Universidad de Cuenca

Facultad de Ingeniería

Carrera de Ingeniería en Ciencias de la Computación

Generación de un grafo de conocimiento de periódicos antiguos del Ecuador a través de procesos OCR.

Trabajo de titulación previo a la obtención del título de Ingeniero en Ciencias de la Computación

Autores:

Raul Sebastian Torres Cordero

Jonnathan Andrés Valdez Llivisaca

Director:

Víctor Hugo Saquicela Galarza

ORCID: 0000-0002-2438-9220

Cuenca, Ecuador

2023-07-26

Resumen

La historia nos revela la existencia de una multitud de eventos que se desarrollan en el mundo día a día, dejando una huella en el tiempo. Antiguamente, la transmisión de ese conocimiento se realizaba de manera oral y se mantenía vivo a través de generaciones. No obstante, el avance de la tecnología ha revolucionado la forma en que accedemos a la información y nos ha permitido explorar registros históricos en una escala sin precedentes.

En este contexto, surge un desafío, gran parte de esa información yace dormida en periódicos antiguos, los cuales se encuentran en un estado de deterioro y son difíciles de tratar. Estos periódicos contienen relatos de eventos de la historia del Ecuador en los siglos XIX y XX, pero acceder a esa información de manera rápida y eficiente es un desafío.

Para abordar este problema, en este trabajo de titulación, se propone una solución basada en la digitalización de texto, el procesamiento texto y las tecnologías de la web semántica. El objetivo principal es extraer la información de los periódicos antiguos, organizarla de manera estructurada y generar un grafo de conocimiento que represente los eventos ocurridos en Ecuador durante ese período histórico.

La solución propuesta implica la automatización de cada uno de los pasos del proceso. Para lograrlo, se han construido varios widgets en Orange, que permite realizar tareas específicas en cada etapa del proceso. Estos widgets trabajan en conjunto para extraer la información, identificar entidades y relaciones, obtener Word Embendings y generar un grafo de conocimiento.

Palabras clave: Periódico Digitales, Incrustación de palabras, Ontología.

Abstract

History reveals to us the existence of a multitude of events that unfold in the world day by day, leaving a footprint in time. In the past, the transmission of this knowledge was done orally and kept alive through generations. However, the advancement of technology has revolutionized the way we access information and has allowed us to explore historical records on an unprecedented scale.

In this context, a challenge arises: a large portion of this valuable information lies dormant in old newspapers, which are in a state of deterioration and are difficult to handle. These newspapers contain detailed accounts of events that marked Ecuador's history in the 19th and 20th centuries, but accessing that information quickly and efficiently has become a challenge.

To address this problem, this thesis proposes a solution based on text digitization, text processing, and semantic web technologies. The main objective is to extract information from old newspapers, organize it in a structured manner, and generate a knowledge graph that represents the events that occurred in Ecuador during that historical period. As part of this solution, a prototype search engine has also been developed that utilizes the generated knowledge graph. This search engine is one of the many ways to exploit the graph and allows users to make specific queries and searches related to historical events, people, places, and topics in the context of old newspapers.

The proposed solution involves the automation of each step of the process. To achieve this, several widgets have been built in Orange, a visual data analysis platform, that allows for specific tasks to be performed at each stage of the process. These widgets include text digitization tools, text processing techniques, and semantic web algorithms that work together to extract relevant information, identify entities and relationships, obtain Word Embeddings, and generate a knowledge graph enriched with historical events.

Keywords: Digital Newspapers, Word Embeddings, Ontology.

Índice de contenidos

Índice de figuras

Índice de tablas

Agradecimientos

Quiero agradecer a mis padres, por su apoyo a lo largo de toda mi vida estudiantil, sin ellos todo esto no hubiese sido posible. A mi tutor y profesor, Víctor Saquicela, por todas sus enseñanzas a lo largo de estos últimos años. También agradecer a mis amigos, compañeros, colegas, profesores, hermanos y al resto de mi familia, cada uno ha sido parte de este camino. Este trabajo está dedicado a todos ellos. Este logro ha sido gracias a todos ustedes.

Eternamente agradecido con todos.

Raúl Sebastián Torres Cordero

Agradecimientos

Primeramente quiero agradecer a mis padres, Carlos e Isabel, por ser un pilar fundamental en este camino. A mi hermano César, por ser mi apoyo incondicional. A mi director de tesis, Victor, por su paciencia, dedicación y por haber compartido conmigo su experiencia y conocimientos. A mis compañeros y amigos que me ha dado esta trayectoria, ya que cada uno de ellos ha sido importante para lograr esta meta. A ellos, les dedico este trabajo, como una muestra de mi agradecimiento y mi cariño. Sin su apoyo, no habría sido posible alcanzar este logro.

Jonnathan Andrés Valdez Llivisaca

1. Introducción

Según [2] los periódicos históricos son una valiosa fuente de información sobre el pasado, ya que pueden proporcionar datos sobre acontecimientos sociales, políticos y económicos, así como sobre la vida cotidiana de personas de distintas épocas. Sin embargo, trabajar con periódicos históricos puede resultar difícil debido a su antigüedad y estado, lo que los hace a menudo difíciles de leer, y puede llevar mucho tiempo buscar información específica.

En los últimos años, ha habido un creciente interés por automatizar el proceso de extracción de información de fuentes históricas [3], gracias a la creciente disponibilidad de fuentes digitalizadas y al desarrollo de nuevas tecnologías para el tratamiento de textos y el procesamiento del lenguaje natural.

La digitalización ha hecho posible almacenar y acceder a las fuentes históricas de forma más eficiente, lo que ha suscitado un renovado interés por la investigación histórica. Ahora, los estudiosos pueden acceder a un abanico de fuentes más amplio que nunca.

Las nuevas tecnologías de tratamiento de textos y de procesamiento del lenguaje natural permiten extraer información de las fuentes históricas de forma más eficaz y precisa. Estas tecnologías pueden utilizarse para identificar y extraer palabras clave, fechas y otra información importante de los textos históricos.

La combinación de la digitalización y las nuevas tecnologías de tratamiento de textos y procesamiento del lenguaje natural está permitiendo automatizar el proceso de extracción de información de fuentes históricas como se muestra en [4]. Se trata de un avance significativo, puesto que permite acceder a la información histórica y utilizarla de formas nuevas e innovadoras.

Los periódicos históricos digitalizados son una fuente rica de información que ofrece una visión única de la historia y la cultura de una sociedad [2]. A través del análisis de estos documentos, los investigadores pueden obtener información sobre la vida cotidiana, los acontecimientos históricos, la política, la economía, la cultura y la sociedad de una época pasada.

Sin embargo, el acceso a esta información puede ser un desafío debido a la naturaleza de estos documentos. Los periódicos históricos a menudo fueron impresos en papel de baja calidad y se almacenaron en condiciones no ideales, lo que ha llevado a la degradación del texto y las imágenes. Además, la información puede estar desestructurada, lo que dificulta la búsqueda y el análisis de los datos.

Otro desafío es la necesidad de convertir las imágenes de los periódicos en texto editable mediante técnicas avanzadas de reconocimiento óptico de caracteres (OCR). Sin embargo, la calidad del texto OCR generado se ve afectada debido a la baja calidad del texto original y las condiciones en las que se han almacenado los periódicos históricos. Además, la naturaleza no estructurada de la información hace que la identificación y extracción de entidades relevantes en los textos sea difícil, lo que puede ser necesario para analizar temas específicos a lo largo del tiempo.

A pesar de estos desafíos, los periódicos históricos digitalizados ofrecen una fuente invaluable de información para los investigadores y estudiosos. Gracias a las técnicas avanzadas de procesamiento de texto y análisis de datos, los investigadores pueden descubrir nuevas perspectivas sobre la historia y la cultura, así como profundizar en la comprensión de los eventos pasados. Además, la digitalización de estos periódicos permite su preservación y acceso a un público más amplio, lo que contribuye a la preservación de la memoria histórica y cultural.

La solución propuesta pretende automatizar el proceso de extracción de información de periódicos históricos digitalizados y crear un grafo de conocimiento relacionado con diferentes tipos de acontecimientos históricos ocurridos en Ecuador durante los siglos XIX y XX. La solución utiliza tecnologías semánticas diseñadas para representar datos en un formato estructurado y legible por máquina, lo que permitirá una fácil integración e intercambio de datos entre diferentes sistemas y aplicaciones.

Para lograr este objetivo, la solución consta de varios pasos. El primero consiste en extraer el contenido y los metadatos de cada registro periódico histórico utilizando el protocolo OAI-PMH [5]. Este protocolo proporciona una forma estándar de acceder a los metadatos de los repositorios digitales, garantizando la coherencia y la estructura de los datos extraídos. En

este caso el repositorio digital del cual se recuperan estos documentos es la Hemeroteca Nacional digital de Ecuador, denominada Casa de la Cultura Ecuatoriana.

A continuación, se realizará un proceso de OCR en todos los archivos descargados de cada página, y el texto resultante se limpia y preprocesa mediante diversas técnicas para garantizar que los datos estén listos para su posterior análisis. Luego, se aplican técnicas de reconocimiento de entidades(NER) y detección de palabras clave para localizar y clasificar los nombres, nombres propios y palabras clave relevantes dentro de un contexto histórico.

La información obtenida mediante NER y la detección de palabras clave es aprovechada para generar un grafo de conocimiento. Para ello, se extrae información de periódicos históricos y se analizan las entidades presentes en los textos, y se amplían los vocabularios semánticos buscando sus significados en DBpedia [6]. Además, como parte del proceso, se obtendrán los Word Embeddings de cada texto. Esto permite comprender tanto la sintaxis como la semántica de los textos, proporcionando una representación numérica que refleja su significado y relaciones contextuales.

Por último, se propone un prototipo de aplicación web para visualizar los resultados extraídos del grafo de conocimiento. Esta aplicación web permitirá a los usuarios buscar información histórica relacionada con diferentes tipos de eventos ocurridos en Ecuador durante los siglos XIX y XX.

En resumen, el trabajo de titulación tiene como objetivo automatizar todo el proceso de tratar los periódicos históricos y no optimizar fases individuales. Una vez automatizado el proceso, se realizará una evaluación para verificar la validez de la automatización con expertos del dominio. En general, la solución propuesta presenta un enfoque integral para extraer y analizar datos históricos, ofreciendo una valiosa contribución al campo de las humanidades digitales.

1.1. Objetivos

A continuación se presentan tanto el objetivo general como los objetivos específicos de este trabajo de titulación.

1.1.1. Objetivo general

Automatizar el proceso de extracción, almacenamiento, descripción y visualización de datos de periódicos históricos digitalizados, que permita generar un grafo de conocimiento relacionado con diferentes tipos de eventos ocurridos en Ecuador en los siglos XIX-XX.

1.1.2. Objetivos específicos

1. Implementar un proceso automático de extracción de contenido de metadatos de periódicos históricos digitalizados utilizando el protocolo OAI-PMH.
2. Describir la información extraída de los periódicos históricos utilizando tecnologías semánticas para crear un grafo de conocimiento.
3. Almacenar el grafo de conocimiento generado de manera estructurada y accesible para su posterior consulta y reutilización.
4. Crear un visualizador que permita la búsqueda efectiva de información histórica a partir del grafo de conocimiento generado.

2. Marco Teórico y Trabajo Relacionado

A continuación se presentan los principales conceptos teóricos para entender los diferentes componentes que constituyen la propuesta del trabajo de titulación, así como un análisis de los trabajos relacionados con la propuesta.

2.1. Marco Teórico

En esta sección, se explorarán y analizarán los conceptos, teorías y modelos relevantes que respaldan el trabajo de titulación.

2.1.1. Ontologías

Las ontologías desempeñan un papel fundamental en la web semántica al proporcionar una especificación formal y explícita de la conceptualización de un dominio específico [7]. Constituyen una representación estructurada y coherente de los elementos clave en un dominio de conocimiento, lo que facilita la comprensión y el intercambio de información entre sistemas informáticos.

Una ontología se compone de varios elementos fundamentales que contribuyen a su estructura y contenido [8]. En primer lugar, se encuentran los conceptos, que son definiciones abstractas de elementos del dominio que representan entidades o propiedades comunes. Estos conceptos capturan las características esenciales y las relaciones dentro del dominio, brindando una base sólida para la organización y el razonamiento sobre la información.

Además de los conceptos, las ontologías incluyen relaciones, que establecen conexiones entre dos conceptos específicos, describen cómo se relacionan y se vinculan los diferentes elementos del dominio, lo que permite capturar de manera más precisa la semántica y las interacciones entre ellos.

Otro componente importante son las instancias, que son elementos específicos de un concepto. Al igual que en el paradigma orientado a objetos, las instancias representan casos individuales o ejemplos concretos dentro del dominio. Proporcionan ejemplos prácticos y concretos de los conceptos definidos en la ontología, enriqueciendo así la comprensión y la

utilidad de la representación.

Por último, los axiomas son restricciones que fortalecen y enriquecen la información contenida en la ontología. Estas restricciones pueden ser reglas lógicas, restricciones de cardinalidad u otras expresiones que definen condiciones y propiedades adicionales sobre los conceptos y las relaciones. Los axiomas ayudan a garantizar la coherencia y la integridad de la ontología, proporcionando un marco sólido para el razonamiento automático y la inferencia lógica.

Para modelar ontologías, existen varios lenguajes disponibles, pero uno de los estándares más ampliamente utilizado es el RDF (Resource Description Framework), aprobado por el World Wide Web Consortium (W3C). RDF utiliza un formato de **tripletas** para representar la información, donde cada tripleta consiste en un sujeto, un predicado y un objeto. Estas tripletas permiten representar entidades y sus relaciones de una manera estructurada y semánticamente rica. Con estos elementos se permite modelar ontologías a modo de grafos dirigidos como se puede ver en la Figura 2.1.

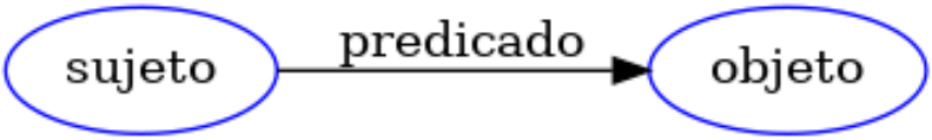

Figura 2.1: Tripleta RDF.

Al modelar ontologías con RDF, se puede visualizar la estructura resultante como un grafo dirigido, donde los conceptos se representan como nodos y las relaciones como arcos que conectan los nodos correspondientes. Este enfoque gráfico facilita la comprensión y la navegación dentro de la ontología, ya que muestra claramente las interconexiones y las dependencias entre los elementos del dominio.

Aunque RDF permite representar los datos en forma de grafo, no suele ser suficiente para representar un esquema. Debido a esta limitación surge RDF Schema o RDFS, una extensión de RDF. RDFS permite inferir sobre los datos agregando significado a los elementos de la ontología según lo recopilado en [9]. Además, todas las definiciones proporcionadas por RDFS, que representan relaciones específicas entre los términos, suelen ser descritas

mediante el lenguaje Web Ontology Language o OWL, el cual está específicamente diseñado para representar el conocimiento, siendo el recomendado por W3C y el más utilizado con este propósito como se menciona en [9].

2.1.1.1. Ontologías para la representación de Periódicos Históricos

Las plataformas para la producción, distribución y consumo de noticias deben aprovechar la cantidad de datos digitales disponibles a través de la web. Estos datos proceden de distintas fuentes y formatos, por ello una solución para juntar toda la información heterogénea es utilizar ontologías como el medio para estandarizar el conocimiento. Dentro de la web existen varias ontologías de dominio específico sobre periódicos, no obstante, la mayoría no están interconectadas entre sí. Una de las más populares y es ampliamente utilizada es la ontología de schema.org[1], la cual es una fuente de conocimiento que busca interconectar los datos, sobre todo, del propio internet. Contiene esquemas comunes de diferentes formatos de documentos incluyendo periódicos y es reconocida frente a diferentes herramientas del mercado de grandes empresas como Google y Microsoft [10]. Por ello, es una opción apropiada para el modelado de información del dominio de esta clase de documentos históricos y es la ontología que se utilizará para estructurar la información dentro del flujo de trabajo.

2.1.2. Ontologías de propósito general

También existen otras ontologías que no tienen un objetivo específico, sino que recopilan diferentes términos de diversas áreas y las interrelacionan entre sí. Este tipo de ontologías se suelen utilizar como diccionarios de términos y se encuentran disponibles a nivel de la web. Una de estas ontologías es DBpedia.

DBpedia es una base de conocimiento semántico extraída de Wikipedia y enlazada con otras fuentes de datos enlazados. Proporciona información estructurada sobre una amplia variedad de entidades, como personas, lugares, eventos, obras de arte, entre otros. Cada entidad en DBpedia tiene una representación semántica detallada que incluye propiedades, relaciones y enlaces a otros recursos relacionados.

Periódicos históricos como fuentes de información Los periódicos históricos han desem-

[1] https://schema.org/

peñado un papel fundamental como fuentes de información en la investigación histórica y social [11]. Estos registros impresos capturan la vida y los sucesos de épocas pasadas, proporcionando una visión valiosa y detallada de los eventos, las tendencias y los puntos de vista que moldearon la sociedad en ese momento específico.

Al examinar los periódicos históricos, los investigadores tienen la oportunidad de sumergirse en el pasado y explorar las realidades sociales, políticas, económicas y culturales de una época determinada[12]. Los periódicos antiguos ofrecen una perspectiva única sobre los sucesos cotidianos, los hitos históricos, las discusiones públicas, las campañas políticas, los cambios sociales y mucho más. Además, reflejan los debates, las creencias y las percepciones de la sociedad de ese momento, lo que permite comprender cómo se formaron y evolucionaron las opiniones públicas a lo largo del tiempo.

La riqueza de información contenida en los periódicos históricos es invaluable para los estudiosos interesados en la historia, la sociología, la comunicación y otras disciplinas relacionadas [13]. Estos documentos se convierten en ventanas hacia el pasado, proporcionando detalles concretos y testimonios directos de los eventos y las experiencias de la época. Además, los periódicos históricos a menudo cubren una amplia gama de temas, desde política y economía hasta cultura y entretenimiento, lo que permite una comprensión integral y multidimensional de la sociedad en un momento dado.

Sin embargo, acceder y utilizar eficientemente la información contenida en los periódicos históricos plantea desafíos significativos. La naturaleza física de estos documentos, muchas veces deteriorados o almacenados en archivos difíciles de explorar, dificulta su manejo y consulta. Además, la cantidad masiva de contenido presente en los periódicos históricos requiere métodos eficientes y automatizados para su extracción, análisis y organización.

Para el caso de estudio, se tomarán en cuenta los periódicos digitalizados que se encuentran en la Casa de la Cultura Ecuatoriana, los cuales conforman una colección de 15.679 registros asociados a periódicos antiguos. Estos recursos de información están disponibles en un sitio web público[2]. Los mismos que pertenecen a tres periódicos específicos: **El grito del pueblo** de Guayaquil, **El tiempo** de Guayaquil y **El Comercio** de Quito, cuyas fechas

[2]http://repositorio.casadelacultura.gob.ec/handle/34000/1534

de publicación abarcan desde 1860 hasta 1920.

2.1.3. Word Embedding

Los seres humanos han demostrado siempre una habilidad destacada en la comprensión de los idiomas, siendo sencillo para una persona entender la relación existente entre las palabras, no obstante, el mismo trabajo resulta en una tarea compleja para un computador. Por ejemplo, los seres humanos somos capaces de reconocer que las palabras **periódico** y **revista** tienen una conexión especial entre ellas pero bicicleta no. Sin embargo, ¿cómo puede una computadora resolver este tipo de relaciones?.

Word Embendings son, en esencia, una forma de representar las palabras que combina la comprensión humana del lenguaje con la capacidad de procesamiento de una máquina tal y como se puede ver en la Figura 2.2 . Estas representaciones aprenden a describir el texto en un espacio multidimensional, donde palabras con significados similares tienen representaciones cercanas. Esto significa que dos palabras que son similares se representan con vectores casi idénticos y se ubican muy cerca en un espacio vectorial. Estas embeddings son fundamentales para abordar la mayoría de los desafíos en el procesamiento del lenguaje natural.

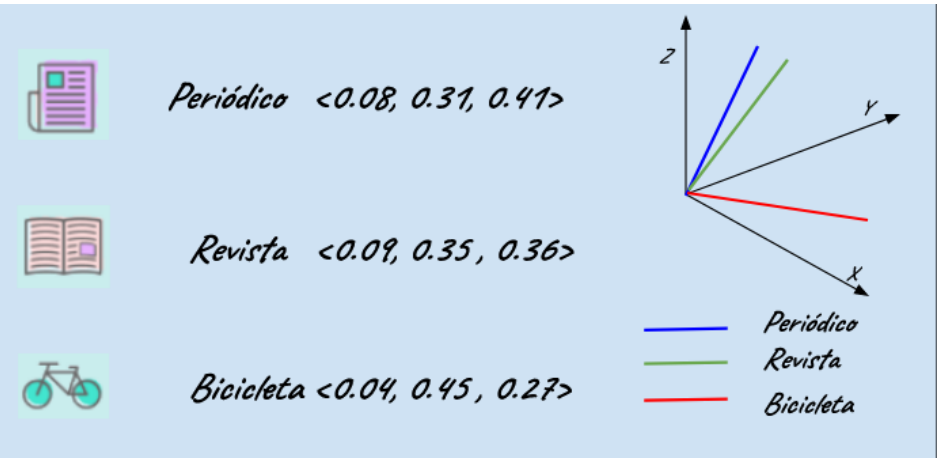

Figura 2.2: Representación de Word Embendings.

Word Embedding, también conocido como incrustación de palabras, es un campo importante dentro del procesamiento del lenguaje natural que se dedica a convertir palabras y frases en vectores de números reales [14]. Estos vectores representan de manera semánti-

ca el significado de las palabras y permiten a las máquinas comparar textos de manera más efectiva, encontrando similitudes y diferencias entre ellos. Estas representaciones pueden ser útiles para abordar tareas relacionadas con el análisis de texto, como clasificación de documentos, búsqueda de información y agrupación de documentos por temas similares.

Se toma como ejemplo un escenario en el que se dispone de un conjunto de datos compuesto por documentos de texto, por ejemplo, artículos de noticias, y se desea identificar temas similares o relacionados entre ellos. Para lograrlo, se emplean word embeddings.

El primer paso consiste en asignar a cada palabra en el conjunto de datos un vector numérico de alta dimensión. Supongamos que se tienen tres palabras: **perro**, **gato** y **pelota**. A cada una de ellas se le asigna un vector numérico, por ejemplo:

perro: [0.2, 0.8, -0.5] **gato**: [-0.4, 0.7, 0.2] **pelota**: [0.9, -0.3, 0.6]

Estos vectores representan las palabras en un espacio vectorial, y su disposición refleja las relaciones semánticas y sintácticas entre ellas. Una vez que se han asignado los embeddings a todas las palabras en los documentos, es posible calcular la similitud entre los vectores de las palabras para encontrar documentos que abordan temas similares.

Por ejemplo, supongamos que se tienen dos documentos:

Documento 1: **Se observó a un perro jugando con una pelota en el parque.**

Documento 2: **Los gatos destacan por su habilidad para cazar ratones.**

Para determinar la similitud entre los documentos, se puede calcular el promedio de los embeddings de las palabras presentes en cada documento y luego calcular la distancia entre los vectores resultantes. El promedio de los embeddings para el Documento 1 sería: [0.55, 0.25, 0.05] y el promedio de los embeddings para el Documento 2 sería: [-0.2, 0.7, 0.2].

Posteriormente, se puede emplear una medida de distancia, como la distancia euclidiana, para calcular la distancia entre estos vectores. Si la distancia resultante es baja, indica que los documentos son más similares en términos temáticos. En este caso, la distancia entre los vectores es relativamente alta, lo cual sugiere que los documentos abordan temas

diferentes.

Los embeddings capturan las relaciones semánticas entre las palabras, de manera que palabras relacionadas, como **perro** y **gato**, tendrán vectores más cercanos entre sí en el espacio vectorial.

En la actualidad, los enfoques principales en Word Embedding se basan en el uso de redes neuronales y modelos probabilísticos. Empresas líderes en inteligencia artificial como OpenAI han desarrollado algoritmos de embeddings que buscan abordar una amplia gama de problemas, como la búsqueda, la agrupación de documentos, las recomendaciones, la detección de anomalías, la diversidad y la clasificación.

Word embeddings internammente utilizan transformers, los transformers son una arquitectura de redes neuronales desarrollada para tareas de procesamiento del lenguaje natural y otros problemas secuenciales [15]. Fueron introducidos en 2017 por Vaswani et al. y desde entonces han demostrado un gran rendimiento en una amplia gama de aplicaciones.

A diferencia de las arquitecturas recurrentes tradicionales, como las redes LSTM (Long Short-Term Memory) o GRU (Gated Recurrent Unit), los Transformers no dependen de la recursión o de una estructura de secuencia lineal. En su lugar, se basan en un mecanismo de atención (attention) que les permite capturar las relaciones de largo alcance entre las palabras o elementos de una secuencia.

El funcionamiento básico de los Transformers se puede dividir en los siguientes componentes:

- Codificador (Encoder): El codificador se encarga de procesar la secuencia de entrada, como una oración, y extraer información contextualizada de cada elemento. Consiste en una serie de capas, cada una de las cuales aplica dos subcapas principales: la atención multi-cabeza y una red neuronal feed-forward (FFN, por sus siglas en inglés).
- Atención multi-cabeza: Es el corazón de la arquitectura del Transformer. Consiste en calcular la similitud entre todas las palabras de entrada y asignar pesos a esas similitudes. Esto permite que el modelo .ªtendiendo.ª diferentes partes de la secuencia durante el procesamiento. La atención multi-cabeza se realiza mediante una transfor-

mación lineal de las palabras de entrada en consultas, claves y valores, que luego se utilizan para calcular la similitud y obtener una representación ponderada de cada palabra.

- Red Neuronal Feed-Forward (FFN): Después de la capa de atención, se aplica una red neuronal feed-forward a cada posición de la secuencia por separado. Esto agrega flexibilidad y capacidad de modelado no lineal al codificador.
- Decodificador (Decoder): El decodificador es similar al codificador, pero tiene algunas diferencias clave. Además de las capas de atención multi-cabeza y FFN, el decodificador también incorpora una capa adicional de atención denominada .ªtención enmascarada"(masked attention). Esta atención enmascarada se utiliza para garantizar que el modelo no tenga acceso a la información futura durante la generación del texto, ya que se entrena utilizando el esquema de aprendizaje supervisado.
- Posicionamiento codificado: En las arquitecturas de los Transformers, se agrega información posicional a las palabras o elementos de la secuencia para capturar su orden. Se utilizan codificaciones de posición para indicar la ubicación relativa de cada elemento en la secuencia.

Durante el entrenamiento, los Transformers se optimizan mediante técnicas de descenso de gradiente estocástico y la retropropagación del error. El objetivo es minimizar una función de pérdida que mide la discrepancia entre las predicciones del modelo y los valores reales.

En el caso específico de la indexación de periódicos históricos, los modelos de Word Embedding desempeñan un papel crucial, permitiendo mejorar la capacidad de búsqueda por términos en documentos digitales, incluso cuando las entidades o palabras clave no se mencionan directamente en el texto. Al capturar el contexto y la semántica de las palabras, estos algoritmos pueden relacionar conceptos y facilitar la recuperación de información relevante en los documentos históricos.

La principal ventaja de utilizar word embeddings radica en que ofrecen una representación densa y continua del lenguaje, a diferencia de las representaciones basadas en matrices de ocurrencias de palabras. Esto implica que los word embeddings son capaces de capturar

matices y sutilezas semánticas que resultan difíciles de obtener mediante enfoques más tradicionales [16]. Al incorporar estas representaciones, es posible obtener un mayor poder de representación y una mejor comprensión de los textos que se analizan, provenientes de diferentes periódicos históricos.

2.2. Marco Tecnológico

Esta sección proporciona el contexto necesario para comprender las herramientas y tecnologías utilizadas en el desarrollo de la solución propuesta, se explorarán las tecnologías y plataformas relevantes que han sido seleccionadas para llevar a cabo la automatización del proceso.

2.2.1. OAI-PMH

La Iniciativa de Archivos Abiertos - Protocolo de recolección de métodos (OAI-PMH) ha surgido como una solución eficiente y de bajo costo para la difusión de contenidos en Internet. A diferencia de otros protocolos más complejos, OAI-PMH se enfoca en facilitar la búsqueda y recuperación de información de manera sencilla y accesible. Este protocolo fue desarrollado a finales de los años 90 en respuesta a la necesidad de la comunidad científica de compartir y transmitir documentos de forma efectiva [5]. Su objetivo principal es establecer un estándar para la interoperabilidad entre repositorios digitales, permitiendo la consulta y recuperación de contenidos de manera descentralizada.

Hoy en día, el OAI-PMH se ha convertido en uno de los protocolos más ampliamente utilizados en repositorios digitales, tanto en ámbitos generales como científicos [17]. Su popularidad se debe en gran parte a su flexibilidad y adaptabilidad, lo que ha permitido su implementación en diferentes contextos y para diversos tipos de recursos.

Una de las características distintivas del OAI-PMH es su enfoque en la simplicidad y la utilización de métodos HTTP como Post y Get para realizar las solicitudes de información. Esto significa que no requiere de tecnologías o infraestructuras complejas, lo que lo convierte en una opción asequible y fácil de implementar para instituciones y organizaciones que deseen compartir sus recursos en línea.

En el caso particular del repositorio utilizado para este trabajo de titulación, utiliza el protocolo

OAI-PMH como parte de su estrategia de difusión y acceso a sus documentos, incluyendo valiosos periódicos históricos. Al utilizar este protocolo, la institución puede compartir de manera eficiente y estandarizada su colección de periódicos históricos, asegurando que estos recursos estén disponibles para su consulta y estudio por parte de investigadores, académicos y el público en general.

2.2.2. Reconocimiento óptico de caracteres (OCR)

El Reconocimiento óptico de caracteres (OCR) es un conjunto de algoritmos o técnicas que permiten recuperar el texto de una imagen y puede ser utilizado por una máquina [18]. Normalmente los procesos OCR conllevan preprocesamiento para mejorar la imagen inicial, reconocimiento de texto para obtener una versión inicial, coincidencia con patrones para mejorar el resultado relacionando el texto bruto con conocimiento previo y post procesamiento que consiste en llevar el texto a algún formato específico.

Es importante destacar que el OCR ha encontrado aplicaciones en una amplia variedad de campos [19], como la digitalización de archivos históricos, la automatización de procesos de oficina, la extracción de datos en aplicaciones de reconocimiento de formularios, la accesibilidad para personas con discapacidad visual y la indexación de contenido para motores de búsqueda.

Los avances en OCR han permitido que las organizaciones y los usuarios individuales puedan convertir rápidamente grandes cantidades de documentos físicos en formatos digitales, lo que facilita su almacenamiento, búsqueda y análisis. Sin embargo, es importante tener en cuenta las limitaciones del OCR, especialmente cuando se trata de fuentes de baja calidad[20], textos manuscritos o idiomas con estructuras complejas.

2.2.2.1. Tesseract OCR

Tesseract es un software ampliamente utilizado en la industria y la comunidad de investigación para OCR [21]. Desarrollado originalmente en los laboratorios de Google, Tesseract se distribuye bajo la licencia Apache, lo que lo hace de código abierto y libre para su uso y modificación.

A lo largo de los años, Tesseract ha experimentado importantes mejoras y evoluciones en su

enfoque. Inicialmente, se basaba en métodos tradicionales de reconocimiento de patrones y técnicas de segmentación de texto para extraer el contenido textual de las imágenes. Sin embargo, con los avances en el campo del aprendizaje automático, Tesseract ha adoptado enfoques más sofisticados, como el uso de redes neuronales de tipo LSTM (Long Short-Term Memory).

Las redes neuronales LSTM son una arquitectura especializada para el procesamiento de secuencias y han demostrado ser altamente efectivas en tareas de reconocimiento de texto [19]. Estas redes permiten capturar y modelar las dependencias a largo plazo en las secuencias de caracteres, lo que mejora significativamente la precisión y la capacidad de reconocimiento de Tesseract. Además, Tesseract ha sido entrenado y optimizado para admitir hasta 116 idiomas diferentes, lo que lo convierte en una herramienta versátil y adaptable a una amplia gama de escenarios de OCR.

La implementación de Tesseract en el lenguaje de programación Python es sencilla y conveniente. La biblioteca de Tesseract proporciona una interfaz fácil de usar que permite a los desarrolladores integrar fácilmente la funcionalidad de OCR en sus aplicaciones. Al importar la biblioteca de Tesseract en un proyecto de Python, los investigadores tienen acceso a potentes funciones y métodos que les permiten procesar imágenes, extraer texto y realizar tareas de análisis y manipulación de datos.

Se seleccionó Tesseract como la herramienta principal para llevar a cabo el reconocimiento óptico de caracteres en este estudio. Esta elección se basó en una cuidadosa revisión de las opciones disponibles y en la reputación positiva que Tesseract ha ganado en la comunidad científica y académica [22].

En primer lugar, se investigaron y exploraron diversas alternativas y motores de OCR para encontrar la solución más adecuada. Sin embargo, tras un análisis, se determinó que Tesseract ofrecía el equilibrio adecuado entre precisión, flexibilidad y accesibilidad requerido para los objetivos de este estudio. La revisión de la comunidad científica reveló que Tesseract es ampliamente utilizado y altamente valorado por su capacidad para realizar el reconocimiento óptico de caracteres de manera efectiva [22].

Además de su reconocimiento y aceptación en la comunidad científica, otro factor importante que influyó en la elección de Tesseract fue su disponibilidad como una solución de código abierto. Esta característica permitió un acceso más amplio y la posibilidad de adaptar y personalizar la herramienta según las necesidades específicas del estudio. Por otro lado en [22] se realizaron evaluaciones experimentales cualitativas y cuantitativas utilizando cuatro servicios OCR reconocidos: Google Docs OCR, Tesseract, ABBYY FineReader y Transym. Se analizó la precisión y confiabilidad de estos paquetes de OCR utilizando un conjunto de datos que incluía 1227 imágenes pertenecientes a 15 categorías diferentes.

Los resultados y las conclusiones obtenidas en esta evaluación respaldan la elección de Tesseract como la tecnología OCR en el presente estudio. Las evaluaciones experimentales cualitativas y cuantitativas proporcionaron una base sólida para afirmar que Tesseract es una opción confiable y eficiente para el reconocimiento óptico de caracteres. Al considerar la relevancia y los resultados obtenidos en este estudio, se pudo respaldar la elección de Tesseract como una tecnología OCR adecuada para el propósito de este estudio, ya que ha demostrado su precisión y confiabilidad en la conversión de imágenes y documentos electrónicos en texto legible.

2.2.3. Large Language Models

Large Language Model (LLM) o modelo de lenguaje grande se refiere a un campo de la inteligencia artificial. Estos modelos son algoritmos de aprendizaje automático diseñados para comprender y generar texto en lenguaje natural. Un LLM es capaz de procesar y comprender grandes cantidades de texto y aprender patrones lingüísticos complejos. Estos modelos utilizan técnicas de aprendizaje profundo, como redes neuronales, para capturar la estructura y las características del lenguaje humano [23].

Un LLM puede ser entrenado en una amplia variedad de tareas relacionadas con el lenguaje, como la generación de texto, la traducción automática, la respuesta a preguntas, la corrección de texto y mucho más. Estos modelos son capaces de producir resultados impresionantes en términos de coherencia y calidad del texto generado [24].

Una de las principales ventajas de los LLMs, como el modelo ”text-davinci-002”, radica en

su capacidad para comprender y analizar el lenguaje natural. Estos modelos han sido entrenados con grandes volúmenes de texto y han adquirido una comprensión profunda de los patrones y contextos lingüísticos [25]. Esta capacidad les permite capturar sutilezas y relaciones semánticas en el texto, lo que resulta beneficioso para identificar y clasificar entidades con nombre de manera precisa y consistente.

Otra ventaja destacada de utilizar un LLM es su capacidad de automatizar el reconocimiento de entidades. Estos modelos son capaces de analizar grandes cantidades de texto de manera rápida y precisa, lo que ahorra tiempo y recursos [26]. Al emplear un enfoque basado en LLM, se evita la necesidad de implementar reglas y patrones manuales, lo que simplifica y agiliza el proceso de detección de entidades en comparación con métodos tradicionales.

Sin embargo, es importante tener en cuenta algunas consideraciones al utilizar un LLM para el reconocimiento de entidades. Estas incluyen la necesidad de un conjunto de datos de entrenamiento adecuado y la comprensión de los posibles sesgos inherentes al modelo. Además, se requiere un proceso de evaluación y refinamiento continuo para garantizar la calidad y la confiabilidad de los resultados obtenidos.

Los LLMs también se destacan por su adaptabilidad y escalabilidad [27]. Estos modelos tienen la capacidad de abordar una amplia gama de tareas relacionadas con el procesamiento del lenguaje natural, incluido el reconocimiento de entidades. Además, se pueden ajustar y afinar para adaptarse a dominios específicos o conjuntos de datos particulares, lo que mejora aún más su desempeño y precisión en contextos especializados.

Además, los LLMs, como el modelo ”text-davinci-002”[3], ofrecen soporte para múltiples idiomas. Esto resulta especialmente valioso en entornos multilingües, donde se requiere el reconocimiento de entidades en diferentes idiomas. Estos modelos pueden adaptarse a diversos contextos lingüísticos y proporcionar resultados precisos y coherentes en cada idioma.

Es importante tener en cuenta que los LLMs, como el modelo ”text-davinci-002”, se benefician de la mejora continua a través de la retroalimentación constante y los datos actualizados. Estos modelos se entrenan continuamente con nuevos datos y técnicas de aprendizaje

[3] Un modelo de lenguaje pre-entrenado que puede entender y generar lenguaje natural

automático, lo que conduce a mejoras en su rendimiento a lo largo del tiempo. Como resultado, utilizar un LLM para el reconocimiento de entidades garantiza una solución actualizada y en constante evolución que se mantiene al día con los avances en el campo del procesamiento del lenguaje natural.

2.2.3.1. Chat GPT

Chat Generative Pre-Trained Transformer (Chat GPT) o Chat de tipo Transformer Generativo Pre-Entrenado es un chatbot basado en LLM que se enfoca en el procesamiento del lenguaje natural [28]. Esta herramienta de vanguardia, desarrollada por OpenAI[4], ha revolucionado la forma en que las personas interactúan con las máquinas, brindando una experiencia de conversación más natural y fluida.

Chat GPT se distingue por su capacidad para comprender y generar respuestas coherentes y contextualmente relevantes [29]. Esto se logra gracias a su arquitectura basada en los modelos Transformer, que ha demostrado ser altamente efectiva en tareas de procesamiento del lenguaje natural. El modelo es pre-entrenado utilizando grandes cantidades de texto de diferentes fuentes, lo que le permite adquirir conocimiento lingüístico y capturar patrones de lenguaje complejos.

Una de las ventajas clave de Chat GPT es su capacidad para resolver una amplia gama de problemas. Además de generar respuestas conversacionales, este software es capaz de realizar traducciones automáticas, ayudar en la programación y analizar el sentimiento de un texto. La versatilidad de Chat GPT ha sido posible gracias a su entrenamiento con una amplia diversidad de datos [30] y su capacidad para aprender de patrones y estructuras lingüísticas.

Otra característica destacada de Chat GPT es su capacidad para sintetizar texto, es decir puede generar texto de manera coherente y con un estilo similar al de las fuentes de entrenamiento. Esto se ha utilizado para generar contenido automatizado, como resúmenes de texto, respuestas a preguntas frecuentes y narraciones de historias.

[4]Empresa de investigación y desarrollo enfocada en inteligencia artificial. Mas información: https://openai.com/about.

2.2.3.2. Chat GPT como postprocesamiento de texto

Para el presente trabajo, se ha aprovechado la capacidad multifuncional de Chat GPT para mejorar los resultados de los algoritmos OCR (Reconocimiento Óptico de Caracteres). La combinación de estas dos tecnologías puede ser altamente efectiva en la mejora de la calidad y precisión de la extracción de texto de documentos históricos. Según [31], evaluaron diferentes LM(modelos de lenguaje) preentrenados en dos conjuntos de datos y encontraron ganancias significativas en escenarios realistas con una mejora de hasta un 15 % en corrección de texto tras un proceso OCR.

La API de Chat GPT se utiliza como un complemento para los resultados obtenidos mediante OCR. Esta integración permite enviar los resultados de OCR a la herramienta, lo que facilita la detección y eliminación de posibles errores presentes en los documentos históricos heterogéneos. Dado que estos documentos a menudo presentan desafíos, como la presencia de manuscritos, baja calidad de la imagen o estructuras de lenguaje complejas, obtener resultados aceptables únicamente con los algoritmos OCR puede resultar difícil.

Una vez que los resultados de OCR son procesados por Chat GPT, los textos presentan mejoras en la sintaxis y la semántica del texto recuperado por OCR. La capacidad de Chat GPT para comprender el contexto y generar respuestas contextualmente relevantes permite superar las limitaciones inherentes de los algoritmos OCR y mejorar la calidad de los resultados. Además, su entrenamiento en grandes cantidades de texto de diversas fuentes le proporciona un amplio conocimiento lingüístico y la capacidad de capturar patrones de lenguaje complejos, lo que contribuye a una mejor reconstrucción del texto original.

2.2.3.3. Chat GPT como herramienta de detección de entidades

El Reconocimiento de Entidades Nombradas (NER, por sus siglas en inglés) es una tecnología clave en el procesamiento del lenguaje natural que se utiliza para identificar y clasificar entidades significativas en un texto, como nombres de personas, organizaciones, ubicaciones, fechas, cantidades, entre otros. Estos algoritmos de extracción de información desempeñan un papel fundamental en la comprensión automática del lenguaje humano y tienen aplicaciones en diversas áreas, como la búsqueda de información, el análisis de sentimien-

tos, la traducción automática y la generación de resúmenes [32]. Según [33] el uso de un LM es efectivo para el reconocimiento de entidades nombradas, es por ello que Chat GPT también se puede aprovechar para llevar a cabo tareas de NER.

La comprensión contextual de Chat GPT es una característica clave que permite realizar una identificación precisa y una clasificación adecuada de las entidades en el texto. Además, su habilidad para generar texto coherente y estructurado facilita la presentación de los resultados en diferentes formatos como JSON, XML o texto plano. Esta flexibilidad en la presentación de los resultados permite una fácil integración con otros sistemas y un procesamiento posterior eficiente de la información extraída.

La versatilidad de Chat GPT como una API flexible para el reconocimiento de entidades nombradas le brinda a los desarrolladores la posibilidad de adaptar su funcionamiento según las necesidades específicas de sus aplicaciones. Puede ser utilizado para extraer información relevante de documentos, analizar grandes volúmenes de texto en tiempo real o mejorar la precisión de los sistemas existentes de reconocimiento de entidades.

El uso de Chat GPT en tareas de reconocimiento de entidades nombradas combina las ventajas de su capacidad de comprensión del lenguaje natural con la flexibilidad y la facilidad de integración de una API, lo que lo convierte en una herramienta poderosa para la extracción precisa y eficiente de información de texto no estructurado [34].

2.2.4. Herramientas de Análisis y Minería de Datos

Las herramientas de análisis y minería de datos son aplicaciones o software diseñados para ayudar a los usuarios a descubrir patrones, tendencias y relaciones ocultas en conjuntos de datos complejos. Estas herramientas permiten la extracción de información valiosa y conocimientos significativos a partir de grandes volúmenes de datos, lo que resulta fundamental en la toma de decisiones informadas.

Estas herramientas se utilizan en una variedad de industrias y disciplinas, como la investigación científica, la medicina, el marketing, las finanzas y más. Proporcionan una serie de técnicas y algoritmos de análisis, como la clasificación, la regresión, el agrupamiento y la detección de anomalías, que permiten explorar y comprender mejor los datos.

Las herramientas de análisis y minería de datos suelen ofrecer interfaces visuales y amigables, lo que facilita su uso por parte de usuarios con diferentes niveles de experiencia en programación o estadística. Además, muchas de estas herramientas también brindan la capacidad de realizar análisis avanzados utilizando lenguajes de programación como Python o R.

Algunas de las características comunes de estas herramientas incluyen la capacidad de importar y procesar datos de diferentes fuentes, realizar transformaciones y limpieza de datos, generar visualizaciones interactivas, construir modelos predictivos y evaluar su rendimiento, y automatizar tareas repetitivas. Entre las principales se tiene RapidMiner, KNIME y Orange Data Mining.

RapidMiner es una herramienta de minería de datos y análisis predictivo que ofrece una interfaz gráfica intuitiva y fácil de usar. Permite a los usuarios construir flujos de trabajo de análisis de datos de manera visual al conectar nodos interactivos que representan diferentes operaciones y algoritmos [35]. Con RapidMiner, los usuarios pueden importar, preprocesar y explorar datos, así como aplicar algoritmos de aprendizaje automático para realizar análisis predictivos y descubrir patrones ocultos en los datos. La herramienta también ofrece opciones de visualización de datos y capacidades de evaluación de modelos para validar y mejorar el rendimiento de los resultados. Con su enfoque en la facilidad de uso y la flexibilidad, RapidMiner es ampliamente utilizado en diversos campos, como el comercio, la salud, el marketing y la investigación.

KNIME es una plataforma de análisis de datos y minería visual que permite a los usuarios manipular, analizar y modelar datos de manera intuitiva y eficiente [36]. Con KNIME, los usuarios pueden construir flujos de trabajo de análisis mediante la interconexión de nodos que representan diversas operaciones y algoritmos de análisis de datos. La plataforma ofrece una amplia gama de herramientas y funcionalidades para la importación y preprocesamiento de datos, así como para el modelado y la evaluación de algoritmos de aprendizaje automático. KNIME también cuenta con una comunidad activa que proporciona una amplia biblioteca de nodos y extensiones, lo que brinda a los usuarios aún más opciones y flexibilidad en sus análisis. Con su enfoque en la visualización y la automatización, KNIME es

utilizado en diversos campos, como la investigación científica, la industria farmacéutica, el análisis financiero y más.

Orange Data Mining es una herramienta poderosa y versátil en el campo de la minería de datos y el análisis predictivo [37]. Desarrollado como un software gráfico y de línea de comandos, Orange ofrece a los usuarios una interfaz intuitiva y fácil de usar para realizar diversas tareas de análisis de datos. Su capacidad para crear flujos de trabajo facilita el procesamiento de datos complejos y la ejecución de algoritmos de minería de datos de manera eficiente.

Una de las características destacadas de Orange es su capacidad para trabajar con widgets interconectados, es por esta razón por la que se ha decidido utilizar esta herramienta para este trabajo de titulación. Estos widgets actúan como bloques de construcción y permiten a los usuarios crear flujos de trabajo personalizados al conectarlos entre sí, esto brinda flexibilidad y personalización, ya que los usuarios pueden construir y adaptar sus flujos de trabajo según sus necesidades específicas. Además, los widgets de Orange ejecutan scripts de Python, lo que brinda la posibilidad de extender las funcionalidades del software utilizando el poderoso lenguaje de programación. Además es importante resaltar que Orange no está limitado a los widgets que vienen ya predefinidos, sino que también permite crear fácilmente componentes propios e interconectables con los ya existentes.

Uno de los aspectos clave de Orange es su capacidad para manejar y procesar datos a través de frames o matrices de datos. Esto permite la importación, manipulación y transformación de conjuntos de datos de manera eficiente. Los usuarios pueden realizar diversas operaciones en los datos, como filtrado, selección de atributos, normalización y agregación, entre otros, lo que facilita el procesamiento de grandes volúmenes de información.

Además de su capacidad para trabajar con datos, Orange también ofrece una amplia gama de opciones de visualización. Los usuarios pueden explorar y analizar sus datos en tiempo real, lo que les permite comprender mejor la estructura y distribución de los datos. Esta capacidad de visualización en tiempo real ayuda a identificar patrones, tendencias y relaciones ocultas en los datos.

2.2.5. Herramientas de almacenamiento de tripletas

Dentro del mundo de las ontologías, estas no son nada sin herramientas que permiten trabajar con ellas y, sobre todo, disponibilizarlas para su accesibilidad. Es por ello, que para el presente trabajo se requiere una herramienta que permite tanto subir información a ella como consultar la que contenga. Una de las herramientas que puede realizar correctamente este trabajo es Apache Jena[5].

Apache Jena, de manera general, es un framework de código abierto que permite trabajar con información ligada dentro de la web semántica. Esta última permite trabajar desde diferentes campos, pero el más relevante para este caso es el almacenamiento de tripletas en Fuseki. Esta última se encarga de almacenar tripletas de manera óptima utilizando las APIs que su propia librería incluye, además de añadir SPARQL **endpoints** que permiten consultar la información a través de servicios de tipo REST. Todo eso en un servidor facilmente despegable y manejable desde una aplicación web tambien contenida [38]. Todas estas características hacen de Fuseki una de las opciones principales a tener en cuenta al momento de trabajar con ontologías, sobre todo en un entorno académico.

2.3. Trabajos Relacionados

La preservación digital de archivos históricos, incluidos los periódicos antiguos, ha despertado un gran interés en el ámbito de la investigación [39]. Los esfuerzos por comprender y analizar los fenómenos históricos se remontan a los primeros testimonios escritos de la humanidad. En este contexto, la digitalización y preservación de los archivos de periódicos antiguos se ha convertido en una forma efectiva de salvar la información en peligro de extinción y crear bases de datos digitales que permitan la extracción de conocimiento para la toma de decisiones mediante el uso de tecnologías actuales.

Varios trabajos de investigación se han enfocado en abordar los desafíos relacionados con el tratamiento de datos provenientes de documentos físicos, especialmente de periódicos históricos [40],[39] o [41]. Algunos de estos trabajos han propuesto enfoques basados en el reconocimiento óptico de caracteres (OCR) para detectar columnas y bloques de texto en

[5]Disponible en: https://jena.apache.org/documentation/

los periódicos como se muestra en [42]. Si bien estos enfoques han mostrado resultados prometedores, no han tenido en cuenta las dificultades inherentes a los periódicos antiguos, como sesgos y deformaciones. Por otro lado autores como [41] y [43], presentan enfoques que combinan diversas técnicas con OCR para abordar los desafíos del procesamiento de periódicos históricos, logrando la identificación de segmentos de las páginas, como títulos, textos e imágenes, provenientes de diversas fuentes.

Un ámbito en donde se ha tratado de recuperar información ha sido el meteorológico, esto se puede apreciar en [44] en donde se han realizado trabajos previos para rescatar y digitalizar datos sobre ciudades específicas, como los registros de presión atmosférica desde fechas tempranas. Estos esfuerzos tienen como objetivo analizar el clima de la ciudad, mejorar la toma de decisiones, realizar investigaciones y garantizar la calidad de los datos. Hulme [45], describió un análisis de datos meteorológicos del condado de Norfolk, Inglaterra, que se centró en eventos históricos, como las olas de calor. Estos estudios han utilizado técnicas de minería de datos para proporcionar información precisa y veraz sobre estos eventos. Otros trabajos como [46], [47] y [48] también han propuesto el análisis de datos de periódicos antiguos relacionados con eventos meteorológicos, con el objetivo de crear bases de datos para la toma de decisiones en relación con el clima.

En un campo relacionado, se destaca el trabajo [49] que se centra en el procesamiento de archivos médicos utilizando grafos de conocimiento para mejorar la recuperación de información y la toma de decisiones en el ámbito de la medicina. Los autores presentan un marco de trabajo que combina técnicas de procesamiento de lenguaje natural, aprendizaje automático y creación de grafos de conocimiento para extraer información relevante de los registros médicos y representarla de forma estructurada y semántica. Este enfoque ha demostrado ser efectivo en la identificación de relaciones relevantes entre entidades y conceptos médicos, lo que facilita una mejor comprensión de la información y una toma de decisiones más precisa.

En [50] se investiga como generar una base de conocimientos de crímenes basado en relaciones de entidades para extraer e integrar datos de texto e imagen relacionados con el crimen de periódicos en línea, con un enfoque en reducir la duplicidad y la pérdida de in-

formación en la base de conocimientos. El sistema propuesto utiliza un enfoque basado en reglas para extraer las entidades de los datos de texto y los títulos de las imágenes. Las entidades extraídas de los datos de texto se correlacionan utilizando medidas de similitud contextual y semántica, mientras que las entidades de imagen se correlacionan utilizando características de imagen de bajo y alto nivel.

Es importante destacar que este trabajo de titulación se fundamenta en el estudio realizado en [51], donde se aborda todas las fases del proceso de manera manual. Este estudio proporciona una base sólida para comprender las diversas etapas involucradas en el proceso y los desafíos que surgen al realizarlo de forma manual. Su enfoque ha permitido identificar las tareas clave, los flujos de trabajo y los criterios para lograr resultados precisos. Sin embargo, debido a la naturaleza laboriosa y propensa a errores de este enfoque manual, surge la necesidad de desarrollar una solución automatizada.

Por lo antes descrito, el propósito de este trabajo de titulación es diseñar e implementar un sistema automatizado que pueda realizar las diferentes fases del proceso de manera eficiente y precisa. Esto implica la aplicación de técnicas de aprendizaje automático, procesamiento de imágenes, análisis de datos y otras tecnologías semánticas relevantes para lograr la automatización deseada.

La automatización del proceso proporcionará numerosos beneficios, como la reducción del tiempo y esfuerzo requeridos, la mejora de la consistencia y calidad de los resultados, la capacidad de procesar grandes volúmenes de datos, realizar análisis avanzados para obtener información adicional. Además, al eliminar la dependencia del enfoque manual, se espera minimizar los errores y aumentar la escalabilidad del proceso.

Como se describe en [52] es importante destacar la importancia de implementar procesos de integración y publicación de datos en formatos abiertos en la web, lo que permite que la información sea reutilizable e interoperable para la comunidad. En este sentido en [53], se han presentado diferentes enfoques para la transformación de datos históricos a formatos RDF (Resource Description Framework) con el objetivo de crear repositorios de acceso libre y estandarizado para los usuarios web. Sin embargo, hasta donde se sabe, no se ha abordado la digitalización de eventos, la extracción de conocimiento y la generación de grafos

de conocimiento a partir de periódicos para ilustrar los antiguos fenómenos históricos en Ecuador.

3. Automatización del proceso para la obtención de un grafo del Conocimiento

Este trabajo trata con periódicos históricos pertenecientes a la Hemeroteca Nacional digital de Ecuador, denominada Casa de la Cultura Ecuatoriana. Esta organización cuenta con una colección de 15.679 registros asociados a periódicos antiguos, y cada uno de estos registros tiene asociados metadatos y un archivo en formato PDF de los periódicos. Estos recursos de información están disponibles en un sitio web público. Los periódicos seleccionados fueron El grito del pueblo de Guayaquil, El tiempo de Guayaquil y El Comercio de Quito, cuyas fechas están entre 1860 y 1920. Estos periódicos presentan algunas características que provocan que, al digitalizarse como imágenes para posteriormente ser tratados con un software de OCR, retornen ruido o errores. Las características son las siguientes:

- Algunos periódicos históricos digitalizados no tienen estructura ni secciones.
- Las noticias aparecen sin títulos, en el mejor de los casos, diversas noticias se separan de otras usando guiones.
- Las colecciones de periódicos suelen estar incompletas y presentan problemas de conservación relevantes debido a las condiciones de preservación y a las características relacionadas con el material impreso.
- La digitalización de estos periódicos históricos presenta diferentes cuestiones de calidad asociadas a los problemas de conservación, a la falta de concienciación y a los recursos humanos y tecnológicos.

En la Figura 3.1 se presenta la solución propuesta en [51], la cual explica los diferentes pasos que se llevan a cabo para tratar los periódicos y solucionar los problemas descritos. Esta propuesta es el proceso base que se automatizará permitiendo así tratar este tipo de documentos. El resultado se encuentra disponible en el repositorio digital del trabajo de titulación[1].

Para comprobar el funcionamiento de cada una de las fases del proceso de automatización, se ejemplificará utilizando un periódico, el cual se encuentra dentro del repositorio de la

[1] https://github.com/Jonathan2703/TesisGrafoConocimiento

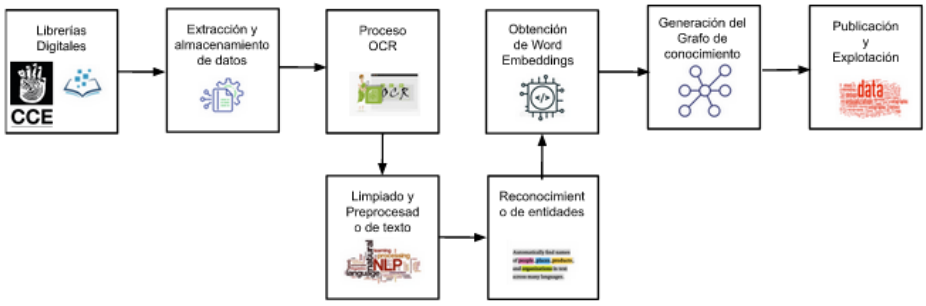

Figura 3.1: Gráfico conceptual de la solución

Casa de la Cultura Ecuatoriana[2] en formato PDF. Además, se proporciona la imagen de la primera página del periódico en la Figura 3.2 para un mejor entendimiento del proceso.

3.1. Extracción y almacenamiento de datos

En este trabajo de titulación, se aborda el desafío de acceder y aprovechar los periódicos históricos disponibles en un repositorio digital. Estos periódicos históricos contienen una valiosa fuente de información que puede contribuir significativamente al estudio y análisis de eventos pasados, así como a la comprensión de la evolución de la sociedad y la cultura.

Para lograr este objetivo, se diseñó e implementó un proceso automatizado que permite extraer de manera eficiente los metadatos de cada registro de periódico histórico. La extracción de metadatos es fundamental para comprender y organizar la información contenida en los registros, ya que proporciona detalles importantes como la fecha de publicación, el título, el autor, las palabras clave y otros atributos relevantes.

El proceso de extracción de metadatos se realizó utilizando el protocolo OAI-PMH, para recuperar los metadatos de los registros almacenados en el repositorio digital. Con el uso del protocolo OAI-PMH, se accedió de manera eficiente a los registros de periódicos históricos, lo que permitió obtener una visión general de los contenidos disponibles en el repositorio.

Sin embargo, surgió un desafío adicional durante el proceso de extracción de metadatos. Se descubrió que los archivos PDF de los periódicos no podían descargarse directamente mediante el protocolo OAI-PMH. Para superar esta limitación, se implementó un proceso adicional utilizando técnicas de web scraping (extracción de datos web).

[2] http://repositorio.casadelacultura.gob.ec/bitstream/34000/11101/1/AND_164.pdf

LOS ANDES

INTERDIARIO INDEPENDIENTE

Año II — Riobamba—Ecuador—Miércoles 23 de Enero de 1918 — No. 164

Ejemplo de patriotismo

Los pueblos septentrionales de la República, en su afán de llevar a cabo la construcción de una obra de la magnitud e importancia del ferrocarril de Quito a Esmeraldas, nos están estimulando con el ejemplo más elocuente de patriotismo práctico q' darse puede.

Nunca ha presenciado el país acción tan unánime y de positivos resultados, explosión más grande de patriotismo, manifestación más clara de la idiosincracia de un pueblo, que la que ahora observamos complacidos en el laudable empeño de nuestros compatriotas del Norte.

La noble actitud de los imbabureños y carchenses que, unidos a los habitantes del Pichincha, trabajan con entusiasmo indeclinable por ernzar su vasto y rico territorio con las paralelas de acero, ha tenido resonancia en la prensa extranjera, la cual ha comentado favorablemente, señalando como un hecho digno de la mayor alabanza, esa insólita decisión de los ecuatorianos para impulsar las obras ferroviarias, venciendo todo género de obstáculos que, en otra tierra y con otros hombres, habrían producido el fracaso del salvador proyecto.

La prensa nacional no necesita ilustrar el criterio de las masas, para llevarles el convencimiento de que el silbato de la locomotora es anuncio de prosperidad y un bello despertar a la vida modernizada; bien comprendido lo tienen, cuando así tan espontáneamente acuden a prestar su contingente personal todos cuantos pueden ser útiles en esa obra de redención. Nuestros gobernantes saben también que facilidad de comunicación es la que hace falta para que el pueblo ecuatoriano se conozca más a sí mismo y que, resuelto el problema de vialidad, se abrirán más vastos horizontes para el trabajo a la bienhechora sombra de la paz.—Tampoco hace al caso ponderar las grandes ventajas de la obra que así ha despertado el interés y patriotismo de toda una importante región, o sea de tres provincias que, debido a su propio esfuerzo, van a incorporarse a la parte privilegiada del país mediante la cinta de hierro que, en estos momentos, es lazo de fraternidad, estrecho abrazo de los que juntos, por iguales aspiraciones y derechos, deben marchar en la senda del progreso. Basta saber que, con ello, la zona interandina del Ecuador tendrá dos salidas al Pacífico: Guayaquil y Esmeraldas.

La elocuencia del ejemplo ha conmovido a todos los habitantes [illegible] secciones [illegible] cesa- dos actualmente [illegible] cons- trucción [illegible] y si bre [illegible] gan [illegible] la li- nea Si[illegible] aquí es tamos [illegible] ávidamente la conclusión de los estudios de delineación, para cumplir lo que siempre hemos ofrecido: la obra de terraplenes la haremos nosotros.

Se está desarrollando un vasto plan de caminos de hierro y quizá sin comprenderlo nos estamos preparando a la formidable lucha económica q'sobrevendrá tan luego como termine la guerra. Secunde el Gobierno este general entusiasmo y hará labor de previsión y patriotismo.

Rectificación de la línea férrea

Como el Ingeniero, Mr. Carlos Brown, necesitara un nivel de mano y una cinta métrica para practicar ciertos estudios y observaciones que requieren el cometido que debe llenar y que lo está llevando a cabo con acierto, el Sr. Gobernador Dávalos se dirigió en seguida al Departamento del Ramo en demanda de los útiles antes nombrados. El Sr. Ministro de Obras Públicas ordenó sin demora que el Sr. Ingeniero en Jefe del F. C. de Quito a Esmeraldas remita el nivel, el cual vino por correo, con fecha 17 de los corrientes. Así mismo, dió la respectiva autorización para que la cinta seda pidiera a Guayaquil y, en esta virtud, la Gobernación la adquirió en dicha forma.

El telegrama que copiamos a continuación, contiene la or[illegible] den de pago solicitada por el Sr. Colector Especial, para el abono del sueldo que pertenece al Sr. Brown y más gastos que ocasionen los estudios, como también la que se refiere a la inversión de una apreciable cantidad en los trabajos. El interés que demuestran el Gobierno y las autoridades, es una prueba de las buenas intenciones de que están animados en orden a la pronta y feliz realización de la obra y de que ella no tiene obstáculos de parte de quienes están llamados a cumplir y hacer cumplir la Ley.

He [illegible] el telegrama: "Quito, [illegible] 19.—Gobernador.—Sírvase Ud. comunicar al Sr. Colector de fondos de la rectificación de la línea férrea que, a la presentación de los vales respectivos, pague los sueldos del Ingeniero Brown y los gastos que se ocasionen en los estudios de la línea, desde el 1o de este mes hasta nueva orden. Al Ministro de Hacienda he pedido que ordene la entrega al Colector expresado de la cantidad de $ 3.000 mensuales para los gastos que se ocasionaren en los trabajos. La cinta métrica que necesita el Sr. Brown, debe ser pagada

Figura 3.2: Primera Página del Periódico El Tiempo de 1918

El proceso de web scraping permitió obtener las URL de los archivos PDF asociados a cada registro de periódico histórico, para lo que se utilizó la librería BeautifulSoup[3] con el fin de analizar el contenido HTML de las páginas web correspondientes a cada registro y extraer las URL de los archivos PDF. De esta manera, se pudo obtener acceso directo a los periódicos completos y descargarlos para su posterior análisis.

No obstante, durante la implementación del proceso de web scraping, surgieron desafíos técnicos, uno de los más significativos es que el servidor que almacena los periódicos históricos limitaba el número de peticiones realizadas desde una misma dirección IP en un período de tiempo determinado, por lo tanto para evitar ser bloqueados por el servidor, se tuvo que gestionar cuidadosamente el número de solicitudes y establecer pausas estratégicas durante el proceso de extracción.

Como resultado de este proceso de extracción de metadatos y descarga de archivos PDF, se logró recuperar exitosamente la gran mayoría de los registros del repositorio digital. De un total de 15.679 registros disponibles, se recuperaron de manera automática y precisa 15.670 registros. Este alto nivel de éxito en la recuperación de registros demuestra la eficacia y fiabilidad del proceso implementado, descargando un total de 41,78 GB.

Además, con el objetivo de facilitar la interacción con el proceso y permitir a los investigadores realizar descargas selectivas de archivos PDF y explorar los metadatos, se ha desarrollado un widget personalizado llamado PDFDownloaderWidget. Este widget, diseñado utilizando la librería Orange, ofrece una interfaz gráfica intuitiva que permite configurar los parámetros de descarga, visualizar los resultados y exportar los datos para su posterior análisis.

En la Figura 3.3 se muestra la interfaz final del widget, donde el usuario puede ingresar el URL del repositorio, especificar el número de periódicos que desea descargar y, opcionalmente, seleccionar un checkbox que indica si desea descargar todo el repositorio.

Para ilustrar el funcionamiento del widget, se presenta un ejemplo de cómo utilizarlo. Se toma como entrada el repositorio digital de la Casa de la Cultura, cuya URL es la siguien-

[3] Librería con métodos de web scraping en archivos HTML y XML.

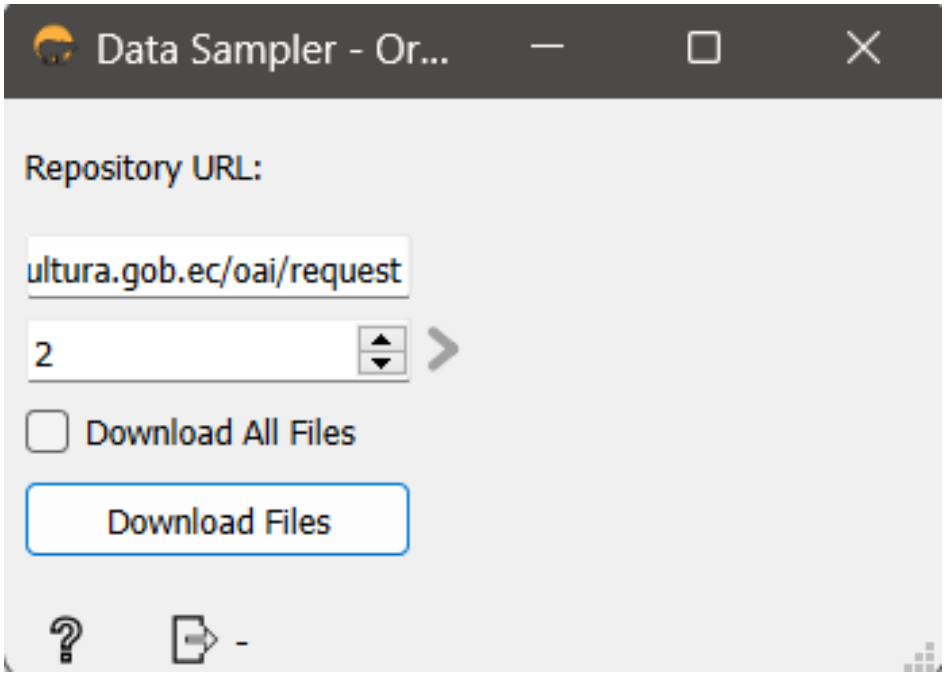

Figura 3.3: Interfaz grafica del widget para la extracción y almacenamiento de datos

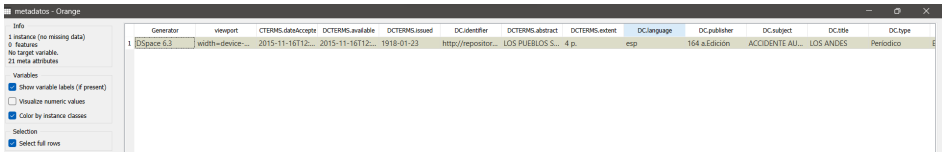

Figura 3.4: Metadatos Obtenidos por el widget de extracción y almacenamiento de datos

te: http://repositorio.casadelacultura.gob.ec/oai/request. Debido a limitaciones de tiempo de procesamiento, se descargará únicamente un periódico.

Una vez configurado todos los atributos de entrada, se puede ejecutar el widget haciendo clic en el botón ”Download Files”. Como resultado de esta acción, se generarán dos tablas. La primera tabla, denominada Tabla de Metadatos (ver Figura 3.4), mostrará los metadatos extraídos del periódico seleccionado. Por otro lado, la segunda tabla, denominada como Tabla de PDFs (ver Figura 3.5), estará relacionada con los archivos PDF obtenidos mediante el proceso de webscraping. Este ejemplo permite visualizar de manera práctica el funcionamiento y los resultados generados por el widget.

Para continuar con el flujo de trabajo, es necesario convertir cada una de las páginas del periódico en imágenes. Con este fin, se ha creado un nuevo widget que toma un archivo PDF y lo transforma en una serie de imágenes. Dentro de la interfaz, el usuario tiene la

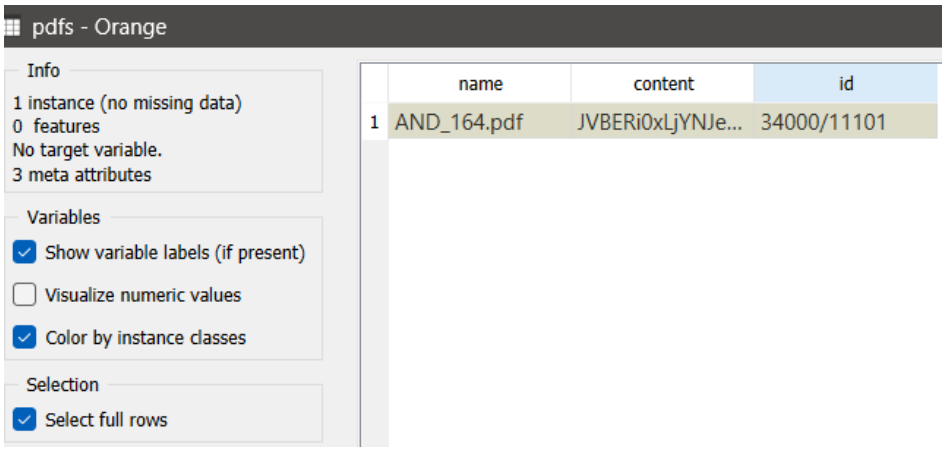

Figura 3.5: PDFs Obtenidos por el widget de extracción y almacenamiento de datos

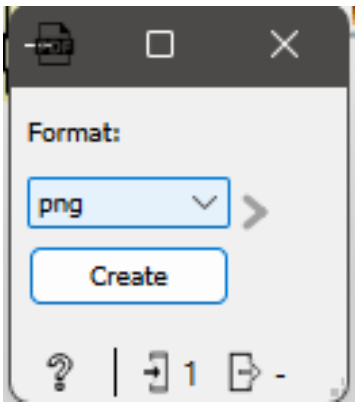

Figura 3.6: Interfaz gráfica del widget para transformar un PDF a imágenes

posibilidad de elegir entre dos formatos: PNG y TIFF, la interfaz se puede apreciar en la Figura 3.6.

El widget de conversión a imágenes genera una imagen por cada página del PDF. Además de la imagen en formato base64, proporciona información adicional como el nombre del periódico, el número de página, el ID del periódico y el ID de la página. Esta información facilita la identificación y organización de las imágenes resultantes, permitiendo un manejo más eficiente de los datos en el posterior análisis y procesamiento, los datos resultantes del periódico escogido para este ejemplo se pueden visualizar en la Figura3.7. Con este widget, se logra una integración fluida entre las diferentes etapas del flujo de trabajo, asegurando una transformación precisa y eficiente del contenido del periódico en imágenes, lo cual es fundamental para las siguientes fases del proceso de automatización.

En resumen, esta implementación presenta un enfoque como parte de un proceso para

	name	image	id	idpage
1	AND_164.pdf Page 1	iVBORw0KGgo...	34000/11101	34000/11101 Page 1
2	AND_164.pdf Page 2	iVBORw0KGgo...	34000/11101	34000/11101 Page 2
3	AND_164.pdf Page 3	iVBORw0KGgo...	34000/11101	34000/11101 Page 3
4	AND_164.pdf Page 4	iVBORw0KGgo...	34000/11101	34000/11101 Page 4

Figura 3.7: Resultados obtenidos tras utilizar el widget para transformar un PDF a imágenes

acceder y aprovechar los recursos de periódicos históricos disponibles en un repositorio digital. El proceso desarrollado permite la extracción de los metadatos de los registros y la descarga de los archivos PDF asociados, superando las limitaciones iniciales del protocolo OAI-PMH.

3.2. Proceso OCR

Dentro del proceso desarrollado en este trabajo de titulación, se incluye una etapa crucial que involucra el reconocimiento óptico de caracteres (OCR). El OCR desempeña un papel fundamental al convertir imágenes bidimensionales de texto, ya sea impreso a máquina o escrito a mano, en texto legible por máquina.

En el contexto específico de este trabajo de titulación, se implementó un enfoque basado en la herramienta OCR denominada Tesseract. Tesseract es un motor de OCR de código abierto desarrollado por HP [54], y ha ganado reconocimiento y adopción en la comunidad científica debido a su precisión y rendimiento. Sin embargo, durante la implementación del OCR en este proceso, se identificaron desafíos y limitaciones que requerían una cuidadosa consideración.

Uno de los desafíos encontrados fue la confusión de columnas en las imágenes, lo cual afectaba la correcta extracción y estructuración del texto. Las imágenes de los periódicos históricos pueden presentar diseños complejos, con múltiples columnas de texto, encabezados y otros elementos gráficos. Esto dificulta la tarea de separar y reconocer adecuadamente el texto de cada columna. Para abordar este desafío, se estan realizando esfuerzos adicionales en otro trabajo de titulación en donde se investiga sobre el preprocesamiento de las imágenes, como la segmentación de columnas y la eliminación de elementos no deseados.

Además, se observó que Tesseract tenía dificultades para reconocer ciertos signos de es-

	name	text	id	idpage
1	AND_164.pdf P...	ye"be ' 23 by —_ ra ic ™| Bee INTERDIARIO INDEPRENDIEN...	34000/11101	34000/11101 P...
2	AND_164.pdf P...	por esa Tesoreria; cuando sesepa el valor exacto se dic'tara ...	34000/11101	34000/11101 P...
3	AND_164.pdf P...	MisiVA Con el mayor plager pallco, & continnacion, a atta c...	34000/11101	34000/11101 P...
4	AND_164.pdf P...	LOSANDESTeatro MaldonadoEl jueves 24 de los corrientes,s...	34000/11101	34000/11101 P...

Figura 3.8: Resultados obtenidos tras utilizar el widget para transformar un PDF a imágenes

critura, como caracteres manuscritos o caligrafía antigua. Esto puede ser atribuido a las diferencias en la forma y estilo de la escritura a lo largo de la historia, así como a la calidad de las imágenes digitalizadas. Para mejorar la precisión del OCR en estos casos, se investigaron técnicas de mejora de imagen, como el aumento del contraste y la reducción del ruido, antes de aplicar el algoritmo de reconocimiento de Tesseract.

Otro aspecto a considerar es el alto tiempo de procesamiento requerido por Tesseract, especialmente al tratar con conjuntos de datos grandes que contienen numerosas imágenes. La extracción de texto de cada imagen puede llevar un tiempo significativo, lo cual puede afectar la eficiencia del proceso general.

El widget diseñado para OCR no requiere de una interfaz de configuración, ya que no recibe parámetros adicionales, dado que, únicamente se ha realizado una instanciación o implementación específica del algoritmo. Simplemente se le suministra la imagen en formato base64, el nombre de la imagen, el identificador del periódico y el identificador de la página. Una vez que estos datos son proporcionados, el widget comienza a trabajar de manera inmediata. Como resultado, se obtiene el nombre de la imagen procesada, el texto extraído de cada una de las imágenes utilizando Tesseract, así como el identificador del periódico y el identificador de la página asociados a cada imagen, el resultado final se puede visualizar en la Figura 3.8.

El código implementado para el procesamiento de OCR se basa en la librería de Orange, así como, la integración de OpenCV y pytesseract, para lo cual se desarrollaron funciones específicas que permiten cargar las imágenes en formato de tabla de datos de Orange, decodificar las imágenes en base64, realizar el procesamiento de imágenes utilizando técnicas de OpenCV y aplicar el OCR mediante la interfaz de pytesseract. Estas funciones se integraron en un widget personalizado denominado **OcrWidget**, que permite la interacción con

el usuario y el procesamiento eficiente de las imágenes.

En resumen, el proceso de OCR implementado en este estudio utiliza el motor Tesseract junto con técnicas de preprocesamiento de imágenes y optimización del rendimiento. A pesar de las limitaciones y desafíos encontrados, se seleccionó Tesseract debido a su capacidad para abordar una amplia gama de desafíos en el reconocimiento de texto. El código desarrollado proporciona una solución para la extracción de texto legible por máquina a partir de imágenes de periódicos históricos, contribuyendo así al objetivo general del estudio de aprovechar los recursos de periódicos históricos disponibles en un repositorio digital.

3.3. Limpieza y preprocesamiento

El proceso de limpieza y preprocesamiento es esencial en el estudio de los textos extraídos de los periódicos históricos, ya que permite preparar los documentos para su posterior análisis y facilita la extracción de información relevante. En este trabajo de titulación, se han aplicado diversos métodos de preprocesamiento con el objetivo de mejorar la calidad y la utilidad de los textos.

Una de las técnicas utilizadas en el preprocesamiento es la conversión de los textos a minúsculas. Esto se realiza para homogeneizar el formato de los textos y evitar ambigüedades en análisis posteriores. Al convertir los textos a minúsculas, se eliminan las diferencias entre mayúsculas y minúsculas, lo que facilita la identificación y el análisis de las palabras clave y las entidades relevantes.

Además, se lleva a cabo la eliminación de los signos de puntuación, los números y las palabras vacías. Los signos de puntuación y los números no suelen ser relevantes para el análisis de los textos históricos, por lo que eliminarlos simplifica el proceso y reduce el ruido en los datos. Asimismo, las palabras vacías, como artículos y preposiciones, carecen de significado contextual y no aportan información relevante, por lo que su eliminación mejora la calidad de los textos procesados.

Otra actividad importante en el preprocesamiento es la eliminación de espacios en blanco innecesarios. Los espacios en blanco adicionales pueden introducir ruido en los textos y dificultar el análisis y la identificación de entidades. Por lo tanto, es crucial eliminar los espacios

Antes de aplicar el Proceso	Después de aplicar el Proceso
Acta del primer Conce- jo de Riobamba	Acta del primer Concejo de Riobamba

Figura 3.9: Ejemplo de unir las palabras que se separan por **-**

Antes de aplicar el Proceso	Después de aplicar el Proceso
ye""be ‘ 23 by —_ ra ic ™| Bee INTERDIARIO INDEPRENDIENT©| aao ir { © RiobumbuCovater—H Mitrenies 23°	yebe by ra ic Bee INTERDIARIO INDEPRENDIENT aao ir RiobumbuCovaterHMit renies

Figura 3.10: Ejemplo de eliminar caracteres especiales

en blanco innecesarios para obtener textos más coherentes y legibles.

Es importante destacar que la elección y la aplicación de las técnicas de preprocesamiento dependen del contexto y los objetivos específicos del estudio. Cada conjunto de datos puede requerir enfoques diferentes para garantizar la calidad y la utilidad de los textos procesados. Por lo tanto, es fundamental adaptar las técnicas de preprocesamiento según las características de los documentos y las necesidades de análisis.

En este caso, para realizar este proceso se creó inicialmente un widget en Orange con los siguientes objetivos:

- Unir las palabras que se separan por el signo **-**: Esto asegura que las palabras que están separadas por un guión sean consideradas como una sola palabra, se puede visualizar un ejemplo en la Figura 3.9.
- Eliminar caracteres especiales: Se utilizan expresiones regulares para eliminar cualquier carácter que no sea una letra, un número o un espacio en blanco, incluyendo caracteres especiales, se puede visualizar un ejemplo en la Figura 3.10.
- Convertir a minúsculas: Todos los textos se convierten a minúsculas para homogeneizar el formato y evitar ambigüedades en el análisis posterior.
- Tokenizar[4] el texto en palabras: Se utiliza la biblioteca NLTK (Natural Language Toolkit) para dividir el texto en palabras individuales, teniendo en cuenta los espacios en

[4]Técnica del procesamiento del lenguaje natural que separa las palabras y los signos de puntuación en entidades conocidas como tokens para su análisis.

blanco.

- Corregir la ortografía de cada palabra: Se utiliza la biblioteca SpellChecker para identificar y corregir palabras mal escritas en el texto. Si una palabra tiene una sugerencia de corrección, se reemplaza por la sugerencia. En caso contrario, se mantiene la palabra original.
- Unir las palabras corregidas en un solo texto: Después de corregir la ortografía de las palabras, se vuelven a unir en un solo texto, separadas por espacios en blanco.

Como una alternativa al algoritmo utilizado en el anterior widget, se planteó crear otra versión aprovechando un LLM. Para este widget en específico, se utilizó el LLM de OpenAI en el modelo GPT-3.5-turbo, una version de Chat GPT, ya que puede considerarse como una opción muy adecuada para la limpieza y corrección de textos debido a sus capacidades avanzadas de procesamiento del lenguaje natural. Aquí hay algunas razones por las que es una buena idea utilizar este modelo:

- Amplio conocimiento lingüístico: GPT-3.5-turbo ha sido entrenado con una gran cantidad de datos de texto y tiene un conocimiento profundo del lenguaje. Esto le permite comprender de manera efectiva la estructura y el significado de los textos en diversos dominios y contextos.
- Capacidad para corregir errores ortográficos y gramaticales: El modelo puede detectar y corregir errores ortográficos y gramaticales en los textos. Esto es especialmente útil para asegurar la precisión y calidad del texto.
- Contextualización de las correcciones: GPT-3.5-turbo tiene la capacidad de analizar el contexto y realizar correcciones coherentes y contextualmente apropiadas [55]. Esto significa que no solo se limita a hacer cambios automáticos basados en reglas, sino que también tiene en cuenta el sentido y la coherencia del texto en su totalidad.
- Flexibilidad y adaptabilidad: Es posible ajustar y personalizar las solicitudes al modelo para que se ajusten a las necesidades específicas del texto. Esto permite controlar el proceso de limpieza y corrección de textos según tus preferencias y requisitos particulares.

- Eficiencia y ahorro de tiempo: Utilizar un modelo de IA como GPT-3.5-turbo para la limpieza y corrección de textos puede ahorrarte una cantidad considerable de tiempo y esfuerzo. El modelo puede procesar grandes volúmenes de texto de manera rápida y precisa, lo que te permite centrarte en otras tareas importantes.

Sin embargo, siempre es importante considerar que, si bien GPT-3.5-turbo es un modelo poderoso, se recomienda revisar y verificar manualmente las correcciones realizadas para asegurarte de que se ajusten y no alteren la información.

Por limitación de tiempo y porque el objetivo principal de este trabajo de titulación no está relacionado con la corrección y análisis de texto, no se realizaron pruebas exhaustivas de los resultados arrojados por el LLM utilizado. Sin embargo, a continuación se muestra un ejemplo de los resultados obtenidos tras procesar con el modelo GPT-3.5-turbo utilizando la primera pagina del periódico cuya imagen se puede ver en la Figura 3.2.

Texto obtenido por el Proceso OCR

```
"ye""be ‘ 23 by -_ ra ic  | Bee INTERDIARIO INDEPRENDIENT©| aao ir { ©
RiobumbuCovater-HMitrenies 23° le Snero te 19/8<=patriotismoLos pueblos
septentrionade la Repiblica, en su ade lleyar a calio la constrircién de una
obra de Ja magjag de lam obva que asidéSpertado el intergs 5ti$mo de toda una
importan’ Gobitéhrevion, 0 tes prot s'ha q‘sobrevendrrno este general
entleLido a supro* visién y patriotismo.tila parte. privs mediante la
cint+odeRecetificaciénoy hard labor de pre:znitud ¢ importancia del
ferro*earril de Quito a Esmeraldanos estan estimulando conJad, estrechozo de
los que juntos, por ijemplo mas slocuentede pa’ Sater feedtismo practice q‘
arse puede. dehen marchar en’ |. senileNunca hapresenciailo el pais del
progreso. Basta saberacsitivos resultados, explosionmas grande de
patmotismo,manifestacién mas clara dela idiosincracia de un pueblo, La
clocuencique la que ahora observamos ha conmoyil) 1complacidos en ¢l
laudable, habitants \empefo de nuestros eompa- seecioitriotas del Norte.La
noble actitud de los imbaburefios vy carchenses que,unidos a los habitantes
detin tan undnime y depot que, con ello, ona inte’randina del Eeuador
```

tendracilicosGuayaquil y EB:Pichincha, trabajan con en‘ -nea Si etusiasmo indeelinalle por ern tamos indo avidamentezar su vasto y rico territorio'la conelu-ion de los estudiosde delineacion, para cumplirlo que siempre hemos ofreci:do: la obra ‘te terraplenes lacon Jas paralelas de acero,ha tenido resonancia en laprensa extranjera, la cual hacomentado fayorablemente, haremos nosotros.senalando como un hecho dig’ Se: desarrollando unno dela mayor alabanza, vasto plan de caminos de hieins6lita decision de los ecuatorianos para impulsar las obras ferroviarias, yenciendorro y quizd sin comprenderlohus estamos preparando alaformidable lucha econémica- nombrados,+ de Obras Publicas fordend sin"dela linea férreaComo el Ingenicro,n para practitos estudios y obsnes que requieren el cometi*do que debe lenar y que loA leyando a cabo con acierto, "el Sr, Gobernador Da* ....."

Texto procesado por el modelo GPT-3.5-turbo

" El Interdiario Independiente. Año III. Riobamba, Ecuador. Martes 13 de enero de 1885. Los pueblos septentrionales de la República, en su afán de llevar a cabo la construcción de una obra de gran magnitud, han despertado el interés no sólo de una importante corporación, sino también de toda la nación. Esta obra, que ha despertado el entusiasmo de todos, es la línea ferroviaria de Quito a Esmeraldas. Está estimulando conjuntamente a los que juntos por, ejemplo, más sólidamente practiquen la lealtad y el patriotismo para marchar en un a sola senda. Este país nunca había presenciado el progreso. Basta saber el activo resultado: la explosión más grande de patriotismo, manifestación más clara de la idiosincrasia de un pueblo. La eloquencia con la que ahora se observa ha conmovido complacida en la noble actitud de los imbabureños, los carchenses, que unidos a los habitantes del Tungurahua, tan unánimes y de potencia para con ello iniciar la gran obra del Ecuador. Tendrá cilicios Guayaquil y Pichincha, trabajan con enérgico entusiasmo, indeclinable por empearse in-duavidamente en la realización de su vasto y rico territorio. La conclusión de los estudios de delineación para cumplir lo que siempre hemos prometido, la obra de los terraplenes, la construcción de las paralelas de acero han tenido resonancia en la prensa extranjera, la cual ha comentado favorablemente. Haremos nosotros señalando como un hecho, digno de grandes alabanzas: el vasto plan de los caminos de hierro. La decisión de los ecuatorianos

para impulsar las obras ferroviarias está encurtando y quizá sin comprenderlo estamos preparándonos para una formidable lucha económica. No solamente el ferrocarril, sino también el ingeniero en jefe para prácticos estudios y observaciones que requieren el cometido que deben llenar y que los leyendo a cabo con acierto. El señor gobernador de ..... "

Como se puede observar en el ejemplo presentado, los resultados parecen ser prometedores. Por lo tanto, es necesario llevar a cabo pruebas más exhaustivas para evaluar cómo el modelo GPT-3.5-turbo de OpenAI puede resultar útil en la corrección y limpieza del texto obtenido mediante el proceso de OCR.

3.4. Reconocimiento de Entidades Nombradas

El reconocimiento de entidades con nombre (NER) es un desafío esencial en el campo del procesamiento del lenguaje natural (NLP) que implica identificar y clasificar nombres y entidades relevantes en un texto determinado. En diversos contextos, como en la extracción de información de noticias, la detección precisa de nombres de personas, organizaciones y lugares adquiere una importancia crucial [32]. El NER desempeña un papel fundamental en diversas aplicaciones, como el análisis de sentimientos, la recuperación de información y la construcción de sistemas de pregunta-respuesta automatizados.

En este trabajo de titulación, para obtener las entidades de texto limpio y preprocesado, se ha implementado un código personalizado que hace uso de la API de OpenAI y el modelo de lenguaje ”text-davinci-002”.

Los LLMs, como el modelo ”text-davinci-002”, pueden representar una solución util para el reconocimiento de entidades en el procesamiento del lenguaje natural. Su capacidad de comprensión del lenguaje, eficiencia y automatización, adaptabilidad y escalabilidad, mejora continua y soporte para múltiples idiomas los convierten en una opción valiosa para abordar el desafío del reconocimiento de entidades en diversas aplicaciones de NLP.

Una vez obtenidas las entidades de los textos, estas tendrán dos fines. El primero de ellos será utilizarlas para poblar el grafo del conocimiento en una instancia de una propiedad denominada "palabras clave". El segundo proceso que se realiza con las entidades extraídas es de gran importancia, ya que busca aprovechar la riqueza de conocimiento y la estructura

semántica de DBpedia para enriquecer el análisis de los textos.

Al buscar la representación semántica de las entidades extraídas en DBpedia, se obtiene información adicional valiosa que puede complementar y enriquecer el análisis de los textos. Algunas de las ventajas de este enfoque son:

- Ampliación del contexto: Al acceder a la información en DBpedia relacionada con una entidad en particular, se obtiene un contexto más completo y detallado sobre dicha entidad. Esto puede incluir detalles biográficos, ubicaciones geográficas, relaciones con otras entidades, eventos históricos relacionados y mucho más. Esta información adicional puede brindar una comprensión más profunda y precisa de las entidades mencionadas en los textos analizados.
- Enlaces a recursos relacionados: DBpedia enlaza las entidades con recursos relacionados a través de propiedades semánticas. Al explorar estos enlaces, se puede acceder a una red de conocimiento interconectada, lo que permite descubrir relaciones más amplias y establecer conexiones con otros conceptos relevantes. Esto ayuda a contextualizar las entidades dentro de un marco más amplio y facilita el descubrimiento de información adicional relacionada.
- Consistencia y normalización de datos: DBpedia proporciona una estructura semántica bien definida y una nomenclatura estandarizada para las entidades y sus propiedades. Esto ayuda a garantizar la consistencia y la normalización de los datos obtenidos, lo que facilita su integración y comparación con otros conjuntos de datos y sistemas de información.
- Integración con otras fuentes de datos: DBpedia se basa en los principios de datos enlazados, lo que permite su integración con otras fuentes de datos enlazados. Esto significa que se puede combinar la información de DBpedia con otras bases de conocimiento o conjuntos de datos, lo que amplía aún más las posibilidades de análisis y descubrimiento de información.

Por último, para concluir, se mostrará el resultado obtenido al utilizar el texto de la primera página del periódico después de aplicarle el proceso de limpieza y preprocesamiento (ver

Entidad Encontrada	URIs
Riobamba	”http://dbpedia.org/resource/Riobamba”
Ecuador	”http://dbpedia.org/resource/Ecuador”
El Interdiario Independiente	”http://dbpedia.org/resource/Independiente”
Martes 13 de enero de 1885	No encontrada
Los pueblos septentrionales de la República	”http://es.dbpedia.org/resource/Pueblos”, ”http://dbpedia.org/resource/Republic”
línea ferroviaria de Quito a Esmeraldas	”http://es.dbpedia.org/resource/Línea”, ”http://es.dbpedia.org/resource/Ferroviaria”, ”http://es.dbpedia.org/resource/Categoría: Quito”, ”http://es.dbpedia.org/resource/Categoría: Esmeraldas”
lealtad	”http://es.dbpedia.org/resource/Lealtad”
patriotismo	”http://es.dbpedia.org/resource/Categoría: Patriotismo”
idiosincrasia	”http://es.dbpedia.org/resource/Idiosincrasia”
imbabureños	”http://es.dbpedia.org/resource/Categoría: Imbabureños”
carchenses	”http://es.dbpedia.org/resource/Categoría: Carchenses”
habitantes del Tungurahua	”http://es.dbpedia.org/resource/Habitantes”, ”http://es.dbpedia.org/resource/Tungurahua”
Guayaquil	”http://es.dbpedia.org/resource/Categoría: Guayaquil”
Pichincha	”http://dbpedia.org/resource/Pichincha”
Carlos Brown	No encontrada
ecuatorianos	”http://es.dbpedia.org/resource/Ecuatorianos”
equilibristas	”http://es.dbpedia.org/resource/ MIR_(serie_de_televisión)__22__1”
juglares	”http://es.dbpedia.org/resource/Juglares”

Tabla 3.1: Tabla de Entidades.

sección 3.3). Además, se puede observar el contenido de la tabla 3.1 que presenta los resultados obtenidos.

3.5. Obtención de Word Embeddings

Como parte adicional de este trabajo de titulación, se incorporo la obtención de word embeddings de los periódicos históricos. Esta elección se fundamenta en el potencial que los word embeddings ofrecen para la automatización y enriquecimiento de los resultados en la explotación de la información. Los word embeddings son representaciones numéricas de palabras o frases que no solo capturan su forma, sino también su significado semántico y las relaciones existentes entre ellas. Estas representaciones se obtienen mediante modelos

Texto	Embedding
El Interdiario Independiente Año III Riobamba Ecuador Martes 13 de enero de 1885 Los pueblos septentrionales de la República en su afán de llevar a cabo la construcción de una obra de gran magnitud han despertado el interés no sólo de una importante corporación sino también de toda la nación Esta obra que ha despertado el entusiasmo de todos es la línea ferroviaria de	[-0.03163599595427513, -0.0029353511054068804, 0.007983611896634102, -0.02079673297703266, -0.005938532296568155, ...]
respectiva autorización para que la cinta se fuera pedida a Guayaquil En esta virtud la gobernación la adquirió en dicha forma El telegrama que copiamos a continuación contiene el ordenamiento Carlos Brown necesitará un nivel de plano y una cinta Has el envió de pago solicitada por el patrio Termine la guerra secunde Olmos El señor colector especial para el abono del sueldo que pertenece al señor Brown y más gastos que ocasionen los estudios como también la que se refiere a la inversión de una aparentidad en los trabajos	[-0.026859600096940994, 0.00053850666154176, -0.004297940526157618, -0.0262258630245924, -0.012843256816267967, 0.014710754156112671, ...]
10 de este mes hasta nueva orden Al ministro de Hacienda he pedido que ordene la entrega al colector expresado de la cantidad necesaria para los gastos que se ocasionen en los trabajos La cinta métrica que necesita el señor Brown debe ser pagada Todo género de obstáculos que en otra tierra y con otros hombres habrían producido el fracaso del salvador proyecto\\\\La prensa nacional no necesita ilustrar el criterio de las masas para llevarles el convencimiento de que el silbato de la locomotora es anuncio de prosperidad y un hallo despertar a la vida modernizada.....	[-0.02253476344048977, -0.017509929835796356, 0.0034694436471909285, -0.006655453238636255, 0.005927622318267822, 0.0006162942736409605, ...]

Tabla 3.2: Tabla de Embendings.

de lenguaje avanzados, entrenados en grandes volúmenes de datos textuales, lo que les permite capturar de manera efectiva la semántica subyacente del lenguaje [56].

Para obtener los word embeddings, se implemento una función en Python. Esta función tiene como objetivo principal obtener el embedding correspondiente a un texto dado utilizando la API de OpenAI. Antes de llamar a esta función, se realiza un preprocesamiento para garantizar un texto claro, eliminando caracteres especiales y secuencias de escape que podrían afectar la calidad del embedding resultante. Este paso es fundamental para asegurar que el modelo de OpenAI pueda generar una representación precisa y coherente del texto.

Es importante tener en cuenta que, hasta el momento, las librerías y herramientas disponibles para obtener word embeddings tienen ciertos límites en cuanto a la longitud del texto

que pueden procesar de manera eficiente. Por esta razón, se incorporo una verificación de la longitud del texto y, en caso de que supere un límite máximo predefinido, se divide el texto en secciones más pequeñas para obtener embeddings individualmente. El resultado final se presenta como una lista de diccionarios, donde cada diccionario contiene el texto original y su correspondiente embedding.

Al utilizar esta tecnología, se mejora la calidad y eficacia del análisis, así como también una comprensión más profunda de los textos históricos que se estan investigando. Los word embeddings permiten identificar similitudes y patrones ocultos en los datos, lo que facilita la exploración y el descubrimiento de información relevante. Además, la capacidad de comparar y contrastar diferentes textos de manera más efectiva, abre nuevas perspectivas y posibilidades para futuras investigaciones en este campo.

Al extraer los embeddings utilizando el código creado para este proceso, se obtendrán los textos individuales junto con sus respectivos embeddings. El propósito de estos datos es utilizarlos posteriormente para poblar el grafo del conocimiento. En la tabla 3.2 se puede observar el resultado obtenido de la primera página del periódico utilizado en cada uno de los ejemplos de los procesos anteriores.

3.6. Generación del grafo del conocimiento

Una vez obtenida toda la información, ésta debe ser almacenada en algún lugar para ser utilizada y, al igual que en [51], se optó por una ontología. Las ontologías permiten no solo almacenar la información sino también relacionarla con la que ya se encuentra disponible en la web, reutilizando lo establecido así como interconectando y dando más significado a los resultados del proceso. De esta manera para implementar la ontología se realizaron los siguientes pasos:

- Modelar y encontrar recursos a nivel de la web que representen la información de las tablas de la herramienta, modelando los que no se logren encontrar.
- Desplegar la ontología y poblarla.

3.6.1. Modelación y búsqueda de recursos a nivel de la web

El primer paso es identificar los sujetos o actores principales del modelo. En este caso se parte desde el objeto periódico y sus metadatos. En este caso los metadatos son aquellos que se encuentran disponibles dentro de la web de la casa de la cultura y definidos dentro del protocolo OAI-PMH, los cuales son:

- Name (Nombre)
- Generator (Herramienta generadora)
- Viewport (Dimensiones)
- Date Accepted (Fecha de aceptación)
- Date Available (Fecha de habilitación)
- Date Issued (Fecha en la que ocurrió)
- Identifier (Identificador)
- Abstract (Resumen)
- Extent (Extensión o número de páginas)
- Language (Idioma)
- Publisher (Editor)
- Subject (Tema)
- Title (Título)
- Type (Tipo)
- Citation Keywords (Palabras clave para citación)
- Citation title (Título para citación)
- Citation Publisher (Editor para citación)
- Citation Language (Idioma para citación)
- Citation PDF Url (Url para citación)

- Citation Date (Fecha para citación)
- Citation Abstract HTML Url (Url del resumen para citación)

Dado que varios de estos campos hacen referencia a una “Citación” se optó por modelar un nuevo sujeto con ese nombre y así todos los datos relacionados van a este último, el cual se encuentra directamente relacionado con “Periódico”.

Cabe resaltar que dentro de los metadatos se encuentra “Type”, el cual representa el tipo de documento al que se hace referencia, sin embargo como se parte de la base de que el documento ya es un periódico, este campo se suprime. El modelo resultante se puede apreciar en la Figura 3.11.

Con el modelo base definido, el siguiente paso consta de añadir la información recuperada del proceso. A continuación se listan los factores extra a considerar:

- Paginas
- Texto
- Entidades
- Embeddings

Para modelar estos datos primero se ha definido la jerarquía entre las clases de la siguiente manera: Un periódico contiene páginas; Las páginas contienen textos y tienen entidades; Los textos contienen tanto su valor literal como sus embeddings. El resultado se puede apreciar en la Figura 3.12.

Con el modelo de ontología definido, el siguiente paso es encontrar recursos a nivel de la web que definan la mayoría de términos asociados, es así que se han buscado y vinculado al vocabulario de schema.org. Como resultado se emparejaron los términos como se ve en la tabla 3.3. Además para los casos donde no existe una traducción directa se crearon los recursos dentro del espacio de nombres ”http://newsont.com/”que representa todos los términos propios. En resumen, este punto consiste en tomar los términos de schema.org y ampliarlos con los términos que sean necesarios, almacenando estos junto a las instancias.

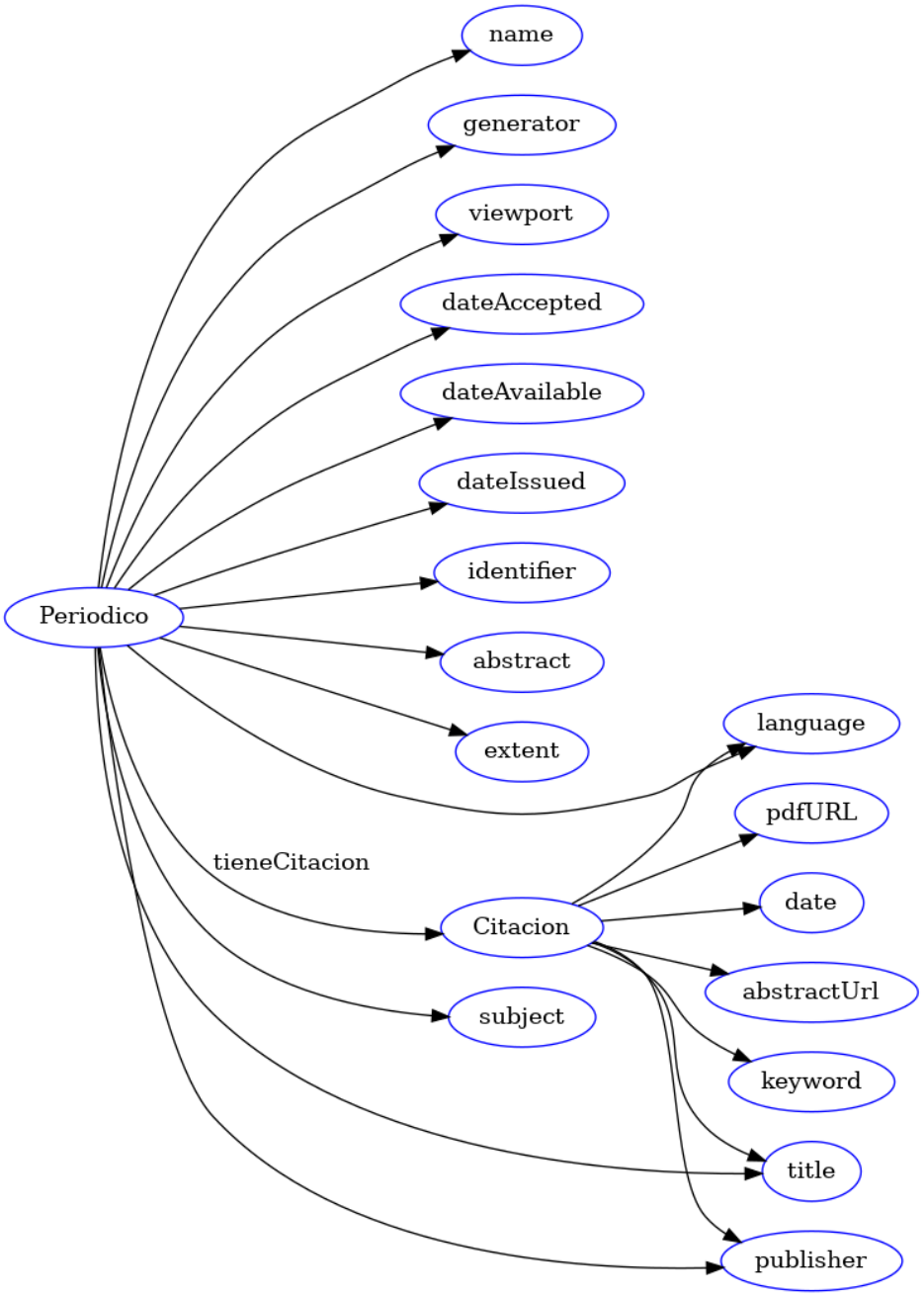

Figura 3.11: Modelo base.

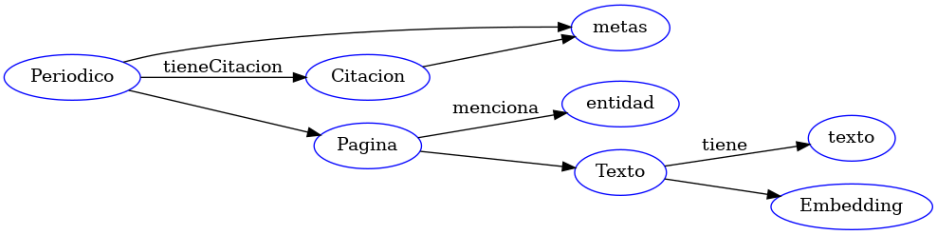

Figura 3.12: Modelo ampliado con metadatos comprimidos.

Término	Recurso
Periódico	https://schema.org/Newspaper
Citación	http://newsont.com/Citation
Generator (Herramienta generadora)	http://newsont.com/generator
Viewport (Dimensiones)	https://schema.org/size
Date Accepted (Fecha de aceptación)	http://newsont.com/dateAccepted
Date Available (Fecha de habilitación)	https://schema.org/datePublished
Date Issued (Fecha en la que ocurrió)	https://schema.org/sdDatePublished
Identifier (Identificador)	https://schema.org/url
Abstract (Resumen)	https://schema.org/abstract
Extent (Extensión)	https://schema.org/materialExtent
Language (Idioma)	https://schema.org/inLanguage
Publisher (Editor)	https://schema.org/version
Subject (Tema)	https://schema.org/about
Title (Título)	https://schema.org/headline
Citation Keywords (Palabras clave para citación)	https://schema.org/keywords
Citation title (Título para citación)	https://schema.org/headline
Citation Publisher (Editor para citación)	https://schema.org/version
Citation Language (Idioma para citación)	https://schema.org/inLanguage
Citation PDF Url (Url para citación)	https://schema.org/url
Citation Date (Fecha para citación)	https://schema.org/sdDatePublished
Citation Abstract HTML Url (Url del resumen para citación)	http://newsont.com/citationAbstractUrl
Página	http://newsont.com/Page
Entidad	https://schema.org/mentions
Texto (Concepto)	http://newsont.com/Text
texto (literal)	https://schema.org/text
Embedding	http://newsont.com/embedding

Tabla 3.3: Términos emparejados.

3.6.2. Despliegue y población

Una vez que la ontología ha sido definida a nivel de RDF, es necesario crear el espacio donde toda esta información será alojada, partiendo desde la estructura base y preparada para recibir instancias del esquema. Como se explicó en el capítulo de Fundamentos Tecnológicos 3, la herramienta elegida para este trabajo es Fuseki. Por ello se alojó el esquema para posteriormente añadir sus instancias a través del SPARQL *endpoint* que tiene incorporado, esta herramienta se aloja como un servidor local.

Para poblar la ontología desde el flujo en la herramienta, se desarrolló un script en python, que se encarga de tomar los datos de cada página y convertirlos en tripletas que se insertan

a través del endpoint de Fuseki. Primero se hace un recorrido por los metadatos, posteriormente las páginas con sus entidades y para finalizar los textos con sus embeddings. Cabe resaltar que dada la forma de reconocer la estructura de las peticiones a través de los servicios REST no es posible enviar cualquier tipo de carácter especial, ya que estos pueden arruinar la estructura y provocar una excepción. Por ello se suprimieron los caracteres especiales reservados del formato ttl (Tripletas Turttle RDF) del texto. El modelo del script funciona como un insert sql, es decir, no crea la ontología desde el inicio cada que se ejecuta el algoritmo, sino que simplemente añade información nueva al servidor.

3.7. Explotación

El grafo de conocimiento generado necesita ser explotado de manera efectiva para garantizar que el proceso realizado sea verdaderamente útil en el contexto del trabajo de Titulación. Con este objetivo en mente, se tomó la decisión de crear un buscador web utilizando la librería de Python llamada Gradio. Gradio es una potente librería de Python que facilita la creación de interfaces de usuario interactivas y personalizadas para modelos de aprendizaje automático [57]. Con Gradio, es posible construir rápidamente interfaces intuitivas que permiten a los usuarios interactuar con los modelos y obtener resultados en tiempo real. Esta librería simplifica en gran medida el proceso de desarrollo de aplicaciones basadas en modelos de aprendizaje automático, ya que no se requiere conocimiento profundo en programación de interfaces de usuario. Gracias a su facilidad de uso y flexibilidad, Gradio ha ganado popularidad en la comunidad de ciencia de datos y se ha convertido en una herramienta muy valorada para la creación de aplicaciones y demostraciones interactivas de modelos de aprendizaje automático. El diseño final del buscador, representado en la Figura 3.13, refleja el las funcionalidades implementadas.

El funcionamiento de este buscador se basa en una serie de pasos. En primer lugar, se recibe la consulta del usuario, que puede ser una frase o una palabra clave relacionada con los textos históricos de interés. A continuación, se utiliza la técnica de word embedding para obtener la representación numérica de dicha consulta. Los word embeddings, que capturan no solo la forma de las palabras, sino también su significado semántico y las relaciones entre ellas, se obtienen mediante modelos de lenguaje avanzados tal y como se explica en

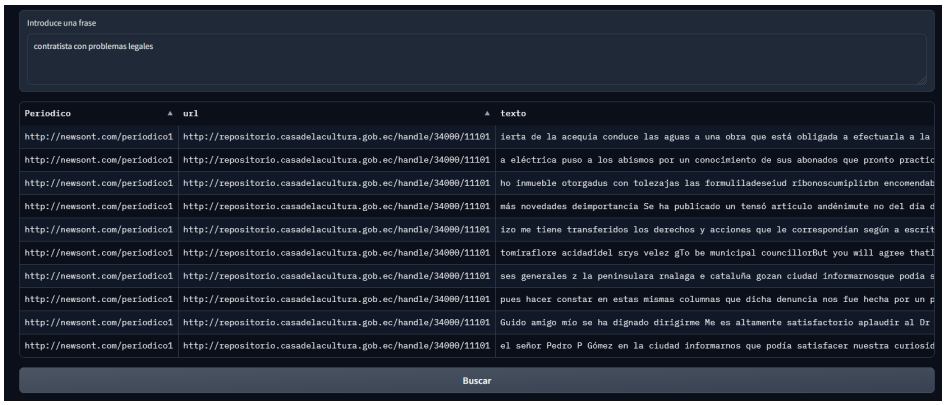

Figura 3.13: Prototipo de buscador para explotar el grafo del conocimiento

la sección 3.5. Este enfoque permite capturar de manera efectiva la semántica subyacente en el lenguaje y obtener una representación densa y continua del mismo.

Una vez obtenido el word embedding de la consulta, se procede a comparar con los word embeddings de los textos presentes en grafo del conocimiento. Aquí es donde entra en juego la técnica de similitud de coseno[5]. La similitud de coseno calcula el ángulo entre dos vectores en un espacio multidimensional y devuelve un valor que indica su grado de similitud. Al calcular la similitud de coseno entre el word embedding de la consulta y los word embeddings de los textos obtenidos del grafo del conocimiento, se puede determinar cuáles son los textos más parecidos a la consulta en términos semánticos.

Para este trabajo de titulación se utilizo la técnica de similitud de coseno, pero existen algunas más y seria de suma importancia hacer un estudio para ver cual seria la mejor opción para obtener mejores resultados. La simulitud conseno en primer lugar, es eficiente computacionalmente, lo que permite realizar búsquedas rápidas incluso en grandes conjuntos de datos. Además, la similitud de coseno no solo tiene en cuenta la presencia de palabras clave en los textos, sino también la similitud semántica entre ellos. Esto significa que el buscador puede identificar similitudes y patrones ocultos que van más allá de las palabras individuales, lo que facilita la exploración y el descubrimiento de información relevante en los textos históricos.

Una vez realizada la comparación de similitud de coseno, los resultados se presentan al usuario de manera ordenada, mostrando los textos que tienen una mayor similitud con la

[5]Alogirtmo que mide la similitud entre dos vectores al calcular el coseno del ángulo entre ellos

consulta. Estos resultados pueden incluir extractos de los textos relevantes, así como enlaces a las fuentes originales para acceder a información más detallada. Esta presentación clara y concisa de los resultados ayuda al usuario a navegar y explorar los textos históricos de manera eficiente y efectiva.

La utilización de la técnica de similitud de coseno en combinación con los word embeddings en el buscador proporciona una capacidad de búsqueda más avanzada y precisa en comparación con los métodos tradicionales basados únicamente en palabras clave o en la coincidencia exacta de términos. Al considerar la similitud semántica entre la consulta y los textos en la base de datos, el buscador puede encontrar resultados relevantes incluso cuando las palabras utilizadas no coincidan exactamente.

El buscador web desarrollado en este proyecto combina la potencia de los word embeddings y la técnica de similitud de coseno para ofrecer una experiencia de búsqueda enriquecida y efectiva en el ámbito de los textos históricos. Esta tecnología permite a los investigadores explorar y descubrir información relevante de manera más eficiente, mejorando la calidad y la eficacia de sus análisis. Además, el uso de word embeddings y la técnica de similitud de coseno abren nuevas posibilidades en futuras investigaciones al permitir la identificación de patrones ocultos y la comparación y contrastación de diferentes textos de manera más efectiva.

4. Evaluación de la validez de la herramienta

En este capítulo se evalúa la validez de la herramienta creada para la automatización del proceso de extracción, almacenamiento, descripción y visualización de datos de periódicos históricos digitalizados. Para ello, se han llevado a cabo la técnica SUS con una pequeña modificación que permitirá conocer la percepción de los expertos sobre la validez de la herramienta. Con el análisis de los resultados obtenidos, se podrán identificar si los expertos perciben valida la herramienta.

En este contexto, se reconoce que los usuarios que pueden llegar a utilizar dicha herramienta conforman un grupo pequeño con experiencia previa, ya sea el conocimiento técnico en ciencias de la computación o afines a hemerotecas. Es por esta razón que se ha decidido utilizar el método de evaluación conocido como juicio de expertos[58]. El juicio de expertos implica la participación de personas con experiencia y conocimientos especializados en el campo de estudio, en este caso, la extracción de información de periódicos históricos[59].

La utilización del juicio de expertos en esta evaluación proporciona varias ventajas. En primer lugar, permite obtener evaluaciones y opiniones fundamentadas por parte de expertos reconocidos en el campo. Su experiencia y conocimientos en el tema aseguran que las evaluaciones sean confiables y basadas en un criterio sólido.

Además, el juicio de expertos aporta una perspectiva especializada que complementa la visión de los usuarios finales. Los expertos pueden evaluar aspectos técnicos, funcionales y conceptuales de la herramienta, identificando posibles mejoras y optimizaciones desde su conocimiento profundo del tema.

Otra ventaja del juicio de expertos es su capacidad para proporcionar una validación adicional a los resultados obtenidos. Al involucrar a expertos en la evaluación de la herramienta, se refuerza la confiabilidad de los hallazgos y se aumenta la credibilidad de los resultados de la investigación.

En cuanto al grupo reducido de usuarios que pueden llegar a utilizar la herramienta, es importante destacar que contar con la opinión de expertos en el campo es especialmente valioso

en este caso. Dado que los expertos poseen un conocimiento profundo y especializado, su evaluación puede ser exhaustiva y detallada, identificando aspectos y consideraciones que podrían pasar desapercibidos para usuarios menos familiarizados con el tema.

4.1. Diseño de la evaluación

En el campo de la evaluación de sistemas y productos, el System Usability Scale (SUS) es una herramienta ampliamente utilizada para medir la usabilidad percibida [60]. Sin embargo, en ciertos contextos, es igualmente importante evaluar la validez de un sistema, es decir, la capacidad de un sistema para cumplir su propósito y proporcionar resultados precisos y confiables.

Para abordar esta necesidad de evaluar la validez, se propone una modificación del SUS denominada System Validity Scale (SVS). La SVS se basa en el mismo enfoque de evaluación subjetiva mediante una serie de afirmaciones que los usuarios deben calificar según su grado de acuerdo o desacuerdo.

La SVS se compone de una serie de ítems que abordan directamente la validez percibida del sistema. Estos ítems se centran en aspectos como la exactitud de los resultados, la integridad de la información proporcionada, la confiabilidad de las funcionalidades y la consistencia de las respuestas del sistema.

Al utilizar la SVS, los usuarios evalúan la validez del sistema en función de su experiencia y conocimiento sobre el dominio en el que opera la herramienta. Esto permite obtener información valiosa sobre la confianza que los usuarios depositan en los resultados y la precisión de la información proporcionada por el sistema.

La SVS se administra después de que los usuarios hayan interactuado con el sistema y hayan tenido la oportunidad de experimentar su funcionalidad y resultados. Los usuarios califican cada afirmación en una escala de acuerdo, desde ”Totalmente en desacuerdo”hasta ”Totalmente de acuerdo”.

La evaluación de la validez se ha llevado a cabo con el fin de recopilar las opiniones y experiencias de expertos en la interacción con la herramienta. Estos expertos han sido seleccionados cuidadosamente para asegurar una representación diversa de conocimientos

y perspectivas. Su experiencia en el campo de los repositorios digitales y su conocimiento sobre periódicos antiguos los convierten en evaluadores idóneos para identificar fortalezas y debilidades en la validez de la herramienta.

4.1.1. Objetivo de la evaluación

El objetivo de esta evaluación de validez es obtener información detallada y fundamentada sobre la experiencia de los expertos al interactuar con la herramienta e identificar áreas que necesiten mejoras y obtener retroalimentación para poder realizar ajustes que conduzcan a una mejor experiencia.

4.1.2. Preguntas de investigación de la evaluación

Para llevar a cabo la evaluación de la validez de la herramienta para la automatización del proceso, se creyó conveniente plantear la siguiente pregunta:
$P1_1$ La herramienta ¿es percibida por los expertos como válida?

4.1.3. Variables

La variable independiente en este estudio es la herramienta para la automatización. Mientras que la variable dependiente es la validez percibida por los expertos, para medir esta variable se utiliza el cuestionario SVS, que se detalla en la sección 4.1.5

4.1.4. Selección de la muestra

Los participantes seleccionados para esta evaluación son profesionales especializados en el campo de las ciencias de la computación, con amplia experiencia en el desarrollo y uso de repositorios digitales para bibliotecas. Además, también se ha contado con personas expertas en el tema de periódicos antiguos, quienes poseen un profundo conocimiento sobre el tema. Se conto con un total de 4 expertos de los cuales 3 son ingenieros y uno es experto en hemerotecas. Cabe destacar que uno de los profesionales seleccionados es uno de los autores del trabajo base [51]. Su contribución en el desarrollo de la herramienta y su conocimiento detallado del contexto de investigación han sido fundamentales para el proceso de evaluación.

4.1.5. Cuestionario

Dentro de la evaluación se utilizará un cuestionario para medir la validez percibida de la herramienta. El cuestionario es validado y reconocido en el campo de la validez, como el cuestionario SVS.

Para la evaluación de la validez, se utiliza un cuestionario basado en la Escala de Usabilidad del Sistema (SUS, por sus siglas en inglés - System Usability Scale) con una variación específica para medir la Validez (SVS - System Validity Scale). Este cuestionario consta de diez preguntas que se califican en una escala de Likert de cinco puntos. Estas preguntas están diseñadas de manera general y aplicable a cualquier tipo de sistema, producto o servicio, con el objetivo de medir la validez del mismo.

Las preguntas del cuestionario se listan a continuación:

1. En general, considero que la herramienta es relevante en el campo de la búsqueda e indexación de Periódicos Históricos.
2. En general, percibo que la herramienta no automatiza eficientemente el proceso de extracción de información de periódicos históricos.
3. En general, considero que la herramienta automatiza eficientemente el proceso de indexación de información de periódicos históricos.
4. Creo que el flujo de trabajo presentado no es más eficiente que el proceso manual que se ha estado utilizando hasta el día de hoy.
5. En general, considero útil la optimización de los algoritmos individuales presentados para mejorar los resultados en futuros trabajos.
6. En general, considero útil añadir nuevos algoritmos para mejorar los resultados en futuros trabajos.
7. En general, percibo que la herramienta no tiene la capacidad de automatizar eficientemente el proceso de extracción de información de periódicos históricos.
8. En general, percibo que el flujo de trabajo presentado no es más eficiente en compa-

ración con el proceso manual que se ha estado utilizando hasta el día de hoy.

La escala de Likert se presenta con las siguientes opciones:

1. Completamente en desacuerdo
2. En desacuerdo
3. Neutro
4. De acuerdo
5. Completamente de acuerdo

4.2. Ejecución de la evaluación

Una vez seleccionados los participantes y el cuestionario a utilizar en la evaluación, se procedió a su ejecución. La evaluación se llevó a cabo en dos sesiones: una sesión de capacitación y seguidamente de una sesión de evaluación.

4.2.1. Sesión de capacitación

Antes de llevar a cabo la evaluación, se llevaron a cabo sesiones de capacitación con los expertos seleccionados. Durante estas sesiones, se presentó inicialmente la forma manual descrita en [51], con el objetivo de brindar a los expertos un entendimiento claro de dicho enfoque. Una vez que los expertos adquirieron conocimientos sobre la forma manual, se procedió a presentar la herramienta de automatización que sería objeto de evaluación.

En estas sesiones, se proporcionaron detalles exhaustivos sobre el funcionamiento y las características de la herramienta. Se destacaron sus capacidades y se explicó cómo podía facilitar el proceso de validación. Además, se brindó orientación detallada sobre cómo completar los cuestionarios y responder adecuadamente a las preguntas planteadas.

4.2.2. Sesión de evaluación

La mayoría de las sesiones de evaluación se llevaron a cabo en la biblioteca del campus Central de la Universidad de Cuenca, mientras que una de ellas se realizó a través de la plataforma Zoom con la participación de uno de los autores del artículo [51]. Durante las sesiones, se les proporcionó a los participantes un cuestionario para recopilar sus opiniones

y criterios de evaluación. Para facilitar este proceso, se preparó una guía en formato de formulario en línea utilizando Google Forms.

4.3. Análisis de resultados

En esta sección se presenta un análisis de los resultados obtenidos en la evaluación de la validez de la herramienta, según la percepción de los expertos. Se examinaron los datos recopilados a través del cuestionario completado por los 4 expertos participantes.

4.3.1. Análisis de la validez

Para evaluar el cuestionario SVS (System Validity Scale), se registra el puntaje obtenido por cada experto. Al igual que en el cuestionario SUS, el puntaje final del SVS es un valor numérico que varía entre 0 y 100, y refleja el nivel de validez percibido por los expertos. El cálculo del puntaje SVS se realiza de manera similar a como se calcula el puntaje SUS, utilizando las respuestas proporcionadas en las preguntas específicas de la escala de validez.

1. Se suman las puntuaciones de las preguntas impares, luego se resta 5 del total.
2. Se suman las puntuaciones de las preguntas pares, luego se resta ese total de 25.
3. Se suman ambos resultados y se multiplica por 2,5.

Se utilizará la misma escala que la del SUS para evaluar el SVS. La escala de puntuación del SVS se representa en la Figura 4.1, donde se establecen diferentes rangos de puntuación y se les asigna un adjetivo correspondiente. Los puntajes inferiores a 50 se consideran ”No válidos”, mientras que las puntuaciones entre 51 y 70 se califican como ”Marginales”. Por otro lado, las puntuaciones superiores a 71 se consideran ”Válidos”. Cada puntuación se asocia con un adjetivo específico según su rango: ”Muy Baja” para el rango de 0-25, ”Baja” para el rango de 25-50, ”Moderada” para el rango de 50-70, ”Alta” para el rango de 70-80, ”Muy Alta” para el rango de 80-90 y ”Excelente” para valores superiores a 90.

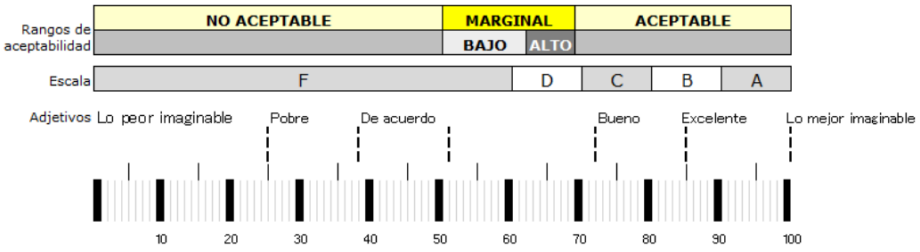

Figura 4.1: Escala de puntuaciones SUS (Tomado de [1]).

Se procedió a calcular el puntaje SVS para cada uno de los expertos, y los resultados obtenidos se presentan en la Figura 4.2. Se observó que el puntaje mínimo registrado fue de 90, mientras que el puntaje máximo alcanzó los 97,5. En cuanto a las medidas de tendencia central, se obtuvo una media de 94.37. El puntaje final del SUS se estableció en base a la media, lo cual indica que los expertos perciben la herramienta de automatización como valida y se sitúa en la categoría de **Aceptable**. La mayoría de los participantes obtuvieron una puntuación superior a la media, lo que sugiere que la aplicación es percibida de manera positiva en términos de validez.

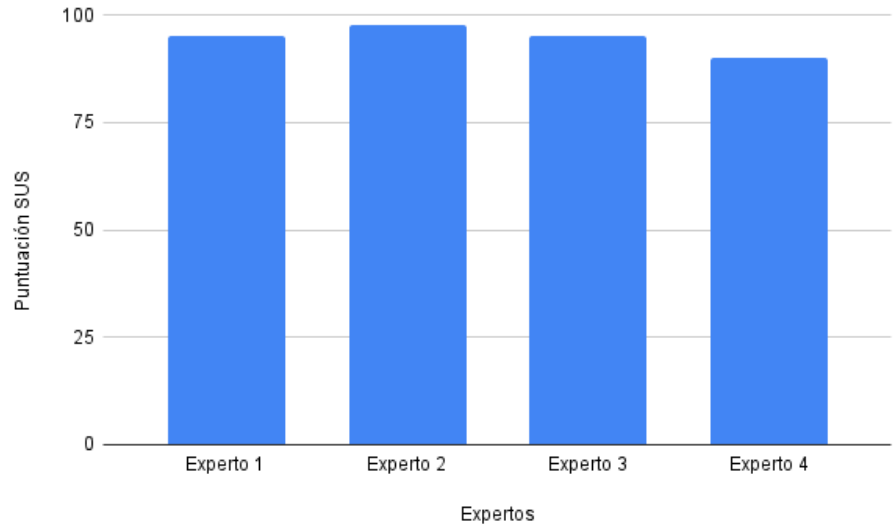

Figura 4.2: Puntuaciones obtenidas en el SVS.

5. Conclusiones

En este capítulo, se expondrán las conclusiones derivadas del trabajo de titulación llevado a cabo, donde se evaluarán los logros alcanzados en relación a los objetivos generales y específicos establecidos para el desarrollo de la herramienta de automatización para la obtención de un grafo del conocimiento y posibles trabajos futuros.

5.1. Conclusiones

El objetivo general planteado para este trabajo de titulación es: **Automatizar el proceso de extracción, almacenamiento, descripción y visualización de datos de periódicos históricos digitalizados, que permita generar un grafo de conocimiento relacionado con diferentes tipos de eventos ocurridos en Ecuador en los siglos XIX-XX.** El propósito principal de este objetivo fue generar un grafo de conocimiento relacionado con diversos tipos de eventos ocurridos en Ecuador durante los siglos XIX-XX.

Este objetivo se cumplió gracias a que se logró desarrollar una herramienta que permite la extracción automática de datos de los periódicos históricos digitalizados de la Casa de la Cultura del Ecuador. Mediante el uso de técnicas de procesamiento de lenguaje natural, webscraping, reconocimiento de patrones, se logró extraer información relevante de los textos, como fechas, nombres de personas, lugares y eventos y obtención de embendings. Asimismo, se implementó un grafo del conocimiento que permite la organización y estructuración de los datos extraídos. Esta solución garantiza un acceso rápido y fácil a la información, facilitando la realización de consultas y búsquedas específicas.

La descripción de los datos se realizó mediante la asignación de etiquetas y metadatos que permiten una mejor comprensión y clasificación de la información. Estos descriptores permiten identificar y relacionar los eventos ocurridos en Ecuador durante los siglos XIX-XX, estableciendo conexiones y patrones significativos en el grafo de conocimiento generado.

Por último, se desarrolló una interfaz visual que permite la visualización intuitiva y dinámica de los datos extraídos y organizados. Esta interfaz brinda a los usuarios la posibilidad de explorar y navegar por el grafo de conocimiento, descubriendo relaciones, tendencias y

eventos históricos relevantes.

En definitiva, la herramienta desarrollada cumple con todos los objetivos planteados y es válida según el juicio de los expertos y los autores del proceso original que se logró automatizar. Además permite ser mejorado y ampliado en diferentes puntos que se expondrán a continuación.

5.2. Trabajos futuros

En futuros trabajos, se recomienda realizar un análisis exhaustivo de técnicas en imágenes que puedan mejorar los resultados obtenidos por OCR. Estas técnicas podrían incluir la aplicación de algoritmos de mejora de imagen, segmentación de texto y reconocimiento de estructuras visuales, con el objetivo de optimizar la precisión y la calidad de los textos extraídos de imágenes además de permitir su configuración directa desde la herramienta.

Además, sería importante investigar si los textos corregidos por la API de OpenAI introducen alguna alteración en la noticia original. Esto requeriría comparar los textos originales con los textos corregidos y evaluar la coherencia semántica y el sentido global de la noticia. Se podrían emplear técnicas de procesamiento de lenguaje natural y análisis de sentimientos para determinar si las correcciones realizadas por la API afectan el significado o la intención original del texto.

Otro aspecto relevante a considerar es el análisis de diferentes algoritmos de similitud para obtener word embeddings. Sería interesante comparar y evaluar algoritmos como Word2Vec[61], GloVe[62] y FastText[63], entre otros, en términos de precisión, velocidad y capacidad para capturar la semántica de las palabras. Esto permitiría seleccionar el algoritmo más adecuado para generar embeddings de palabras que se ajusten a las necesidades específicas del sistema.

Adicionalmente, se podría explorar más el grafo del conocimiento y la inferencia de conocimientos. Esto implica utilizar técnicas de razonamiento y lógica para inferir nuevas relaciones y conocimientos a partir de los datos existentes en el grafo. Esto podría enriquecer la comprensión y la capacidad de búsqueda del sistema, permitiendo obtener información más completa y precisa.

Además, sería beneficioso desarrollar un chatbot que utilice la información extraída del grafo de conocimiento. Este chatbot podría interactuar con los usuarios, responder preguntas relacionadas con eventos históricos, proporcionar detalles sobre personas, lugares y temas específicos, y brindar una experiencia interactiva en la exploración del conocimiento histórico de Ecuador. Esta adición permitiría a los usuarios acceder a la información de manera más conversacional y facilitaría aún más la comprensión y el descubrimiento de eventos relevantes.

Por último, se sugiere la implementación de grafos del conocimiento más específicos y especializados en dominios particulares. Esto implica la creación de grafos temáticos que abarquen áreas específicas de conocimiento, como historia, ciencia, literatura, etc. Estos grafos del conocimiento específicos permitirían realizar consultas y búsquedas más precisas y contextualizadas en cada dominio, mejorando la calidad y relevancia de los resultados obtenidos.

Referencias

[1] Ponce, J. Pincay, J. Herrera, y W. Delgado, “La usabilidad y la escala diferencial de emociones en aplicaciones para android. un estudio de caso,” *MIKARIMIN Revista Multidisciplinaria*, vol. 7, num. 1, pp. 79–86, 2021.

[2] C. Neudecker y A. Antonacopoulos, “Making europe’s historical newspapers searchable,” in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016, pp. 405–410.

[3] A. Cybulska y P. Vossen, “Historical event extraction from text,” 06 2011, pp. 39–43.

[4] M. Hallo, S. Luján-Mora, A. Maté, y J. Trujillo, “Current state of linked data in digital libraries,” *Journal of Information Science*, vol. 42, num. 2, pp. 117–127, 2016. [En línea]. Disponible: https://doi.org/10.1177/0165551515594729

[5] C. Lagoze y H. Van de Sompel, “The making of the Open Archives Initiative Protocol for Metadata Harvesting,” *Library Hi Tech*, vol. 21, num. 2, pp. 118–128, Ene. 2003, publisher: MCB UP Ltd. [En línea]. Disponible: https://doi.org/10.1108/07378830310479776

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, y Z. Ives, “Dbpedia: A nucleus for a web of open data,” in *The Semantic Web*, K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, y P. Cudré-Mauroux, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 722–735.

[7] T. BERNERS-LEE, J. HENDLER, y O. LASSILA, “The semantic web,” *Scientific American*, vol. 284, num. 5, pp. 34–43, 2001. [En línea]. Disponible: http://www.jstor.org/stable/26059207

[8] S. Ismail y T. Shaikh, “A Literature Review on Semantic Web - Understanding the Pioneers’ Perspective,” in *Computer Science & Information Technology (CS & IT)*.

Academy & Industry Research Collaboration Center (AIRCC), Sep. 2016, pp. 15–28. [En línea]. Disponible: http://airccj.org/CSCP/vol6/csit65802.pdf

[9] M. Peñalfiel y D. Seaman, “Aprovisionamiento de recursos para la ejecución de experimentos basados en machine learning a través de las guías mlops,” 2023. [En línea]. Disponible: http://dspace.ucuenca.edu.ec/handle/123456789/41969

[10] Schema.org, “Welcome to schema.org,” 2023. [En línea]. Disponible: https://schema.org/

[11] E. Bunout, M. Ehrmann, y F. Clavert, Eds., *Reflections on Tools, Methods and Epistemology*. Berlin, Boston: De Gruyter Oldenbourg, 2023. [En línea]. Disponible: https://doi.org/10.1515/9783110729214

[12] R. Franzosi, “The press as a source of socio-historical data: Issues in the methodology of data collection from newspapers,” *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 20, num. 1, pp. 5–16, 1987. [En línea]. Disponible: https://doi.org/10.1080/01615440.1987.10594173

[13] L. Albrecht y P. H. Thibodeau, “Historical newspapers as a research source: The case of chronicling america,” *Journal of the Association for Information Science and Technology*, vol. 70, num. 10, pp. 1034–1045, 2019.

[14] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, y F. Rudzicz, “A survey of word embeddings for clinical text,” *Journal of Biomedical Informatics*, vol. 100, p. 100057, 2019. [En línea]. Disponible: https://www.sciencedirect.com/science/article/pii/S2590177X19300563

[15] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, y A. Dosovitskiy, “Do vision transformers see like convolutional neural networks?” in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, y J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 12 116–12 128. [En línea]. Disponible: https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf

[16] G. Grand, I. A. Blank, F. Pereira, y E. Fedorenko, “Semantic projection recovers rich human knowledge of multiple object features from word embeddings,” *Nature Human Behaviour*, vol. 6, num. 7, pp. 975–987, Jul. 2022. [En línea]. Disponible: https://doi.org/10.1038/s41562-022-01316-8

[17] S. Kapidakis, “Consistency and Interoperability on Dublin Core Element Values in Collections Harvested using the Open Archive Initiative Protocol for Metadata Harvesting:,” in *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Budapest, Hungary: SCITEPRESS - Science and Technology Publications, 2020, pp. 181–188. [En línea]. Disponible: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010112001810188

[18] S. Mori, C. Suen, y K. Yamamoto, “Historical review of ocr research and development,” *Proceedings of the IEEE*, vol. 80, num. 7, pp. 1029–1058, 1992.

[19] T. Malathi, D. Selvamuthukumaran, C. S. D. Chandar, V. Niranjan, y A. K. Swashthika, “An experimental performance analysis on robotics process automation (rpa) with open source ocr engines: Microsoft ocr and google tesseract ocr,” *IOP Conference Series: Materials Science and Engineering*, vol. 1059, num. 1, p. 012004, feb 2021. [En línea]. Disponible: https://dx.doi.org/10.1088/1757-899X/1059/1/012004

[20] J. Smith y E. Johnson, “Challenges in ocr: Addressing limitations in low-quality sources, handwritten texts, and complex languages,” *International Journal of Document Analysis and Recognition*, vol. 25, num. 3, pp. 123–145, 2021.

[21] M. Koistinen, K. Kettunen, y J. Kervinen, “How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine – Final Notes on Development and Evaluation,” in *Human Language Technology. Challenges for Computer Science and Linguistics*, Z. Vetulani, P. Paroubek, y M. Kubis, Eds. Cham: Springer International Publishing, 2020, pp. 17–30.

[22] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, y P. Peissig, “Ocr as a service: An experimental evaluation of google docs ocr, tesseract, abbyy finereader, and transym,”

in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, y T. Isenberg, Eds. Cham: Springer International Publishing, 2016, pp. 735–746.

[23] S. Zhang, M. Diab, y L. Zettlemoyer, “Democratizing access to large-scale language models with opt-175b,” *Meta AI*, 2022.

[24] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, y D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, y H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [En línea]. Disponible: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[25] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, y G. Wang, “Gpt-ner: Named entity recognition via large language models,” 2023.

[26] C. Sun, Z. Yang, L. Wang, y et al., “Deep learning with language models improves named entity recognition for pharmaconer,” *BMC Bioinformatics*, vol. 22, num. Suppl 1, p. 602, 2021.

[27] J. Devlin, M.-W. Chang, K. Lee, y K. Toutanova, “Scaling up named entity recognition with large language models,” *arXiv preprint arXiv:2101.07285*, 2021.

[28] S. Shahriar y K. Hayawi, “Let’s have a chat! a conversation with chatgpt: Technology, applications, and limitations,” *arXiv preprint arXiv:2302.13817*, 2023. [En línea]. Disponible: https://arxiv.org/pdf/2302.13817.pdf

[29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, y I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, 2019. [En línea]. Disponible: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[30] A. Radford, K. Narasimhan, T. Salimans, y I. Sutskever, “Improving language understanding by generative pre-training,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2018. [En línea]. Disponible: https://www.aclweb.org/anthology/P18-1001.pdf

[31] H. Gupta, L. Del Corro, S. Broscheit, J. Hoffart, y E. Brenner, “Unsupervised multi-view post-OCR error correction with language models,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8647–8652. [En línea]. Disponible: https://aclanthology.org/2021.emnlp-main.680

[32] I. Zitouni, *Natural Language Processing of Semitic Languages*. Morgan & Claypool Publishers, 2014.

[33] C. Jia, X. Liang, y Y. Zhang, “Cross-domain NER using cross-domain language modeling,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2464–2474. [En línea]. Disponible: https://aclanthology.org/P19-1236

[34] Y. Hu, I. Ameer, X. Zuo, X. Peng, Y. Zhou, Z. Li, Y. Li, J. Li, X. Jiang, y H. Xu, “Zero-shot Clinical Entity Recognition using ChatGPT,” May 2023, arXiv:2303.16416 [cs]. [En línea]. Disponible: http://arxiv.org/abs/2303.16416

[35] “RapidMiner | Amplify the Impact of Your People, Expertise & Data.” [En línea]. Disponible: https://rapidminer.com/

[36] “Open for Innovation.” [En línea]. Disponible: https://www.knime.com/open-for-innovation-0

[37] B. L. Ljubljana, University of, “Data Mining.” [En línea]. Disponible: https://orangedatamining.com/

[38] T. A. S. Foundation, “Apache jena fuseki,” 2023. [En línea]. Disponible: https://jena.apache.org/documentation/fuseki2/index.html

[39] B. Gatos, S. Mantzaris, S. Perantonis, y A. Tsigris, “Automatic page analysis for the creation of a digital library from newspaper archives,” *International Journal on Digital Libraries*, vol. 3, pp. 77–84, 01 2000.

[40] G. Rundblad y H. Chen, “Advice-giving in newspaper weather commentaries,” *Journal of Pragmatics*, vol. 89, pp. 14–30, 11 2015, epub 2015 Oct 23.

[41] D. Hebert, T. Palfray, S. Nicolas, P. Tranouez, y T. Paquet, “Automatic article extraction in old newspapers digitized collections,” in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 3–8.

[42] H. Wijfjes, “Digital humanities and historical newspaper research,” *Tijdschrift Voor Mediageschiedenis*, vol. 20, pp. 4–24, 2017.

[43] C. An, D. Yin, y H. Baird, “Document segmentation using pixel-accurate ground truth,” in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 245–248.

[44] F. Le Blancq, “Rescuing old meteorological data,” *Weather*, vol. 65, num. 10, pp. 277–280, 2010.

[45] M. Hulme, “‘telling a different tale’: literary, historical and meteorological readings of a norfolk heatwave,” *Climatic Change*, vol. 113, 07 2012.

[46] E. Garnier, P. Ciavola, T. Spencer, O. Ferreira, C. Armaroli, y A. McIvor, “Historical analysis of storm events: Case studies in France, England, Portugal and Italy,” *Coastal Engineering*, vol. 134, pp. 10–23, 2018. [En línea]. Disponible: https://www.sciencedirect.com/science/article/pii/S0378383917300686

[47] J. P. Corella, G. Benito, X. Rodriguez-Lloveras, A. Brauer, y B. L. Valero-Garcés, “Annually-resolved lake record of extreme hydro-meteorological events since ad 1347 in ne iberian peninsula,” *Quaternary Science Reviews*, vol. 93, pp. 77–90, 2014.

[48] C. Boussalis, T. Coan, y M. Poberezhskaya, “Measuring and modeling russian newspaper coverage of climate change,” *Global Environmental Change*, vol. 41, 11 2016.

[49] S. Liu, H. Yang, J. Li, y S. Kolmanič, “Preliminary study on the knowledge graph construction of chinese ancient history and culture,” *Information*, vol. 11, num. 4, 2020. [En línea]. Disponible: https://www.mdpi.com/2078-2489/11/4/186

[50] K. Srinivasa y P. Santhi Thilagam, “Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers,” *Information Processing & Management*, vol. 56, num. 6, p. 102059, 2019. [En línea]. Disponible: https://doi.org/10.1016/j.ipm.2019.102059

[51] V. Saquicela, L. M. Vilches-Blázquez, y M. Espinoza, “Building a Knowledge Graph from Historical Newspapers: A Study Case in Ecuador,” in *Smart Technologies, Systems and Applications*, F. R. Narváez, F. Urgilés, T. F. Bastos-Filho, y J. P. Salgado-Guerrero, Eds. Cham: Springer Nature Switzerland, 2023, pp. 134–145.

[52] C. Vassilakis, K. Kotis, D. Spiliotopoulos, D. Margaris, V. Kasapakis, C.-N. Anagnostopoulos, G. Santipantakis, G. A. Vouros, T. Kotsilieris, V. Petukhova, A. Malchanau, I. Lykourentzou, K. M. Helin, A. Revenko, N. Gligoric, y B. Pokric, “A semantic mixed reality framework for shared cultural experiences ecosystems,” *Big Data and Cognitive Computing*, vol. 4, num. 2, 2020. [En línea]. Disponible: https://www.mdpi.com/2504-2289/4/2/6

[53] G. E. Modoni, M. Sacco, y W. Terkaj, “A survey of rdf store solutions,” in *2014 International Conference on Engineering, Technology and Innovation (ICE)*. IEEE, 2014, pp. 1–7.

[54] R. Smith, “An overview of the tesseract ocr engine,” in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007, pp. 629–633.

[55] A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundage, y I. Sutskever, “Better language models and their implications,” *OpenAI Blog https://openai. com/blog/better-language-models*, vol. 1, num. 2, 2019.

[56] S. Selva Birunda y R. Kanniga Devi, “A review on word embedding techniques for text classification,” in *Innovative Data Communication Technologies and Application*, J. S.

Raj, A. M. Iliyasu, R. Bestak, y Z. A. Baig, Eds. Singapore: Springer Singapore, 2021, pp. 267–281.

[57] G. Team, “Gradio Docs.” [En línea]. Disponible: https://gradio.app/docs

[58] J. J. Ryan, T. A. Mazzuchi, D. J. Ryan, J. Lopez De La Cruz, y R. Cooke, “Quantifying information security risks using expert judgment elicitation,” *Computers & Operations Research*, vol. 39, num. 4, pp. 774–784, Abr. 2012. [En línea]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0305054810002893

[59] M. Majszak y J. Jebeile, “Expert judgment in climate science: How it is used and how it can be justified,” *Studies in History and Philosophy of Science*, vol. 100, pp. 32–38, 2023. [En línea]. Disponible: https://www.sciencedirect.com/science/article/pii/S0039368123000857

[60] P. Vlachogianni y N. Tselios, “Perceived usability evaluation of educational technology using the system usability scale (sus): A systematic review,” *Journal of Research on Technology in Education*, vol. 54, num. 3, pp. 392–409, 2022. [En línea]. Disponible: https://doi.org/10.1080/15391523.2020.1867938

[61] T. Mikolov, K. Chen, G. Corrado, y J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[62] J. Pennington, R. Socher, y C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [En línea]. Disponible: https://aclanthology.org/D14-1162

[63] P. Bojanowski, E. Grave, A. Joulin, y M. Tomas, “Enriching word vectors with subword information,” *arXiv preprint arXiv:1607.04606*, 2017.