risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 16/09/2023
Aceitação/Acceptance: 19/11/2023

Web scraping: Uso de plataformas de extracción de datos aplicadas a un sitio web sobre profesiones en México

Felipe Cocón[1], Dámaris Pérez-Cruz[1], José Ángel Pérez-Rejón[1], Patricia Zavaleta-Carrillo[1], Ulises Barradas-Arenas[1], Rubí Gómez-Ramón[1], José Alonso Pérez Cruz[1].

{jcocon, dperez, japerez; pzavaleta; ubarradas; rgramon; japcruz}@pampano.unacar.mx

[1] Facultad de Ciencias de la Información, Universidad Autónoma del Carmen, 24180, Campeche, México.

DOI: 10.17013/risti.52.61–73

Resumen: Este artículo realiza una revisión exhaustiva de las principales herramientas de *web scraping* disponibles en el mercado y compara sus características y funcionalidades. Se selecciona una herramienta específica para demostrar su uso en la obtención de datos sobre porcentajes de egresados en diversas carreras en México, así como la distribución relacionada con género y salarios en varios estados del país. El artículo tiene como objetivo principal ilustrar cómo se pueden recopilar datos mediante una herramienta de extracción. Además, se destaca la importancia de acceder a fuentes de datos confiables y se proporciona una descripción detallada del proceso de extracción de datos utilizando la herramienta WebHarvy. En última instancia, se destaca la importancia del *web scraping* como una técnica poderosa y ética profesional para recopilar datos valiosos de la web de manera eficaz y responsable.

Palabras-clave: Educación; extracción; laboral; mercado; scraping.

Web scraping: Use of data extraction platforms applied to a website about professions in Mexico

Abstract: This article provides a thorough review of the main web scraping tools available on the market and it is comparing their features and functionalities. A specific tool is selected to demonstrate it is use in obtaining data on percentages of graduates in various careers in Mexico, as well as the distribution related to gender and salaries in several states of the country. The main objective of the article is to illustrate how data can be collected using a data extraction tool. Additionally, the importance of accessing reliable data sources is highlighted and a detailed description of the data extraction process using the WebHarvy tool is provided. Ultimately, it is highlighting the importance of web scraping as a powerful technique and professional ethical to collect valuable data from the web to effectively and responsibly.

Keywords: Education; extraction; employment; scraping; professions.

1. Introducción

Realizar una búsqueda de información en la *World Wide Web* se ha vuelto cada vez más problemático, esto debido a la creciente disponibilidad y diversidad de documentos que proporciona, ya el proceso de revisión de la información, el análisis y la extracción de manera manual demanda tiempo de modo que, se vuelve un desafío para los usuarios de la web.

El *web scraping* es una técnica que consiste en extraer o raspar datos de sitios web de forma automática mediante scripts o programas (Sinche & Torres, 2021, Mitchel, 2018). Aplicar esta técnica ofrece múltiples aplicaciones y beneficios, como la recopilación de información relevante, la automatización de tareas repetitivas, el ahorro de tiempo y dinero, y la generación de contenidos de calidad (Ribeiro et al., 2022)

Existen diversas herramientas de web scraping que facilitan la extracción de datos de la web sin requerir conocimientos técnicos. Estas herramientas suelen ofrecer una interfaz gráfica que permite seleccionar y nombrar los elementos deseados, así como opciones para guardar y exportar los datos en diferentes formatos.

En este sentido, el *Web Scraping* (raspado) permite extraer (*scrapear*) y recopilar información de páginas web y PDF de forma automatizada. Este procedimiento funciona a través del uso de programas o scripts, también conocidos como *scrapers*, capaces de "navegar por múltiples sitios web" y así "identificar y extraer información relevante de acuerdo con criterios preestablecidos" (Kinsta, 2022). Como se aprecia en la Figura 1 la información que se extrae se recopila y exporta a un formato que sea útil para el usuario.

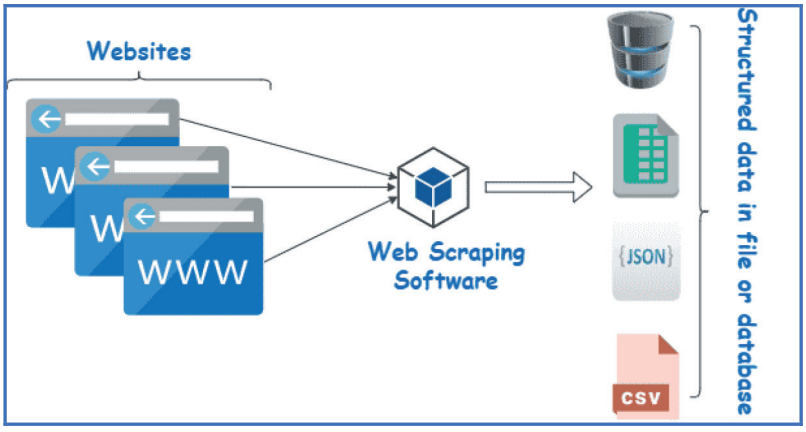

Figura 1 – Representación de Web scraping, fuente: Obtenido de (Kinsta, 2022).

El campo de aplicación para esta técnica es amplio, entre los usos comunes se mencionan los siguientes:

- Investigación de mercado para recopilar datos de precios, características de productos y reseñas de productos en línea, para realizar un análisis más detallado de sus competidores.

- Extraer opiniones y comentarios de usuarios en redes sociales, foros y blogs para evaluar la percepción pública de una marca, producto o servicio.
- Recopilar información de contacto de posibles clientes en directorios o sitios web de empresas, para estudiar tendencias del mercado, demanda y competidores mediante la recopilación y análisis de información en línea.
- Noticias y contenido.
- También se emplea para crear bases de datos para inteligencia artificial y aprendizaje automático.
- Extracción de datos de manera ilegal o no ética, para obtener información personal o confidencial de los usuarios de un sitio web sin su consentimiento.

Para agilizar el trabajo de los *scrapers*, los *Crawlers* o arañas navegan por la web buscando e indexando (Rama-Rico, 2022; Gilabert, 2021). De igual modo, los motores de búsqueda como Google utilizan rastreadores para actualizar los índices y las clasificaciones de los sitios web.

En este artículo se presenta una revisión de las principales herramientas de *web scraping* disponibles en el mercado, se comparan algunas de sus características y funcionalidades, con la intención de seleccionar una herramienta scraping para buscar y extraer datos sobre porcentajes de egresados de las diferentes carreras ofertadas en México, así como, la distribución con respecto a género y salario acorde con las profesiones en distintos estados de la República Mexicana.

2. Problemática

Para presentar la utilidad de las técnicas extracción de datos, se planteó obtener datos sobre la empleabilidad de los egresados en México. La educación superior es un factor clave para el desarrollo económico y social de un país, ya que contribuye a la formación de capital humano calificado, a la generación de conocimiento y a la innovación (Medina et al, 2021). Sin embargo, existen diversos desafíos y desigualdades que afectan al sistema educativo y al mercado laboral en México, tales como la baja cobertura, la calidad y pertinencia de la oferta educativa, la inserción laboral de los egresados, las condiciones de trabajo y los salarios de los profesionistas, y las brechas de género y territoriales (Fernández-Fassnach, 2017).

Para estudiar estos aspectos, es necesario contar con información estadística confiable y actualizada que permita conocer el perfil de los egresados de nivel superior por carreras, la cantidad de mujeres y hombres empleados, las carreras mejor y peor pagadas, las carreras con mayor porcentaje de empleos, la diferencia salarial de empleos por estados, las áreas profesionales con mayor y menor número de profesionistas, entre otros indicadores. Estos datos ayudan a identificar las fortalezas y debilidades del sistema educativo superior y del mercado laboral, así como las oportunidades y necesidades de los egresados y los empleadores.

Para obtener o recuperar la información, es necesario realizar un proceso de extracción de datos en los sitios web que contienen los datos del comportamiento de las profesiones en México. Por tanto, los datos recuperados pueden responder a interrogantes como las que se mencionan a continuación:

- ¿Qué carreras tienen mayor demanda y oferta laboral en el país y en cada entidad federativa?
- ¿Qué carreras tienen mayor retorno económico y social para los egresados?
- ¿Qué carreras presentan mayor equidad de género en términos ocupación y salario?

Por lo anterior, es preciso obtener información que permitan la generación de datos analizables. Para lo cual, en primer lugar, se necesita encontrar la fuente de información que satisface el estudio requerido. En segundo lugar, revisar las posibles herramientas *Web Scrapers,* finalmente, realizar la extracción de los datos con una herramienta seleccionada.

3. Fuentes para la obtención de datos

Para lograr la extracción de datos planteada, primero, se requiere la recopilación de información de empleabilidad de los egresados de las distintas carreras que se ofrecen en México, por lo tanto, es necesario realizar una búsqueda de sitios oficiales que administran información de este tipo. Dependiendo del sitio *web*, la información disponible se encuentra en publicaciones periódicas o en anuarios estadísticos, para lo cual se realiza un primer filtrado de datos para consensuar si la información satisface los requisitos planteados.

A continuación, se mencionan algunas fuentes oficiales más destacadas que proporcionan información sobre egresados en México:

- **Sistema Nacional de Información Estadística Geografía (INEGI).** Es un organismo público, autónomo responsable de normar y coordinar el Sistema Nacional de Información Estadística y Geográfica, así como de captar y difundir información de México en cuanto al territorio, los recursos, la población y economía, que permita dar a conocer las características de nuestro país y ayudar a la toma de decisiones. El INEGI a través del levantamiento de datos en la Encuesta Nacional de Ocupación y Empleo (ENOE) recopila y publica información referente a: empleo y ocupación, hogares y vivienda, población, salud y seguridad social, tecnologías de la información y comunicaciones. En el caso de las características educativas de la población mexicana, proporciona información sobre la población que asiste a la escuela, así como alfabetismo y nivel de escolaridad. También, incluye información sobre el número de alumnos inscritos en los diferentes niveles educativos, indicadores de eficiencia y sobre los recursos humanos del sistema educativo nacional (Instituto Nacional de Estadística y Geografía, 2023).
- **Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES).** La Asociación Nacional de Universidades e Instituciones de Educación Superior, en constante esfuerzo por mejorar el sistema de información de la educación superior en México, presenta el Anuario Estadístico en su versión digital, el cual contiene información de la población escolar y del personal docente de los tipos de educación media superior y educación superior en los niveles técnico superior universitario, licenciatura universitaria y

tecnológica, licenciatura en educación normal y posgrado. La ANUIES, plantea la necesidad de contar con un *sistema* consolidado *de seguimiento de egresados universitarios que oriente sobre la demanda laboral,* que permita obtener información confiable que oriente a las instituciones de educación superior sobre los profesionales que demanda el mundo laboral (Asociación Nacional de Universidades e Instituciones de Educación Superior, 2023).

- **Observatorio Laboral (OLA).** Es un servicio público de información confiable y gratuito que la Secretaría del Trabajo y Previsión Social (STPS), a través del Servicio Nacional de Empleo (SNE) ofrece, sobre las principales carreras profesionales del país, con la finalidad de que los jóvenes, los estudiantes, y los padres de familia, cuenten con información confiable y veraz que les permita tomar decisiones sobre qué carrera elegir y como insertarse en el mundo del trabajo (Observatorio Laboral, 2023).

Existen otras fuentes oficiales de información desde el portal del Gobierno de México (2018), pero también se basan en datos de la INEGI y en los anuarios de la ANUIES. Consecuentemente, para el caso de estudio propuesto se determina que la fuente de información disponible en el portal del Observatorio Laboral es factible y suficiente para realizar un *scraping* y generar la información requerida.

3.1. El futuro de las profesiones según OLA

Sobre el futuro de las carreras en las universidades, el Observatorio Laboral (OLA) menciona:

- Carreras que tienen mayor captación de empleo son las relacionadas con:
  - Nuevas tecnologías, Internet y el sector digital.
- Las carreras más prometedoras tienen que ver con Tecnologías:
  - La informática, la telemática, la telefonía celular, la ingeniería genética, la biotecnología, la biónica, la realidad virtual, la información multimedia y los nuevos materiales cerámicos.
- Serán prometedoras las profesiones relacionadas con:
  - Los cuidados a distancia para la tercera edad y la infancia, la teleasistencia sanitaria, los cultivos acuáticos, la robótica, la domótica, los sistemas de seguridad pública y la inteligencia artificial.
- La tendencia internacional (según la OCDE):
  - Ingeniería Molecular, Nanotecnología, Biomedicina, Investigación Espacial, Cibernética, Mecatrónica, Ciencias de la Tierra.
- Deja fuera las carreras relacionadas con las Ciencias Sociales, Humanidades y Artes.

3.2. Herramientas disponibles para web scraping

El web *scraping* es el proceso de automatizar la extracción de datos de sitios web de manera estructurada. Permite recopilar grandes cantidades de información de forma eficiente y automatizada, lo que ahorra tiempo y esfuerzo considerable. Las herramientas de *scraping* desempeñan un papel fundamental en este proceso al proporcionar una interfaz o biblioteca que simplifica la extracción de datos.

Estas herramientas utilizan diferentes métodos y técnicas para extraer datos de sitios web. Algunas herramientas se basan en técnicas de web *scraping* tradicionales, como el análisis de HTML y el uso de expresiones regulares para extraer información específica de las páginas web. Otras herramientas más avanzadas utilizan tecnologías como la automatización de navegadores o la inteligencia artificial para realizar la extracción (Jezreel & Ramirez, 2016).

Existen diversas herramientas de *scraping* disponibles, desde bibliotecas de programación, hasta aplicaciones con interfaz gráfica. Algunas de las herramientas más utilizadas incluyen: BeautifulSoup, Scrapy, Selenium, Octoparse, Import.io, dexi.io, WebHarvy, ParseHub, Web Scraper Chrome Extension, Apify, OutWit Hub, visual scraper, scraping hub, UiPath entre otras. Cada herramienta tiene sus propias características, ventajas y limitaciones, por lo que es importante elegir la herramienta adecuada según las necesidades y requerimientos del problema o proyecto.

Para seleccionar una herramienta de *scraping*, es importante considerar criterios como el modelo de negocio (de pago o libre), los formatos de documentos a scrapear (ej. html o pdf), los formatos de exportación de datos (ej. CSV, XLS o JSON), la cantidad de *scrapers* permitidos, la compatibilidad con sistemas operativos y otras características adicionales ofrecidas por cada herramienta. En la Tabla 1 se muestra el resultado de la revisión de 11 herramientas de *scraping*, teniendo en cuenta los criterios antes mencionados y la compatibilidad con sistemas operativos (SO).

Herramienta	Modelo de negocio	Fuente	Exportación	Scrapers permitidos	Compatibilidad con SO
BeautifulSoup	Libre y de código abierto	HTML y XML	Combinación con Phyton para exportar CSV, JSON, XML	Sin límite	Windows, macOS y Linux
Scrapy	Libre y de código abierto	HTML y XML	JSON, CSV y XML	Sin límite	Windows, macOS y Linux
Selenium	Libre y de código abierto	HTML	Combinación con Phyton para exportar CSV, JSON	Sin límite	Windows, macOS y Linux
Octoparse	Gratuitas y de pago	HTML y PDF	CSV, Excel, JSON y base de datos SQL	Gratuita: hasta 10 scrapers Pago: sin límite	Windows
Import.io	Gratuitas y de pago	HTML	CSV, Excel, JSON y base de datos SQL	Gratuita: 1 scraper Pago: sin límite	Windows, macOS y Linux
ParseHub	Gratuitas y de pago	HTML	CSV, Excel, JSON y base de datos SQL	Gratuita: hasta 5 proyectos. Pago: sin límite.	Windows, macOS y Linux
WebHarvy	Gratuitas y de pago	HTML	CSV, Excel, JSON y base de datos SQL	Sin límite	Windows

Herramienta	Modelo de negocio	Fuente	Exportación	Scrapers permitidos	Compatibilidad con SO
Apify	Gratuitas y de pago	HTML	JSON, CSV y Excel	Gratuita: hasta 10 actores Pago: sin límite	Windows, macOS y Linux
OutWit Hub	Gratuitas y de pago	HTML	CSV, Excel y XML	Sin límite	Windows y macOS
Data Miner	Gratuitas y de pago	HTML	CSV, Excel y Google Sheets	No limitada	Compatible con Google Chrome como una extensión
Web Scraper Chrome Extension	Gratuita	HTML	CSV y JSON	Sin límite	Compatible con Chrome en Windows, macOS y Linux

Tabla 1 – Comparación de herramientas *scrapers*. Elaboración propia.

Como se muestra en la Figura 2, para efectos de este trabajo, se realizó una prueba de conocimiento con tres herramientas que se ajustaron a las condicionantes de la investigación: acceso gratuito, fácil instalación, facilidad de uso, varios sitios para procesar y el formato de exportación en csv (https://www.octoparse.com/, https://www.webharvy.com/index.html, https://webscraper.io/). En la Figura 2, la ventana superior es una vista de escrapeado del sitio OctoParse, la ventana a la izquierda corresponde a la ejecución en Web Harvy y la ventana derecha corresponde al sitio web scraper.

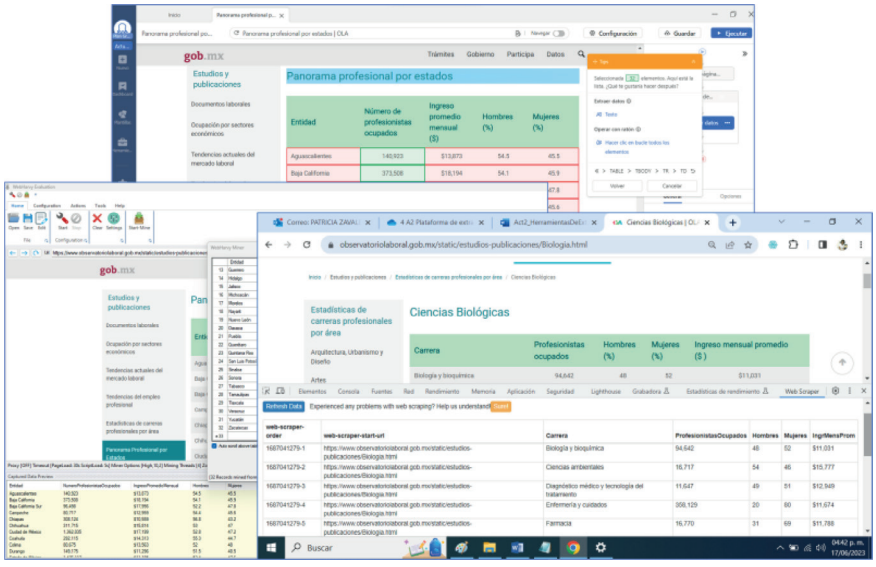

Figura 2 – Escrapeado con octoparse, web escraper y webharvy

En la Figura 3, se muestra una arquitectura de *entrada-proceso-salida* para representar un *web scraping* exitoso siguiendo la siguiente secuencia de pasos:

1. Especificar las URLs de los sitios web y las páginas que se desean scrapear.
2. Acceder a la página web mediante una solicitud HTTP.
3. Analizar el contenido HTML de la página para identificar la información deseada. Utilizar localizadores como expresiones regulares para extraer la información.
4. Extraer la información y almacenarla en una base de datos o en un formato estructurado, como CSV, JSON u otro formato disponible.

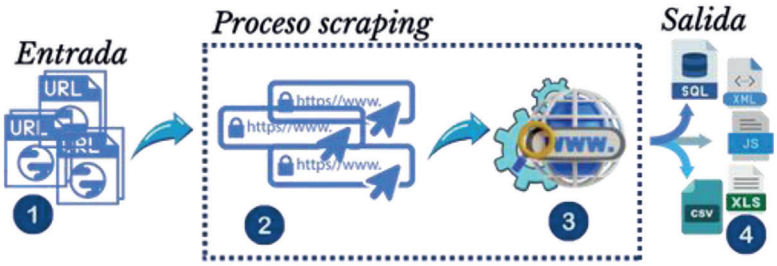

Figura 3 – Arquitectura web scraping. Elaboración propia.

3.3. Extracción empleabilidad de egresados con la herramienta WebHarvy

Para el caso de la obtención de datos de egresados planteado como problemática para extracción, se eligió la herramienta de WebHarvy, ya que, durante las pruebas de raspado las tres herramientas (Octoparse, WebHarvy y Web scraper), WebHarvy arrojó mejores resultados, tanto en la extracción como en la exportación de los datos. A continuación, se describe el proceso seguido.

3.3.1. Paso 1. URL del portal web para la fuente de información

En este paso se proporciona el sitio que contiene los datos, es el más sencillo, y corresponde a proporcionar el enlace (*link)* o *URL*, para iniciar el raspado, como se puede observar en la Figura 4.

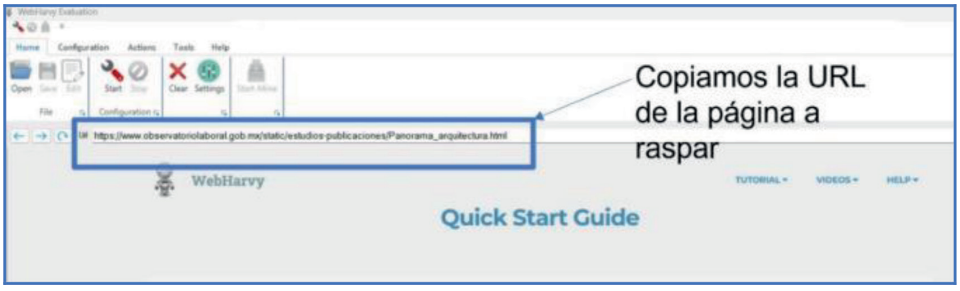

Figura 4 –URL de la página para su raspado.

3.3.2. Paso 2. Proceso de scrapeado (raspado)

Posteriormente de ingresar el URL, se oprime el ícono de "*Start*". Como se puede observar en la Figura 5, es necesario seleccionar un campo de la columna deseada, posteriormente oprimir "*Capture Text*", en la ventana emergente colocar el nombre que se le dará a la columna del dato indicado, finamente presionar el botón "*OK*".

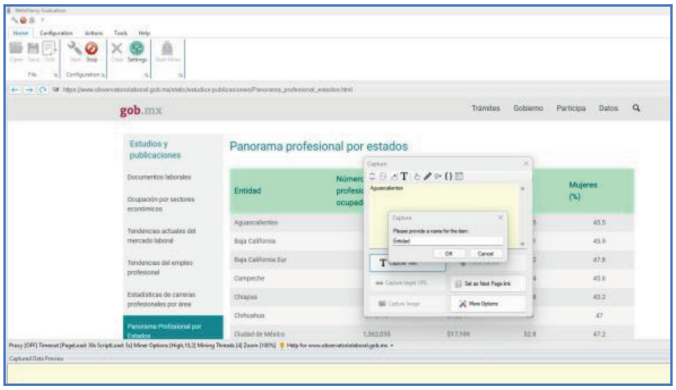

Figura 5 – Ejemplo de columnas para el raspado.

En la Figura 6 se puede observar que en la parte inferior se recuperan los datos de la columna seleccionada, identificada con el nombre que se escribió como encabezado; se repetirán estos pasos para completar todas y cada una de las columnas deseadas. Al terminar de raspar (scrapear) toda la tabla, se oprime el botón "*Stop*" de la parte superior, para detener el proceso.

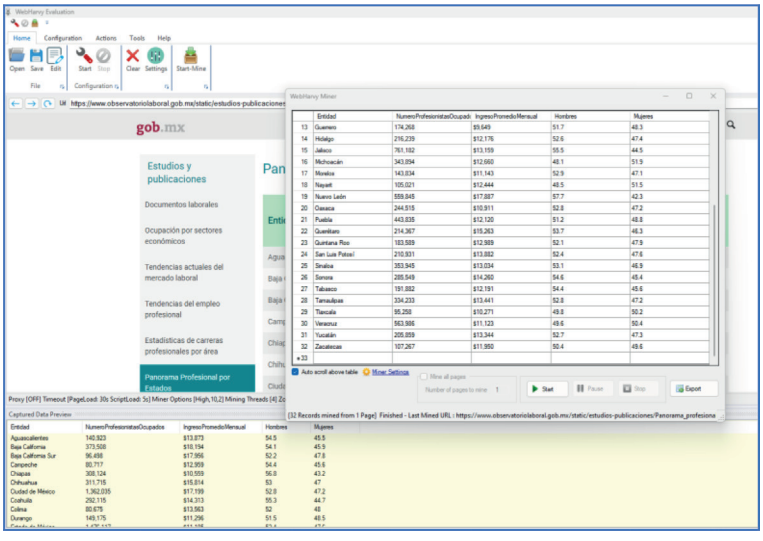

Figura 6 – Raspado de datos en WebHarvy.

Este proceso finaliza oprimiendo el icono "*Start-Mine*", posteriormente presionar el botón "*Start*" para proceder a generar todos los datos antes seleccionados y almacenarlos en una tabla, para su exportación.

En este caso los datos se guardaron en un archivo de hoja de cálculo permitiendo elegir el destino y el nombre del archivo. De igual modo, esta opción permite seguir guardando en el archivo, sobrescribirlo (*Overwrite*), agregar registros (*Append*) o actualizar el documento (*Update*).

3.3.3. Paso 3. Resultados: datos obtenidos

Una vez terminado el proceso anterior, abrimos el documento de hoja de cálculo y aparecen los datos de todas las páginas raspadas, acomodados en columnas, como se observa en la Figura 7. Para efectos de la muestra del raspado, solo se muestran algunas filas de la información obtenida.

A	B	C	D	E	F	G
Area	**Carrera**	**ProfesionistasOcupa**	**Hombres**	**Mujeres**	**Ingresopromediomensual**	
Arquitectura, Urbanismo y Dise	Arquitectura y urbanismo	240,510	71	29	$15,031	
Artes	Artes, programas multidisciplinarios o generales	9,936	26	74	$11,756	
Artes	Bellas artes	18,580	40	60	$8,869	
Artes	Diseño	44,114	29	71	$12,440	
Artes	Música y artes escénicas	31,205	63	37	$9,420	
Artes	Técnicas audiovisuales y producción de medios	156,300	53	47	$13,889	
Ciencias Biológicas	Biología y bioquímica	94,642	48	52	$11,031	
Ciencias Biológicas	Ciencias ambientales	16,717	54	46	$15,777	
Ciencias Biológicas	Diagnóstico médico y tecnología del tratamiento	11,647	49	51	$12,949	
Ciencias Biológicas	Enfermería y cuidados	358,129	20	80	$11,674	
Ciencias Biológicas	Farmacia	16,770	31	69	$11,788	
Ciencias Biológicas	Medicina	324,198	54	46	$18,569	
Ciencias Biológicas	Odontología	162,107	46	54	$14,809	
Ciencias Biológicas	Psicología	357,044	25	75	$11,395	
Ciencias Biológicas	Química	31,468	40	60	$13,278	
Ciencias Biológicas	Salud pública	14,692	46	54	$18,698	
Ciencias Biológicas	Terapia y rehabilitación	107,851	27	73	$10,209	
Ciencias Biológicas	Veterinaria	77,656	75	25	$12,378	
Ciencias Físico Matemáticas	Ciencias de la tierra y de la atmósfera	14,428	63	37	$16,742	
Ciencias Físico Matemáticas	Estadística	8,557	62	38	$17,232	
Ciencias Físico Matemáticas	Física	9,967	82	18	$12,723	
Ciencias Físico Matemáticas	Matemáticas	36,313	61	39	$13,664	
Ciencias Sociales	Biblioteconomía	4,877	47	53	$10,179	
Ciencias Sociales	Ciencias políticas	77,835	45	55	$14,479	
Ciencias Sociales	Ciencias sociales y estudios del comportamiento	12,692	58	42	$14,125	
Ciencias Sociales	Comunicación y periodismo	198,551	47	53	$12,425	
Ciencias Sociales	Criminología	58,163	52	48	$12,028	
Ciencias Sociales	Derecho	308,694	61	39	$13,123	
Ciencias Sociales	Sociología y antropología	46,252	56	44	$14,218	
Ciencias Sociales	Trabajo y atención social	93,296	10	90	$10,018	
Económico Administrativas	Administración y gestión de empresas	1,091,842	49	51	$13,137	
Económico Administrativas	Comercio	281,002	49	51	$15,780	
Económico Administrativas	Contabilidad y fiscalización	803,231	51	49	$13,418	
Económico Administrativas	Economía	81.937	58	42	$15.256	

Figura 7 – Documento en hoja de cálculo con datos raspados.

3.4. Recomendaciones sobre la extracción obtenida

Cabe mencionar que para realizar el raspado de varias páginas web, cómo fue el caso planteado, se debe realizar página por página, de manera individual, es decir, que se realiza el raspado de la primera página, se genera un documento en Excel y se guarda para exportar el archivo. Posteriormente con la opción *Append* se repite el procedimiento para la página siguiente y anexarlo al final del último registro.

El proceso anterior es iterativo para todas las páginas; se debe considerar que las otras páginas pueden tener un orden diferente a las columnas de información extraídas, por ello, es importante que cada vez que se realiza el raspado, se verifique el orden con respecto al nombre de las columnas para garantizar que los registros coincidan con el encabezado previamente seleccionado.

Una utilidad que se resalta de WebHarvy es la facilidad para seleccionar las columnas deseadas y el orden en que se requieren, permitiendo la asignación del identificador (nombre) para cada columna.

En los inconvenientes se mencionan que, no se puede scrapear otra URL de manera continua. Otro inconveniente es que la licencia tiene límite de tiempo de 16 días.

La técnica *Scraping* es de gran utilidad, sin embargo, acceder a la información conlleva a un proceso de ética y responsabilidad sobre el acceso y protección de datos, sobre todo aquellos que no están disponibles para uso de dominio público o aquellos que son utilizados en perjuicio de terceros.

4. Conclusiones

Este trabajo presentó a las herramientas *Scrapers* para la obtención de datos de sitios *web,* las cuales otorgan facilidades cuando se requiere descargar y tratar considerables cantidades de información, que pueden apoyar en la toma de decisiones.

La selección de una herramienta *scraping* requiere de un proceso de selección en el que el usuario debe dar importancia a la determinación de criterios que permitan elegir una herramienta que mejor se adapte tanto a las necesidades de obtención y procesamiento datos, al modelo de negocio esperado (de pago o libre), así como también a los requisitos de compatibilidad con sistemas operativos ente otras características adicionales ofrecidas por cada herramienta.

La herramienta *WebHarvy* utilizada para el proceso de raspado del sitio "Observatorio Laboral", ofrece grandes ventajas a la hora de raspar los datos, con la opción de *Append,* para anexar los datos al final del archivo y la posibilidad de omitir columnas innecesarias. Así también presentó algunos inconvenientes, como lo fueron: el orden diferente de las columnas en las otras páginas, la necesidad de cerrar las ventanas abiertas para cambiar de URL y la duración limitada de la licencia. Al realizar el proceso de manera repetida, la herramienta se vuelve mecánica y sencilla de trabajar.

Como continuación a este trabajo, se propone contrastar los datos presentados por el Observatorio Laboral, con los resultados escrapeados llevados a un segundo nivel, utilizando una herramienta de inteligencia de negocios.

Por último, el *scraping* es una técnica que permite recoger datos de la *web* de forma rápida y automatizada. Con estos datos se puede obtener información útil para diversos fines, como estudios de mercado, comparación con la competencia, captación de clientes y más. Si se respeta la ética y los términos de servicio de los

sitios *web*, el *scraping* puede ser una herramienta eficaz para extraer conocimientos valiosos de la gran cantidad de datos que hay en la web.

Referencias

Asociación Nacional de Universidades e Instituciones de Educación Superior. (2023). Anuarios Estadísticos de Educación Superior. http://www.anuies.mx/informacion-y-servicios/informacion-estadistica-de-educacion-superior/anuario-estadistico-de-educacion-superior

Dirección General de Educación Superior (DGESUI). (2023). Consultas. https://dgesui.ses.sep.gob.mx/

El comercio Perú. (2023). Web scraping: qué es y cómo funciona la herramienta que extrae información de los sitios web. (M. M. Saravia, Editor) https://elcomercio.pe/tecnologia/ciberseguridad/web-scraping-que-es-y-como-funciona-la-herramienta-que-extrae-informacion-de-los-sitios-web-espana-mexico-usa-noticia/

Escuela de datos. (2023). Introducción a la extracción de datos de sitios web: scraping. https://escueladedatos.online/introduccion-a-la-extraccion-de-datos-de-sitios-web-scraping/

Fernández-Fassnach, E. (2017). Una mirada a los desafíos de la educación superior en México. Innovación Educativa, 17(74), 183-207. Recuperado el 20 de octubre 2023, de https://www.scielo.org.mx/pdf/ie/v17n74/1665-2673-ie-17-74-00183.pdf

Gilabert-Perea, X. (2021). Diseño e implementación de un extractor de noticias automatizado. Trabajo de Fin de Grado en Ingeniería Informática. Universitat Politècnica de València, España. Recuperado de https://riunet.upv.es/bitstream/handle/10251/174252/Gilabert%20-%20Diseno%20e%20implementacion%20de%20un%20extractor%20de%20noticias%20automatizado.pdf

Gobierno de México. (2018). Educación por niveles. https://www.gob.mx/sep/acciones-y-programas/educacion-por-niveles?state=published

Instituto Nacional de Estadística y Geografía. (2023). México en Cifras. https://www.inegi.org.mx/app/areasgeograficas/#collapse-Resumen

Jezreel, M., & Ramirez, H. (2016). Estableciendo controles y perímetro de seguridad para una página web de un CSIRT. RISTI - Revista Ibérica de Sistemas y Tecnologias de Información, 17, 1-14. https://www.risti.xyz/issues/risti17.pdf

Kinsta. (2022). ¿Qué Es el Web Scraping? Cómo Extraer Legalmente el Contenido de la Web. https://kinsta.com/es/base-de-conocimiento/que-es-web-scraping/

Medina Delgado, B., Palacios Alvarado, W. Camargo Ariza, L. L. (2021). Economía del conocimiento en la educación superior: factor clave en la calidad educativa. REDIPE 10(7). https://doi.org/10.36260/rbr.v10i7.1347

Mitchell, R. (2018). Web scraping with Python: Collecting more data from the modern web. .O'Reilly Media, Inc..

Observatorio Laboral (2023). Estadísticas de carreras profesionales por área. https://www.observatoriolaboral.gob.mx/static/estudios-publicaciones/Ola_indice_estadisticas_area.html

Papeles de Inteligencia. (2018). 10 herramientas de web scraping para extraer datos online de forma automática. https://papelesdeinteligencia.com/herramientas-de-web-scraping/#:~:text=Las%20herramientas%20de%20web%20scraping%20est%C3%A1n%20especialmente%20dise%C3%B1adas%20para%20extraer,datos%20de%20una%20p%C3%A1gina%20web.

Rama Rico, J. A. (2022). Desarrollo de un software para la búsqueda de apuestas seguras. Trabajo Fin de Grado. Grado en Ingeniería en Tecnologías Industriales. Universidad de Sevilla, España. Recuperado de: https://idus.us.es/bitstream/handle/11441/140647/TFG4331_Rama%20Rico.pdf?sequence=1&isAllowed=y

Ribeiro, A. J., Mendes, R., & Duarte, M. do C. (2022). Requisitos para a ciência de dados: Analisando anúncios de vagas de emprego com mineração de texto. RISTI - Revista Ibérica de Sistemas y Tecnologias de Información, 46, 54-70. https://www.risti.xyz/issues/risti46.pdf

Secretaría del Trabajo y Previsión Social. (2023). 90 Datos en datos.gob.mx. Catálogo de Datos Abiertos del Gobierno de la República: https://datos.gob.mx/busca/organization/stps

Sinche, J. U., & Torres, J. C. (2021). Análisis de sentimientos en los mensajes recibidos en el entorno virtual de aprendizaje de la modalidad abierta y a distancia de la UTPL. RISTI - Revista Ibérica de Sistemas y Tecnologías de la Información, 41, 98-113. http://www.risti.xyz.

Sitelabs. (2016). Qué es el Web scraping? Introducción y herramientas. (M. Marq, Editor) https://sitelabs.es/web-scraping-introduccion-y-herramientas/