WEB SCRAPING

Y LAS BASES DE DATOS OFUSCADAS

Abril 2024

Capgemini

QUÉ APRENDERÁS HOY

“Defensa” y riesgos del Web scraping

- Algunas estrategias

Testing enfocado a web scraping

¿QUIÉN SOY?

Pepe Fabra Valverde
Líder y Arquitecto de Front

Haciendo web scraping desde mis inicios, a nivel personal, académico y ligeramente profesional

AVISO PARA NAVEGANTES

- Los recursos se compartirán al final de la sesión
- La sesión **se grabará**, aparecerá el enlace cuando termine
- Es parte de una serie de charlas acerca de Web Scraping

1

DEFENSA ANTE WEB SCRAPING

¿CÓMO OFUSCAMOS NUESTROS DATOS?

“Defensa” ante web scraping

QUÉ OCURRE CUANDO HACEMOS WEB SCRAPING PARA EL OTRO LADO

Extracción de datos

- Clientes potenciales -> bien
- Competidores -> no tan bien

DDoS involuntario

Extracción de información confidencial (posibles exploits)

TÉCNICAS DE “OFUSCACIÓN”

Throttling

TÉCNICAS DE “OFUSCACIÓN”

Throttling

- Peticiones con un mismo patrón de por medio (cada X segundos) son un robot, así que fuera

DDoS protection (Cloudfare por ejemplo, delays)

TÉCNICAS DE “OFUSCACIÓN”

Throttling

- Peticiones con un mismo patrón de por medio (cada X segundos) son un robot, así que fuera

DDoS protection (Cloudfare por ejemplo, delays)

- Delay en la información y registro de sesión

Captcha (¿eres un robot?)

TÉCNICAS DE “OFUSCACIÓN”

Throttling

- Peticiones con un mismo patrón de por medio (cada X segundos) son un robot, así que fuera

DDoS protection (Cloudfare por ejemplo, delays)

- Delay en la información y registro de sesión

Captcha (¿eres un robot?)

- Requiere interacción avanzada (Computer visión + Interacción programática)

Ofuscamiento de classNames (nuestra guía cambiante)

TÉCNICAS DE “OFUSCACIÓN”

Throttling

- Peticiones con un mismo patrón de por medio (cada X segundos) son un robot, así que fuera

DDoS protection (Cloudfare por ejemplo, delays)

- Delay en la información y registro de sesión

Captcha (¿eres un robot?)

- Requiere interacción avanzada (Computer visión + Interacción programática)

Ofuscamiento de classNames (nuestra guía cambiante)

- Si extraemos información por classNames, autogenerarlos con hashes nos dificultaría la faena

Agente web o User-Agent

TÉCNICAS DE “OFUSCACIÓN”

Throttling

- Peticiones con un mismo patrón de por medio (cada X segundos) son un robot, así que fuera

DDoS protection (Cloudfare por ejemplo, delays)

- Delay en la información y registro de sesión

Captcha (¿eres un robot?)

- Requiere interacción avanzada (Computer visión + Interacción programática)

Ofuscamiento de classNames (nuestra guía cambiante)

- Si extraemos información por classNames, autogenerarlos con hashes nos dificultaría la faena

Agente web o User-Agent

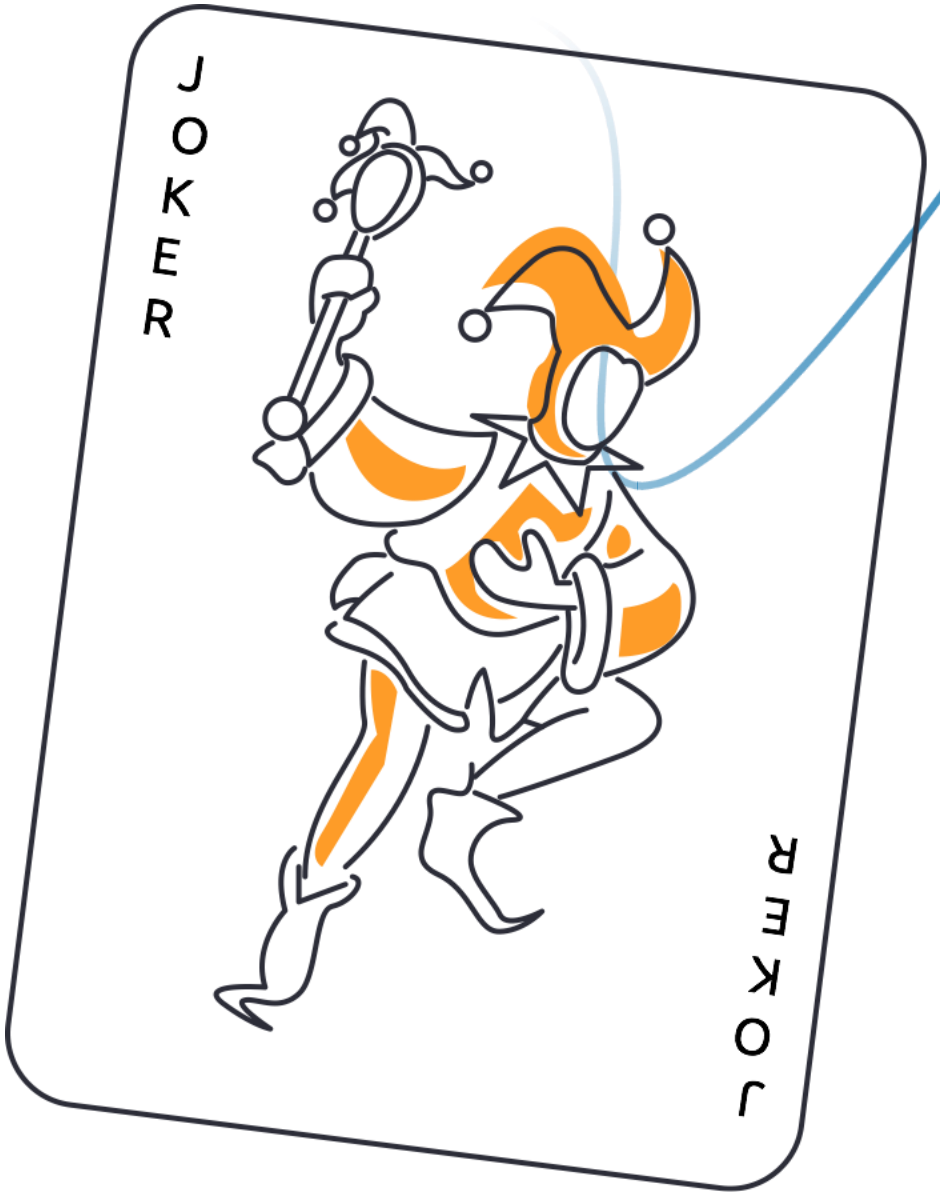
- Cabecera de peticiones, información de los navegadores

DEMO: SIN HEADLESS

Páginas más “ofuscadas”, (in)voluntariamente, y los problemas que atañen

Tiempo aproximado: **5 minutos**

Target: **ZARA**



2

TESTING

TESTING

¿Qué es el testing? ¿Y qué propósito cumple?

TESTING

¿Qué es el testing? ¿Y qué propósito cumple?

- La validación funcional de nuestro código
- Cumple el propósito de comprobar las funcionalidades una única vez

¿Es viable?

TESTING

¿Qué es el testing? ¿Y qué propósito cumple?

- La validación funcional de nuestro código
- Cumple el propósito de comprobar las funcionalidades una única vez

¿Es viable?

- Todo código puede ser testeado (y automatizado)
- Qué partes deberían testearse y mantenerse es la clave

¿QUÉ TIPOS DE TESTS CONOCÉIS?

¿Quién se ánima a enumerar tipos de tests?

TIPOS DE TESTS Y HERRAMIENTAS

Estáticos (Tipados, Clases Abstractas, Interfaces, Structs)

- IntelliSense

Unitarios

- JUnit, Mockito, Vitest, Enzyme

Integración

- JUnit, Testing Library, Vitest, Enzyme

End-to-end (e2e)

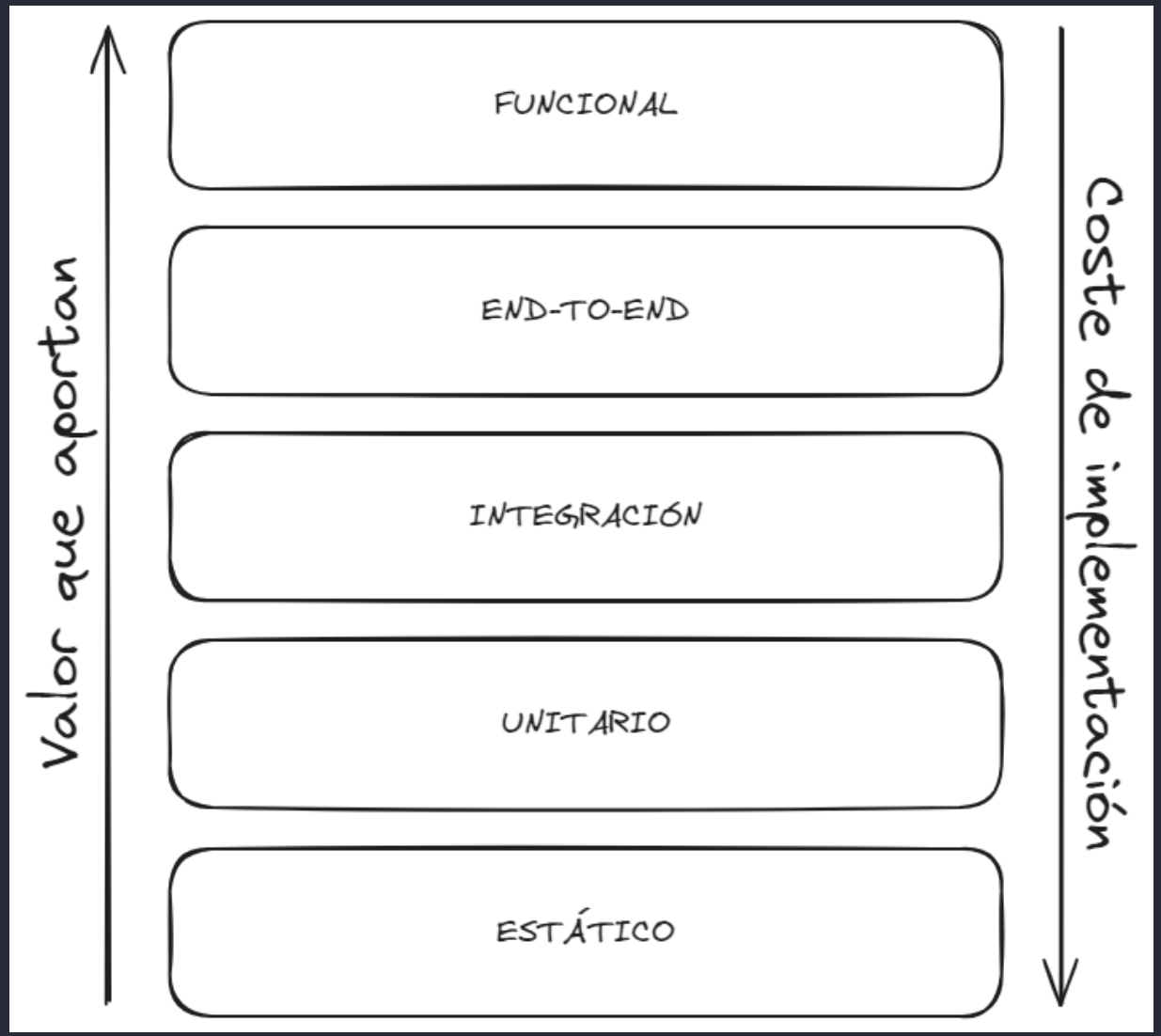
- Selenium, Cypress, PlayWright, Puppeteer, Postman

Smoke (Comprobaciones de sistema)

- ping, Kubernetes + Istio

…y unos cuantos más,

JERARQUÍA DE TESTING

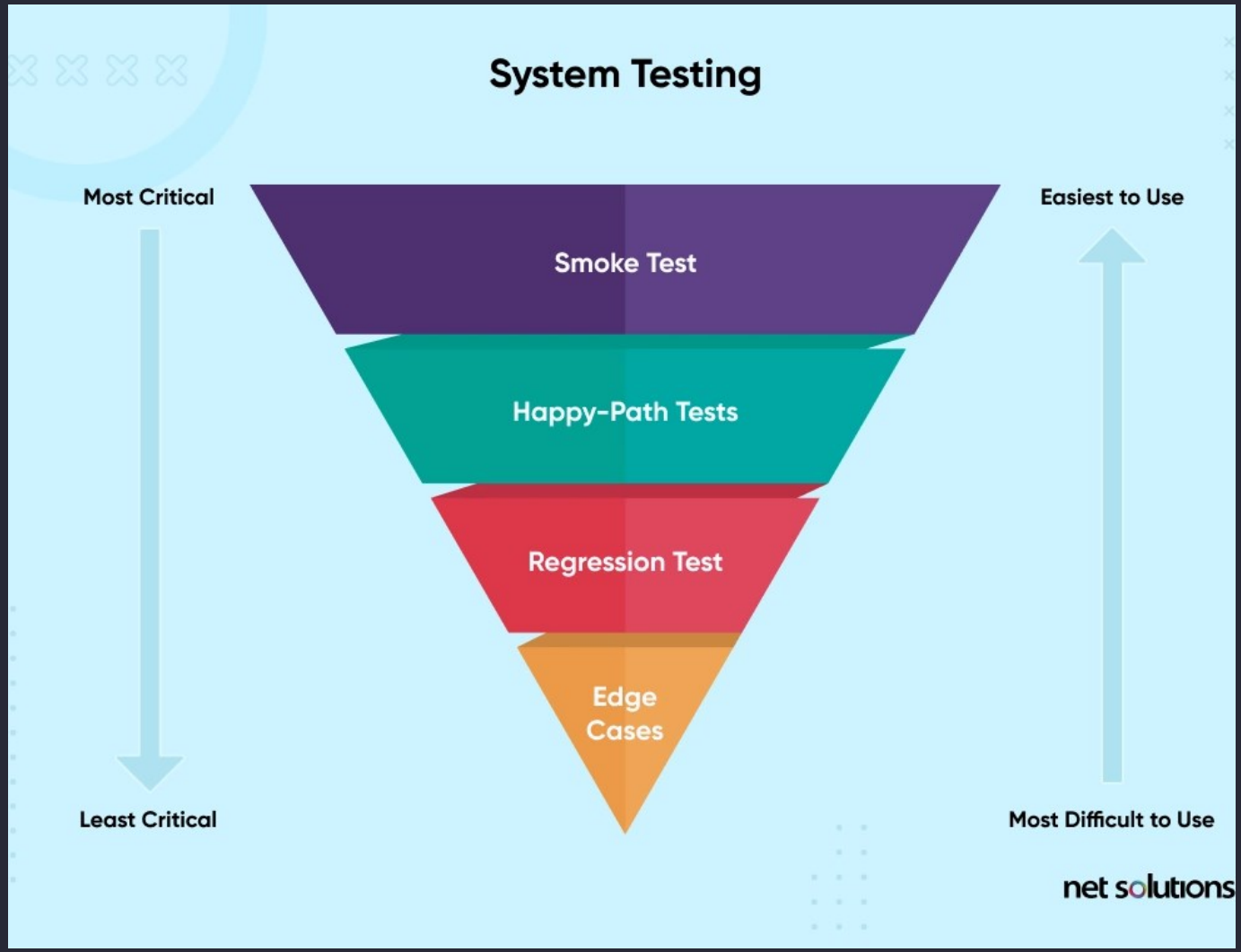


END-TO-END

Cubrir una prueba de extremo a extremo

Son las más fiables y más costosas de mantener

Qué testearíamos y qué no



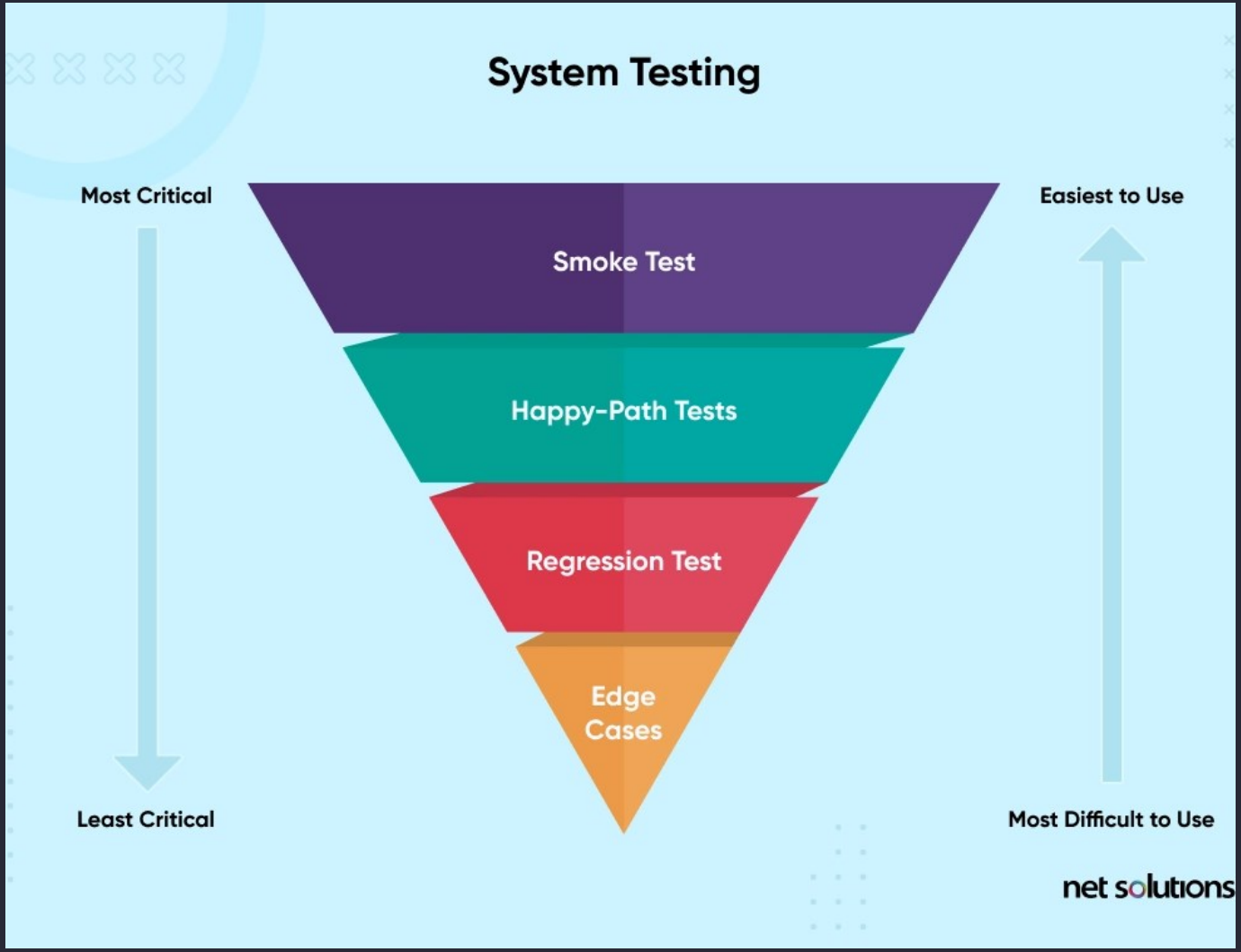
END-TO-END

Cubrir una prueba de extremo a extremo

Son las más fiables y más costosas de mantener

Qué testearíamos y qué no

- El sistema funciona (HTTP GET)
- Página de producto fiable
- Recuperar los enlaces de categorías
- Página inexistente



¿SON NECESARIOS?

Los malos tests

¿SON NECESARIOS?

Los malos tests

- Siempre será mejor no tener tests, que tests incompletos, o flaky

Mantenimiento de los tests

¿SON NECESARIOS?

Los malos tests

- Siempre será mejor no tener tests, que tests incompletos, o flaky

Mantenimiento de los tests

- Los más valiosos serían e2e, si el proyecto y/o equipo es pequeño, tal vez no compense

Empresas dedicadas

¿SON NECESARIOS?

Los malos tests

- Siempre será mejor no tener tests, que tests incompletos, o flaky

Mantenimiento de los tests

- Los más valiosos serían e2e, si el proyecto y/o equipo es pequeño, tal vez no compense

Empresas dedicadas

- Si es la fuente principal de financiación, son necesarios, e incluso obligatorios
- Si una de las páginas de las que haces seguimiento cambia
  - Mejor enterarte tú por un test que rompe
  - Que al mes y no tener información

REFERENCIA DE TESTING PARA WEB SCRAPING

https://webscraper.io/test-sites

Para los tests con web scraping, se recomienda usar las librerías que se utilizarían para un e2e, y para los tests unitarios lo mismo, en caso de Java, Junit, en el caso de Python, usar el *built-in* **assert**

RECAPITULANDO

RECAPITULANDO...

Qué hemos visto

RECAPITULANDO…

1. Defensa ante Web scraping

RECAPITULANDO…

1. Defensa ante Web scraping
2. Riesgos del Web scraping

RECAPITULANDO…

1. Defensa ante Web scraping
2. Riesgos del Web scraping
3. Testing enfocado

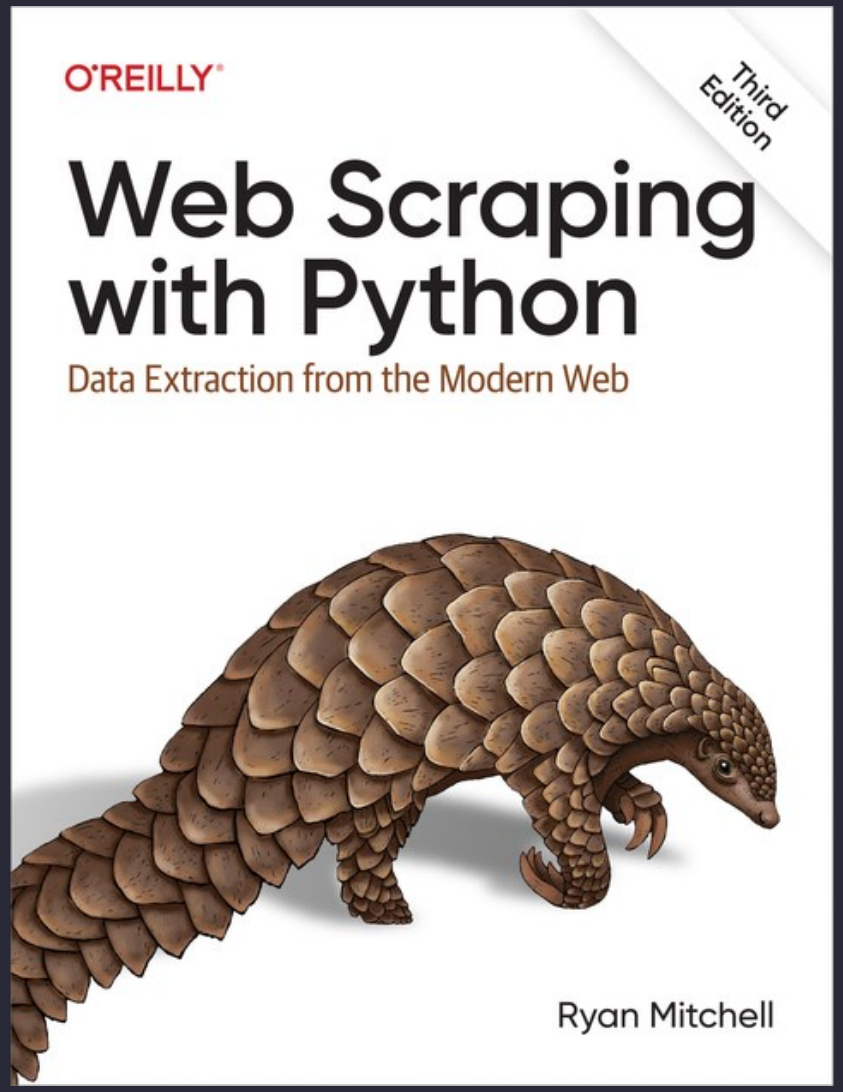
CRÉDITOS

LIBRO RECOMENDADO

https://www.amazon.com/-/es/Fernando-Rosa/dp/8409363801

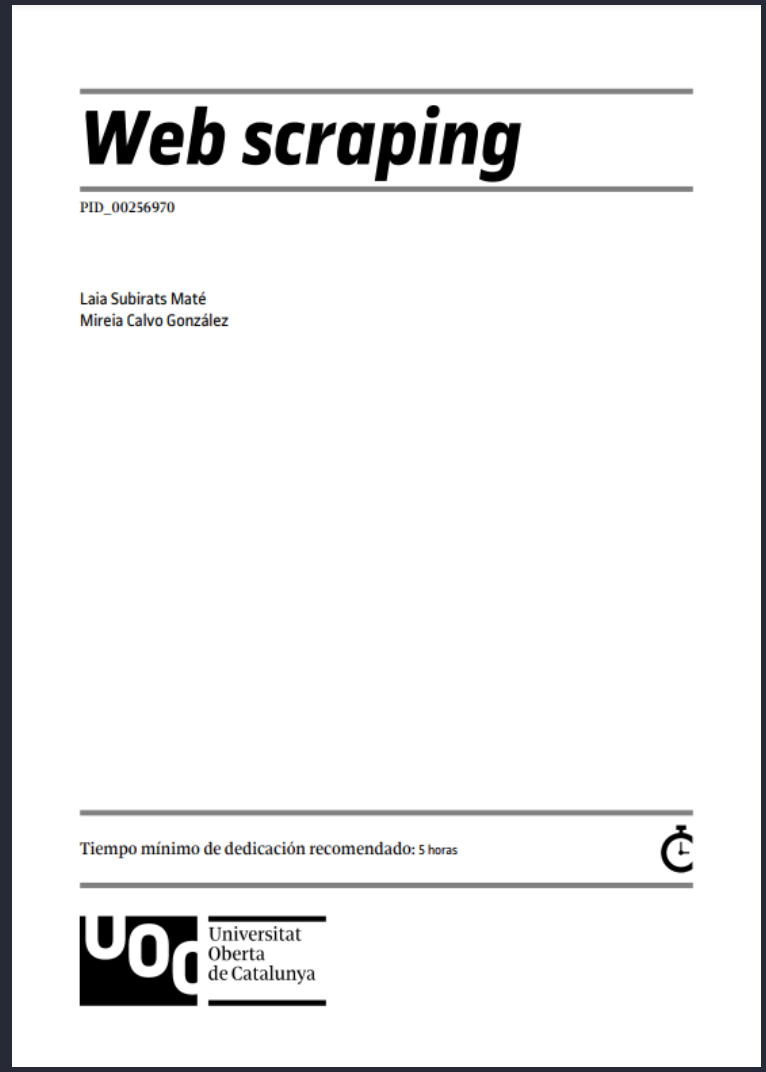
LIBRO RECOMENDADO

https://www.amazon.es/web-scraping-python-extraction-modern/dp/1098145356



LIBRO RECOMENDADO

https://openaccess.uoc.edu/bitstream/10609/147437/1/webscraping_modulo1_webscraping.pdf



Web scraping

PID_00256970

Laia Subirats Maté
Mireia Calvo González

Tiempo mínimo de dedicación recomendado: 5 horas

Universitat Oberta de Catalunya

PREGUNTAS

¡¡GRACIAS!!

About Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided everyday by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of 270,000 team members in nearly 50 countries. With its strong 50 year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fuelled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2020 global revenues of €16 billion.

Get the Future You Want | www.capgemini.com