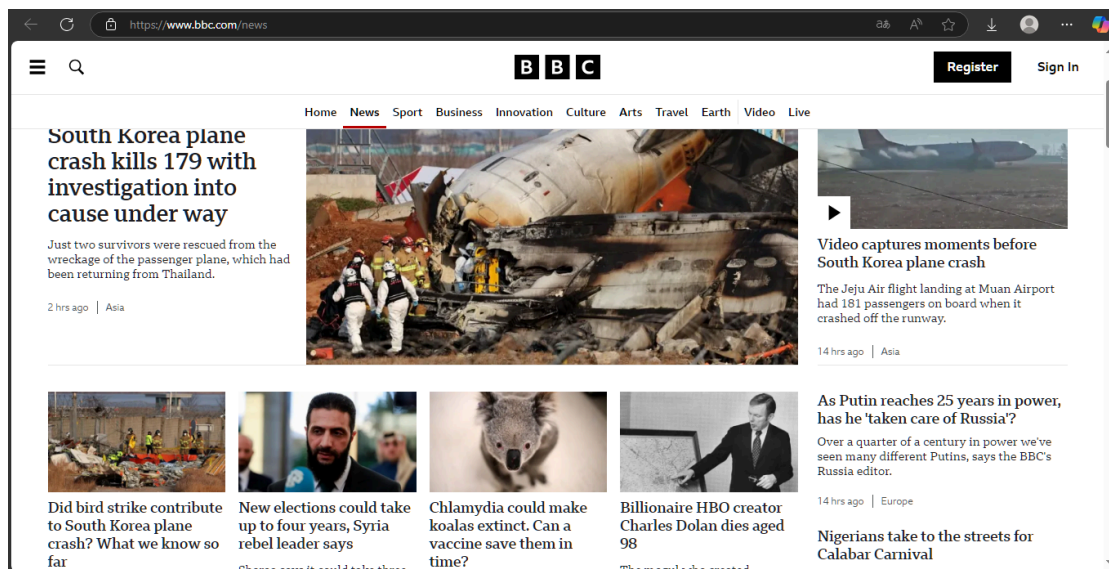
Teniendo todo esto en cuenta se ejecuto el spider: `scrapy crawl titulares -o titulares.json`

Esto generó un archivo .json que contiene los titulares extraídos.



```
[
  {
    "title": "South Korea plane crash kills 179 with investigation into cause under way"
  },
  {
    "title": "Video captures moments before South Korea plane crash"
  },
  {
    "title": "Billionaire HBO creator Charles Dolan dies aged 98"
  },
  {
    "title": "As Putin reaches 25 years in power, has he 'taken care of Russia'?"
  },
  {
    "title": "Nigerians take to the streets for Calabar Carnival"
  },
  {
    "title": "Video captures moments before South Korea plane crash"
  },
  {
    "title": "Did bird strike contribute to South Korea plane crash? What we know so far"
  },
  {
    "title": "New elections could take up to four years, Syria rebel leader says"
  }
]
```

Se comprobó que los resultados obtenidos coinciden con el de la página web de BBC News.



## Conclusiones

El web scraping es una técnica eficaz para extraer y estructurar información de la web. Gracias a las herramientas disponibles, como BeautifulSoup y Scrapy en Python, es posible implementar soluciones potentes que facilitan la automatización de tareas repetitivas y la obtención de datos valiosos. Sin embargo, es crucial realizar estas actividades de manera ética y respetuosa con las normas establecidas, asegurando un equilibrio entre el aprovechamiento de los recursos web y el cumplimiento de los principios legales.

## Bibliografía

- Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. Briefings in bioinformatics, 15(5), 788-797. Obtenido de: [Web scraping technologies in an API world](#)
- Yuste, G. (2022, agosto 17). ¿Qué es web scraping? KeepCoding Bootcamps. <https://keepcoding.io/blog/que-es-web-scraping/>
- Chanda, S. (2022, noviembre 28). Web Scraping Usando Python: Guía paso a paso. Geekflare Spain. <https://geekflare.com/es/web-scraping-with-python/>

## Librerías usadas

- BeautifulSoup: <https://beautiful-soup-4.readthedocs.io/en/latest/>
- Scrapy: <https://scrapy.org/>
- Requests: <https://docs.python-requests.org/en/latest/index.html>