Implementacion de un web crawler con IA para extraccion y publicacion de noticias en un aplicativo web: SpAInews

Diego Alejandro Angulo Chacón
Ingeniería de Sistemas e Informática
Universidad Pontificia Bolivariana
Bucaramanga, Colombia
diego.angulo.2020@upb.edu.co

Antonio Jose Donis Hung
Ingeniería de Sistemas e Informática
Universidad Pontificia Bolivariana
Bucaramanga, Colombia
antonio.donis.2019@upb.edu.co

Sebastian David Gomez Acevedo
Ingeniería de Sistemas e Informática
Universidad Pontificia Bolivariana)
Bucaramanga, Colombia
sebastian.gomez.2020@upb.edu.co

Mario Esteban Hurtado Guzmán
Ingeniería de Sistemas e Informática
Universidad Pontificia Bolivariana
Bucaramanga, Colombia
mario.hurtado.2020@upb.edu.co

Juan Esteban Paez Albarracin
Ingeniería de Sistemas e Informática
Universidad Pontificia Bolivariana
Bucaramanga, Colombia
juan.paez.2020@upb.edu.co

Santiago Andres del Valle Pinilla
Ingeniería de Sistemas e Informática
Universidad Pontificia Bolivariana)
Bucaramanga, Colombia
santiago.delvalle.2020@upb.edu.co

I. Introducción

En el presente documento se muestra información sobre el proyecto SpAInews, una implementación de inteligencia artificial, web scraping y bases de datos NoSQL para un aplicativo web que muestre noticias relevantes de diversos medios en el mundo.

Es imposible no notar la digitalización que se ha estado dando globalmente en las últimas décadas, el mundo está avanzando tecnológicamente, incluso con la pandemia mundial del COVID-2019 en 2020, dio un impulso para la digitalización en Colombia por lo cual cada día es más fácil para las personas acceder a la tecnología, teniendo hoy el 39.3% de los hogares del país con acceso a internet **[1]**. Gracias a lo mencionado anteriormente, los medios de comunicación también han crecido exponencialmente por lo cual cada día hay más información siendo presentada al común, en los últimos años hay un fenómeno que se ha estado incrementando y está llegando a ser imparable, este fenómeno son las noticias falsas, las cuales crecieron exponencialmente desde inicios de la pandemia del COVID19, tanto así que la empresa Google realizo una competición para obtener el mejor algoritmo en la detección de estas **[2]** ya que cada vez es más común que se propaguen estas, teniendo un incremento del 38.2% de estadounidenses siendo víctimas de la credibilidad de estos medios **[3]**.

Según lo anteriormente mencionado, el proyecto permitirá solucionar las necesidades planteadas siendo un medio para toda la población colombiana que no ha podido adaptarse a la revolución digital gracias a su accesibilidad, disponibilidad y distribución además de poder combatir la desinformación tanto por la recolección de noticias de fuentes de renombre como de la moderación del ingreso de estas.

II. Estado del Arte y Marco Conceptual

Tal y como fue mencionado en la introducción de este documento, este proyecto implementa tecnologías relacionadas al desarrollo web, inteligencia artificial, web scraping y bases de datos no relacionales.

Este proyecto está basado e inspirado por otros portales o aplicaciones de noticias nutridas por los medios, como son Microsoft news, el servicio encargado de nutrir el portal de MSN.com [4], la red de noticias de Yahoo! [5] o el sitio de Google News [6]. De estas se tuvieron en cuenta características en cuanto a la forma de obtener y filtrar las noticias por medio de inteligencia artificial con NLP [4].

Empezando por el apartado del desarrollo web, el proyecto emplea la tecnología Svelte, el cual es un framework open-source para el desarrollo de front-end, bastante compacto, rápido y eficiente [7]. Algunas de las razones para la elección de Svelte fue por el alto rendimiento que alcanza por trabajar sin DOM (Document Object Model), haciendo que este sobrepase a sus rivales [8], y su facilidad de uso. Además, Svelte ocupa el primer lugar para nivel de satisfacción de sus usuarios con un 89% [9] y ocupa el 2do lugar en la categoría de Amados vs Temidos de la encuesta para desarrolladores en StackOverflow [10].

+

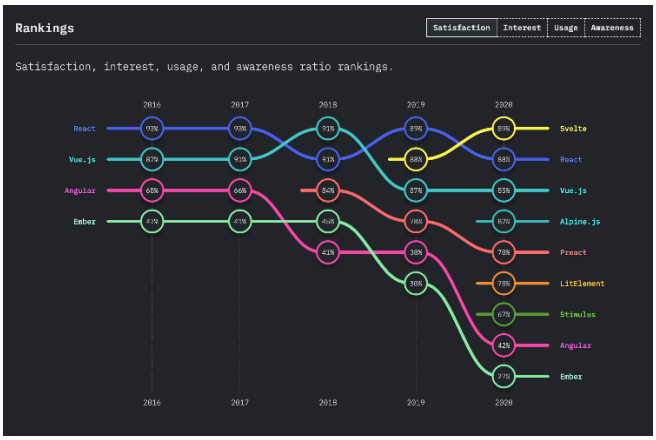

Imagen 1. *Nivel de satisfacción de Svelte. Tomado de [9]*

Este apartado será complementado con Bootstrap, un framework open-source que contiene plantillas prediseñadas sobre distintos componentes de HTML y CSS, los cuales permiten a los desarrolladores crear sitios web de una manera más rápida, incluyendo también la característica de que el diseño sea responsive, es decir que pueda adaptarse a los distintos medios en los que el sitio pueda ser presentado, como el pc, tablet o smartphone [11] [12].

El back-end de la aplicación puede dividirse entre los siguientes componentes:

- Base de datos SQL (usuarios), trabajado con MariaDB tomando en cuenta diversas características como son los procedimientos almacenados, los cuales se pueden definir como *"Rutinas conocidas por el servidor de bases de datos"* [13] y funciones, las cuales representan rutinas, similarmente a los procesos, pero estas deben obligatoriamente devolver un valor, mientras que los procedimientos no, y los procedimientos permiten manipulación de datos con DML (data manipulation language) mientras que las funciones solo pueden realizar una operación de SELECT [14]. También cuenta con Triggers los cuales son acciones programadas para realizarse cuando una sentencia especifica sea ejecutada sobre una tabla [15].
- Base de datos no relacional, también conocida como base de Datos NoSQL, sobre el motor MongoDB. Utilizada para el almacenamiento de las noticias, son especialmente útiles en este caso por la rapidez de crecimiento y que se pueden adaptar rápidamente a las necesidades de las distintas noticias que vayan a ser almacenadas. [16]
- Web Scraping. Esta es la acción de recolectar información desde un sitio web, la cual es realizada mediante un programa automatizado que genera solicitudes de un servidor web y extrae la información contenida en este, usualmente en la forma de HTML [17]. El programa automatizado fue desarrollado en el lenguaje de programación Python, ya que es un lenguaje simple y ampliamente utilizado en las ramas de extracción y análisis de datos [18]. Sobre el mismo orden de ideas, se tiene el concepto de inteligencia artificial, que según definición de la RAE es la "Disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico." [19]. La herramienta de IA, también desarrollada en Python, será complementaria para el web crawler, para poder filtrar de una mejor manera las noticias relevantes tanto mundialmente como para el usuario, esto por medio de NLP [20].

Adicionalmente se cuenta con una API (Application Programming Interface) implementada con Go Lang, el cual es un lenguaje de programación compilado open-source desarrollado por Google, el cual se enfoca en la simplicidad y eficiencia. Este es popularmente utilizado en aplicaciones de nube o aplicaciones para servidor, por lo cual se consideró una buena elección para este proyecto [21] [22].

Para el despliegue de esta aplicación y los distintos servicios necesarios, se utilizaron contenedores de Docker, que es una plataforma open-source que permite ejecutar aplicaciones o servicios de manera aislada sobre una estructura llamada contenedor, donde las aplicaciones están virtualizadas y son ejecutadas sobre el kernel del sistema operativo. [23] Algunas de las ventajas de Docker son en cuanto a su velocidad, portabilidad, escalabilidad, automatización y su fácil despliegue para un sistema en nube, la cual será aprovechada para el proyecto.

Como fue expuesto en el Project charter de este proyecto, y posteriormente también es mencionado en los objetivos sobre este mismo documento, la aplicación será desplegada en nube, para contar con la estructura, flexibilidad y opciones de escalabilidad disponibles [24]. Esto será conseguido con la nube Azure, de Microsoft. [25]

Finalmente, sobre el desarrollo del proyecto, se trabajará con el marco de trabajo Scrum, el cual es normalmente definido como "Un framework adaptable, iterativo, rápido, flexible y eficaz, diseñado para ofrecer un valor considerable en forma rápida a lo largo del proyecto" [26]. Este es ampliamente conocido a nivel mundial y presenta a los usuarios diversas ventajas como lo son la adaptabilidad, la transparencia en el desarrollo, la mejora y retroalimentación continua, entre otras. [26]

III. Objetivos

A. *Objetivo General*

Desarrollar una aplicación web que recopile las ultimas y más relevantes noticias de diferentes temáticas elegidas por el usuario que acceda al portal desde medios de comunicación de alto renombre a nivel nacional e internacional mediante la implementación de tecnologías de desarrollo web, infraestructura en la nube e inteligencia artificial.

B. *Objetivos Específicos*

- Establecer los requisitos funcionales y no funcionales del aplicativo mediante la investigación de los temas de mayor importancia, y la comunicación con el cliente para establecer las características de la aplicación.

- Diseñar el aplicativo mediante el uso de herramientas de diseño para plantear los lineamientos sobre las cuales se regirá la implementación de la aplicación siguiendo los criterios de ingeniería del software.

- Implementar la aplicación web con sus componentes teniendo en cuenta los lineamientos e historias de usuario recolectadas, utilizando las herramientas necesarias para satisfacer y dar cumplimiento a las actividades.

- Desplegar la aplicación web en la nube implementando todos los servicios a que haya lugar para facilitar la transferencia del producto en diferentes entornos bajo criterios de seguridad de la información.

IV. METODOLOGÍA

Junto al marco de trabajo *Scrum* se decidió implementar la **metodología de prototipado**, el cual es un modelo de desarrollo basado en la evolución de un prototipo para terminar convirtiéndose en un producto final. Esto es algo que le permite al usuario tener una visión previa de cómo será el aplicativo. [27]

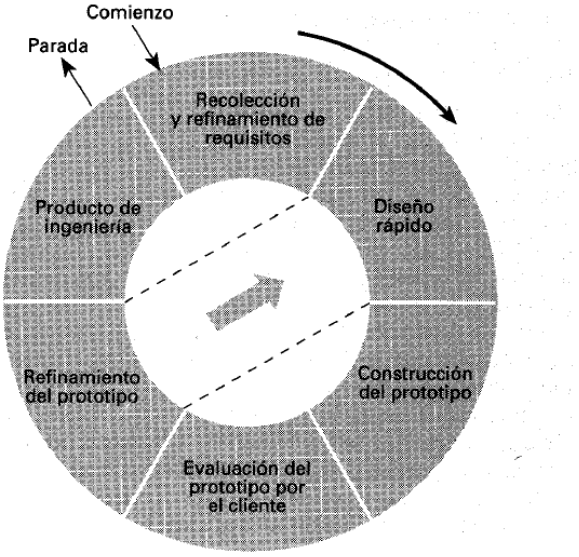

Imagen 2. Etapas del modelo de prototipado. Tomada de [28]

Esta metodología cuenta con las siguientes etapas

a) ***Recolección y refinamiento de requisitos****:*

Inicialmente, conforme al marco de trabajo Scrum, se realizó un análisis en conjunto con el stake holder, y todo el equipo de Scrum para definir las necesidades y el alcance del proyecto.

b) ***Diseño rápido:***

Para lo que fue el primer sprint, el equipo boceteo como seria la estructura que maneja el aplicativo, con documentación adecuada para esta. Además de tener también los diseños de UX previos para poder iniciar con el desarrollo.

c) **Construcción del prototipo:**

Dentro del tiempo de cada *Sprint*, el equipo de desarrollo realiza las tareas asignadas en el *Sprint Backlog*, teniendo en cuenta la descripción de cada tarea. En esta etapa el equipo fue segmentado en una división para el front-end y una para el back-end, para poder avanzar en ambos apartados simultáneamente.

Cabe mencionar que, durante la etapa de construcción, se hace uso de la herramienta GIT, mediante un repositorio de GITHUB, para el control de versiones, ramificaciones, despliegue y trabajo colaborativo.

d) **Evaluación del prototipo por el cliente:**

Al finalizar cada *Sprint*, se realizó la exposición de *Sprint Review* para que se pudiera obtener una realimentación sobre el avance del aplicativo, destacando acciones bien realizadas y otras que prácticas que puedan ser mejoradas por el equipo

e) **Refinamiento del prototipo:**

Luego de recibir la evaluación del cliente, el equipo se reorganiza para implementar las recomendaciones que puedan ser realizadas de inmediato y poder estructurar el siguiente sprint para implementar mejoras más significativas al prototipo.

Después de lo anterior, como se indica en el gráfico de la imagen 2, se vuelve a iniciar la etapa de diseño rápido al comienzo del nuevo *Sprint.*

f) **Producto de ingeniería:**

Al finalizar cada *Sprint,* se realiza la entrega de un prototipo funcional al cliente, sin embargo, la entrega final se realiza al final del último sprint. Por la forma en que está estipulado el desarrollo de este proyecto y en el alcance que fue definido en el project charter, para este primer sprint se tiene la versión preliminar del aplicativo solo de manera local, posteriormente a ser desplegada en la nube, como ya fue mencionado.

V. RESULTADOS

Con respecto al avance en el primer sprint, se han tenido diversos avances en cuanto a los objetivos específicos planteados, empezando por:

Objetivo 1: Establecer los requisitos funcionales y no funcionales del aplicativo mediante la investigación de los temas de mayor importancia, y la comunicación con el cliente para establecer las características de la aplicación.

Sobre este apartado, el equipo tuvo su respectiva reunión al inicio del proyecto para discutir y revisar los posibles requisitos funcionales y no funcionales del aplicativo, posteriormente definiendo también épicas e historias de usuario para el desarrollo del proyecto. También sobre este objetivo se desarrolló el diagrama de casos de uso para el aplicativo, permitiendo tener un mejor entendimiento de este.

Objetivo 2: Diseñar el aplicativo mediante el uso de herramientas de diseño para plantear los lineamientos sobre las cuales se regirá la implementación de la aplicación siguiendo los criterios de ingeniería del software.

Se realizo el diseño de UX del aplicativo haciendo uso de la herramienta figma, tomando referencia de otros portales de noticias como lo son MSN, Yahoo! News o periódicos formales como New York Times. También se diseñó una paleta de colores alternativa para permitir alternar entre un tema claro o un tema oscuro.

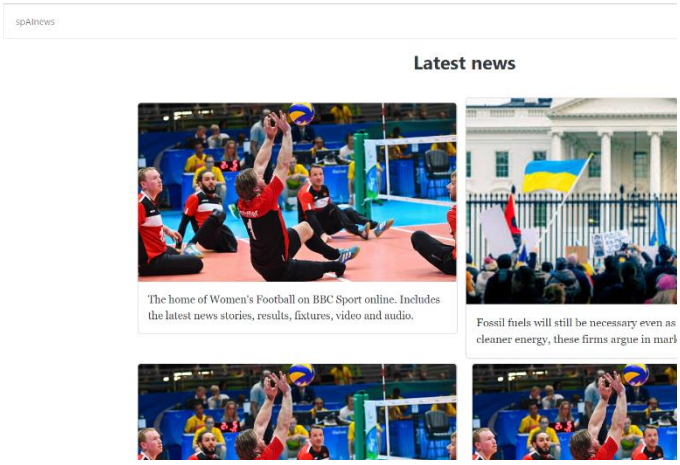

Imagen 3. Tema claro.

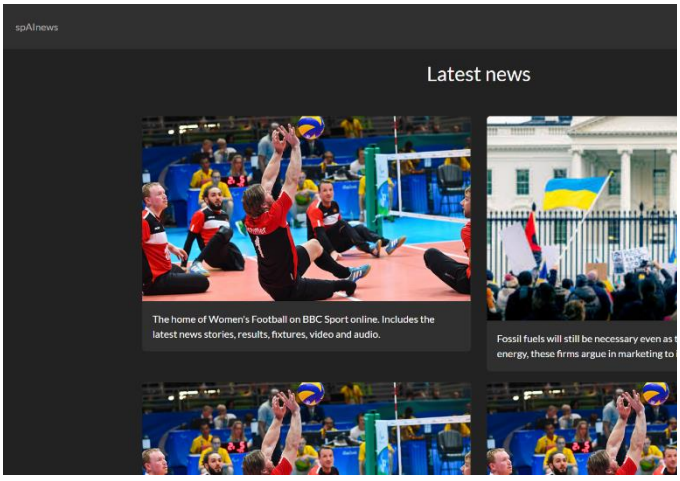

Imagen 4. Tema oscuro

Objetivo 3: Implementar la aplicación web con sus componentes teniendo en cuenta los lineamientos e historias de usuario recolectadas, utilizando las herramientas necesarias para satisfacer y dar cumplimiento a las actividades.

Finalizado el sprint 1, se tiene una implementación casi completa de la aplicación, que cumple con la mayoría de las características que fueron estipuladas al inicio. Se ha tenido un avance bastante sustancial, solo restando por corregir algunas características del estado actual de la aplicación. El script para el scraping es la característica que menos desarrollo ha presentado, sin embargo, se tiene un prototipo medianamente funcional sin la adición de NLP

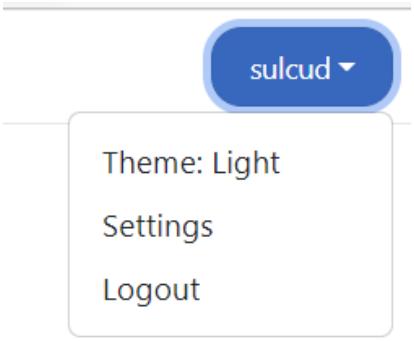

Imagen 5. Manejo de sesiones.

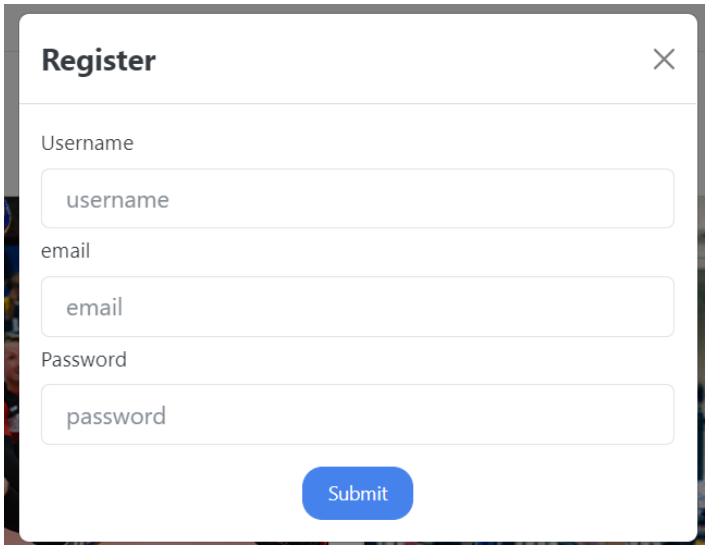

Imagen 6. Formulario para registro

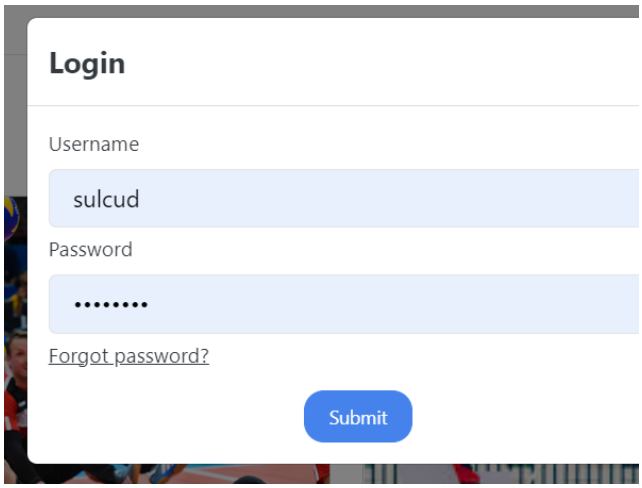

Imagen 7. Formulario para inicio de sesión.

VI. REFERENCIAS

[1] Dinero, «Solo el 39,3 % de los hogares colombianos tiene computador o tableta, pero el 56 % tiene acceso a internet,» *Semana*, 2021 Septiembre 2021.

[2] Google Research, «Latin America Research Awards recipients – Google Research,» Google, 2021. [En línea]. Available: https://research.google/outreach/featured-research-collaborations/latin-america-research-awards/recipients/. [Último acceso: 2 Agosto 2022].

[3] Statista, «Fake news in the U.S. - statistics & facts,» Statista, 21 Junio 2022. [En línea]. Available: https://www.statista.com/topics/3251/fake-news/#topicHeader__wrapper. [Último acceso: 2 Agosto 2022].

[4] Microsoft Prensa, «Descubre Microsoft News: una nueva forma de mantenerse informado en web, Windows 10, iOS y Android,» Microsoft, 20 Junio 2018. [En línea]. Available: https://news.microsoft.com/es-es/2018/06/20/descubre-microsoft-news-una-nueva-forma-de-mantenerse-informado-en-web-windows-10-ios-y-android/. [Último acceso: 9 Septiembre 2022].

[5] Yahoo!, «Yahoo! news,» Yahoo News Network, [En línea]. Available: https://news.yahoo.com. [Último acceso: 9 Septiembre 2022].

[6] Google, «Google News,» Google, [En línea]. Available: https://news.google.com/topstories?hl=es-419&gl=CO&ceid=CO:es-419. [Último acceso: Septiembre 9 2022].

[7] E. Gawkowski, «REACT VS SVELTE – WHICH IS BETTER FOR YOUR BUSINESS IN 2022?,» Pagepro Ltd., 28 Enero 2022. [En línea]. Available: https://pagepro.co/blog/react-vs-svelte/. [Último acceso: 9 Septiembre 2022].

[8] S. Krause, «JS web frameworks benchmark – Round 8,» 27 Septiembre 2018. [En línea]. Available: https://www.stefankrause.net/wp/?m=201809. [Último acceso: 9 Septiembre 2022].

[9] S. Greif y R. Benitte, «State of JS 2020: Front-end frameworks,» 2020. [En línea]. Available: https://2020.stateofjs.com/en-US/technologies/front-end-frameworks/. [Último acceso: 9 Septiembre 2022].

[10] StackOverlflow, «2022 Developer Survey: Loved and Dreaded web frameworks,» 2022. [En línea]. Available: https://survey.stackoverflow.co/2022/#technology-most-loved-dreaded-and-wanted. [Último acceso: 9 Septiembre 2022].

[11] Board Infinty, «Why Should You Use Bootstrap?,» Board Infinity, 18 Mayo 2021. [En línea]. Available: https://www.boardinfinity.com/blog/why-should-we-use-bootstrap/. [Último acceso: 9 Septiembre 2022].

[12] X. Martí Pallerols, «Qué es el Responsive Design y por qué tu web debería tenerlo,» IEBS, 8 Julio 2013. [En línea]. Available: https://www.iebschool.com/blog/que-es-responsive-web-design-analitica-usabilidad/. [Último acceso: 9 Septiembre 2022].

[13] MySQL, «13.1.17 CREATE PROCEDURE and CREATE FUNCTION Statements,» [En línea]. Available: https://dev.mysql.com/doc/refman/8.0/en/create-procedure.html. [Último acceso: 25 Marzo 2022].

[14] S. Chauhan, «Difference between Stored Procedure and Function in SQL Server,» DotNet Tricks, 31 Agosto 2022. [En línea]. Available: https://www.dotnettricks.com/learn/sqlserver/difference-between-stored-procedure-and-function-in-sql-server. [Último acceso: 9 Septiembre 2022].

[15] MariaDB, «Create Trigger,» MariaDB, 21 Mayo 2021. [En línea]. Available: https://mariadb.com/kb/en/create-trigger/. [Último acceso: 9 Septiembre 2022].

[16] PandoraFMS, «NOSQL vs SQL. Key differences and when to choose each,» Pandora FMS , 18 Febrero 2022. [En línea]. Available: https://pandorafms.com/blog/nosql-vs-sql-key-differences/. [Último acceso: 9 Septiembre 2022].

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

[18] A. Visus, «¿Para qué sirve Python? Razones para utilizar este lenguaje de programación,» ESIC, Octubre 2020. [En línea]. Available: https://www.esic.edu/rethink/tecnologia/para-que-sirve-python. [Último acceso: 10 Septiembre 2022].

[19] Real Academia Española, «Diccionario de la lengua española, 23.ª ed.,» [En línea]. Available: https://dle.rae.es. [Último acceso: 10 Septiembre 2022].

[20] P. Majumder, «Web Scraping a News Article and performing Sentiment Analysis using NLP,» 29 Octubre 2021. [En línea]. Available: https://www.analyticsvidhya.com/blog/2021/11/web-scraping-a-news-article-and-performing-sentiment-

analysis-using-nlp/. [Último acceso: 10 Septiembre 2022].

[21] Redaccion Keep Coding, «Lenguaje de programación Go y sus características,» KeepCoding, 13 enero 2022. [En línea]. Available: https://keepcoding.io/blog/lenguaje-de-programacion-go-caracteristicas/. [Último acceso: 9 Septiembre 2022].

[22] W. Boyd, «What is Go? An intro to Google's Go programming language (aka Golang),» A Cloud Guru, 25 Mayo 2021. [En línea]. Available: https://acloudguru.com/blog/engineering/what-is-go-an-intro-to-googles-go-programming-language-aka-golang. [Último acceso: 9 Septiembre 2022].

[23] B. Bashari Rad, H. John Bhatti y M. Ahmadi, «An Introduction to Docker and Analysis of it´s performance|,» *International Journal of Computer Science and Network Security*, vol. 17, nº 3, pp. 228-235, 2017.

[24] R. Sauerwalt, «Beneficios de la computación en la nube,» IBM, [En línea]. Available: https://www.ibm.com/co-es/cloud/learn/benefits-of-cloud-computing. [Último acceso: 10 Septiembre 2022].

[25] Microsoft, «What is Azure?,» Microsoft, [En línea]. Available: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-azure/. [Último acceso: 10 Septiembre 2022].

[26] CRUMstudy, CUERPO DE CONOCIMIENTO DE SCRUM (GUÍA SBOK), 3 ed., VMEdu Inc, p. 428.

[27] Hostingplus, «Modelo de prototipos: ¿qué es y cuáles son sus etapas?,» Hostingplus, 6 Julio 2021. [En línea]. Available: https://www.hostingplus.com.co/blog/modelo-de-prototipos-que-es-y-cuales-son-sus-etapas/. [Último acceso: 10 Septiembre 2022].

[28] C. d. EcuRed, «Modelo de prototipos,» 29 Agosto 2019. [En línea]. Available: https://www.ecured.cu/Modelo_de_prototipos. [Último acceso: 10 Septiembre 2022].