

# Tarea extra

Boris Garcés

## Tabla de Contenidos

<b>Web scraping</b>	<b>1</b>
¿Qué es? . . . . .	1
Realizar una prueba en python para dos librerías diferentes . . . . .	1
Requests y BeautifulSoup . . . . .	1
Selenium . . . . .	3
Realizar scraping de un sitio web de su elección . . . . .	4

## Web scraping

### ¿Qué es?

Consiste en el proceso de extraer información de sitios web de forma automatizada, el nombre se debe a que se utilizan programas scripts capaces de acceder a uno o varios sitios o páginas web, identifican las estructuras del contenido, extraen y almacenan la información deseada en una forma útil para futuros análisis, sus principales usos se encuentran en la investigación de mercado, crear sitios o aplicaciones que reúnan datos de distintas fuentes, obtener grandes volúmenes de información para ser posteriormente analizados, entre otros.

### Realizar una prueba en python para dos librerías diferentes

#### Requests y BeautifulSoup

La librería Requests es una librería que permite enviar solicitudes HTTP, por lo que habilitará la interacción con el contenido de páginas web, mientras que la librería BeautifulSoup permite la extracción de datos HTML y XML por lo que se utilizará para mostrar los resultados de las solicitudes realizadas.

```

import requests
from bs4 import BeautifulSoup
url = "https://www.chess.com/"
try:
    response = requests.get(url)
    response.raise_for_status()
    soup = BeautifulSoup(response.text, "html.parser")
    titulo = soup.find("title").text
    print("Título de la página:", titulo)
    enlaces = soup.find_all("a")
    for enlace in enlaces:
        texto = enlace.text.strip()
        url_enlace = enlace.get("href")
        print(f"Enlace: {texto} -> {url_enlace}")
except requests.exceptions.RequestException as e:
    print("Error durante la petición:", e)

```

Título de la página: Chess.com - Play Chess Online - Free Games

Enlace: Home -> <https://www.chess.com/>

Enlace: Play -> <https://www.chess.com/play>

Enlace: Puzzles -> <https://www.chess.com/puzzles/rated>

Enlace: Learn -> <https://www.chess.com/learn>

Enlace: Watch -> <https://www.chess.com/watch>

Enlace: News -> <https://www.chess.com/today>

Enlace: Social -> <https://www.chess.com/social>

Enlace: -> <https://www.chess.com/search>

Enlace: Sign Up -> <https://www.chess.com/register?returnUrl=https://www.chess.com/>

Enlace: Log In -> [https://www.chess.com/login\\_and\\_go?returnUrl=https://www.chess.com/](https://www.chess.com/login_and_go?returnUrl=https://www.chess.com/)

Enlace: -> <https://www.chess.com/>

Enlace: Sign Up -> <https://www.chess.com/register?returnUrl=https://www.chess.com/>

Enlace: Log In -> [https://www.chess.com/login\\_and\\_go?returnUrl=https://www.chess.com/](https://www.chess.com/login_and_go?returnUrl=https://www.chess.com/)

Enlace: Play Bots

Play vs customizable training bots -> <https://www.chess.com/play/computer>

Enlace: -> <https://www.chess.com/puzzles/rated>

Enlace: Solve Puzzles -> <https://www.chess.com/puzzles/rated>

Enlace: -> <https://www.chess.com/lessons>

Enlace: Start Lessons -> <https://www.chess.com/lessons>

Enlace: Follow the 2024 FIDE World Championship LIVE with the BEST coverage. -> <https://www.chess.com/live>

Enlace: Fedoseev, Lazavik, Sindarov, Bortnyk Keep Weissenhaus Hopes Alive

Colin\_McGourty -> <https://www.chess.com/news/view/2025-freestyle-chess-grand-slam-weissen>  
Enlace: It's Tactics Time! The Chess.com Puzzles Championship Starts On January 16

CHESScom -> <https://www.chess.com/news/view/announcing-chesscom-puzzles-championship-2025>  
Enlace: Queen Sacrifices: From Obvious To Impossible

Gserper -> <https://www.chess.com/article/view/queen-sacrifices-from-obvious-to-impossible>  
Enlace: Rare Fourth Moves

GM

JanistanTV -> <https://www.chess.com/video/player/rare-fourth-moves>  
Enlace: Chess Today -> <https://www.chess.com/today>  
Enlace: Support -> <https://chess.com/support>  
Enlace: Chess Terms -> <https://www.chess.com/terms>  
Enlace: About -> <https://www.chess.com/about>  
Enlace: Jobs -> <https://www.chess.com/jobs>  
Enlace: Developers -> <https://www.chess.com/club/chess-com-developer-community>  
Enlace: User Agreement -> <https://www.chess.com/legal/user-agreement>  
Enlace: Privacy Policy -> <https://www.chess.com/legal/privacy>  
Enlace: Privacy Settings -> [https://www.chess.com/legal/privacy#privacy\\_settings](https://www.chess.com/legal/privacy#privacy_settings)  
Enlace: Fair Play -> <https://www.chess.com/fair-play>  
Enlace: Partners -> <https://www.chess.com/partners>  
Enlace: Compliance -> <https://www.chess.com/legal/compliance>  
Enlace: Chess.com © 2025 -> <https://www.chess.com/>  
Enlace: -> <https://www.chess.com/play/apps/ios>  
Enlace: -> <https://www.chess.com/play/apps/android>  
Enlace: -> <https://www.tiktok.com/@chess>  
Enlace: -> <https://twitter.com/chesscom>  
Enlace: -> <https://www.youtube.com/user/wwwChesscom>  
Enlace: -> <https://www.twitch.tv/chess>  
Enlace: -> <https://www.instagram.com/wwwchesscom>  
Enlace: -> <https://discord.gg/3VbUQME>

## Selenium

Selenium es un entorno de pruebas de software cuyo objetivo es validar aplicaciones web, por lo que permite controlar navegadores imitando el comportamiento humano.

```

from selenium import webdriver
from selenium.webdriver.common.by import By
import time
driver = webdriver.Chrome()
driver.get("https://www.chess.com/")
time.sleep(2)
titulo = driver.find_element(By.TAG_NAME, "title").get_attribute("innerText")
print("Título de la página (Selenium):", titulo)
try:
    elemento_h1 = driver.find_element(By.TAG_NAME, "h1")
    print("Texto dentro de <h1>:", elemento_h1.text)
except:
    print("No se encontró ningún elemento <h1> en la página.")
driver.quit()

```

Título de la página (Selenium): Chess.com - Play Chess Online - Free Games  
 Texto dentro de <h1>: Play Chess  
 Online  
 on the #1 Site!

## Realizar scraping de un sitio web de su elección

```

from selenium import webdriver
from selenium.webdriver.common.by import By
driver = webdriver.Chrome()
driver.get("https://aulasvirtuales.epn.edu.ec/login/index.php")
results = {}
try:
    results['title'] = driver.title
    headers = driver.find_elements(By.XPATH, "//h1 | //h2 | //h3")
    results['headers'] = [header.text for header in headers if header.text]
    links = driver.find_elements(By.TAG_NAME, "a")
    results['links'] = \
    [link.get_attribute('href') for link in links if link.get_attribute('href')]
    images = driver.find_elements(By.TAG_NAME, "img")
    results['images'] = \
    [img.get_attribute('src') for img in images if img.get_attribute('src')]
    body_text = driver.find_element(By.TAG_NAME, "body").text

```

```

results['body_text_sample'] = body_text[:1000]
for key, value in results.items():
    print(f"{key}:")
    if isinstance(value, list):
        print(f" {value[:5]}...")
    else:
        print(f" {value}")
    print("\n")

finally:
    driver.quit()

```

```

title:
    Aula Virtual - EPN: Log in to the site

```

```

headers:
    []...

```

```

links:
    ['https://aulasvirtuales.epn.edu.ec/login/index.php#maincontent', 'https://aulasvirtuales.

```

```

images:
    ['https://aulasvirtuales.epn.edu.ec/LOGO_EPN1.svg']...

```

```

body_text_sample:
    Skip to main content
English (es)
    Preguntas Frecuentes
Username
Password
Remember username
Log in
asistencia.tecnica@epn.edu.ec
(+593) 2 2976 300 ext 1402 / 1404
Escuela Politécnica Nacional © Todos los Derechos Reservados.2024

```