

Universidad de Oriente

Sede “Julio Antonio Mella”

Facultad de Ingeniería en Telecomunicaciones, Informática y Biomédica

Trabajo de Diploma

En opción al título de Ingeniera en Informática

**Título:** “Herramienta de Procesamiento de Lenguaje Natural para el reconocimiento de entidades”.

**Autor:** Luis Andrés Licea Berenguer.

**Tutor:** Dr. C. Dionis López Ramos.

Santiago de Cuba,

“Año de la Revolución”

Resumen

En la actualidad, el reconocimiento de entidades nombradas desempeña un papel fundamental en diversos contextos, desafiando la capacidad de las aplicaciones para identificar y comprender información clave en grandes volúmenes de datos no estructurados. La habilidad de discernir entidades, como personas, lugares, organizaciones y otros elementos específicos en el texto, es un componente crítico en la extracción de conocimiento y toma de decisiones.

Este estudio se enfoca en la creación de un modelo de reconocimiento de entidades mediante la implementación de la biblioteca SpaCy y el lenguaje de programación Python. El modelo no solo tiene la capacidad de identificar entidades, sino que también incorpora una función de reentrenamiento supervisado para abordar la aparición de nuevas y desconocidas entidades.

Para lograr este propósito, se aplican técnicas avanzadas de procesamiento de lenguaje natural y aprendizaje automático. Se utiliza un conjunto de datos etiquetados para entrenar el modelo, cuyo desempeño se evalúa mediante métricas de precisión, exhaustividad y F1-score. Los resultados obtenidos demuestran que el modelo es altamente preciso y exhaustivo en su capacidad para reconocer entidades.

La característica distintiva de este modelo radica en su funcionalidad de reentrenamiento supervisado, lo que permite una adaptación dinámica a nuevas entidades, garantizando un rendimiento óptimo en tiempo real. En resumen, este trabajo aporta significativamente al campo del procesamiento de lenguaje natural al proporcionar un modelo preciso y adaptable que satisface la creciente importancia del reconocimiento de entidades en el entorno contemporáneo.

**Palabras clave:** Reconocimiento de entidades, SpaCy, Python, reentrenamiento supervisado

*ABSTRACT*

In the present day, named entity recognition plays a fundamental role across various contexts, challenging applications to identify and comprehend critical information within extensive volumes of unstructured data. The ability to discern entities, such as people, places, organizations, and other specific elements within text, is a critical component in knowledge extraction and decision-making.

This study focuses on the development of a named entity recognition model through the implementation of the SpaCy library and the Python programming language. The model not only possesses the capability to identify entities but also incorporates a supervised retraining function to address the emergence of new and unfamiliar entities.

To achieve this objective, advanced techniques in natural language processing and machine learning are employed. A labeled dataset is used to train the model, and its performance is evaluated using precision, recall, and F1-score metrics. The results demonstrate that the model exhibits a high degree of accuracy and comprehensiveness in its entity recognition capabilities.

The distinctive feature of this model lies in its supervised retraining functionality, allowing dynamic adaptation to novel entities, ensuring optimal real-time performance. In summary, this work significantly contributes to the field of natural language processing by providing a precise and adaptable model that meets the growing importance of named entity recognition in the contemporary environment.

**Keywords*:***Named Entity Recognition, SpaCy, Python, Supervised Retrainingíndice

[INTRODUCCIÓN 1](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868136)

[CAPITULO 1 . MARCO TEÓRICO (u otro nombre de capítulo) 7](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868137)

[1.1 Como redactar la revisión de la literatura 7](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868138)

[1.2 Segunda división (o sub-epígrafe) dentro del primer epígrafe 8](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868139)

[1.3 Ponga aquí el segundo epígrafe 8](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868140)

[1.3.1 Primera división (o sub-epígrafe) dentro del segundo epígrafe 8](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868141)

[CAPITULO 2 . MATERIALES Y MÉTODOS (u otro nombre de capítulo) 9](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868142)

[2.1 Primer epígrafe del segundo capítulo 9](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868143)

[2.2 Primera división (o sub-epígrafe) dentro del primer epígrafe 9](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868144)

[2.3 Ponga aquí el segundo epígrafe 9](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868145)

[2.3.1 Primera división (o sub-epígrafe) dentro del segundo epígrafe 9](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868146)

[CAPITULO 3 . RESULTADOS Y DISCUSIÓN (u otro nombre de capítulo) 10](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868147)

[3.1 Primer epígrafe del tercer capítulo 10](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868148)

[3.1.1 Primera división (o sub-epígrafe) dentro del primer epígrafe 11](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868149)

[3.2 Ponga aquí el segundo epígrafe 12](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868150)

[3.2.1 Primera división (o sub-epígrafe) dentro de este epígrafe 12](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868151)

[3.3 Análisis económico 12](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868152)

[3.4 Conclusiones del capitulo 13](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868153)

[CONCLUSIONES Y RECOMENDACIONES 14](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868154)

[REFERENCIAS BIBLIOGRÁFICAS 15](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868155)

[GLOSARIO DE TÉRMINOS 17](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868156)

[ANEXOS 18](file:///D:\TesisEscrita\TesisArr.docx#_Toc404868157)

**INTRODUCCIÓN**

El término “entidad nombrada” fue usado por primera vez en la Message Understanding Conference(MUC-6), organizada por Grishman y Sundheim (1996), y se refirió a la tarea de identificar nombres de organizaciones, personas y ubicaciones geográficas en textos, así como expresiones monetarias de tiempo y porcentajes. Desde la conferencia MUC-6, el interés en NER fue incrementándose hasta hacerse presente en varios eventos científicos.[1]

Ya en la actualidad elReconocimiento de Entidades Nombradases una técnica basada en el procesamiento de lenguaje natural (NLP) que se utiliza para extraer, identificar y clasificar información en documentos de texto. Detecta entidades (es decir, partes de la oración) y las clasifica en una categoría predeterminada, como nombre o código de país.[2]

Las categorías NER pueden ser genéricas e indicar, por ejemplo, palabras que significan una organización, una persona o una época. Sin embargo, también pueden personalizarse en función de un caso de uso específico. Por poner un ejemplo, el modelo NER puede crearse para reconocer categorías como “nombre del paciente” y “fecha de nacimiento” en documentos médicos o “nombre del comerciante” y “fecha de compra” en facturas. Las posibilidades son infinitas.[2]

La primera fase del reconocimiento de entidades generalmente se reduce a un problema de segmentación: los nombres son una secuencia contigua de tokens, sin solapamiento ni anidamiento, de modo que Banco de la Nación Cuba es un nombre único, a pesar del hecho de que dentro de este nombre aparezca la subcadena Cuba que es a su vez el nombre de un país. La segunda fase se trata de asignar una categoría, de entre un conjunto predeterminado, a cada una de las entidades previamente reconocidas en la fase uno.[3]

Desde la aparición de distintas tecnologías, como la Internet, la cantidad de información disponible en distintos formatos y fuentes ha crecido a pasos agigantados. El tamaño de las colecciones almacenadas dificulta su manejo y organización, así como la posibilidad de encontrar información específica, ya sea, en un solo documento o en un conjunto de documentos. Además, la información puede no tener una estructura definida, como es el caso de la información textual, donde se encuentran principalmente secuencias de palabras. [4].

Dadas estas dificultades el desarrollo de herramientas que permitan administrar y permitir la búsqueda de ciertos elementos en un documento se ha vuelto necesario y de suma importancia. Por lo que herramientas como Elasticsearch un motor de búsqueda y análisis de datos altamente escalable, Spacy un modelo de procesamiento de lenguaje natural y Llama2 un modelo generador de oraciones más rápido que los modelos de GPT-3.5 y GPT-4, aprovechando al máximo el uso de estas se construiría una poderosa y robusta herramienta de procesamiento de lenguaje natural para la detección de entidades de Elasticsearch, mejorando aún más su búsqueda y siendo capaz de reentrenarse de manera supervisada cuando surja una nueva entidad.

**Problema de investigación:**

La empresa de desarrollo de software CERPAMID-DATYS recibe un flujo de información de diferentes empresas del país, estos datos son indexados en Elasticsearch por las facilidades de búsqueda que el mismo posee. El centro le surge la problemática de que, al ser un volumen de información demasiado grande de datos, es humanamente imposible analizar esta información. Por lo que la empresa propone que una vez indexado los datos se haga un análisis de estos que permita conocer de quien se habla, lugares, fechas u organizaciones. Al no contar con el personal especializado en el procesamiento del lenguaje natural, y sería muy costoso mantener un personal especializado que procesara manual esta información, se propuso hacer una extracción de entidades una vez indexadas. Pero el gestor de la base de datos de Elasticsearch, Kibana nos brinda la posibilidad de alojar modelos de inteligencia artificial pero solo por un mes de prueba. Por lo tanto, se propuso la creación de un software que permita realizar el reconocimiento de las entidades de la base de datos y que devuelva dicha información para su posterior estudio, así como la capacidad de reentrenarse al no reconocer o surgir una nueva entidad.

**Objeto de Estudio:** Reconocimiento de Entidades Nombradas

**Campo de Estudio:** Herramientas de Procesamiento del Lenguaje Natural.

**Objetivo General:**

Desarrollar un programa que reciba los datos de Elasticsearch y conjunto con el modelo de procesamiento de lenguaje natural SPACY haga el reconocimiento de las entidades, en caso de que el modelo no reconozca una entidad, reentrenar el modelo de manera supervisada, para el reconocimiento de esta nueva entidad.

**Objetivo Específicos:**

1. **Extracción de Datos de Elasticsearch:** El programa se encargará de establecer la conexión con Elasticsearch y extraer los datos almacenados en la base de datos, asegurando la integridad y calidad de la información recuperada.
2. **Reconocimiento de Entidades (NER):** Se desarrollará una funcionalidad de NER utilizando el modelo de Spacy para la identificación y clasificación automática de entidades presentes en los datos extraídos. Esto incluye la detección de nombres de personas, organizaciones, ubicaciones, fechas, entre otros.
3. **Aprendizaje Supervisado:** En caso de que el modelo Spacy no reconozca una entidad específica en los datos, se implementará un proceso de aprendizaje supervisado. El programa permitirá a los usuarios etiquetar manualmente las nuevas entidades, reentrenando así el modelo para reconocer estas entidades en el futuro.
4. **Almacenamiento de Resultados en Elasticsearch:** Una vez identificadas las entidades, el programa guardará los resultados de manera estructurada en Elasticsearch, permitiendo un acceso eficiente y consultas posteriores.
5. **Evaluación de Rendimiento:** Se realizará una evaluación del rendimiento del programa, incluyendo la precisión y la eficiencia del reconocimiento de entidades, así como la efectividad del proceso de aprendizaje supervisado.
6. **Optimización Continua:** El objeto de investigación incluye la posibilidad de optimizar y mejorar el programa a medida que se recopilen más datos y se identifiquen nuevas entidades.

**Campo de Acción:**

Aplicación de modelos de procesamiento del lenguaje natural, como Spacy, para detectar entidades, proponiendo el reentrenamiento del modelo cuando no se reconocen correctamente las entidades. Además, se integra una base de datos en Elasticsearch para almacenar y analizar los resultados del reconocimiento de entidades.

**Hipótesis:**

La implementación conjunta de Spacy, Elasticsearch y Python en un sistema de procesamiento de lenguaje natural permitirá la identificación y clasificación eficiente de entidades en grandes volúmenes de datos, facilitando el estudio lingüístico. Esta hipótesis plantea que la combinación de estas tecnologías tendrá un impacto positivo en la capacidad de la empresa para gestionar y aprovechar datos masivos con fines de estudio y análisis lingüístico, y se buscará probar esta hipótesis a través de la implementación y evaluación de la solución propuesta.

**Métodos de investigación:**

1. Método histórico-lógico: se aplicó al realizar el análisis de la existencia de otros sistemas que den solución al problema en cuestión.
2. Método de análisis y síntesis: se aplicó al realizar el análisis de todo el proceso llevado a cabo en el proyecto y sintetizar las ideas que fueron surgiendo; extrayendo los elementos comunes al objeto de estudio.

**Interrogantes Científicas:**

1. ¿Cómo se puede mejorar la precisión y exhaustividad del modelo de reconocimiento de entidades utilizando técnicas de procesamiento de lenguaje natural y aprendizaje automático?
2. ¿Cómo se puede implementar una funcionalidad de reentrenamiento supervisado para adaptar el modelo a nuevas entidades en tiempo real?
3. ¿Cuál es el impacto de utilizar la biblioteca Spacy y el lenguaje de programación Python en el desarrollo y desempeño del modelo de reconocimiento de entidades?

**Aportes de la Investigación:**

* Un sistema que reconozca las entidades de los textos de la base de datos de elasticsearch.
* Capacidad de reentrenamiento de forma supervisada por parte del usuario o la entidad en caso de surgir o no detectar una nueva entidad.
* Generacion de datos de entrenamiento y testing de la entidad que no reconozca el modelo.
* Capacidad de retroceder en caso de el modelo perder conocimientoa al realizar el reentrenamiento.

1. **MARCO TEÓRICO (u otro nombre de capítulo)**

En este capítulo se explican los principales aspectos teóricos, los conceptos básicos de las tecnologías y la caracterización de las herramientas computacionales utilizadas.

## Breve historia de los modelos de procesamiento del lenguaje natural para la detección de entidades

El término “entidad nombrada” fue usado por primera vez en la Message Understanding Conference (MUC-6), organizada por Grishman y Sundheim (1996), y se refirió a la tarea de identificar nombres de organizaciones, personas y ubicaciones geográficas en textos, así como expresiones monetarias de tiempo y porcentajes. Desde la conferencia MUC-6, el interés en NER fue incrementándose hasta hacerse presente en varios eventos científicos.

Con respecto a la definición de entidades nombradas, G. Petasis la expresa como: “Una entidad nombrada es un nombre propio, sirviendo como un nombre para algo o alguien”, dicha definición se justifica debido al porcentaje tan significativo de nombres propios que encontramos en un corpus. Nadeau y Sekine afirman que la palabra “nombrada” restringe la tarea a sólo aquellas entidades para las cuales uno o muchos designadores rígidos, representan el referente. A pesar de las múltiples definiciones de las entidades nombradas, investigadores han logrado un acuerdo común de los tipos de entidades nombradas a reconocer.[5]

**1990**:

Los primeros enfoques para el NER se basaban en reglas y patrones en el texto. Estos enfoques utilizaban conjuntos de reglas predefinidas para identificar y extraer entidades nombradas [6]

**2000**:

A medida que los sistemas de procesamiento de lenguaje natural (NLP) se volvieron más sofisticados, se desarrollaron enfoques basados en diccionarios. Estos enfoques utilizaban diccionarios de palabras y frases para identificar y extraer entidades nombradas [6]

**2010**:

Con la aparición del machine learning, los enfoques basados en diccionarios dieron paso a enfoques basados en machine learning. Estos enfoques utilizaban algoritmos de machine learning para aprender de los datos y extraer entidades nombradas [6]

**2018**:

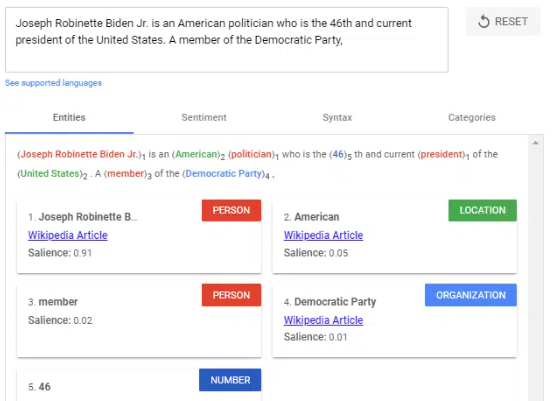
En los últimos años, los enfoques basados en deep learning han ganado popularidad. Estos enfoques utilizan redes neuronales profundas para aprender representaciones jerárquicas de las palabras y extraer entidades nombradas [6]

**2020**:

Los modelos de lenguaje preentrenados como BERT han revolucionado el NER al permitir que los modelos aprendan representaciones jerárquicas de las palabras a partir de grandes cantidades de texto. Estos modelos pueden ser ajustados para tareas específicas de NER, lo que los hace muy eficaces[6]

## Estado del Arte de los marcos de procesamiento del lenguaje natural para la detección de entidades

**Google NLP**



Google Cloud NLP es una API de Inteligencia Artificial (IA) que se centra en el procesamiento del lenguaje natural (NLP). Esta API es parte de la suite de productos digitales de Google, que están en constante crecimiento y amplían las capacidades de la tecnología para las empresas[7]

La API de Google Cloud NLP se enfoca en la creación de contenido que se asemeje lo más posible a lo que un humano podría crear, proporcionando un análisis de texto a un nivel superior que fluye de manera más clara que lo que muchos programas de IA actuales son capaces de proporcionar. Esta API es la opción más compleja de Google, lo que significa que se dirige a empresas más grandes que su producto básico AutoML[7]

Entre las ventajas de Google Cloud NLP, destaca su reconocimiento de entidades nombradas multilingüe y su eficiencia en comparación con la contratación de humanos con habilidades de idioma para manejar proyectos similares. Esta API ofrece servicios en 10 idiomas diferentes y proporciona características como la resumación de texto, la generación de texto, la traducción, la detección de idioma, la tokenización, la lematización y el análisis de sentimientos

Sin embargo, Google Cloud NLP tiene algunas limitaciones. Aunque es una opción fuerte en las áreas en las que proporciona, la API actualmente es más limitada en los tipos de información que es capaz de manejar. Muchos usuarios han encontrado que esta API necesita una formación significativamente mayor para ser capaz de realizar menos tareas que muchas de las otras opciones en el mercado. Además, Google Cloud NLP es una opción más costosa que muchas de sus competidoras, lo que significa que puede ser más difícil de experimentar que otras opciones. Aunque esta API proporciona créditos gratuitos a nuevos clientes, el costo más alto de este programa puede no valer la pena una vez que su negocio utiliza estos créditos y necesita pasar a una suscripción continua.

**ChatGPT**

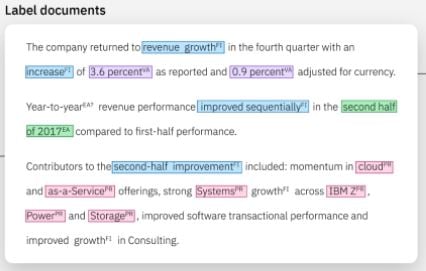


ChatGPT es, probablemente, el referente inmediato para millones de personas sobre lo que es un sistema de procesamiento natural de lenguaje. Esta tecnología está equipada con un potente motor de análisis de lenguaje conocido como GPT-4.

Esta herramienta ganó popularidad a principios del 2023 ya que es una de las pioneras en poner al servicio de las personas una herramienta capaz de responder a demandas concretas con un alto nivel de éxito. Además, esta plataforma lanzó su propia API (interfaz de programación de aplicaciones, por sus siglas en inglés) que puede ser integrada en programas para la creación de chatbots mucho más potentes, útiles y precisos con las respuestas que ofrecen a sus usuarios.

Hoy en día, esta herramienta está detrás de miles de programas y aplicaciones, como son tiendas online, buscadores, aplicaciones de aprendizaje y hasta plataformas de redes sociales. Además, es un gran apoyo si quieres investigar sobre un tema altamente especializado, si deseas reescribir textos o si necesitas ideas creativas para redactar tus propios contenidos. [8]

**IBM Watson Discovery**

****

Watson es la plataforma de inteligencia artificial creada por IBM que tiene como objetivo acercar el uso de datos a los negocios de hoy. De entre sus herramientas, Discovery representa el mayor avance en aplicación del procesamiento de lenguaje natural.

Esta herramienta es, esencialmente, un procesador de documentos que evalúa, interpreta y extrae información de archivos de texto con el fin de hacer más simple la consulta y lectura de materiales escritos. Mediante la creación de etiquetas, el sistema aprende a detectar información de relevancia, así como a distinguir entre tablas, títulos y anotaciones.

Gracias a su potente motor de IA, esta tecnología es incluso capaz de traducir imágenes en texto mediante el reconocimiento de caracteres, así como crear mapas o constelaciones de conceptos para que no dejes fuera ninguna conexión entre elementos de tus documentaciones. [8]

**Amazon Comprehend**

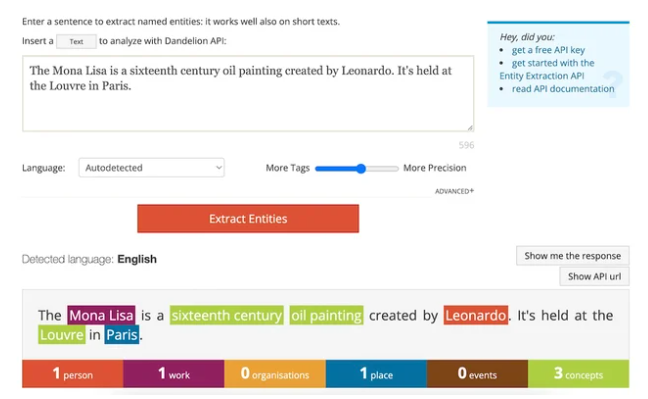
****

Otra excelente herramienta para el procesamiento de texto en documentos es Amazon Comprehend, que forma parte del universo de Amazon Web Services. Este sistema es capaz de extraer información tanto de documentos como de tickets de atención al cliente, reseñas, correos electrónicos y redes sociales.

Por estas funciones es que Comprehend resulta tan útil para tiendas en línea o empresas con gran actividad en medios de comunicación masiva. Con algunos simples pasos, puedes saber qué opinión tienen los clientes de ti, qué piensan de tus productos y hasta el tono de tus interacciones con ellos a través de tus canales de atención.

También, hay empresas que usan este sistema para la administración de enfoques legales y procesamiento de documentos financieros, ya que pueden detectar detalles que hasta tu equipo legal y de ventas puede pasar por alto en los términos y condiciones de alguna operación.[8]

**Dandelion**



Dandelion es una API que se centra en el análisis semántico y de texto para ayudar a los usuarios a entender si los documentos que se están visualizando son principalmente positivos, negativos o neutrales. Esto permite a los usuarios aprender más sobre las opiniones generales sobre ciertos temas, lo que puede ayudarles a determinar qué ángulo sería el mejor para trabajar con la información que aprenden.[7]

Esta API se dirige principalmente a las empresas que están interesadas en investigar lo que su público objetivo está diciendo sobre varios temas que están relacionados con su industria, así como cómo sus competidores están manejándolos. Aunque no se adentra tanto en lo que los usuarios pueden hacer con esta información, Dandelion proporciona una visión general sólida de qué información relevante está actualmente disponible[7]

## Herramientas, Lenguaje de programación y Tecnologías

Las herramientas son objetos elaborados a fin de facilitar la realización de una tarea. Se diseñan y fabrican para cumplir uno o más propósitos específicos, por lo que son generalmente artefactos con una función técnica. Un lenguaje de programación es un lenguaje formal que especifica una serie de instrucciones para que una computadora produzca diversas clases de datos. Los lenguajes de programación pueden usarse para crear programas que pongan en práctica algoritmos específicos los cuales controlan el comportamiento físico y lógico de una computadora. La tecnología es la ciencia aplicada a la resolución de problemas concretos. Constituye un conjunto de conocimientos científicamente ordenados, que permiten diseñar y crear bienes o servicios que facilitan la adaptación al medio ambiente y la satisfacción de las necesidades humana.

**Python**

Python en su versión 3.10.8 es un lenguaje de programación ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el machine learning (ML). Los desarrolladores utilizan Python porque es eficiente y fácil de aprender, además de que se puede ejecutar en muchas plataformas diferentes. El software Python se puede descargar gratis, se integra bien a todos los tipos de sistemas y aumenta la velocidad del desarrollo [14].

Se utilizó Python porque ofrece los siguientes beneficios:

* Los desarrolladores pueden leer y comprender fácilmente los programas de Python debido a su sintaxis básica similar a la del inglés.
* Python permite que los desarrolladores sean más productivos, ya que pueden escribir un programa de Python con menos líneas de código en comparación con muchos otros lenguajes.
* Python cuenta con una gran biblioteca estándar que contiene códigos reutilizables para casi cualquier tarea. De esta manera, los desarrolladores no tienen que escribir el código desde cero.
* Los desarrolladores pueden utilizar Python fácilmente con otros lenguajes de programación conocidos, como Java, C y C++.
* La comunidad activa de Python incluye millones de desarrolladores alrededor del mundo que prestan su apoyo. Si se presenta un problema, puede obtener soporte rápido de la comunidad.
* Hay muchos recursos útiles disponibles en Internet si desea aprender Python. Por ejemplo, puede encontrar con facilidad videos, tutoriales, documentación y guías para desarrolladores.
* Python se puede trasladar a través de diferentes sistemas operativos de computadora, como Windows, macOS, Linux y Unix.

**Visual Studio Code**

Visual Studio Code (VS Code) es un editor de código fuente ligero pero potente desarrollado por Microsoft. Está disponible para Windows, macOS y Linux y es gratuito para descargar y utilizar. VS Code combina la simplicidad de un editor de código fuente con potentes herramientas para desarrolladores, como la finalización de código IntelliSense y la depuración [15]

Algunas de las características clave de VS Code:

1. Soporte integrado para varios lenguajes de programación: VS Code viene con soporte integrado para JavaScript, TypeScript y Node.JS, y tiene un rico ecosistema de extensiones para otros lenguajes como C++, C#, Java, Python, PHP, Go, y para tiempo de ejecución como .NET y Unity [15].
2. Interfaz de línea de comandos potente: VS Code tiene una potente interfaz de línea de comandos que te permite controlar cómo lanzas el editor. Puedes abrir diferentes archivos, instalar extensiones, e incluso cambiar el idioma de visualización al inicio [15].
3. Integración con Git: VS Code viene con una integración con Git que te permite hacer commit, pull y push de tus cambios de código a un repositorio Git remoto [15].
4. Extensibilidad: Puedes elegir entre miles de extensiones para personalizar tu IDE. Esto te permite agregar funcionalidades adicionales a VS Code para adaptarlo a tus necesidades específicas [15].

**Elasticsearch**

Esta base de datos es un motor de analítica y búsqueda de RESTful distribuido que almacena de forma centralizada tus datos para que puedas buscar, indexar y analizar datos de todas formas y tamaños[16].

Algunas características de Elasticsearch son:

* Desarrollado en Java
* Open Source
* Distribuido
* Escalable
* Basado en Lucene

Lucene es una librería de búsqueda de texto[17]:

* Desarrollado en java
* Open Source
* Escalable
* Con alto rendimiento
* Basada en índices invertidos



Fig 1.0

Como se muestra en la figura 1.0 un índice directo en una base de datos es que a cada texto o indexación a la base de datos se guarda con un id, mientras que en índices invertidos se le asigna un id a cada palabra y lo que se guarda en la base de datos es el conjunto de ubicaciones donde se encuentra esta palabra, haciendo así una búsqueda más rápida.

**Git**

Git es un sistema de control de versiones distribuido que se utiliza para rastrear cambios en archivos y coordinar el trabajo entre programadores. Es el sistema de control de versiones más popular en el mundo y se utiliza para una amplia gama de proyectos, desde software hasta documentación [18].

Git permite a los usuarios realizar un seguimiento de los cambios en los archivos de un proyecto, lo que facilita la colaboración y la resolución de conflictos. Además, Git permite a los usuarios revertir a una versión anterior de un proyecto si se produce un error o un bug. Esto es especialmente útil en un entorno de desarrollo de software, donde los errores son comunes y es necesario poder volver atrás rápidamente [18].

Git también es compatible con todos los sistemas operativos disponibles y puede acceder directamente a otros repositorios de control de versiones remotos. Esto significa que los usuarios pueden cambiar fácilmente a Git sin tener que mover sus archivos de esos repositorios al repositorio de Git. Además, Git es altamente escalable y más rápido que otros sistemas de control de versiones, lo que permite manejar eficientemente proyectos grande [18].

**Spacy**

SpaCy es una biblioteca de procesamiento de lenguaje natural (NLP) para Python que proporciona una serie de herramientas para el procesamiento de texto, incluyendo tokenización, etiquetado de partes del discurso, reconocimiento de entidades nombradas y más. spaCy es conocido por su velocidad y eficiencia, y es especialmente útil para tareas de NLP que requieren un análisis rápido y preciso del texto [12].

En este trabajo se utilizó la biblioteca Spacy para realizar el reconocimiento de entidades de la base de datos de Elasticsearch.

**Llama2**

Llama2 es una versión de un modelo de lenguaje grande desarrollado por Facebook. Este modelo ha sido entrenado para generar texto en un estilo similar al de un humano, y es especialmente útil para tareas de generación de texto y procesamiento del lenguaje natural. Llama2 puede generar respuestas a las preguntas, continuar las historias y realizar una variedad de tareas de generación de texto[19].

En este trabajo se utilizó para la generación de oraciones para la posterior conversión de datos de entrenamiento del modelo Spacy en caso de no reconocer una entidad.

**Como redactar la revisión de la literatura**

El error más frecuente que se comete en este numeral es colocar demasiadas referencias que nada o poco tienen que ver con los objetivos del trabajo. La norma para enfrentar esto es tener en cuenta solamente aquellas referencias que le ayuden a reforzar la justificación, a adoptar procedimientos y equipos, ó a explicar y apoyar los resultados. *Se requiere seleccionar sólo las más importantes y recientes, y que además estén directamente vinculadas con nuestro planteamiento del problema de investigación*.

Recomendaciones para evitar las citas excesivas:

No respaldes una aseveración con más de tres citas.

Si puedes respaldar tus aseveraciones con varias citas, usa las más recientes y las más importantes.

No cites tesis de grado, resúmenes, ni  informes de proyectos si la información se ha publicado en una revista científica.

No cites información publicada en revistas locales o de escasa distribución si la información se ha publicado en una revista internacional.

No respaldes aseveraciones que toda la audiencia conoce (*vox populi*).

Si consultó 100 libros; pero solo cinco hacen referencia al tema, solamente tenga en cuenta estos cinco y no piense que perdió el tiempo, pues alguna cosa aprendió de todos ellos.

**Segunda división (o sub-epígrafe) dentro del primer epígrafe**

Inicie a escribir aquí

**Ponga aquí el segundo epígrafe**

Inicie a escribir aquí.

**Primera división (o sub-epígrafe) dentro del segundo epígrafe**

Inicie a escribir aquí.

Ponga aquí el tercer epígrafe

Inicie a escribir aquí.

Puede añadir tantos acápites como le sean necesarios pero *no es una buena práctica utilizar más de tres niveles de profundidad ( #.#.#) en la numeración de los epígrafes.* Puede añadir alguna idea al final de los capítulos que motive la lectura del siguiente, dando así fluidez a la lectura.

1. **RESULTADOS Y DISCUSIÓN (u otro nombre de capítulo)**

El error más común en la redacción de este epígrafe es no prever cómo se van a anotar los datos a medir lo cual está estrechamente relacionado con el método de análisis, ó sea con la estadística necesaria para establecer la confiabilidad y exactitud de los resultados En este capítulo debe hacerse el análisis de los resultados alcanzados, realizándose comparaciones con los objetivos planteados, con otros trabajos similares realizados en el centro o fuera de este, con estándares internacionales, etc.

Es importante incluir además el análisis económico, si procede

**Primer epígrafe del tercer capítulo**

A continuación se muestra ejemplo de una tabla. Las principales normas a seguir son: el título se colocará en la parte superior de la tabla a continuación del término Tabla, solamente seguido del número del capítulo, punto y el número de orden de la tabla en ese capítulo. No se debe emplear la abreviatura No. ni el signo #; el espaciado del título de la tabla debe ser sencillo; cada columna de la tabla llevará su título, procurando no hacer abreviaturas; los títulos de las columnas se encerrarán entre dos líneas horizontales sencillas. La tabla se cerrará también con una línea horizontal sencilla. En caso que la tabla no sea de su autoría, debe ser agregada la Fuente de referencia de donde fue tomada la misma.

*Tabla 3.1. Escribir aquí el título de la tabla (Fuente: sitio de la UIT, 2014)*

|  |  |  |  |
| --- | --- | --- | --- |
| Título de la columna uno\* | Título de la columna dos | Título de la columna tres | Título de la columna cuatro |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

\* Las llamadas para explicar algo en la tabla se deberán hacer con asteriscos u otro símbolo, y las notas se colocarán al pie de la tabla y no al pie de la página.

Las ecuaciones deben ser enumeradas siguiendo el mismo esquema de las tablas, y la etiqueta numérica debe ser colocada a la derecha tratando de uniformizar esa ubicación para todas las ecuaciones que se utilicen, Por ejemplo en (3.1) se calcula el área de un círculo.

|  |  |
| --- | --- |
|  | **(3.1)** |

Ahora veamos un ejemplo de figura, en este caso una gráfica experimental, por ser de las más utilizadas. Las principales características a respetar son: el número y título deberán aparecer en la parte inferior de la figura, a continuación del término Figura; colocar nombre, símbolo y/o unidades para las variables sobre los ejes correspondientes y se usarán líneas y/o símbolos diferentes para distinguir las diferentes curvas que se puedan dibujar en una misma figura. No se admiten las palabras "foto" ni "gráfica", debe utilizarse la palabra Figura o Fig. seguida por un espacio, el número del capítulo, punto y el número de orden dentro del capítulo al cual pertenece, Debe utilizarse Arial 10 en itálica. En caso que la figura no sea de su autoría, debe colocarse la Fuente de referencia de los datos en el pie de la misma.

2000

2002

2004

80

100

120

140

160

180

200

220

240

Eje x

Eje Y

Curva 1

Curva 2

*Figura 3.1. Escribir aquí el título de la figura (Fuente: [3]).*

Continúe con el texto.

**Primera división (o sub-epígrafe) dentro del primer epígrafe**

Inicie a escribir aquí. Si alguno de los subtítulos hubiera quedado en la última línea de una página, sería necesario trasladarlo a la primera línea de la página siguiente

**Ponga aquí el segundo epígrafe**

Inicie a escribir aquí.

**Primera división (o sub-epígrafe) dentro de este epígrafe**

1. **CONCLUSIONES Y RECOMENDACIONES**

**Conclusiones**

Las conclusiones deben estar en correspondencia con los objetivos planteados. Ellos deben referirse al problema de investigación planteado, de hecho la primera conclusión de un estudio es evaluar qué ocurrió con el planteamiento. Deben ser objetivas, breves, precisas y convincentes. Es importante destacar que las conclusiones no son un recuento de lo que se realizó en el trabajo. La conclusión constituye una generalización o comentario aplicado,.relativa a los resultados obtenidos

Para una mejor comprensión pueden ser numeradas y si algún caso lo requiere, pueden ser comentadas. Ejemplos

Se comprobó que el efecto del jitter sobre la BER no es significativo cuando la velocidad de transmisión no excede los 4 Mbps

Se verificó que …

Se demostró que …

Conclusión 3

**Recomendaciones**

Escriba lo que usted recomienda a otros investigadores para mejorar su trabajo o para determinar otros datos que usted no pudo lograr, sugerir qué hacer con sus resultados y aportes.

Recomendación 1

Recomendación 2

Recomendación 3

No debería haber muchas recomendaciones pues si uno recomienda muchas acciones la lectura sería que no se avanzó mucho en la investigación.

La extensión de estos aspectos no debe exceder de cinco páginas.

1. **REFERENCIAS BIBLIOGRÁFICAS**

Este capítulo no se enumera. Incluyen los materiales que fueron utilizados durante el desarrollo del trabajo.

Deberá usarse el estilo de citación de la IEEE (orden numérico. ubicando Al citar usando este estilo, las referencias deben ser ordenadas en el orden que fueron citadas.

Recuerde que utilizando un gestor bibliográfico como el que viene embebido en Microsoft Word o en *Endnote*, el ordenamiento de las referencias se hará automáticamente. Debe utilizar las normas Harvard.

Las referencias se señalan cuando se hace alusión directa o indirecta a un autor o a su obra, sobre todo si se le cita textualmente. Recuerde que la utilización de palabras o conceptos de otro autor sin señalar la fuente constituye un delito (PLAGIO) sancionado. Denota además cierta falta de ética.

|  |  |
| --- | --- |
| [1] | R. H. Sampieri, C. F. Collado y P. B. Lucio, Metodología de la Investigación, Istapalapa: Mc Graw Hill, 2006.  Autor(es) (Año). Título del Artículo, Título de la Revista, Número del Volumen, (Número de Edición): Páginas.  Autor(es) (Año). Título del libro en itálica. Número de la edición, Editorial, Lugar de Publicación.  Autor(es) (Año). Título del Artículo o del Capítulo en el Libro, En: Nombre del Libro, Número de la Edición, Número del Capítulo, Editores, Editorial, Lugar de la Publicación. |

Se sugiere utilizar los gestores bibliográficos[[1]](#footnote-1) para la gestión de las referencias bibliográficas, de esta forma ellos harán todo el proceso de organización y formateo por el autor.. El uso del gestor posibilita la exportación de un archivo (base de datos) con una lista detallada de las referencias utilizadas, la cual debe ser entregada al/los tutor/es. Esta base de datos es de utilidad para la continuidad de las investigaciones en la institución académica )

**BIBLIOGRAFIA (**opcional**)**

[1] R. Viana, «Reconocimiento de entidades nombradas: Una breve introducción», SoldAI. Accedido: 8 de noviembre de 2023. [En línea]. Disponible en: https://medium.com/soldai/reconocimiento-de-entidades-nombradas-una-breve-introducci%C3%B3n-ad6e2a3f911a

[2] V. Osorio, «¿Qué es el Reconocimiento de Entidades Nombradas?», Klippa. Accedido: 25 de octubre de 2023. [En línea]. Disponible en: https://www.klippa.com/es/blog/informativo/reconocimento-entidades-nombradas/

[3] «Haag, K. Y. Reconocimiento de entidades nombradas en texto de dominio legal.pdf».

[4] «SanchezPCR.pdf».

[5] R. Viana, «Reconocimiento de entidades nombradas: Una breve introducción», SoldAI. Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://medium.com/soldai/reconocimiento-de-entidades-nombradas-una-breve-introducci%C3%B3n-ad6e2a3f911a

[6] «Qué es el reconocimiento de entidad nombrada (NER): definición, ejemplos, tipos y aplicaciones». Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://es.shaip.com/blog/named-entity-recognition-and-its-types/

[7] «Top 5 named entity recognition APIs to use in your app». Accedido: 17 de diciembre de 2023. [En línea]. Disponible en: https://datavid.com/blog/named-entity-recognition-apis

[8] D. Santos, «Procesamiento de lenguaje natural: qué es, ejemplos y herramientas». Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://blog.hubspot.es/marketing/procesamiento-de-lenguaje-natural

[9] «Cómo funciona BERT, el modelo de procesamiento del lenguaje natural de Google». Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://wwwhatsnew.com/2023/03/21/como-funciona-bert-el-modelo-de-procesamiento-del-lenguaje-natural-de-google/

[10] «Named Entity Recognition», CoreNLP. Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://stanfordnlp.github.io/CoreNLP/ner.html

[11] A. Chauhan, «Training Custom NER Model Using Flair», TheCyPhy. Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://medium.com/thecyphy/training-custom-ner-model-using-flair-df1f9ea9c762

[12] «Probamos SpaCy: mucho más que una librería para crear proyectos reales de procesamiento del lenguaje natural | datos.gob.es». Accedido: 7 de octubre de 2023. [En línea]. Disponible en: https://datos.gob.es/es/blog/probamos-spacy-mucho-mas-que-una-libreria-para-crear-proyectos-reales-de-procesamiento-del

[13] «Alternativas de SpaCy para la extracción de entidades (NER)». Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://nlpcloud.com/es/spacy-alternatives-for-entity-extraction-ner.html

[14] «¿Qué es Python? - Explicación del lenguaje Python - AWS», Amazon Web Services, Inc. Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://aws.amazon.com/es/what-is/python/

[15] «Why Visual Studio Code?» Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://code.visualstudio.com/docs/editor/whyvscode

[16] «Lista de características de Elasticsearch | Elastic». Accedido: 7 de octubre de 2023. [En línea]. Disponible en: https://www.elastic.co/es/elasticsearch/features

[17] «Qué es Lucene. Qué es Lucene y qué relación tiene con Solr.» Accedido: 7 de octubre de 2023. [En línea]. Disponible en: https://solrtutorial.es/que-es-lucene.html

[18] B. Marijan, «What Is Git and What Is It Used For? {+ Main Features}», Knowledge Base by phoenixNAP. Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://phoenixnap.com/kb/what-is-git

[19] K. Luzniak, «6 main differences between Llama 2, GPT-3.5 & GPT-4», Neoteric. Accedido: 25 de octubre de 2023. [En línea]. Disponible en: https://neoteric.eu/blog/6-main-differences-between-llama2-gpt35-and-gpt4/

Es opcional, puede incluir aquí los libros que fueron leídos, pero que no fueron referidos en el trabajo.

1. **GLOSARIO DE TÉRMINOS**

Se utiliza cuando durante el desarrollo se han utilizado palabras, conceptos, siglas, etc., que no expresan una idea clara, precisa y concisa, o que puedan tener múltiples interpretaciones

Cuando aparece una sigla por primera vez, ella se escribe normalmente y entre paréntesis se pone el significado de sus términos en el idioma original y a su vez se pone en el Glosario de términos, cuando aparece nuevamente no es necesario poner entre paréntesis el significado de la sigla.

El glosario de términos se anexa después de la bibliografía y antes de los anexos y se escribe en orden alfabético, Se recomienda no abusar de los anglicismos y si se utilizan deben escribirse en itálicas

1. **ANEXOS**

Anexos: se colocan al final del trabajo, no se numeran como capítulos, pero sí se les coloca un número de orden luego de la palabra ANEXO; en ellos se incluye información que no es necesaria para la presentación del trabajo, pero sí útil al lector. Exigen que se hagan referencia de ellos en el contenido o cuerpo de la obra y deben aparecer en el mismo orden en que han sido citados y deben tener un pie de grabado o título que enuncie lo que ilustra el mismo.

Los anexos deben recoger aquellos aspectos del trabajo que, por su longitud, complejidad o que no era adecuado ponerlos, no se incluyen en el texto de la tesis pero que coadyuvan a una mejor comprensión de lo que se expone en ella. Sin embargo, es necesario dejar bien claro que los anexos no pueden usarse para extender la tesis más allá del límite establecido.

Es opcional su uso

**Anexo I** Inserte título del primer anexo

Inserte aquí el primer anexo.

**Anexo II** Inserte título del segundo anexo

Puede añadir tantos anexos como le sean necesarios.

1. [↑](#footnote-ref-1)