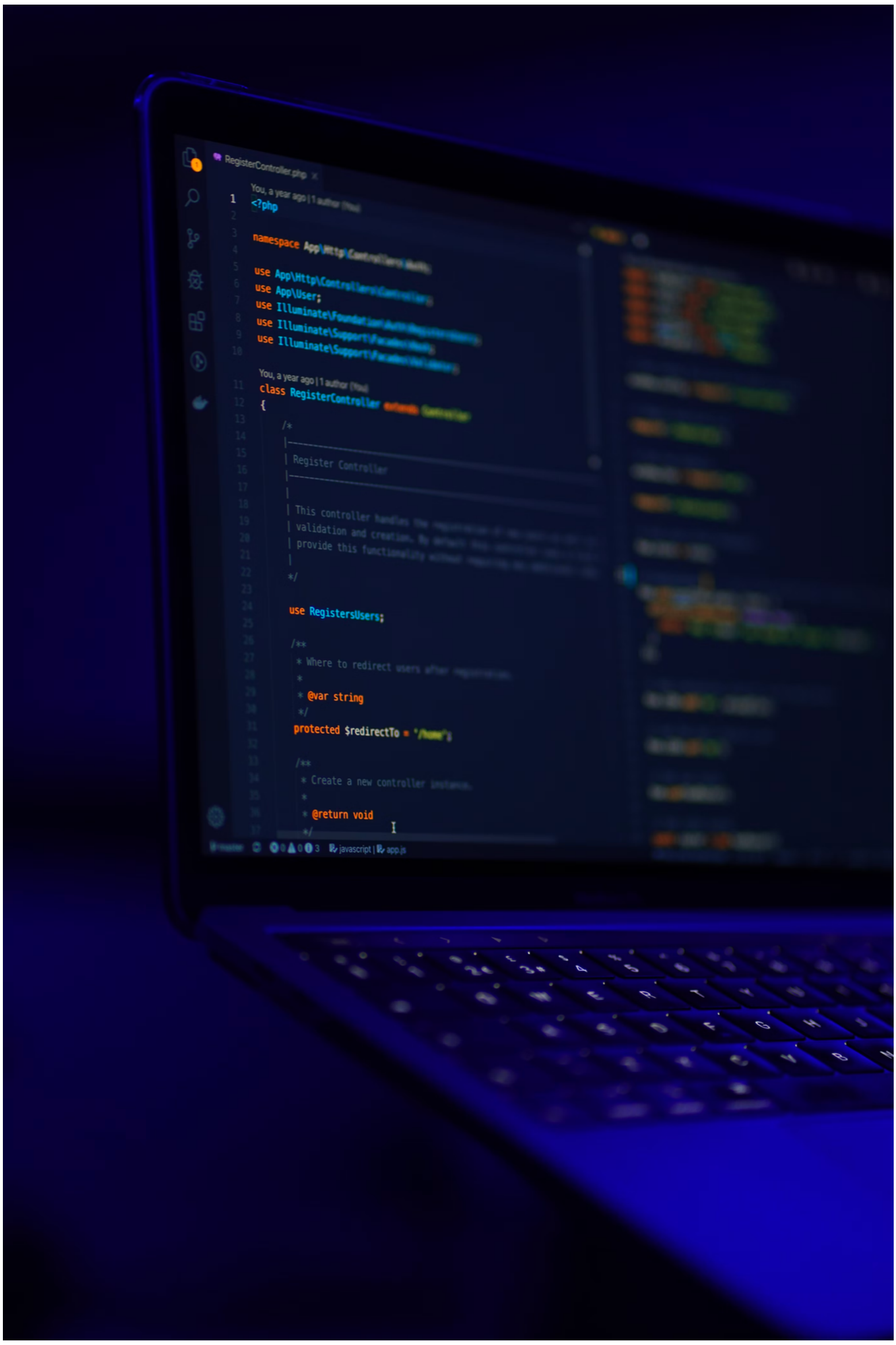
Lección 3: DML y DDL desde Python



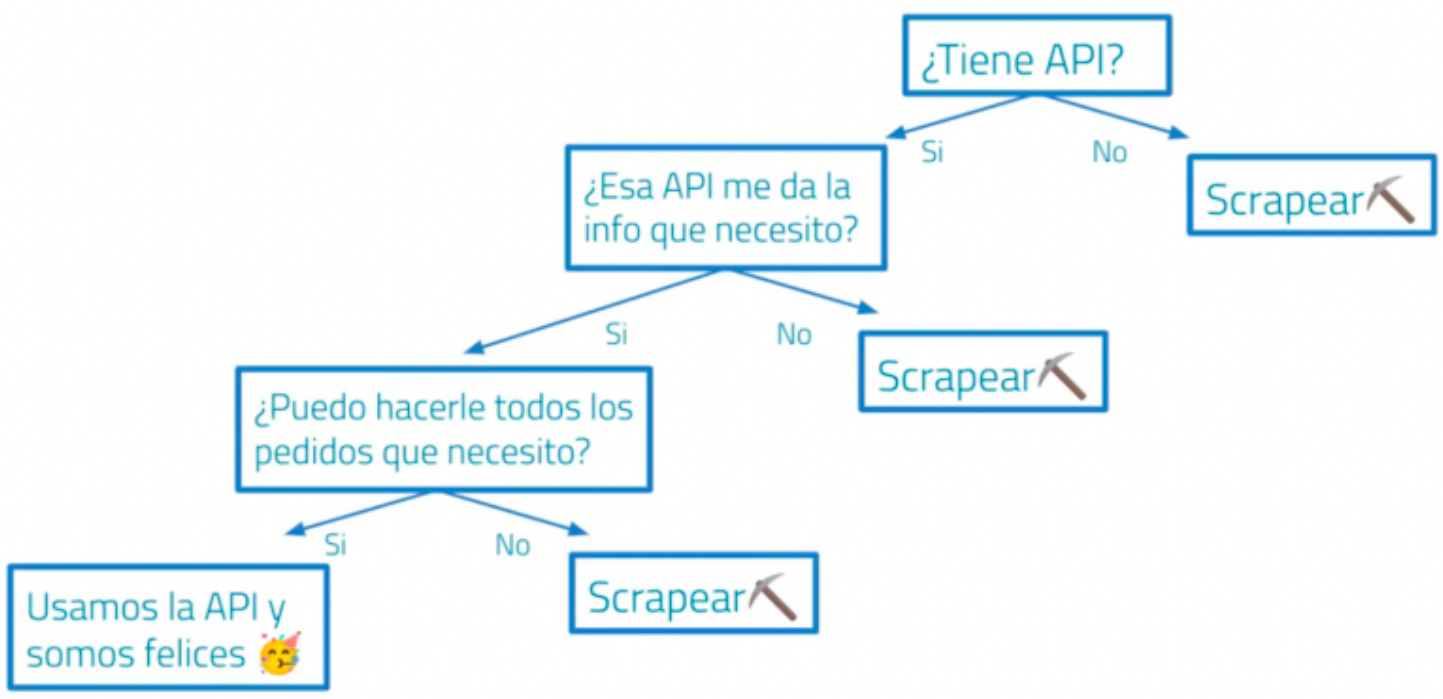
Tiempo de ejecución: 4 horas

Planteamiento de la sesión

Materiales

- Web scraping con Python: primeros pasos, tutorial y herramientas - IONOS
- Learn Web Scraping with Python and Beautiful Soup in Spanish – Course for Beginners
- ¿Qué es una API? - Explicación de interfaz de programación de aplicaciones - AWS.
- DOM - Glosario de MDN Web Docs: Definiciones de términos relacionados con la Web).
- Document Object Model - Wikipedia
- https://scrapepark.org/courses/spanish/
- web_scraping.ipynb

Para finalizar la unidad, se hará una introducción a una técnica conocido como Web Scraping, que es una técnica para extraer datos de un sitio web, que pueden no estar disponibles a través del uso de la API del desarrollador, o no contar con API, o que su uso limita la cantidad de peticiones.



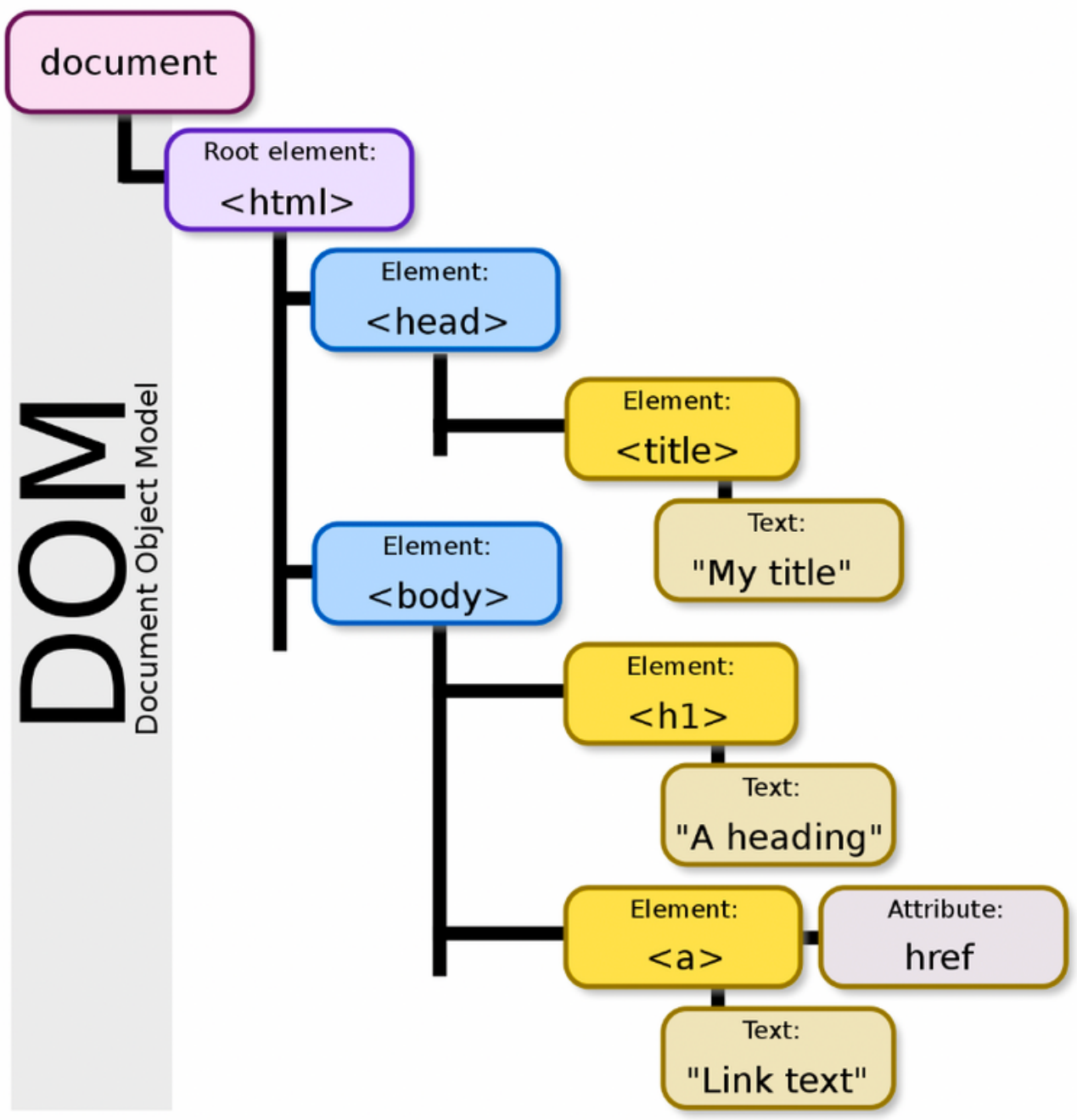
Fuente: freeCodeCamp, consultado en marzo de 2024, disponible en internet, Learn Web Scraping with Python and Beautiful Soup in Spanish – Course for Beginners

La técnica tiene, como se indica en la gráfica, una serie de condiciones en donde se recomiendo utilizar y otra en la que hacer el procedimiento de manera manual o utilizando la API del desarrollador, es más que suficiente.

Una API, muy básicamente, es un mecanismo que permite la comunicación o traspaso de información entre sistemas, definiendo formatos y protocolos, por ejemplo, desde una aplicación que yo desarrolle, puedo agregar un buscador de películas y series que me entregue la información de IMDB o si quiero agregar el clima de una ubicación específica, también puede consumir los datos desde otro sistema que ofrezca una interfaz (API) y así ahorrar mucho tiempo de desarrollo.

Python, cuenta con algunas herramientas que permiten desarrollar esta técnica para la que como prerrequisito es necesario conocer un par de elementos de la WEB, entre ellos, el más importante es el DOM.

El DOM es un modelo de documento que se carga en el navegador web y que representa el documento como un árbol de nodos, en donde cada nodo representa una parte del documento (puede tratarse de un elemento, una cadena de texto o un comentario).



Fuente: Document Object Model - Wikipedia, consultado en marzo de 2024, disponible en internet, Document Object Model - Wikipedia

Lo que se presenta en la imagen, son las etiquetas HTML de una página web sencilla que, a partir del documento que lee e interpreta el navegador web, construyen el árbol del DOM.

Un concepto adicional para funcionar con la técnica, es conocer cómo funciona la web, de manera general, cuando un cliente, ya sea un navegador web u otro sistema, hace una solicitud a un recurso en la web, lo que genera es una petición que debe atender un servidor, a ese petición se le conoce como el request. El servidor web genera una respuesta conocida como response, que puede llevar lo que pedía el usuario en el request o un mensaje de error.

Todo esto sucede bajo las definiciones de un protocolo conocido como HTTP, que cuenta con mensajes y códigos que resumen el estado de la response, respuestas exitosas, problemas en la petición o problemas en el servidor.

Fuente: freeCodeCamp, consultado en marzo de 2024, disponible en internet, Learn Web Scraping with Python and Beautiful Soup in Spanish – Course for Beginners

El DOM se forma con los archivos HTML pero también puede contener información de CSS o de JS, insertando bloques al árbol después de una determinada acción del usuario.

Como resumen, las peticiones que hacemos a una página web, se compone de recursos que nuestro navegador interpreta y presenta como el sitio web que solicitamos, esos recursos son archivos HTML, CSS y JS que contienen toda la información necesaria para atender el request.

De esas respuestas, principalmente el HTML, es donde, utilizando Python y una librería llamada Beautifulsoup, vamos a extraer la información.

Ahora sí, manos a la obra, a continuación, se presenta un notebook the jupyter desarrollado por freeCodeCamp y Gustavo Juantorena donde se exploran las herramientas indicadas, una breve explicación de cada una, su uso y los resultados de aplicar la técnica sobre el sitio web de pruebas.

Notebook: web_scraping.ipynb
Web: https://scrapepark.org/courses/spanish/