



Instituto Superior Universitario Tecnológico del Azuay
Tecnología Superior en Big Data

Guía Práctica N°1 - Captura y adquisición de datos

Integrantes:

Eduardo Mendieta
Freddy Montalván

Materia:

Introducción a Big Data

Docente:

MSc. Ing. Carmen Tacuri Vintimilla

Ciclo:

Primer Ciclo

Fecha:

29 de julio de 2024

Periodo Académico:

Abril 2024 - Agosto 2024

Índice

1.	Introducción	2
2.	Objetivos	3
2.1.	Objetivo general	3
2.2.	Objetivos específicos	3
3.	Paso a paso	4
3.1.	Fórmula ImportHTML	4
3.2.	Table capture	6
3.3.	Tabula	7
3.4.	Hunter.io	9
3.5.	Octoparse	11
4.	Conclusiones y recomendaciones	14
4.1.	Conclusiones	14
4.2.	Recomendaciones	14
5.	Bibliografía	15

Guía Práctica N°1 - Captura y adquisición de datos

1. Introducción

En el análisis de datos, la captura y adquisición de datos juegan un papel crucial al proporcionar la base sobre la cual se construyen los insights y las decisiones informadas. Estos procesos comprenden la identificación, recolección e integración de información relevante, que puede provenir de una variedad de fuentes, como sistemas, sensores, encuestas o archivos. La captura se enfoca en recoger datos brutos, mientras que la adquisición se centra en organizar y almacenar estos datos para su procesamiento posterior.

Una técnica destacada en la captura y adquisición de datos es el web scraping, que permite la extracción automatizada de información de sitios web. Utilizando scripts o herramientas especializadas, el web scraping facilita la obtención de datos específicos de páginas web, incluso cuando estos datos no están disponibles en formatos estructurados o accesibles. Esta técnica resulta invaluable para recolectar grandes volúmenes de información de manera eficiente, facilitando así el análisis de tendencias, la investigación de mercado y la toma de decisiones basada en datos actualizados.

Entre las herramientas más útiles para el web scraping se encuentra IMPORTHTML en Google Sheets, que permite importar tablas o listas desde páginas web y mantener los datos actualizados automáticamente. Adicionalmente, Table Capture es una extensión para Chrome que simplifica la extracción de datos de tablas HTML y su exportación a formatos como Google Sheets, Excel o CSV. Para extraer datos de tablas en archivos PDF, Tabula es una aplicación de escritorio compatible con múltiples sistemas operativos que convierte estos datos en formatos editables como CSV o Excel. Hunter.io, por su parte, se especializa en la recopilación de correos electrónicos de sitios web, facilitando la ampliación de listas de contactos. Finalmente, Octoparse es una herramienta de web scraping intuitiva y con una prueba gratuita, que permite a usuarios sin experiencia en programación extraer datos de cualquier sitio web con facilidad y exportarlos en diversos formatos.

Estas herramientas y técnicas son fundamentales para realizar una captura y adquisición de datos efectiva, asegurando que la información recolectada sea precisa, completa y lista para su análisis y aplicación.

2. Objetivos

2.1. Objetivo general

Explorar las herramientas de scraping presentadas en la guía, adquiriendo un conocimiento básico sobre su uso para la captura y adquisición de datos.

2.2. Objetivos específicos

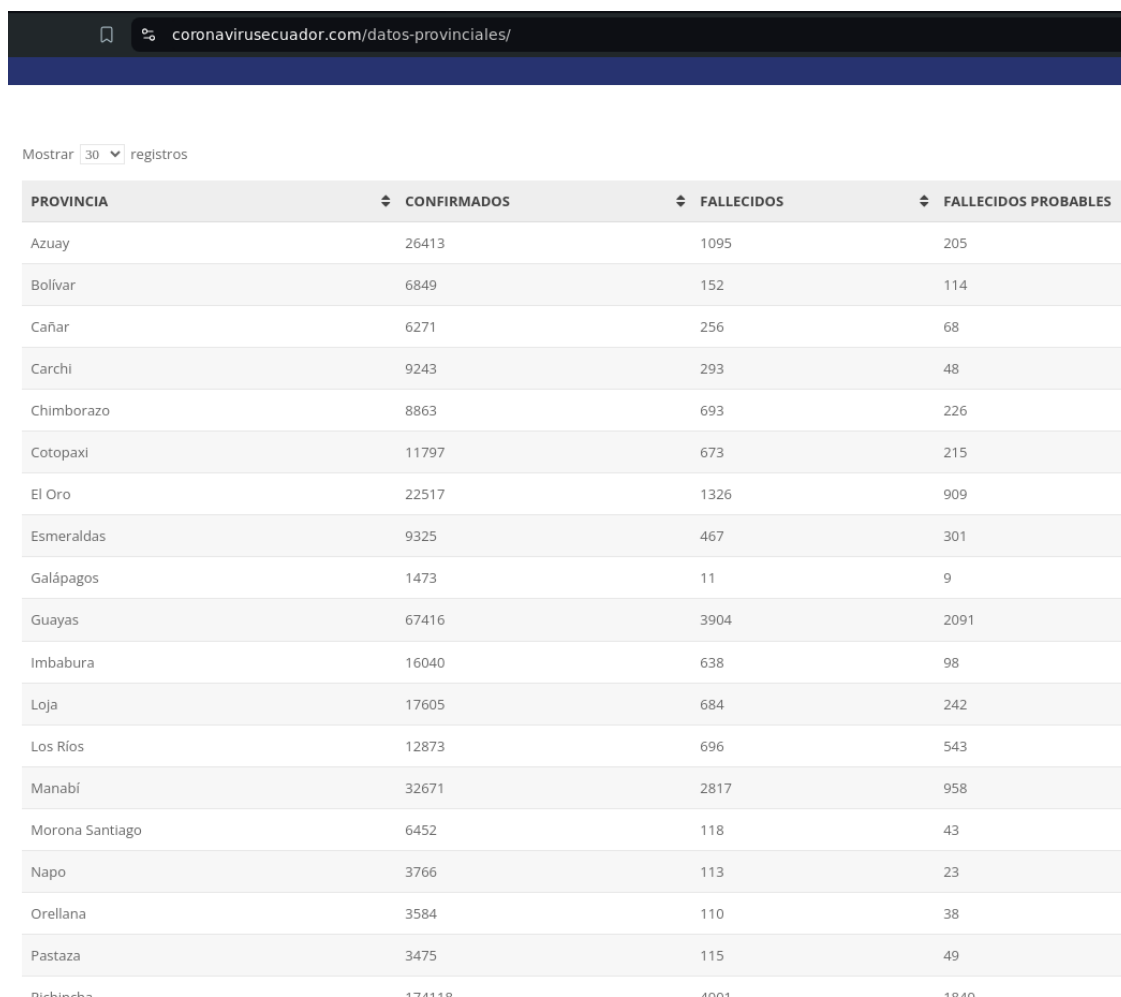
- Extraer los datos de la web oficial con información actual sobre el COVID-19 utilizando las herramientas propuestas en la guía.
- Examinar las interfaces que ofrecen cada una de estas aplicaciones y extensiones para la captura de datos.
- Concluir sobre las ventajas, similitudes y diferencias entre estas herramientas.

3. Paso a paso

3.1. Fórmula ImportHTML

En Google Sheets, la función *IMPORTHTML* permite importar tablas o listas desde páginas web externas, actualizando automáticamente los datos en la hoja de cálculo cuando se realizan cambios en la fuente original. La sintaxis de la función es *=IMPORTHTML(url, consulta, índice)*, donde *url* es la dirección del sitio web, *consulta* especifica si se desea importar una tabla o una lista, y *índice* indica el número de la tabla o lista a importar.

Por ejemplo, para extraer datos estadísticos del Covid-19 de una página específica, se puede usar la fórmula *=IMPORTHTML("https://www.coronavirusecuador.com/datos-provinciales";table";1)* para capturar la primera tabla disponible en la web, y los datos se mostrarán automáticamente en la hoja de cálculo.



Mostrar 30 registros

PROVINCIA	CONFIRMADOS	FALLECIDOS	FALLECIDOS PROBABLES
Azuay	26413	1095	205
Bolívar	6849	152	114
Cañar	6271	256	68
Carchi	9243	293	48
Chimborazo	8863	693	226
Cotopaxi	11797	673	215
El Oro	22517	1326	909
Esmeraldas	9325	467	301
Galápagos	1473	11	9
Guayas	67416	3904	2091
Imbabura	16040	638	98
Loja	17605	684	242
Los Ríos	12873	696	543
Manabí	32671	2817	958
Morona Santiago	6452	118	43
Napo	3766	113	23
Orellana	3584	110	38
Pastaza	3475	115	49
Dakshin	174110	4001	1040

Figura 1: Datos estadísticos del Covid- 19

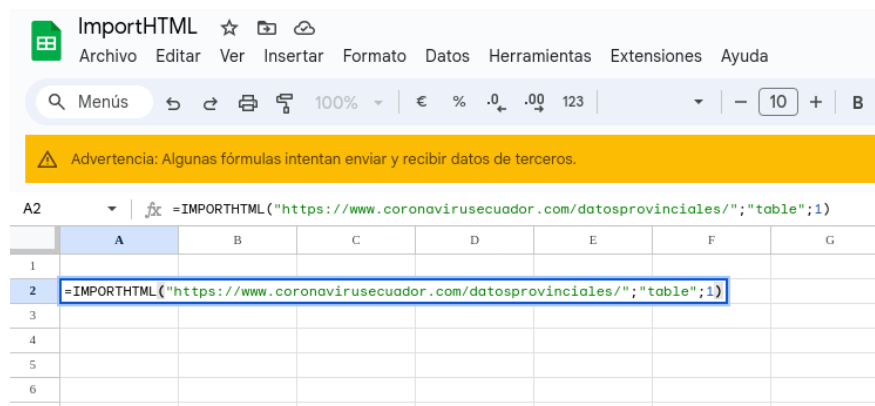


Figura 2: Fórmula para importar los datos del Covid- 19 en la hoja de cálculo de Google

ImportHTML

Archivo Editar Ver Insertar Formato Datos Herramientas Ex

Menús 100% 123 P

A3 =IMPORTHTML("https://www.coronavirusecuador.com/datosprovinciales/";"table";1)

	A	B	C	D	E
1					
2	PROVINCIA	CONFIRMADOS	FALLECIDOS	FALLECIDOS PROBABLES	
3	Azuay	26413	1095	205	
4	Bolívar	6849	152	114	
5	Cañar	6271	256	68	
6	Carchi	9243	293	48	
7	Chimborazo	8863	693	226	
8	Cotopaxi	11797	673	215	
9	El Oro	22517	1326	909	
10	Esmeraldas	9325	467	301	
11	Galápagos	1473	11	9	
12	Guayas	67416	3904	2091	
13	Imbabura	16040	638	98	
14	Loja	17605	684	242	
15	Los Ríos	12873	696	543	
16	Manabí	32671	2817	958	
17	Morona Santiago	6452	118	43	
18	Napo	3766	113	23	
19	Orellana	3584	110	38	
20	Pastaza	3475	115	49	
21	Pichincha	174118	4901	1849	
22	Santa Elena	4446	476	443	
23	Sto. Domingo Tsá	11873	783	287	
24	Sucumbíos	5664	221	100	
25	Tungurahua	15014	877	457	
26	Zamora Chinchipe	2972	126	35	
27	Total General	480720	21446	9351	

Figura 3: Datos importados en la hoja de cálculo de Google

3.2. Table capture

Table Capture es una extensión para el navegador Chrome que facilita la extracción de datos de tablas HTML en páginas web, permitiendo su exportación a formatos como Google Sheets, Excel o CSV, similar a la función *IMPORTHTML*.

Tras instalar la extensión y acceder a la página de datos del Covid-19 en Ecuador, se puede utilizar la opción *Captura de tabla* del menú contextual para copiar la información al portapapeles o abrirla directamente en Google Sheets. Aunque las funciones avanzadas de exportación y almacenamiento en la nube están disponibles solo en la versión Pro, la versión gratuita permite copiar y modificar los datos antes de su uso.

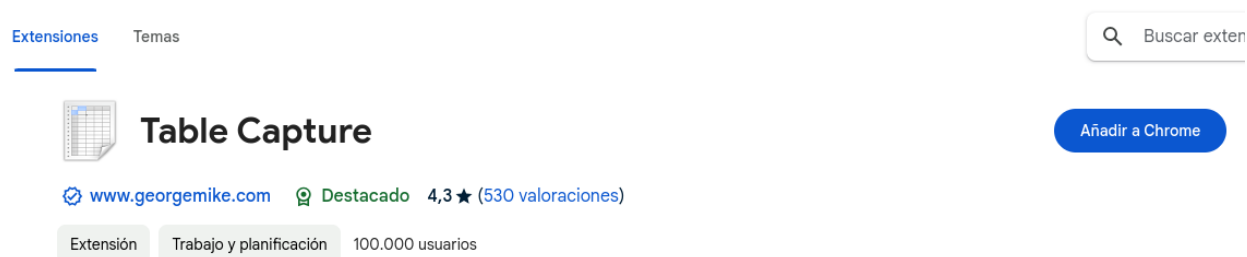


Figura 4: Añadiendo Table Capture a Chrome

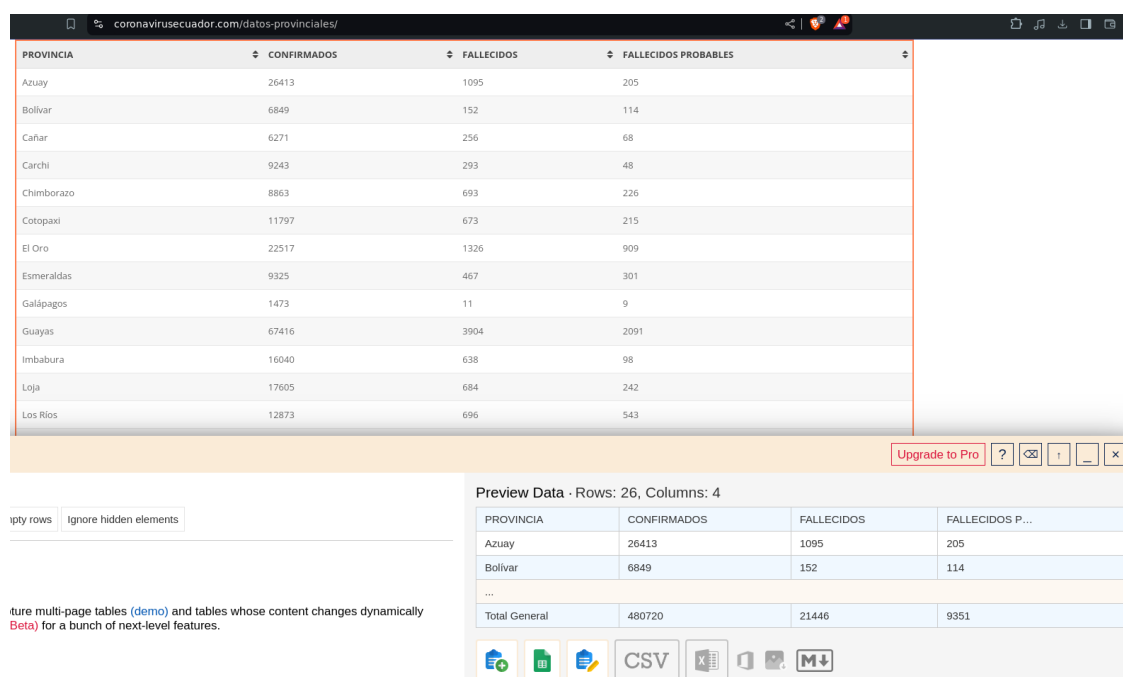


Figura 5: Seleccionando los datos y exportando a una hoja de cálculo de Google

	A	B	C	D	E
1	PROVINCIA	CONFIRMADOS	FALLECIDOS	FALLECIDOS PROBABLES	
2	Azuay	26413	1095	205	
3	Bolívar	6849	152	114	
4	Cañar	6271	256	68	
5	Carchi	9243	293	48	
6	Chimborazo	8863	693	226	
7	Cotopaxi	11797	673	215	
8	El Oro	22517	1326	909	
9	Esmeraldas	9325	467	301	
10	Galápagos	1473	11	9	
11	Guayas	67416	3904	2091	
12	Imbabura	16040	638	98	
13	Loja	17605	684	242	
14	Los Ríos	12873	696	543	
15	Manabí	32671	2817	958	
16	Morona Santiago	6452	118	43	
17	Napo	3766	113	23	
18	Orellana	3584	110	38	
19	Pastaza	3475	115	49	
20	Pichincha	174118	4901	1849	
21	Santa Elena	4446	476	443	
22	Sto. Domingo Tsá	11873	783	287	
23	Sucumbíos	5664	221	100	
24	Tungurahua	15014	877	457	
25	Zamora Chinchipe	2972	126	35	
26	Total General	480720	21446	9351	
27					
28					

Figura 6: Pegando los datos obtenidos con Table Capture

3.3. Tabula

Tabula es una aplicación de escritorio compatible con Windows, Mac OSX y Linux, que simplifica la extracción de datos de tablas en archivos PDF a formatos editables como CSV o Microsoft Excel, siendo especialmente útil en el periodismo de datos. Para instalarla, se debe descargar el archivo correspondiente desde [su pagina web oficial](#), seguir las instrucciones específicas según el sistema operativo. Tras la instalación, se carga el archivo PDF con la tabla deseada en Tabula, se usa la opción *Autodetect Tables* para identificar y extraer los datos, y se puede revisar la información extraída antes de exportarla a un archivo CSV o Excel. La aplicación permite también la visualización y modificación de los datos antes de su exportación.

Download & Install Tabula

Windows & Linux users will need a copy of [Java](#) installed. You can [download Java here](#). (Java is included in the Mac version.)

1. Download the version of Tabula for your operating system:

- **Windows:** [tabula-win.zip](#)
- **Mac OS X:** [tabula-mac.zip](#)
- **Linux/Other:** [tabula-jar.zip](#), view README.txt inside for instructions

Figura 7: Descarga de Tabula

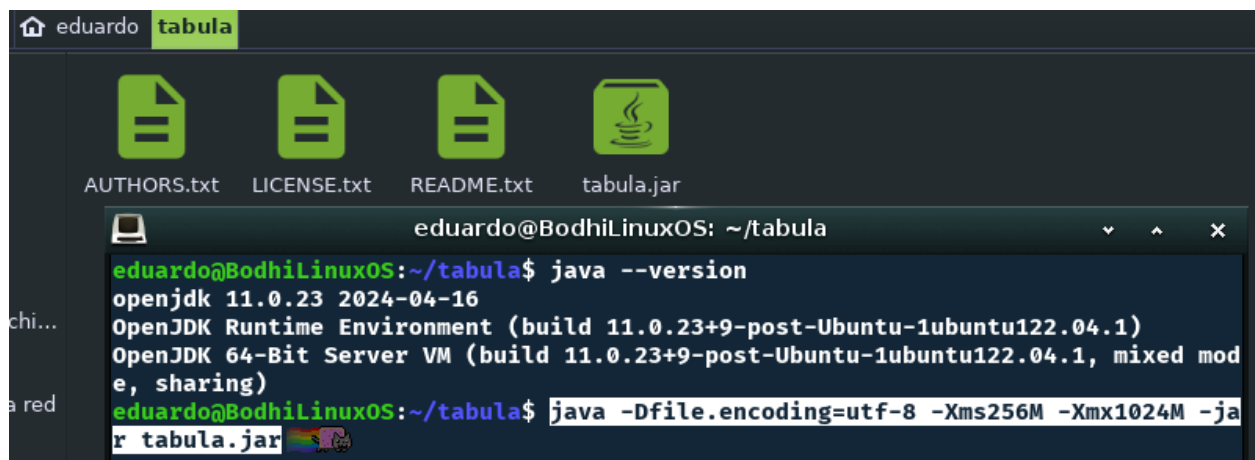


Figura 8: Ejecución de Tabula

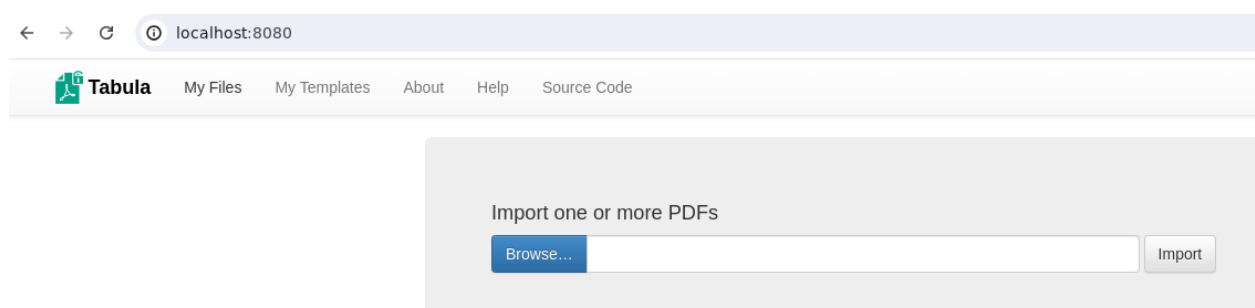


Figura 9: Importando PDF con datos sobre el Covid-19

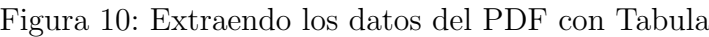


Figura 11: Datos importados en formato CSV

Hunter.io es una herramienta en línea diseñada para la recopilación de correos electrónicos a partir de sitios web, funcionando como un buscador especializado en encontrar direcciones de correo electrónico. Ideal para ampliar listas de contactos empresariales y realizar campañas de marketing por correo, *Hunter.io* permite buscar emails asociados a un dominio específico, como *colinealcorp.com*. Al ingresar el dominio, se muestra una lista de correos electrónicos encontrados en la web, con detalles adicionales disponibles al hacer clic en cada ítem. Aunque el uso básico es anónimo y gratuito, también es posible crear una cuenta para exportar los datos en formato CSV y acceder a funcionalidades adicionales.

A screenshot of a web application titled "Domain Search". The main heading reads "Find email addresses from any company name or website." Below this, there are three tabs: "Find email by company" (which is selected), "Find email by name", and "Verify email". In the "Find email by company" tab, the input field contains the text "colinealcorp.com" with a red squiggly underline. To the right of the input field is a button labeled "Find email addresses".

Figura 12: Buscando el dominio *colinealcorp.com*

7 results for your search	
<div> <div></div> <div>@colinealcorp.com</div> <div>94%</div> </div>	20+ sources ▾
http://comprarenpanama.com/muebles	Mar 08, 2018
http://marcasecuador.club/colineal-corporation	Nov 16, 2022
http://colineal.pa/pages/nuestros-locales	Feb 13, 2023
http://colineal.pa/blogs/tips-e-ideas	Feb 19, 2023
http://infurma.com/company/furniture/frames_and_moldings/colineal/181455.html	Aug 28, 2023
http://colineal.pa/collections/espejos-y-tocadores	Oct 16, 2023
http://cipem.org.ec/afiliados	Nov 03, 2023
http://colineal.pa/collections/sillones-reclinables	Nov 13, 2023
http://infurma.com/company/furniture/chest_of_drawers/colineal/181455.html	Nov 29, 2023
http://colineal.pa/collections/vitrinas	Feb 05, 2024
http://colineal.com/products/alfombra-fiji	Feb 16, 2024
http://colineal.com/collections/servicio-de-mesa	Feb 25, 2024
http://infurma.com/company/furniture/accessories/colineal/181455.html	Mar 02, 2024
http://colineal.com/products/cama-ibiza	May 25, 2024
http://colineal.com/products/bowl-de-servir-shell-c-negro	May 25, 2024
http://colineal.com/collections/todos-los-colchones/products/colchon-cambiadorl-cuna-cambiador-anghel...	Jun 15, 2024

Figura 13: lista de correos asociados al domino



Donde comprar muebles en Panamá

Esta nación es tendencia y referencia no solo por sus enormes rascacielos de última generación en cuanto a infraestructura y diseño se refiere, sino también por el canal de Panamá atrae a comerciantes de todo tipo lo que facilita la importación de mercancía de todos los continentes y ponerlas a disposición de sus coterráneos. La empresa de bienes inmuebles **para el hogar u oficinas** comerciales y negocios no se queda atrás ofreciendo a sus clientes la posibilidad de **comprar muebles en Panamá** desde modelos minimalistas hasta las mas sobrias colecciones en una infinita gama de tiendas a tu disposición tanto en su ciudad capital como en sus provincias a nivel nacional.

En las **tiendas de muebles** puedes encontrar desde una almohada o un adorno para tu mesa de centro como fascinantes juegos de muebles para sala y comedor asesorados e instalados por personal altamente calificado en diseños de interiores y exteriores, no te quedes atrás, por eso te ofrecemos una selecta lista de las seis tiendas más reconocidas en este ramo sin ningún orden específico:

Contenido [\[Ocultar\]](#)

Figura 14: Desplegando uno de los items

3.5. Octoparse

Octoparse es una herramienta de web scraping gratuita y fácil de usar, diseñada para usuarios sin experiencia en programación. Permite extraer datos de cualquier sitio web mediante su función de detección automática, evitando los complicados procesos de construcción de crawlers. Los datos extraídos pueden exportarse en formatos como Excel, CSV, JSON, HTML, o a bases de datos. Para utilizar Octoparse, se debe registrar una cuenta en su página web y descargar el cliente de escritorio desde un archivo comprimido.

Tras la instalación, se accede con las credenciales registradas y se pueden usar plantillas predefinidas o personalizadas para capturar datos de diversas fuentes, como Google Search. Al ejecutar una búsqueda, como sobre el Covid-19 en Ecuador, la aplicación realiza la extracción y permite exportar los datos obtenidos. Aunque la prueba gratuita requiere una tarjeta de crédito, la guía permite comprender el funcionamiento de la herramienta sin problemas.

Paso 1/3

Háblanos de ti para sugerencias precisas

¿Cuál es tu nombre de usuario preferido?

eduardo

¿Qué te trae por aquí hoy?

Mi trabajo

Educación

Uso personal

posición actual?

Estudiante

Docente

Otros

Figura 15: Creando una cuenta gratuita en Octoparse



Figura 16: Descarga de la aplicación de escritorio

Figura 17: Seleccionando el motor de búsqueda

Figura 18: Solicitud de ingreso de tarjeta

4. Conclusiones y recomendaciones

4.1. Conclusiones

1. La función IMPORTHTML de Google Sheets permite importar y actualizar automáticamente datos de tablas o listas desde páginas web, integrando los datos directamente en una hoja de cálculo sin necesidad de herramientas adicionales. Esta función es ideal para usuarios que desean mantener los datos sincronizados con la fuente web en tiempo real, facilitando el análisis y la visualización de información directamente dentro de Google Sheets.
2. Table Capture, como extensión para el navegador Chrome, ofrece una forma sencilla de copiar datos de tablas visibles en páginas web y exportarlos a formatos como Google Sheets, Excel o CSV. Aunque la versión gratuita proporciona funcionalidades básicas, la versión Pro amplía las opciones y soporta características avanzadas. Esta herramienta es útil para usuarios que prefieren una solución rápida y directa para capturar datos desde la web sin tener que realizar configuraciones complejas.
3. Por otro lado, Tabula está diseñada específicamente para extraer datos de tablas dentro de archivos PDF, convirtiéndolos en formatos editables como CSV o Excel. A diferencia de las otras herramientas mencionadas, Tabula trabaja con documentos locales en lugar de datos web, lo que la hace esencial para quienes necesitan transformar contenido de PDF en datos estructurados para su análisis o procesamiento.
4. Hunter.io se centra en la búsqueda y recopilación de correos electrónicos desde sitios web, facilitando la creación de listas de contactos para marketing y otras aplicaciones. Mientras que la funcionalidad básica es gratuita, las características más avanzadas y la capacidad para exportar datos requieren una suscripción paga. Es una herramienta especializada para quienes buscan expandir y gestionar listas de contactos empresariales a partir de información disponible en la web.
5. Finalmente, Octoparse proporciona una solución completa para el web scraping, permitiendo la extracción de datos de cualquier sitio web y exportándolos en diversos formatos como Excel, CSV, JSON o bases de datos. Su interfaz amigable y capacidades de detección automática la hacen accesible para usuarios sin experiencia en programación. Sin embargo, la versión gratuita puede estar limitada en cuanto a funcionalidades y puede requerir una tarjeta de crédito para pruebas extensivas, lo que la diferencia de otras herramientas más simples o específicas.

4.2. Recomendaciones

1. Como única recomendación, se sugiere buscar herramientas de código abierto como alternativas a algunas de las herramientas propuestas en la guía, ya que estas pueden tener limitaciones en comparación con sus versiones de pago.

5. Bibliografía

- Guía practica proporcionada por la docente.