Desarrollar una herramienta de web scraping que permita recopilar y analizar información de precios de competidores.

Develop a web scraping tool to collect and analyze price information from competitors.

Juan Felipe Lindo [1]
Juan.lindo00@usc.edu.co

Bernardo García Palacios [1]
Bernardo.garcia00@usc.edu.co

Jefferson Herrera Achinte [1]
Jefferson.herrera00@usc.edu.co

Gustavo Adolfo Alomia Peñafiel, M.Sc [1]
Gustavo.alomia00@usc.edu.co

Universidad Santiago de Cali, Facultad de Ingeniería, Programa de [Ingeniería de sistemas] (1)

Resumen

La recopilación y análisis de información precisa y actualizada sobre las empresas que representan potencial competencia en un mercado específico, como por ejemplo las ventas en línea, que es el enfoque de esta investigación, resulta fundamental para alcanzar el éxito empresarial. Aunque esta tarea puede requerir mucha dedicación de forma manual, existen métodos automatizados como el web scraping, una solución eficiente que permite extraer datos de múltiples sitios web de forma rápida y precisa.
El propósito de este proyecto es crear una herramienta de web scraping que permita obtener datos relevantes, como precios, descripciones y características de productos, de competidores, y presentarlos en un dashboard personalizado que se ajuste a las necesidades específicas de cada negocio. Con el propósito de ayudar a las empresas a tomar decisiones según el comportamiento del mercado, mejorando su estrategia de precios y obteniendo información valiosa sobre las preferencias de los consumidores.

Palabras Clave: Raspado web; Python; Procesamiento de datos; Estudio de mercado

Abstract

Gathering and analyzing accurate and up-to-date information about companies that potentially represent competition in a specific market, such as online sales, which is the focus of this research, is of utmost importance for achieving business success. Although this task can require a lot of manual effort, there are automated methods such as web scraping, an efficient solution that allows for the quick and precise extraction of data from multiple websites.
The purpose of this project is to create a web scraping tool that allows obtaining relevant data such as prices, descriptions, and product features from competitors, and presenting them in a customized dashboard that meets the specific needs of each business. The goal is to help companies make decisions based on market behavior, improve their pricing strategy, and obtain valuable information about consumer preferences.

Keywords: Web scraping; Python; Data mining; Market study

1. INTRODUCCION

En el actual mercado altamente competitivo de las ventas en línea, la recopilación y análisis de información precisa y actualizada sobre los competidores es un factor crítico para el éxito empresarial. Sin embargo, esta tarea manual puede resultar tediosa y demandando tiempo y recursos valiosos. Para enfrentar este desafío, una solución automatizada y eficiente es el uso de herramientas de web scraping, las cuales permiten extraer datos de múltiples sitios web de manera rápida y precisa.

Web scraping es una técnica que ha sido utilizada en diversos campos de investigación. Por ejemplo, la utilidad del web scraping es el estudio llevado a cabo en la extracción de datos y análisis de financiamiento de viviendas de alquiler para optimizar los procesos y ayudar a los planificadores a comprender mejor el alcance financiero en el sector de vivienda (St-Hilaire et al., 2023). Asimismo, en el campo de la hostelería, se llevó a cabo la investigación "Web Scraping for Hospitality

Research: Overview, Opportunities, and Implications" En dicha investigación, se creó una guía para la recolección de datos de hoteles disponibles en línea. (Han & Anderson, 2021). A pesar de que el web scraping se ha empleado en diversas áreas, esta investigación se distingue por su objetivo de desarrollar una solución efectiva que aproveche las ventajas de las tecnologías de la información, la inteligencia de negocios y la automatización de procesos. Su propósito es lograr resultados efectivos que justifiquen la inversión realizada y generen ganancias. Esta herramienta de web scraping responde a la necesidad de proporcionar a los negocios emergentes un estudio de mercado que les permita competir en las ventas por internet y fortalecer la toma de decisiones.

El enfoque metodológico que se utilizará para llevar a cabo el proyecto será el Design Thinking. Esta metodología es efectiva para desarrollar productos y resolver problemas de manera óptima, aprovechando las habilidades del equipo. No obstante, es importante aplicarla de manera específica para cumplir con el objetivo general. Por tanto, será necesario realizar algunas modificaciones en la ejecución de la metodología para adecuarla a dichos objetivos.

En el caso concreto de este proyecto, cada cliente tiene necesidades específicas relacionadas con su negocio principal. Sin embargo, se ha conseguido estandarizar el proceso de manera que se puedan seguir los mismos pasos en todos los casos, sin comprometer la calidad del resultado final ni el rendimiento del equipo. Gracias a esta metodología iterativa y estandarizada, se logra obtener resultados de alta calidad para los clientes.

La eficacia y utilidad de las herramientas de web scraping se evidencia por su amplia aplicabilidad para la recopilación de información de sitios web con diversos propósitos (Gallagher & Beveridge, 2022). En este contexto, este trabajo se enfoca en desarrollar una herramienta automatizada de extracción de datos a precios, descripciones de productos como licores, alimentos y bebidas de competidores, con el objetivo de presentar esta información en un dashboard que ofrezca una buena experiencia de usuario para la toma de decisiones empresariales (Ashouri et al., 2022). El propósito de esta herramienta es facilitar nuevas soluciones que asistan a las empresas en la formulación de estrategias más informadas y precisas, optimizando de esta manera su estrategia de precios y obteniendo datos valiosos acerca de las preferencias de los consumidores en México.

El presente artículo se organiza en varias secciones. En la Sección 2 se proporciona una revisión de algunos términos y elementos necesarios para comprender las siguientes secciones. En la Sección 3 se describe detalladamente las herramientas y bibliotecas utilizadas para llevar a cabo el proceso de Web Scraping y la automatización de la extracción de información de sitios web. Además, se presentan los resultados obtenidos a partir de la investigación, describiendo los procesos y pasos seguidos para extraer y consolidar la información de manera estructurada. Por último, en la Sección 4 se presentan las conclusiones del estudio.

1.1 ESTADO DEL ARTE

Esta investigación ha permitido recopilar una gran cantidad de información sobre los diversos estudios y aplicaciones de la tecnología de Web Scraping en todo el mundo. Son numerosos los ejemplos de su uso y cómo se ha podido explotar su funcionalidad en diversos ámbitos, como el de la salud, financiero y entretenimiento, entre otros. Un ejemplo claro de esto se expone en la investigación."Web Scraping for Hospitality Research: Overview, Opportunities, and Implications", en la que los investigadores construyeron una guía sobre cómo recopilar de manera más eficiente los datos de hoteles en línea disponibles públicamente (Han & Anderson, 2021). Esto sugiere que el Web Scraping tiene un gran potencial para mejorar el comercio turístico.

Como se mencionó anteriormente, esta tecnología posee un gran potencial para la minería de datos, pudiendo automatizar la recolección de grandes cantidades de información para una posterior toma de decisiones. En el campo de la medicina, también existen importantes aportes del uso de esta tecnología. Dos ejemplos de esto se encuentran en el artículo " Hybrid machine learning approach for Arabic medical web page credibility assessment ", En este estudio se propone un método que automatiza la evaluación de la fiabilidad de las páginas médicas en línea, basándose en características tanto textuales como no textuales, a través de web scraping, algoritmos de aprendizaje automático y aprendizaje profundo se identificaron los patrones para detectar páginas web poco fiables para la consulta de temas médicos (Alasmari et al., 2022).

1.1.1 Aprovechando el potencial del Web Scraping en la investigación académica y el monitoreo gubernamental

La utilización de la tecnología de Web Scraping ha abierto puertas no solo en el ámbito empresarial, sino también en el académico y gubernamental. En la investigación "Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar", llevada a cabo por (Rahmatulloh & Gunawan, 2020), se demuestra el uso de Web Scraping para recopilar datos de Google Scholar y posteriormente resumir la información de publicaciones de artículos científicos mediante la aplicación de tecnología de raspado web. Asimismo, otro ejemplo del uso de esta tecnología con fines académicos es "Open Editors: A dataset of scholarly journals' editorial board positions", una aplicación que analiza el papel de los editores de revistas académicas y su impacto en el sistema de publicación científica, según (Nishikawa-Pacher et al., 2022).

Además, se han desarrollado aplicaciones gubernamentales que explotan todas las ventajas que ofrece el Web Scraping, como en el caso de "Development aid contracts database: World Bank, Inter-American Development Bank, and EuropeAid", donde se construyó una base de datos global de contratos gubernamentales financiados por el Banco Mundial, según (Fazekas et al., 2022).

1.1.2 Extracción de datos con Web Scraping: Una revisión de las tecnologías y aplicaciones utilizadas en diversos campos, desde la ciencia de los materiales hasta el aprendizaje automático supervisado

Existen múltiples campos en los que se ha utilizado el Web Scraping. Se ha demostrado que las principales tecnologías utilizadas por los investigadores son la librería Python Selenium, el lenguaje de extracción xpath, el lenguaje de etiquetado HTML, los archivos xml para la extracción de datos y las bases de datos relacionales como MySQL.

Entre otros ejemplos en diferentes campos, se puede mencionar el artículo "Web Scraping data labeling system on liquid chromatography-mass spectrometry of rodent tuber for efficiency of supervised learning preprocessing" que aborda la dificultad del etiquetado de datos para el aprendizaje supervisado (Binanto et al., 2022). Asimismo, en "Auto-generating databases of Yield Strength and Grain Size using ChemDataExtractor", se utilizó esta técnica para la recolección de datos en el campo de la ingeniería de materiales y construir bases de datos de valores de límite elástico y tamaño de grano (Kumar et al., 2022).

1.1.3 Estudio de casos sobre la recopilación de información de productos en línea y análisis de precios en la industria a través de técnicas de raspado web

Como se puede observar en el resumen, hay muchos trabajos similares utilizando esta técnica. Un ejemplo muy similar a esta investigación se puede encontrar en "Web Scraping for Hospitality Research: Overview, Opportunities, and Implication" el este artículo se presentan herramientas y técnicas bien conocidas que son utilizadas para el raspado de precios. Se describe un escenario de ataque de raspado de precios a través del web scraping y se explican los pasos de ejecución del escenario de ataque de forma sistemática (Rahman & Tomar, 2021).

1.2 PLANTEAMIENTO DEL PROBLEMA

Un reciente estudio llevado a cabo por la consultora de negocios McKinsey & Company ha revelado que el 90% de las empresas que no realizan un análisis de sus competidores presentan una tasa de fracaso del 35%. Además, estas empresas sufren una disminución del 10% en sus ingresos anuales en comparación con aquellas que sí llevan a cabo este tipo de

análisis (Schmieder-Ramirez & Mallette, 2021).

Por otra parte, otro informe publicado por la firma de investigación de mercado IBISWorld ha puesto de manifiesto que aquellas empresas que no llevan a cabo una investigación adecuada de sus competidores son más propensas a experimentar un aumento en los costos de producción y una disminución en los márgenes de beneficio (IBISWorld, s. f.)

Los estudios realizados destacan la importancia crítica de analizar a los competidores y sugieren que las empresas que no lo hacen corren un mayor riesgo de fracasar y sufrir pérdidas financieras significativas. La investigación está fundamentada en un problema real que fue presentado en una empresa multinacional con una de sus sedes en Ciudad de México. Parte del equipo de este proyecto está compuesto por trabajadores de dicha compañía, quienes gracias a su conocimiento del negocio identificaron la necesidad latente y comprobaron los efectos negativos de no analizar a los competidores. La empresa en cuestión proporciona una plataforma tecnológica que mejora la experiencia de compra de sus clientes al permitirles adquirir sus productos principales (Cervezas, licores y abarrotes) de manera sencilla, así como ofrecer productos alternativos de empresas aliadas. No obstante, se enfrenta a un problema habitual en las ventas en línea, la obtención de datos fiables para la regulación de los precios. Según un estudio de (Khan et al., 2020), "las empresas afrontan diversos desafíos en la recolección de datos, como la calidad de los datos, la falta de acceso a datos relevantes y la incapacidad para integrar datos de múltiples fuentes". Esto puede resultar en que las empresas inicien proyectos con bases de datos poco fiables que dificulten la toma de decisiones y la elaboración de una estrategia eficaz. ¿Qué alternativa puede ser ofrecida a la empresa para satisfacer su necesidad de información confiable y rápida, que les permita diseñar una estrategia de ventas basada en datos precisos de precios?

A continuación, se definen las metas para el proyecto

1. Consolidar listado de empresas que sean competidores directos, con el fin de obtener información detallada acerca de los precios de sus productos.
2. Desarrollar una herramienta de web scraping que permita extraer los precios de los competidores seleccionados de forma automatizada y regular.
3. Realizar la limpieza y normalización de los datos extraídos.
4. Crear un panel de control de inteligencia empresarial utilizando modelos de diseño que brinden una experiencia de usuario placentera.

1.3 JUSTIFICACIÓN

La razón que motiva el desarrollo de la presente herramienta de web scraping reside en la necesidad de suministrar a la empresa en cuestión, así como a los negocios emergentes, un análisis de la estrategia de precios de sus competidores, que sea de fácil comprensión, robusto en su capacidad para informar la toma de decisiones y que les permita competir en el mercado de ventas en línea de México. Las empresas emergentes tienen el desafío de competir con compañías consolidadas. Un estudio reciente de 2021 destaca que la innovación requiere una amplia variedad de recursos, los cuales son difíciles de obtener para las empresas jóvenes con limitaciones económicas. Para superar estas limitaciones y lograr un crecimiento sostenible, las nuevas empresas buscan apoyo en los actores del ecosistema de innovación para aprovechar sus capacidades (Marcon & Ribeiro, 2021). Sin embargo, la falta de herramientas efectivas como el web scraping puede poner a las empresas emergentes en una posición menos competitiva en el mercado. Si bien los métodos tradicionales para realizar estudios de mercado pueden ofrecer resultados valiosos, estos métodos son más prolongados y requieren la dedicación de un profesional en la materia. Por lo tanto, la herramienta de web scraping que se está desarrollando será una solución efectiva que pueda suplir una necesidad evidente de las empresas que ofertan sus productos en línea y convertirse en aliados estratégicos generándoles valor.

Web scraping es una técnica que ha sido utilizada en diversos campos de investigación. Por ejemplo, en la investigación "Medical informatics labor market analysis using web crawling, web scraping, and text mining" se empleó el web scraping para realizar un estudio de mercado en el campo de la informática médica (Schedlbauer et al., 2021). Asimismo, en el campo de la hostelería, se llevó a cabo la investigación "Web Scraping for Hospitality Research: Overview, Opportunities, and

Implications" en la que los investigadores construyeron una guía para recopilar y raspar los datos de hoteles disponibles en línea (Han & Anderson, 2021). Aunque el web scraping ha sido utilizado en diversas áreas, esta investigación se distingue por su objetivo de desarrollar una solución eficaz que integre las ventajas de las tecnologías de la información, la inteligencia de negocios y la automatización de procesos. Asimismo, se persigue la generación de resultados efectivos que permitan recuperar la inversión y obtener beneficios económicos.

En conclusión, el desarrollo de la herramienta de web scraping se justifica por la necesidad de brindar a la empresa en cuestión una solución efectiva para realizar estudios sobre la estrategia de precios de competidores centrados en las ventas en línea en México, mejorando de esta manera su competitividad y posición en el mercado.

2. MATERIALES Y MÉTODOS / METODOLOGÍA

En las secciones siguientes se llevará a cabo un análisis del aspecto legal del proyecto, seguido de una explicación detallada de los materiales y herramientas empleados en el estudio, desde la descripción de los sitios de comercio electrónico seleccionados hasta la selección y aplicación de las herramientas de web scraping. Esto tiene como objetivo permitir una fácil comprensión del estudio y facilitar la reproducibilidad de los resultados.

2.1 MARCO LEGAL

Antes de emprender un proyecto de Web Scraping, es fundamental considerar tanto la legalidad como la ética del uso de los datos obtenidos de la Web. Una investigación exhaustiva de artículos de revistas que contengan términos como "web scraping" o "data scraping" en el título, sugiere que la recuperación y organización automatizada de datos de la Web sigue siendo un fenómeno relativamente nuevo y emergente (Murray State University et al., 2020). Es importante destacar que la mayoría de los artículos que tratan estos temas son recientes, con un período de publicación que abarca sólo unos pocos años.

La mayoría de los sitios web tienen términos y condiciones que incluyen acuerdos de licencia de usuario final en sus páginas. Estos términos explícitamente mencionan el acceso a su sitio web a través de scrapers y establecen un contrato entre el propietario del sitio web y el "scraper", con la intención de crear una responsabilidad por incumplimiento de contrato. Sin embargo, publicar estos términos en un sitio web puede no ser suficiente para demostrar que un "scraper" ha incumplido las condiciones, ya que no hay una aceptación activa por parte del "scraper". Por lo tanto, parece más efectivo utilizar una casilla de verificación explícita o un enlace "Acepto", en el que el "scraper" tenga que hacer clic activamente para aceptar las condiciones. Lo mismo ocurre con las aplicaciones de "scraping" que se registran en un sitio para acceder a un área no pública, ya que la creación de una cuenta de usuario en un sitio web también suele incluir la aceptación explícita de las condiciones (vanden Broucke & Baesens, 2018). Este requisito ha sido destacado en casos como el de Alan Ross Machinery Corp. contra Machinio Corp en 2018[1] y en el caso de Facebook, Inc. contra Power Ventures, Inc. en 2016[2].

Por consiguiente, la exclusiva prohibición del Web Scraping en un sitio web puede no ser suficiente para evitar que alguien lo rastree legalmente. Asimismo, el sitio web tendría que evidenciar daños económicos derivados del incumplimiento de las "Condiciones de uso" o las "Condiciones del servicio" para presentar una demanda exitosa por incumplimiento contractual (Murray State University et al., 2020).

Las medidas de seguridad implementadas por el sitio web también representan un aspecto a tener en cuenta en la aplicación de esta práctica, el sitio web debe contar con medidas de seguridad adecuadas para evitar que un usuario despliegue herramientas automatizadas con el fin de recolectar datos u otros fines. Es necesario verificar si se han implementado medidas razonables como el uso de captchas, bloqueo de accesos masivos desde una misma dirección IP u otros mecanismos que el propietario considere pertinentes (Sanabria De Luque, 2021). En relación con el proyecto en desarrollo, es importante destacar que los sitios de raspado seleccionados (véase **Tabla 1**) no implementan este tipo de medidas de seguridad o limitantes contra bots en su navegación.

[1] Alan Ross Machinery Corp. v. Machinio Corp, 2018 WL 6018603 (N.D. Ill. Nov. 16, 2018).

[2] Facebook, Inc. v. Power Ventures, Inc. et. al., 844 F.3d 1058 (9th Cir. 2016).

Asimismo, el contexto del presente proyecto se han considerado las recomendaciones primordiales generales presentadas en los diversos artículos relacionados. Se hace notar que los datos que se pretenden recopilar son de acceso público y no restringido. Además, se asegura que el scraping se realizara con propósitos lícitos y no con la intención de cometer actividades fraudulentas o de competencia desleal.

2.2 IDENTIFICACIÓN DE FUENTES DE DATOS Y RECOGIDA DE DATOS

2.2.1 Identificación de empresas

Con el fin de identificar empresas con un alto potencial de venta online, se decidió apoyarse en fuentes especializadas. Para la obtención de la información necesaria se utilizaron los servicios de Ecommercedb.com, empresa dedicada a brindar información completa, así como cifras claves y actuales sobre las tiendas en línea más destacadas del mundo. Además, a partir de la profundización en la categorización del tipo de productos que buscamos licores, alimentos y bebidas, nos permitió compilar una lista de las empresas más populares en México.

A continuación, en la Tabla 1 se presentan los 15 sitios escogidos ordenados en orden alfabético. De esta forma, se ofrece una visión más completa de los operadores de comercio electrónico más grandes del país.

Portal
https://abarrotero.com
https://alcca.mx
https://autoserviciolaplaya.com
https://cavadelduero.com
https://comercializadorajm.com
https://corpovino.com.mx
https://ibarramayoreo.com
https://lacastellana.com
https://mx.frubana.com
https://www.bodegaaurrera.com.mx
https://www.bodegasalianza.com
https://www.chedraui.com.mx
https://www.consuvino.com.mx
https://www.etiendas3b.com.mx
https://www.laeuropea.com.mx

Tabla 1. Listado de páginas web-empresas - Creación propia.

2.2.2 Lenguaje Python

Se trata de un lenguaje de programación de alto nivel, interpretado y orientado a objetos que es conocido por su sintaxis simple y fácil de leer, lo que lo hace ideal para este proyecto de programación. Este lenguaje es utilizado en una amplia variedad de aplicaciones, como desarrollo web, análisis de datos, inteligencia artificial y automatización de tareas. Así mismo Python nos ofrece una gran productividad en todas las fases del ciclo de vida del software: análisis, diseño, creación de prototipos, codificación, pruebas, depuración, puesta a punto, documentación y por supuesto, mantenimiento (Martelli et al., 2023).

2.2.3 Selección Scrapy

La elección de la plataforma para el desarrollo de los bots en este proyecto fue determinada bajo la premisa de garantizar

un alto grado de libertad en su desarrollo, rapidez y potencia. Por consiguiente, se estableció la utilización de Scrapy como framework de trabajo, en virtud de tratarse de una herramienta de código abierto, que corre sobre Python y que ofrece una alta configurabilidad. Además, por ser reconocido gracias a su capacidad para raspar grandes cantidades de datos de manera eficiente, lo que lo hace una opción altamente adecuada para el alcance del presente proyecto.

2.2.4 Visión general de la arquitectura de Scrapy

La arquitectura de Scrapy se fundamenta en un diseño modular y escalable, lo que otorga un elevado grado de flexibilidad. Se compone de varios elementos fundamentales, tales como el motor Scrapy, el gestor de nuestras arañas, el sistema de pipelines y el middleware. A continuación, en la **Figura 1**. Se muestra el funcionamiento e interacción de los componentes, seguido de una descripción detallada de los mismos.

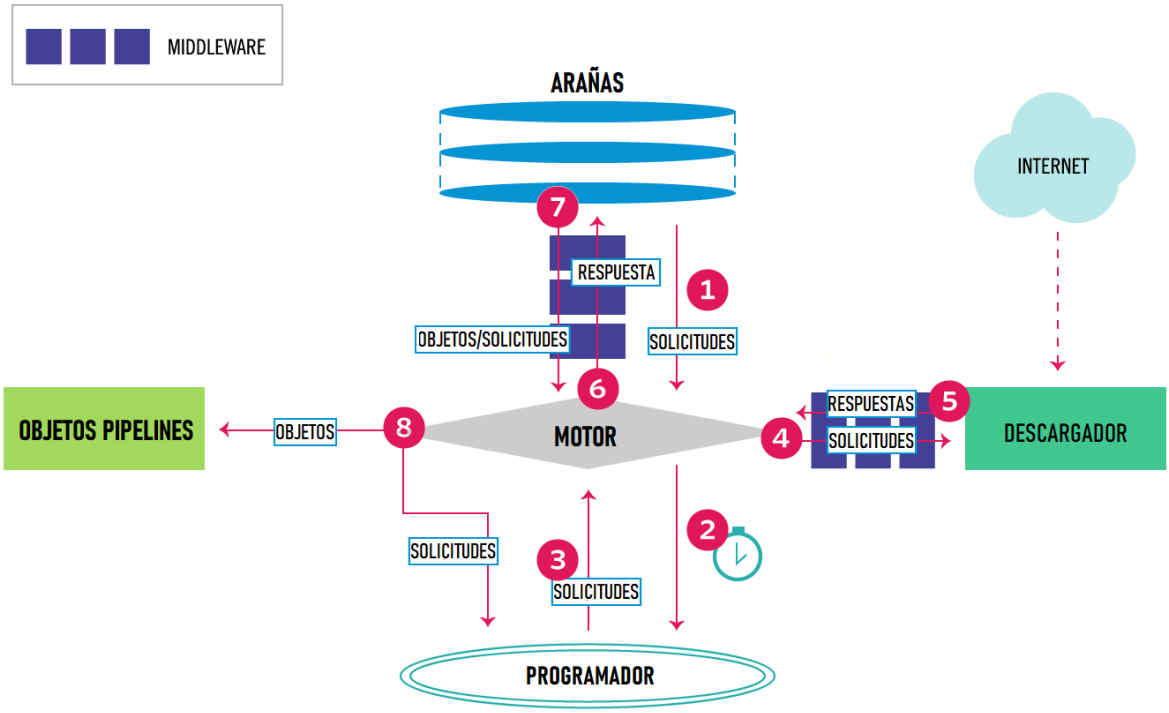

Figura 1. Arquitectura de Scrapy (*Architecture overview — Scrapy 2.8.0 documentation*, s. f.)

El flujo de datos en Scrapy está controlado por el motor de ejecución, el cual sigue el siguiente proceso:

1. Las solicitudes iniciales para llevar a cabo el rastreo son recibidas por el motor a través de las arañas.
2. El motor de ejecución de Scrapy programa las solicitudes en el programador y solicita la exploración de las próximas solicitudes.
3. El programador envía al motor las solicitudes siguientes.
4. El motor de Scrapy se encarga de dirigir las solicitudes hacia el descargador, el cual a su vez atraviesa los Middlewares del descargador mediante una función “process_request()”.
5. Después de que la página web se completa su descarga, el descargador procede a crear una respuesta que incluye dicha página y la envía al motor de Scrapy. Antes de llegar al motor, la respuesta pasa por los Middlewares del descargador, lo cual se puede observar en la función “process_response()”.
6. Después de que el descargador obtiene la respuesta, el motor la recibe y la dirige hacia el Spider para su procesamiento. En este proceso, la respuesta también atraviesa el Middleware de la araña, mediante la función "process_spider_input()".
7. Al recibir la respuesta, la araña procesa los datos y envía los elementos extraídos junto con las solicitudes adicionales a seguir al motor de ejecución. Este proceso implica la intervención del middleware de la araña a través de la función "process_spider_output()". De esta manera, el motor de ejecución de Scrapy puede coordinar eficazmente el flujo de datos y controlar el proceso de web scraping de manera efectiva.
8. Una vez que los ítems son procesados por el motor, estos son transferidos a los objetos pipelines. Asimismo, las

solicitudes procesadas son enviadas al programador y se indaga acerca de posibles solicitudes a seguir.

9. El proceso continúa iterando desde el tercer paso hasta que no haya más solicitudes del programador. En otras palabras, la repetición de este proceso depende de la existencia de peticiones del programador para realizar la extracción de datos en Scrapy.

2.2.5 Componentes Scrapy

- **Motor de scrapy**

El control del flujo de datos entre los diferentes componentes del sistema y la activación de eventos cuando se ejecutan ciertas acciones son responsabilidad del motor.

- **Programador**

La función de esta entidad es recibir solicitudes del motor y programarlas para su posterior procesamiento, de modo que puedan ser proporcionadas al motor cuando éste las requiera.

- **Desacargador**

Su responsabilidad es la de adquirir las páginas web y remitirlas al motor, el cual se encarga de transmitirlas a las arañas correspondientes.

- **Arañas**

Los usuarios de Scrapy pueden crear sus propias clases personalizadas, conocidas como "arañas", que se encargan de analizar las respuestas recibidas y extraer los elementos relevantes, así como realizar solicitudes adicionales según sea necesario.

- **Objetos Pipeline**

El proceso encargado de manejar los ítems extraídos por las arañas es conocido como el "Objetos Pipeline". Su función principal es procesar los datos para asegurar que estén limpios, validados y almacenados correctamente en una base de datos u otro medio de persistencia.

- **Descargador middlewares**

Los middlewares de descarga son componentes específicos que se ubican entre el motor y el descargador, y se encargan de procesar tanto las solicitudes que pasan del motor al descargador como las respuestas que pasan del descargador al motor.

- **Araña middlewares**

Los middlewares de araña son componentes especializados que actúan como intermediarios entre el motor de Scrapy y las arañas, y tienen la capacidad de gestionar tanto las entradas (respuestas) como las salidas (elementos y solicitudes) que éstas producen.

2.3 METODOLOGÍA

La metodología que se usará para el desarrollo del proyecto será Design Thinking, esta metodología facilita el desarrollo de un producto y la solución de problemas explotando al máximo las capacidades del equipo. Sin embargo, el deber es enfocar esta metodología al cumplimiento de los objetivos específicos y así poder cumplir el objetivo general, por ello será necesario alterar algunos pasos en el ejercicio de la metodología para adaptarla a dichos objetivos.

Durante la primera fase del proceso de Design Thinking, conocida como "Empatizar", se puso un enfoque especial en comprender las necesidades de los funcionarios que utilizaban el mercado en línea. Se identificó que la problemática principal estaba relacionada con la regulación y selección de los precios adecuados para los productos. A través de un análisis detallado, se determinó que, aunque las necesidades de cada mercado de productos eran diferentes, podían ajustarse a un marco inicial. Este marco incluía la definición de los sitios web objetivos, la identificación de palabras o etiquetas clave que se deseaban extraer de los sitios, y finalmente, la utilización de iteraciones para ajustar y socializar con los clientes los resultados de las campañas de raspado de los sitios web, siguiendo la metodología del Design Thinking.

Una vez que se han comprendido claramente los requerimientos específicos de cada cliente, recolectados y analizados en la fase de empatía del Design Thinking, y considerando que el primer objetivo es consolidar sitios web y verificar la legalidad de la práctica de scraping, se procede a la fase de definir. Durante esta etapa, es necesario realizar un análisis exhaustivo de los sitios web objetivo que se van a raspar, revisando detenidamente los términos y condiciones de cada sitio para identificar posibles restricciones o políticas que prohíban la extracción de datos y evitar así problemas legales.

Además, en esta fase se definen los sitios web o competidores específicos de cada mercado al cual está dirigida cada campaña. Se determinan los ítems principales que se desean extraer de cada sitio, como el nombre del producto, el precio, la descripción, entre otros. Es importante realizar una cuidadosa planificación y definición de los elementos que se van a raspar, para asegurar que se obtenga la información relevante de manera efectiva y eficiente. Esta fase sienta las bases para el desarrollo de las campañas de scraping, asegurando que se cumplan los objetivos establecidos y se eviten posibles conflictos legales.

La fase de idear en la metodología Design Thinking propone abordar los problemas mediante técnicas como lluvias de ideas, mapas mentales y otras, por lo que utilizamos este espacio para enfrentar el siguiente objetivo, que es definir la librería óptima ajustada al lenguaje Python para el desarrollo del proyecto.

Es importante tener en cuenta que existen varias herramientas y enfoques para el raspado de datos con Python, pero debemos considerar los recursos informáticos y el presupuesto disponible para implementar todos los componentes necesarios, así como los requerimientos específicos de cada cliente. Las diferentes librerías de Python enfocadas en el raspado web ofrecen soluciones para una amplia variedad de problemas, dependiendo de cómo se combinen.

Como se mencionó en el estado del arte, una de las librerías más utilizadas para esta práctica es Selenium. Esta librería contiene un conjunto de herramientas y scripts especializados en el raspado de datos, y ayuda a seleccionar el script que mejor se adapte a las necesidades del usuario. Todos los scripts de Python en Selenium son ejemplos reales listos para probar, y se explican detalladamente en las declaraciones de problemas (Raghavendra, 2021).

La elección de la librería, los scripts y las pruebas de uso corresponden a las etapas de ideación y prototipado de la metodología Design Thinking.

Una vez que se haya practicado lo suficiente con los scripts de la librería seleccionada y se hayan recolectado varias muestras de datos extraídos de páginas web, se procede a la etapa de pruebas (testing). En esta etapa, se realiza un proceso manual para comprobar la veracidad de toda la información recolectada, cumpliendo así con el objetivo de analizar los resultados y la precisión de la extracción de datos.

Esta etapa demanda un tiempo considerable, pero proporciona información valiosa para el resto del proceso. El objetivo es identificar posibles fallas y errores que puedan ocurrir al extraer datos con determinados scripts de la librería seleccionada. Además, es crucial tener en cuenta la estructura de los datos obtenidos. Aunque esta etapa sea de pruebas, se ha definido aplicar un proceso de normalización de los datos en este punto.

La normalización implica estandarizar la información, suprimiendo o reemplazando caracteres irregulares, como acentos, combinaciones de minúsculas y mayúsculas, o la separación de miles con comas o puntos. Este proceso de normalización garantiza que los datos puedan ser interpretados correctamente por las herramientas de inteligencia de negocio y utilizados de manera efectiva en análisis posteriores.

Continuando con la metodología propuesta para esta investigación se ha llegado al punto de presentar los resultados de tal forma que se genere una buena experiencia de usuario para los clientes, fácil de usar, y fácil de interpretar, dando cumplimiento al objetivo construir un prototipo de Dashboard mediante herramientas de inteligencia empresarial para la visualización de datos. Para ello se usarán las herramientas populares de inteligencia de negocio BI, por sus siglas en ingles Business Intelligence. Un ejemplo de una de las herramientas más usadas para inteligencia de negocios es Power BI de Microsoft. Para tener éxito, los tableros interactivos deben satisfacer a los usuarios y producir una experiencia de usuario agradable (Muppidi et al., 2022). Se cumplirá el objetivo mediante la aplicación de un modelo propuesto en la construcción del Dashboard que garantice una experiencia de usuario óptima. Los criterios que se definieron para la construcción del Dashboard fueron:

- Identificar la naturaleza del mercado y clasificar los tableros en función de su propósito, como tableros estratégicos, operativos o analíticos, para asegurar que se ajusten a las necesidades específicas del negocio (Sorapure, 2023).
- Elegir el tipo de visualizaciones para que el tablero resulte eficiente, teniendo en cuenta que la eficiencia de un tablero se ocupa de la adquisición de información, no de su interpretación (Dalpiaz et al., 2020). Algunos ejemplos de visualizaciones incluyen, tablas, gráficos, indicadores.
- Crear un diseño intuitivo para el Dashboard que permita a los usuarios aprender rápidamente cómo interactuar

con él. Se propone una disposición lógica de los componentes que facilite la identificación de la información relevante por parte de los usuarios.

- Diseñar una interfaz flexible que permita una interacción intuitiva y sencilla con los datos.
- Proporcionar opciones prediseñadas y personalizar los diseños para adaptarse a las necesidades y preferencias de los clientes.

La metodología aplicada en este proyecto es iterativa, cada cliente tiene unos requerimientos específicos del core de su negocio, sin embargo, se ha logrado estandarizar el proceso de una forma que se puedan aplicar los mismos pasos, sin comprometer el rendimiento y obteniendo resultados de calidad para los clientes. A continuación, en la **Figura 2** se presenta un diagrama de flujo que describe la metodología de forma iterativa.

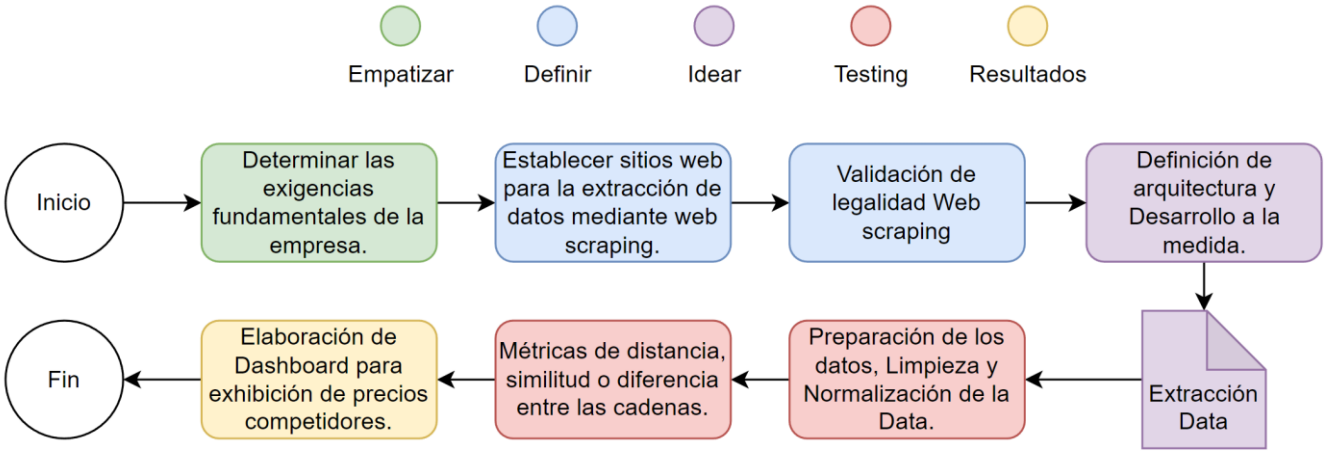

Figura 2. Diagrama de flujo – Metodología Creación propia

Se puede apreciar que inmersos en el diagrama están los objetivos específicos de esta investigación y así se dará cumplimiento al objetivo general desarrollar herramienta de Web Scraping para proporcionar tableros de datos confiables a las empresas de Ecommerce respecto a sus competidores.

2.4 DESARROLLO

Una vez establecido el alcance del proyecto y elaborado el listado de portales objetivo para el raspado de información, se dio inicio al desarrollo de este. En una primera instancia, se llevó a cabo en un entorno local de un ordenador personal bajo el editor de código fuente Visual Studio Code. A continuación, en la **Figura 3**, se presenta el flujo de proceso que muestra las 5 etapas del proceso del bot desarrollado. Después, se describen detalladamente cada una de estas etapas.

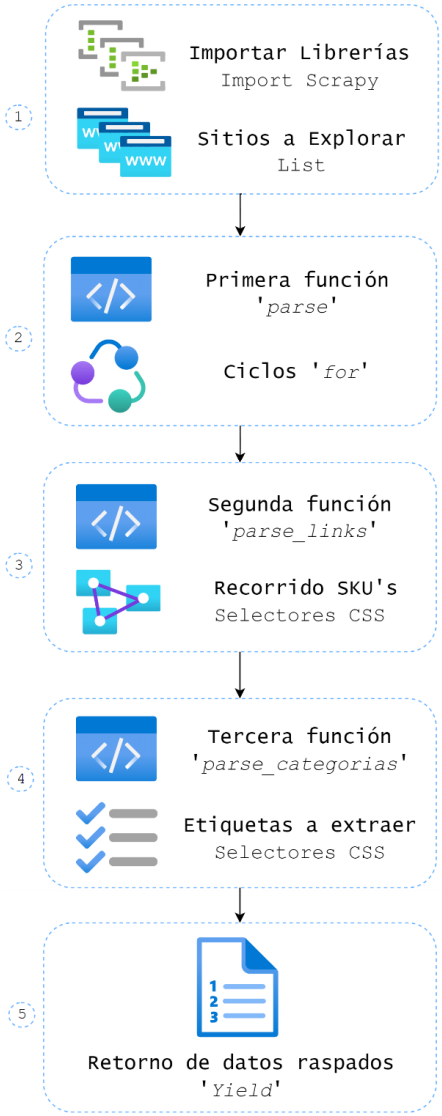

Figura 3. Diagrama de flujo del desarrollo - Creación propia

En la etapa 1 de la **Figura 3**, se realiza inicialmente la importación de las librerías necesarias para el desarrollo del proyecto. Seguidamente, se define una función denominada "URL", la cual contiene una variable de tipo 'array' donde se almacena el conjunto de páginas que se van a explorar mediante el uso del "spyder".

Seguidamente 2, 3 y 4 de la **Figura 3** se han creado tres métodos, a saber: "parse", "parse_links" y "parse_categorias". Es importante destacar que Scrapy dispone de varios métodos por defecto que pueden ser utilizados como base para ajustarse a las necesidades específicas de un proyecto. En seguida se describen dichos métodos ajustados a la medida de la necesidad.

2.4.1 Métodos de Scrapy

2.4.1.1 **Primera función** parse(self, response)

En la etapa 2 de la **Figura 3**, el presente método constituye la función por defecto que Scrapy invoca para procesar las respuestas extraídas en este caso a la función "parse_links". En aquellos casos en que las solicitudes no especifiquen otra función. En tal sentido, resulta imperativo que este método retorne un objeto que contenga uno o varios valores pertinentes.

En el caso presente, se ha optado por una solución que implica la inclusión de dos ciclos for. Esta solución se ha planteado como una respuesta a la necesidad de recorrer el array de URLs previamente definido, así como para abordar la

cuestión de la paginación correspondiente.

2.4.1.2 Segunda función Parse_links(self, response)

En la etapa 3 de la **Figura 3**, el método en cuestión ha sido desarrollado con el propósito de proporcionar una solución para listar cada URL de producto como se evidencia enseguida en la **Figura 4**.

Figura 4. Producto especifico con ítems importantes en la extracción (Grupo Abarrotero Punto Com SA de CV & Martínez, 2022)

Dicho producto se encuentra contenido en la cuadricula de productos sobre una URL principal que se está explorando como se muestra en la **Figura 5**, esta extracción de links se realiza a través de un selector CSS donde su respuesta será enviada a la función "parse_categorias".

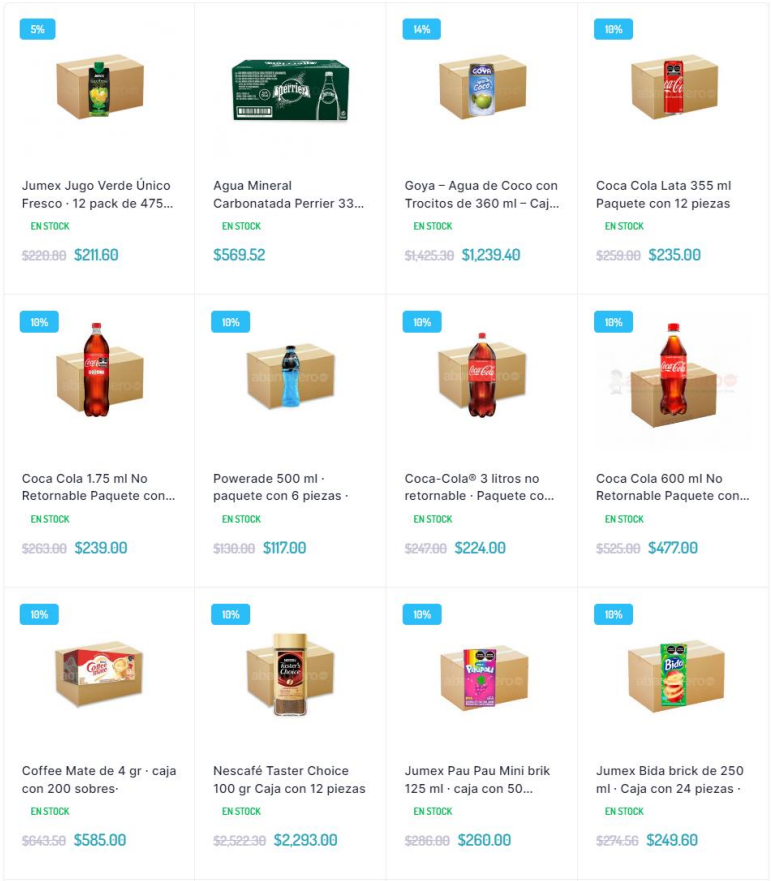

Figura 5. Productos y Categorías para el raspado de datos (Grupo Abarrotero Punto Com SA de CV & Martínez, 2022)

2.4.1.3 Tercera función Parse_categorias(self, response)

En la etapa 4 de la **Figura 3** es donde se elevan los selectores CSS de acuerdo con los patrones descubiertos en la exploración manual del código HTML del sitio. De esta manera, se puede acceder a los elementos específicos tales como el nombre, la marca, el identificador del producto, el precio y la URL. Además, se incorporan las sentencias "Try" y "Except" para el manejo de excepciones en caso de que los elementos buscados se encuentren en otra etiqueta del código.

Finalmente, en la etapa 5 de la **Figura 3**, a través de la expresión "yield", se retorna la información y se pausa la ejecución de la función. El estado de la función se guarda hasta que sea llamada nuevamente a través de los ciclos "for".

2.4.2 Proceso de matching de productos

Para llevar a cabo la comparación entre el catálogo de productos y el catálogo de información extraído de los portales correspondientes, es necesario realizar en primer lugar un proceso de emparejamiento entre ambos listados. Este proceso permitirá una comparación efectiva de los datos. Para lograr este objetivo, se lleva a cabo una preparación de los datos y normalización, seguida de un proceso de comparación de cadenas mediante métricas de distancia.

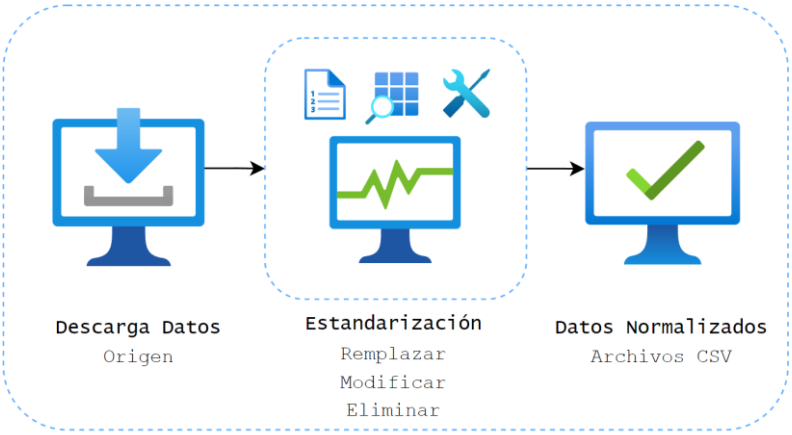

Figura 6. Diagrama de flujo del desarrollo - Creación propia

2.4.3 Preparación de los datos

La limpieza y normalización de datos es crucial en la preparación para su análisis, ya que los datos no normalizados y sucios pueden producir resultados inexactos e inconsistentes. Para asegurar la calidad de los datos, es importante garantizar que estén en un formato coherente y fácilmente interpretable para el análisis posterior.

Como se muestra en la **Figura 6**. Dentro del proceso de preparación de datos, se utilizan diversas técnicas para limpiar y normalizar cadenas de texto, las cuales son parte integral del desarrollo y se aplican tanto sobre la información descargada de los sitios web como del catálogo de productos a comparar.

- Remplazo de mayúsculas a minúsculas.
- Eliminación de los signos de puntuación, caracteres especiales, acentos, etc.
- Eliminar palabras vacías.
- Eliminar tokens duplicados.
- Homogeneizar unidades de medida.
- Homogeneizar abreviaturas.

2.4.4 Métricas de distancia

Para llevar a cabo el análisis de cadenas, se hizo uso de métricas de distancia para medir la del catálogo de productos y la información extraída. Estas métricas son funciones matemáticas que para este caso toman estas dos entradas como input y producen un output de valor numérico que representa la distancia entre ellas. Para el presente proyecto se usaron dos métricas descritas a continuación.

2.4.4.1 Distancia de Levenshtein

También conocida como distancia de edición, mide el número mínimo de operaciones de edición (inserción, eliminación o sustitución) necesarias para transformar una cadena en otra.

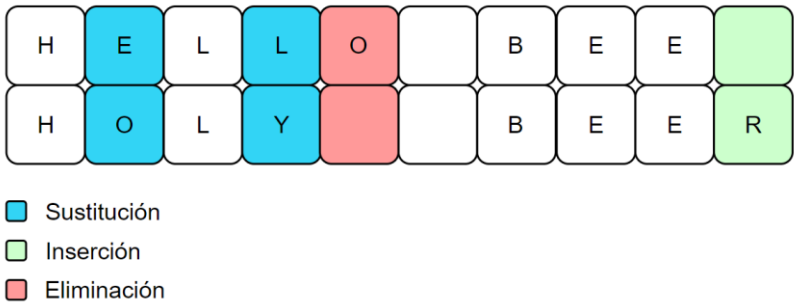

Figura 7. Demostración Levenshetein - Creación propia.

Como se observa en la **Figura 7**. Para la transformación de la cadena 'Hello Bee' a 'Holy Beer' hicieron falta 4 pasos.

2.4.4.2 Métrica Dice-Sorensen

También conocido como coeficiente de Dice o índice de Sørensen Dice. Es una medida de similitud utilizada en el análisis de cadenas y otros campos científicos. Esta métrica nos sirve para medir la similitud de dos conjuntos de elementos y se define como la intersección de los conjuntos dividida por el doble de la suma de los tamaños de los conjuntos.

En el contexto del proyecto para el análisis de cadenas, la métrica Dice-Sorensen se utiliza para medir la similitud entre las cadenas del catálogo de productos y la información extraída. Se calcula como el doble del número de caracteres comunes en ambas cadenas dividido por la suma de las longitudes de las dos cadenas.

El coeficiente de Dice-Sorensen varía entre 0 y 1, donde un valor de 1 indica que las dos cadenas son idénticas y un valor de 0 indica que las dos cadenas no tienen caracteres en común.

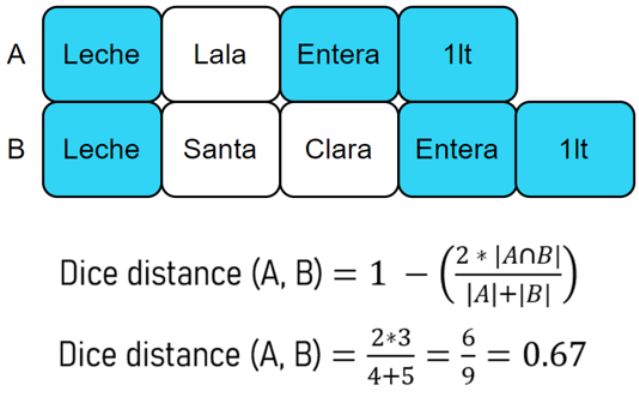

Figura 8. Demostración Dice-Sorensen - Creación propia.

Conforme se puede apreciar en la **Figura 8**, la medición de similitud entre las cadenas de texto "Leche Lala Entera 1lt" y "Leche Santa Clara Entera 1lt" es de 0.67. Este ejemplo exhibe que las cadenas se acercan más a la identidad que a la ausencia de semejanza.

3. RESULTADOS

Para el éxito empresarial es fundamental la recopilación y análisis de información precisa y actualizada sobre los competidores. Aunque esta tarea puede requerir mucha dedicación de forma manual, existen métodos automatizados como el web scraping, una solución eficiente que permite extraer datos de múltiples sitios web de forma rápida y precisa.

El objetivo de este trabajo es desarrollar una herramienta de web scraping que extraiga datos de precios, descripciones y características de productos de competidores para presentarlos en un dashboard. Con el propósito de ayudar a las empresas a tomar decisiones según el comportamiento del mercado, mejorando su estrategia de precios y obteniendo información valiosa sobre las preferencias de los consumidores.

Como resultado de la investigación, se logró consolidar varios sitios web y verificar su legalidad para la práctica de scraping. Se diseñó un plan de acción que incluyó la recopilación de los sitios a raspar y la obtención de una fuente de datos única y precisa. También se evaluaron diferentes librerías de Python y se seleccionó Scrapy como la más adecuada, integrándola al proyecto y verificando su correcto funcionamiento. Además, se analizó la calidad y veracidad de los datos extraídos, resolviendo problemas en la extracción de datos y verificando su precisión y fiabilidad para su uso en el proyecto. Finalmente, se construyó un prototipo de dashboard mediante herramientas de inteligencia empresarial que permitió visualizar los datos de manera efectiva y brindar información valiosa y relevante para los usuarios.

3.1 Consolidar listado de empresas que sean competidores directos, con el fin de obtener información detallada acerca de los precios de sus productos.

Se ha reconocido que evaluar los precios de los productos ofrecidos por competidores directos es una labor compleja que requiere dedicación y tiempo, al punto de que existen servicios especializados dedicados a esta tarea. Durante el curso de esta investigación, se ha tenido en cuenta la relevancia de los competidores seleccionados para llevar a cabo la práctica de Scraping web. Con este fin, se ha optado por utilizar un servicio en línea especializado, que ha proporcionado un catálogo de 15 sitios web que garantizan la obtención de datos valiosos a través de dicho proceso. Como resultado, se ha llegado a la conclusión de que el estudio y clasificación de los métodos utilizados por la competencia son una inversión justificada de tiempo y recursos, ya que proporcionan los fundamentos necesarios para llevar a cabo una práctica efectiva de web scraping.

3.2 Desarrollar una herramienta de web Scraping que permita extraer los precios de los competidores seleccionados de forma automatizada y regular.

El estado actual de la investigación revela que se han llevado a cabo numerosos proyectos similares utilizando técnicas de raspado de datos, y se ha identificado que el lenguaje de programación Python es una herramienta clave en esta práctica. Python se utiliza ampliamente en diversos campos, como inteligencia artificial y automatización, lo que ha llevado al desarrollo de librerías robustas como Scrapy, específicamente diseñadas para el web scraping.

Como resultado de este enfoque, se ha desarrollado una solución altamente efectiva para el web scraping, cuyo funcionamiento se puede entender mejor mediante el siguiente diagrama de caja negra.

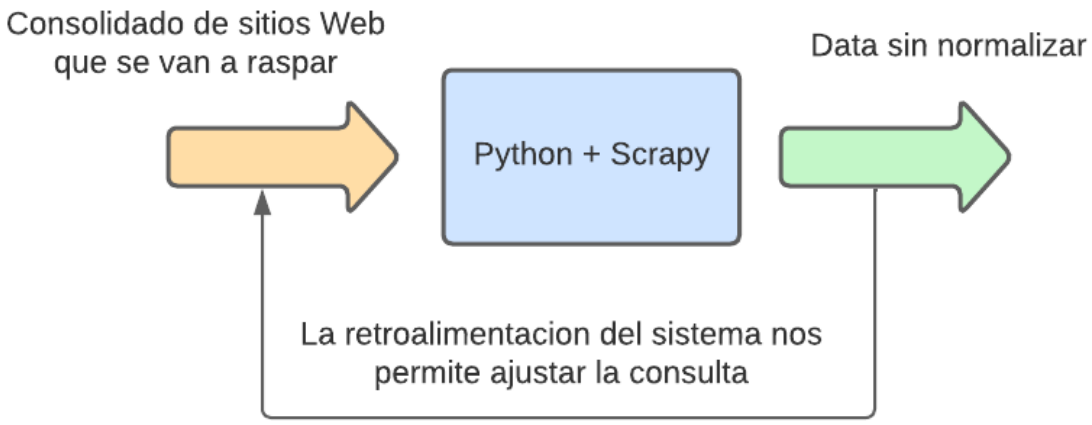

Figura 9. Diagrama de caja negra - Creación propia.

3.3 Transformar los datos obtenidos mediante el web scraping en una estructura definida que facilite su manipulación con herramientas de inteligencia de negocios.

El proceso de raspado web es solo el primer paso para extraer información relevante de los sitios seleccionados. Sin embargo, los datos extraídos deben someterse a un proceso de normalización y limpieza para obtener información valiosa a partir de ellos.

Este objetivo ha permitido identificar y consolidar un proceso que se aplica a cualquier tipo de campaña de raspado web, como la eliminación de palabras vacías y acentos, así como la homogeneización de las unidades de medida. Es importante destacar que estos son aspectos básicos y aplicables en la mayoría de los casos, pero en algunas situaciones, el proceso de normalización puede requerir el uso de técnicas más complejas para darle la forma necesaria a los datos.

Type	Store	Consultation	Category	Product_id_	Product_id_	Desc	Unit_pr	Quantity	Price	Url	Img_Site	Img_Product
Online	Chedraui	10/04/2023	Agua Natural Epura Garraf		3041555				34.50	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Natural Bonafont en		3041562				47.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Natural Epura 6 Bote		3467356				32.60	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Mineral Penafiel 2L		3062312				23.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Natural Epura Solo L		3041567				35.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Mineral Topo Chico		3001984				18.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Bonafont Agua Natural 6x		3679056				27.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Mineral Penafiel Sif		3361842				18.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Purificada Selecto B		3698841				75.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Natural Bonafont co		3732475				45.90	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Bonafont Agua Natural 6L		3163754				31.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua de Manantial Santa I		3041598				42.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Mineral Topo Chico		3001982				22.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Mineral Penafiel 1L		3062314				16.50	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Natural Gerber Bote		3041533				29.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Natural Epura 12 Bot		3532003				39.50	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Natural Purificada C		3277687				41.50	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua Mineral Penafiel Na		3062316				13.50	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Bonafont Agua Natural 8x		3435664				36.50	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Agua de Manantial Santa I		3041530				27.00	https://www.ched	Img_Site	https://ched
Online	Chedraui	10/04/2023	Bonafont Agua Natural 10		3698080				42.00	https://www.ched	Img_Site	https://ched

Figura 10. Resultados extraídos - Creación propia.

En la **Figura 10**, se puede observar que la información tiene una estructura uniforme. Las palabras están formateadas como nombres propios, los signos de miles no se mezclan con punto y coma, y las fechas siguen el formato día/mes/año. Esto asegura que los datos puedan ser interpretados fácilmente por herramientas de inteligencia de negocio.

3.4 Elaborar un Dashboard que sea fácil de interpretar y que exhiba de manera precisa los precios ofrecidos por los competidores, con el fin de facilitar el proceso de toma de decisiones sobre productos.

Como resultado de la cuidadosa selección de las herramientas de inteligencia empresarial más adecuadas para el proyecto, se pudo diseñar y desarrollar un prototipo de dashboard que permitió visualizar los datos extraídos de manera efectiva. Se trabajó para asegurarse de que el dashboard fuera fácil de usar y ofreciera información valiosa y relevante para los usuarios.

% Confidence	MKP SKU Code	MKP SKU	Competitor SKU	Competitor	Competitor Price	MKP Price	Price Diff	Price Ratio
100 %	3009513	COCA COLA 355ML - 12PZS	Coca Cola Lata 355 ml Paquete con 12 piezas Coca Cola	Abarrotero	235,00	162,01	-72,99	0,69
100 %	3009505	CLAMATO 296ML - 24 PZS	Clamato 24/296ml Clamato	Alcca	3.323,00	314,58	-3.008,42	0,09
100 %	3009523	ELECTROLIT UVA 625ML - 12PZS	Electrolit Uva 625 ml Caja con 12 Electrolit	Alcca	209,00	208,85	-0,15	1,00
100 %	3009529	SOPA MARUCHAN CAMARON - 12 PZS	Sopa Maruchan Camaron 12/64g Maruchan	Alcca	159,00	160,00	1,00	1,01
100 %	3009536	SOPA MARUCHAN RES - 12 PZS	Sopa Maruchan Res 12/64g Maruchan	Alcca	159,00	160,00	1,00	1,01
100 %	3009559	ELECTROLIT FRESA 625ML - 12PZS	Electrolit Fresa 625 ml Caja con 12 Electrolit	Alcca	209,00	208,85	-0,15	1,00
100 %	3009560	ELECTROLIT COCO 625ML - 12PZS	Electrolit Coco 625 ml Caja con 12 Electrolit	Alcca	209,00	208,85	-0,15	1,00
100 %	3009581	PEPSI COLA REFRESCO 2.5L - 8 PZAS	Pepsi 8/2.5L Pepsi	Alcca	207,00	204,00	-3,00	0,99
100 %	3009771	ELECTROLIT MORA AZUL 625 ML - 12 PZS	Electrolit Mora Azul 625 ml Caja con 12 Electrolit	Alcca	209,00	208,85	-0,15	1,00
100 %	3010259	ACEITE NUTRIOLI 800 ML - 12 PZS	Aceite Nutrioli 12/800ml Nutrioli	Alcca	523,50	562,80	39,30	1,08
100 %	3010260	ACEITE NUTRIOLI 400 ML - 12 PZS	Aceite nutrioli 12/400ml	Alcca	289,50	291,60	2,10	1,01
100 %	3010262	ACEITE AVE 800 ML - 12 PZS	Aceite Ave 12/800 ml	Alcca	489,00	458,40	-30,60	0,94
100 %	3010276	LALA LECHE ENTERA 1L - 12 PZS	Leche Lala Entera 12/1L Lala	Alcca	252,00	262,00	10,00	1,04
100 %	3010278	LALA LECHE DESLACTOSADA 1L - 12 PZS	Leche Lala Deslactosada 12/1L Lala	Alcca	252,00	261,30	9,30	1,04
100 %	3009782	JOSE CUERVO TEQUILA TRAD PLATA 950ML-1PZ	Tequila Cuervo Tradicional Plata 950 ml	Alianza	370,00	328,96	-41,04	0,89
100 %	3009794	BUCHANANS 12 ANOS WHISKY 750 ML - 1PZ	Whisky Buchanans 12 Anos 750ml	Alianza	945,00	686,72	-258,28	0,73
100 %	3009802	JACK DANIELS WHISKY 700ML-1PZ	Whiskey Jack Daniel's 700 ml	Alianza	465,00	379,29	-85,71	0,82
100 %	3009810	CHIVAS REGAL WHISKY 12 ANOS 750ML-1PZ	Whisky Chivas Regal 12 Anos 750 ml	Alianza	815,00	610,00	-205,00	0,75
100 %	3009795	CAZADORES TEQUILA REPOSADO 950 ML - 1PZ	Tequila Cazadores Reposado 950 ml	Bodega_Aurrera	499,00	259,00	-240,00	0,52
100 %	3009798	GRAN CENTENARIO TEQUILA PLATA 950ML-1PZ	Tequila Gran Centenario Plata 950 ml Gran Centenario Plata	Bodega_Aurrera	498,99	302,46	-196,53	0,61
100 %	3009810	CHIVAS REGAL WHISKY 12 ANOS 750ML-1PZ	Whisky Chivas Regal 12 anos 750ml CHIVAS REGAL Botella 750ml	Bodega_Aurrera	549,00	610,00	61,00	1,11
100 %	3009557	WHISKY BLACK & WHITE 700ML - 1 PZ	Whisky Black&White 700ml	Cava_del_Duero	214,00	186,99	-27,01	0,87
100 %	3009578	BACARDI RON BLANCO 980 ML - 1 BOTELLA	Ron Bacardi Carta Blanca 980ml	Cava_del_Duero	218,00	162,00	-56,00	0,74
100 %	3009674	RON BACARDI CARTA BLANCA 750ML - 1PZ	Ron Bacardi Carta Blanca 750ml	Cava_del_Duero	186,00	138,01	-47,99	0,74

Figura 11. Dashboard para el monitoreo de productos - Creación propia.

4. CONCLUSIONES

4.1 Conclusión primer objetivo

El primer objetivo determina la importancia de los competidores, con el fin de obtener información valiosa que pueda ser aprovechada para desarrollar estrategias efectivas. Es crucial tener presente que la competencia es un factor fundamental y que conocerla detalladamente puede tener un gran impacto en la obtención de los datos. Esto también puede permitir identificar oportunidades de crecimiento, expansión en el mercado al conocer las necesidades y deseos de los clientes para elaborar estrategias de marketing.

4.2 Conclusión segundo objetivo

En el desarrollo se comprobó que Python es uno de los lenguajes mejor optimizados para trabajar en esta implementación, ya que cuenta con una gran cantidad de librerías adecuadas para todo tipo de propósitos, lo que posibilita resolver los problemas de manera flexible y efectiva.

4.3 Conclusión tercer objetivo

La transformación de los datos obtenidos mediante el web scraping en una estructura definida es un paso fundamental para facilitar su manipulación y análisis con herramientas de inteligencia de negocios. Al estructurar los datos de manera adecuada, se puede asegurar que sean más accesibles y comprensibles para el análisis y la toma de decisiones. La transformación también permite la integración de datos de múltiples fuentes, lo que proporciona una vista más completa y

precisa de la información.

4.4 Conclusión cuarto objetivo

La elaboración de un Dashboard puede ayudar a las empresas a tomar decisiones sobre sus estrategias e identificar tendencias en los precios de la competencia. En general, un Dashboard bien diseñado puede ser una herramienta valiosa para cualquier empresa que busque mejorar su posición en el mercado mediante la toma de decisiones de precios y productos.

4.5 Recapitulación de resultados

En general la implementación de la solución propuesta puede tener un gran impacto en la toma decisiones informadas y estratégicas en cuanto a precios y marketing. Al automatizar el proceso de monitoreo y análisis de precios, la empresa puede obtener información valiosa en tiempo real sobre el mercado y su competencia, lo que le permitirá ajustar sus estrategias de precios y marketing de manera más efectiva en la toma de decisiones más informadas y estratégicas para alcanzar sus objetivos comerciales. Además, esta solución puede ahorrar tiempo y recursos valiosos para la empresa al eliminar la necesidad de realizar tareas de monitoreo manualmente.

REFERENCIAS

Alasmari, A., Alhothali, A., & Allinjawi, A. (2022). Hybrid machine learning approach for Arabic medical web page credibility assessment. *Health Informatics Journal*, *28*(1), 146045822110709. https://doi.org/10.1177/14604582211070998

Architecture overview—Scrapy 2.8.0 documentation. (s. f.). Recuperado 20 de abril de 2023, de https://docs.scrapy.org/en/latest/topics/architecture.html

Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C., & Cunningham, S. (2022). Indicators on firm level innovation activities from web scraped data. *Data in Brief*, *42*, 108246. https://doi.org/10.1016/j.dib.2022.108246

Binanto, I., Warnars, H. L. H. S., Sianipar, N. F., & Budiharto, W. (2022). *Webscraping Data Labeling System on Liquid Chromatography-Mass Spectrometry of Rodent Tuber for Efficiency of Supervised Learning Preprocessing* (N.º 01). ICIC International 学会. https://doi.org/10.24507/icicelb.13.01.107

Dalpiaz, F., Zdravkovic, J., & Loucopoulos, P. (Eds.). (2020). *Research Challenges in Information Science: 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23–25, 2020, Proceedings* (Vol. 385). Springer International Publishing. https://doi.org/10.1007/978-3-030-50316-1

Fazekas, M., Abdou, A., Kazmina, Y., & Regős, N. (2022). Development aid contracts database: World Bank, Inter-American Development Bank, and EuropeAid. *Data in Brief*, *42*, 108121.

https://doi.org/10.1016/j.dib.2022.108121

Gallagher, J. R., & Beveridge, A. (2022). Project-Oriented Web Scraping in Technical Communication Research. *Journal of Business and Technical Communication*, *36*(2), 231-250. https://doi.org/10.1177/10506519211064619

Grupo Abarrotero Punto Com SA de CV, & Martínez, P. (2022, febrero 4). *Abarrotero.com*. Mayoristas de Abarrotes. https://abarrotero.com/

Han, S., & Anderson, C. K. (2021). Web Scraping for Hospitality Research: Overview, Opportunities, and Implications. *Cornell Hospitality Quarterly*, *62*(1), 89-104. https://doi.org/10.1177/1938965520973587

IBISWorld, I. (s. f.). *IBISWorld—Industry market research, reports, and statistics*. Recuperado 1 de marzo de 2023, de https://help.ibisworld.com/s/article/industry-research-reports

Khan, S., Miah, S. J., & Gao, J. (2020). Big data analytics in manufacturing: A systematic review. *Journal of Business Research*, *118*, 354-366.

Kumar, P., Kabra, S., & Cole, J. M. (2022). Auto-generating databases of Yield Strength and Grain Size using ChemDataExtractor. *Scientific Data*, *9*(1), 292. https://doi.org/10.1038/s41597-022-01301-w

Marcon, A., & Ribeiro, J. L. D. (2021). How do startups manage external resources in innovation ecosystems? A resource perspective of startups' lifecycle. *Technological Forecasting and Social Change*, *171*, 120965. https://doi.org/10.1016/j.techfore.2021.120965

Martelli, A., Ravenscroft, A. M., Holden, S., & McGuire, P. (2023). *Python in a Nutshell*. O'Reilly Media, Inc.

Muppidi, A., Hashim, A. S. B., & Hasan, M. H. B. (2022). *Proposed User-Experience Model for the Design and Development of BI Dashboards*. 23-28. Scopus. https://doi.org/10.1109/ICICyTA57421.2022.10037904

Murray State University, Krotov, V., Johnson, L., Murray State University, Silva, L., & University of Houston. (2020). Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems*, *47*, 539-563. https://doi.org/10.17705/1CAIS.04724

Nishikawa-Pacher, A., Heck, T., & Schoch, K. (2022). Open Editors: A dataset of scholarly journals' editorial board positions. *Research Evaluation*.

Raghavendra, S. (2021). *Python testing with Selenium: Learn to implement different testing techniques using the Selenium Webdriver / Sujay Raghavendra.* (1st ed. 2021.). Apress. https://doi.org/10.1007/978-1-4842-6249-8

Rahman, R. U., & Tomar, D. S. (2021). Threats of price scraping on e-commerce websites: Attack model and its detection using neural network. *Journal of Computer Virology and Hacking Techniques*, *17*(1), 75-89.

https://doi.org/10.1007/s11416-020-00368-6

Rahmatulloh, A., & Gunawan, R. (2020). Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar. *Indonesian Journal of Information Systems*, *2*(2), 95-104. https://doi.org/10.24002/ijis.v2i2.3029

Sanabria De Luque, J. J. (2021). *SECTOR PRIVADO Y LIBRE COMPETENCIA: IMPLICACIONES JURÍDICAS DEL WEB SCRAPING*.

Schedlbauer, J., Raptis, G., & Ludwig, B. (2021). Medical informatics labor market analysis using web crawling, web scraping, and text mining. *International Journal of Medical Informatics*, *150*, 104453. https://doi.org/10.1016/j.ijmedinf.2021.104453

Schmieder-Ramirez, J., & Mallette, L. A. (2021). *Communication in Organizations*. Oxford University Press.

Sorapure, M. (2023). User Perceptions of Actionability in Data Dashboards. *Journal of Business and Technical Communication*, 10506519231161612. https://doi.org/10.1177/10506519231161611

St-Hilaire, C., Brunila, M., & Wachsmuth, D. (2023). High Rises and Housing Stress. *Journal of the American Planning Association*, 1-15. https://doi.org/10.1080/01944363.2022.2126382

vanden Broucke, S., & Baesens, B. (2018). *Practical Web Scraping for Data Science*. Apress. https://doi.org/10.1007/978-1-4842-3582-9