# Desarrollo de una herramienta de adquisición automática de datos de fuentes externas, para el sistema de gestión de información de la Vicerrectoría de Investigaciones de la Universidad CESMAG, mediante crawlers y web scraping.

Daniel Clemente Gualteros Sinsajoa, ✉ dgualteros637@gmail.com
Andres Sebastian Trejo Quintero, ✉ andrestrejo7777@gmail.com
David Esteban Finlay Estrella, ✉ definlay.8875unicesmag.edu.co

Universidad CESMAG
Facultad de Ingeniería
Ingeniería de Sistemas
Pasto - Colombia
2024

# Desarrollo de una herramienta de adquisición automática de datos de fuentes externas, para el sistema de gestión de información de la Vicerrectoría de Investigaciones de la Universidad CESMAG, mediante crawlers y web scraping.

Daniel Clemente Gualteros Sinsajoa, ✉ dgualteros637@gmail.com
Andres Sebastian Trejo Quintero, ✉ andrestrejo7777@gmail.com
David Esteban Finlay Estrella, ✉ definlay.8875unicesmag.edu.co

Trabajo de grado como requisito para optar el título de Ingeniero de Sistemas

Asesor: Héctor Andrés Mora Paz

Universidad CESMAG
Facultad de Ingeniería
Ingeniería de Sistemas
Pasto - Colombia
2024

NOTA DE ACEPTACIÓN

_____

_____

_____

_____

_____

_____

Firma del jurado

San Juan de Pasto, 3 de mayo de 2024

NOTA DE EXCLUSIÓN

"El pensamiento que se expresa en esta obra es exclusiva responsabilidad de sus autores y no compromete la ideología de la Universidad CESMAG".

# DEDICATORIA

A mi amada familia, el motor de mi vida, cuyo apoyo incondicional y sacrificio constante han sido mi mayor fortaleza y guía en este viaje académico. Sus palabras de ánimo, su constante respaldo y su presencia han sido la luz que ha iluminado mi camino en los momentos más difíciles y oscuros. Cada esfuerzo que han hecho en silencio, cada gesto de amor desinteresado, ha sido un recordatorio constante de su inmenso cariño y dedicación hacia mí. No encuentro palabras suficientes para expresar mi profunda gratitud por todo lo que han hecho por mí. A mis amigos, cómplices de risas, consuelo en las adversidades y compañeros de travesía, gracias por su amistad, por estar en cada paso del camino y por compartir conmigo los altibajos de esta nueva experiencia.

A todos aquellos que han formado parte de mi vida y han contribuido a mi crecimiento personal y académico, les dedico este logro con profundo agradecimiento y afecto.

Con cariño,

Andrés Sebastián Trejo Quintero.

DEDICATORIA

A mi familia, mi más grande amor y motivación para seguir adelante en mi vida. por todo el apoyo, por todas las palabras de aliento en momentos difíciles, por todo su esfuerzo invertido a lo largo de toda mi carrera y, sobre todo, por creer en mí. A todas las personas que de alguna forma aportaron a mi crecimiento personal y profesional tanto compañeros como amigos. A mí, por no rendirme y continuar luchando contra cualquier adversidad y Finalmente, pero no menos importante, a mi amada compañera gatuna, una luz de esperanza en momentos de total oscuridad.

**Daniel Clemente Gualteros Sinsajoa**

DEDICATORIA

Este proyecto se lo dedico a Dios, quien me dio la fuerza y la sabiduría necesaria para superar cada obstáculo en este camino.

A mis queridos padres, este logro académico es un reflejo del incansable esfuerzo que han invertido para brindarme una educación sólida. Cada sacrificio que han hecho, cada día de trabajo duro y cada decisión que tomaron en mi nombre son el fundamento de mi éxito. Su dedicación y compromiso con mi educación son un regalo que valoro más allá de las palabras. Este proyecto de grado es un testimonio de su sacrificio y amor, y me llena de orgullo honrarlos de esta manera. Gracias por ser los faros en mi vida, por iluminar el camino hacia el conocimiento y por inculcarme la importancia del trabajo duro y la educación. Los amo profundamente.

**David Esteban Finlay Estrella.**

# AGRADECIMIENTOS

Agradezco sinceramente a Dios por otorgarme la fuerza y la motivación para alcanzar este hito en mi vida académica. Expreso mi profunda gratitud a mi familia por su apoyo moral y económico constante a lo largo de mi carrera universitaria, así como por sus sabios consejos que han sido fundamentales para mi desarrollo. Agradezco el compromiso y el apoyo de mi grupo de trabajo, cuya colaboración ha sido indispensable en este camino. Asimismo, extiendo mi agradecimiento a todas las personas que han contribuido de alguna manera a mi formación, brindándome enseñanzas valiosas que han enriquecido mi experiencia. Este logro no habría sido posible sin el respaldo y la colaboración de todos ustedes. Finalmente, expreso gran gratitud a mi segundo hogar, la universidad CESMAG y a todos los docentes que aportaron sus conocimientos y experiencias que me llevaré para siempre.

CONTENIDO

La investigación está conformada de la siguiente manera:

Capítulo 1: Introducción
Se presenta la introducción a la investigación, abordando la línea y sub línea de investigación, el problema planteado, los objetivos, la justificación y la delimitación del estudio.

Capítulo 2: Marco Teórico
Se desarrolla el marco teórico, revisando antecedentes, exponiendo supuestos teóricos, definiendo variables y formulando hipótesis.

Capítulo 3: Metodología
Se describe la metodología utilizada, incluyendo paradigma, enfoque, método, tipo y diseño de investigación, población y muestra, técnicas e instrumentos de recolección de datos.

Capítulo 4: Resultados de la Investigación
Se presentan los resultados obtenidos durante la investigación.

Capítulo 5: Análisis de Resultados
Se analizan e interpretan los resultados obtenidos, respondiendo a los objetivos planteados.

Capítulo 6: Conclusiones y Recomendaciones
Se presentan las conclusiones derivadas del análisis de resultados, destacando los hallazgos más relevantes y su significado. Además, se ofrecen recomendaciones para futuras investigaciones, identificando posibles áreas de desarrollo y sugerencias para mejorar el estudio.

RESUMEN ANALITICO DE ESTUDIO R.A.E

| | |
|---|---|
| FACULTAD: | INGENIERIA |
| PROGRAMA: | INGENIERÍA DE SISTEMAS |
| FECHA DE ELABORACIÓN: | MAYO DE 2024 |
| AUTORES DE INVESTIGACIÓN: | DANIEL CLEMENTE GUALTEROS SINSAJOA<br>ANDRES SEBASTIAN TREJO QUINTERO<br>DAVID ESTEBAN FINLAY ESTRELLA |
| ASESOR DE LAINVESTIGACIÓN: | MSC. HÉCTOR ANDRÉS MORA PAZ |
| TÍTULO DE LA INVESTIGACIÓN: | DESARROLLO DE UNA HERRAMIENTA DE ADQUISICIÓN AUTOMÁTICA DE DATOS DE FUENTES EXTERNAS, PARA EL SISTEMA DE GESTIÓN DE INFORMACIÓN DE LA VICERRECTORÍA DE INVESTIGACIONES DE LA UNIVERSIDAD CESMAG, MEDIANTE CRAWLERS Y WEB SCRAPING. |

PALABRAS O FRASES CLAVES: WEB CRAWLER, SCRAPING, REPOSITORIOS EDUCATIVOS, EXTRACCIÓN DE INFORMACIÓN, BASES DE DATOS, DESARROLLO WEB, FRAMEWORKS DE DESARROLLO, INTELIGENCIA ARTIFICIAL, LIBRERÍAS DE PROGRAMACIÓN, METODOLOGÍA CRISP-DM.

**TABLA DE CONTENIDO**

INTRODUCCIÓN ..... 16

1. PROBLEMA DE INVESTIGACIÓN ..... 17

1.1 Objeto o Tema de Investigación. ..... 17

1.2 Línea de Investigación. ..... 17

1.3 Sublínea de Investigación. ..... 17

1.4 Planteamiento del problema. ..... 17

1.5 Formulación del problema. ..... 18

1.6 Objetivos ..... 19

1.6.1 Objetivo general ..... 19

1.6.2 Objetivo específicos ..... 19

1.7 Justificación ..... 19

1.8 Viabilidad ..... 20

1.8.1 Operativa ..... 20

1.8.2 Técnica ..... 20

1.8.3 Económica ..... 20

1.9 Delimitación ..... 21

2. MARCO TEÓRICO ..... 21

2.1 Antecedentes ..... 21

2.2 Supuestos teóricos de la investigación ..... 27

2.2.1 La estructura del sitio web ..... 28

2.2.2 Protocolo de la web ..... 29

2.2.3 crawlers y web scraping ..... 30

2.2.4 Tipos de web scraping ..... 33

2.2.5 Aplicativos para el uso de crawlers y web scraping ..... 34

2.2.6 Lenguajes de programación ..... 38
2.2.7 Librerías para aplicar crawlers y web scraping ..... 38
2.2.8 Framework utilizado para el desarrollo del proyecto ..... 39
2.2.9 Bases de datos ..... 41
2.3 Variables de la investigación. ..... 41
2.4 Definición nominal de las variables. ..... 42
2.5 Definición operativa de las variables ..... 43
2.6 Formulación de hipótesis ..... 45
2.6.1 Hipótesis de la investigación ..... 45
2.6.2 Hipótesis nula ..... 45
2.6.3 Hipótesis alterna ..... 45
3. METODOLOGÍA ..... 45
3.1 Paradígma ..... 45
3.2 Enfóque ..... 45
3.3 Método ..... 46
3.4 Tipo de Investigación ..... 46
3.5 Diseño de investigación ..... 46
3.6 Población y muestra ..... 46
3.7 Técnicas de recolección de información ..... 47
3.8 Validez de las técnicas de recolección de información ..... 48
3.9 Confiabilidad de las técnicas de recolección ..... 48
3.10 Instrumentos de recolección de información ..... 49
4. RESULTADOS DE LA INVESTIGACIÓN ..... 50
4.1 contextualización ..... 50
4.2 Procesamiento y Recolección de la Información ..... 50

4.2.1 Desarrollo de rutinas de Crawlers y Web Scraping. ..... 50
4.2.1.2 Definición de patrones de búsqueda. ..... 51
4.2.1.3Evaluación de los repositorios académicos. ..... 52
4.2.1.4 Resultados. ..... 52
4.2.2 Construcción de la base de datos. ..... 53
4.2.2.1 Diseño de la base de datos. ..... 54
4.2.2.2 Resultados. ..... 56
4.2.3 Validación del Módulo de Recolección de Datos ..... 56
4.2.3.1 Preparación del Cuestionario de Evaluación ..... 57
4.2.3.2 Ejecución del Proceso de Validación ..... 59
4.2.3.3 Resultados ..... 61
4.2.3.4 Documento de Aceptación por los Líderes de Investigación ..... 65
4.3 aspectos propios de la metodología. ..... 65
4.3.1 Metodología SCRUM. ..... 65
4.3.2 Análisis del sistema. ..... 66
4.3.2.1 Requerimientos funcionales. ..... 66
4.3.2.2 Requerimientos no funcionales. ..... 67
4.3.2.3 Sprint. ..... 71
5. ANÁLISIS DE RESULTADOS ..... 76
CONCLUSIONES ..... 78
RECOMENDACIONES ..... 80
REFERENCIAS BIBLIOGRÁFICAS ..... 82
ANEXOS ..... 89

## LISTA DE FIGURAS

Fig 1. Esquema del funcionamiento de una araña web o crawler[28]. .....30
Fig 2. Esquema del funcionamiento del web scraping[29]. .....31
Fig 3. INTERFAZ DE OCTOPARSE [42]. .....35
Fig 4. Modelo MVC DJANGO.[57] .....40
Fig 5.Interfaz de búsqueda académica. .....72
Fig 6. Interfaz de muestra de resultados. .....72
Fig 7. Interfaz de muestra de resultados (barra de paginación). .....73
Fig 8. Resultados de más repositorios (Recolecta). .....73
Fig 9. Base de datos creada. .....74
Fig 10. Prueba de registro de datos exitoso. .....75
Fig 11. Verificación de registros ingresados. .....75

## LISTA DE TABLAS

Tabla 1. Clasificación de las aplicaciones beneficios/planes de pago. ..... 37
Tabla 2. Resultados De la encuesta aplicada. ..... 61
Tabla 3. Roles En metodología SCRUM. ..... 65
Tabla 4. Responsabilidades en el desarrollo de SCRUM. ..... 65
Tabla 5. Requerimientos funcionales. ..... 66
Tabla 6. Requerimientos no funcionales. ..... 67
Tabla 7 Historias de usuario. ..... 68
Tabla 8. Sprint 1. ..... 71
Tabla 9. Sprint 2. ..... 74
Tabla 10. Sprint 3. ..... 75

# INTRODUCCIÓN

La Vicerrectoría de Investigaciones de la Universidad CESMAG requiere de un sistema de gestión de información que le permita recopilar y organizar de manera eficiente los datos necesarios para el desarrollo de sus investigaciones. Para ello, se propone el desarrollo de una herramienta de adquisición automática de datos de fuentes externas utilizando técnicas de crawlers y web scraping. El objetivo principal de este trabajo de grado es diseñar, implementar y evaluar un módulo de adquisición automática de datos que permita la recolección y organización de la información necesaria para la gestión de información en la Vicerrectoría de Investigaciones. Esta herramienta se basó en la utilización de técnicas de crawlers y web scraping para la adquisición automática de datos de fuentes externas relevantes.

Las variables a trabajar en el proyecto, las cuales fueron autor, descripción, fuente externa de la investigación, fecha de publicación, enlace del documento, número de citas del documento, tipo de documento consultado, cantidad de versiones del documento, palabras clave y número de descargas del documento. Todas estas variables son importantes para la gestión de información en la Vicerrectoría de Investigaciones y permitirán tener una visión más completa y detallada de las investigaciones en curso.

Es importante destacar que, para garantizar la validez y confiabilidad de los datos recolectados, se utilizaron técnicas y herramientas de recolección de información confiables y validadas en el campo de los crawlers y web scraping. Además, se realizarán pruebas y ajustes necesarios para asegurar la eficacia y precisión de la herramienta de adquisición automática de datos.

El desarrollo de esta herramienta de adquisición automática de datos utilizando técnicas de crawlers y web scraping, es un proyecto adscrito al proyecto de investigación profesoral "**Desarrollo de un módulo KDD, para la exploración y análisis robusto de los datos generados por el sistema de gestión de información de la Vicerrectoría de Investigaciones de la Universidad CESMAG, mediante algoritmos de data mining y machine learning**" mediante la modalidad de instancia en línea. El cual permitirá a la Vicerrectoría de Investigaciones de la Universidad CESMAG contar con una herramienta eficiente y precisa para la recolección y organización de la información necesaria para sus investigaciones. Además, la utilización de técnicas y herramientas validadas garantizará la confiabilidad y validez de los datos recolectados.

# 1. PROBLEMA DE INVESTIGACIÓN

## *1.1 Objeto o Tema de Investigación.*

Desarrollo de rutinas de crawlers y Web Scraping utilizando el lenguaje de programación Python y sus librerías, con el objetivo de recolectar información de sitios web de repositorios académicos externos a los de la universidad CESMAG y almacenarla en una base de datos para su posterior análisis y procesamiento.

## *1.2 Línea de Investigación.*

Gestión desarrollo e innovación en las TIC.

## *1.3 Sublínea de Investigación.*

Inteligencia Artificial.

## *1.4 Planteamiento del problema.*

La Universidad CESMAG realiza investigaciones en diferentes áreas, desde ciencias sociales, pasando por investigaciones tecnológicas, hasta ciencias naturales, y recopila grandes cantidades de datos a través de los registros obtenidos en sus propias investigaciones. Si bien, la recolección de información interna es útil como un repositorio fuente para las investigaciones de la universidad, es muy importante reconocer el contexto externo de las investigaciones desarrolladas fuera de la universidad CESMAG en diferentes repositorios como MinCiencias[1] el cual incluye plataformas como GrupLAC[2], entre otros, y como la recolección de esta información es de gran utilidad para los investigadores[3]. Estos últimos enfrentan el desafío de extraer información valiosa de grandes cantidades de datos no estructurados que se encuentran en diversas fuentes en línea.

En la actualidad, el proceso de extracción manual de datos es costoso y lento, por lo que la automatización de este proceso a través del uso de herramientas como crawlers y web scraping se ha vuelto cada vez más importante[4].

Los crawlers son herramientas de software que permiten extraer datos de una página web, mientras que el web scraping, son programas que automatizan la navegación web para recopilar la información que se necesita. La combinación de ambas herramientas permite la extracción eficiente de grandes volúmenes de datos de varias fuentes web[5].

Sin embargo, existen problemas al obtener información de Internet. Algunas fuentes pueden no proporcionar datos estructurados, lo que dificulta su procesamiento. Además, la información obtenida puede ser incompleta, inexacta o irrelevante. La calidad de los datos es fundamental para la eficiencia del proceso de recuperación de datos, por lo que se deben aplicar técnicas de limpieza de datos y garantizar su calidad[6].

Con la ayuda de crawlers y web scraping, la universidad CESMAG puede recuperar información relevante de muchas fuentes en línea externas a la información ya obtenida en los propios procesos investigativos, como revistas científicas y demás información de repositorios de interés. Además, el uso de estas herramientas permite recolectar toda esta valiosa información y almacenarla en una base de datos para su posterior estudio y revisión. A estos datos se le pueden aplicar técnicas de minería de datos para descubrir patrones y relaciones, lo que permite a los investigadores tomar decisiones informadas y precisas, así como también encontrar patrones y tendencias de investigación para estar a la vanguardia a nivel tecnológico e investigativo.

Sin embargo, existen desafíos en la recopilación de datos en línea en el campo de la investigación universitaria. Por ejemplo, es posible que cierta información no esté disponible públicamente o no tenga derechos de autor. Además, la calidad de la información puede variar según la fuente y el formato. Si no se implementa una herramienta que utilice crawlers y web scraping, la universidad podría estar en desventaja frente a otras instituciones académicas que sí cuenten con este tipo de tecnología, lo que podría afectar su reputación y competitividad en el campo de la investigación.

Por lo tanto, el desarrollo de una herramienta de adquisición automática de datos de fuentes externas utilizando crawlers y web scraping se vuelve necesario para simplificar y automatizar el proceso de extracción de datos de varias fuentes web que pueden usarse en varios campos. Esta herramienta debe ser capaz de recolectar datos para obtener información valiosa que pueda ser utilizada para la toma de decisiones. Es necesario asegurar la calidad y exactitud de la información obtenida, para que el proceso de búsqueda de información sea efectivo y permita a los investigadores obtener resultados acertados.

### *1.5 Formulación del problema.*

¿Cómo se puede desarrollar de manera eficiente una herramienta utilizando crawlers y web scraping para recolectar y procesar grandes cantidades de datos de diferentes fuentes académicas

externas y en diferentes formatos, y luego almacenarlos como información útil, contextualizándolos al contexto local de la Universidad CESMAG?

## *1.6 Objetivos*

### *1.6.1 Objetivo general*

Desarrollar una herramienta eficiente de recolección de datos mediante crawlers y web scraping con el fin de obtener información confiable y precisa de repositorios académicos webs externos de manera efectiva.

### *1.6.2 Objetivo específicos*

- Generar rutinas de crawlers y web scraping para la recolección de información en la web.
- Crear una base de datos con la información recolectada.
- Validar la herramienta de recolección de datos en función de los criterios de precisión y calidad establecidos.

## *1.7 Justificación*

El proyecto beneficia a la comunidad universitaria CESMAG, en especial a los programas académicos que requieren información de investigaciones en el contexto nacional e internacional para renovar su registro calificado. Además, será útil para estudiantes, docentes e investigadores que necesitan acceder a información actualizada y relevante sobre ciencia, tecnología e investigación.

El proyecto tiene un gran interés para la universidad CESMAG, ya que la recopilación y conexión de información externa es vital para el desarrollo de investigaciones y proyectos académicos. Asimismo, la posibilidad de contar con datos actualizados y relevantes a nivel nacional e internacional permite a la universidad mantenerse actualizada y competitiva en el ámbito académico e investigativo.

Como novedad del proyecto, cabe mencionar que la creación de una herramienta de conexión con la información externa a la universidad CESMAG es una innovación que permite acceder a datos y conocimientos del contexto externo en tiempo real. Este enfoque de trabajo permitirá a la universidad CESMAG mejorar la toma de decisiones en la planificación de proyectos académicos

y de investigación, así como mantenerse actualizada en cuanto a tendencias y avances en ciencia y tecnología. La adquisición automática de datos del contexto externo, permite la recopilación automatizada de información en tiempo real, lo que reduce el tiempo y esfuerzo necesarios para obtener y procesar datos relevantes para la universidad.

## *1.8 Viabilidad*

### *1.8.1 Operativa*

El software que se va a utilizar debe funcionar correctamente para recopilar los datos de manera efectiva. En este caso, el proyecto de crawlers y web scraping se define al asegurarse de que el software o herramienta utilizado para el web scraping sea compatible con los sitios web a los que se están recopilando y buscando datos mediante los crawlers. Los crawlers y web scraping deberán funcionar correctamente para recopilar los datos necesarios de forma confiable.

### *1.8.2 Técnica*

El software es operativo para la universidad CESMAG y de gran ayuda, ya que se dispone del uso de software libre que es una excelente opción, porque proporciona herramientas gratuitas y de código abierto que se pueden utilizar para este proyecto. Librerías de uso libre, como Scrapy o Beautiful Soup, son altamente configurables y personalizables, lo que permite adaptarlos a las necesidades específicas para la herramienta de crawlers y web scraping. En definitiva, se dispone de los recursos y oportunidades necesarios para la ejecución eficaz y eficiente del proyecto, es decir que la viabilidad técnica es muy positiva.

### *1.8.3 Económica*

El proyecto cuenta con una viabilidad económica favorable, ya que los costos estimados son relativamente bajos en comparación con los beneficios potenciales que se pueden obtener como recurso económico por parte de la universidad y como mejora en la calidad de la investigación y la eficiencia en la gestión de la información. Además, el uso de equipos propios reduce los costos operativos del proyecto, lo que aumenta su rentabilidad a largo plazo.

### *1.9 Delimitación*

El proyecto consiste en el desarrollo de una herramienta para el sistema de gestión de información de la vicerrectoría de investigación de la Universidad CESMAG, una herramienta de adquisición automática de datos de fuentes externas a través de crawlers y web scraping. El objetivo es tomar información de investigaciones externas y contextualizarla con información dentro del sistema, permitiendo su posterior análisis y búsqueda de patrones. El proyecto tuvo viabilidad técnica, operativa y financiera, y se espera que sus resultados sean de utilidad a la universidad en la toma de decisiones y desarrollo de investigaciones, el proyecto se llevó a cabo entre el primer periodo del 2023 hasta finales del primer periodo del 2024. Esta información proporciona una referencia temporal para la ejecución y finalización del desarrollo de la herramienta de adquisición automática de datos.

## 2. MARCO TEÓRICO

### *2.1 Antecedentes*

En los últimos años, el uso de crawlers y web scraping ha aumentado significativamente debido al creciente interés en el análisis y la minería de datos. Esto se debe en gran parte a la creciente cantidad de información disponible en línea y al aumento de la capacidad computacional y la tecnología de almacenamiento.

En 2016 se presenta un artículo de investigación en la Universidad CESMAG en la ciudad de Pasto en Nariño cuyo objetivo principal fue caracterizar el espacio web colombiano, desde las páginas hasta los sitios web y los dominios, para comprender su estructura y peculiaridades en comparación con las páginas web de otros países. Este proyecto contribuye al conocimiento y comprensión de la Red en Colombia y permite obtener una comprensión más clara de su estructura y componentes, así como sus similitudes y diferencias con otras redes alrededor del mundo[7]. Además, un factor diferenciador de este proyecto es que se utilizaron técnicas de crawlers y web scraping para recopilar información de los sitios web en Colombia, lo que asegura la exactitud y confiabilidad de la información obtenida. Además, se llevó a cabo un análisis exhaustivo de varios parámetros característicos de la red, lo que le dió más detalle y profundidad a la investigación. Este proyecto tiene un gran valor científico y académico, porque contribuye al conocimiento y comprensión de la red en Colombia, permite identificar oportunidades para mejorar su estructura y uso, y crea una base sólida para futuras investigaciones sobre este tema.

En 2021, en la Universidad Javeriana de Bogotá, se desarrolló un web scraping para optimizar los procesos de mercadeo en la línea "Dolor" de Abbott Colombia. Se buscaba una solución a las debilidades del actual panorama de control de productos de los portales digitales.[8]

El objetivo principal de la herramienta fue recopilar información sobre productos para identificar información importante que ayudaría en la toma de decisiones de cartera digital. Para cumplirlo, hicieron uso de la tecnología de crawlers y web scraping, delimitando las variables a utilizar (precio, molécula, etc), realizando un mapeo de las URL a las que la herramienta tendría acceso y de esta manera extraer la información específica necesaria para cada gerente o línea del producto.

Una vez que se han extraído los datos, se pueden guardar en archivos de formato CSV, lo cual es cercano a la estructura de datos que tendrá el proyecto. Después de abrir el archivo, se realizan los cambios necesarios para mejorar la presentación de los datos. Por otro lado, las imágenes recopiladas a través de la captura web se almacenan en la carpeta creada previamente por el desarrollador sin cambios posteriores[8]. Finalmente, estos datos son mostrados al usuario mediante un DASHBOARD, el cual le muestra la información de los productos como disponibilidad, cantidad en stock, etc.

El uso de las técnicas usadas para la recolección de datos haciendo uso de crawlers y web scraping como el mapeo de las URL en Python, establecen la importancia de la delimitación de los sitios a los cuales se desea acceder y de la información que será extraída de estos. Esto último demuestra que, al hacer uso de esta tecnología, un buen mapeo de sitios web es muy importante para la obtención de información de calidad.

En el 2020 se realizó un trabajo de investigación en la Universidad Pedagógica y Tecnológica de Colombia de Sogamoso, que tuvo como objetivo principal crear ScraCOVID-19 una plataforma web de contenido digital dedicada a acceder a las noticias actualizadas y de manera rápida. Como caso de estudio se manejaron cuatro medios digitales con licencia a nivel nacional. Las noticias se presentan de manera resumida para permitir a los lectores, en función de su interés, leer las noticias mediante algunos filtros como: desempleo, educación, maltrato, corrupción y discriminación; ScraCOVID-19 se creó utilizando la tecnología de extracción Scraping de BeautifulSoup, una biblioteca que permite extraer información en formato HTML de varios sitios web utilizando el lenguaje de programación Python. Describe un modelo para realizar clasificación que extrae información útil para clasificar información haciendo referencia a URLs utilizando técnicas de extracción combinadas con herramientas de almacenamiento de datos no estructurados, la

información se obtiene de diferentes páginas web, y toda esta información se gestiona a partir de datos recopilados en la misma, generada dinámicamente en el sitio web[9].

El aporte de esta investigación es el manejo de crawlers y web scraping a través de la librería BeautifulSoup, ya que aporta conocimientos necesarios para entender mejor la librería y de esta manera, realizar una buena recolección de la información para su posterior almacenamiento en una base de datos. La diferencia de este trabajo con el proyecto está en el acoplamiento de información de 3 sitios web de los cuales solo se tiene en cuenta un pequeño grupo de información la cual es relevante para poder brindar uso datos nuevos y actualizados en un contexto más amplio. En el desarrollo de la herramienta para la adquisición de datos mediante crawlers y web scraping se hizo aplicación de estas técnicas a muchos otros sitios web o repositorios.

En junio de 2019 se presentó un proyecto de grado en la Universidad del Sinú Elías Bechará Zainúm seccional Cartagena, el cual se propuso resolver el siguiente problema: ¿Cómo desarrollar un prototipo para extraer, consultar y centralizar la información en la web de portales de empleos y búsquedas de vacantes en Bolívar? De esta manera, su enfoque consistió en la investigación y análisis de patrones que identifican y centralizan la búsqueda de información en las bolsas de empleo, permitiendo la extracción de los datos, su clasificación y almacenamiento, siendo un medio para unificar las páginas web que ofertan vacantes de trabajo y de esta manera lograr que las personas se les facilite encontrar todos los cargos disponibles en un solo lugar de acuerdo a sus especificaciones[10].

El aporte de este trabajo investigativo es la recolección de información y su acoplamiento para generar una web con la información relevante de los sitios que proporcionan vacantes de trabajo como LinkedIn y Computrabajo, hallando nuevos patrones que brindan los lineamientos para establecer una publicación acertada correspondiente a las búsquedas indicadas de las necesidades de los usuarios. Este proyecto se diferencia en que está desarrollado para la captura de información correspondiente a dos sitios web acordes a la publicación de vacantes de empleo.

En mayo de 2020 se desarrolló un prototipo de recopilador web en la Corporación Universitaria Minuto de Dios Vicerrectoría Regional Orinoquía Sede Villavicencio (Meta), el cual tenía como finalidad desarrollar un prototipo de recopilador web enfocado en las redes sociales que permita traer datos personales de un usuario específico, para la generación de un microinforme de actividad reciente que pueda poner en evidencia información que represente una vulnerabilidad informática si estos son almacenados sin consentimiento y son visibles al público en internet. De esta manera,

se establecieron unos objetivos concernientes para cumplir lo anteriormente mencionado, los cuales fueron: 1) Enfocar la búsqueda de información en las redes sociales más usadas en Colombia, como Facebook, Google y YouTube. 2) Recopilar información personal de un usuario específico. 3) Proporcionar tres métodos externos de recomendación para la toma de acciones en caso de que la búsqueda sea positiva para el usuario y así proteger gran parte de la información expuesta. Este trabajo contribuyó a la reflexión de los usuarios sobre la cantidad de información personal que proporcionan en línea y cómo los terceros pueden usar esa información. De igual manera, ayuda a mejorar las políticas de privacidad y seguridad en línea de diferentes plataformas con respecto a la información personal de los usuarios[11].

En el año 2022 se desarrolló en la Universidad Autónoma de Bucaramanga – UNAB, en la facultad de ingeniería de sistemas, un estudio sobre la gran demanda que existe hoy en día respecto a la tecnología, productos de canasta familiar, productos vehiculares, etc. En donde el teléfono celular, al ser una de las herramientas fundamentales de comunicación, entretenimiento, conexión a internet y productividad de las personas, puede brindar un acceso inmediato a las búsquedas de diferentes productos. De esta forma, se generó el objetivo correspondiente al desarrollo de un prototipo de plataforma tecnológica de búsqueda en internet mediante el uso de técnicas de crawlers y web scraping en distintas plataformas de comercio electrónico, que ayude al usuario a encontrar las mejores opciones de precio. Para cumplirlo, se siguieron los siguientes pasos:

1. Identificar las características principales y variantes de crawlers y web scraping en la web haciendo búsquedas en bases de datos bibliográficas.
2. Desarrollar una API de búsqueda en Django que implemente técnicas de crawlers y web scraping para ayudar al usuario a encontrar las mejores opciones de precio.
3. Implementar el prototipo de aplicación móvil que consuma la API desarrollada en Django para que sea una herramienta de búsqueda interactiva y que obtenga los resultados realizados a páginas web.
4. Realizar pruebas de usabilidad, verificando que el prototipo de aplicación móvil cumple con su propósito de sugerir productos basándose en su precio.

El resultado de esta investigación contribuyo a tener en cuenta el desarrollo de mejores y más eficientes tecnologías de búsqueda que puedan mejorar la interacción de las personas con la información en Internet, lo que puede tener un impacto positivo en diferentes áreas, como en este caso, el área de investigación[12].

En 2023 en la Universidad Privada Antenor Orrego en facultad de Ingeniería en Trujillo–Perú se usó técnicas de web scraping para el análisis de datos de jugadores profesionales del fútbol peruano para el periodo 2021. Las técnicas para web scraping en los datos de los jugadores profesionales tiene como objetivo analizar la información disponible sobre los jugadores de la Liga 1 en el fútbol peruano. Para saber la condición física del jugador y su valoración comercial, la solución al problema planteado consistió en utilizar la técnica de web scraping para recopilar datos de páginas web futbolísticas y crear un dashboard en Power BI con información sobre los encuentros deportivos, jugadores destacados, faltas deportivas, edad de los jugadores y nivel económico de los equipos en la Liga 1 del fútbol peruano estos fueron algunos rangos, se utilizó la extensión de Scraper de Google Chrome debido a su facilidad de uso, rapidez y gratuidad. Se procedió a guardar la información en formato CSV o extensión .xlsx y se importó al Power BI para su posterior procesamiento y creación de los dashboards[13].

El proyecto es un buen punto de referencia para la investigación en el desarrollo de la herramienta de adquisición automática de datos ya que se quiere recopilar los datos de manera automática para la adquisición de datos externos mediante crawlers y web scraping. En particular al desarrollar el proyecto, el uso de herramientas de web scraping como la extensión Scraper de Google Chrome y Power BI para recopilar y analizar datos de fútbol peruano son una propuesta para el desarrollo del proyecto. Ya que estas herramientas de web scraping y análisis de datos como Power BI se pueden utilizar en conjunto con crawlers para adquirir datos de diferentes fuentes externas.

En 2020 se presentó en la Universidad de Harvard en Cambridge Massachusetts, una tesis para el grado de Licenciatura en Artes y Ciencias con Honores en el tema de Ciencias de la Computación, donde se utilizó el web scraping para obtener datos de varias páginas web y crear un conjunto de datos etiquetados que se utilizaron para entrenar un clasificador en el análisis de datos de transacciones de criptomonedas. Fue útil cuando se utilizaron técnicas de aprendizaje semi-supervisado para extraer señales basadas en sectores de datos de transacciones pseudo-anónimas.[14].

Construyendo un pequeño conjunto de datos etiquetados: Como se mencionó en la sección anterior, la mayoría de las criptomonedas, incluyendo Ethereum, son seudónimas. Por lo tanto, se recopilaron datos de Etherscan y Twitter para identificar cuentas conocidas, que luego se clasificaron en uno de los siguientes seis sectores: intercambio descentralizado (DEX), intercambio, juego, oferta inicial de monedas (ICO), individuo y minero. Etherscan es uno de los

exploradores de bloques más conocidos de Ethereum. También contiene una variedad de métricas e información útil. En particular, contiene una lista de cuentas conocidas y sus propietarios. Sin embargo, esta lista carece de cuentas individuales que conforman una parte importante del ecosistema de Ethereum. Por lo tanto, se utilizó la información extraída de Twitter para complementar el conjunto de datos. La tesis de Tancredi Castellano Pucci di Barsento aporta en buenas prácticas para llevar a cabo técnicas de crawlers y web scraping de forma ética y responsable, respetando los términos y condiciones de uso de las páginas web y protegiendo la privacidad de los usuarios como lo son algunas páginas que no permiten estas técnicas. En general, la tesis podría ser una buena fuente de información para ampliar el conocimiento sobre estos temas y para aprender de las experiencias y metodologías utilizadas para este proyecto de investigación.

En 2022 en la Universidad César Vallejo en Lima – Perú se establece el desarrollo de un Framework apoyado en web scraping y geolocalización para la identificación y selección de productos en supermercados. La solución que se propone en el proyecto es el desarrollo de un framework que, a través de técnicas de crawlers y web scraping y geolocalización, permita a los usuarios identificar y seleccionar los productos que necesitan de manera eficiente y efectiva. Al utilizar web scraping, el framework puede recopilar información actualizada sobre la disponibilidad de productos y sus precios en tiempo real en las distintas tiendas. La geolocalización permite a los usuarios encontrar los productos que necesitan en la tienda más cercana, evitando la necesidad de visitar múltiples tiendas en busca de lo que necesitan.[15].

La similitud de este proyecto con el desarrollado, es que se utiliza técnicas de crawlers y web scraping y recolección de datos para lograr sus objetivos. Esto puedo traer beneficios a este proyecto obteniendo información útil sobre las técnicas y herramientas utilizadas para la recolección y gestión de datos, especialmente si se trata de adquirir información sobre proveedores o productos de investigación que podrían estar disponibles en plataformas y bibliotecas.

En general, el proyecto de identificación y selección de productos en supermercados puede ofrecer conocimientos y habilidades valiosas en términos de web scraping, geolocalización y manejo de datos que podrían aplicarse en otros proyectos relacionados con la recolección y análisis de información de fuentes externas.

En 2021 en Madrid se desarrolló un proyecto para la titulación de máster en big data: tecnología y analítica avanzada que describe la aplicación de técnicas de web scrpaing, se utilizó para obtener de manera automática la información de una página web de empleo. El objetivo era extraer las

valoraciones o comentarios que los usuarios habían dejado sobre sus experiencias laborales en diferentes empresas. Esto proporciona una herramienta para comprender la satisfacción de los empleados y la reputación de las empresas en función de las opiniones expresadas en las valoraciones. Al realizar el web scraping, se pudo obtener un conjunto de datos con las valoraciones de los usuarios, que posteriormente se utilizaron para desarrollar los modelos de clasificación basados en el análisis de sentimiento. El análisis de sentimiento utiliza técnicas de procesamiento del lenguaje natural para clasificar las valoraciones como positivas o negativas. La técnica de comparación de modelos consiste en evaluar y contrastar los resultados obtenidos por los modelos desarrollados en este proyecto con modelos pre-entrenados ya existentes. Esto permite obtener una perspectiva más completa y amplia sobre las capacidades de análisis de sentimiento utilizadas. El proyecto exploró las aplicaciones de los modelos de clasificación en contextos empresariales. Esto significa que los resultados y conclusiones obtenidos podrían ser utilizados por empresas y organizaciones para evaluar la satisfacción de sus empleados, identificar áreas de mejora y tomar decisiones informadas relacionadas con la gestión del talento.

Para el beneficio de este proyecto, se ofrecen soluciones tecnológicas que automatizan la adquisición y análisis de datos de fuentes externas, mejorando la eficiencia, la calidad y la disponibilidad de la información en sistemas existentes. Esto proporciona una base sólida para la toma de decisiones informadas y la mejora de procesos en diversos ámbitos, como la gestión de empleo y la investigación universitaria[16].

### *2.2 Supuestos teóricos de la investigación*

Históricamente, los crawlers y web scraping se remontan a la década de 1990, cuando el uso de la web se volvió común. En ese momento, la mayoría de las páginas web eran estáticas y no contenían contenido dinámico, por lo que era relativamente fácil crear crawlers simples para recopilar información de los sitios web.

Hoy en día, el uso de crawlers y web scraping se ha vuelto aún más sofisticado y se utiliza en una amplia variedad de aplicaciones, desde datos de precios y análisis de mercado hasta investigación académica y ciberseguridad. A medida que la web sigue evolucionando, es probable que el uso de crawlers simples siga siendo una herramienta valiosa para recopilar y analizar datos en línea. La extracción de datos a través de técnicas de crawlers y web scraping implica la extracción de datos de sitios web de forma automatizada. El web scraping concierne al proceso de extraer información

estructurada de las páginas web, mientras que los crawlers son programas que se utilizan para navegar por la web y recopilar datos automáticamente[17].

La extracción de datos utiliza técnicas que reconocen la estructura de las páginas web y extraen información relevante. Estos datos se pueden almacenar en diferentes formatos como CSV, JSON o bases de datos. Los crawlers y el web scraping se utilizan en muchas aplicaciones, como la recopilación de datos para análisis de mercado, la extracción de datos para investigación académica, el seguimiento de precios y la recopilación de datos para informes[18].

Para comprender mejor el funcionamiento crawlers y web scraping, es importante tener en cuenta conceptos tales como:

### *2.2.1 La estructura del sitio web*

La arquitectura del sitio web puede ser un factor determinante en la efectividad y precisión del rastreo de sitios web y la tecnología de rastreo. Una buena arquitectura del sitio puede facilitar la extracción de datos, mientras que una mala arquitectura puede dificultar o incluso imposibilitar la extracción. Por ejemplo, una arquitectura de sitio web bien estructurada y organizada podría incluir etiquetas HTML claras y coherentes, nombres de atributos y clases coherentes y una jerarquía lógica de elementos. Esto puede ayudar a los crawlers y web scraping a identificar y usar los datos de manera más efectiva[19]. Este es un ejemplo sencillo de la estructura HTML:

```
1 <h1> Título del proyecto </h1>
2 <p> Párrafo </p>
3 <a href="http://tusitio.com">Visita nuestro sitio</a>
4 <img src="imagen.jpg">
```

El código HTML de una página web consta de etiquetas que definen la estructura y el contenido de la página[19]. El encabezado principal usa la etiqueta h1, mientras que los encabezados secundarios usan h2, h3, etc. La etiqueta p se usa para definir párrafos de texto y la etiqueta se usa para crear enlaces a otras páginas web. La etiqueta img también se usa para agregar imágenes a una página web. Es importante comprender cómo se estructura una página web mediante etiquetas para poder extraer la información deseada mediante técnicas de crawlers y web scraping.

Por otro lado, una arquitectura de sitio web desigual y confusa, combinada con etiquetas HTML inconsistentes o una falta de estructura lógica, puede dificultar la extracción de datos y generar resultados incompletos o inexactos. En algunos casos, los sitios web pueden incluso implementar medidas de seguridad como CAPTCHA o bloques de IP para evitar la extracción automática de datos.

En términos generales, los sitios web se dividen en dos categorías: estáticos y dinámicos. Los sitios web estáticos son sitios web creados con un lenguaje de marcado estático (como HTML) y no utilizan tecnologías avanzadas (como JavaScript o bases de datos)[20]. Estos sitios son más fáciles de indexar porque la información está en archivos HTML estáticos y es fácilmente accesible.

Por otro lado, los sitios web dinámicos utilizan tecnologías avanzadas como Javascript[20], Ajax y bases de datos, lo que dificulta su indexación. Los crawlers y web scraping pueden tener dificultades para acceder a los datos de sitios web dinámicos debido a su diseño y al uso de técnicas avanzadas. Además, los sitios web dinámicos suelen tener sistemas de seguridad más sofisticados para evitar el acceso no autorizado[21].

Por lo tanto, la arquitectura de un sitio web puede afectar en gran medida la capacidad de los crawlers y web scraping para recopilar información de manera efectiva. Un sitio web bien diseñado y estructurado puede facilitar el seguimiento de la información, mientras que un sitio web mal diseñado o mal estructurado puede hacer que esta tarea sea mucho más difícil.

### *2.2.2 Protocolo de la web*

El conocimiento de los diversos formatos y protocolos utilizados para transmitir datos en la web es fundamental para que estas técnicas de web scraping y crawlers puedan aplicarse y sean efectivas. Elegir el formato y el protocolo correctos es importante para que la extracción de datos sea precisa y eficiente, y los datos se puedan procesar de manera eficiente. A continuación, se muestran los protocolos y estándares que son necesarios para el funcionamiento y uso del servicio en web.

• **XML:** Este es un lenguaje de etiqueta que se usa para crear documentos con estructura jerárquica, es decir, están organizados en forma de árboles. Se utiliza para almacenar y transmitir datos independientes del software o hardware utilizado, y generalmente se usa para el intercambio de datos en la red[22].

• **JSON:** Este es un formato de intercambio de datos para enviar y recibir información entre diferentes sistemas o aplicaciones. Es fácil de leer y escribir, y puede explicarse por muchos lenguajes de programación[23].

• **SOAB:** Este es un acuerdo utilizado para intercambiar información en un entorno distribuido. Como un conjunto de reglas, define cómo transmitir dos aplicaciones, y generalmente se usa para el desarrollo de servicios web[24].

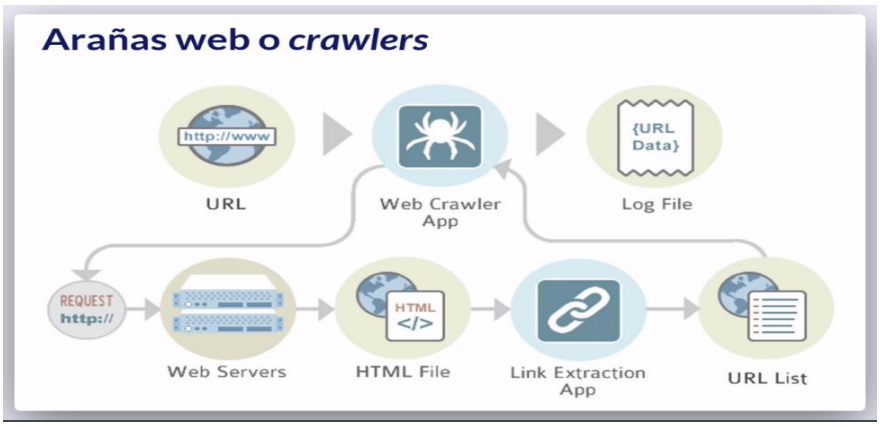
• **REST:** es la arquitectura del sistema web, y el enfoque es crear servicios web escalables y flexibles. Se basa en el concepto de recursos y se puede reconocer a través de la URL. Es muy utilizado en el desarrollo de aplicaciones móviles y web modernas[25].

• **WSDL:** Este es un lenguaje utilizado para describir los servicios web y cómo acceder a ellos. Proporcione información sobre las interfaces de servicios web, los métodos disponibles y cómo llamarlos. Es muy útil en el desarrollo de la integración del sistema y los servicios web[26].

### *2.2.3 crawlers y web scraping*

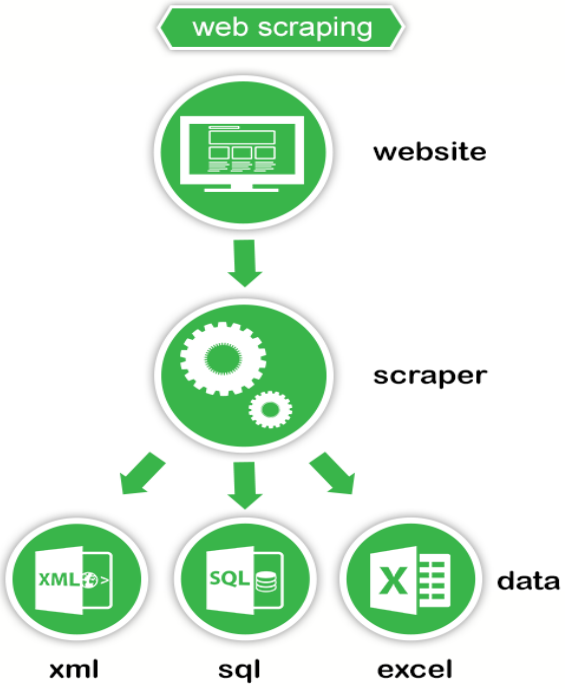
Un Crawler, también conocido como araña o robot, es un programa automatizado que se mueve sistemáticamente a través de la web, visita páginas web y recopila datos para su posterior procesamiento y análisis. Los motores de búsqueda como Google hacen un uso extensivo de los rastreadores para indexar el contenido web y facilitar su búsqueda. Por otro lado, el web scraping es una técnica de extracción de datos que implica el uso de un software para recopilar automáticamente datos de una página web. Esta tecnología permite obtener información específica y relevante del sitio web[27].

Fig 1. Esquema del funcionamiento de una araña web o crawler[28].



El web scraping se realiza examinando y analizando el código fuente de una página web, utilizando herramientas y técnicas que permiten extraer automáticamente la información deseada. Es importante mencionar que, si bien el web scraping es una técnica legal, su uso indebido puede violar las políticas de privacidad de los sitios web y generar problemas legales.

Fig 2. Esquema del funcionamiento del web scraping[29].




Dentro de los usos actuales para los crawlers y el web scraping se encuentran:

**Análisis de precios:** Es posible conocer los costos de bienes particulares en diferentes áreas del país utilizando como herramienta web scraping en sitios web de minoristas en línea y otros mercados, de esta manera las empresas podrán establecer precios competitivos para sus productos y tomar en cuenta algunos otros que se encuentren a un mejor precio, ayudando así a maximizar su rentabilidad y de igual forma conocer los intereses y comportamientos del consumidor[30].

**Monitorización de noticias locales**: La información constantemente varía y es por esto que existen muchos medios que difunden y brindan el servicio de publicación de información concerniente a los acontecimientos diarios más importantes e impactantes que se presentan en el país, es por esto que surge la necesidad de conocer aquellos sitios más confiables de plataformas como sitios web, redes sociales y periódicos del área para realizar una obtención de esta mediante el web scraping y

generar la unificación de la información. Esta información sería de utilidad para periodistas, especialistas en marketing y el público en general[31].

**Investigación de la competencia:** Al mirar los sitios web de los competidores, los dueños de negocios locales pueden obtener información útil sobre sus estrategias de precios, estrategias de mercadeo y otros datos comerciales pertinentes que los ayudarán a mejorar sus productos y servicios y desarrollar nuevas estrategias de mercadeo[32].

**Análisis de opiniones de los clientes:** Cuando se trata de mejorar la calidad de sus servicios y construir una mejor reputación en línea, las empresas de servicios como restaurantes y hoteles pueden encontrar que recopilar opiniones y comentarios de los clientes en línea puede ayudarlos a comprender mejor las necesidades y preferencias de sus clientes en un contexto geográfico particular[33].

**Recopilación de datos gubernamentales:** Se puede encontrar información sobre iniciativas de construcción, esquemas de subvenciones y otras facetas cruciales de la política local haciendo uso de la herramienta de web scraping en sitios web del gobierno y todo esto para que las personas y organizaciones de la zona que deseen participar en la toma de decisiones políticas y la planificación del desarrollo local encuentren un medio seguro y confiable que les permita informarse y realizar su respectiva participación[34].

**Estudio de mercado:** Se puede obtener una investigación de mercado a profundidad de los sitios web de empresas y asociaciones teniendo en cuenta aspectos importantes como: los precios de la mercancía, conocer a su público objetivo, identificar sus necesidades, saber cuáles son las debilidades y fortalezas, medir campañas y estrategias de marketing, para una vez identificados realizar un web scraping para la obtención de la información la cual ayudará a los propietarios de negocios locales a evaluar la demanda de bienes o servicios o encontrar nuevas oportunidades comerciales.[35].

**Análisis de datos meteorológicos:** Las personas que cuentan con comercios, negocios o responsabilidades que dependen del clima pueden beneficiarse a través de la recolección de información en sitios web relacionados con el clima, ya que de esta manera podrán tomar decisiones respecto a eventos que se generen al aire libre analizando las estadísticas o resultados de la información recolectada[36].

**Análisis de redes sociales:** Los sitios web de redes sociales como Facebook, Twitter e Instagram pueden brindar datos importantes sobre la opinión pública, identificación de problemas y

oportunidades, identificación de influenciadores y análisis de sentimientos, lo cual puede ayudar para que las personas con negocios locales comprendan mejor su mercado objetivo y puedan crear campañas de marketing más exitosas para tener mayor acogida y mejores ganancias[37].

**Recopilación de información turística:** La extracción de información en sitios web de viajes puede ayudar a visualizar cuáles son los lugares más visitados y de esta manera establecer estrategias para hacer que los lugares turísticos tengan mayor reconocimiento y por ende mayor cantidad de visitantes incrementando de esta forma las ventas, la economía local, diversificación de clientes, entre otras, lo cual resulta muy beneficioso para hoteles, restaurantes y otras empresas de los sitios turísticos[38].

### *2.2.4 Tipos de web scraping*

Existen diferentes tipos de web scraping, entre ellos se denotan dos muy importantes y además muy usados en estas técnicas, estos son:

**El web scraping a una sola página web:** Es la extracción de datos de una única página web. Este enfoque es adecuado cuando se necesitan datos específicos de una sola página web, como información de contacto, reseñas de productos, precios, etc. Para llevar a cabo el web scraping de una sola página, se utiliza una biblioteca de Scraping como BeautifulSoup o Scrapy. Estas bibliotecas permiten al usuario navegar por la estructura de la página web y extraer los datos necesarios[39].

**El web scraping de varias páginas web:** Se utiliza cuando se necesita extraer datos de múltiples páginas web. En este caso, el proceso es más complejo ya que se necesita navegar por varias páginas web y extraer datos de cada una de ellas. Se utilizan técnicas de Crawler para explorar y extraer datos de varias páginas web. Los crawlers pueden ser configurados para seguir ciertas reglas, como el tiempo de espera entre cada solicitud o el número máximo de páginas a explorar [40].

**Crawler vertical:** Esta es una técnica utilizada para extraer información de una sola página web, pero de forma exhaustiva. Significa que se extrae toda la información posible de la página, incluidos sus enlaces internos. De esta manera, se puede recopilar información detallada sobre un tema en particular o un sitio en particular.

**Crawler horizontal:** También conocida como Oruga de Unidad de Ancho. Esta es una tecnología utilizada para extraer información de múltiples páginas web en un sitio web. El Crawler comienza en la página de inicio y luego sigue los enlaces para recuperar información de todas las páginas relacionadas en el sitio. Este enfoque es útil si desea obtener una visión general de todo el sitio.

### *2.2.5 Aplicativos para el uso de crawlers y web scraping*

Hay varias herramientas web que brindan servicios de crawlers y web scraping, cada una con sus propias características y beneficios. Algunas herramientas son de pago y ofrecen una amplia gama de funciones, mientras que otras son gratuitas y tienen funciones limitadas. Es importante evaluar cuidadosamente las opciones disponibles y elegir la herramienta adecuada en función de las necesidades específicas de uso. Estas herramientas suelen ofrecer diferentes formatos de salida, como CSV, Excel o XML, e incluso la posibilidad de introducir datos en bases de datos. Algunas de estas herramientas se muestran a continuación:

**Scrapy**

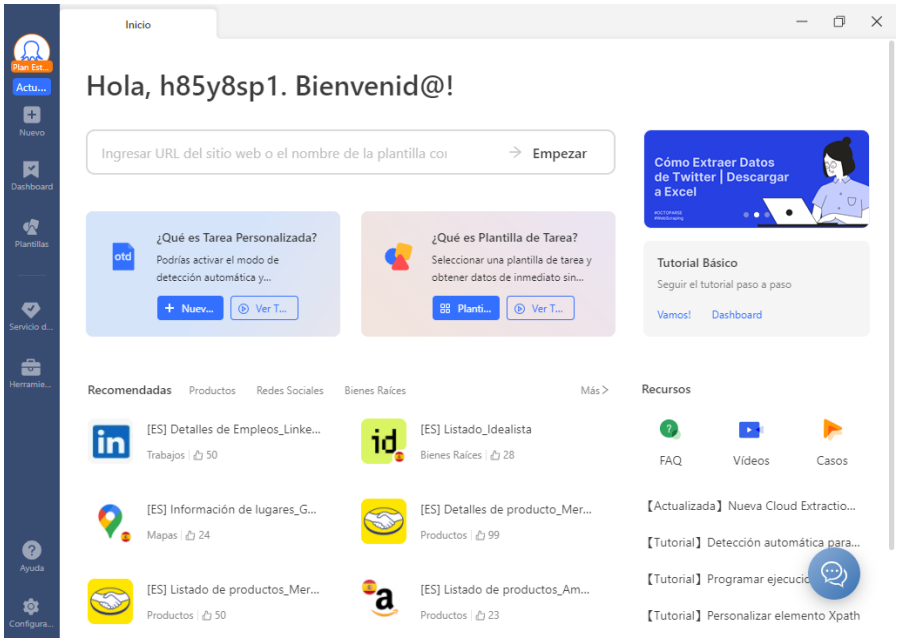
Scrapy es un framework de web scraping escrito en Python y ampliamente utilizado en el ámbito de la extracción de datos en la web. Proporciona una arquitectura escalable y flexible para construir aplicaciones de web scraping eficientes y rápidas. Scrapy se basa en la estructura de código abierto Twisted, que permite una implementación asíncrona de solicitudes de red y un alto rendimiento en la extracción de datos[41].

Scrapy tiene una arquitectura modular que consta de varias partes, incluyendo el motor central, los spiders, los middlewares, los ítems y los pipelines. El motor central es el componente principal de Scrapy que coordina todos los procesos de extracción de datos. Los spiders son los encargados de realizar la extracción de los datos de las páginas web y definir cómo se debe procesar la información extraída. El middleware es el responsable de realizar operaciones adicionales en las solicitudes y respuestas de la red. Los ítems son las estructuras de datos que almacenan la información extraída, mientras que los pipelines procesan y almacenan los datos extraídos en diferentes formatos, como archivos CSV o bases de datos. Además de su arquitectura modular, Scrapy también tiene otras características útiles, como la gestión de cookies, la manipulación de formularios y la compatibilidad con proxis. Scrapy también permite la integración con otras herramientas y bibliotecas de Python, lo que lo hace aún más poderoso[41].

**Octoparse:** Es una herramienta de web scraping visual fácil de usar que permite extraer datos de sitios web sin tener que escribir código. Con Octoparse, los usuarios pueden seleccionar y extraer información de sitios web mediante una interfaz gráfica de usuario (GUI) que funciona como un navegador. Octoparse admite una variedad de sitios web, incluidos sitios web dinámicos y sitios web que requieren autenticación.

Una de las características únicas de Octoparse es su capacidad para manejar el web scraping de datos en sitios web con tecnología JavaScript. Esto significa que puede extraer información de sitios web que cargan datos dinámicamente a través de Ajax o JavaScript. La herramienta también es compatible con la extracción de datos de múltiples páginas y puede exportar los datos extraídos a diferentes formatos, como Excel, CSV y bases de datos. Además de la versión de escritorio, Octoparse también ofrece una versión basada en la nube que permite a los usuarios ejecutar tareas de web scraping en la nube sin necesidad de descargar o instalar nada en su computadora.[42].

Fig 3. INTERFAZ DE OCTOPARSE [42].




**ParseHub:** ParseHub es una herramienta de web scraping que permite extraer datos de múltiples sitios web de manera automatizada y estructurada. Ofrece una interfaz gráfica de usuario para crear proyectos de extracción de datos utilizando técnicas de selección visual y arrastrar y soltar. La

plataforma también admite la programación y automatización de proyectos, así como la integración con otras herramientas y servicios a través de API. ParseHub es ampliamente utilizado por empresas, investigadores y otros profesionales para extraer datos de la web de manera eficiente y rápida[43].

**Hunter.io:** Es una herramienta en línea que se utiliza para encontrar y verificar direcciones de correo electrónico. Esta herramienta cuenta con una base de datos que incluye millones de correos electrónicos, la cual se actualiza constantemente mediante la extracción de información de diferentes fuentes. Hunter.io también ofrece la posibilidad de verificar la validez de los correos electrónicos, lo que es muy útil para el envío de correos electrónicos masivos. Además, ofrece otras funciones como la búsqueda de nombres de dominio y la extracción de datos de contacto de sitios web[44].

**Datafiniti:** La plataforma de Datafiniti permite a los usuarios buscar y descargar datos de varias fuentes en una variedad de formatos, lo que facilita la integración de los datos en otros sistemas y aplicaciones. Datafiniti también ofrece servicios de extracción de datos personalizados y a medida para satisfacer las necesidades específicas de los clientes[45].

**80legs.com:** Proporciona una gama de servicios que incluyen rastreo web, extracción de datos y análisis de datos, todos los cuales están disponibles a través de una interfaz o API basada en la web. La plataforma es altamente personalizable, lo que permite a los usuarios especificar qué sitios web rastrear, qué datos extraer y con qué frecuencia rastrear[46].

**Scraper API:** Ofrece una API simple para recuperar datos web de cualquier sitio web y entrega la respuesta en formato JSON. Scraper API admite varios lenguajes de programación como Python, Ruby, Java y PHP, entre otros, y se puede integrar con varios marcos de raspado web como BeautifulSoup, Scrapy y Selenium. El servicio ofrece un plan de prueba gratuito y varios planes pagos con diferentes opciones de precios según la cantidad de llamadas API requeridas[47].

**Mozenda.com:** Es una herramienta de raspado web que permite a los usuarios recopilar datos de varias fuentes web y transformarlos en formatos estructurados como CSV, XML o JSON. Ofrece una interfaz visual de apuntar y hacer clic para crear agentes de raspado web, así como una función de programación para automatizar las tareas de extracción de datos. Mozenda también proporciona una API para integrarse con otros sistemas y aplicaciones, así como un sistema de almacenamiento y gestión de datos para organizar y analizar los datos raspados[48].

**Diffbot.com:** Es una plataforma de extracción de datos y web scraping que utiliza IA para extraer y estructurar automáticamente datos de páginas web. Ofrece una gama de herramientas para la extracción de datos, incluidas API personalizadas, extensiones de navegador y un panel basado en la web. La plataforma está diseñada para trabajar con sitios web complejos y dinámicos, lo que permite a los usuarios extraer datos de fuentes a las que sería difícil acceder utilizando métodos tradicionales de crawlers y web scraping.

Diffbot.com ofrece una serie de características, incluida la extracción de datos de múltiples páginas web, la extracción de datos estructurados de fuentes no estructuradas, la normalización automática de datos, la deduplicación de datos y la integración con herramientas populares de análisis de datos. La plataforma está diseñada para ser altamente escalable, por lo que es adecuada para su uso por empresas de todos los tamaños[49].

**Dexi.io:** Es una plataforma de extracción y automatización de datos basada en la nube que permite a las empresas extraer datos de diversas fuentes, transformarlos en un formato estructurado e integrarlos en sus sistemas. La plataforma utiliza herramientas visuales de apuntar y hacer clic para crear crawlers, que pueden extraer datos de sitios web, archivos PDF y otras fuentes.

Dexi.io también ofrece procesamiento y transformación automatizados de datos, lo que permite a los usuarios limpiar y estructurar sus datos antes de exportarlos a otras plataformas, como Excel o Salesforce[50].

Tabla 1. Clasificación de las aplicaciones beneficios/planes de pago.

Fuente: Investigación propia.

| WEB | CRAWLING PERSONALIZADO | INTERFAZ GRÁFICA | IA | REPOSITORIOS PÚBLICOS | DOCUMENTACIÓN DE CALIDAD | PLANES DE PAGO | |
|---|---|---|---|---|---|---|---|
| | | | | | | Mas barato | Mas caro |
| **Scrapy** | SI | NO | NO | NO | SI | $ 0 USD | $ 0 USD |
| **Octoparse** | NO | SI | NO | NO | NO | $ 75 USD /MES | $ 4899 USD /MES |
| ParseHub | SI | SI | NO | NO | SI | $ 149 USD /MES | $ 499 USD /MES |
| **Hunter.io** | NO | NO | NO | NO | SI | $ 34 USD /MES | $ 399 USD /MES |
| Datafiniti | NO | SI | NO | NO | NO | $ 34 USD /MES | $ 399 USD /MES |
| 80legs.com | SI | NO | NO | NO | SI | $ 29 USD /MES | $ 299 USD /MES |
| Scraper API | SI | NO | NO | NO | SI | $ 29 USD /MES | $ 249 USD /MES |
| Mozenda.com | SI | SI | NO | NO | SI | $ 250 USD /MES | $ 450 USD /MES |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Diffbot.com | SI | NO | SI | NO | SI | $ 299 USD /MES | $ 399 USD /MES |
| Dexi.io | SI | SI | NO | NO | SI | $ 119 USD /MES | $ 699 USD /MES |

### *2.2.6 Lenguajes de programación*

Para el desarrollo de este proyecto, se utilizaron distintos lenguajes de programación. Teniendo en cuenta que los procesos de crawlers y web scraping cuentan con numerosas librerías para ser trabajados en Python, se optó por utilizar este lenguaje para implementar la herramienta de adquisición automática de datos de la web.

**PYTHON:** Es un lenguaje de programación interpretado y de alto nivel, diseñado para ser fácil de leer y escribir. Es muy popular en la comunidad de la ciencia de datos y en la inteligencia artificial debido a su sintaxis simple y la gran cantidad de bibliotecas y herramientas disponibles, lo que lo hace ideal para el análisis de datos y la automatización de tareas[52].

Algunas de las características de Python son:

- Es un lenguaje interpretado, lo que significa que el código se ejecuta línea por línea en lugar de tener que compilar el programa completo antes de ejecutarlo.
- Tiene una sintaxis clara y legible, lo que lo hace fácil de entender y escribir para los programadores.
- Es un lenguaje de alto nivel, lo que significa que se ocupa más de la abstracción del problema que de la gestión de la memoria y otros detalles de bajo nivel.
- Es multiplataforma, lo que significa que se puede ejecutar en diferentes sistemas operativos.
- Tiene una gran cantidad de bibliotecas y herramientas disponibles, lo que lo hace ideal para tareas de ciencia de datos y automatización[53].

### *2.2.7 Librerías para aplicar crawlers y web scraping*

Resultó crucial familiarizarse con las principales librerías disponibles para desarrollar nuestros propios procesos de crawlers y web scraping, las cuales fueron indispensables para el trabajo en la

herramienta. Estas representaban solo algunas de las librerías más empleadas en proyectos de web scraping y crawling en Python:

**Beautiful Soup:** Es una biblioteca de Python que se utiliza para analizar y extraer datos de HTML y XML. Es especialmente útil para extraer datos de sitios web dinámicos y para web scraping en general. Beautiful Soup proporciona formas de navegar por el árbol del documento HTML/XML, encontrar elementos específicos por su etiqueta, clase, identificador, texto, etc., y extraer el contenido de los mismos. Es una de las herramientas más populares y ampliamente utilizadas para web scraping en Python[29].

**Selenium:** Permite simular la interacción de un usuario con un navegador web y, de esta manera, realizar tareas como el llenado de formularios, la selección de elementos y la navegación por diferentes páginas. También se puede utilizar para la extracción de información de páginas web, ya que permite acceder al código fuente y buscar elementos mediante expresiones XPath o selectores CSS.

Selenium puede ser utilizado en varios lenguajes de programación, incluyendo Python, Java, C#, Ruby, JavaScript, entre otros. Es una herramienta muy útil para el desarrollo de aplicaciones web y la automatización de pruebas de software[54].

**Requests:** Es una de las más populares en Python para realizar solicitudes HTTP y trabajar con respuestas de servidores web. Con ella se pueden realizar tareas como la obtención de contenidos de páginas web, el envío de formularios, la gestión de sesiones de usuario, entre otros. Requests permite enviar y recibir datos en diferentes formatos, como JSON, XML o HTML[56].

**BeautifulSoup4:** Es una biblioteca de Python que se usa comúnmente para fines de web scraping. Está diseñado para facilitar la extracción de información de archivos HTML o XML. Con esta biblioteca, es posible navegar, buscar y modificar el árbol de análisis de documentos HTML y XML. BeautifulSoup4 proporciona muchas funciones integradas para analizar archivos HTML y XML. También tiene soporte para varios analizadores populares, incluidos lxml y html5lib[57].

### *2.2.8 Framework utilizado para el desarrollo del proyecto*

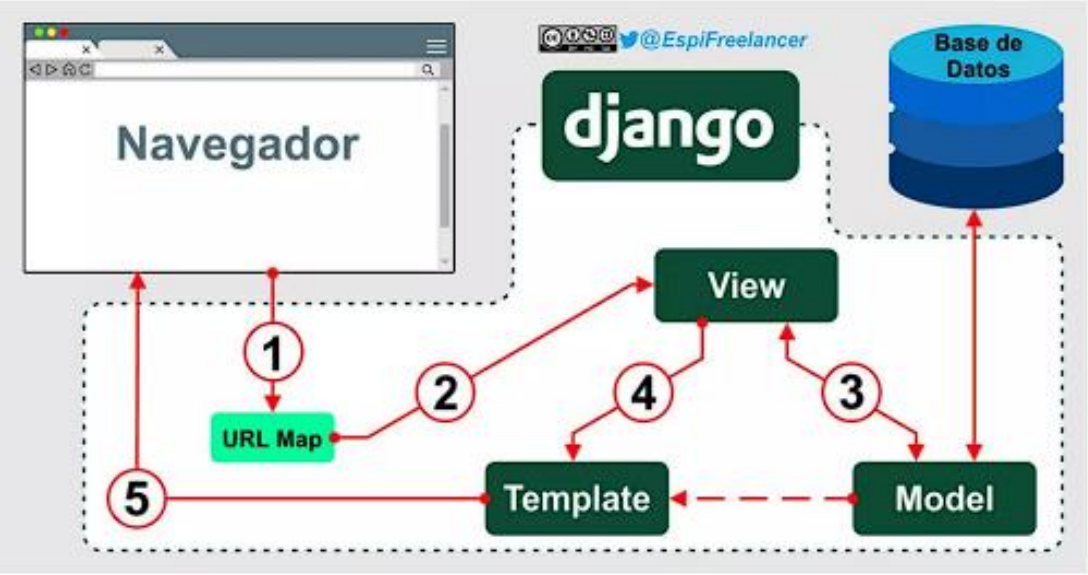
**Django:** Es un framework web avanzado de código abierto escrito en Python. Fue creado para facilitar el desarrollo de aplicaciones web complejas y escalables. Django sigue el patrón de diseño

Model-View-Controller (MVC), donde la vista es el componente que gestiona la lógica empresarial de la aplicación, el modelo es el componente que interactúa con la base de datos y el controlador es el componente responsable del procesamiento. preferencias de usuario, vista y modelo[58].

Las principales ventajas de Django son:

- **Velocidad de desarrollo:** Django tiene una gran cantidad de herramientas y bibliotecas listas para usar que permiten a los desarrolladores crear aplicaciones de manera rápida y eficiente.
- **Seguridad:** Django se encarga de administrar la seguridad de la aplicación, lo que permite a los desarrolladores concentrarse en la funcionalidad de la aplicación. Escalabilidad: Django está diseñado para escalar para manejar grandes cantidades de datos y tráfico de red.
- **Flexibilidad:** Django es altamente configurable y se puede personalizar
- **Documentación:** Django tiene una documentación muy completa y fácil de seguir que facilita el aprendizaje del marco y el desarrollo de aplicaciones.

Fig 4. Modelo MVC DJANGO.[57]



#### *2.2.9 Bases de datos*

Hay diferentes tipos de bases de datos que se pueden usar en proyectos de Crawler y web scraping, que incluyen:

**Bases de datos relacionales**: Este tipo de bases de datos utilizan tablas para almacenar información y realizar operaciones en la base de datos utilizando el lenguaje SQL. Ejemplos de bases de datos relacionales son MySQL, PostgreSQL y Oracle[59].

**Bases de datos NoSQL:** A diferencia de las bases de datos relacionales, las bases de datos NoSQL no utilizan tablas y no requieren esquemas predefinidos para almacenar información. Estas bases de datos son más flexibles y escalables que las bases de datos relacionales. Ejemplos de bases de datos NoSQL son MongoDB, Cassandra y Redis[60].

**Base de datos orientada a gráfos:** Este tipo de base de datos está diseñada para almacenar información en forma de gráfico, donde los nodos representan entidades y los bordes representan relaciones entre ellos[61].

La elección de la base de datos dependerá del tipo de información recopilada y de los objetivos del proyecto. Por lo general, para proyectos pequeños o medianos, una base de datos relacional como MySQL o PostgreSQL puede ser suficiente. Sin embargo, para proyectos más grandes o que requieran mayor escalabilidad y flexibilidad, las bases de datos orientadas a gráficos o NoSQL pueden ser más adecuadas[62].

### *2.3 Variables de la investigación.*

Las variables establecidas en este proyecto de crawlers y web scraping fueron encontradas durante la etapa de preprocesamiento, ya que la información en la página de destino podía cambiar con el tiempo. Por ejemplo, las etiquetas HTML, la estructura de la página web, los nombres de los elementos (como el título de la investigación), y otras propiedades podían cambiar, lo que afectaría la forma en que se extraían los datos y las variables utilizadas en el proceso. Además, era posible que durante el desarrollo del proyecto surgieran nuevas características o requisitos que requerían la inclusión de nuevas variables o la eliminación de algunas existentes. Por lo tanto, el proceso de preprocesamiento fue continuo y las variables podían cambiar a medida que avanzaba el desarrollo del proyecto. Esto garantizaba que el proceso de aplicación de los crawlers y el web scraping fuera preciso y completo, y que la información necesaria se obtuviera de manera eficiente. Por lo tanto, inicialmente se establecieron las siguientes variables:

### *2.4 Definición nominal de las variables.*

**Título de la investigación:** El nombre o título descriptivo de la investigación o documento consultado. Es una variable independiente.

**Autor:** El nombre o nombres de los autores o responsables del documento consultado. Es una variable independiente.

**Descripción:** Una breve explicación o resumen del contenido del documento o investigación. Es una variable independiente.

**Fuente externa de la investigación:** El nombre o identificación de la fuente externa de donde se obtuvo el documento o información. Es una variable independiente.

**Fecha de publicación:** La fecha en la que el documento fue publicado por primera vez. Es una variable independiente.

**Link o enlace del documento:** La dirección web o URL que lleva al documento o información completa. Es una variable independiente.

**Número de citas del documento:** El número de veces que el documento ha sido citado por otros autores o investigadores. Es una variable independiente.

**Tipo de documento consultado:** La categoría o tipo de documento consultado, como artículo de revista, tesis, libro, informe, etc. Es una variable independiente.

**Cantidad de versiones del documento:** El número de versiones o revisiones que ha tenido el documento a lo largo del tiempo. Es una variable independiente.

**Palabras clave:** Las palabras o términos que describen el contenido o tema principal del documento. Es una variable independiente.

**Número de descargas del documento:** El número de veces que el documento ha sido descargado o accedido por usuarios. Es una variable independiente.

**Campo de investigación:** El área temática o disciplina a la que pertenece la investigación o documento consultado. Es una variable independiente.

**Calidad de los datos adquiridos:** La medida o evaluación de la precisión, confiabilidad y completitud de los datos obtenidos mediante los métodos de crawlers y web scraping. Es una variable dependiente.

### *2.5 Definición operativa de las variables.*

**Título de la investigación:** Capturar el título del proyecto de investigación para su posterior análisis y registro en la base de datos del módulo de adquisición automática de datos. Es una variable de naturaleza cuantitativa.

**Indicador:** Palabras clave relevantes incluidas en el título son un indicador de la calidad y relevancia del título de la investigación.

**Autor:** Identificar el autor principal de la investigación para poder analizar su trayectoria y enfoque investigativo. Es una variable de naturaleza cuantitativa nominal.

**Indicador:** Número de publicaciones realizadas por cada autor.

**Descripción:** Obtener una descripción detallada de la investigación para entender de qué trata y poder categorizarla adecuadamente. Es una variable de naturaleza cuantitativa nominal.

**Indicador:** Número de palabras en la descripción de la investigación. Nivel de coherencia y claridad de la descripción.

**Fuente externa de la investigación:** Conocer la fuente de donde se obtuvo la investigación para poder determinar su fiabilidad y credibilidad. Es una variable de naturaleza cuantitativa nominal.

**Indicador:** Número de publicaciones realizadas en una fuente externa específica. Frecuencia de actualización de la fuente externa.

**Fecha de publicación:** Obtener la fecha en la que se publicó la investigación para poder contextualizarla y analizar su relevancia temporal. Es una variable de naturaleza cuantitativa discreta.

**Indicador:** Distribución de las publicaciones en el tiempo. Frecuencia de actualización de las publicaciones.

**Link o enlace del documento:** Obtener el enlace o URL del documento para poder acceder a la investigación y tener acceso a su contenido completo. Es una variable de naturaleza cuantitativa nominal.

**Indicador:** Número de clicks en los enlaces de cada documento. Tiempo promedio de permanencia en cada documento.

**Número de citas del documento:** Conocer el número de veces que la investigación ha sido citada por otros autores para poder evaluar su impacto y relevancia en la comunidad científica. Es una variable de naturaleza cuantitativa discreta.

**Indicador:** Número de veces que un documento ha sido citado por otros autores. Frecuencia de citas a lo largo del tiempo.

**Tipo de documento consultado:** Identificar el tipo de documento que se está consultando (PDS, LIBRO, ETC). Es una variable de naturaleza cuantitativa nominal.

**Indicador:** Distribución de los tipos de documentos consultados. Frecuencia de acceso a cada tipo de documento.

**Cantidad de versiones del documento:** Identificar la cantidad de versiones o revisiones que ha tenido el documento a lo largo del tiempo para poder evaluar su evolución y actualización. Es una variable de naturaleza cuantitativa discreta.

**Indicador:** Número de versiones de cada documento. Frecuencia de actualización de las versiones.

**Palabras clave:** Identificar las palabras clave que se utilizaron en la investigación para poder clasificarla adecuadamente y hacer búsquedas más efectivas. Es una variable de naturaleza cuantitativa nominal.

**Indicador:** Frecuencia de uso de cada palabra clave.

**Número de descargas del documento:** Conocer el número de descargas que ha tenido la investigación para poder evaluar su impacto y relevancia en la comunidad científica. Es una variable de naturaleza cuantitativa discreta.

**Indicador:** Frecuencia de uso de cada palabra clave. Número de descargas realizadas de cada documento. Frecuencia de descargas a lo largo del tiempo.

**Campo de investigación:** Ayuda definir y delimitar claramente el área temática en la que se centrará la recolección de información. Es una variable de naturaleza cualitativa.

**Indicador:** Número de publicaciones científicas relevantes en el campo de interés en un período determinado.

**Calidad de los datos adquiridos:** Establecer criterios y estándares para garantizar la calidad de los datos adquiridos y asegurar que sean relevantes y útiles para los fines del proyecto. Es una variable de naturaleza continua.

**Indicador:** porcentaje de datos recopilados que cumplen con ciertos criterios de calidad, como la exactitud, la completitud y la consistencia.

### *2.6 Formulación de hipótesis*

#### *2.6.1 Hipótesis de la investigación*

El empleo de un módulo de recolección de datos eficiente a través de crawlers y web scraping permitirá obtener información confiable y precisa de repositorios web externos, en comparación con los métodos tradicionales de recolección de datos.

#### *2.6.2 Hipótesis nula*

No existe una relación entre el uso de las herramientas de recolección de información crawlers y web scraping y la obtención de información confiable, lo que resulta en la generación de información de baja calidad.

#### *2.6.3 Hipótesis alterna*

El uso de crawlers y web scraping ofrece beneficios significativos al permitir la obtención rápida de grandes volúmenes de información de sitios web en períodos de tiempo reducidos.

## 3. METODOLOGÍA

### *3.1 Paradígma*

La investigación tuvo como base epistemológica el positivismo como guía para su elaboración. La elección de este paradigma se debió a que se adaptaba a las características y necesidades de la investigación. El paradigma positivista asumió una postura ontológica objetiva, considerando que existía una realidad objetiva y observable que podía ser estudiada científicamente. En este caso, se reconoció la existencia de datos externos que podían ser recopilados y analizados de manera objetiva. El paradigma positivista buscaba la verificación de las hipótesis planteadas.

### *3.2 Enfóque*

El proyecto propuso utilizar una metodología cuantitativa, considerada la más adecuada dentro de lo planteado en el paradigma del positivismo. Se emplearon herramientas y técnicas para analizar los datos recopilados de manera objetiva y rigurosa. El enfoque cuantitativo valoró la objetividad

en la medición y el análisis de los datos, lo cual buscaba minimizar la influencia de sesgos personales y asegurar que los resultados pudieran ser replicados por otros investigadores.

### *3.3 Método*

Este proyecto utilizó el método científico como enfoque para la investigación. En particular, se seleccionó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) para guiar el desarrollo del proyecto. CRISP-DM es un proceso estándar ampliamente utilizado en la minería de datos que consta de diversas etapas, incluyendo la comprensión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación y la implementación. Esto garantizó un enfoque estructurado y riguroso para la investigación, permitiendo obtener resultados confiables y significativos.

### *3.4 Tipo de Investigación*

El tipo de investigación para este proyecto fue descriptiva analítica. Este tipo de investigación utilizado en la investigación permitió obtener una descripción detallada y precisa de los datos recopilados mediante crawlers y web scraping.

### *3.5 Diseño de investigación*

Teniendo en cuenta que se realizaron distintos tipos de pruebas y se obtuvieron resultados mediante los algoritmos, se propuso un diseño experimental. Esto permitió controlar las variables que influyeron en los resultados de las pruebas y experimentos, así como evaluar el efecto de dichas variables en la recolección de información de los repositorios externos.

### *3.6 Población y muestra*

Es importante aclarar que este proyecto no pertenecía al paradigma theory driven approach, si no al paradigma data-driven approach[63], que se fundamenta en la toma de decisiones en función de los datos disponibles. En este caso, el objetivo del proyecto era automatizar la recolección de datos de fuentes externas para la gestión de la información del rectorado de actividad científica de la Universidad CESMAG. El objetivo principal era trabajar con todos los datos disponibles de estas fuentes externas, sin importar si existían o no conjuntos o modelos definidos. Además, en algunos casos, la fuente de datos externa podría carecer de una población bien definida, por lo que el

objetivo del proyecto era recopilar todos los datos disponibles para su posterior análisis. En lugar de seleccionar una muestra representativa, se recopilaron todos los datos disponibles para obtener una imagen completa y precisa de la información. Por lo tanto, el enfoque basado en datos de este proyecto justificaba el uso de la recopilación de datos sin la necesidad de establecer una población o muestra definida.

### *3.7 Técnicas de recolección de información*

Una de las técnicas a utilizar para la recolección de información fue la técnica de revisión documental. Consistió en examinar y analizar de manera sistemática y exhaustiva diversos documentos relevantes relacionados con el tema de investigación, ya que estos documentos podían incluir artículos científicos, tesis, informes técnicos, libros, entre otros.

Se utilizó técnicas de scraping y crawling para la recolección de información. Es importante destacar que el proyecto también trabajó con conjuntos de datos (data sets). Esto implicó la utilización de datos ya existentes, los cuales podían ser recopilados previamente o provenir de fuentes internas o externas. Los conjuntos de datos podían contener información estructurada en forma de tablas, archivos CSV u otros formatos, y se utilizaron para complementar y enriquecer la información obtenida a través del scraping y crawling. Al trabajar con conjuntos de datos, se pudó aplicar diversas técnicas de análisis de datos, como el preprocesamiento, la limpieza, la transformación y la visualización de los datos. Estas técnicas permitieron obtener información más significativa y comprensible, lo que facilitó el análisis y la toma de decisiones en la Vicerrectoría de Investigaciones. Se utilizo un crawler para extraer información de diferentes repositorios académicos y científicos como Google Académico, Directory of Open Access Journals (DOAJ), utilizando diferentes herramientas y bibliotecas útiles como Scrapy, BeautifulSoup, Selenium, Puppeteer, entre otras, las cuales facilitaron la extracción, clasificación y estructuración de la información. Se aplicó la herramienta de crawling y scraping para recolectar información de diferentes repositorios y se filtraron los datos para obtener solo aquellos relacionados con el tema de interés o investigado en el momento. Finalmente, se analizó la calidad de la información recolectada.

El objetivo fue recolectar información sobre artículos de revistas científicas, tesis de posgrado, libros, informes técnicos, comunicaciones a congresos, entre otros, que ayuden a robustecer las investigaciones realizadas en la Vicerrectoría de Investigaciones. El enfoque basado en datos del

proyecto justificó la recolección de información sin la necesidad de establecer una población o muestra definida, por lo que se recopilaron todos los datos disponibles para obtener una imagen completa y precisa de la información.

De igual manera se realizaron encuestas y entrevistas a expertos y usuarios relevantes para complementar la información recopilada mediante crawlers y web scraping.

### *3.8 Validez de las técnicas de recolección de información*

La efectividad de los métodos de recolección de datos en un proyecto depende de la calidad y confiabilidad de los datos obtenidos. En el caso de los crawlers y el web scraping, su eficacia depende de la calidad de los datos disponibles en el sitio y de la capacidad del software de extracción de datos para extraerlos de forma precisa y completa. Es importante comprobar la fuente y la fecha de los datos obtenidos para garantizar su validez.

Por lo tanto, fue necesario realizar un monitoreo continuo y riguroso de la calidad de los datos recolectados, con el fin de garantizar su validez y utilidad para la toma de decisiones informadas. Todo esto está encaminado a la captura de información útil y de repositorios confiables para los investigadores, tomando en cuenta las observaciones de expertos que dieron su aprobación para continuar con el almacenamiento de estos datos recolectados.

### *3.9 Confiabilidad de las técnicas de recolección*

Para este proyecto, la confiabilidad del método dependió principalmente de la calidad de la programación y configuración de los crawlers y los scripts de web scraping. Al haber sido diseñados correctamente, los resultados obtenidos con estos métodos fueron fiables y precisos. Además, fue muy importante haber establecido repositorios veraces y útiles con datos avalados por entidades, por ejemplo, la web de datos abiertos, que estaba avalada por el gobierno colombiano o la plataforma GRUPLAC que estaba respaldada por el MINCIENCIAS, para realizar la extracción de datos, estableciendo lineamientos para obtener información de interés. Sin embargo, hubo varios factores que afectaron la confiabilidad de estos métodos. Por ejemplo, si las fuentes de datos cambiaron constantemente su estructura o formato, era posible que los crawlers y scripts de web scraping no pudieran adaptarse y proporcionaran datos inexactos o incompletos. Además, si las

herramientas utilizadas no se actualizaban periódicamente, podían presentar problemas técnicos o errores que afectaran la exactitud de los datos obtenidos.

### *3.10 Instrumentos de recolección de información*

Para la recolección de información en el proyecto de adquisición automática de datos mediante crawlers y web scraping, se utilizaron principalmente los siguientes instrumentos:

**crawlers y web scraping:** se utilizaron herramientas y programas especializados en la extracción de datos de fuentes externas. Estos programas permitieron acceder a la información de forma automatizada, lo que redujo el error humano y aumentó la eficiencia en la recolección de datos.

**Bases de datos:** se utilizaron sistemas de gestión de bases de datos para almacenar los datos obtenidos mediante los crawlers y web scraping. Estos sistemas permitieron la organización y almacenamiento eficiente de grandes volúmenes de datos.

# 4. RESULTADOS DE LA INVESTIGACIÓN

## *4.1 contextualización*

En respuesta a la creciente necesidad de la Universidad CESMAG de acceder y aprovechar la vasta información contenida en repositorios académicos externos, el proyecto "Desarrollo de una herramienta para la adquisición automática de datos de fuentes externas mediante Crawlers y Web Scraping" se presenta como una iniciativa estratégica. La universidad, reconocida por su activa participación en la investigación en diversas disciplinas, enfrentaba el desafío de limitado acceso a datos externos cruciales que podrían enriquecer significativamente sus procesos de investigación.

## *4.2 Procesamiento y Recolección de la Información*

### *4.2.1 Desarrollo de rutinas de Crawlers y Web Scraping.*

En la fase inicial del proyecto, se concentraron esfuerzos en el desarrollo de rutinas de Crawlers y Web Scraping como parte del primer objetivo específico: Generar rutinas de crawlers y web scraping para la recolección de información en la web. Esta etapa crítica tuvo como propósito superar las limitaciones de acceso a datos externos que enfrentaba la Universidad CESMAG en sus procesos de investigación.

La caracterización del proceso implicó un análisis detallado de las fuentes de información externa, identificando repositorios académicos clave, como IEEexplore, entre otros, que se convirtieron en los principales objetivos de las rutinas de Crawlers. Estos instrumentos de software fueron diseñados para emular la navegación humana en la web, permitiendo la extracción sistemática y eficiente de datos valiosos.

En el marco de la recolección de datos, se aplicaron técnicas de Web Scraping para extraer información específica de las páginas web identificadas como fuentes relevantes. Estas técnicas incluyeron la identificación y selección de elementos clave, como títulos de investigación, nombres de autores, descripciones y otros metadatos relevantes.

Es fundamental destacar que este proceso no se limitó a una única fuente, sino que se diseñaron y adaptaron diferentes rutinas de Crawlers y Web Scraping para diversas fuentes web, asegurando una cobertura exhaustiva de repositorios externos.

Durante la implementación de estas rutinas, se priorizó la eficiencia y la precisión. Las herramientas de Crawling y Web Scraping se calibraron para optimizar la velocidad de recolección sin comprometer la calidad de los datos. Se establecieron procedimientos para gestionar la estructura variable de las páginas web y manejar posibles obstáculos, como medidas de seguridad y variaciones en la presentación de datos.

#### *4.2.1.2 Definición de patrones de búsqueda.*

En el desarrollo de las rutinas de Crawlers y Web Scraping, se adoptó un enfoque estratégico que implicó la definición de patrones de búsqueda específicos. Estos patrones se diseñaron considerando las variables fundamentales para la investigación, identificadas mediante un análisis exhaustivo de la estructura de los repositorios académicos objetivo. Este análisis reveló que, en su mayoría, estos repositorios compartían características comunes que permitieron establecer un conjunto consistente de variables, a saber:

**Título de la investigación:** Nombre descriptivo del documento o investigación.
**Autor:** Nombre o nombres de los autores.
**Descripción:** Resumen o explicación del contenido.
**Fuente externa de la investigación:** Identificación de la fuente externa.
**Fecha de publicación:** Fecha de publicación del documento.
**Link o enlace del documento:** URL que dirige al documento completo.
**Número de citas del documento:** Veces que el documento ha sido citado.
**Tipo de documento consultado:** Categoría del documento, como artículo, tesis, libro, etc.
**Cantidad de versiones del documento:** Número de revisiones del documento.
**Palabras clave:** Términos que describen el contenido.
**Número de descargas del documento:** Veces que el documento ha sido descargado.
**Campo de investigación:** Área temática a la que pertenece la investigación.

**Calidad de los datos adquiridos:** Medida de precisión, confiabilidad y completitud de los datos.

#### *4.2.1.3Evaluación de los repositorios académicos.*

Cada uno de estos repositorios fue evaluado meticulosamente en términos de calidad de resultados y utilidad académica. Se verificó su trayectoria, se consideraron los reconocimientos que hayan recibido y se analizaron las publicaciones de autores de renombre presentes en dichos repositorios. Además, se evaluó el tipo de contenido y de publicaciones que ofrecen. Este proceso de revisión garantizó la inclusión de repositorios de alta calidad académica en la herramienta, brindando a los usuarios la certeza de acceder a información confiable y valiosa.

Cabe resaltar el desafío adicional que representó la necesidad de realizar un estudio minucioso de los términos y condiciones de cada sitio web, además de revisar el archivo robots.txt donde el sitio mostraba si estaba disponible para la extracción automática de datos. Algunos de estos sitios no permiten la extracción automática de datos, planteando implicaciones legales significativas para la universidad. Para superar este desafío, se llevó a cabo una revisión exhaustiva de las políticas de cada sitio, asegurando que la extracción de datos cumpliera con sus términos y no generara conflictos legales. Este proceso garantizó una implementación ética y legal de las rutinas de Crawlers y Web Scraping en cada uno de los sitios objetivo, contribuyendo a la integridad y sostenibilidad del proyecto.

#### *4.2.1.4 Resultados.*

El resultado de esta etapa inicial no solo consiste en el conjunto de datos recolectado de manera automatizada y sistemática, abordando así la problemática inicial de acceso limitado a información externa, sino que también destaca el desarrollo de las rutinas de Crawlers y Web Scraping. Estas rutinas, diseñadas estratégicamente para emular la navegación humana en la web, han demostrado su eficacia al recolectar datos de repositorios académicos clave.

Este corpus de datos no solo representa una solución a la limitación de acceso, sino que también ha sido implementado en una interfaz de búsqueda intuitiva. En esta interfaz, los usuarios tienen la capacidad de realizar extracciones de datos utilizando las rutinas desarrolladas en todos los repositorios mencionados simultáneamente. Los parámetros de búsqueda incluyen términos de búsqueda específicos, la posibilidad de filtrar por autor, tipo de documento (como PDF, revista, artículo), así como establecer rangos de años inicial y final.

Cabe resaltar que el enfoque no solo se centra en la recolección de datos, sino en proporcionar a los usuarios una herramienta poderosa y versátil para la adquisición de datos externos. La implementación de una interfaz visual para la búsqueda, donde se ingresan los parámetros, y otra interfaz de resultados, donde se muestran los resultados de la búsqueda con los parámetros establecidos, añade un componente interactivo y amigable al proceso de extracción de datos.

Con estos cimientos sólidos, que incluyen tanto la eficaz recolección de datos como la implementación de una interfaz amigable, se avanza a la siguiente fase del proyecto con la certeza de haber superado un obstáculo clave en la adquisición de datos externos para la Universidad CESMAG.

### *4.2.2 Construcción de la base de datos.*

Considerando las necesidades fundamentales de los investigadores y los componentes esenciales de recopilación de datos, se ha establecido un objetivo específico: crear una base de datos con la información recolectada. Este proceso se ha desarrollado en varias etapas, cada una diseñada para garantizar la integridad, eficiencia y utilidad de la base de datos resultante.

En primer lugar, se llevó a cabo un análisis exhaustivo de las necesidades de los investigadores y de los requisitos de recopilación de datos. Esto incluyó identificar los tipos de información que los investigadores necesitan acceder y registrar durante sus actividades de búsqueda, así como los formatos y estructuras de datos más adecuados para almacenar y gestionar esta información de manera eficiente.

#### *4.2.2.1 Diseño de la base de datos.*

La base de datos está diseñada para mantener una relación entre las búsquedas realizadas por los usuarios y los resultados obtenidos de esas búsquedas. Esto permite un seguimiento preciso de las interacciones de los investigadores con la plataforma y facilita la generación de análisis y métricas útiles para comprender mejor las dinámicas de búsqueda y el uso de la información dentro de la comunidad investigativa.

Una vez definido el esquema de la base de datos, se procedió a la implementación técnica utilizando MySQL como gestor de base de datos. Se crearon las tablas necesarias según el diseño establecido, definiendo cuidadosamente los tipos de datos, las restricciones de integridad y las relaciones entre las tablas para garantizar la coherencia y la integridad de los datos almacenados.

La base de datos cuenta con dos tablas principales:

**scrapingapp_busqueda:** Esta tabla almacena los detalles principales de cada búsqueda realizada por los usuarios, incluyendo información como el término de búsqueda, la fecha y hora de la búsqueda, y cualquier otro dato relevante relacionado con la consulta. Esta cuenta con los siguientes campos:

**idBusqueda (int):** Identificador único de la búsqueda.
**fec_bus (date):** Fecha de la búsqueda.
**tem_bus (varchar):** Tema de la búsqueda.
**aut_bus (varchar):** Autor de la búsqueda.
**year_ini (varchar):** Año inicial de la búsqueda.
**year_fin (varchar):** Año final de la búsqueda.
**tipo_doc (varchar):** Tipo de documento consultado (opcional).

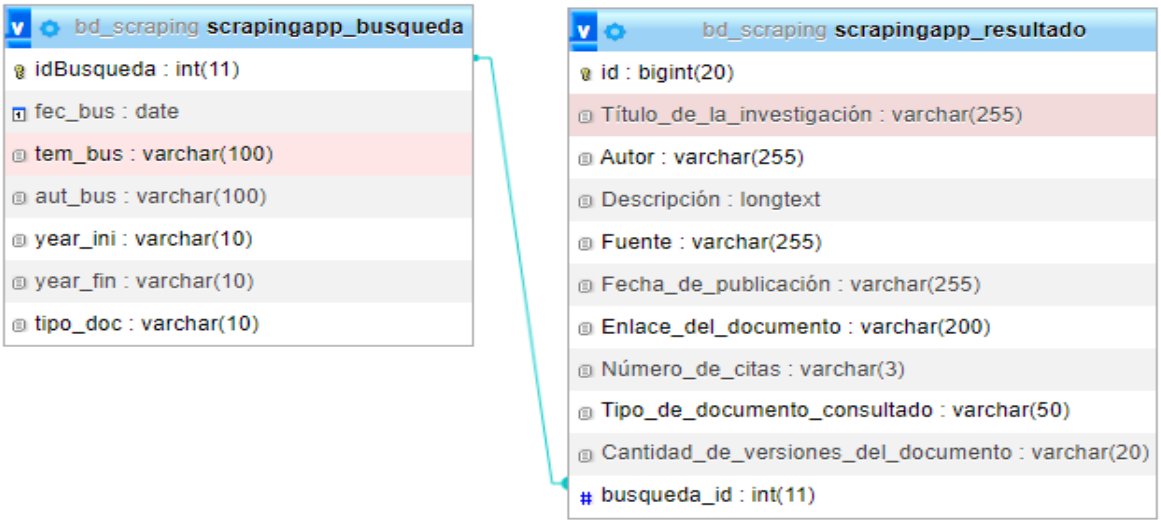
**scrapingapp_resultado**: En esta tabla se registran los resultados obtenidos de cada búsqueda realizada. Se incluyen detalles como el título del resultado, la URL asociada, la relevancia

del resultado, y cualquier otro dato pertinente relacionado con los resultados obtenidos. Esta cuenta con los siguientes campos:

**id (bigint):** Identificador único del resultado.
**Título_de_la_investigación (varchar):** Título de la investigación.
**Autor (varchar):** Autor del resultado.
**Descripción (longtext):** Descripción del resultado.
**Fuente (varchar):** Fuente del resultado.
**Fecha_de_publicación (varchar):** Fecha de publicación del resultado.
**Enlace_del_documento (varchar):** Enlace al documento del resultado.
**Número_de_citas (varchar):** Número de citas del resultado.
**Tipo_de_documento_consultado (varchar):** Tipo de documento consultado del resultado.
**Cantidad_de_versiones_del_documento (varchar):** Cantidad de versiones del documento del resultado.
**busqueda_id (int):** ID de la búsqueda asociada a este resultado (clave foránea, opcional), esta clave relaciona los resultados con la búsqueda especifica realizada.
Esto último, se puede comprender mucho mejor mediante el siguiente diseño de la base de datos:

Fig 5. Diseño de la base de datos.

Fuente: Investigación propia.



Después de definir los campos necesarios tanto para el almacenamiento de los parámetros de búsqueda como para los resultados de la búsqueda más relevantes para la comunidad académica, se logró completar e integrar la base de datos.

#### *4.2.2.2 Resultados.*

Como resultado de este proceso exhaustivo de creación de la base de datos, se ha logrado desarrollar una plataforma completamente funcional y robusta que cumple con los objetivos establecidos. La base de datos, implementada utilizando MySQL como gestor, está diseñada de manera relacional y cuenta con dos tablas principales: scrapingapp_busqueda y scrapingapp_resultado.

La base de datos está operativa y lista para almacenar los datos cruciales tomados de las búsquedas realizadas por los investigadores, así como los resultados obtenidos de esas búsquedas. Cada componente de la base de datos ha sido cuidadosamente diseñado y probado para garantizar su funcionamiento correcto y su capacidad para manejar grandes volúmenes de datos de manera eficiente.

Finalmente, los datos recolectados serán útiles para su posterior análisis en búsqueda de patrones en las líneas de investigación. Este análisis permitirá a la universidad comprender qué líneas de investigación o tendencias están siendo llevadas a cabo en las mismas. Este conocimiento será invaluable para la toma de decisiones estratégicas y la planificación futura de actividades de investigación en la universidad CESMAG.

### *4.2.3 Validación del Módulo de Recolección de Datos*

Para cumplir con el objetivo de validar la herramienta de recolección de datos en función de los criterios de precisión y calidad establecidos, se llevó a cabo un proceso detallado de evaluación que implicó la participación activa de los líderes de investigación de la Vicerrectoría Académica de la universidad CESMAG. A continuación, se describe en detalle cómo se llevó a cabo este proceso:

#### *4.2.3.1 Preparación del Cuestionario de Evaluación*

En primer lugar, se diseñó un cuestionario exhaustivo que abarcaba todos los criterios necesarios para evaluar la precisión y calidad de los datos obtenidos mediante el módulo de recolección de datos. Este cuestionario incluía preguntas detalladas sobre cada uno de los siguientes criterios de calidad:

**Claridad del título de la investigación:** Este criterio evaluaba si el título del documento obtenido era claro y descriptivo, lo que facilitaría su comprensión y relevancia para los investigadores.

**Correcta identificación de los autores:** Se verificaba si los nombres de los autores estaban correctamente identificados en los resultados obtenidos, lo que contribuiría a la atribución adecuada de la autoría en la investigación.

**Detalle y precisión de la descripción del contenido:** Se evaluaba si la descripción del contenido proporcionada era suficientemente detallada y precisa para comprender el alcance y la relevancia del documento.

**Claridad en la identificación de la fuente:** Este criterio buscaba determinar si se identificaba de manera clara y precisa la fuente externa de la cual se extrajeron los datos, lo que añadiría credibilidad y transparencia a la información obtenida.

**Relevancia de la fecha de publicación del documento:** Se analizaba si la fecha de publicación del documento consultado era relevante para la investigación, lo que ayudaría a contextualizar la información en el tiempo.

**Accesibilidad del enlace del documento:** Se verificaba si al hacer clic en el título de la investigación proporcionado, se llevaba al documento o información completa de manera adecuada, lo que garantizaría el acceso a la fuente original.

**Pertinencia del número de citas del documento:** Se evaluaba si el número de citas del documento consultado era relevante para la investigación, lo que indicaría su impacto y relevancia en el ámbito académico.

**Adecuación del tipo de documento consultado:** Se analizaba si el tipo de documento consultado se ajustaba a las necesidades de investigación, lo que aseguraría la idoneidad de la información obtenida.

**Importancia de la cantidad de versiones del documento:** Se evaluaba si la cantidad de versiones del documento consultado era un factor relevante para la investigación, lo que podría indicar su evolución o actualización a lo largo del tiempo.

**Calidad de los datos adquiridos en términos de precisión:** Se examinaba si la calidad de los datos adquiridos cumplía con las expectativas en términos de precisión, lo que garantizaría la fiabilidad de la información obtenida.

**Confiabilidad de los datos adquiridos:** Se verificaba si la confiabilidad de los datos adquiridos era satisfactoria para la investigación, lo que aseguraría su validez y consistencia.

**Suficiencia de la información obtenida para llevar a cabo una investigación adecuada:** Se analizaba si la información obtenida proporcionaba todos los datos necesarios para llevar a cabo una investigación de manera adecuada, lo que garantizaría su utilidad y relevancia.
Utilidad de los resultados obtenidos para avanzar en la investigación: Se evaluaba si los resultados obtenidos eran útiles para avanzar en la investigación, lo que indicaría su relevancia y aplicabilidad en el contexto investigativo.

**Agilidad del proceso de obtención de datos mediante el uso de crawlers y web scraping:** Se verificaba si la aplicación de crawlers y web scraping agilizaba el proceso de obtención de datos en comparación con métodos manuales, lo que aumentaría la eficiencia del proceso.

**Experiencia de usuario satisfactoria en términos de usabilidad:** Se analizaba si la herramienta proporcionaba una experiencia de usuario satisfactoria en términos de usabilidad, lo que aseguraría su eficacia y facilidad de uso.

**Recomendación de la herramienta desarrollada para la adquisición de datos externos a otros investigadores:** Se evaluaba si los usuarios recomendarían la herramienta desarrollada para la adquisición de datos externos a otros investigadores, lo que indicaría su utilidad y valor en el ámbito académico y científico.

#### *4.2.3.2 Ejecución del Proceso de Validación*

Se convocó a los líderes de investigación de la Vicerrectoría Académica de la universidad CESMAG incluyendo al vicerrector de investigación y extensión **DOCTOR JAVIER ALEJANDRO JIMÉNEZ TOLEDO** para participar en el proceso de validación de la herramienta de recolección de datos. Cada líder recibió acceso a la herramienta y se le proporcionó orientación sobre cómo utilizarla para llevar a cabo sus actividades de búsqueda. Posteriormente, se les solicitó que completaran el cuestionario de evaluación basado en su experiencia utilizando la herramienta.
Este cuestionario, cada pregunta contaba con las siguientes opciones de respuesta:

**Muy en desacuerdo:** Indica que no estás de acuerdo en absoluto con la afirmación. Crees que la respuesta a la pregunta es totalmente negativa o que la situación descrita no se cumple en lo más mínimo.

**En desacuerdo:** Sugiere que no estás completamente de acuerdo con la afirmación. Puede significar que crees que la respuesta a la pregunta es en su mayoría negativa, aunque con algunos aspectos que podrían mejorar.

**De acuerdo:** Implica que estás conforme con la afirmación. Consideras que la respuesta a la pregunta es en general positiva, y que la situación descrita se ajusta a tus necesidades o expectativas en cierta medida.

**Muy de acuerdo:** Indica que estás completamente de acuerdo con la afirmación. Sugiere que la respuesta a la pregunta es totalmente positiva y supera tus expectativas en gran medida.:
La encuesta se puede revisar en el apartado de anexos:

Fig 6. Presentación de la herramienta a lideres de investigación.

Fuente: Investigación propia.

Fig 7. Presentación de la herramienta a lideres de investigación.

Fuente: Investigación propia.

#### *4.2.3.3 Resultados*

Una vez recopiladas las respuestas de los líderes de investigación, se procedió a analizar los resultados obtenidos en el cuestionario de evaluación. Se obtuvieron los siguientes resultados:

Tabla 2. Resultados De la encuesta aplicada.

Fuente: Investigación propia.

| **Resultados de la encuesta** | | | | |
|---|---|---|---|---|
| | **Muy en desacuerdo:** | **En desacuerdo:** | **De acuerdo:** | **Muy de acuerdo:** |
| 1. El título de la investigación obtenido es claro y descriptivo. | | | 38.46% | 61.54% |
| 2. Los nombres de los autores están correctamente identificados en los resultados obtenidos. | | | 38.46% | 61.54% |
| 3. La descripción del contenido proporcionada es suficientemente detallada y precisa. | | | 46.15% | 53.85% |
| 4. Se identifica la fuente externa de la cual se extrajeron los datos de manera clara y precisa. | | | 30.77% | 69.23% |

| | | | | |
|---|---|---|---|---|
| 5. La fecha de publicación del documento consultado es relevante para su investigación. | | | 30.77% | 69.23% |
| 6. Al hacer clic en el título de la investigación proporcionado, lleva al documento o información completa de manera adecuada. | | | 30.77% | 69.23% |
| 7. El número de citas del documento consultado es relevante para su investigación. | | | 30.77% | 69.23% |
| 8. El tipo de documento consultado se ajusta a sus necesidades de investigación (PDF, ARTICULO, REVISTA, ETC). | | | 30.77% | 69.23% |
| 9. La cantidad de versiones del documento consultado es un factor relevante para su investigación. | | | 38.46% | 61.54% |

| | | | | |
|---|---|---|---|---|
| 10. La calidad de los datos adquiridos cumple con las expectativas en términos de precisión. | | | 38.46% | 61.54% |
| 11. La confiabilidad de los datos adquiridos es satisfactoria para su investigación. | | | 30.77% | 69.23% |
| 12. La información obtenida proporciona todos los datos necesarios para llevar a cabo una investigación de manera adecuada. | | | 30.77% | 69.23% |
| 13. Encuentra útiles los resultados obtenidos para avanzar en su investigación. | | | 38.46% | 61.54% |
| 14. La aplicación de Crawlers y Web Scraping en este módulo de recolección automática de datos agiliza el proceso de obtención de datos en comparación con métodos manuales. | | | 23.08% | 76.92% |

| | | | | |
|---|---|---|---|---|
| 15. La herramienta proporciona una experiencia de usuario satisfactoria en términos de usabilidad. | | | 30.77% | 69.23% |
| 16. ¿Recomendaría la herramienta desarrollada para la adquisición de datos externos a otros investigadores? | | | 30.77% | 69.23% |

Los resultados de la validación de la herramienta de recolección de datos muestran una recepción positiva por parte de los usuarios, lo que confirma el logro del objetivo de desarrollar una herramienta efectiva y útil para los procesos de investigación. La mayoría de los usuarios expresaron satisfacción con varios aspectos de la herramienta, incluida la claridad del título de la investigación, la identificación precisa de los autores y la relevancia de la información recopilada.

Es alentador observar que los usuarios encontraron útil la aplicación de Crawlers y Web Scraping, destacando su capacidad para agilizar el proceso de obtención de datos en comparación con métodos manuales. Esta retroalimentación valida la eficacia de la herramienta en cumplir con su propósito de mejorar la eficiencia en la recolección de datos para investigaciones.

Además, las felicitaciones recibidas por parte de los usuarios subrayan el valor percibido de la herramienta y su importancia en los procesos de investigación. Este reconocimiento positivo no solo valida la efectividad de la herramienta, sino que también resalta su contribución significativa para facilitar y mejorar la calidad de los trabajos de investigación.

#### *4.2.3.4 Documento de Aceptación por los Líderes de Investigación*

Para validar formalmente el uso y la aceptación de la herramienta de recolección de datos, se solicitó a los 17 líderes de investigación que firmaran un documento de aceptación. En este documento, los líderes de investigación verificaron que habían hecho uso de la herramienta y completado el cuestionario de evaluación. Su firma representaba su aprobación, validación y aceptación de la herramienta de recolección de datos para su uso en actividades de investigación futuras. Este documento se encuentra en el apartado de anexos.

### *4.3 aspectos propios de la metodología.*

#### *4.3.1 Metodología SCRUM.*

Scrum es un marco de trabajo ágil que facilita la colaboración y la adaptabilidad en el desarrollo de proyectos. En nuestro caso, aplicaremos Scrum para garantizar una ejecución eficiente y satisfacer las necesidades de la Vicerrectoría de Investigaciones.

Principalmente se ha definido los Roles según la metodología Scrum para asignar responsabilidades para determinar las funcionalidades de cada uno.

Tabla 3. Roles En metodología SCRUM.

Fuente: Investigación propia.

| **Roles** | **Responsable** |
|---|---|
| SCRUM MASTER | Daniel Clemente Gualteros Sinsajoa |
| PRODUCT OWNER | Daniel Clemente Gualteros Sinsajoa |
| DESARROLLADOR DE SOFTWARE | Andres Sebastian Trejo Quintero |
| DESARROLLADOR DE SOFTWARE | David Esteban Finlay Estrella |

La siguiente tabla describe las responsabilidades de SCRUM:

Tabla 4. Responsabilidades en el desarrollo de SCRUM.

Fuente: Investigación propia.

| **Roles** | **Responsable** |
|---|---|

| | |
|---|---|
| SCRUM MASTER | Encargado de asegurar la correcta implementación de la metodología Scrum, así como facilitar la comunicación y colaboración del equipo. |
| PRODUCT OWNER | Responsable de definir y priorizar los elementos del backlog, así como de tomar decisiones que maximicen el valor del producto. |
| DESARROLLADOR DE SOFTWARE | Encargado de desarrollar las rutinas de crawling, web scraping y otros aspectos de la implementación del sistema. |

### *4.3.2 Análisis del sistema.*

#### *4.3.2.1 Requerimientos funcionales.*

Tabla 5. Requerimientos funcionales.

Fuente: Investigación propia.

| **No** | **Nombre** | **Descripción** |
|---|---|---|
| RF1 | Programación del scraping | El sistema debe ser capaz de realizar scraping de manera efectiva y eficiente en los repositorios académicos especificados. |
| RF2 | Motor de búsqueda | El motor de búsqueda debe indexar de manera eficiente las fuentes de datos extraídas de los sitios web para garantizar la rápida recuperación de resultados relevantes. |

| | | |
|---|---|---|
| RF3 | Filtros de búsqueda | El sistema debe permitir a los usuarios aplicar filtros personalizados a sus consultas de búsqueda para obtener resultados más relevantes y específicos según sus necesidades. |
| RF4 | Filtros de resultados | El sistema debe permitir ordenar y filtrar resultados según diferentes criterios. |
| RF5 | Visualización de datos | El sistema debe proporcionar una interfaz de usuario intuitiva y clara para visualizar los resultados de búsqueda en forma de tabla, facilitando la comprensión y el análisis de los datos. |
| RF5 | Almacenamiento de datos | El sistema debe ser capaz de almacenar los datos recopilados de los repositorios académicos. |

#### *4.3.2.2 Requerimientos no funcionales.*

Tabla 6. Requerimientos no funcionales.

Fuente: Investigación propia.

| **No** | **Nombre** | **Descripción** |
|---|---|---|
| RN1 | Usabilidad | El sistema debe tener una interfaz de usuario que sea fácil para los usuarios. |
| RN2 | Rendimiento | El sistema debe realizar el scraping y la indexación de manera eficiente para proporcionar resultados en tiempo real. |
| RN3 | Mantenibilidad | El código del sistema debe estar bien estructurado y documentado para facilitar futuras actualizaciones y mantenimiento. |

Tabla 7 Historias de usuario.

Fuente: Investigación propia.

| **Historias de usuario** | | | | | | |
|---|---|---|---|---|---|---|
| Identificador (ID) de la Historia | Nombre de la actividad | Historia de Usuario | Estado | Dimensión / Esfuerzo | Iteración (Sprint) | Prioridad |
| **HUWS-001** | Evaluación e identificación de repositorios de importancia | Como usuario, requiero que el sistema me brinde la capacidad de buscar información externa en repositorios y sitios web, permitiéndome así obtener datos relevantes para mis procesos y estudios de investigación. Esta funcionalidad es esencial para facilitar la adquisición de información pertinente de manera eficiente, potenciando mis actividades de investigación y análisis dentro de la plataforma. | Terminado | 2 semanas | 1 | Media |

| | | | | | | |
|---|---|---|---|---|---|---|
| HUWS-002 | Extracción de información de con rutinas de web crawleres y web scraping. | Como un Usuario, requiero que esta información tenga detalles específicos como título, autor, fecha, descripción, fuente, fecha de publicación, enlaces, numero de citas tipo de documento y cantidad de versiones de estos sitios web. | Terminado | 2 semanas | 1 | Alta |
| HUWS-003 | Busqueda de información externa | Como usuario, deseo tener un campo de búsqueda intuitivo que me permita realizar consultas en tiempo real y, además, quiero la capacidad de aplicar filtros avanzados, como la búsqueda por autor, establecer un rango de fechas (inicial y final), y seleccionar el tipo de documento, como artículos, revistas o archivos PDF. Esto garantizará que mis búsquedas sean precisas y personalizadas, mejorando significativamente mi experiencia al buscar información en los repositorios académicos. | Terminado | 2 semanas | 3 | Alta |

| | | | | | | |
|---|---|---|---|---|---|---|
| HUWS-004 | Visualizacion eficiente de resultados | Como usuario requiero que el sistema me proporcione una visualización clara y detallada de los resultados de búsqueda. Esta funcionalidad me permitirá examinar de manera eficiente la información relevante y tomar decisiones informadas sobre qué documentos explorar con mayor detalle. | Terminado | 2 semanas | 3 | Alta |
| HUWS-005 | Almacenamiento de datos eficiente | Como usuario requiero que toda la información relevante recolectada sea almacenada en una base de datos para su posterior análisis. | Terminado | 2 semanas | 2 | Alta |

### *4.3.2.3 Sprint.*

Tabla 8. Sprint 1.

Fuente: Investigación propia.

| **Objetivo específico 1:** Generar rutinas de crawleres y web scraping para la recolección de información en la web | | |
|---|---|---|
| **Sprint 1** | | |
| **Historia de usuario** | **Actividad** | **Prioridad** |
| **HUWS-001** | Evaluación e identificación de repositorios de importancia | Media |
| **HUWS-002** | Extracción de información con rutinas de crawlers y web scraping | Alta |
| **HUWS-004** | Visualización Eficiente de resultados | Alta |
| **Criterios de aceptación** | | |
| • Tablas de análisis con los repositorios más relevantes tomando en cuenta sus políticas de seguridad y su importancia investigativa.<br>• Rutinas de crawlers y web scraping.<br>• La interfaz gráfica permite realizar búsquedas. Agregando filtros avanzados, incluyendo opciones como autor, año de finalización y tipo de documento, artículos y revistas.<br>• La plataforma presenta los resultados de búsqueda en tarjetas, con información detallada y correcta a los parámetros de búsqueda que incluye el título del articulo o revista, datos del autor, fecha de publicación, tipo de documento, nombre del repositorio, numero de citas, cantidad de versiones y un resumen (abstract). La presentación adopta un diseño visual minimalista y fácil de comprender para el usuario. | | |

Desarrollo de una herramienta de adquisición automática de datos de fuentes externas, en el sistema de gestión de información de la Vicerrectoría de Investigaciones de la Universidad CESMAG, mediante crawlers y web scraping.

| | |
|---|---|
| **Estado de implementación:** | **Terminado** |

## Resultados del desarrollo del Sprint 1:

Fig 8.Interfaz de búsqueda académica.

Fuente: Investigación propia.

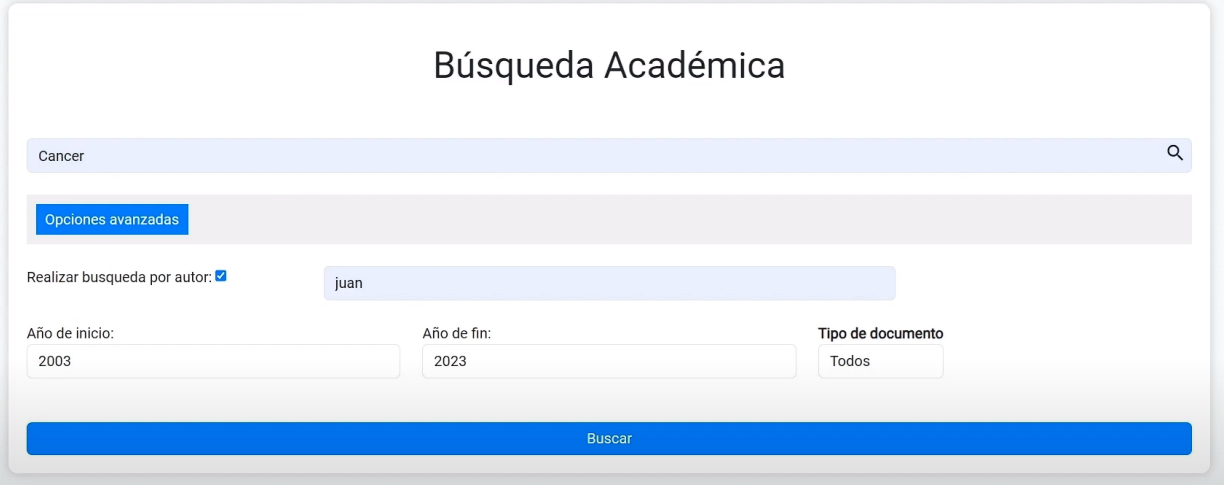



Fig 9. Interfaz de muestra de resultados.

Fuente: Investigación propia.

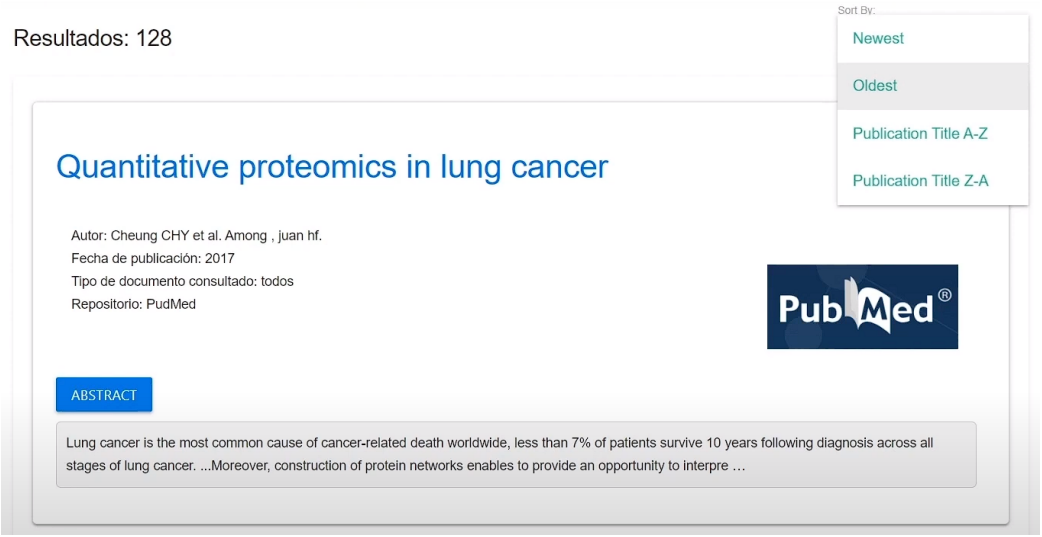


Fig 10. Interfaz de muestra de resultados (barra de paginación).

Fuente: Investigación propia.

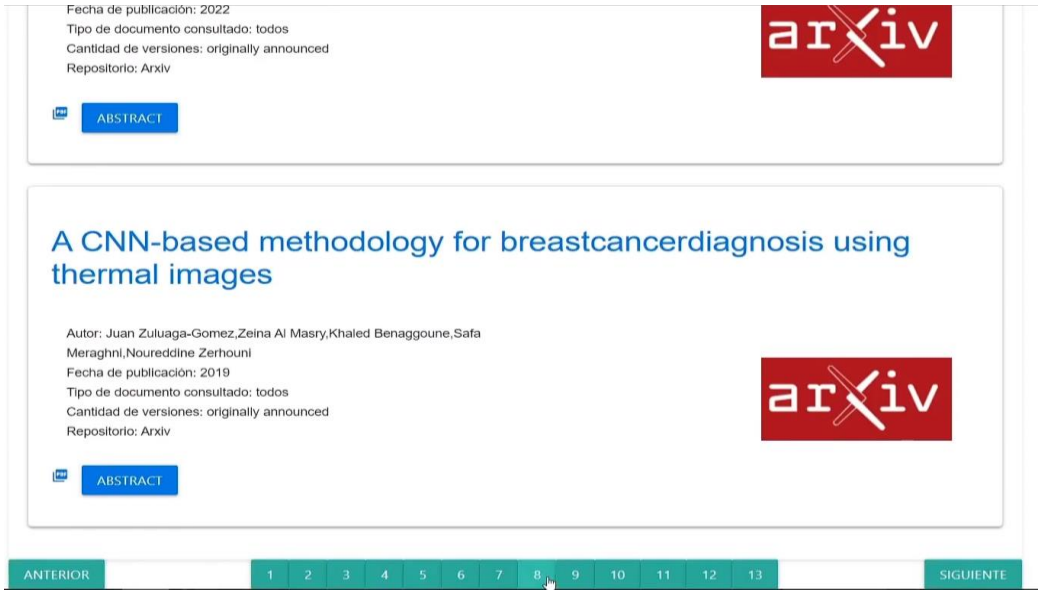



Fig 11. Resultados de más repositorios (Recolecta).

Fuente: Investigación propia.

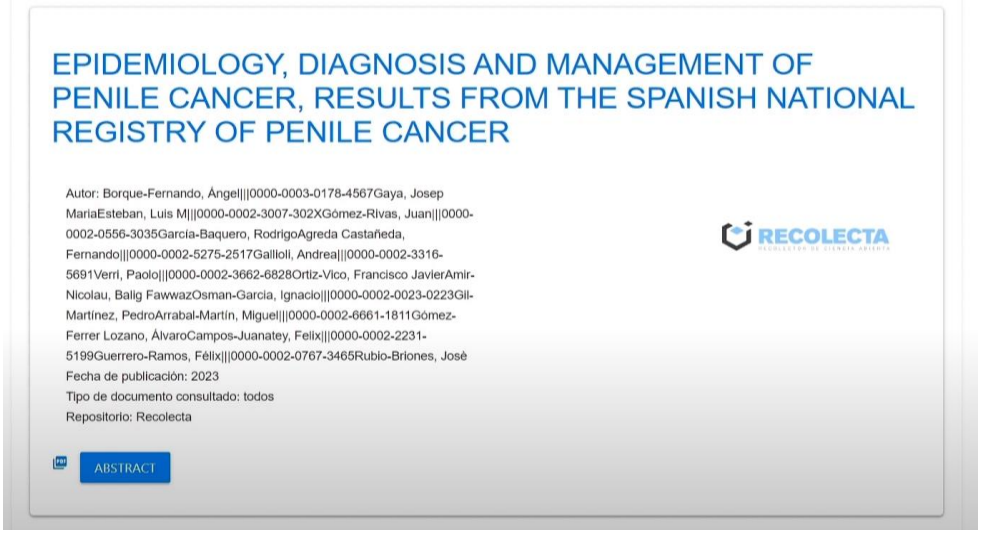


Tabla 9. Sprint 2.

Fuente: Investigación propia.

| **Objetivo específico 2:** Crear una base de datos con la información recolectada | | |
|---|---|---|
| **Sprint 2** | | |
| **Historia de usuario** | **Actividad** | **Prioridad** |
| **HUWS-005** | Almacenamiento de datos eficientes | Alta |
| **Criterios de aceptación** | | |
| • Se establece una base de datos relacional que captura los datos más relevantes de las búsquedas y los resultados | | |
| **Estado de implementación:** | | **Terminado** |

**Resultados del desarrollo del sprint 2:**

Después de definir los campos necesarios tanto para el almacenamiento de los parámetros de búsqueda como para los resultados de la búsqueda más relevantes para la comunidad académica, se logró completar e integrar la base de datos. La información contenida y establecida se presenta de la siguiente manera:

Fig 12. Base de datos creada.

Fuente: Investigación propia.

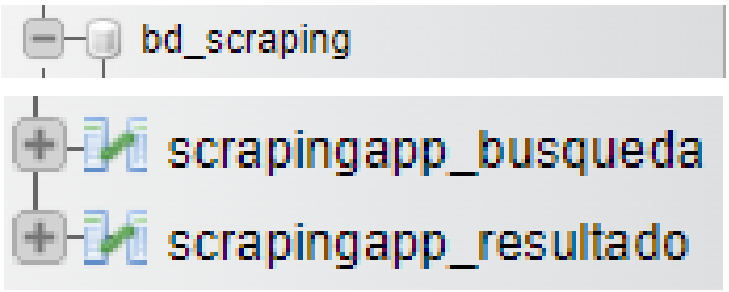


Fig 13. Prueba de registro de datos exitoso.

Fuente: Investigación propia.

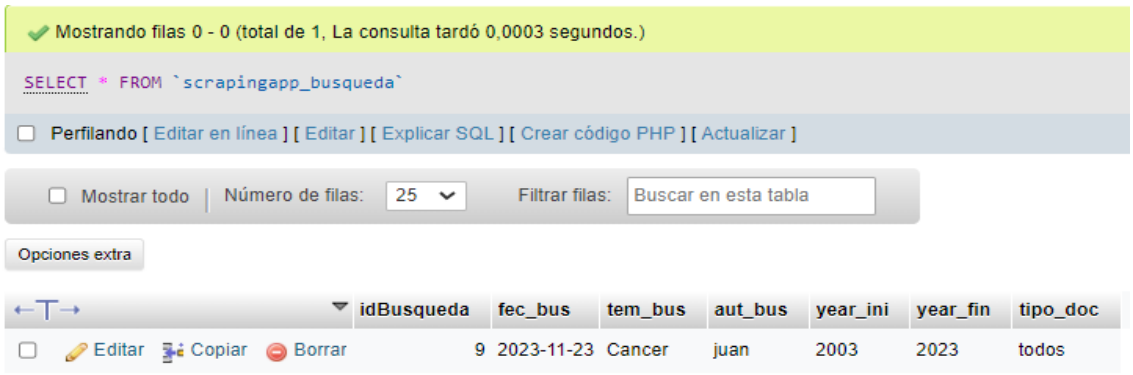
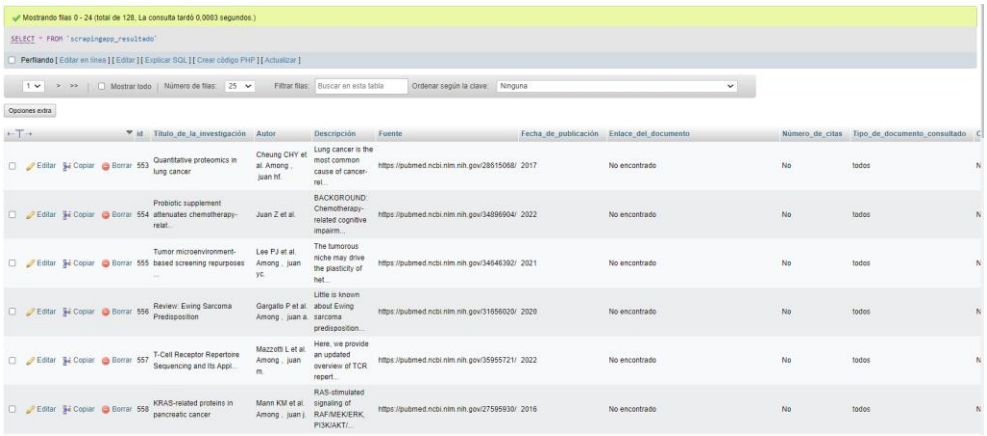



Fig 14. Verificación de registros ingresados.

Fuente: Investigación propia.



Esta estructura de base de datos ha permitido una organización clara y efectiva de la información, facilitando tanto el registro de las consultas realizadas como el almacenamiento de los resultados obtenidos, lo que contribuye a optimizar el proceso de investigación para la comunidad académica.

Tabla 10. Sprint 3.

Fuente: Investigación propia.

| **Objetivo específico 3:** Validar la herramienta de recolección de datos en función de los criterios de evaluación de precisión y calidad establecidos | | |
|---|---|---|
| **Sprint 3** | | |
| **Historia de usuario** | **Actividad** | **Prioridad** |

| | | |
|---|---|---|
| **HUWS-003** | Búsqueda de información externa | Alta |
| **HUWS-004** | Visualización Eficiente de resultados | Alta |
| **Criterios de aceptación** | | |
| • La interfaz gráfica permite realizar búsquedas. Agregando filtros avanzados, incluyendo opciones como autor, año de finalización y tipo de documento, artículos y revistas.<br><br>• La herramienta presenta los resultados de búsqueda en tarjetas, con información detallada y correcta a los parámetros de búsqueda que incluye el título del articulo o revista, datos del autor, fecha de publicación, tipo de documento, nombre del repositorio, numero de citas, cantidad de versiones y un resumen (abstract). La presentación adopta un diseño visual minimalista y facil de comprender para el usuario. | | |
| **Estado de implementación:** | | **Terminado** |

## 5. ANÁLISIS DE RESULTADOS

Durante la validación del producto de software desarrollado, se observaron resultados alentadores que respaldan su utilidad y efectividad para la adquisición automatizada de datos externos. Las encuestas realizadas revelaron un alto nivel de satisfacción por parte de los usuarios, quienes destacaron la facilidad de uso de la herramienta y su capacidad para agilizar el proceso de obtención de datos. Específicamente, un 81.25% de los encuestados expresaron un alto grado de satisfacción con la herramienta, lo que indica una recepción positiva por parte de los Lideres de investigacion. Además, se identificaron áreas de mejora durante el proceso de validación. Por ejemplo, algunos investigadores mencionaron que la interfaz de usuario podría ser más intuitiva y atractiva visualmente, lo que sugiere la necesidad de realizar ajustes en el diseño para mejorar la experiencia del usuario. Asimismo, se destacó la importancia de del uso de estas técnicas de recolección automática de datos, para la vicerrectoría académica.

Los resultados obtenidos en la validación del producto se alinean estrechamente con el marco teórico establecido para la investigación. Los antecedentes proporcionados muestran que el uso de técnicas de crawlers y web scraping para la recopilación automatizada de datos ha sido ampliamente explorado en la literatura académica y profesional. El producto de software

desarrollado contribuye a este cuerpo de conocimientos al proporcionar una solución práctica y eficiente para la adquisición de datos externos, alineándose con las mejores prácticas y tendencias actuales en el campo.

La mayoría de lideres de investigación de la universidad CESMAG coincidieron en que la herramienta cumple con sus expectativas en términos de precisión en la recopilación de datos, lo que indica que el producto se alinea con las expectativas teóricas establecidas en la investigación.

Los antecedentes presentados revelan una amplia gama de proyectos e investigaciones que han utilizado técnicas de crawlers y web scraping para diversos propósitos. Estos estudios proporcionan un contexto relevante para la investigación actual y destacan la importancia y la versatilidad de estas técnicas en diferentes contextos y aplicaciones.

Además, al comparar los resultados obtenidos con los antecedentes presentados, se observa una coherencia entre la naturaleza y los objetivos del proyecto actual y los proyectos previamente realizados en la misma área de estudio. Por ejemplo, el proyecto desarrollado en la Universidad CESMAG en Pasto, Colombia, proporcionó una base sólida para entender la estructura de la web colombiana, mientras que el trabajo en la Universidad Javeriana de Bogotá se centró en la optimización de procesos de mercadeo mediante la recopilación automatizada de información de productos. Estos antecedentes respaldan y contextualizan la relevancia y el alcance del proyecto actual en el campo de la adquisición automatizada de datos.

los resultados de la investigación en su totalidad están alineados con los referentes teóricos y antecedentes presentados, lo que demuestra la coherencia y la contribución del proyecto al cuerpo de conocimiento existente en el área de estudio. El proyecto no solo valida la utilidad práctica de las técnicas de crawlers y web scraping, sino que también enriquece la comprensión teórica y la aplicación práctica de estas técnicas en el ámbito de la investigación científica y académica.

## CONCLUSIONES

El empleo de una herramienta de recolección de datos eficiente a través de crawlers y web scraping garantiza la obtención de información confiable y precisa de repositorios web externos, superando en efectividad a los métodos tradicionales de recolección de datos.

La implementación de rutinas de Crawlers y Web Scraping representa un paso crucial para superar las limitaciones de acceso a datos externos y aprovechar la información valiosa contenida en repositorios académicos clave. En un entorno donde la cantidad y diversidad de datos disponibles en la web continúa creciendo exponencialmente, estas técnicas se han convertido en herramientas indispensables para investigadores y organizaciones que buscan acceder a información relevante de manera sistemática y eficiente. El diseño estratégico de estas rutinas es fundamental para garantizar una cobertura exhaustiva de la información relevante. Esto implica identificar y adaptar las rutinas de Crawlers y Web Scraping a diversas fuentes web, incluidos repositorios académicos, bases de datos en línea, sitios web de revistas científicas, entre otros. Al personalizar las rutinas para cada fuente específica, se puede garantizar una extracción precisa y completa de la información deseada, maximizando así el valor de los datos obtenidos.

El diseño y la implementación de la base de datos han sido pilares fundamentales para proporcionar un repositorio centralizado y estructurado que almacene y gestione de manera eficiente la información recolectada. Este repositorio no solo ha facilitado el almacenamiento de los datos obtenidos mediante las rutinas de Crawlers y Web Scraping, sino que también ha asegurado su integridad y coherencia, lo que es esencial para su posterior análisis y utilización por parte de los investigadores. Este repositorio centralizado proporciona a los investigadores un acceso rápido y fácil a los datos almacenados, facilitando así su análisis y utilización en el avance de la investigación y la generación de conocimientos significativos.

La evaluación exhaustiva de la herramienta de recolección de datos por parte de los líderes de investigación de la Vicerrectoría Académica de la Universidad CESMAG confirmó su eficacia y utilidad para los procesos de investigación. Los resultados reflejaron una recepción positiva, destacando la claridad del título de la investigación obtenido, la identificación precisa de los autores, y la relevancia de la información recopilada. Además, se reconoció la capacidad de las

rutinas de Crawlers y Web Scraping para agilizar significativamente el proceso de obtención de datos en comparación con métodos manuales, lo que contribuye a mejorar la eficiencia en la recolección de datos para investigaciones académicas en la universidad.

# RECOMENDACIONES

Se recomienda implementar medidas adicionales de seguridad al utilizar crawlers y web scraping para garantizar el acceso ético y legal a los datos, dado que muchos repositorios pueden bloquear el acceso debido a estas prácticas. Esto podría incluir la revisión y el cumplimiento estricto de los términos de servicio de cada sitio web, así como la exploración de opciones de autorización o consentimiento para la extracción de datos.

Se recomienda considerar el uso de APIs (Interfaces de Programación de Aplicaciones) siempre que sea posible para la recolección de información, ya que proporcionan una forma más estructurada y segura de acceder a los datos de manera autorizada. La utilización de APIs puede ofrecer una alternativa más confiable y sostenible para la obtención de datos en comparación con crawlers y web scraping.

Se recomienda realizar un estudio más profundo de los datos almacenados en la base de datos en el futuro. Esto podría incluir análisis de tendencias, identificación de patrones, y evaluación de la calidad y relevancia de los datos para garantizar su utilidad y validez continua en el contexto de la investigación académica en la Universidad CESMAG. Este estudio adicional permitirá optimizar el uso de la base de datos y maximizar su valor en futuras investigaciones.

Desarrollo de una herramienta de adquisición automática de datos de fuentes externas, en el sistema de gestión de información de la Vicerrectoría de Investigaciones de la Universidad CESMAG, mediante crawlers y web scraping.

## REFERENCIAS BIBLIOGRÁFICAS

[1] "Plataforma SCIENTI - Colombia | Minciencias." https://minciencias.gov.co/scienti (accessed Apr. 15, 2023).

[2] "GrupLAC." https://scienti.minciencias.gov.co/gruplac/jsp/index.jsp (accessed Apr. 15, 2023).

[3] VERA VERA DAYANNA MELINA, "MODELO BASADO EN WEB SCRAPING Y CHATBOT PARA CONOCER LAS HABILIDADES DIGITALES MÁS DEMANDADAS EN SECTOR DE TI," Sep. 2022. http://repositorio.ug.edu.ec/bitstream/redug/64139/1/VERA%20VERA%20DAYANNA%20MELINA.pdf (accessed Mar. 28, 2023).

[4] M. En and C. De Datos, "UNIVERSIDAD POLITÉCNICA DE MADRID ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS TRABAJO FIN DE MÁSTER".

[5] I. R. Daniel and M. Espinosa, "Tecnológico Nacional de México Tesis de Maestría Presentado por".

[6] X. Alejandro, A. Pesantes, D. Jamil, and L. Baquerizo, "UNIVERSIDAD DE GUAYAQUIL AUTORES".

[7] A. Eraso Torres, "Caracterización de la web Colombia mediante la herramienta Wire," 2016, Accessed: Apr. 30, 2023. [Online]. Available: http://repositorio.unicesmag.edu.co:8080/xmlui/handle/123456789/82

[8] "Elaboración de herramienta web scraping para optimizar procesos del área de marketing en la línea 'dolor' de Abbott Colombia." https://repository.javeriana.edu.co/handle/10554/62634 (accessed May 04, 2023).

[9] "Vista de ScraCOVID-19: Plataforma informativa de contenido digital mediante Scraping y almacenamiento NoSQL." https://revistascientificas.cuc.edu.co/ingecuc/article/view/3280/3018,%C2%BB%20INGE%20CUC,%2017%2009%202020 (accessed May 04, 2023).

[10] L. F. Gómez Estrada and V. M. Orozco Puello, "Desarrollo de un prototipo de aplicación web que permita la extracción de las ofertas laborales de las principales plataformas que postulan empleos en la región caribe, usando la técnica web SCRAPING.," Jun. 2019,

Accessed: May 04, 2023. [Online]. Available: http://repositorio.unisinucartagena.edu.co:8080/xmlui/handle/123456789/94

[11] C. Stecben and M. Sterling, "Prototipo de recopilador web para información personal en diferentes redes sociales.," Jun. 2020, Accessed: Apr. 15, 2023. [Online]. Available: https://repository.uniminuto.edu/handle/10656/11557

[12] L. Santiago, D. Ramirez, D. Jady, and R. Blanco, "Quaesitor-Plataforma Tecnológica de búsqueda interactiva soportada en dispositivos móviles".

[13] R. F. y E. Monzón Laiza, I. N. Vargas Ulloa, and R. F. y E. Monzón Laiza, "Uso de técnicas de web scraping para el análisis de datos de jugadores profesionales del futbol peruano para el periodo 2021," *Universidad Privada Antenor Orrego*, 2023, Accessed: May 04, 2023. [Online]. Available: https://repositorio.upao.edu.pe/handle/20.500.12759/10390

[14] "Heartbeat of a Crypto-Economy:Transaction Information in a World withCentral Bank Digital Currencies Citation Permanent link Terms of Use Share Your Story", Accessed: May 04, 2023. [Online]. Available: https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364736

[15] M. D. P. Mezarina Cerna and M. Samame Vega, "Framework apoyado en web scraping y geolocalización para la identificación y selección de productos en supermercados," *Repositorio Institucional - UCV*, 2022, Accessed: May 04, 2023. [Online]. Available: https://repositorio.ucv.edu.pe/handle/20.500.12692/111049

[16] E. Y. Evaluación De, G. Valle Gutiérrez Director, and L. Pita-Romero Rodríguez Madrid, "APLICACIÓN DE TÉCNICAS DE WEB SCRAPING Y PROCESAMIENTO DEL LENGUAJE NATURAL PARA LA," 2021.

[17] A. Elamin, A. Mohamed, N. Ishag, and M. Saeed, "Web crawling in data engineering and analysis with web application".

[18] P. D. W. Fabián, "Software Para web scraping Desde Las Apis De Repositorios De Código," Jul. 2021, Accessed: Apr. 15, 2023. [Online]. Available: http://localhost/xmlui/handle/123456789/2605

[19] R. Tabarés, "HTML5 and the evolution of HTML; tracing the origins of digital platforms," *Technol Soc*, vol. 65, p. 101529, May 2021, doi: 10.1016/J.TECHSOC.2021.101529.

[20] S. Dutta and S. Roy, "Complex Network Visualisation Using JavaScript: A Review," *Lecture Notes in Networks and Systems*, vol. 431, pp. 45–53, 2022, doi: 10.1007/978-981-19-0901-6_5/COVER.

[21] "María del Carmen Santiago Díaz APLICACIONES DE LAS CIENCIAS COMPUTACIONALES DURANTE LA PANDEMIA COVID19".

[22] K. V. S. Ramos, "XML," *Nextia*, no. 6, pp. 26–29, doi: 10.1145/3132269.

[23] "Formato JSON (JavaScript Object Notation) - Documentación de IBM." https://www.ibm.com/docs/es/baw/20.x?topic=formats-javascript-object-notation-json-format (accessed Apr. 15, 2023).

[24] "SOAP - Documentación de IBM." https://www.ibm.com/docs/es/rsas/7.5.0?topic=standards-soap (accessed Apr. 15, 2023).

[25] X. M. Rea-Peñafiel, T. B. Mancero-Menoscal, D. C. Rosero-Rea, and D. E. Imbaquingo-Esparza, "Web Services REST: una revolución en la forma de acceso a datos".

[26] "WSDL (Web Services Description Language) - Documentación de IBM." https://www.ibm.com/docs/es/rsas/7.5.0?topic=standards-web-services-description-language-wsdl (accessed Apr. 15, 2023).

[27] M. A. Khder, "web scraping or Web Crawling: State of Art, Techniques, Approaches and Application," *Int. J. Advance Soft Compu. Appl*, vol. 13, no. 3, 2021, doi: 10.15849/IJASCA.211128.11.

[28] "Diferencias entre Scraping, Crawling y Parsing | OpenWebinars." https://openwebinars.net/blog/diferencias-entre-scraping-crawling-y-parsing/ (accessed May 06, 2023).

[29] "Beautiful Soup Documentation — Beautiful Soup 4.12.0 documentation." https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed Apr. 30, 2023).

[30] A. Felipe *et al.*, "APLICACIÓN DEL WEB SCRAPING Y EL ANÁLISIS AUTOMATIZADO A LOS MERCADOS DE DIVISAS Y DE ACCIONES," *Encuentros con semilleros*, vol. 3, no. 1 (3), pp. 8–17, Nov. 2022, doi: 10.15765/ES.V3I1.

[31] S. Paipilla *et al.*, "ScraCOVID-19: Plataforma informativa de contenido digital mediante Scraping y almacenamiento NoSQL," *INGE CUC*, vol. 16, no. 2, pp. 229–237, Oct. 2020, doi: 10.17981/INGECUC.16.2.2020.18.

[32] S. Álvarez Pesce, A. Bertoletti Laclau, and C. Pintos Conde, "Spycy," 2022, Accessed: Apr. 16, 2023. [Online]. Available: https://dspace.ort.edu.uy/handle/20.500.11968/4837

[33] C. Agesto and J. B. Lee, "FACULTAD DE INGENIERÍA Y ARQUITECTURA ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS".

[34] R. Mijangos-Espinosa, A. Martínez-Rebollar, H. Estrada-Esquivel, and Y. Hernández-Pérez, "Use of web scraping Techniques for Automatic Capturing of Databases Located in the Web".

[35] A. Felipe *et al.*, "APLICACIÓN DEL WEB SCRAPING Y EL ANÁLISIS AUTOMATIZADO A LOS MERCADOS DE DIVISAS Y DE ACCIONES," *Encuentros con semilleros*, vol. 3, no. 1 (3), pp. 8–17, Nov. 2022, doi: 10.15765/ES.V3I1.

[36] C. DE Ingeniería En Sistemas Y Computación, E. Jhaldyr Males Flores Diana Elizabeth Villacrés Bonilla, M. Basantes Valverde, and R. -Ecuador, "Desarrollo del sistema web de análisis de información Climatológica para la central meteorológica Yanarumi utilizando el software estadístico R.," Jun. 2021, Accessed: Apr. 16, 2023. [Online]. Available: http://dspace.unach.edu.ec/handle/51000/7761

[37] S. Macías Sánchez, "EL ANÁLISIS DE SENTIMIENTO APLICADO A LA EMPRESA", Accessed: Apr. 16, 2023. [Online]. Available: https://www.iic.uam.es/inteligencia/que-

[38] N. Villabona, D. J. Garcés, and R. J. Martelo, "Caracterización de contenido de sitios web turísticos mediante scraping y minería web para contribuir a la satisfacción de turista Characterization of content of tourism websites through web scraping and web mining to contribute to tourist satisfaction," vol. 41, no. 36, p. 2020, Accessed: Apr. 16, 2023. [Online]. Available: https://www.revistaespacios.com

[39] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola, "Web scraping technologies in an API world," *Brief Bioinform*, vol. 15, no. 5, pp. 788–797, Sep. 2014, doi: 10.1093/BIB/BBT026.

[40] V. Krotov, L. Johnson, and L. Silva, "Tutorial: Legality and Ethics of web scraping," *Faculty & Staff Research and Creative Activity*, vol. 47, pp. 539–563, Dec. 2020, doi: https://doi.org/10.17705/1CAIS.04724.

[41] "Scrapy at a glance — Scrapy 2.8.0 documentation." https://docs.scrapy.org/en/latest/intro/overview.html (accessed Apr. 16, 2023).

[42] "web scraping Guide | Octoparse Hello World." https://dataservice.octoparse.com/octoparse-hello-world (accessed Apr. 16, 2023).

[43] "ParseHub | Free web scraping - The most powerful web scraper." https://www.parsehub.com/ (accessed Apr. 25, 2023).

[44] "Find email addresses in seconds • Hunter (Email Hunter)." https://hunter.io/ (accessed Apr. 25, 2023).

[45] "Datafiniti: Instant Access to the Data You Need." https://www.datafiniti.co/?gad=1&gclid=CjwKCAjwxr2iBhBJEiwAdXECw3fHuOIfGOoFenjoraCdM5fquXBIlHnHK4-IZC3ohOlqBDgQxrxYTBoCrjEQAvD_BwE (accessed Apr. 30, 2023).

[46] "80legs – Customizable web scraping." https://80legs.com/ (accessed Apr. 30, 2023).

[47] "ScraperAPI - The Proxy API For web scraping." https://www.scraperapi.com/ (accessed Apr. 30, 2023).

[48] "Mozenda - Scalable Web Data Extraction Software & Services." https://www.mozenda.com/ (accessed Apr. 30, 2023).

[49] "Diffbot | Knowledge Graph, AI Web Data Extraction and Crawling." https://www.diffbot.com/ (accessed Apr. 30, 2023).

[50] "Dexi.io - Digital Commerce Intelligence, Retail, Brands & E-Commerce." https://www.dexi.io/ (accessed Apr. 30, 2023).

[51] "Java | Oracle." https://www.java.com/es/ (accessed Apr. 30, 2023).

[52] "Welcome to Python.org." https://www.python.org/ (accessed Apr. 30, 2023).

[53] "idUS - Clasificación de imágenes usando redes neuronales convolucionales en Python." https://idus.us.es/handle/11441/89506 (accessed Apr. 30, 2023).

[54] "Selenium." https://www.selenium.dev/ (accessed Apr. 30, 2023).

[55] "jsoup: Java HTML parser, built for HTML editing, cleaning, scraping, and XSS safety." https://jsoup.org/ (accessed Apr. 30, 2023).

[56] "requests · PyPI." https://pypi.org/project/requests/ (accessed Apr. 30, 2023).

[57] "beautifulsoup4 · PyPI." https://pypi.org/project/beautifulsoup4/ (accessed Apr. 30, 2023).

[58] "The web framework for perfectionists with deadlines | Django." https://www.djangoproject.com/ (accessed Apr. 30, 2023).

[59] J. Parinango, E. Alfredo, M. Bravo Ruiz, and J. Arturo, "Análisis de los sistemas de gestión de base de datos relacionales con marcos de trabajo para procesamiento de datos masivos," *Repositorio Institucional - USS*, 2022, Accessed: May 03, 2023. [Online]. Available: http://repositorio.uss.edu.pe//handle/20.500.12802/10234

[60] L. Marrero *et al.*, "Un estudio de procesos de diseño de bases de datos NoSQL," 2022, Accessed: May 03, 2023. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/149452

[61] T. Gómez, E. Tutor, and M. Herranz, "Introducción a las bases de datos NoSQL. Sistemas de bases de datos orientados a grafos," Sep. 2022, Accessed: May 03, 2023. [Online]. Available: https://riunet.upv.es/handle/10251/186175

[62] B. Campoverde Vega, V. Andree, M. O. Moreno, and R. Liliana, "Análisis comparativo de rendimiento en gestores de bases de datos relacionales y no relacionales," *Repositorio Institucional - USS*, 2022, Accessed: May 03, 2023. [Online]. Available: http://repositorio.uss.edu.pe//handle/20.500.12802/9211

[63] P. C. Humphreys *et al.*, "A data-driven approach for learning to control computers." PMLR, pp. 9466–9482, Jun. 28, 2022. Accessed: May 06, 2023. [Online]. Available: https://proceedings.mlr.press/v162/humphreys22a.html

[64] "DIAL: download document." https://dial.uclouvain.be/downloader/downloader.php?pid=boreal:168996&datastream=PDF_01 (accessed Apr. 27, 2023).

[65] J. C. Alvarado-Pérez, D. H. Peluffo-Ordóñez, and R. Therón, "Bridging the gap between human knowledge and machine learning," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 4, no. 1, pp. 54–64, Oct. 2015, doi: 10.14201/ADCAIJ2015415464.

[66] D. H. Peluffo-OrdÓÑez, J. C. Alvarado-PÉrez, and A. E. Castro-Ospina, "On the spectral clustering for dynamic data," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9108, pp. 148–155, 2015, doi: 10.1007/978-3-319-18833-1_16/COVER.

[67] "Vista de Creatividad e innovación: competencias genéricas o transversales en la formación profesional." https://revistavirtual.ucn.edu.co/index.php/RevistaUCN/article/view/620/1155 (accessed Apr. 27, 2023).

[68] R. I. Hernández-Arteaga, J. C. Alvarado-Pérez, and J. A. Luna, "Responsabilidad social en la relación universidad-empresa-Estado," *Educación y Educadores*, vol. 18, no. 1, pp. 95–

110, Jun. 2015, Accessed: Apr. 27, 2023. [Online]. Available: https://educacionyeducadores.unisabana.edu.co/index.php/eye/article/view/4424

Desarrollo de una herramienta de adquisición automática de datos de fuentes externas, en el sistema de gestión de información de la Vicerrectoría de Investigaciones de la Universidad CESMAG, mediante crawlers y web scraping.

## ANEXOS

Anexo 1. Encuesta de Validación de la herramienta de adquisición automática de datos.

EVALUACIÓN DEL MÓDULO DE ADQUISICIÓN AUTOMÁTICA DE DATOS (google.com)

Anexo 2. Resultados de la encuesta.

EVALUACIOÓN SOBRE EL MÓDULO DE ADQUISICIÓN AUTOMÁTICA DE DATOS - Google Forms

Anexo 3. Documento de Aceptación por los Líderes de Investigación.

https://docs.google.com/document/d/1q3feFjHqNY_0JeSEI-t5hI9PcSW-sN_9/edit?usp=sharing&ouid=102418213515618188350&rtpof=true&sd=true

Anexo 4. Manual de usuario de la herramienta.

https://drive.google.com/drive/folders/1SZ5NiL1HsoJcqX8s4FQSR3gSs80oA7XE?usp=sharing

| UNIVERSIDAD CESMAG NIT. 800.109.387-7 | CARTA DE ENTREGA TRABAJO DE GRADO O TRABAJO DE APLICACIÓN – ASESOR(A) | CÓDIGO: AAC-BL-FR-032 |
|---|---|---|
| | | VERSIÓN: 1 |
| | | FECHA: 09/JUN/2022 |

San Juan de Pasto, 3 Seltembre del 2024

Biblioteca
**REMIGIO FIORE FORTEZZA OFM. CAP.**
Universidad CESMAG
Pasto

Saludo de paz y bien.

Por medio de la presente se hace entrega del Trabajo de Grado / Trabajo de Aplicación denominado Desarrollo de una herramienta de adquisición automatica de datos de fuentes externas... presentado por el (los) autor(es) Daniel Gualteros, Sebastian Trejo, David Finlay y _____ del Programa Académico Ing. de Sistemas al correo electrónico biblioteca.trabajosdegrado@unicesmag.edu.co. Manifiesto como asesor(a), que su contenido, resumen, anexos y formato PDF cumple con las especificaciones de calidad, guía de presentación de Trabajos de Grado o de Aplicación, establecidos por la Universidad CESMAG, por lo tanto, se solicita el paz y salvo respectivo.

Atentamente,

---

**NOMBRE Y APELLIDOS DEL ASESOR(A)**
Numero de documento: 1085257779
Programa académico: Ing Sistemas
Teléfono de contacto: 3172537641
Correo electrónico: hamola@unicesmag.edu.co

| UNIVERSIDAD CESMAG | AUTORIZACIÓN PARA PUBLICACIÓN DE TRABAJOS DE GRADO O TRABAJOS DE APLICACIÓN EN REPOSITORIO INSTITUCIONAL | CÓDIGO: AAC-BL-FR-031 |
|---|---|---|
| | | VERSIÓN: 1 |
| | | FECHA: 09/JUN/2022 |

| INFORMACIÓN DEL (LOS) AUTOR(ES) | |
|---|---|
| **Nombres y apellidos del autor:** Andres Sebastian Trejo O. | **Documento de identidad:** 1233194140 |
| **Correo electrónico:** andrestrejo7777@gmail.com | **Número de contacto:** 3188222747 |
| **Nombres y apellidos del autor:** Daniel Clemente Gualteros Sinisajoa | **Documento de identidad:** 1114389682 |
| **Correo electrónico:** dgualteros637@gmail.com | **Número de contacto:** 3219567349. |
| **Nombres y apellidos del autor:** David Esteban Finlay Estrella | **Documento de identidad:** 1085328875 |
| **Correo electrónico:** definlay8875@unicesmag.edu.co | **Número de contacto:** 3163243434 |
| **Nombres y apellidos del autor:** | **Documento de identidad:** |
| **Correo electrónico:** | **Número de contacto:** |
| **Nombres y apellidos del asesor:** Hector Andres Mora | **Documento de identidad:** 1085257779 |
| **Correo electrónico:** hamora@unicesmag.edu.co | **Número de contacto:** 3172537641 |
| **Título del trabajo de grado:** Desarrollo de una herramienta de adquisición automatica de datos de fuentes externas para el sistema de gestión de información de la Vicerrectoria de investigación de cesmag mediante crawlers y web scraping. | |
| **Facultad y Programa Académico:** Ing. Sistemas. | |

En mi (nuestra) calidad de autor(es) y/o titular (es) del derecho de autor del Trabajo de Grado o de Aplicación señalado en el encabezado, confiero (conferimos) a la Universidad CESMAG una licencia no exclusiva, limitada y gratuita, para la inclusión del trabajo de grado en el repositorio institucional. Por consiguiente, el alcance de la licencia que se otorga a través del presente documento, abarca las siguientes características:

a) La autorización se otorga desde la fecha de suscripción del presente documento y durante todo el término en el que el (los) firmante(s) del presente documento conserve (mos) la titularidad de los derechos patrimoniales de autor. En el evento en el que deje (mos) de tener la titularidad de los derechos patrimoniales sobre el Trabajo de Grado o de Aplicación, me (nos) comprometo (comprometemos) a informar de manera inmediata sobre dicha situación a la Universidad CESMAG. Por consiguiente, hasta que no exista comunicación escrita de mi(nuestra) parte informando sobre dicha situación, la Universidad CESMAG se encontrará debidamente habilitada para continuar con la publicación del Trabajo de Grado o de Aplicación dentro del repositorio institucional. Conozco(conocemos) que esta autorización podrá revocarse en cualquier momento, siempre y cuando se eleve la solicitud por escrito para dicho fin ante la Universidad CESMAG. En estos eventos, la Universidad CESMAG cuenta con el plazo de un mes después de recibida la petición, para desmarcar la visualización del Trabajo de Grado o de Aplicación del repositorio institucional.

b) Se autoriza a la Universidad CESMAG para publicar el Trabajo de Grado o de Aplicación en formato digital y teniendo en cuenta que uno de los medios de publicación del repositorio institucional es el internet, acepto(amos) que el Trabajo de Grado o de Aplicación circulará con un alcance mundial.
c) Acepto (aceptamos) que la autorización que se otorga a través del presente documento se realiza a título gratuito, por lo tanto, renuncio(amos) a recibir emolumento alguno por la publicación, distribución, comunicación pública y/o cualquier otro uso que se haga en los términos de la presente autorización y de la licencia o programa a través del cual sea publicado el Trabajo de grado o de Aplicación.
d) Manifiesto (manifestamos) que el Trabajo de Grado o de Aplicación es original realizado sin violar o usurpar derechos de autor de terceros y que ostento(amos) los derechos patrimoniales de autor sobre la misma. Por consiguiente, asumo(asumimos) toda la responsabilidad sobre su contenido ante la Universidad CESMAG y frente a terceros, manteniéndose indemne de cualquier reclamación que surja en virtud de la misma. En todo caso, la Universidad CESMAG se compromete a indicar siempre la autoría del escrito incluyendo nombre de(los) autor(es) y la fecha de publicación.
e) Autorizo(autorizamos) a la Universidad CESMAG para incluir el Trabajo de Grado o de Aplicación en los índices y buscadores que se estimen necesarios para promover su difusión. Así mismo autorizo (autorizamos) a la Universidad CESMAG para que pueda convertir el documento a cualquier medio o formato para propósitos de preservación digital.

**NOTA**: En los eventos en los que el trabajo de grado o de aplicación haya sido trabajado con el apoyo o patrocinio de una agencia, organización o cualquier otra entidad diferente a la Universidad CESMAG. Como autor(es) garantizo(amos) que he(hemos) cumplido con los derechos y obligaciones asumidos con dicha entidad y como consecuencia de ello dejo(dejamos) constancia que la autorización que se concede a través del presente escrito no interfiere ni transgrede derechos de terceros.

Como consecuencia de lo anterior, autorizo(autorizamos) la publicación, difusión, consulta y uso del Trabajo de Grado o de Aplicación por parte de la Universidad CESMAG y sus usuarios así:

- Permito(permitimos) que mi(nuestro) Trabajo de Grado o de Aplicación haga parte del catálogo de colección del repositorio digital de la Universidad CESMAG por lo tanto, su contenido será de acceso abierto donde podrá ser consultado, descargado y compartido con otras personas, siempre que se reconozca su autoría o reconocimiento con fines no comerciales.

En señal de conformidad, se suscribe este documento en San Juan de Pasto a los____días del mes de ____del año _____

| | |
|---|---|
| Firma del autor<br>Nombre del autor: Andrey Sebastian Puyo G. | Firma del autor<br>Nombre del autor: Daniel Bonilla Ortiz |
| Firma del autor<br>Nombre del autor: David Esteban Finlay E. | Firma del autor<br>Nombre del autor: |
| Firma del asesor<br>Nombre del asesor: Hector Andres Mora. | |