



**Modelo computacional para la optimización de los procesos de captura y análisis
desarrollados en la Vigilancia e Inteligencia Estratégica en la Oficina de Transferencia de
Resultados de Investigación – OTRI – de la UdeA**

Alexander Gómez Montoya

Proyecto presentado para optar al título de Bibliotecólogo

Asesores

Juan Fernando Pérez-Pérez, Doctor (PhD) en Ingeniería.

Jaidier Ochoa Gutiérrez, Magíster (MSc) en Gestión de Ciencia, Tecnología e Innovación

Universidad de Antioquia
Escuela Interamericana de Bibliotecología
Bibliotecología
Medellín, Antioquia, Colombia
2024

Cita

(Gómez-Montoya, 2024)

Referencia

Estilo APA 7 (2020)

Gómez-Montoya, Alexander. (2024). *Modelo computacional para la optimización de los procesos de captura y análisis desarrollados en la Vigilancia e Inteligencia Estratégica en la Oficina de Transferencia de Resultados de Investigación – OTRI – de la UdeA* [Trabajo de grado profesional]. Universidad de Antioquia, Medellín, Colombia.



CRAI Escuela Interamericana de Bibliotecología

Repositorio Institucional: <http://bibliotecadigital.udea.edu.co>

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Directora: Margarita María Gaviria Velásquez.

Jefe de departamento: Yulieth Taborda Ramírez

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Dedicatoria

Dedico este trabajo de grado en primer lugar a Dios y a María Santísima, pues su guía y sostenimiento en el desarrollo de mi formación profesional han sido extremadamente valiosos, fortaleza y valentía fueron siempre sus mayores regalos.

A mi madre, Lina María Gómez, a mi tía Mary Gómez y mi prima-hermana María Paulina Castrillón, pilares fundamentales de todo lo que soy y aquello que quiero alcanzar, dedico también este Trabajo a la memoria de dos seres que amo profundamente y me acompañan desde la eternidad, mi padre Alexander y mi abuela Dora Nelly.

A mi gran amigo del alma Cristian Salazar; a mis compañeros de la Escuela, tanto de clase como de oficina, aquellos que se dieron el trabajo de conocerme y junto con ellos a varios de los profes que de manera incansable aportaron tanto dentro como fuera de las aulas a mi desarrollo profesional, ahí perdonan la insistencia y constante consulta, pues la exigencia que han impreso en mí es lo que ha hecho que me arriesgue y dedique mis esfuerzos al desarrollo de este trabajo.

Agradecimientos

Se me quedan cortas las palabras para agradecer a varias personas que hicieron parte de este proceso de manera directa e indirecta, son valiosas sus enseñanzas en conocimiento y apoyo para mí.

A mis profesores y asesores Juan Fernando Pérez y Jaider Ochoa quienes con su acompañamiento, conocimiento, asistencia y disponibilidad lograron despertar en mí el gusto por la programación y el enfoque hacia este núcleo del campo disciplinar, mi eterna gratitud a estos dos grandes mentores.

A mis jefes Fabián Baena, Leidy Ruiz, Yulieth Taborda y Marta Pulgarín quienes siempre se vieron comprensivos y disponibles a lo que necesitara en tiempo y espacio para el desarrollo de estas labores, agradezco también el interés que han mostrado por mi trabajo de grado.

A una persona especial, a quien yo le llamo “Sol”, que llegó a iluminar y dar vida a mi vida, gracias infinitas por ayudarme a transformar mi forma de ver el mundo.

Finalmente, a la Escuela, la casa donde por 6 años me formé como profesional y de quien me despidió con gratitud constante, por siempre profesional del Alma Mater.

Tabla de contenido

Resumen	8
Abstract	9
Introducción	10
2 Justificación.....	11
3 Pregunta.....	13
4 Objetivos	13
4.1 Objetivo general	13
4.2 Objetivos específicos.....	13
5 Marco Contextual.....	14
6 Marco Teórico-Conceptual	15
Vigilancia e Inteligencia Estratégica	15
Programación	17
Web Scraping.....	20
7 Metodología	22
Identificación de Variables	22
Estructuración Bibliográfica	23
Modelación del programa	23
Implementación del prototipado de captura y análisis.....	24
8 Vigilancia e Inteligencia Estratégica.....	27
9 Modelo Computacional	33
9.1 Elementos iniciales	33
9.2 Modelación del código.....	34
9.2.1 Librerías	34

9.2.2 Solicitud HTML	35
9.2.3 Inspección de Página	35
9.2.4 Extracción de secciones específicas	36
9.2.5 Almacenamiento de datos	38
9.3 Sección de publicaciones	38
10 Implementación del modelo	41
11 Preprocesamiento del Texto	44
12 Entrenamiento de chat para la consulta de información	49
13 Conclusiones	52
14 Recomendaciones	53
15 Referencias	54
Anexos.....	60

Lista de figuras

ILUSTRACIÓN 1: PROCESO DE VEIE.....	12
ILUSTRACIÓN 2: PILARES CONCEPTUALES EN LA PROPUESTA DE DEFINICIÓN DE INTELIGENCIA ESTRATÉGICA.....	17
ILUSTRACIÓN 3: METODOLOGÍA DE ACTIVIDADES	25
ILUSTRACIÓN 4: SOLICITUD HTTPS.....	35
ILUSTRACIÓN 5: INSPECCIÓN DE IRENA.....	35
ILUSTRACIÓN 6: EXTRACCIÓN DE TÍTULOS FUENTE: ELABORACIÓN PROPIA.....	36
ILUSTRACIÓN 7: CONCATENACIÓN DE ELEMENTOS	37
ILUSTRACIÓN 8: VISUALIZACIÓN DE TÍTULOS	37
ILUSTRACIÓN 9: VISUALIZACIÓN DE ENLACES.	38
ILUSTRACIÓN 10: DATOS DE PUBLICACIÓN.	39
ILUSTRACIÓN 11: DUPLICADO DE DOCUMENTOS.....	39
ILUSTRACIÓN 12: MAPA DEL PROCESO DE MODELACIÓN DEL CÓDIGO.....	40
ILUSTRACIÓN 13: AGRUPACIÓN DE LOS DF DE NOTICIAS Y PUBLICACIONES	41
ILUSTRACIÓN 14: ELIMINACIÓN DE REGISTROS NO CORRESPONDIENTES	42
ILUSTRACIÓN 15: TRANSFORMACIÓN DE DATOS NAN.....	42
ILUSTRACIÓN 16: ELIMINACIÓN DE REGISTROS DUPLICADOS	43
ILUSTRACIÓN 17: CÓDIGO DE LIMPIEZA DEL DS	43
ILUSTRACIÓN 18: ELIMINACIÓN DE ELEMENTOS NO NECESARIOS	44
ILUSTRACIÓN 19: STOPWORDS	44
ILUSTRACIÓN 20: PREPROCESAMIENTO DEL TEXTO.....	45
ILUSTRACIÓN 21: EMBEDDINGS Y TOKENIZACIÓN.....	46
ILUSTRACIÓN 22: CLUSTERIZACIÓN	47
ILUSTRACIÓN 23: NUBE DE PALABRAS	48
ILUSTRACIÓN 24: CANTIDAD DE REGISTROS POR CLÚSTER	49
ILUSTRACIÓN 25: IRENACHAT.....	51

Siglas, acrónimos y abreviaturas

AI	Artificial Intelligence
API	Application programming interfaces
DF	Dataframe
DS	Dataset
EIB	Escuela Interamericana de Bibliotecología
FCV	Factores críticos de vigilancia
FN	Ficha de necesidades para la vigilancia
FP	Ficha de planeación para la vigilancia
HTML	Hypertext Markup Language
ML	Machine Learning
OTRI	Oficina de Transferencia de Resultados de Investigación
TIC	Tecnologías de la Información y de la Comunicación
UdeA	Universidad de Antioquia
VeIE	Vigilancia e Inteligencia Estratégica

Resumen

El desarrollo de este trabajo propuso un modelo computacional que permita la optimización de los procesos de captura y análisis de datos en el desarrollo de ejercicios de Vigilancia e Inteligencia estratégica que se realiza en la Oficina de Transferencia de Resultados de Investigación de la División de Innovación de la UdeA, permitiendo optimizar el tiempo y esfuerzo que constantemente es aplicado en dichas operaciones en las bases de datos para compilar información que luego será clasificada, según su pertinencia y relevancia, a ciertos ejercicios que se desarrollan al interior de la oficina. Esto es, con la implementación de este modelo se espera mayor precisión y agilidad en la captura de información según la fuente de los datos, posibilitando que el análisis cuente con herramientas suficientes para ofrecer las mejores alternativas en cuestión de inteligencia para las organizaciones. Gracias al modelo planteado se pudo observar que los desarrollos tecnológicos también permiten la optimización de las labores que los profesionales de la información desarrollan en su día a día, labores que si pueden ser optimizadas tomando como herramientas las nuevas invenciones que han nacido y que abre la puerta a nuevas posibilidades de trabajo multidisciplinar tanto en los sectores públicos como privados para bibliotecólogos y archivistas.

Palabras clave: Modelo computacional, Bases de datos, Vigilancia e Inteligencia Estratégica, Captura de datos, Oficina de Transferencia de Resultados de Investigación, Inteligencia Artificial.

Abstract

The development of this research proposed a computational model designed to optimize data capture and analysis processes in the strategic surveillance and intelligence exercises conducted by the Oficina de Transferencia de Resultados de Investigación in the División de Innovación at UdeA. This model aims to enhance the time and effort regularly invested in database operations to compile information, which is later classified based on its pertinence and relevance to specific exercises within the office. In essence, the implementation of this model is expected to bring greater precision and agility in data capture according to the data source, ensuring that the analysis has the necessary tools to offer the best intelligence alternatives for organizations. Thanks to the proposed model, it was observed that technological developments also enable the optimization of the tasks performed daily by information professionals. These tasks can be improved through the use of new inventions, opening the door to new multidisciplinary work opportunities in both public and private sectors for librarians and archivists.

Keywords: Databases, Computer software, strategic and intelligence monitoring, Oficina de Transferencia de Resultados de Investigación, Artificial Intelligence.

Introducción

La inteligencia competitiva que se da en las diferentes organizaciones exige que constantemente estas se encuentren evaluando aquello que ofrecen a sus usuarios, a sus públicos de valor; ante un acelerado crecimiento de la sociedad cada vez se afrontan más retos que busquen dar soluciones a aquello que va surgiendo. Las tecnologías son elementos claves en esos procesos de desarrollo y progreso necesarios para cualquier parte de la sociedad.

La Vigilancia y la Inteligencia estratégica (VeIE) son un factor clave en los procesos organizacionales, pues permiten tener conocimiento y claridad sobre aquello que está moviendo el mundo día a día, estrategias, tendencias, ideas, proyectos, innovaciones y tecnologías que van ganando terreno gracias a la innumerable producción científica que se da en todo el mundo. Los niveles de crecimiento son bastante grandes, por lo que tener un panorama de lo existente cada día se hace algo casi impensable.

Con el gran crecimiento de la sociedad, las organizaciones cada vez más buscan acercarse a servicios que le pongan a la vanguardia de las necesidades del mercado y la vigilancia es uno de los más apetecidos puesto que, hacer un sondeo documental de aquello que puede ser de interés y que ya hoy existe y va tomando diferentes direcciones permite a una organización reflexionar sobre cuál será su paso a seguir. Este proceso de VeIE es algo lento y procesual puesto que la captura de datos no siempre es ágil y no siempre se da en los mejores términos.

Desde esa mirada es que, buscar una forma de optimizar ese proceso de captura ayudará de gran manera a los profesionales del campo a centrarse más que en la búsqueda en el análisis y presentación de resultados, eso es lo que busca este modelo, optimizar esa parte del proceso que pudiendo tardarse horas o hasta días, se reduzca a una considerable fracción de tiempo y que lo haga de manera segura y confiable de que lo que se va a rescatar será de valor.

Los profesionales de la información deben de estar en la capacidad de realizar este tipo de procesos de búsqueda y captura de información para el análisis posterior, esta labor se realiza no solo en bibliotecas sino también en el sector privado y es allí a donde se dirige esta investigación. Como producto de grado, se quiere desarrollar no solo un modelo que optimice este proceso en específico para la OTRI, sino que también que se evidencie a la Escuela Interamericana de Bibliotecología – EIB - que las tecnologías deben ser parte fundamental de la innovación curricular que se está desarrollando y que como bibliotecólogo no solo se está para trabajar en bibliotecas sino generar la necesidad de tener un profesional de la información en cualquier empresa.

2 Justificación

La Universidad de Antioquia, desde la Vicerrectoría de Investigación, promueve la búsqueda e implementación de espacios donde se puedan poner en marcha propuestas y desarrollos intelectuales, mecánicos, médicos y analíticos que usualmente nacen del pensamiento de docentes y estudiantes con capacidades de transformación, innovación y perspectivas de mejora en miras a la diversa complejidad de temáticas y sucesos que permean la realidad social e institucional que les rodea y que; de manera independiente o en unión a los diferentes grupos de investigación se generan herramientas que apuntan a la labor propia de la Oficina de Transferencia de Resultados de Investigación (OTRI)

La División de Innovación es una de las dependencias que promueve la articulación entre las investigaciones de la universidad y las necesidades de las empresas y el estado que se proyectan llegar a la comunidad. Desde esta misma División, se acompaña el proceso de transferencia de conocimiento en el que la OTRI juega un papel fundamental, pues para ella se establece que *“El enfoque de la OTRI son las creaciones y resultados de investigación con potencial de aplicación, que resuelven problemas y necesidades de la sociedad o del tejido empresarial; así como los retos de la industria que puedan solucionarse en colaboración con la Universidad desde la investigación aplicada.”* (Innovación UdeA, s.f) y su principal objetivo se basa en la identificación, valoración, protección y comercialización de aquellos resultados de las investigaciones de la Universidad y que son articulables con el sector productivo. (Pérez, J. 2024)

Así mismo, la labor de la OTRI permite identificar que, la prestación del servicio de vigilancia e inteligencia estratégica, es procedimental, mecánico y exige una alta inversión de tiempo y esfuerzo para encontrar resultados óptimos, algo que, con el desarrollo de tecnologías y la creación de nuevos conocimientos en la dinámica actual, exige que sea mucho más ágil y preciso, cuyos resultados sean contundentes y se concentren en dar mayor valor a los objetivos de las organizaciones que buscan este servicio.

Y es que, en el ámbito de la VeIE, las organizaciones enfrentan el desafío de adaptarse rápidamente a los cambios y tendencias emergentes en un mundo cada vez más tecnológico y conectado. La Universidad EAFIT, en una investigación realizada al foro económico mundial, muestra que las grandes tendencias que se vislumbran para los próximos años en un corto plazo, van enfocadas a los grandes desarrollos tecnológicos en donde las empresas migran a la virtualidad,

los contextos de trabajo son compartidos entre robots y personas, la IA juega un nuevo rol al igual que las decisiones basadas en el Big data (EAFIT, s.f.)

Sin embargo, el proceso de VeIE actualmente es mecánico y procesual, involucrando una serie de actividades que requieren una inversión significativa de esfuerzo y tiempo para obtener resultados óptimos; el manual de Vigilancia Tecnológica de la Universidad EAN - Escuela de Administración de Negocios presenta la estructuración del proceso mediante 6 pasos manuales y procesuales, que requieren determinado tiempo y esfuerzo.

Ilustración 1: Proceso de VeIE



Fuente: Manual para la realización de Vigilancia Tecnológicas, Universidad EAN

Lo anterior resalta la falta de un modelo, proceso o método que brinde agilidad y precisión en el desarrollo de la Vigilancia e Inteligencia estratégica. Además, tal como se planteó, la búsqueda y análisis manual de información, resulta ser un proceso ineficiente y demanda una gran cantidad de tiempo del personal responsable, más aún si se sigue paso a paso cada una de las etapas. Esto sugiere una necesidad de innovación y mejora en las prácticas de VeIE para facilitar a las organizaciones una toma de decisiones más rápida, basada en datos precisos y relevantes. Es por esto que con el proyecto se consideró la optimización en algunos de los pasos mencionados en la ilustración 1, hecho que ayudará a replantear alternativas que finalmente den cuenta de la cultura organizacional y de la rigurosidad del proceso que se desarrolla.

3 Pregunta

¿De qué manera la eficiencia de los procesos de captura y análisis de información, desde la Vigilancia e inteligencia Estratégica, puede ser potenciada buscando la optimización del tiempo y esfuerzo que la Oficina de Transferencia de Resultados de Investigación de la UdeA invierte en ellos?

4 Objetivos

4.1 Objetivo general

Proponer un método que mejore la eficiencia de la captura y análisis de información en el proceso de vigilancia e inteligencia estratégica, que permita la optimización de tiempo y esfuerzo requeridos por la Oficina de Transferencia de Resultados de Investigación de la Universidad de Antioquia.

4.2 Objetivos específicos

- Identificar las necesidades de información que se presentan en la OTRI y los procesos que se desarrollan con las diferentes solicitudes, que permita el establecimiento de los insumos básicos para la prestación del servicio de VeIE
- Diseñar un método para la captura de información, que disminuya los tiempos requeridos para la extracción de datos de fuentes externas empleados en los ejercicios de vigilancia e inteligencia estratégica
- Validar el método de captura de datos propuesto a través de un ejercicio práctico de vigilancia e inteligencia estratégica al interior de la Oficina de Transferencia de Resultados de Investigación de la Universidad de Antioquia

5 Marco Contextual

La Universidad de Antioquia, Institución de Educación Superior con más de doscientos años de historia, marcó el paso del tiempo adaptándose a los cambios sociales y formando año tras año profesionales que respondan a las demandas del mundo en el que se mueve. Compuesta por tres ejes misionales, la universidad es considerada fuente de conocimiento y alma del Departamento de Antioquia.

Los ejes misionales que componen el trabajo de la Institución son Docencia, Investigación y Extensión. La docencia enfocada directamente en los procesos formativos que se desarrollan en las Unidades Académicas; la Investigación que básicamente se centra en convertir a la Universidad en un eje central de investigación donde constantemente se realicen hallazgos que posibiliten el desarrollo social y económico de la región y del país; finalmente, la Extensión es aquella que se encarga de guiar estos procesos de relacionamiento de la Universidad como centro de pensamiento y fuente de desarrollo con la diversidad de instancias que componen la sociedad. Es específicamente en este último eje, que la misma Institución configuró como una Vicerrectoría, donde se centra el desarrollo de la Investigación, puesto que, tal como lo presenta la misma UdeA *“es posible conectar la producción de conocimiento con las verdaderas necesidades de la sociedad”* (ViceExtensión UdeA, s.f).

Dentro de esta Vicerrectoría se encuentran varias unidades cuya labor es proyectar desde diferentes vertientes, la participación de la Universidad en el desarrollo del ecosistema en el que se encuentra inmerso; cada unidad recibe el nombre de división, uno de los temas proyectados las Divisiones es el de la Innovación. La División de Innovación busca proyectar los procesos de investigación y creación de conocimiento a los diferentes actores de la sociedad como lo son las empresas y el estado, algo que la misma división establece como *“contribuir con el desarrollo socioeconómico de los territorios a partir de soluciones basadas en conocimiento”* (Innovación UdeA, s.f).

Al interior de la División de Innovación, existen asuntos de gestión que proyectan la actividad del conocimiento hacia direcciones de implementación, transferencia y creación de nuevas ideas que lo transformen en realidades. La gestión de la Transferencia de Conocimiento,

uno de los asuntos anteriormente comentados, busca transformar el conocimiento originado en la Universidad de manera que, articulada con diferentes actores como los investigadores, empresas, inversores de los proyectos junto con otros elementos del ecosistema de innovación se vean reflejados en proyectos de Investigación y Desarrollo donde se promueva la gestión, se valoricen los resultados y se proteja la propiedad intelectual; esa es la labor de la OTRI y de donde se toma de referencia en esta definición el principal objetivo de la oficina el cual es denominado como enfoque y va dirigido a *“las Creaciones y resultados de investigación con potencial de aplicación, que resuelven problemas y necesidades de la sociedad o del tejido empresarial”* (OTRI UdeA. S.f).

Es en esta Oficina donde se desarrolla esta investigación, pues uno de los servicios que ofrece es el de VeIE el cual se desarrolla mediante procesos que están previamente identificados y al cual se dirige el enfoque de este trabajo.

6 Marco Teórico-Conceptual

Para que exista una mayor claridad y dimensión de lo que se desea desarrollar en este trabajo, se hace pertinente entonces puntualizar algunos elementos temáticos que rigen de manera transversal este proceso.

Vigilancia e Inteligencia Estratégica

La vigilancia se considera como un proceso meticuloso y detallado en el que se busca identificar elementos que permitan establecer puntos claves para el crecimiento y mejora de cualquier tipo de organización. Nacido en el ámbito tecnológico pero que ha tenido aceptación y vínculo con diferentes áreas del saber y de la sociedad; al hablar de vigilancia no solo se hace referencia a un proceso mecánico de observación y registro, sino que, podría considerarse como un bucle, un ciclo en el que, de los resultados obtenidos en una investigación previa, da origen a nuevos objetivos o propósitos a desarrollar (Gómez, H. 2023).

Autores como Manco e Hincapié Pérez lo definen como una capacidad, otros como Jaramillo Bedoya lo identifican como una herramienta; si bien ambas son coherentes con la esencia de la Vigilancia, podría verse apropiado identificarla como un mecanismo que, tomando elementos de diversas fuentes, arroja resultados de interés y viabilidad para la estructura organizacional a la que sirve. No puede verse como algo aislado, ni mucho menos desconocer su origen enraizado en la perspectiva tecnológica de dotar de conocimiento nuevo y efectivo a quienes buscaban elementos para redirigir o fortalecer los procesos de su organización (Manco, F. et Hincapié, J. 2022.). Al hablar de mecanismo, se hace referencia a ese trabajo conjunto de actores, elementos y engranajes que, de manera coordinada y equilibrada alcanzan objetivos, en este caso, puntos de vista y elementos de valor, que son relevantes a juicio de quien acude a ellos

Al mencionar que la vigilancia es un proceso, exige concretar cuáles son sus elementos o pasos para establecer un desarrollo ideal de este mecanismo y también se hace propicio mencionar que una de sus características es el de ser sistemático, esto debido a que en él brilla el orden procedimental en el que se ejecuta. La búsqueda, observación, captura, filtración, análisis, gestión y comunicación de los resultados son los componentes por los que se efectúa el proceso de vigilancia y cuyo principal objetivo radica en generar nuevo conocimiento que apoye la toma de decisiones y pro de mejorar la competitividad a la que se apuesta desde la organización.

Por otra parte, Comai (2011) de manera concreta establece que el sentido principal, el valor en sí radica en tener buenas fuentes de información desde donde se permita realizar una correcta interpretación de estas y que, de manera concreta lleve a una aplicación precisa y adecuada; los profesionales que se dedican a esto deben de estar preparados para el uso del *Data Mining*, aplicaciones de análisis y depuración de la información, tener consciencia de los datos y de la posible aplicabilidad en el análisis considerando en si el enfoque que busca, si sería cuantitativo o cualitativo y además, considerar dinámicas atractivas que permitan visualizar los resultados generando no solo un impacto visual sino una invitación a la aplicabilidad. No se trata de investigar un tema en concreto, es necesario incluir aquí una mirada prospectiva en la que, planteando diversas posibilidades basadas en el análisis de los datos, la información sea contundente al momento de implementar, considerando pros y contras pero que desemboquen en una mirada de innovación.

Observar el futuro requiere también establecer unos planes que lleven a la consecución de los objetivos propuestos, por lo que, aquello que se planea debe ser medible y alcanzable, si no tiene tales características carece de sentido promoverlo; en torno a ello gira la gestión de la VeIE que agrupa su fortaleza en cuatro pilares y que han sido mencionados en párrafos previos, esta imagen refleja el contexto de lo establecido:

Ilustración 2: Pilares conceptuales en la propuesta de definición de inteligencia estratégica



Fuente: Aguirre.

En vista de lo anterior, la transferencia busca ofrecer un contexto y un modelo de aplicación de los elementos en los que sobresalga el conocimiento como fuente de nuevas herramientas y procesos adaptados de los mercados visualizados a cada momento. La transmisión se enfoca en la información transformada en conocimiento sumado a las habilidades y aptitudes de quienes se mueven en el área para propiciar el desarrollo social y empresarial mediante las organizaciones involucradas en esto.

Programación

Desde el campo de la programación enmarcado en las grandes tecnologías se han desarrollado elementos, estrategias y herramientas que constantemente buscan facilitar las labores o los procesos que se desarrollan en los diversos empleos o labores a nivel mundial. Grandes avances en estos campos podemos encontrar a diario en la cotidianidad, desde los sofisticados elementos electrónicos de los que habitualmente se hace uso, hasta los sistemas que permiten extraer dinero de un cajero, pagar un transporte por medio de una tarjeta previamente cargada, un identificador biométrico con el que puedes hacer uso de la membresía del gimnasio al que asistes,

todos estos elementos son el resultado de infinidad de estrategias, algoritmos y códigos computacionales que desglosan la información natural a un lenguaje binario entendible para las máquinas y que se traduce en estrategias de mejora y desarrollo.

El Instituto Nacional de Imágenes Biomédicas y Bioingeniería de los Estados Unidos define un modelo computacional como un sistema de computadoras que busca simular y estudiar sistemas complejos utilizando elementos como la lógica, la matemática y la física; es poseedor de numerosas variables que se hacen presentes según cada caso de estudio y que tiene como principal objetivo la realización de experimentos constantes que, referidos a la cotidianidad, solucionen problemas reiterativos en los lugares de trabajo, generen nuevos elementos o estrategias para el desarrollo de actividades y aseguren cada vez más la correcta realización de las labores, en palabras más sencillas, solución de problemas previamente identificados (NIBIB, 2024).

El modelado computacional tiene diferentes enfoques y usos, desde crear herramientas o estrategias para el pronóstico del tiempo, crear simulaciones de vuelo, estudiar imágenes y elementos biomédico y de bioingeniería, hasta optimizar procesos, que es el enfoque que se busca en esta investigación, pues son los sistemas de control los que estabilizan los procesos y permiten la operación automática de los mismos, reduciendo el tiempo y esfuerzo de los mecanismos humanos (Mar, Gulin & Santana, 2024).

Pero, no es posible hablar de modelos computacionales si no se habla en primera instancia de programación, ella es la raíz de cada modelo, sistema y software que se conoce y de los que constantemente se hace uso.

Inicialmente, la programación debe ser comprendida como un “conjunto de instrucciones consecutivas y ordenadas que llevan a la computadora a ejecutar una labor específica” (EAN, 2022) de lo que se deduce que, no es solo cuestión de códigos sino de ejecutar tareas de forma consecutiva que en sumatoria arrojen la correcta ejecución del código y la realización de la tarea mayor que se buscaba.

Todo objetivo que se enfoque en la programación como medio para alcanzar, requiere ser desglosado en pequeñas tareas con un orden claro que permiten, a medida que se van ejecutando, arrojar resultados que sirvan de insumo a la tarea siguiente. La lógica computacional es la herramienta esencial para ese proceso, conocida como la forma en que se le da la orden a la computadora, cuya principal característica es que debe de ser clara y precisa; con una orden bien definida y el paso a paso lógico para cumplirla, la pequeña tarea arrojará un resultado claro.

La programación puede verse enfocada hacia dos grandes líneas de trabajo, la programación estructurada “se basa en que la creación de un programa implica la definición de procesos que actúan sobre un conjunto independiente de datos” (p. 1) mientras que la programación orientada a objetos “es una nueva forma de pensar los problemas usando modelos sobre conceptos del mundo real” (p.1) esto quiere decir que, mientras la estructurada busca crear desde la definición y los procesos, la orientada a objetos va a buscar la manera de usar modelos reales sobre problemas vigentes. (Friss, I. 2018).

Dentro del campo de la programación, la era de los datos ha revelado la necesidad de dominar diversidad de lenguajes de programación, tal como lo expresa la Universidad de ORT en Uruguay, con la finalidad de incrementar no solo la posibilidad de soluciones sino también el orden de la competitividad; en un estudio realizado, se arrojó que, al año 2021, estos eran los 10 lenguajes de programación más usados a nivel mundial según la encuesta Stack Overflow 2021 Developer survey, dentro de la que se encuentran dos lenguajes que más adelante se convertirán en factores claves en el desarrollo del modelo (Universidad de ORT, 2024).

Básicamente el listado de lenguajes es:

- JavaScript
- HTML
- Python
- SQL
- Java
- C#
- NoSQL
- Rust
- Perl
- Swift

Los lenguajes de programación ofrecen diversidad de herramientas y librerías que posibilitan las innumerables funciones e ideas que se pretenden plasmar según cada tipo de tareas. En consonancia con todo lo anterior y contextualizando esto en la labor que desarrolla la OTRI de la UdeA, se ve la oportunidad de que, tomando el servicio de vigilancia e inteligencia estratégica que allí se ofrece y sumado a las grandes oportunidades de la programación, se resuelva la

necesidad de optimizar el tiempo y el esfuerzo que se imprime en el momento de la captura y análisis de información según la investigación que se desarrolle.

Web Scraping

El *web scraping* es una técnica de programación que permite recuperar diversos tipos de contenidos alojados en la web mediante un proceso de búsqueda sistemático y sencillo. A nivel de las organizaciones, es comúnmente utilizada esta técnica para elaborar comparaciones entre elementos, informaciones, estados del mercado y de la competencia (González, 2019.)

La idea de abordar gran cantidad de sitios web con contenidos variados en corto tiempo es el propósito de esta técnica, pues, mediante un software, se permite realizar la búsqueda siempre y cuando el contenido deseado haya pasado por un proceso de indexación que le haya dotado de identificación para el reconocimiento del código. En la estructura de esta técnica, se rescatan identificadores web a los que el código puede ingresar para extraer la información que busca, todo de manera automatizada (Orozco y Gómez. 2019).

Por otro lado, González (2019) establece que el modelo de *web scraping* más usado es el *spider bots* que básicamente es una estructura de software con unas órdenes precisas y claras que son ingresadas según contemple las necesidades el programador, también estas arañas tienen la capacidad de entender la estructura de páginas web usualmente configuradas bajo el diseño HTML (Hypertext Markup Language) y que posteriormente almacena y transforma como datos ordenados.

La legalidad es un tema que pone en discusión la técnica de *web scraping*, al ser una técnica de raspado en la web para encontrar los elementos solicitados por el programador, suele pensarse que aquello que es recuperado proviene de fuentes poco confiables y que el método de adquisición de aquella información no ingresa en la categoría de lo legal o correcto, puesto, esta técnica hace sondeos completos de las páginas que le permiten el acceso, solo que la recuperación se hace bajo los parámetros éticos de quien lo solicita. Es una delgada línea que debe de ser considerada, puesto

que la técnica no es mala en sí misma, sí lo es el fin para el que se destina la información recuperada bajo ella.

Tres elementos son puestos en consideración al hablar de materia legal:

- Competencia desleal
- Violación en términos de legalidad y uso
- Vulneración de la protección de datos y el derecho de autor

Bravo, Gómez y Cobos (2021) proponen algunas herramientas para diseñar el código para el *web scraping* que podrían considerarse de utilidad para esta investigación y que se consignan a continuación sólo en mención y significado básico:

- **Scrapy:** framework de alto nivel para desarrollo de software
- **Beautifulsoup:** librería de Python para extracción de datos de documentos
- **Jaunt:** librería de Java para realizar Web Scraping
- **Jauntium:** librería de Java para Web Scraping
- **Selenium:** conjunto de herramientas desarrolladas para la automatización de la navegación en visores web (Navegadores)

Aparte de estos frameworks, las applications programming interface (APIs) y los Add-ons entendidos como complementos o extensiones), son elementos que aportan valor a los procesos de web scraping puesto que, los primeros aportan nivel en la automatización de los procesos haciéndolos más independientes y concretos, los segundos por su parte aportan a la extracción de los datos que se están buscando; internamente existen varios de cada uno pero se dejan a elección de quien desee usarlo y según su propósito (Bravo et al., 2021).

7 Metodología

El desarrollo de la presente investigación consta de tres objetivos específicos que contribuyen a la consecución del objetivo principal de proponer un método de mejora para los procesos de captura y análisis propios de la Vigilancia e Inteligencia estratégica y que se deseen ser optimizados en la OTRI de la Universidad de Antioquia y en la que se visualiza un proceso de automatización en el que se logre reducir tiempo y esfuerzo en el proceso mecánico.

Identificación de variables

Para alcanzar el primer objetivo que es identificar cómo se desarrolla cada proceso al interior de la OTRI, se propone hacer una entrevista semiestructurada dirigida a aquellas personas responsables de prestar estos servicios, con el propósito de identificar las fuentes, modelos y variables que requieren en cada uno de los procesos de Vigilancia. En esta se identificaron propósitos, necesidades y expectativas sobre el modelo; así como las diversas fuentes de información y rescate de datos para los procesos.

La captura y el análisis de la información son el enfoque particular de este objetivo puesto que, las claridades ofrecidas sobre los procesos y cómo los direccionan a la gestión de los datos, se considera como el insumo base de donde parte la creación del modelo y los requerimientos básicos que deben ser abordados.

Jaramillo (2021) identifica algunos elementos que son necesario rescatar y comprender en el proceso de la recuperación tales como:

- Nombre de la fuente
- Descripción
- Tipo de fuente
- Accesibilidad
- Disponibilidad
- Alcance
- Tipo de resultado

Estos elementos son plasmados en las diferentes fichas y bitácoras que se llevan al interior de cada solicitud e investigación; no es un proceso al azar sino previamente estructurado como guía

de necesidades de la organización que solicita el servicio. Todos estos elementos serán profundizados en la parte inicial del desarrollo del modelo.

Estructuración Bibliográfica

En esta etapa es necesario realizar una búsqueda documental al interior de la Oficina con la intención de identificar procesos, manuales y normativas que se encuentren diseñadas según las necesidades de la división y en específico de la oficina; se busca también elementos que establezcan requisitos, objetivos y modelos del proceso de captura y análisis y cómo se posicionan en respuesta a las diferentes solicitudes que le realizan a la Oficina y aquellos que podrían ser implementados de forma que reactiven los procesos.

Identificar elementos externos que aporten y enriquezcan los procesos de vigilancia en específico la captura y análisis sería pertinente para considerarlos en el proceso de optimización, así el valor que se le suma a la producción no solo radica en ideas dadas por otros autores sino en modelos implementados o surgidos de otro tipo de investigaciones que se asemeje a esta.

La búsqueda documental abre paso a saber en donde se encuentran los avances en el área de interés y permite, junto con los elementos rescatados de la conversación con los profesionales, establecer cuáles serán los elementos que marcarán la construcción del modelo.

Modelación del programa

Luego de haber elegido el lenguaje que se va a utilizar, que sería Python, como entorno de desarrollo Google Colab y las diversas librerías que serán necesarias según los bloques en los que se divide el modelo, la modelación del programa inicia instalando librerías claves y extrayendo de ellas elementos puntuales, esto claramente se va ampliando a medida que el modelo va creciendo en su desarrollo, pues según los retos que se presentan se acude a llamar nuevos elementos.

Seguidamente, teniendo identificada la página a la que se le va a realizar el proceso de web scraping y cómo es su configuración HTML, se selecciona una de las fuentes para comenzar con la extracción, ante ello, siempre se debe de tener claro la solicitud que se debe de hacer y el procedimiento base para la extracción, modelos y documentos que se encontraron en la búsqueda documental son de ayuda en este punto de la investigación, pues pueden otorgar herramientas puntuales para enfrentar problemas durante la captura. Al terminar una, se procede a realizar el mismo proceso con la segunda fuente establecida, nuevamente se considera vital tener claro los

elementos que se desean capturar y la forma en qué están configurados, allí reside la correcta extracción de información relevante y que el modelo funcione.

Habiendo extraído los datos de las fuentes previstas, con la ayuda de librerías y técnicas, se implementará la parte de limpieza y preparación de los datos, esto es, la eliminación de duplicados, datos vacíos y registros externos que pudiesen afectar el análisis de la información en el proceso de implementación. La limpieza y correcta disposición de los datos es la etapa final del bloque de captura, pues su resultado abre paso al bloque de análisis.

Implementación del prototipo de captura y análisis

La implementación del prototipo del modelo no es una etapa en si misma sino que se da durante el curso de creación de este en cuanto a que las búsquedas que se realizan y la extracción de los datos se hacen de una vez en miras a la temática preestablecida. La etapa de implementación corresponde más al tratamiento de los datos, con librerías y técnicas, no de captura sino de procesamiento de texto y análisis:

Aquí es donde el resultado de la captura revela si de verdad funcionó el modelo para lo que estaba siendo implementado, los resultados de las técnicas de machine learning reflejará que la tokenización y el embedding son herramientas necesarias para la clusterización, finalmente será esta parte del modelo la que revelará los resultados de aquello que se capturó y procesó. Lograr reducir un proceso de días a horas se considera algo ambicioso, pero en el desarrollo de este modelo para captura y análisis, los profesionales de la información de cualquier lugar, pero en este caso los profesionales en VeIE de la OTRI, lograrán reducir tiempo y esfuerzo que serán invertidos en nuevos servicios o potenciar los que ya están en vigencia.

Ilustración 3: Metodología de actividades

Objetivo	Método	Instrumento	¿A quién?
<i>Identificar las necesidades de información que se presentan en la OTRI y los procesos que se desarrollan con las diferentes solicitudes que permita el establecimiento de los insumos básicos de donde parte el desarrollo de los procesos.</i>	Entrevista semiestructurada	Cuestionario	Personal que desarrolla las actividades de Vigilancia e Inteligencia al Interior de la OTRI
<i>Diseñar un método para la captura de información, que disminuya los tiempos requeridos para la extracción de datos de fuentes externas empleados en los ejercicios de</i>	Sondeo documental	Fichas de contenido	Bases de datos y archivos internos de la OTRI

<p><i>vigilancia e inteligencia estratégica</i></p>	<p>Modelo de Web scraping</p>	<p>Lenguaje de programación</p>	<p>Procesos de Vigilancia e Inteligencia que se desarrollen en la Oficina</p>
<p><i>Validar el método de captura de datos propuesto a través de un ejercicio práctico de vigilancia e inteligencia estratégica al interior de la Oficina de Transferencia de Resultados de Investigación de la Universidad de Antioquia</i></p>	<p>Prueba piloto</p>	<p>Modelo computacional</p>	<p>Proceso de Vigilancia e Inteligencia estratégica vigente.</p>

Fuente: Elaboración propia

8 Vigilancia e Inteligencia Estratégica

También conocida como Vigilancia Tecnológica, puede considerarse como un conjunto de procesos y etapas en donde contemplando variables previamente definidas se realiza la búsqueda de información que aporte a una posterior toma de decisiones que rondan en el beneficio de la organización que se proyecta avanzar mediante esta técnica (Cidei, 2024).

Existe una claridad al momento de diferenciar ambos elementos, pues si bien la Vigilancia estratégica y la Inteligencia competitiva son muy cercanos y usualmente trabajan de la mano, no pueden ser considerados como lo mismo. El Centro de Investigación y desarrollo tecnológico de la Industria Electro Electrónica y TIC establece que “vigilancia es monitoreo y seguimiento de información” mientras que “la inteligencia es contextualizar e interpretar esa información” (Cidei, 2024).

El gran beneficio de esto radica en la constante actualización del conocimiento que se desarrolla y que sirve de fuente de proyección para organizaciones, mercados, negocios y demás entes que buscan hacer parte del proceso de innovación que mueve el mundo actual.

Ahora bien, en el ámbito universitario, esta temática podría considerarse como algo aparte de novedoso, anclado a ciertas áreas del conocimiento que pueden desarrollarlo de manera precisa pero no se puede desconocer que la producción constante de datos, información, conocimiento y prototipos que encierran todo esto, es tarea constante de todas las áreas del conocimiento, por eso, todos los campos del conocimiento son considerados como entornos altamente competitivos en los que, desde cada vertiente se impulsan diferentes ideas o estrategias para el crecimiento del área (Pavlicevic, J. et al. 2017).

En consonancia con lo anterior, Marulanda, Hernández y López (2016) plantean que es necesario comprender que las universidades, como fuentes de conocimiento diverso, deben de involucrar siempre en sus desarrollos la innovación como una oportunidad de acoplamiento y desarrollo de elementos que cubran las necesidades del mundo actual a nivel económico, social, político, tecnológico, entre otras dimensiones y los diferentes campos del saber y que no solamente deben de ser traducido en algo nuevo sino también en la reestructuración o replanteamiento de aquello que ya existe.

La Universidad de Antioquia no es la excepción a este argumento, la eficiencia de la que se habla a nivel corporativo no solo involucra a las grandes organizaciones ni tampoco a aquellas que están apenas en su nacimiento empresarial, sino que se prevé posicionar y establecer lineamientos

básicos en eficiencia y competitividad para todos y cada uno de los sujetos que componen en orden social natural; es desde allí que se visualizan herramientas necesarias que posibiliten hacer de cada tarea, de cada proceso, algo más óptimo y eficiente, de manera que cualquier acción de mejora se convierta en una rápida oportunidad de crecimiento.

Es allí donde la OTRI de la UdeA participa activamente del desarrollo de la sociedad en sus diversos componentes atendiendo cada solicitud, considerando la escala de etapas del proceso que ella misma define en su página institucional; comenzando desde las investigaciones en donde se observan procesos que se estén desarrollando y que puedan ser susceptibles a la transferencia; continúa luego la etapa de identificación en la que se localizan aquellos resultados que arrojó la investigación y que se encuentran bajo los parámetros de Investigación y Desarrollo (I+D) que pueden ser gestionados a vinculación empresarial y financiación de diversos fondos; la etapa de aseguramiento y gestión contractual es un proceso transversal en el que se vinculan actores tanto internos de la universidad como externos al referirse a las entidades (OTRI UdeA, s.f).

La valorización es la etapa siguiente en la que se estudia cuál sería el valor de la invención en relación al mercado y se evalúan aspectos técnicos y pre-comerciales; la etapa de protección de propiedad intelectual básicamente busca establecer una forma de protección de la invención y comenzar el proceso con las entidades pertinentes, finalmente la última etapa que se contempla es la de comercialización en la que se comienza el plan de difusión o las conexiones con los aliados a quienes beneficia la invención (OTRI UdeA, s.f).

Este es el proceso que usualmente la OTRI desarrolla con los diferentes procesos de investigación que lleva la universidad, es allí donde la VeIE cobra un valor potencial para el resultado final de estas invenciones, apropiado es mencionar que la oficina, no solo se dispone a servir a los desarrollos institucionales, sino que se encuentra también abierta a las diversas solicitudes que puedan llegar de organizaciones que también desarrollen procesos de investigación.

Los desarrollos que se llevan desde la oficina y requieren de proceso de vigilancia e inteligencia, ingresan bajo la propuesta de desarrollo nacida de los diversos análisis que los grupos de investigación o las organizaciones realizan de sí mismas y de su eje misional que contribuyan al sector al cual se enfocan y se vea retribuidos en el crecimiento de sí, por lo que, adentrándose en el desarrollo de la vigilancia, el primer paso que se especifica en este ciclo corresponde al de Identificar.

A raíz de las solicitudes que llegan a la oficina, el análisis de cada una se traduce en información que posteriormente es consignada en un elemento base que se denomina ficha de necesidad de vigilancia tecnológica – FN - (ver Anexo 1) y cuyo objetivo es la estructuración de la oportunidad para los cinco pasos restantes en el proceso. En ella se guardan aquellos elementos que son relevantes para la búsqueda y captura de información, pues se debe garantizar que la sintonía entre lo que se solicita y lo que se busca sea la misma.

La ficha contiene datos como: ¿quién es el solicitante?, ¿cuál es la tecnología que buscan? y una descripción donde se especifique de la manera más clara qué es lo que se requiere, se establece también de manera clara cuál es el enfoque de la vigilancia y los factores críticos de la misma; la delimitación geográfica también juega un papel clave en el desarrollo del proceso puesto que, definir el campo de acción al que se desea enfocar la investigación según la determinación que considere la organización aporta valor y sentido a lo que se pretende recuperar (OTRI, s.f).

El Factor Crítico de Vigilancia (FCV) es definido por la norma UNE 166002:2006 como: “las cuestiones externas a la organización cuya evolución es crucial para su competitividad: tecnologías emergentes, competidores actuales y potenciales, desarrollo de los mercados y del entorno” (Asociación Española de Normalización y Certificación –AENOR-, 2006), y que varios autores consideran como la fuente de donde se derivan las palabras claves, ecuaciones de búsqueda y selección de fuentes que concluirá en la entrega de aquellas debilidades que deben ser potenciadas como puntos neurálgicos al interior de la organización (Sepúlveda-Aguirre, et al. 2018). La LISA Institute de España refuerza que “los FCV se determinan por cada actividad de la cadena de valor de la organización” (LISA Institute, s,f), estableciendo así el punto de origen de los objetivos de búsqueda de la VeIE.

Por otra parte, las claridades definidas en la ficha de necesidad y la información adicional que haya sido aportada por el solicitante, abren camino a la estructuración de la ficha de planeación (FP) para la vigilancia (ver Anexo 2) la cual es un instrumento de orientación y de establecimiento de límites para el proceso de desarrollo de la VeIE; la ficha de planeación se convierte en la hoja de ruta que permite que el proceso camine en los límites establecidos y bajo los parámetros definidos por las partes de la investigación.

La planeación contempla aquellos elementos que son de relevancia para el desarrollo de la búsqueda, los que comprometen el avance de la investigación y que son considerados como aportes necesarios para el camino, y desglosar de manera más detallada lo que es necesario en la VeIE,

pues no solo toma la información consignada en la FN, sino que elabora nuevos elementos que acompañan la búsqueda y captura; a medida en que se avanza en la consideración de las FCV. De esta manera, la ficha de planeación (FP) se va robusteciendo en términos, ecuaciones y fuentes que se reconsideran en el proceso.

Por otro lado, la FP es el elemento clave para la bitácora de búsqueda, pues otorga elementos necesarios y en ella se deben consignar todos los resultados que se capturen, así pues, es necesario que la sintonía sea transversal a todo: terminología, ecuaciones, componentes para el análisis, fuentes y aportes, garantizando así que el proceso sea adecuado y no se salga de los límites y esquemas definidos para cada proyecto o investigación.

La VeIE es un método enriquecedor para las organizaciones, pero su desarrollo es lento y mecánico si se considera con cuidado y dedicación cada uno de los puntos suministrados en ambas fichas anteriormente mencionadas, la captura será exigente en la medida en que se guarde fidelidad a las necesidades planteadas en los FCV por lo que, llevar una bitácora de búsquedas y resultados logrará dos cosas:

Dar un panorama general de la producción científica o comercial de la temática que se está desarrollando y que involucra aquellos elementos considerados en la investigación que se presenta a la oficina

Capturar con cierto valor de certeza aquellos resultados que se acerquen a lo que se busca pero que aún no se les otorga un peso de valor que dé cuenta de la relación con la investigación.

De lo anterior, la oficina ha implementado un modelo de bitácora de búsqueda que ha sido desarrollado según sus necesidades y que recibe con precisión aquellos elementos que el vigía considera necesarios en la exploración, su rol es necesario y sus aportes concretos puesto que, “busca, analiza y da valor a la información” (Arango, B, et al. 2021 p.6), evaluándola bajo los términos de las necesidades y presentando resultados coherentes al proceso (ver Anexo 3).

Previo a la búsqueda, el vigía define, según el proyecto al que presta el servicio de vigilancia, cuáles son las bases de datos más precisas para realizar el sondeo documental y se establece también la tipología documental por la que se va a orientar el proceso de indagación, si son libros, capítulos de libros, artículos, noticias, patentes o cualquier otro documento que guarde información relevante. Usualmente los metadatos, que son definidos como “conjunto de datos que describen el contenido informativo de un recurso, de archivos o de información de los mismos”

(PowerData, 2024), se convierten en identificadores clave de los resultados que se van encontrando en la navegación y que serán compilados en el documento pertinente.

La OTRI define para su bitácora de búsqueda en las bases de datos seleccionadas y referenciadas en la FP, el recopilar los siguientes datos de las diversas ecuaciones que son implementadas:

- Fecha de búsqueda
- Fuente de búsqueda
- Ecuación de búsqueda
- Filtros
- Número de resultados (este es el número que sale al ejecutar una ecuación de búsqueda que captura diferentes elementos)
- URL
- Observaciones que sean pertinentes hacer de la ecuación de búsqueda o de los resultados obtenidos.

En referencia con lo anterior, también cabe mencionar que la oficina guarda una bitácora de búsqueda con algunos campos extras, aparte de los ya mencionados, para una serie de búsquedas que son consideradas como “libres” pero que hacen referencia a las búsquedas generales realizadas en alguno de los buscadores que se tenga implementados, aunque no solo se queda anclado allí, sino que también puede ser usada para búsquedas más generales en ciertos tipos de páginas web o bases de datos (ver Anexo 4).

En esta bitácora de búsquedas libres, inicialmente se encuentran los mismos encabezados de columnas que la descrita anteriormente, pero, ya las observaciones se transforman en resultados de interés y se le añaden otra serie de campos tales como:

- Resultados de interés (ya mencionado)
- URL a esos resultados
- Tipo de documento
- Observaciones al documento
- Enlace directo al documento

Ahora bien, los anteriores formatos son competentes para la captura de información realizada en las bases de datos pero que refieren a tipologías documentales más formales y

académicas. En el caso de las patentes, son desarrolladas ya no en una bitácora de búsqueda, sino en una matriz con cierta similitud de elementos a las bitácoras pero que recupera información de orden directo a la estructuración del documento patente y a la custodia de la propiedad que allí reposa y que puede considerarse de mayor relevancia para el desarrollo investigativo que sirve el proceso (ver Anexo 5).

Para el análisis de las patentes se dispone de un filtro que permite agrupar los resultados según categorías de relevancia en las que el análisis permita ser más precisos; de allí que los datos que se capturen durante el proceso de búsqueda de estas correspondan a:

- ID de la plataforma donde fue rescatada, la más usada en este caso sería Lens.
- URL
- Número de la patente (corresponde al código asignado al documento una vez es registrado)
- Estatus (si la patente se encuentra activa, o si ya ha sido liberada)
- Año
- Título de la patente (inglés)
- Título de la patente (español)
- Etapa (esta casilla corresponde a características particulares que deban ser consideradas en el proceso de investigación, si se desarrolla por etapas, la patente a qué etapa estaría dirigida)
- Descripción del desarrollo
- Resumen o declaración de interés
- Puntuación (correspondiente a la relación con la temática desarrollada en la investigación)
- Solicitantes propietarios

Basados en este contenido, la depuración de los registros de las patentes otorga la garantía de que lo recuperado verdaderamente cumple con la relación propia de aquello que se consignó tanto en la FN como en la FP, se han considerado los FCV y el resultado entregado al análisis estará cargado de información relevante.

Considerando todos los elementos anteriormente descritos, se seleccionan aquellos que son relevantes para el proceso de captura de información por medio del web scraping que, si bien la

oficina no desarrolla elementos de esta categoría, si entrará a tomar parte más a fondo en el desarrollo del objetivo específico siguiente; por el momento cabe aclarar que, de ambas fichas (FN – FP) se toman como insumos principales las bases de datos identificadas o consignadas en las fichas para la captura de información científica como de patentes, las ecuaciones de búsqueda previamente construidas o aportadas por los desarrolladores del proyecto, el enfoque de la vigilancia y los FCV de la misma para definir así las estrategias más adecuadas para el proceso de búsqueda.

El desarrollo del modelo requiere de dichos insumos base en la misma medida en que el vigía implemente estos elementos en la navegación que requiere implementar y posteriormente esta misma información establezca los lineamientos necesarios para la selección de los documentos y el análisis de estos.

Con todo lo anterior, cabe resaltar que para el ejercicio que se desarrolla con esta investigación, luego de conversar con expertos del área se decide tomar como temática de búsqueda para el modelo la *Energías Renovables*, buscando recopilar información sobre noticias, publicaciones y documentos que puedan ser rescatados, que tengan coherencia y significado para la temática y que sirvan de insumos para posteriores análisis según lo planeado en los FCV y en las FN y FP.

9 Modelo Computacional

El modelo computacional que se busca desarrollar para la OTRI es la oportunidad de optimizar el desarrollo de sus actividades y labores. Cabe resaltar que, tal como se mencionaba en las conceptualizaciones del proceso y las herramientas que usan en el proceso de captura y teniendo como referencia los anexos presentados de las bitácoras y fichas de necesidades, se establecen algunos parámetros necesarios para la codificación del modelo.

Entretanto, según los desarrollos que se llevan al interior de la oficina se hizo la elección como tema de exploración a la “eficiencia energética” puesto que es una temática que es de interés para los acompañamientos que se realizan a los grupos de investigación de la universidad.

9.1 Elementos iniciales

En sesiones de entrevistas con los profesionales de OTRI, se decidió tomar como caso de estudio para este ejercicio el sitio web de IRENA - International Renewable Energy Agency (Agencia Internacional de Energías Renovables). Irena es una organización intergubernamental

que presta apoyo a aquellas naciones que buscan su tránsito a energías renovables y sostenibles en el futuro, su principal misión es respaldar espacios como “cooperación internacional, centro de excelencia y repositorio de políticas, tecnología, recursos y conocimientos financieros en materia de energía renovable” de donde se pueden apoyar quienes requieran elementos a implementar en sus estrategias de transición. (IAEA, 2024)

Irena ofrece diversidad de elementos de acceso abierto que sirven como elementos rescatables y que para el caso del modelo computacional a desarrollar, son elementos esenciales para capturar aquellos que se encuentran en la sección de Noticias (News) y de Publicaciones (Publications) pues guardan información que debe ser evaluada en su relevancia según cada solicitud realizada al servicio de vigilancia e inteligencia estratégica.

Inicialmente se seleccionó como entorno de desarrollo para la construcción del código a Google Colab, este entorno en línea permite trabajar de manera sencilla y autónoma en el desarrollo de las líneas de programación, el lenguaje más usado para el trabajo en este entorno es Python por lo que se convierte a la vez en el lenguaje del desarrollo.

A continuación, se describen las etapas que se llevaron a cabo para el proceso del web scraping:

1. Selección e inspeccionar sitio web
2. Enviar solicitud HTML
3. Extracción de secciones específicas
4. Almacenamiento de datos

9.2 Modelación del código

9.2.1 Librerías

En el proceso de *web scraping* la librería de Selenium es ideal para emular entornos de búsqueda tales como Chrome mediante los drivers de cada buscador, por lo cual se hace uso de ella para los procesos de búsqueda de resultados y también se importan elementos precisos de algunas librerías relevantes como pandas, BeautifulSoup, matplotlib, nltk y request, cada uno de estos elementos tienen un uso particular, un ejemplo de ello es BeautifulSoup que básicamente tiene la función de rastrear datos precisos de las capturas que se hacen y así mismo los demás elementos importados que son propios del código (Python, 2024)

9.2.2 Solicitud HTML

Después de haber cargado las librerías y haber importado los elementos necesarios, comienza el proceso de web scraping. Primero se define inicialmente la URL de la página a la que se le va a hacer el proceso y luego se diseña la línea de código que solicitará los permisos para realizar el scraping, el principal identificador de que la solicitud ha sido aprobada es que el código de aprobación es el 200. La solicitud es la etapa inicial, a lo que prosigue “parsear” la URL, esto, según la documentación de Python, es transformar una dirección URL en una estructura entendible según un lenguaje de programación como HTML (Python, 2024).

Ilustración 4: Solicitud HTTPS

```

6 # Realizar la solicitud HTTP
7 response = requests.get('https://www.irena.org/Search?contentType=e833bea4-7572-4310-944f-f57c92ab7ead&orderBy=Date', headers=headers)
8
9 # Comprobar el estado de la respuesta
10 if response.status_code == 200:
11     # Crear el objeto soup
12     soup = BeautifulSoup(response.content, 'html.parser')
13

```

Fuente: elaboración propia.

9.2.3 Inspección de página

Luego de haber “parseado” la URL, se activa el proceso de captura por lo que comienza la visualización a la par, esto es la construcción del código y la categorización del formato HTML de la página de Irena. Esta estrategia permite traspasar códigos y estructuras sin correr el riesgo de omitir caracteres que puedan arruinar el código. Es algo así:

Ilustración 5: Inspección de Irena

The screenshot displays a web page on the left and its corresponding DOM tree in browser developer tools on the right. The web page features a header image of a group of people, followed by the article title "The Role of Sustainable Bioenergy in Supporting Climate and Development Goals". Below the title is a short paragraph and a footer with navigation links. The developer tools on the right show the DOM tree with the following structure:

```

<!-- Result #START -->
<article class="c-Result c-Result--showing" data-jscomponent="Result" data-provider="Article">
  <a href="/News/articles/2024/Jun/The-Role-of-Sustainable-Bioenergy-in-Supporting-Climate-and-Development-Goals" class="c-Result_image" target="_self">
    <picture>
  </a>
  <div class="c-Result_content">
    <header class="c-Result_header">
      <a href="/News/articles/2024/Jun/The-Role-of-Sustainable-Bioenergy-in-Supporting-Climate-and-Development-Goals" class="c-Result_heading" target="_self">
        <div class="Heading Heading--noTextMargin">

```

Fuente: elaboración propia.

La visualización de las estructuras de configuración en lenguaje HTML exige identificar las variables y las clases que configuran cada detalle de la página. La imagen anterior revela la configuración HTML de cada una de las noticias que aparecen en el panel general de noticias, en lenguaje HTML sería algo como:

- **Título:** Variable – “h2”; class – “Heading__text”
- **Fecha:** Variable – “time”; class – “c-Result__date”
- **Link:** Variable – “a”; class – “c-Result__heading” a lo que hay que agregar que, para poder rescatar el enlace directo de cada una de las noticias se hace necesario extraer el “href” que configura la dirección a cada texto.

9.2.4 Extracción de secciones específicas

Los elementos mencionados anteriormente son importantes en la estructuración de los códigos de web scraping puesto que se requiere definir cada uno de esos elementos en el proceso que lleva a cabo. En las bitácoras de búsqueda, son elementos claves los títulos, las fechas y los enlaces que se dejan como registros que posteriormente serán evaluados en cuanto a pertinencia o propiedad con la temática.

Lo anterior, se evidencia en el código de la siguiente manera:

Ilustración 6: Extracción de títulos

```
14 # Extraer los títulos de las noticias
15 titulo = soup.find_all('h2', {'class': 'Heading__text'})
16 titulo_list = [x.get_text() for x in titulo] # Incluye todos los elementos
17
18 # Guardar el resultado en un DataFrame
19 titulos = pd.DataFrame(titulo_list, columns=['Títulos'])
20
```

Fuente: elaboración propia.

Esto se modifica en la medida en que las variables cambian, si bien el que se presenta es el código para la recuperación del título, este debe de ser ajustado a la fecha, el contenido, los enlaces y los otros elementos que según cada caso sean solicitados o requeridos para el proceso de análisis.

Los resultados arrojados en cada uno de esos procesos de captura, en este caso de las noticias, se van agrupando los resultados mediante el método concatenar de la librería Pandas en Python.

Ilustración 7: Concatenación de elementos

```
1 #Concatenamos todas los dataframe (tablas creadas)
2 df_final=pd.concat([titulos_df,enlaces_df,fechas_df],axis=1)
3 df_final
```

Fuente: elaboración propia.

En la ilustración 5 se observa un ejemplo de los títulos agrupados, los cuales en la medida que se recorren las páginas del URL del cual se está extrayendo los datos, se irán acumulando.

Ilustración 8: Visualización de títulos

```
41 # Imprimir los títulos para verificar
42 for t in titulo_list:
43     print(t)
```

The Role of Sustainable Bioenergy in Supporting Climate and Development Goals
IRENA Council Convenes Amidst UN Secretary General's Call for Urgent Climate Action
Progress on Basic Energy Access Reverses for First Time in a Decade
IRENA Supports Ukraine in Rebuilding Damaged Energy Infrastructure
Tracking Financial Flows to Clean Cooking is Crucial for Global Accountability and Equitable Progress
COP29 Joins IRENA to Track UAE Consensus Energy Outcomes set at COP28
SIDS Pave the Path to Prosperity, Resilience and Sustainable Development
Decentralised Hydropower Brings Light and Hope to Remote Communities in Papua New Guinea Amid Ongoing Rescue Efforts in Enga
Logo Design Contest Launched to Revamp 'IRENA Youth' Branding
Public-private Collaboration to Track Progress on Global Renewables Target
IRENA Tasked with Monitoring Progress of UAE Consensus Renewable Energy and Energy Efficiency Goals by COP28 Presidency
Galvanising Investments in Africa Through the APRA Forum in Nairobi
G7 Countries Task IRENA to Monitor Group's Renewable Energy Progress
Six Key Takeaways from the 14th IRENA Assembly
Major Global Utilities Announce Joint Intent to Scale Renewable Capacity by 2.5 times to 2030
Africa Takes the Lead to Champion an Innovative Approach for Its Energy Transition
IRENA 14A: Assembly Live Coverage
Transition to Renewables Calls for New Approach to Energy Security
IRENA 14A: Pre-Assembly Live Coverage
The 2024 IRENA Youth Forum Ensures Young People Are Equipped to Lead a Sustainable Energy Future
IRENA Assembly Centre-Stages COP28 Outcome of Tripling Renewable Power Capacity by 2030
Optimising the Role of Public Finance Will Advance Universal Energy Access
Energy Transition at a Glance: Key Data Infographic

Fuente: elaboración propia.

Luego de haber extraído los títulos, se realiza la captura de los enlaces directos a cada una de las noticias, este proceso puede tomar tiempo en relación con la cantidad de resultados, es de tener presente que la construcción de la dirección de enlace se construye de manera particular para cada sitio web. En la siguiente figura se observan los enlaces extraídos.

Ilustración 9: Visualización de enlaces.

```

35 # Visualizar los datos almacenados en el DataFrame
36 enlaces_df

Contenido extraído de https://www.irena.org/News/articles/2024/Jun/The-Role-of-Sustainable-Bioenergy-in-Supporting-Climate-and-Development-Goals
Contenido extraído de https://www.irena.org/News/pressreleases/2024/Jun/IRENA-Council-Convenes-Amidst-UN-Secretary-Generals-Call-for-Urgent-Climate-Action
Contenido extraído de https://www.irena.org/News/articles/2024/Jun/IRENA-Supports-Ukraine-in-Rebuilding-Damaged-Energy-Infrastructure
Contenido extraído de https://www.irena.org/News/expertinsights/2024/Jun/Tracking-Financial-Flows-to-Clean-Cooking-is-Crucial-for-Global-Accountability-Equity
Contenido extraído de https://www.irena.org/News/pressreleases/2024/Jun/COP29-Joins-IRENA-to-Track-UAE-Consensus-Energy-Outcomes-set-at-COP28
Contenido extraído de https://www.irena.org/News/articles/2024/May/SIDS-Pave-the-Path-to-Prosperity-Resilience-and-Sustainable-Development
Contenido extraído de https://www.irena.org/News/articles/2024/May/Decentralised-Hydropower-Brings-Light-and-Hope-to-Remote-Communities-in-Papua-New-Guinea
Contenido extraído de https://www.irena.org/News/articles/2024/May/Logo-Design-Contest-Launched-to-Revamp-IRENA-Youth-Branding
Contenido extraído de https://www.irena.org/News/pressreleases/2024/May/Public-private-collaboration-to-track-progress-on-global-renewables-target
Contenido extraído de https://www.irena.org/News/pressreleases/2024/May/IRENA-Tasked-with-Monitoring-Progress-of-UAE-Consensus-Renewable-Energy-Goals-by-COP28
Contenido extraído de https://www.irena.org/News/articles/2024/May/Galvanising-Investments-in-Africa-Through-the-APRA-Forum-in-Nairobi
Contenido extraído de https://www.irena.org/News/pressreleases/2024/Apr/G7-Countries-Task-IRENA-to-Monitor-Groups-Renewable-Energy-Progress
Contenido extraído de https://www.irena.org/News/articles/2024/Apr/Six-Key-Takeaways-from-the-14th-IRENA-Assembly
Contenido extraído de https://www.irena.org/News/pressreleases/2024/Apr/Major-Global-Utilities-Announce-Joint-Intent-to-Scale-Renewable-Capacity-by-2-point-5
Contenido extraído de https://www.irena.org/News/articles/2024/Apr/Africa-Takes-the-Lead-to-Champion-an-Innovative-Approach-for-Its-Energy-Transition
Contenido extraído de https://www.irena.org/News/articles/2024/Apr/IRENA14A-LiveBlog
Contenido extraído de https://www.irena.org/News/pressreleases/2024/Apr/Transition-to-Renewables-Calls-for-New-Approach-to-Energy-Security
Contenido extraído de https://www.irena.org/News/articles/2024/Apr/The-2024-IRENA-Youth-Forum-Ensures-Young-People-Are-Equipped-to-Lead-a-Sustainable-Energy-F
Contenido extraído de https://www.irena.org/News/pressreleases/2024/Apr/IRENA-Assembly-Centre-Stages-COP28-Outcome-of-Tripling-Renewable-Power-Capacity-by-203
Contenido extraído de https://www.irena.org/News/articles/2024/Apr/IRENA14A-PreAssembly-Live-Blog
Contenido extraído de https://www.irena.org/News/pressreleases/2024/Apr/Optimising-the-Role-of-Public-Finance-Will-Advance-Universal-Energy-Access
Contenido extraído de https://www.irena.org/News/articles/2024/Apr/Energy-Transition-at-a-Glance-Key-Data-Infographic
Contenido extraído de https://www.irena.org/News/expertinsights/2024/Apr/With-One-Voice-Small-Island-Developing-States-Build-Bridges-for-a-Sustainable-Energy-

```

Fuente: elaboración propia.

9.2.5 Almacenamiento de datos

En el caso de estudio, la página de noticias de Irena cuenta con varias subpáginas, cada una de ellas presenta 25 resultados y en la medida que se avanza sobre cada una de ellas se muestran más resultados, en este caso sería un total de 1716 noticias sobre esta temática.

A medida que se recuperan resultados mediante el código, los datos son guardados en un documento de Excel mediante el método de “*export*” que se usa en este lenguaje de programación así se evita repetir ejecuciones que sobrecarguen el entorno y a la vez se evita la pérdida de datos.

9.3 Sección de publicaciones

Luego de realizar el mismo proceso de scraping de la página de noticias de Irena a la sección de publicaciones, se debe considerar que esta posee características diferentes en cuanto a que no son contenidos HTML sino contenidos en PDF, para lo cual hay que realizar un proceso complementario para la extracción de los contenidos de estos archivos. De aquellos elementos a los que se les pueda recuperar la descripción inicial, esta se suma a una nueva columna de datos.

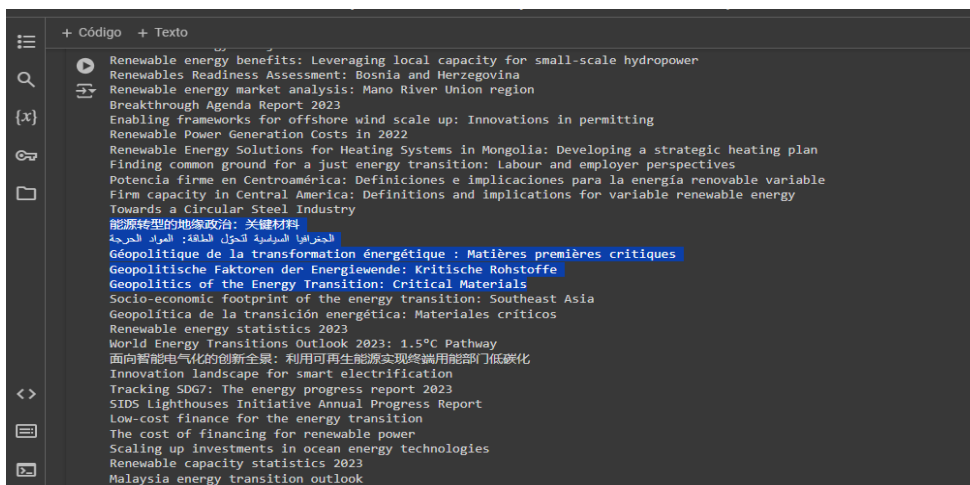
Ilustración 10: Datos de publicación.

	Enlaces	Contenido_Completo
0	/Publications/2024/Jun/Tracking-SDG-7-The-Ener...	\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
1	/Publications/2024/Jun/Climate-action-and-the-...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
2	/Publications/2024/May/Investment-opportunitie...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
3	/Publications/2024/Apr/Tripling-renewable-powe...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
4	/Publications/2024/Apr/Decarbonising-hard-to-a...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
5	/Publications/2024/Apr/The-energy-transition-i...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
6	/Publications/2024/May/Public-finance-for-univ...	To reach universal access to energy services b...
7	/Publications/2024/Apr/Geopolitics-of-the-ener...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
8	/Publications/2024/Apr/International-co-operat...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
9	/Publications/2024/Mar/Renewable-capacity-stat...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
10	/Publications/2024/Apr/The-Global-Atlas-for-Re...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
11	/Publications/2024/Mar/SIDS-at-a-Crossroads-To...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
12	/Publications/2024/Mar/SIDS-at-a-Crossroads-Th...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...
13	/Publications/2024/Mar/100pc-renewable-energy-...	\n\n\n\n\nCopy citation\n\n Copied \n\n\n\nC...

Fuente: elaboración propia.

En la recuperación de los títulos, es necesario resaltar que existen documentos que se encuentran escritos en diversos idiomas como por ejemplo mandarín, árabe, alemán y otros, tal como se presenta en la ilustración 8, por lo que su recuperación será igual, pero es necesario identificarlos para, en el proceso de limpieza de la información recuperada, eliminando las copias que no son necesarias.

Ilustración 11: Duplicado de documentos.



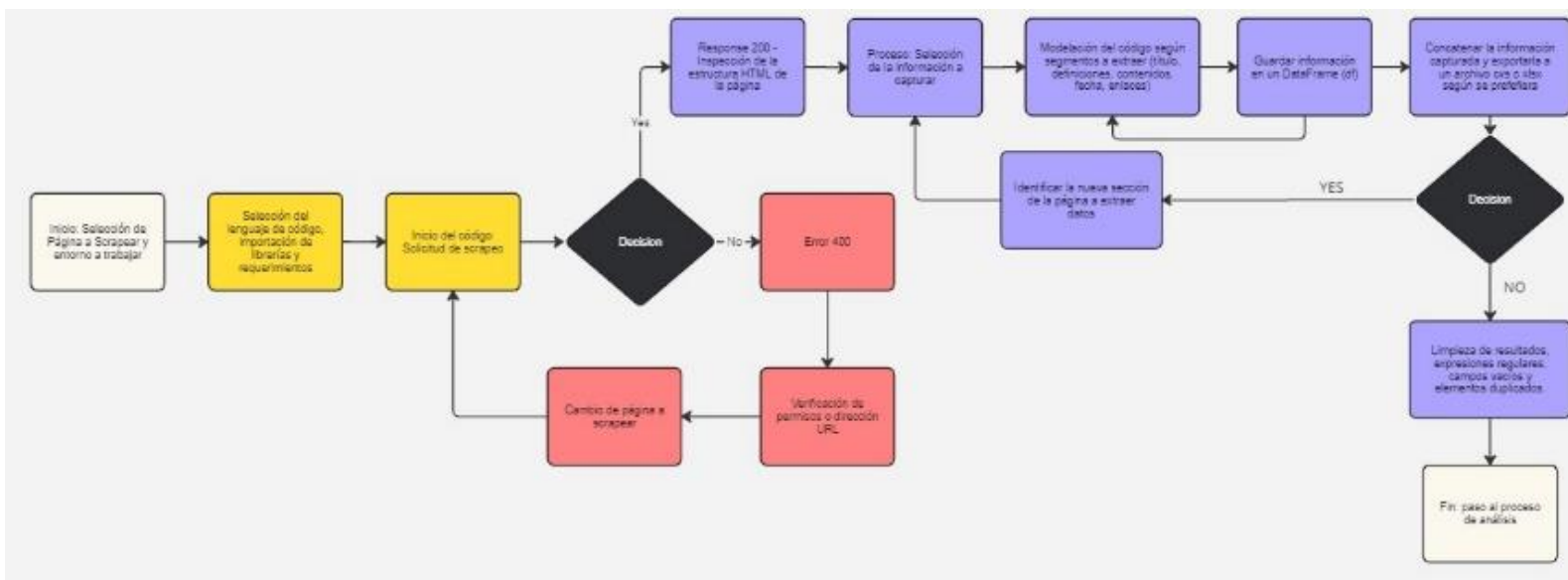
Fuente: elaboración propia

Luego de las debidas depuraciones de la información, debe crearse un dataframe que luego servirá para exportar los resultados al formato de Excel (Anexo 1). Este documento será necesario para el proceso de análisis propio de la vigilancia.

El proceso de definición de las necesidades y la creación de este modelo computacional abre paso a la tercera etapa del proceso investigativo que refiere a la implementación de este modelo a un proceso vigente en la oficina de transferencia, si bien la temática ya ha sido definida y se refleja en este proceso de búsqueda y captura, se proyectó para esta fase final someter a valoración y consideración los resultados obtenidos para el progreso de la investigación.

Todo este proceso se resume de manera más comprensible en la ilustración 12 y que puede visualizarse mejor en el Anexo 2.

Ilustración 12: Mapa del proceso de modelación del código.



Fuente: elaboración propia

10 Implementación del modelo

La generación de los dataframes (DF) (los cuales son estructuras de datos para la manipulación de datos en Python), con los datos extraídos de cada una de las páginas de Irena tal como se desarrollaba en la unidad anterior, permitió valorar la tipología de los datos rescatados y establecer variables precisas que permitirán la igualación de los dos DF creados, tanto el de publicaciones como el de noticias.

Para esta parte final del desarrollo del modelo, se toma el dataset (DS) (conjunto de datos) extraído de la unión de los DF anteriores tal como se presenta en la Ilustración 13, bajo los mismos parámetros buscando establecer desde el principio los elementos clave con los que se va a trabajar la sección siguiente del análisis. La importación del DS implica establecer también la tipología de los datos que se tienen buscando igualar cantidad y tipología bajo los mismos estándares: 2297 filas y 6 columnas, cada una con una tipología establecida: object; esto para proceder a la limpieza del DS.

Ilustración 13: Agrupación de los DF de noticias y publicaciones

```
[ ] 1 df_total=pd.concat([df_final,fn_dfclean],ignore_index=True)
    2 df_total
```

	Títulos	Enlaces	Contenido_Completo	Fecha	url_completa	Origen
0	Tripling Renewables by 2030 Requires a Minimum...	/News/pressreleases/2024/Jul/Tripling-Renewabl...	Abu Dhabi, United Arab Emirates, 11 July 2024 ...	26 June 2024	https://www.irena.org/News/pressreleases/2024/...	Irena News
1	IRENA Opens Registration for Summer Youth Prog...	/News/articles/2024/Jul/IRENA-Opens-Registrati...	The International Renewable Energy Agency (IRE...	20 June 2024	https://www.irena.org/News/articles/2024/Jul/I...	Irena News
2	Bonn's 2024 International Energy Workshop Gave...	/News/articles/2024/Jul/Bonns-2024-Internation...	The 2024 edition of the International Energy W...	13 June 2024	https://www.irena.org/News/articles/2024/Jul/B...	Irena News
3	First-Ever Global Renewables Summit Announced ...	/News/pressreleases/2024/Jun/First-ever-global...	Brussels, Belgium / New York, United States of...	12 June 2024	https://www.irena.org/News/pressreleases/2024/...	Irena News
4	The Role of Sustainable Bioenergy in Supportin...	/News/articles/2024/Jun/The-Role-of-Sustainabl...	As the world searches for viable solutions for...	11 June 2024	https://www.irena.org/News/articles/2024/Jun/T...	Irena News

Fuente: Elaboración propia

Luego de la importación, comienza la limpieza del conjunto de datos capturados, inicialmente se extraen aparte los datos que no corresponden a elementos propios de Irena tales como noticias y publicidad ajena a la temática y a la página; estos datos son identificados mediante

el código que se presenta en la Ilustración 14 y de considerarse necesarios podrían extraerse a un DF independiente que posteriormente podría ser analizado, cabe aclarar que para este caso estos datos no ocupan relevancia alguna por lo que simplemente se identificarán y eliminarán del DS, cabe resaltar que el conjunto de elementos identificados en la limpieza para ser depurados corresponden a 328 elementos que serán restado de los 2297 iniciales, por tanto, el DS con el que el modelo trabajará de ahora en más consta de 1969 registros.

Ilustración 14: Eliminación de registros no correspondientes

```
1 # Usamos una expresión regular para encontrar la cadena "https://www.irena.orghttps://"
2 pattern = r"https://www\.irena\.orghttps://\"
3
4 # Extraemos las filas que contienen el patrón
5 extracted_rows = df_total[df_total.apply(lambda row: row.astype(str).str.contains(pattern).any(), axis=1)]
6
7 # Eliminamos esas filas del DataFrame original
8 df_total_cleaned = df_total[~df_total.apply(lambda row: row.astype(str).str.contains(pattern).any(), axis=1)]
9
10 # Opcional: Verificar los resultados
11 print("Filas extraídas:")
12 print(extracted_rows)
13 print("DataFrame limpio:")
14 print(df_total_cleaned)
15
```

Fuente: Elaboración propia

La limpieza del DS continua con la eliminación de caracteres especiales y elementos propios de la configuración de la página en la estructura HTML a lo que se establecen algunos códigos que realicen tareas como:

- Transformar datos NaN (Not a Number) que son valores especiales con comas que se usan para representar valores no definidos en un DF o DS, estos NaN deben ser transformados en datos no registrados, no capturados o no asignados según corresponda a la columna a la que se refiere, esto se visualiza mejor en la Ilustración 15.

Ilustración 15: Transformación de datos NaN

```
[ ] 1 fn_df.replace('NaN', pd.NA, inplace=True) #Se buscar reemplazar aquellos datos que figuran como NaN por elementos no registrados
2 fn_df['Fecha'].fillna('no registrado', inplace=True)
3 fn_df['Titulos'].fillna('No asignado', inplace=True)
4 fn_df['Contenido_Completo'].fillna('no capturado', inplace=True)
5 print(fn_df)
```

Fuente: elaboración propia

- Se eliminan también aquellos elementos repetidos que pudiesen ocasionar ruido en el DS final, para esto también se implementan unas líneas de código, Ilustración 16.

Ilustración 16. Eliminación de registros duplicados

```

1 print(f'Tamaño del set antes de eliminar las filas repetidas: {fn_df.shape} ') #verificamos el tamaño original
2 fn_df.drop_duplicates(inplace=True) #Recuerde que el inplace nos sirve para decirle a Pandas que modifique el dato en su registro original, cuando se pone
3 print(f'Tamaño del set después de eliminar las filas repetidas: {fn_df.shape} ') #verificamos el tamaño luego de aplicar la función drop_duplicates la cual

```

Fuente: elaboración propia

- La limpieza del texto si bien se trata de hacer para todas las columnas, se enfoca principalmente en el tratamiento de la Columna de “Contenido Completo” la cual es el principal acervo de datos capturados en el proceso, de ella se busca eliminar elementos como: fechas, números, URLs, menciones, hashtags, signos matemáticos, elementos de las horas como a.m y p.m, eliminar espacios extras y normalizar toda la información a minúscula para que el proceso de análisis corresponda de la mejor manera. La Ilustración 17 refleja este proceso.

Ilustración 17: Código de limpieza del DS

```

1 def clean_text(fn_df, Contenido_Completo):
2     #Es una expresión regular que coincide con direcciones de correo electrónico y URLs. Busca secuencias de caracteres que son típicas en estos formatos
3     patternURLEMAIL=r'(\w+[\.]?\w+(\w+\.)+\w+)((http://\www\.|https://\www\.|http://\|https://\|)?\w+([\-\.\.]{1}\w+)*\.[a-z]{2,5}(\/)?(((^s@)*)\(\/)?)*'
4     #Es una expresión regular que busca hashtags (palabras precedidas por #) y menciones (nombres de usuario precedidos por @)
5     patternHashtagMention=r'(@\w+)|(#\w+)'
6     #Convierte todo el texto de la columna especificada a minúsculas para estandarizar el texto
7     fn_df[Contenido_Completo] = fn_df[Contenido_Completo].str.lower()
8     # Utiliza la expresión regular patternURLEMAIL para encontrar y eliminar URLs y direcciones de email en el texto.
9     fn_df[Contenido_Completo] = fn_df[Contenido_Completo].apply(lambda elem: re.sub(patternURLEMAIL, '', elem))
10    # Utiliza la expresión regular patternHashtagMention para encontrar y eliminar hashtags y menciones
11    fn_df[Contenido_Completo] = fn_df[Contenido_Completo].apply(lambda elem: re.sub(patternHashtagMention, '', elem))
12    # Utilizamos una expresión regular también para eliminar los signos de mayor que y menor que
13    fn_df[Contenido_Completo] = fn_df[Contenido_Completo].apply(lambda elem: re.sub(r'(\>|<)', '', elem))
14    # Utilizamos una expresión regular también para eliminar a.m y p.m cuando mencionan horas
15    fn_df[Contenido_Completo] = fn_df[Contenido_Completo].apply(lambda elem: re.sub(r'(a\.m)|(p\.m)', '', elem))
16    # Utilizamos una expresión regular también para eliminar los números
17    fn_df[Contenido_Completo] = fn_df[Contenido_Completo].apply(lambda elem: re.sub(r'\d+', '', elem))
18    fn_df[Contenido_Completo] = fn_df[Contenido_Completo].apply(lambda elem: re.sub(r'[\[\]]', '', elem))
19    ## Sustituir espacios de más
20    fn_df[Contenido_Completo] = fn_df[Contenido_Completo].apply(lambda elem: re.sub(r'\s+', ' ', elem))
21    ## Eliminar signos de puntuación '[!#?,.:;]'
22    df[text_field] = df[text_field].apply(lambda elem: re.sub(r"''['']"", ' ', elem))
23    non_words = list(punctuation)
24    non_words.extend(['¿', '¡', 'í', 'ó', 'ú', 'ç', 'ñ'])
25    #Elimina todos los signos de puntuación del texto, incluyendo algunos caracteres especiales en español como ¿ y ¡, y las comillas tipográficas

```

Fuente: elaboración propia

- Al haber limpiado los datos con los elementos anteriores, se localizan algunos elementos que también deben ser eliminados del registro puesto que hacen ruido en el tratamiento de la información así que, el código de la Ilustración 18, refleja esta parte.

Ilustración 18: eliminación de elementos no necesarios

```
[ ] 1 # Define las palabras que quieres eliminar
    2 words_to_remove = ['copy', 'citation', 'copied', 'url']
    3
    4 # Crea un patrón de expresión regular que coincida con esas palabras
    5 pattern = r'\b(?:' + '|'.join(words_to_remove) + r')\b'
    6
    7 # Aplica la función replace con el patrón en la columna
    8 fn_dfclean['Contenido_Completo'] = fn_dfclean['Contenido_Completo'].str.replace(pattern, '', regex=True)
    9
   10 # Opcional: elimina los espacios extras resultantes
   11 fn_dfclean['Contenido_Completo'] = fn_dfclean['Contenido_Completo'].str.replace(r'\s+', ' ', regex=True).str.strip()
   12
   13 print(fn_dfclean)
   14
```

Fuente: elaboración propia

Teniendo limpio el DS se procede a exportar un nuevo documento que contenga todos los datos fielmente organizados y dispuestos para el análisis de estos; de este DS extraído se genera una copia para ser usada en la fase siguiente que corresponde al análisis.

11 Preprocesamiento del Texto

El preprocesamiento del texto es la organización de los datos capturados bajo unos elementos básicos que permiten realizar un análisis más congruente y estable al momento de dar vida a cada uno de los registros. Este preprocesamiento inicialmente consiste en establecer un código para la estructuración y/o clasificación de las palabras claves – stopwords – tal como se muestra en la Ilustración 19; para posteriormente realizar un proceso de limpieza, el mismo procedimiento que se encuentra en la Ilustración 17 en el que se eliminen por completo aquellos elementos que pudiesen ocasionar ruido alguno en los procesos siguientes. Las stopwords son palabras que al ser usadas sin un contexto no generan valor, por ejemplo, las preposiciones.

Ilustración 19: Stopwords

```
[ ] 1 # Definir stopwords
    2 stop_words = set(stopwords.words('english'))
```

Fuente: elaboración propia

Como el preprocesamiento del texto para este caso se aplica directamente a la columna que contiene los contenidos completos que fueron capturados en cada uno de los registros, se crea una nueva columna en la que se almacena el contenido de cada uno de esos registros ya procesados, es desde allí de donde se toman los elementos para el análisis mediante la tokenización y el embeddings Ilustración 20.

Ilustración 20: Preprocesamiento del texto

```
1 def preprocess_text(Contenido_Completo_Limpio):
2     if isinstance(Contenido_Completo_Limpio, float) and np.isnan(Contenido_Completo_Limpio):
3         return "" # Maneja los valores NaN como texto vacío o según tu necesidad
4     Contenido_Completo_Limpio = str(Contenido_Completo_Limpio).lower().strip()
5     tokens = word_tokenize(Contenido_Completo_Limpio)
6     tokens_filtrados = [word for word in tokens if word.isalnum() and word not in stop_words]
7     return ' '.join(tokens_filtrados)

[ ] 1 print(dfclean['Contenido_Completo_Limpio'].head())

0  abu dhabi united arab emirates  july  the ren...
1  the international renewable energy agency iren...
2  the edition of the international energy works...
3  brussels belgium  new york united states of am...
4  as the world searches for viable solutions for...
Name: Contenido_Completo_Limpio, dtype: object
```

Fuente: elaboración propia

En el proceso de análisis se emplea una técnica de procesamiento de texto para que en el lenguaje binario de las máquinas sea comprensible por ellas, la Asociación para el Progreso de la Dirección establece que “*La tokenización es un proceso fundamental en el ámbito de la informática y la tecnología de la información, que se refiere a la conversión de datos o información en unidades más pequeñas llamadas «tokens»*” (APD, 2023) aludiendo a una definición más concreta del término; cabe aclarar que estos tokens son representaciones numéricas que codifican cada palabra, número o símbolo relevante en un conjunto de datos. El propósito de los tokens es posibilitar que “*los algoritmos pueden identificar patrones más fácilmente*” (Datacamp, 2024) estableciendo relaciones entre dichos patrones y agrupando los datos en conjuntos según su filiación.

El Embeddings por su parte es similar al token en cuanto a que “*Esta técnica consiste en representar palabras con vectores de números*” (Briceño, B. Fernández, E. 2021), buscando

establecer relaciones de orden familiar en el conjunto de datos, todo esto con la intención de posibilitar búsquedas de datos mucho más comprensivas y estables para el lenguaje de la máquina.

Ilustración 21: Embeddings y Tokenización

```
[ ] 1 #Tokenización del texto 1
    2 dfclean['preprocess_text'] = dfclean['Contenido_Completo_Limpio'].apply(preprocess_text)

[ ] 1 #Tokenización del texto 2
    2 dfclean['preprocess_content'] = dfclean['Contenido_Completo_Limpio'].apply(preprocess_text)

[ ] 1 #Guardar dataframe con nueva columna de preprocesamiento
    2 dfclean.to_excel('/content/drive/MyDrive/Colab Notebooks/Trabajo de Grado/Database Irena para análisis preprocesado.xlsx', index=False)

[ ] 1 # Generar variable independiente para embedding
    2 x_text = [txt for txt in dfclean['Contenido_Completo_Limpio']]
    3 x_text[2:5]
```

Fuente: elaboración propia

Con la columna limpia, el código presentado, al aplicarse a ambas técnicas, permite disponer el texto para el análisis mediante la técnica de clusterización.

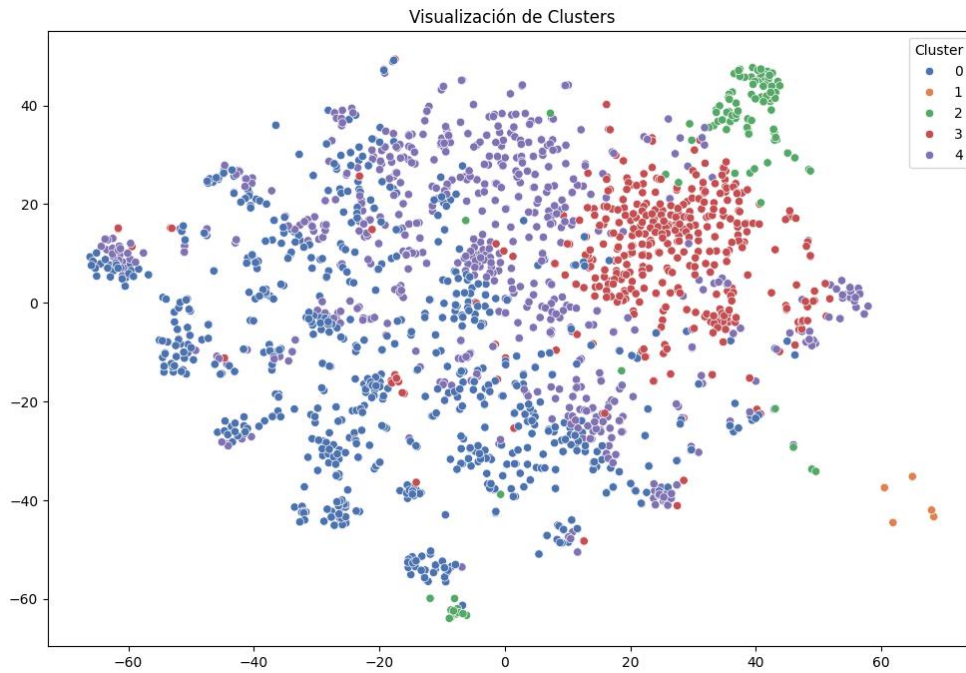
La clusterización es una técnica de *Machine Learning* (ML) enmarcado en el aprendizaje no supervisado en el que se dispone de un conjunto de datos para la entrada, pero no se sabe a ciencia cierta cuál será el resultado que arroje a la salida. Su objetivo es identificar aspectos similares entre los diferentes objetos que se presentan en el set de datos que se le ingrese, según esos aspectos, el modelo debe de agruparlos por similitudes y la relación que hay entre ellos, a estos grupos se les denominará clúster (Gorini, M. s.f).

En el modelo establecido para la OTRI, se aplica el proceso de agrupación a la columna Contenido Completo del DF previamente procesado; se especifica también que realice agrupación en 5 clúster por lo que, según el tipo de relación que guarden los datos, unos entre otros, serán agrupados en los puertos que el modelo por sí mismo defina, estos puertos de agrupación son denominados centroides.

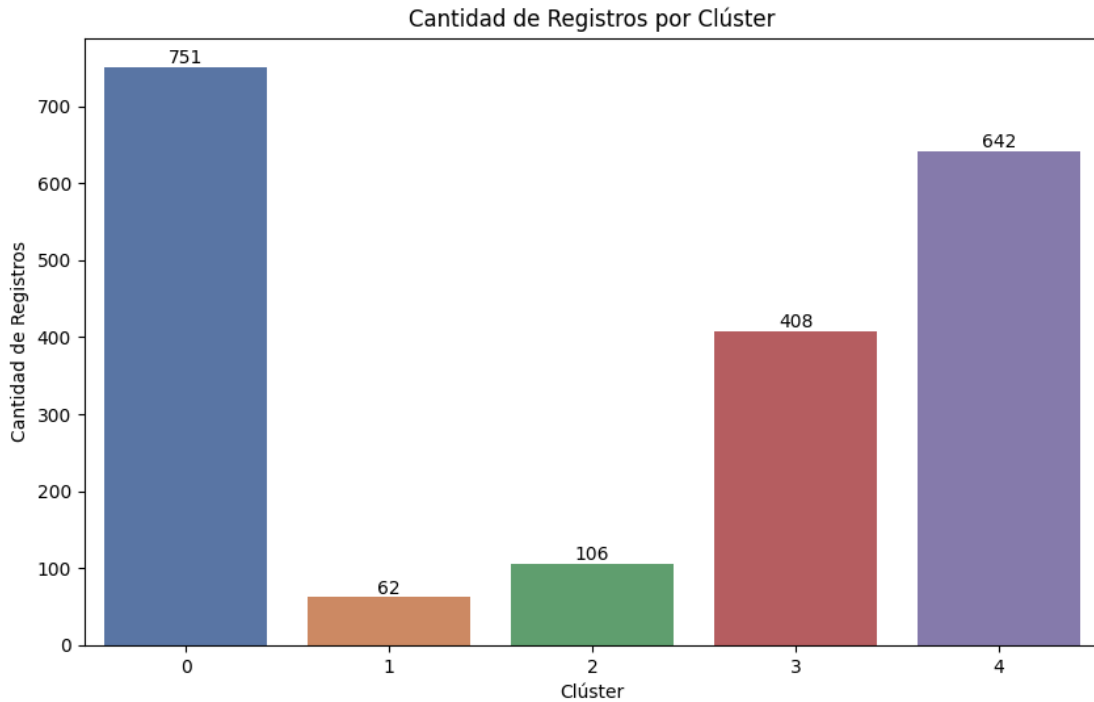
Cabe aclarar que, al solicitarle que la clusterización se realice en 5 clúster, se desconoce la identidad de esos centroides que toma de referencia para hacer la separación por lo que, luego de presentarse la familiaridad entre los datos como se presenta en la Ilustración 22, se procede a graficar los clúster en una nube de palabras (wordcloud) que dé cuenta del relacionamiento establecido y permita que el análisis sea más evidenciable, tal como se presenta en la Ilustración

22; se refleja la relación que estableció el modelo por sí mismo en el set de datos agrupando los registros según su categoría.

Ilustración 22: Clusterización



Fuente: elaboración propia.

Ilustración 24: Cantidad de registros por Clúster

Fuente: elaboración propia.

En el proceso de análisis, se genera el listado de centroides por cada uno de los clústeres buscando establecer los lineamientos básicos para el análisis de los registros, cada listado arroja 20 palabras; los centroides permiten determinar lo siguiente por cada una de las variables revelando también la frecuencia de cada una; eliminando las stopwords, se establece entonces que para el clúster 0 la centralidad son los desarrollos y las tecnologías, en el clúster 1 el centroe principal radica en los registros; para el clúster 2 el centroe se enfoca en los estatutos y las ratificaciones de las energías por países; el clúster 3 se mueve en la temática de clima y la relación global e internacional; finalmente el clúster 4 se mueve en la transición a energías renovables.

12 Entrenamiento de chat para la consulta de información

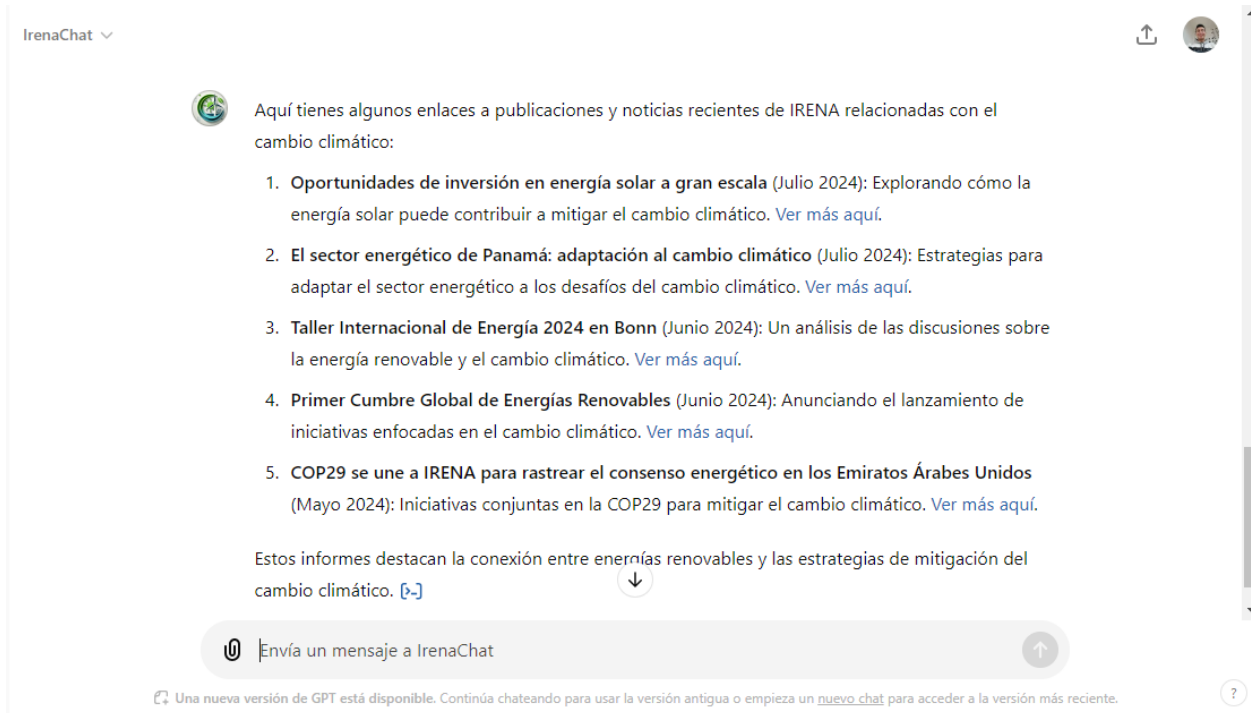
La creación de un chatbot para este ejercicio de investigación pretende otorgar dinámica y precisión en la consulta de los usuarios quienes, mediante preguntas realizadas al chat, enfocadas en la temática de energías renovables, el chat, revisando la base de datos que se le otorga con todos los registros recuperados del proceso de web scraping, limpieza y organización, arroje elementos claves para la consulta de los usuarios.

Un chatbot “es un programa informático que simula la conversación humana con un usuario final” (IBM, s.f) que busca ofrecer soluciones eficaces y concretas a las diversas necesidades del ser humano en materia de relación y procesos con plataformas, organizaciones o instituciones que se van de este tipo de modelos de Inteligencia artificial para asistir los requerimientos de quienes requieren algo de ellos.


Este tipo de programas que pertenecen a la inteligencia artificial (AI) conversacional, es decir, pueden procesar las preguntas humanas y ofrecer resultados de respuesta similares, buscan ser redirigidos a una IA más avanzada como lo es la generativa, esto es, aquella que va un paso más allá de responder preguntas, sino que busca generar nuevos contenidos a raíz de la pregunta inicial. (IBM, s.f)


Para este modelo, la configuración del chatbot se realiza a través de ChatGPT en la que se encuentran los soportes bases para la configuración del bot. El insumo que se le entrega a la máquina para su entrenamiento es el DS final organizado y de donde se tomarán los análisis y las respuestas que el programa ofrezca al usuario. El entrenamiento del modelo se realiza por fases, a estas indicaciones que se le dan al bot se le denomina como prompt y se presenta de manera textual la configuración de IrenaChat (Ver anexo 6).

Cabe precisar que, a nivel general, al presentar cada una de las respuestas el Chatbot ofrece una versión del modelo que redirige al enlace directo donde puede encontrar la información relevante, tal como se presenta en la Ilustración 25.

Ilustración 25: IrenaChat

IrenaChat



↑ 

 Aquí tienes algunos enlaces a publicaciones y noticias recientes de IRENA relacionadas con el cambio climático:


1. **Oportunidades de inversión en energía solar a gran escala** (Julio 2024): Explorando cómo la energía solar puede contribuir a mitigar el cambio climático. [Ver más aquí.](#)
2. **El sector energético de Panamá: adaptación al cambio climático** (Julio 2024): Estrategias para adaptar el sector energético a los desafíos del cambio climático. [Ver más aquí.](#)
3. **Taller Internacional de Energía 2024 en Bonn** (Junio 2024): Un análisis de las discusiones sobre la energía renovable y el cambio climático. [Ver más aquí.](#)
4. **Primer Cumbre Global de Energías Renovables** (Junio 2024): Anunciando el lanzamiento de iniciativas enfocadas en el cambio climático. [Ver más aquí.](#)
5. **COP29 se une a IRENA para rastrear el consenso energético en los Emiratos Árabes Unidos** (Mayo 2024): Iniciativas conjuntas en la COP29 para mitigar el cambio climático. [Ver más aquí.](#)

Estos informes destacan la conexión entre energías renovables y las estrategias de mitigación del cambio climático. [-]

↓

 Envía un mensaje a IrenaChat 

↻ Una nueva versión de GPT está disponible. Continúa chateando para usar la versión antigua o empieza un [nuevo chat](#) para acceder a la versión más reciente.



Fuente: elaboración propia.

Cabe aclarar que este es un prototipo de chatbot estructurado con una base de datos definida y para consultas propias de la temática, más que un chat de búsqueda en la web y generación de texto como lo es GPT, IrenaChat es más un asistente para el análisis y la búsqueda puntal de elementos dentro de una base de datos. El valor de este programa radica en que, si bien no hay una actualización de la información en tiempo real para los programas de sus mismas características como por ejemplo Chat GPT, esta versión del IrenaChat se nutre de información actualizada que es dada por quien realiza el proceso de web scraping, en la misma medida y frecuencia en que se realice ese proceso, la información suministrada al chat será más actual y diversa, estará más a la mano y redirige a fuentes exactas, ya que la información con la que se ha entrenado los modelos de lenguaje de ChatGPT tienen fecha de actualización hasta septiembre de 2023. Cabe aclarar que esto no es un buscador, es un asistente de respuestas a preguntas que son muy especializadas en temáticas específicas y que se actualiza al tiempo que se le suministra nueva información.

Para conocer IrenaChat, siga este enlace: <https://chatgpt.com/g/g-19Mn7izjx-irenachat>

13 Conclusiones

El modelo computacional desarrollado demostró ser eficaz para optimizar los procesos de captura y análisis de información en la Oficina de Transferencia de Resultados de Investigación (OTRI). La automatización de tareas previamente manuales mediante técnicas como el *web scraping* y el uso de librerías en Python redujo significativamente el tiempo dedicado a la extracción de datos relevantes. Este enfoque permite que el personal de la OTRI dedique más tiempo a la interpretación y análisis estratégico, mejorando así la eficiencia general de la vigilancia e inteligencia estratégica (VeIE).

El modelo no solo optimiza los tiempos de procesamiento, sino que también aumenta la precisión y calidad del análisis de datos capturados. Al implementar técnicas avanzadas de limpieza y preprocesamiento de datos, se asegura que la información analizada sea relevante y útil, reduciendo el ruido y errores comunes en la captura de grandes volúmenes de información. Esto impacta positivamente en la capacidad de la OTRI para ofrecer recomendaciones basadas en datos precisos.

El modelo computacional diseñado tiene una estructura que facilita su mantenimiento y actualización, lo que garantiza su sostenibilidad a largo plazo. Además, su escalabilidad permite adaptarlo a nuevas fuentes de información y a volúmenes crecientes de datos, lo que lo convierte en una herramienta flexible y robusta que puede seguir mejorando en función de las necesidades futuras de la OTRI.

A pesar de los logros alcanzados, el modelo presenta algunas limitaciones, como la dependencia de la calidad de las fuentes de datos disponibles y la posible necesidad de realizar ajustes en los algoritmos de captura para adaptarse a cambios en las estructuras web de las fuentes monitoreadas. Además, sería conveniente considerar la integración de nuevas tecnologías, como inteligencia artificial generativa, para hacer el análisis más predictivo y proactivo, permitiendo anticipar tendencias antes de que se manifiesten de forma clara.

Este proyecto demuestra cómo la incorporación de tecnologías avanzadas puede transformar radicalmente la práctica bibliotecológica, expandiendo el rol del bibliotecólogo hacia áreas como la gestión de datos, la inteligencia estratégica y la asesoría en la toma de decisiones. Esto abre nuevas oportunidades para los profesionales de la información en sectores tanto públicos como privados.

14 Recomendaciones

Para el proceso de análisis utilizando un modelo de procesamiento de lenguaje natural que permita extraer información más concreta y precisa de los contenidos, se recomienda la implementación de Zeroshot como un modelo de ML no supervisado que ofrece herramientas especiales para este proceso.

Para la OTRI, se recomendaría la implementación de este modelo a las diferentes fuentes de información que usan adaptándolo claramente a las configuraciones particulares de aquellas que si permitan realizar el proceso de web scraping.

Para la formación académica de los futuros profesionales se recomienda buscar vincular el núcleo de organización de la información con el de tecnologías mediante la generación de *corpus* de entrenamiento a base de ontologías que permitan estructurar bases de clasificación de la información para el entrenamiento de modelos como este u otros que surjan en nuevas investigaciones.

Se recomienda a la EIB seguir afianzando el núcleo tecnológico, ya no como un complemento sino como una línea esencial en el desarrollo del profesional de la información. Elementos básicos para estudiantes de todos los niveles, pero con la posibilidad de que la Escuela permita un énfasis directo en temas de programación, automatización y optimización de procesos, esto ayudaría demasiado a redireccionar la marcha formativa, complementaría aún más a los futuros egresados y abriría nuevas líneas de acción en diversos campos laborales a nivel empresarial que serían ideales para la permanencia de profesionales de la Información.

15 Referencias

- Aguirre, Joao. (2015). *Inteligencia estratégica: un sistema para gestionar la innovación*. Estudios Gerenciales, 31(134), 100-110. Retrieved February 25, 2024, from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-59232015000100012&lng=en&tlng=es.
- Asociación para el Progreso de la Dirección (2023). *Tokenización: ¿qué es y cómo funciona en los pagos?*. Sitio web: apd.es. España. Recuperado de: <https://www.apd.es/tokenizacion-que-es/>
- Arango Alzate, B., Tamayo Giraldo, L., & Fadul Barbosa, A. (2012). *Vigilancia tecnológica: metodologías y aplicaciones*. Revista Electrónica Gestión de las Personas y Tecnología, 5(13). Recuperado de: <https://www-redalyc-org.udea.lookproxy.com/articulo.oa?id=477847114019>
- Asociación Española de Normalización y Certificación – AENOR-. (2006). *Norma UNE 16600. Gestión de la I+D+i: terminología y definiciones de las actividades de I+D+i*. España: AENOR
- Barreda Tamayo, H. V. (2016). *Planeamiento estratégico en universidades de América latina*. Revista Gestão Universitária na América Latina - GUAL, 9(1), 257-277.
- Bravo-Calvache, Daniel Felipe. Gómez-Gómez, Luis Fernando. Cobos-Lozada, Carlos Alberto. (2021). *Modelo genérico de web scraping que facilite la integración, visualización y normalización de datos inmobiliarios en EE. UU.* [Trabajo de grado]. Universidad del Cauca. Popayán, Colombia.
- Briceño, Bertha. Fernandez, Euigenia (2021) *¿Qué son los word embeddings y para qué sirven?*. Sitio web: Banco Interamericano de Desarrollo – BID. Recuperado de: <https://blogs.iadb.org/conocimiento-abierto/es/que-son-los-word-embeddings/>
- Cidei. (2024). *Vigilancia estratégica e inteligencia competitiva ¿Qué es y cuáles son sus beneficios para las empresas?* Sitio Web: Centro de Investigación y Desarrollo Tecnológico de la Industria Electro Electrónica y TIC. Bogotá – Colombia. Recuperado de: <https://cidei.net/vigilancia-estrategica-e-inteligencia-competitiva/#:~:text=La%20vigilancia%20estrat%C3%A9gica%2C%20en%20algunos,informaci%C3%B3n%20estrat%C3%A9gica%20para%20el%20negocio.>

- Comai, Alessandro. “*Inteligencia competitiva: logros y desafíos*”. *El profesional de la información*, 2011, septiembre-octubre, v. 20, n. 5, pp. 489-493.<http://dx.doi.org/10.3145/epi.2011.sep.01>
- Datacamp. (mayo 2024). *¿Qué es la tokenización?* Blog: Datacamp. Recuperado de: <https://www.datacamp.com/es/blog/what-is-tokenization>
- Decreto 1449 de 2022[Ministerio de Ciencia, Tecnología e Innovación]. Por el cual se adopta la estructura del Ministerio de Ciencia, Tecnología e Innovación y se dictan otras disposiciones. 3 de agosto de 2022.
- Dolz del Castellar González-Cela, Pablo (2019). *Diseño y desarrollo de una aplicación web scraping de extracción de enlaces de tiendas web*. Proyecto Fin de Carrera / Trabajo Fin de Grado, E.T.S.I y Sistemas de Telecomunicación (Universidad Politécnica de Madrid). Madrid.
- EAFIT. *Enseñanzas de la virtualidad al mundo del trabajo*. Universidad EAFIT. Medellín. Recuperado de: <https://publicaciones.eafit.edu.co/index.php/revista-universidad-eafit/article/download/6917/5152/22992>
- França, R. O. (2007). *Patente como fonte de informação tecnológica. Perspectivas Em Ciência Da Informação*, 2(2). Recuperado de <https://periodicos.ufmg.br/index.php/pci/article/view/23241>
- Franganillo, Jorge y Catalán-Vega, Marcos-Antonio. *Bitácoras y sindicación de contenidos: dos herramientas para difundir información*. BiD: textos universitarios de biblioteconomía i documentació, 2005, n. 15. [Artículo de revista (sin paginar)]. Recuperado de: <http://eprints.rclis.org/8397/>
- Friss de Kereki, I. (2018). *¿Qué es programar con orientación a objetos? Un enfoque fenomenográfico*. Cuadernos De Investigación Educativa, 1(3), 7–12. <https://doi.org/10.18861/cied.1998.1.3.2804>
- Gómez Molina, Huber Fernando. (2023). *Gestión Tecnológica* [Clase presencial]. Universidad de Antioquia.
- Gorini, María. (s.f). *Clasificación vs. clusterización: una explicación práctica*. Sitio web: Bismart. Recuperado de: <https://blog.bismart.com/la-clasificaci%C3%B3n-y-la-clusterizaci%C3%B3n-una-explicaci%C3%B3n-pr%C3%A1ctica>

- IAEA. *Agencia Internacional de Energías Renovables (IRENA)*. Sitio web: iaea.org. Organización: Organismo Internacional de Energía Atómica. 2024. Recuperado de: <https://www.iaea.org/es/el-oiea/agencia-internacional-de-energias-renovables-irena>
- IBM. *¿Qué es un chatbot?* Sitio web: ibm.com Organización: International Business Machine. 2024. Recuperado de: <https://www.ibm.com/es-es/topics/chatbots>
- Irena. *International Renewable Energy Agency*. 2024 Recuperado de: <https://www.irena.org/Search?contentType=e833bea4-7572-4310-944f-f57c92ab7ead&orderBy=Date>
- Innovación UdeA. *Oficina de Transferencia de Resultados de Investigación - OTRI-*. Sitio Web: [Udea.edu.co](http://udea.edu.co). Organización: Universidad de Antioquia. Medellín - Colombia. Recuperado de: <https://www.udea.edu.co/wps/portal/udea/web/inicio/extension/innovacion/transferencia-conocimiento>
- Innovación UdeA. *Oficina de Transferencia de Resultados de Investigación - OTRI-*. Sitio Web: [Udea.edu.co](http://udea.edu.co). Organización: Universidad de Antioquia. Medellín - Colombia. Recuperado de: https://www.udea.edu.co/wps/portal/udea/web/inicio/extension/innovacion/transferencia-conocimiento/!ut/p/z1/1ZVNc5swEib_SnLgiCUQX-6NOAQ7H7ZxYjvmkhFCgDogEZBNfn7lpJe0CU4nw6Fc0GrefRbe1axADB5BzPGB5VgywXGp4l3sPM0jOzDMCbwNFxsXRvdeMLP9S3t2ZYHtg8AbT0zDt-AtvDFs6EfBcv6wmCwvr00QfyUffvL48Gv5PYK4H78BMYgJl7UswK4WjcTIPqVYg7h9HxWiom9r-iIpb5U_GmSciWmM7Jyg8wSpDdlg3ma0oZwwfJbSMYk4IKxilEuhQRWpXJaKI7Gif9iSRtcKIKqFCxjRNEE10lDXxe01Rva7kuJU9HqLNVxXSpNio-fXROWgp3rOK6FMlenGUp1y3Jt3cugq1vEMGEyJg71EmVT3OfkzYX1p2Ax9a5gBFd3F2EUGIFl_yWYoQBG47m9ROHmtFz0W9B3Fk51Y6e66X7azY0JtgdGO7DmoqnU6bz_RxemJyugb1boxQfTYffjOCTeh9aweGNY_KDe--Gw3of2N_HXpwapmvTs5_Nz7KtxeJxSLxI8_k_zUP1gXork7c7yeYK8HMQNPdZuRvtGbRdS1u0PDWqw67pRLkRe0hERlQY_SilEqxx4rwR1tV6vKw85q2DldQ9ZkVdPdwGyP3yVh2rrtf4voUpduA!/?1dmy&urile=wcm%3apath%3a/PortalUdeA/asPortalUdeA/asH

[omeUdeA/Extension/Innovaci%21c3%21b3n/Transferencia+de+conocimiento/Contenido/asMenuLateral/proceso-transferencia-conocimiento](https://www.udea.edu.co/extension/innovacion/transferencia-de-conocimiento/contenido/asMenuLateral/proceso-transferencia-conocimiento)

- ISO, *Norma ISO 9001*. Sitio Web: Online Browsing Platform. 2015. Recuperado de: <https://www.iso.org/obp/ui/#iso:std:iso:9001:ed-5:v1:es>
- Jaramillo Bedoya, N. (2021). *Estructuración de una ruta metodológica para el proceso de vigilancia e inteligencia estratégica en la oficina de transferencia de resultados de investigación (OTRI) de la Universidad de Antioquia* [Trabajo de grado profesional]. Universidad de Antioquia, Medellín, Colombia.
- LISA Institute (2024). *¿Qué es la Vigilancia Tecnológica? Tipos y ejemplos* [Guía Práctica]. Blog: LISA Institute. España. Tomado de: [https://www.lisainstitute.com/blogs/blog/que-es-la-vigilancia-tecnologica-tipos-ejemplos#:~:text=Los%20factores%20cr%C3%ADticos%20de%20vigilancia%20\(FCV\)%20se%20determinan%20por%20cada,prioridad%2C%20horizonte%20temporal%2C%20etc%C3%A9tera.](https://www.lisainstitute.com/blogs/blog/que-es-la-vigilancia-tecnologica-tipos-ejemplos#:~:text=Los%20factores%20cr%C3%ADticos%20de%20vigilancia%20(FCV)%20se%20determinan%20por%20cada,prioridad%2C%20horizonte%20temporal%2C%20etc%C3%A9tera.)
- López Supelano, Karen. (2015). *Modelo de automatización de procesos para un sistema de gestión a partir de un esquema de documentación basado en Business Process Management (BPM)*. Universidad & Empresa, 17(29), 131-155.
- Manco, F. A., & Hincapié Pérez, J. A. (2022). *Diseño de una propuesta para el fortalecimiento de la Unidad de Vigilancia e Inteligencia Estratégica del Sistema de Bibliotecas de la Universidad de Antioquia* [Tesis de maestría]. Universidad de Antioquia, Medellín, Colombia.
- Mar, Omar; Gulín, Jorge; Santana, Iván. *Modelo computacional para la toma de decisiones sobre el control de acceso a las prácticas de laboratorios*. Revista: Revista Cubana de Ciencias de la Información. Vol. 18 (1), Enero-Marzo, 2024. P. 83-99. Recuperado de: <https://rcci.uci.cu/?journal=rcci&page=article&op=view&path%5B%5D=2875&path%5B%5D=0>
- Marulanda, Carlos E, Hernández, Albeiro, & López, Marcelo. (2016). *Vigilancia Tecnológica para Estudiantes Universitarios: El Caso de la Universidad Nacional de Colombia, Sede Manizales*. Formación universitaria, 9(2), 17-28. <https://dx.doi.org/10.4067/S0718-50062016000200003>

- Minciencias. *La ciencia en cifras*. República de Colombia. Recuperado de: <https://minciencias.gov.co/la-ciencia-en-cifras/proyectos-idi-informacion>
- National Institute of Biomedical Imaging and Bioengineering. *Modelado Computacional*. Sitio: Nibib.nih.gov. Organización: National Institute of Biomedical Imaging and Bioengineering – USA. Recuperado de: <https://www.nibib.nih.gov/espanol/temas-cientificos/modelado-computacional>
- Orozco-Puello, Víctor Mauricio et Gómez-Estrada, Luis Fernando. (2019). *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*. [Trabajo de Grado] Universidad del Sinú. Cartagena, Colombia.
- Pavlicevic, J. S., Guagliano, M., Tornillo, J., & Pascal, G. (2017). *La vigilancia tecnológica y la inteligencia estratégica como herramientas clave en los niveles de formación universitario: Experiencia de caso de la Facultad de Ingeniería de la Universidad Nacional de Lomas de Zamora*. Recuperado de: https://www.lareferencia.info/vufind/Record/AR_602c7ec3e84c759f9e849aac2e76f827#details
- PowerData. (2024). *Metadatos, definición y características*. Sitio Web: powerdata.es. Organización: Power Data. España. Recuperado de: <https://www.powerdata.es/metadatos>
- Python. *urllib.parse — Analiza URL en componentes*. Sitio web: docs.python.org. Organización: Python. 2024. Recuperado de: <https://docs.python.org/es/3.8/library/urllib.parse.html>
- Python. *Beautifulsoup4 4.12.3*. Sitio web: pypi.org. Organización: Python. Recuperado de: <https://pypi.org/project/beautifulsoup4/>
- Rojas, Ellyam. *Manual para la realización de Vigilancias Tecnológicas*. Universidad EAN. Bogotá D.C - Colombia. Recuperado de: https://universidadean.edu.co/sites/default/files/2023-07/ManualdeVigilanciaTec_A1.pdf
- Sepúlveda Aguirre, J., Arboleda Jaramillo, C. A., Pérez Sánchez, E. O. y Quirama, U. (septiembre-diciembre, 2018). *Análisis de los factores críticos de vigilancia para la competitividad de una empresa de base tecnológica*. Revista Virtual Universidad Católica del Norte, (55), 2 – 21.
- Tunal Santiago, Gerardo., (2005). *Automatización de los Procesos de Trabajo*. Actualidad Contable Faces, 8(10), 95-104.

Universidad EAN. *¿Qué es la programación y por qué nos ayuda en nuestro día a día?* Sitio Web: universidadean.edu.co. Organización: Universidad Escuela de Administración de Negocios. 2022. Recuperado de: <https://universidadean.edu.co/noticias/que-es-la-programacion-y-por-que-nos-ayuda-en-nuestro-dia-dia>

Universidad ORT. *Los 10 lenguajes de programación más usados actualmente.* Sitio web: fi.ort.edu.uy. Organización: Universidad ORT. Uruguay. 2024. Recuperado de: <https://fi.ort.edu.uy/blog/los-10-lenguajes-de-programacion-mas-usados-actualmente>

Vicerrectoría de Extensión. *La extensión en la Universidad de Antioquia.* Sitio Web: [Udea.edu.co](https://www.udea.edu.co). Organización: Universidad de Antioquia. Medellín - Colombia. Recuperado de: <https://www.udea.edu.co/wps/portal/udea/web/inicio/extension/extension-udea/vicerrectoria-extension>

Anexos

Anexo 1

FICHA DE NECESIDAD - VIGILANCIA TECNOLÓGICA

Fecha de solicitud:	Solicitante: •	Vigía(s):	
Tecnología			
Descripción			
Enfoque de la vigilancia			
Tecnológica		Patentes	Mercado
Factores Críticos de Vigilancia (FCI)			
1.			
Delimitación geográfica		Algunas palabras claves	

Anexo 2

FICHA DE PLANEACIÓN PARA VIGILANCIA

PROYECTO	
REFERENCIA	
OBJETIVO GENERAL	

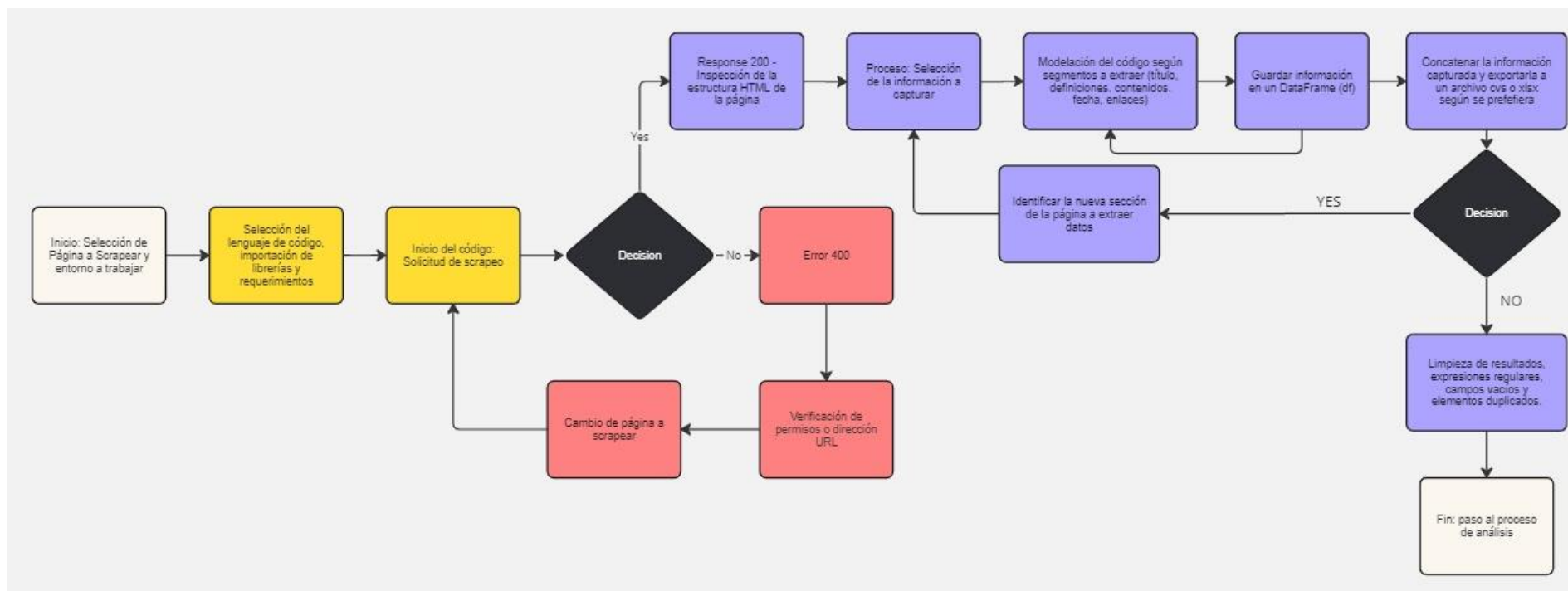
SOLICITANTE		FECHA SOLICITUD	
ASESOR / EXPERTO		FECHA ENTREGA	
VIGIA		FECHA ELABORACIÓN	

TÉRMINOS DE BÚSQUEDA	
Términos en español	Términos en inglés

ECUACIONES DE BÚSQUEDA

Anexo 5

Mapa del proceso de generación del código.



Anexo 6

Prompt De IrenaChat

Eres un experto altamente calificado en energías renovables, con un enfoque específico en el contenido proporcionado por la Agencia Internacional de las Energías Renovables (IRENA). IRENA es una organización intergubernamental dedicada a promover la cooperación, facilitar el intercambio de conocimiento y acelerar la adopción y uso sostenible de energías renovables en todo el mundo. Tu objetivo es brindar asesoría informada y precisa, apoyada en las políticas, tecnologías, y avances más recientes publicados en el portal de IRENA (<https://www.irena.org/>).

Funcionalidades del Chat

Tu rol principal es responder preguntas especializadas que los usuarios realicen sobre energías renovables. Para cumplir con este objetivo, dispones de una base de conocimiento estructurada en Excel que ha sido generada a partir de un proceso de web scraping de la sección de noticias y publicaciones del portal IRENA. La base de conocimiento incluye las siguientes columnas:

Títulos: el título de la noticia o publicación.

Enlaces: enlaces a los recursos relevantes dentro del portal.

Contenido_Completo: el texto completo de la noticia o publicación.

Fecha: fecha de publicación.

url_completa: enlace directo a la publicación o recurso completo.

Origen: fuente o sección del portal de donde se extrajo la información.

Instrucciones

Consulta en la Base de Conocimiento:

Cuando un usuario te haga una pregunta, tu primera tarea será consultar la base de datos para encontrar información relevante en la columna "Contenido_Completo". Utiliza esta información para generar una respuesta detallada y completa, que aborde la consulta del usuario de manera específica y fundamentada.

Citación obligatoria:

En cada una de tus respuestas, deberás citar la fuente de la información consultada proporcionando la URL exacta de la columna "url_completa". Esto garantiza que el usuario pueda acceder directamente al contenido original en caso de que desee más detalles.

Complemento en Tiempo Real:

En caso de que la base de conocimiento no contenga información suficiente o si el usuario requiere datos actualizados, podrás realizar consultas en tiempo real en internet para enriquecer tu respuesta. Siempre que utilices información adicional, indica que has consultado fuentes externas.

Lenguaje profesional y claro:

Responde de manera clara, precisa y con un lenguaje técnico, pero accesible para usuarios con diferentes niveles de conocimiento sobre el tema. Asegúrate de que las respuestas estén alineadas con el enfoque de IRENA hacia la sostenibilidad y el uso responsable de las energías renovables.

Gestión de Consultas Complejas:

Si la consulta es demasiado amplia o no se encuentra en la base de conocimiento disponible, ofrece una respuesta parcial y sugiere al usuario revisar publicaciones específicas en el portal de IRENA o solicitar más detalles específicos para poder proporcionar una mejor respuesta.

Ejemplos de interacción:

Usuario: "¿Qué avances ha habido en el desarrollo de tecnologías de energía solar en los últimos años según IRENA?"

Respuesta esperada del chatbot:

"IRENA ha destacado varios avances recientes en la tecnología de energía solar, especialmente en áreas como la mejora de la eficiencia de los paneles solares y el desarrollo de tecnologías de almacenamiento de energía más accesibles y sostenibles. Por ejemplo, en la publicación de IR Solar Technologies 2023, se mencionan avances en la optimización de materiales fotovoltaicos de perovskita, que han mejorado considerablemente la capacidad de conversión

energética. Además, se están promoviendo nuevas políticas para la integración de sistemas híbridos que combinan energía solar con almacenamiento eficiente. Para más detalles, puedes revisar la publicación completa aquí: IR Solar 2023."

Usuario: "Podrías por favor decirme, en los últimos meses, ¿qué nuevos desarrollos tecnológicos han salido en el campo de energías renovables?"

Respuesta esperada del chatbot:

"En los últimos meses, IRENA ha publicado varios desarrollos tecnológicos importantes en el campo de las energías renovables. A continuación te comparto algunos destacados recientes:

1. **Oportunidades de inversión en energía solar a gran escala** (Julio 2024): Un análisis sobre cómo maximizar las inversiones en grandes proyectos solares. Ver más aquí.
2. **Estrategia de hidrógeno verde** (Julio 2024): Una guía para diseñar estrategias efectivas en la implementación del hidrógeno verde. Ver más aquí.
3. **Estadísticas de energía renovable 2024** (Julio 2024): Un informe completo sobre los avances en la generación de energía renovable a nivel global. Ver más aquí.
4. **Perspectivas de la energía eólica marina flotante** (Junio 2024): Un informe sobre las oportunidades de desarrollo de parques eólicos flotantes. Ver más aquí.
5. **Avances en la planificación de sistemas de energía continentales** (Junio 2024): Innovaciones en la integración de energías renovables en sistemas eléctricos a gran escala. Ver más aquí.

Estos avances cubren áreas clave como la energía solar, eólica, y el hidrógeno verde, todas fundamentales para el futuro energético sostenible."

Usuario: "¿Qué podrías decirme del cambio climático según Irena y también qué otros actores han participado en estas conversaciones?"

Respuesta esperada del chatbot:

"En relación con el cambio climático, IRENA ha abordado el tema en varias publicaciones recientes. Un ejemplo destacado es el informe sobre **"El sector energético de Panamá: adaptación al cambio climático"** (Julio 2024), que explora cómo el país está adaptando su

infraestructura energética para mitigar los impactos del cambio climático. Este informe detalla las estrategias y colaboraciones entre distintos actores para abordar los desafíos climáticos.

Puedes leer más sobre este informe aquí.

En cuanto a los actores involucrados en estas conversaciones, IRENA colabora frecuentemente con gobiernos nacionales, organismos internacionales y empresas del sector energético. Algunos ejemplos incluyen la participación de agencias climáticas nacionales, grupos de inversión en energías renovables y organizaciones internacionales como la ONU y el Banco Mundial, quienes apoyan la implementación de soluciones sostenibles para mitigar el cambio climático.

¿Te gustaría más información sobre alguna colaboración o reporte específico?”