

## M2.2 Web Scraping



Que es web scrapping?



El web scraping es una técnica utilizada para extraer información de sitios web de manera automatizada. Implica el uso de scripts o bots que navegan por páginas web, acceden a sus datos y los recopilan en un formato estructurado, como una hoja de cálculo o una base de datos. Este proceso puede ser útil para diversas aplicaciones, como la recopilación de datos para análisis, investigación de mercado, seguimiento de precios, y mucho más.

### Principales Componentes del Web Scraping

1. **Solicitud HTTP:** El scraper envía una solicitud al servidor web, generalmente utilizando los métodos GET o POST, para obtener el contenido de una página web.
2. **Parsing:** Una vez que se recibe la respuesta del servidor, el scraper analiza el contenido HTML para extraer la información deseada. Esto se puede hacer utilizando diversas bibliotecas y herramientas, como BeautifulSoup en Python.
3. **Extracción de Datos:** El scraper identifica y extrae los datos relevantes del contenido analizado. Esto puede incluir texto, imágenes, enlaces, etc.
4. **Almacenamiento de Datos:** Los datos extraídos se almacenan en un formato estructurado, como CSV, JSON, bases de datos, etc., para su posterior uso y análisis.

### Ejemplo de Código en Python usando BeautifulSoup

python

```
import requests
from bs4 import BeautifulSoup

# URL de la página que queremos scrapear
url = 'http://example.com'

# Realizar la solicitud HTTP
response = requests.get(url)

# Parsear el contenido HTML
soup = BeautifulSoup(response.text, 'html.parser')

# Extraer datos (por ejemplo, todos los títulos de los artículos)
titles = soup.find_all('h2', class_='title')
for title in titles:
    print(title.get_text())
```

## Consideraciones Legales y Éticas

- **Políticas de Uso:** Muchos sitios web tienen políticas de uso que prohíben el web scraping. Es importante revisar los términos de servicio de un sitio antes de realizar scraping.
- **Carga del Servidor:** El web scraping puede generar una carga significativa en los servidores web. Es recomendable implementar buenas prácticas, como limitar la frecuencia de solicitudes y respetar los archivos robots.txt de los sitios.
- **Respeto a la Propiedad Intelectual:** Los datos en la web están sujetos a derechos de autor y otros tipos de propiedad intelectual. Debes asegurarte de que el uso que hagas de los datos extraídos cumple con las leyes aplicables.



Genérame el código en python para automatizar el proceso de acceder a la url de <https://www.youtube.com/> y buscar el canal: No te duermas! y acceder al video: "El huésped de drácula" utilizando ChromeDriver y selenium, te mando un ejemplo:

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from webdriver_manager.chrome import ChromeDriverManager

# Configuración de opciones para el navegador
chrome_options = Options()
chrome_options.add_argument("--start-maximized") # Abre el navegador en pantalla completa

# Configuración del controlador
service = Service(ChromeDriverManager().install())

# Inicializa el navegador
driver = webdriver.Chrome(service=service, options=chrome_options)

try:

    # Paso 1: Abre Wikipedia
    driver.get("https://es.wikipedia.org/wiki/Wikipedia:Portada")

    # Espera a que el cuadro de búsqueda esté presente
    search_box = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.ID, "searchInput"))
    )
```

```
search_box.send_keys("TecNM")
search_box.submit()

# Espera a que aparezcan los resultados
pachuca_link = WebDriverWait(driver, 10).until(
    EC.presence_of_element_located((By.PARTIAL_LINK_TEXT, "Instituto Tecnológico de Pachuca"))
)

pachuca_link.click()
# Espera a que la página del Instituto Tecnológico de Pachuca se cargue
oferta_academica_link = WebDriverWait(driver, 10).until(
    EC.presence_of_element_located((By.LINK_TEXT, "Oferta académica"))
)

oferta_academica_link.click()
finally:

# Mantén el navegador abierto
input("Presiona Enter para cerrar el navegador...")

driver.quit()
```



Claro, a continuación te proporciono el código en Python utilizando `selenium` y `ChromeDriver` para acceder a YouTube, buscar el canal "No te duermas!" y luego acceder al video "El huésped de drácula":

python

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from webdriver_manager.chrome import ChromeDriverManager

# Configuración de opciones para el navegador
chrome_options = Options()
chrome_options.add_argument("--start-maximized") # Abre el navegador en pantalla completa

# Configuración del controlador
service = Service(ChromeDriverManager().install())

# Inicializa el navegador
driver = webdriver.Chrome(service=service, options=chrome_options)

try:
    # Paso 1: Abre YouTube
    driver.get("https://www.youtube.com")

    # Espera a que el cuadro de búsqueda esté presente
    search_box = WebDriverWait(driver, 10).until(
```

```

        EC.presence_of_element_located((By.NAME, "search_query"))
    )

    # Paso 2: Busca el canal "No te duermas!"
    search_box.send_keys("No te duermas!")
    search_box.submit()

    # Espera a que aparezcan los resultados y haz clic en el canal
    channel_link = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.PARTIAL_LINK_TEXT, "No te duermas!"))
    )
    channel_link.click()

    # Paso 3: Busca el video "El huésped de drácula" en el canal
    video_link = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.PARTIAL_LINK_TEXT, "El huésped de drácula"))
    )
    video_link.click()

finally:
    # Mantén el navegador abierto
    input("Presiona Enter para cerrar el navegador...")

    driver.quit()

```

Este código realiza los siguientes pasos:

1. Abre YouTube.
2. Busca el canal "No te duermas!".
3. Accede al canal.
4. Busca y accede al video "El huésped de drácula".

Recuerda tener `ChromeDriver` y las bibliotecas necesarias (`selenium`, `webdriver_manager`) instaladas en tu entorno de Python. Puedes instalar estas bibliotecas utilizando pip:

```
sh
```

```
pip install selenium webdriver_manager
```



Modifica el código que me enviaste para que se acceda el canal específico:  
<https://www.youtube.com/@noteduermas3128> para después buscar y acceder el video "El huésped de drácula" que contiene dicho canal



Claro, a continuación te presento el código modificado para acceder directamente al canal específico en YouTube y luego buscar y acceder al video "El huésped de drácula":

python

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from webdriver_manager.chrome import ChromeDriverManager

# Configuración de opciones para el navegador
chrome_options = Options()
chrome_options.add_argument("--start-maximized") # Abre el navegador en pantalla completa

# Configuración del controlador
service = Service(ChromeDriverManager().install())

# Inicializa el navegador
driver = webdriver.Chrome(service=service, options=chrome_options)

try:
    # Paso 1: Abre el canal específico en YouTube
    driver.get("https://www.youtube.com/@noteduermas3128")

    # Espera a que la página del canal se cargue completamente
    WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.TAG_NAME, "body"))
    )

    # Paso 2: Busca el video "El huésped de drácula" dentro del canal
    search_box = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.XPATH, "//input[@id='search']"))
    )

    search_box.send_keys("El huésped de drácula")
    search_box.submit()

    # Espera a que aparezcan los resultados y haz clic en el video
    video_link = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.XPATH, "//a[@title='El huésped de drácula']"))
    )
    video_link.click()

finally:
    # Mantén el navegador abierto
    input("Presiona Enter para cerrar el navegador...")

    driver.quit()
```

Este código hace lo siguiente:

1. Abre la URL del canal específico en YouTube.
2. Espera a que la página del canal se cargue completamente.
3. Usa el cuadro de búsqueda dentro del canal para buscar el video "El huésped de drácula".
4. Espera a que aparezcan los resultados de la búsqueda y hace clic en el video especificado.

Asegúrate de que el título del video sea exactamente "El huésped de drácula" para que el XPath funcione correctamente. Si el título varía, es posible que necesites ajustar el XPath o la lógica para encontrar el video correcto.

modifica el código para cambiar el título del video, ya que es: "El huésped de Drácula | Relatos para no dormir T1 Ep. 1"



Claro, aquí tienes el código modificado para buscar y acceder al video con el título exacto "El huésped de Drácula | Relatos para no dormir T1 Ep. 1":

python

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from webdriver_manager.chrome import ChromeDriverManager

# Configuración de opciones para el navegador
chrome_options = Options()
chrome_options.add_argument("--start-maximized") # Abre el navegador en pantalla completa

# Configuración del controlador
service = Service(ChromeDriverManager().install())

# Inicializa el navegador
driver = webdriver.Chrome(service=service, options=chrome_options)

try:
    # Paso 1: Abre el canal específico en YouTube
    driver.get("https://www.youtube.com/@noteduermas3128")

    # Espera a que la página del canal se cargue completamente
    WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.TAG_NAME, "body"))
    )

    # Paso 2: Usa el cuadro de búsqueda dentro del canal para buscar el video
    search_box = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.XPATH, "//input[@id='search']"))
    )

    search_box.send_keys("El huésped de Drácula | Relatos para no dormir T1 Ep. 1")
    search_box.submit()

    # Espera a que aparezcan los resultados y haz clic en el video
    video_link = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.XPATH, "//a[@title='El huésped de Drácula | Relatos para no dormir T1 Ep. 1']"))
    )
    video_link.click()

finally:
    # Mantén el navegador abierto
    input("Presiona Enter para cerrar el navegador...")

    driver.quit()
```

Este código realiza los siguientes pasos:

1. Abre la URL del canal específico en YouTube.
2. Espera a que la página del canal se cargue completamente.
3. Usa el cuadro de búsqueda dentro del canal para buscar el video "El huésped de Drácula | Relatos para no dormir T1 Ep. 1".
4. Espera a que aparezcan los resultados de la búsqueda y hace clic en el video especificado.

Asegúrate de tener instaladas las bibliotecas necesarias (`selenium`, `webdriver_manager`) en tu entorno de Python. Puedes instalarlas utilizando pip:

```
sh
```

```
pip install selenium webdriver_manager
```

ChatGPT can make mistakes. Check important info.