WEB SCRAPING

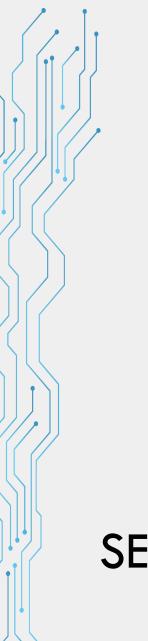
PAULO CÉSAR TUYA



ÍNDICE

- Módulo 1: Introducción a Web Scraping
- Módulo 2: Herramientas para el Análisis de una Página Web
- Módulo 3: Web Scraping con Scrapy
- Módulo 4: Scraping en páginas dinámicas
 - Automatizar interacciones mediante Selenium
 - Rompimiento de Captchas mediante Tesseract
 - Rellenado de Formularios
- Módulo 5: Despliegue de un Spider





SELENIUM



SELENIUM



- Conjunto de herramientas de automatización de pruebas
- Ejemplos de estas herramientas:
 - Selenium IDE
 - Selenium WebDriver
 - Selenium Grid
- WebDriver: nos permite simular el uso del navegador a través de código



SELENIUM WEBDRIVER

- Permite controlar un navegador específico mediante un lenguaje de programación
- Software externo que debe descargarse para ser ejecutado
- Posee una interfaz en Python que también debe descargarse





INCORPORANDO SELENIUM A DINUESTRO SCRAPING

- Selenium no entra en conflicto con BeautifulSoup o
 Scrapy, pues cumple otra función
- Es posible extraer código HTML con Selenium para utilizarlo con BeautifulSoup
- O bien crear un Middleware de Scrapy basado en Selenium para simular interacciones







TESSERACT

- Motor de Reconocimiento Óptico de Caracteres (OCR) de código abierto
- Desarrollado actualmente por Google
- Permite **leer** texto de una imagen
- Funciona incluso si el texto está distorsionado
- Es muy útil para romper captchas basados en texto





ACTIVIDADES MÓDULO 3



