UNIVERSIDAD TÉCNICA DE BABAHOYO

FACULTAD DE ADMINISTRACIÓN, FINANZAS E INFORMÁTICA

PROCESO DE TITULACIÓN

JUNIO - OCTUBRE 2023

EXAMEN COMPLEXIVO DE GRADO O DE FIN DE CARRERA

ESTUDIO DE CASO PREVIO A LA OBTENCIÓN DEL TITULO DE INGENIERO EN SISTEMAS

TEMA:

ESTUDIO COMPARATIVO DE HERRAMIENTAS DE PROCESAMIENTO DE LENGUAJE NATURAL OPEN SOURCE Y SOFTWARE PROPIETARIO PARA EL DESARROLLO DE APLICACIONES BASADAS EN INTELIGENCIA ARTIFICIAL

EGRESADO:

KLEVER EDUARDO ANGAMARCA PEÑA

TUTOR:

ING. CARLOS ALFREDO CEVALLOS MONAR

AÑO 2023

RESUMEN

El objetivo de este estudio comparativo es identificar las principales herramientas de Procesamiento de Lenguaje Natural Open Source y Software Propietario y analizar una herramienta de NLP Open Source y una de Software Propietario en términos de su funcionalidad, eficiencia, precio y facilidad de implementación para el desarrollo de aplicaciones basadas en Inteligencia Artificial (AI).

Empezamos con la recolección de información respecto al Procesamiento de Lenguaje Natural y sus principales aplicaciones. En cuanto a las herramientas Open Source elegimos Spacy, NLTK, Gensim y Spark NLP destacando sus principales características, de la misma manera realizamos este proceso con Google Natural Language API, IBM Watson Natural Language Understanding, Amazon Comprehend y Azure AI API las cuales son herramientas de origen propietario, comparamos las ventajas y desventajas del uso de estos dos tipos de herramientas facilitando así una toma de decisiones informada.

Para el análisis práctico seleccionamos dos herramientas, Google Natural Language API junto con la plataforma de Google Cloud y la plataforma de Google Colab la utilizamos para ejecutar Spacy.

Se concluye que las herramientas Open Source y de Software Propietario ofrecen excelentes resultados en cuanto al análisis de texto y la selección de una herramienta u otra debe basarse en las necesidades y características específicas de cada proyecto además del presupuesto que se disponga como empresa o desarrollador.

Palabras clave: NLP, Spacy, Open Source, Google API, Google Cloud, AI.

ABSTRACT

The goal of this comparative study is to identify the main Open-Source and Proprietary Natural Language Processing (NLP) tools and to analyze one Open-Source NLP tool and one Proprietary tool in terms of their functionality, efficiency, price, and ease of implementation for the development of artificial intelligence-based applications.

To this end, a research was carried out on Natural Language Processing and its main applications. Regarding Open-Source tools, Spacy, NLTK, Gensim, and Spark NLP were selected highlighting their main features. In the same way, this process was carried out with Google Natural Language API, IBM Watson Natural Language Understanding, Amazon Comprehend, and Azure AI API, which are proprietary tools. The advantages and disadvantages of using these two types of tools were compared, thus facilitating informed decision-making.

In terms of practical analysis two tools were selected: Google Natural Language API with the Google Cloud platform and the Google Colab platform which was used to run Spacy.

It is concluded that Open-Source and proprietary tools offer excellent results in terms of text analysis. The selection of one tool or another should be based on the specific needs and characteristics of each project, as well as the available budget as a company or developer.

Keywords: NLP, Spacy, Open Source, Google API, Google Cloud, AI.

INTRODUCCIÓN

Entre la Inteligencia Artificial y las ciencias de la computación emerge uno de los campos más prometedores y desafiantes para el desarrollo del lenguaje por parte de las maquinas, el Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) busca habilitar la comunicación fluida entre humanos y máquinas, el NLP constituye la base sobre el cual se desarrollan las aplicaciones de Inteligencia Artificial.

La información con respecto al tamaño del mercado mundial del Procesamiento de Lenguaje Natural se estimó en 16,08 mil millones de dólares en el año 2022 y se espera que alcance alrededor de 413,11 mil millones de dólares en 2032 con una tasa de crecimiento anual compuesta del 38,4% durante el período previsto de 2023 a 2032 (precedenceresearch, 2023). esto nos indica el potencial que tiene el desarrollo de tecnologías y aplicaciones basadas en (AI).

Al tener múltiples usos dentro de la Inteligencia Artificial el NLP dispone de una variada selección de herramientas que permiten realizar diferentes tareas y procesos, es aquí donde nos planteamos algunas preguntas como: ¿cuál es la herramienta adecuada que necesito para trabajar en mi proyecto?, ¿es rentable usar herramientas de Procesamiento de Lenguaje Natural Propietario?, ¿Qué tan fiables son las herramientas de NLP Open Source?. El desarrollo de aplicaciones basadas en Inteligencia Artificial no solo requiere una comprensión exhaustiva de los fundamentos teóricos del NLP, sino también la elección cuidadosa de las herramientas adecuadas para transformar esos conceptos en soluciones efectivas

El presente estudio comparativo se realizó de acuerdo con las directrices establecidas en la línea de sistemas de información y comunicación, emprendimiento e innovación y como sublinea de investigación las tecnologías inteligentes de software y hardware. Utilizamos el tipo de investigación descriptivo-comparativo para cumplir con el objetivo de identificar las

principales herramientas NLP, documentar sus ventajas, desventajas y analizar una herramienta de origen Open Source y una de Origen Propietario en términos de su funcionalidad, eficiencia, precio y facilidad de implementación.

Actualmente el Procesamiento de Lenguaje Natural es un componente vital en la expansión de la Inteligencia Artificial transformando la forma en que interactuamos con la tecnología y abriendo un sinfín de posibilidades para mejorar la comunicación y el entendimiento entre humanos y máquinas. A medida que la investigación y la innovación continúen avanzando el Procesamiento de Lenguaje Natural seguirá desempeñando un papel fundamental en la construcción de un futuro digital más conectado y eficiente.

DESARROLLO

El lenguaje natural ha sido el primer medio de comunicación entre los humanos desde tiempos inmemoriales, pero las computadoras solo pueden procesar datos binarios 0s y 1s. Mientras nosotros podemos representar los datos en lenguaje binario, como hacemos que las computadoras entiendan el lenguaje. Aquí es donde empieza el Procesamiento del Lenguaje Natural. Todas las aplicaciones inteligentes que trabajan con recursos de voz o texto están altamente relacionadas con el NLP. (Sowmya Vajjala, 2020, págs. 3-4)

Según (Beysolow, 2018, pág. 1) “El Procesamiento del Lenguaje Natural es un subcampo de las ciencias de la computación que se enfoca en permitir que las computadoras entiendan el lenguaje de una manera “natural”, como lo hacen los humanos”.

Este procesamiento de información requiere una combinación de técnicas lingüísticas, estadísticas y de aprendizaje automático (Machine Learning) para procesar y analizar grandes cantidades de datos no estructurados, como pueden ser llamadas telefónicas, informes, sitios web incluso correo electrónico o los chats de su aplicación favorita de mensajería.

En el ámbito empresarial el Procesamiento del Lenguaje Natural tiene diferentes aplicaciones, como la atención al cliente, la traducción de idiomas, el reconocimiento de voz, la generación de textos o la automatización de procesos.

El NLP se subdivide a su vez en NLU y NLG por sus siglas en inglés o Comprensión de Lenguaje Natural (NLU) y Generación de Lenguaje Natural (NLG) los cuales constituyen las principales fases para comprender y producir el lenguaje humano a través de las computadoras principalmente el NLG tiene la capacidad de proporcionar una descripción verbal de lo que ha sucedido esto también se llama "language out"(producción de texto), ya

que resume la información significativa en texto mediante el concepto conocido como "gramática de gráficos". (Oracle, 2023)

Entre las principales tareas que realiza el Procesamiento del Lenguaje Natural tenemos:

Tokenización de palabras: la tokenización es un tipo particular de segmentación de documentos, la segmentación divide el texto en fragmentos o segmentos más pequeños. Los segmentos de texto tienen menos información que el conjunto. Los documentos se pueden segmentar en párrafos, los párrafos en oraciones, las oraciones en frases y las frases en tokens. (Hobson Lane, 2023, pág. 78)

Lematización: proceso mediante el cual se determina la raíz de las palabras, llamada lema.

Identificar palabras vacías: hay palabras que aparecen con mucha frecuencia las cuales pueden filtrarse antes de realizar cualquier análisis estadístico.

Análisis de dependencias: se utiliza para encontrar cómo se relacionan entre sí todas las palabras de la oración.

Etiquetas POS (tagging): significa partes de la oración, que incluyen sustantivo, verbo, adverbio y adjetivo. Una palabra tiene una o más partes de la oración según el contexto en el que se usa.

Reconocimiento de entidad nombrada (NER): es el proceso de detectar la entidad nombrada, como el nombre de la persona, el nombre de la película, el nombre de la organización o la ubicación.

Fragmentación: se utiliza para recopilar información individual y agruparla en oraciones más grandes.

Herramientas de Procesamiento de Lenguaje Natural Open Source.

Algunas de las herramientas de Procesamiento de Lenguaje Natural Open Source tenemos SpaCy, NLTK, Gensim, Spark NLP.

Además de estas tenemos las bibliotecas de aprendizaje profundo y basadas en tensores como Theano, TensorFlow y Keras también son útiles si desea crear modelos avanzados de aprendizaje profundo basados en redes neuronales, convnets (red neuronal convolucional) y modelos basados en LSTM (Long short-term memory). (Sarkar, 2016, pág. 105)

Spark NLP

Spark NLP es una librería de procesamiento de texto para NLP avanzado en lenguajes de programación como Python, Java y Scala. Su objetivo es proporcionar una interfaz de programación de aplicaciones (API) para los pipelines(canales) de Procesamiento de Lenguaje Natural. También ofrece modelos de redes neuronales preentrenadas, así como soporte para la formación de modelos personalizados. (Oracle, 2023)

Spark NLP proporciona precisión, velocidad y escalabilidad de última generación, (microsoft, 2023) es, con mucho, la biblioteca NLP de código abierto más rápida y poderosa disponible en la actualidad. Es esencial para proyectos de Procesamiento de Lenguaje Natural que requieren un alto rendimiento, además su comunidad activa garantiza que siga siendo líder en el campo del NLP.

El siguiente diagrama muestra los principales componentes de un flujo de Spark NLP.

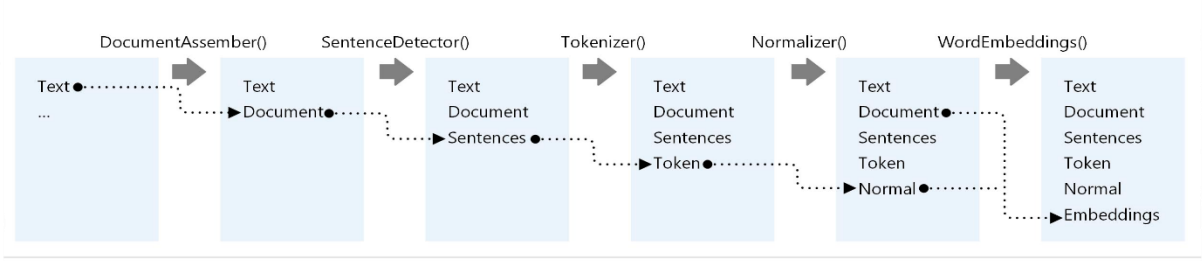

Fig 1. flujo de Spark NLP. Fuente (microsoft, 2023)

NLTK (Natural Language Toolkit)

NLTK es una librería para la creación de programas en Python los cuales trabajan con el procesamiento del lenguaje humano. Proporciona interfaces fáciles de usar para más de 50 corpus y recursos léxicos como WordNet, posee un conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, derivación, etiquetado, análisis y razonamiento semántico, además posee contenedores para bibliotecas de NLP de potencia industrial, y un foro de discusión activo.

NLTK es adecuado para lingüistas, ingenieros, estudiantes, educadores, investigadores y usuarios de la industria por igual. NLTK está disponible para Windows, Mac OS X y Linux. Lo mejor de todo es que NLTK es un proyecto gratuito, de código abierto e impulsado por la comunidad. (Project, 2023)

Gensim

Gensim es una biblioteca Python de código abierto para representar documentos como vectores semánticos. La diferencia principal respecto al resto de librerías de lenguaje natural para Python reside en que Gensim es capaz de identificar automáticamente la temática del conjunto de documentos a tratar. También permite analizar la similitud entre ficheros, algo realmente útil cuando utilizamos la librería para realizar búsquedas. (datos.gob.es, 2022)

La mejor parte de Gensim es que contiene una adaptación de Python del muy popular modelo Word2vec de Google (originalmente disponible como un paquete C), un modelo de red neuronal implementado para aprender representaciones distribuidas de palabras, donde palabras similares (semánticas) aparecen cerca unas de otras. (Sarkar, 2016, pág. 105)

“Gensim se ejecuta en Linux, Windows y Mac OS X, y debería ejecutarse en cualquier otra plataforma que admita Python 3.6+ y Numpy”, (Gensim, 2022).

Spacy

Spacy es una biblioteca gratuita de código abierto para el procesamiento avanzado del lenguaje natural (NLP) en Python.

Está diseñado específicamente para uso en producción y le ayuda a crear aplicaciones que procesan y "comprenden" grandes volúmenes de texto. Se puede utilizar para crear sistemas de extracción de información o de comprensión del lenguaje natural, para preprocesar texto o para un aprendizaje profundo. (Spacy, 2023)

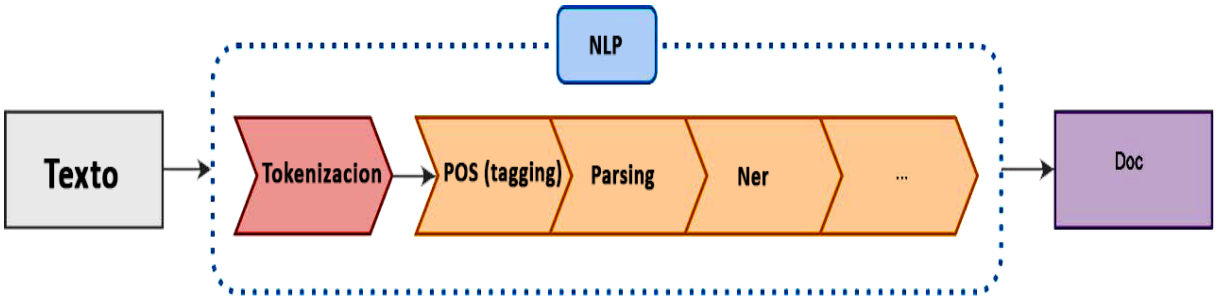

Fig 2. flujo o pipeline de NLP en Spacy inicia con el texto que queremos analizar hasta obtener el resultado final, aplica el mismo procedimiento para analizar sitios web. Fuente: (Spacy, 2023)

Spacy se creó con el objetivo de facilitar la creación de productos reales, la librería contempla los aspectos prácticos de un producto de software real, en el que es necesario tener en cuenta aspectos tan importantes como:

La velocidad de ejecución, puesto que cuando tenemos una aplicación real necesitamos que la experiencia sea lo más fluida posible y no podemos soportar largos tiempos de espera entre ejecuciones de los algoritmos.

Spacy no solo proporciona herramientas de código de bajo nivel, sino que soporta los procesos desde que creamos hasta que integramos esta parte algorítmica con otras partes de la aplicación como las bases de datos o las interfaces de usuario final.

Esta librería esta optimizada para ejecutarse fácilmente en servidores estándar (basados en CPU) sin necesidad de usar procesadores gráficos (GPU)

Spacy posee un gran número de modelos y flujos pre-entrenados (84 flujos) en 25 idiomas diferentes además del soporte para más de 73 idiomas. Tiene una gran operatividad sobre el idioma español lo cual es complicado encontrar en otras librerías y herramientas (Spacy, 2023).

Tabla 1 Características de las herramientas Open Source

Características de las herramientas Open Source				
	SpaCy	Gensim	Spark NLP	NLTK
Idiomas soportados	73+	50+	192+	50+
Escalabilidad	Excelente	Buena	Excelente	Buena
Eficiencia	Buena	Buena	Buena	Buena
Recursos	CPU / GPU	CPU / GPU	CPU / GPU	CPU / GPU
Comunidad	Grande	Media	Grande	Grande

Elaborado por: Klever Angamarca

Spacy, Gensim, Spark NLP y NLTK son algunas de las herramientas de Procesamiento de Lenguaje Natural Open Source muy populares que ofrecen una amplia gama de características y capacidades.

Herramientas de Procesamiento de Lenguaje Natural Propietario.

Google Natural Language API

La API de Cloud Natural Language proporciona a los desarrolladores tecnologías de comprensión del lenguaje natural. Esta API forma parte de la familia más amplia de la API de Cloud Machine Learning. Dispone de bibliotecas cliente en varios lenguajes: C#, Go, Java, Node Js, PHP, Python y Ruby.

Las tareas que se pueden realizar a través de sus APIs son:

- Análisis de entidades: Identifique entidades y etiquetarlas por tipo, como persona, organización, ubicación, eventos, productos y medios.
- Análisis de sentimiento Comprenda el sentimiento general expresado en un bloque de texto.
- Análisis de sentimiento de entidad Comprenda el sentimiento de las entidades identificadas en un bloque de texto.
- Análisis de sintaxis Extraiga tokens y oraciones, identifique partes del discurso (PoS) y cree árboles de análisis de dependencias para cada oración.
- Clasificación de contenido Identifique las categorías de contenido que se aplican a un bloque de texto.
- Moderación de texto Identifique categorías dañinas y sensibles que se aplican a un bloque de texto.

Los precios por el uso de la API de Natural Language se calculan mensualmente en función de las peticiones que realiza la API, por ejemplo, si envía tres solicitudes de análisis

de opinión que contienen 800, 1500 y 600 caracteres respectivamente, se le cobrarán cuatro unidades: una por la primera solicitud (800), dos por la segunda solicitud (1500) y una por la tercera solicitud (600), (Google, 2023).

IBM Watson Natural Language Understanding

IBM Watson Natural Language Understanding es un conjunto de herramientas sofisticadas de Procesamiento de Lenguaje Natural que permite a los usuarios analizar texto y extraer información. Dispone de bibliotecas cliente en varios lenguajes: Android, Java, Node , .NET y Swift. Las tareas que se pueden realizar a través de su API o biblioteca cliente son:

- Clasificación de texto
- Identificación de conceptos de alto nivel
- Análisis de emociones
- Extracción de entidades
- Extracción de palabras clave
- Extracción de metadatos
- Extracción de relaciones
- Análisis semántico
- Análisis de sentimiento
- Extracción de entidades y relaciones con modelos personalizados

Al igual que las demás plataformas, dispone de un plan gratuito hasta los 30,000 elementos al mes y un modelo personalizado incluido.

En el plan de pago tiene tres niveles que van desde los $0.003/elemento de NLU hasta un máximo de 250,000 elementos.

El nivel dos va desde $0.0001/elemento de NLU hasta los 5 millones de elementos.

Por último, el nivel tres de más de 5 millones de elementos a un valor de $0,0002 /elemento de NLU.Los tres niveles incluyen la opción de incluir modelos personalizados por $800/modelo/mes, (ibm, 2023).

Amazon Comprehend

Amazon Comprehend utiliza el procesamiento del lenguaje natural (NLP) para extraer información sobre el contenido de los documentos. Desarrolla conocimientos identificando entidades, frases clave, lenguaje, sentimientos y otros elementos comunes en un documento.

Por ejemplo, con Amazon Comprehend puede buscar en las redes sociales menciones de productos o escanear un repositorio de documentos completo en busca de frases clave.

Puede utilizar los modelos previamente entrenados que proporciona Amazon Comprehend o puede entrenar sus propios modelos personalizados para la clasificación y el reconocimiento de entidades, (Amazon, 2023).

Todas las funciones de Amazon Comprehend aceptan archivos de texto UTF-8 como entrada. Además, la clasificación personalizada y el reconocimiento de entidades personalizadas aceptan archivos de imagen, archivos PDF y archivos de Word como entrada. También puede examinar y analizar documentos en una variedad de idiomas, según la característica específica, (Amazon, 2023) .

Amazon Comprehend tiene un nivel gratuito hasta 50 000 unidades de texto (5 millones de caracteres) por API y por mes. Los precios para los planes de pago varían según las tareas de NLP que se realicen mediante su API, aunque sus precios se encuentran entre 0,00005 USD y 0,003 USD en el plan básico hasta 10 millones de unidades.

Azure AI API

El Lenguaje de Azure AI es un servicio basado en la nube destinado a brindar capacidades de NLP para la comprensión y análisis de texto. Este servicio se usa para ayudar a compilar aplicaciones inteligentes mediante Language Studio basado en web, las API REST y las bibliotecas cliente, (microsoft, 2023).

El servicio de lenguaje también dispone de bibliotecas cliente en varios lenguajes: C#, Go, Java, JavaScript, Python y Ruby.

Las tareas que se pueden realizar a través de sus APIs o bibliotecas cliente son:

- Análisis de sentimiento
- Extracción de frases clave
- Reconocimiento de entidades
- Identificación de idioma

Azure AI ofrecen distintos planes en función del número de transacciones mensuales.

Una transacción corresponde al número de unidades de 1000 caracteres en un documento que se proporcionan como entrada en una solicitud de Text Analytics API, en el plan gratuito disponen de 5000 registros al mes.

Además, con los planes de pago, se pueden realizar más transacciones pagando una cuota por cada 1000 nuevas transacciones. Se incluye soporte técnico gratis de facturación y administración de suscripciones. (microsoft, 2023)

Tabla 2 Características de las herramientas de Software Propietario

Características de las herramientas de Software Propietario				
	Google Natural Language API	Amazon Comprehend	IBM Watson Natural Language Understanding	Azure AI API
Idiomas compatibles	120+	100+	13+	120+
Velocidad	Alta	Alta	Alta	Alta
Coste	Basado en el uso	Basado en el uso	Basado en el uso	Basado en el uso
Documentación	Completa	Completa	Completa	Completa
Soporte	Excelente	Excelente	Excelente	Excelente
Recursos	CPU / GPU	CPU / GPU	CPU / GPU	CPU / GPU
Escalabilidad	Excelente	Excelente	Buena	Buena
Comunidad	Grande	Media	Baja	Media

Elaborado por: Klever Angamarca

Tabla 3 Ventajas y Desventajas de las herramientas Open Source y Software Propietario

Ventajas y Desventajas de las herramientas Open Source y Software Propietario		
	Ventajas	Desventajas
Herramientas Open Source	• Su uso es libre • Acceso al código fuente • Comunidad usuarios siempre activa • Mayor escalabilidad y personalización	• Las actualizaciones dependen de la comunidad • Puede estar limitado en funcionalidades • Su uso puede ser complejo
Herramientas Software Propietario	• Asistencia directa de la empresa • Incluye plataforma para su uso • Seguridad • Variedad de herramientas	• Costoso de mantener • Vínculo con tarjeta bancaria • Sujeto a limitaciones geográficas

Elaborado por: Klever Angamarca

Una vez recolectada y analizada la información sobre las herramientas Open Source y Software Propietario realizamos una prueba de su funcionamiento, para esto seleccionamos por el lado Open Source a Spacy y por el lado propietario seleccionamos Google Natural Language API para determinar de mejor manera sus funcionalidades capacidades y requisitos técnicos evaluando las principales tareas del NLP como son:

- Tokenización
- Análisis de sentimiento
- Reconocimiento de Entidades Nombradas
- Análisis de Sintaxis

Pasos para la ejecución de Google Natural Language API

- Crea una cuenta de Google Cloud, puedes hacerlo en la página de inicio de Google Cloud Platform.
- Activamos la prueba gratuita para acceder a los servicios de Google Cloud sin cargo durante 90 días.
- Crea un proyecto de Google Cloud el cual es un contenedor para tus tareas de Google Cloud.

Una vez creado nuestro proyecto en la plataforma necesitamos activar la API de Google Cloud Natural Language:

- Iniciamos sesión en la consola de Google Cloud.
- En el menú de la izquierda, hacemos clic en APIs y servicios.
- En la página Biblioteca de APIs, buscamos Natural Language.
- Habilitamos la API.
- En la página Configuración, clicamos en el botón Crear clave de API.
- En la página Crear clave de API, selecciona el tipo de clave de API de servicio.
- Pulsamos en el botón Crear.
- Se creará una clave de API y se mostrará una ventana emergente con la información de la clave.
- Copia la información de la clave de API

Para obtener mayores detalles sobre la creación de un proyecto y la activación del API podemos visitar su web oficial. (Google, 2023)

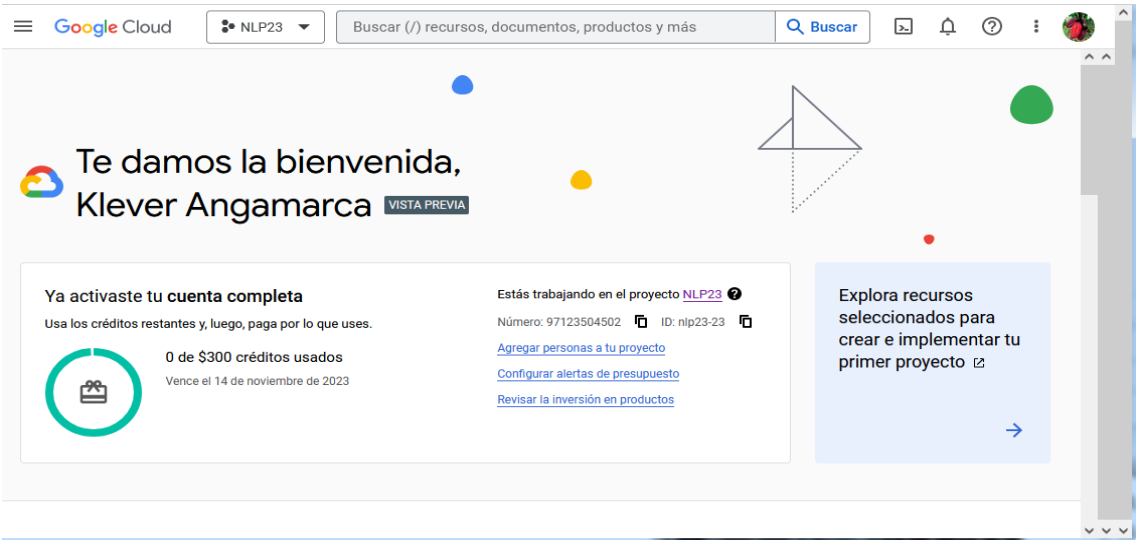

Fig 4. página de inicio de Google Cloud Platform

Una vez activada la API de Natural Language podemos empezar a realizar nuestros primeros análisis en la plataforma.

Activamos el uso del servicio mediante la siguiente línea de comando gcloud services enable language.googleapis.com

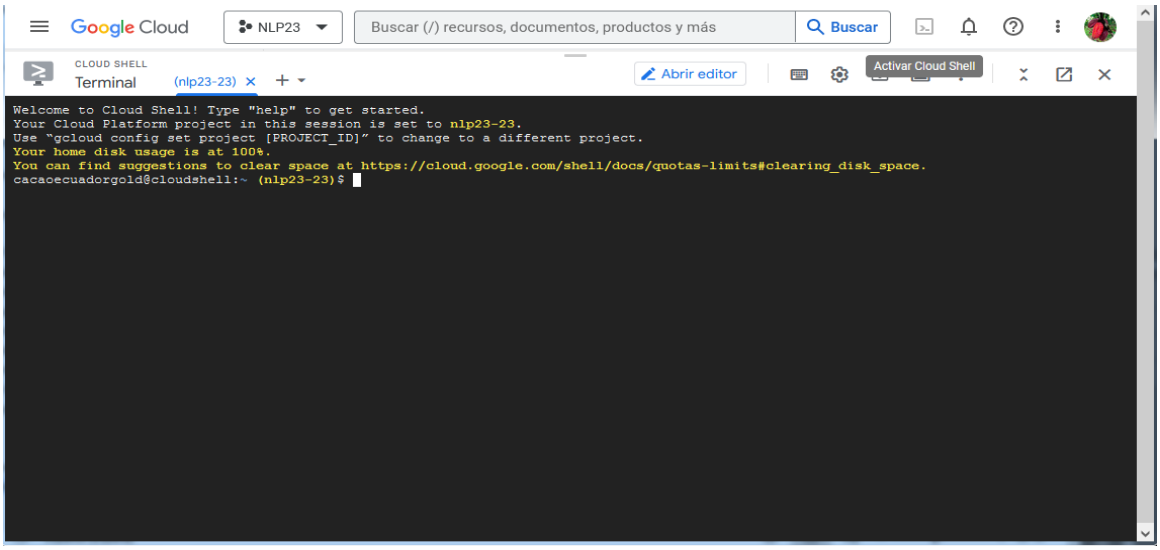

Fig 5. Google Cloud Terminal.

Ejecución de Spacy en la plataforma de Google Colab

Elegimos esta plataforma ya que es de acceso gratuito y nos ofrece un espacio en disco de 100GB y 12GB RAM con lo cual es bastante aceptable para empezar a trabajar con Spacy.

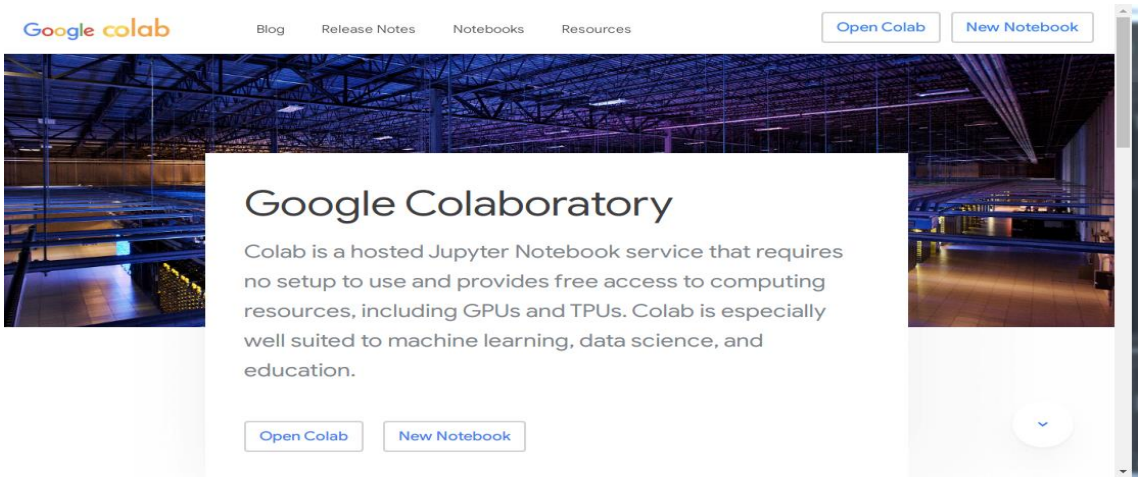

Fig 6. Página de inicio de Google Colab

- Para acceder a los servicios de la plataforma de Google Colab ingresamos con nuestra cuenta de Gmail.
- Creamos una nueva notebook en Google Colab.
- Una vez dentro del notebook descargamos el modelo de lenguaje con el que vamos a trabajar.
- Importamos el módulo de Spacy y podemos empezar a realizar las diferentes tareas de NLP.

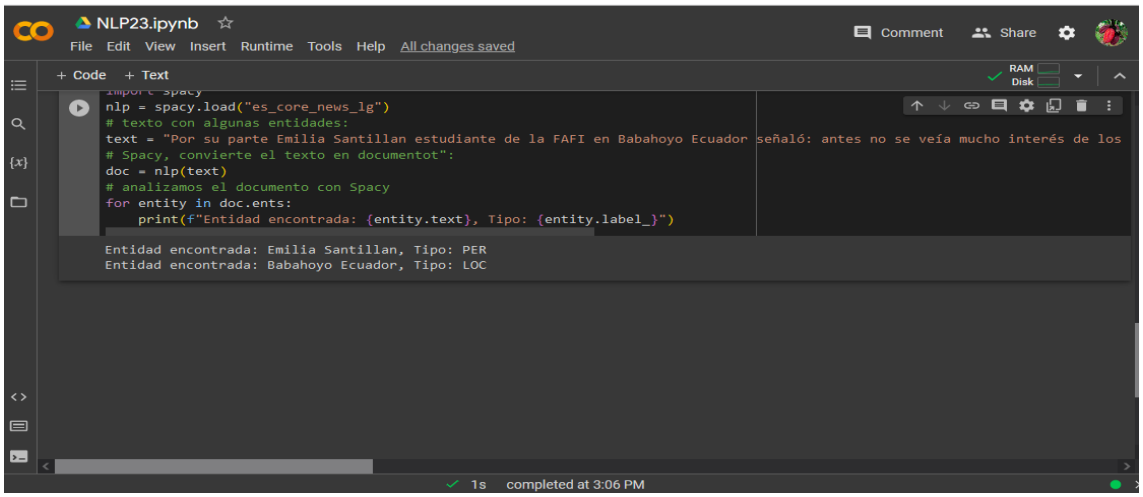

Fig 7. Entorno de desarrollo de Google Colab

Tabla 4 Análisis comparativo entre Google NLP API y SPACY

Análisis comparativo entre Google NLP API y Spacy		
	Google NLP API	Spacy
Velocidad ejecución	2s	3s
Facilidad de Uso	Difícil	Fácil
Precisión	Excelente	Excelente
Precio	$0.000125	$0.0
Lenguajes de Programación	C#, Go, Java, Node Js, PHP, Python y Ruby.	Python, Cython

Elaborado por: Klever Angamarca

El desempeño de las herramientas Open Source es muy eficiente frente a sus competidores de origen Propietario, en la práctica las dos herramientas poseen una gran precisión en las principales tareas de NLP, la compatibilidad con otros lenguajes de programación es superior en las herramientas de origen propietario, si está empezando en el mundo del Procesamiento de Lenguaje Natural o si desea realizar un proyecto a largo plazo las herramientas Open

Source están al mismo nivel en desarrollo y tecnología que las herramientas de origen Propietario.

Las herramientas de código abierto son una opción más económica que las herramientas propietarias, ya que solo requieren una inversión inicial en equipos para su ejecución. En cambio, las herramientas propietarias pueden tener un costo adicional por cada característica, por lo que es importante informarse correctamente antes de realizar una inversión en ellas.

En cuanto a los requisitos técnicos para trabajar con herramientas de NLP en tareas altamente complejas se recomienda un equipo con un procesador multinúcleo de última generación y una memoria de 16GB a 32GB de RAM adicional de una GPU para reducir el tiempo de proceso.

También puede optar por servicios en línea como Google Colab en la cual bajo una suscripción mensual le garantiza la potencia suficiente para trabajar con grandes volúmenes de datos en proyectos de análisis y aprendizaje automático aprovechando su infraestructura sin la necesidad de configurar y administrar hardware localmente.

CONCLUSIONES

Las herramientas Open Source están a la par con la tecnología de herramientas propietarias ya que proporcionan características altamente sofisticadas. Además, ofrecen una excelente flexibilidad al personalizar y adaptar nuestras tareas de procesamiento lo que es crucial en proyectos de Inteligencia Artificial.

Al elegir herramientas de NLP propietarias se debe tener en cuenta que estas poseen funciones específicas y el uso de estos servicios dentro de la plataforma está sujeto a un costo adicional. Esta consideración financiera es vital en la planificación presupuestaria, asegurando que un proyecto esté preparado para cubrir los gastos asociados con el despliegue y el mantenimiento de estas herramientas.

La práctica reveló que la disponibilidad de corpus de texto en español es menor que en el idioma inglés. Esta limitación destaca la necesidad de crear y recopilar recursos lingüísticos en español para fortalecer el desarrollo de herramientas y aplicaciones de Procesamiento de Lenguaje Natural en este idioma.

Se ha evidenciado una mayor compatibilidad de las herramientas de origen propietario hacia otros lenguajes de programación, esto permite a los desarrolladores aprovechar sus habilidades especialmente en proyectos que requieren la interoperabilidad con múltiples tecnologías y plataformas, contribuyendo a una mayor eficiencia y flexibilidad en la implementación de soluciones informáticas.

Considero que las herramientas Open Source son ideales para empezar proyectos de Procesamiento de Lenguaje Natural ya que no requieren licencias o planes mensuales para su

uso, las herramientas de NLP Open Source también son una excelente opción en el ámbito educativo y pueden ser usadas por cualquier empresa o institución interesada en este campo.

La elección final entre herramientas NLP Open Source y Software Propietario depende de factores como las necesidades de nuestros proyectos, el presupuesto, la capacidad de personalización requerida o la compatibilidad con los lenguajes de programación, es importante considerar todos estos detalles antes de tomar una decisión.

BIBLIOGRRAFÍA

Amazon. (2023). *¿Qué es el Procesamiento de lenguaje natural (NLP)?* Recuperado el 22 de 07 de 2023, de Amazon Web Services: https://aws.amazon.com/es/what-is/nlp/#seo-faq-pairs#why-is-nlp-important

Amazon. (2023). *Amazon Comprehend*. Recuperado el 15 de 08 de 2023, de Amazon: https://docs.aws.amazon.com/es_es/comprehend/latest/dg/what-is.html

Beysolow, T. (2018). *Applied Natural Language Processing with Python.* San Francisco, California, USA: Apress, Berkeley, CA. doi:https://doi.org/10.1007/978-1-4842-3733-5_1

datos.gob.es. (02 de 08 de 2022). *10 Librerías populares de procesamiento del lenguaje natural*. Recuperado el 27 de 07 de 2023, de datos.gob.es: https://datos.gob.es/es/blog/10-librerias-populares-de-procesamiento-del-lenguaje-natural

Gensim. (21 de 12 de 2022). *Why Gensim?* Recuperado el 17 de 08 de 2023, de Why Gensim?: https://radimrehurek.com/gensim/

github. (2023). *Spark NLP: State-of-the-Art Natural Language Processing*. Recuperado el 15 de 08 de 2023, de github: https://github.com/JohnSnowLabs/spark-nlp

Google. (2023). *Cloud Natural Language pricing*. Recuperado el 15 de 08 de 2023, de Google: https://cloud.google.com/natural-language/pricing

Google. (2023). *Cómo usar la API de Natural Language desde Documentos*. Recuperado el 28 de 08 de 2023, de Google: https://www.cloudskillsboost.google/focuses/680?locale=es&parent=catalog

Hobson Lane, M. D. (2023). *Natural Language Processing in Action, Second edition* (Vol. 2). Recuperado el 27 de 07 de 2023

ibm. (2023). *Watson Natural Language Understanding*. Recuperado el 15 de 08 de 2023, de ibm: https://www.ibm.com/es-es/cloud/watson-natural-language-understanding/pricing

microsoft. (23 de 07 de 2023). *¿Qué es Lenguaje de Azure AI?* Recuperado el 17 de 08 de 2023, de microsoft: https://learn.microsoft.com/es-es/azure/ai-services/language-service/overview

microsoft. (2023). *Procesamiento de lenguaje natural personalizado a gran escala*. Recuperado el 22 de 08 de 2023, de microsoft: https://learn.microsoft.com/es-es/azure/architecture/solution-ideas/articles/large-scale-custom-natural-language-processing

Oracle. (2023). *¿Qué es el procesamiento de lenguaje natural (NLP)?* Recuperado el 28 de 08 de 2023, de oracle: https://www.oracle.com/mx/artificial-intelligence/what-is-natural-language-processing/

precedenceresearch. (08 de 2023). *Natural Language Processing Market Size, Report 2032* . Obtenido de precedenceresearch: https://www.precedenceresearch.com/natural-language-processing-market

Project, N. (2023). *Natural Language Toolkit*. Recuperado el 27 de 07 de 2023, de NLTK Project: https://www.nltk.org/

Sarkar, D. (2016). *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable.* Bangalore, Karnataka, INDIA. doi:DOI 10.1007/978-1-4842-2388-8

Sowmya Vajjala, B. M. (2020). *Practical Natural Language Processing.* United States of America.: O'Reilly Media, Inc.

Spacy. (2023). *whats-spacy*. Recuperado el 20 de 08 de 2023, de Spacy: https://spacy.io/usage/spacy-101

ANEXOS

Prueba práctica realizada en la plataforma de Google Cloud

Autorizamos el uso de nuestras credenciales para realizar llamadas a la API de Google Cloud

gcloud services enable language.googleapis.com

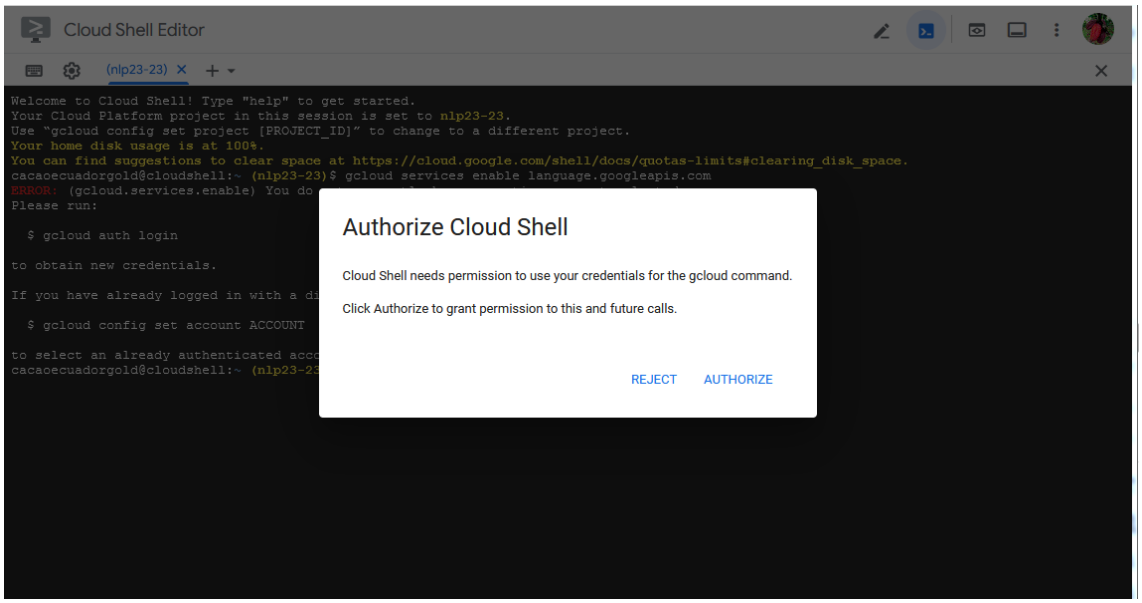

Fig. Terminal de Google Cloud

Una vez que autorizamos el uso del servicio de la API de Procesamiento de Lenguaje Natural podemos empezar a realizar nuestras primeras practicas directamente en la terminal de Google Cloud.

Análisis de sintaxis con la API de Google Natural Language

gcloud ml language analyze-syntax --content="La Universidad Técnica de Babahoyo (UTB) es una universidad Pública ubicada en la provincia de Los Ríos, cuya sede se encuentra en la ciudad de Babahoyo. El 5 de octubre de 1971 fue creada oficialmente por decreto del entonces presidente de la República, José María Velasco Ibarra con las facultades de Ingeniería Agronómica, Medicina, Veterinaria y Ciencias de la Educación."

Fig. Análisis de Sintaxis

Análisis de entidades con la API de Google Natural Language

gcloud ml language analyze-entities --content='El apoyo que recibió Daniel Noboa sorprendió a ingenuos, seguidoras y seguidores por su parte Luisa González no salió a debatir.

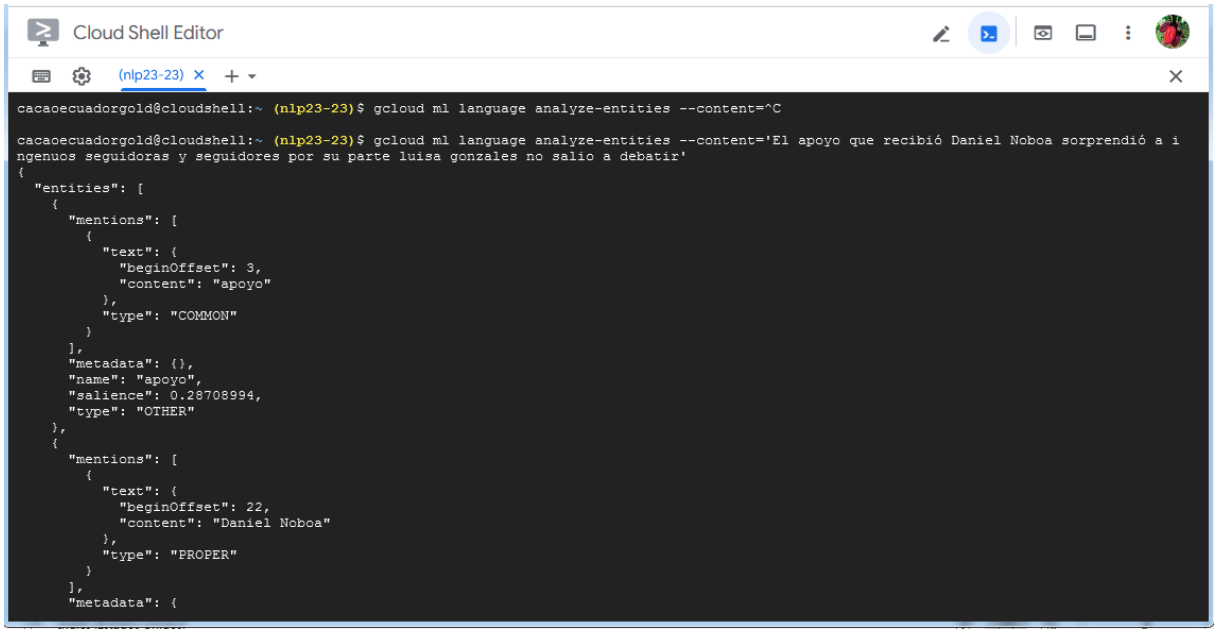

Fig. Análisis de entidades

Análisis de Sentimiento de entidades

gcloud ml language analyze-entity-sentiment --content="Por su parte Emilia Santillán estudiante de la FAFI en Babahoyo Ecuador señaló: antes no se veía mucho interés de los jóvenes pero ahora a nosotros nos interesa dar un voto inteligente y somos más críticos y eso es un reto para los candidatos"

Fig. Análisis de Sentimiento de entidades

sentiment contiene los valores de *nivel de opiniones de la oración* adjuntos a cada oración, el cual contiene score entre -1.0 (negativo) y 1.0 (positivo), y los valores de **magnitude** indica la intensidad general de la emoción entre 0.0 y 1.0.

Análisis práctico realizado en la plataforma de Google Colab con Spacy

Análisis de entidades nombradas en SpaCy

```
nlp = spacy.load('es_core_news_lg')

def ver_ents(doc):

    if doc.ents:

        for ent in doc.ents:

            print(ent.text, '-', ent.label_, '-', str(spacy.explain(ent.label_)))

    else:

        print('No se encontraron entidades.')
```

doc = nlp("La Universidad Técnica de Babahoyo (UTB) es una universidad Pública ubicada en la provincia de Los Ríos, cuya sede se encuentra en la ciudad de Babahoyo. El 5 de octubre de 1971 fue creada oficialmente por decreto del entonces presidente de la República José María Velasco Ibarra con las facultades de Ingeniería Agronómica, Medicina, Veterinaria y Ciencias de la Educación.")

```
doc.text

ver_ents(doc)
```

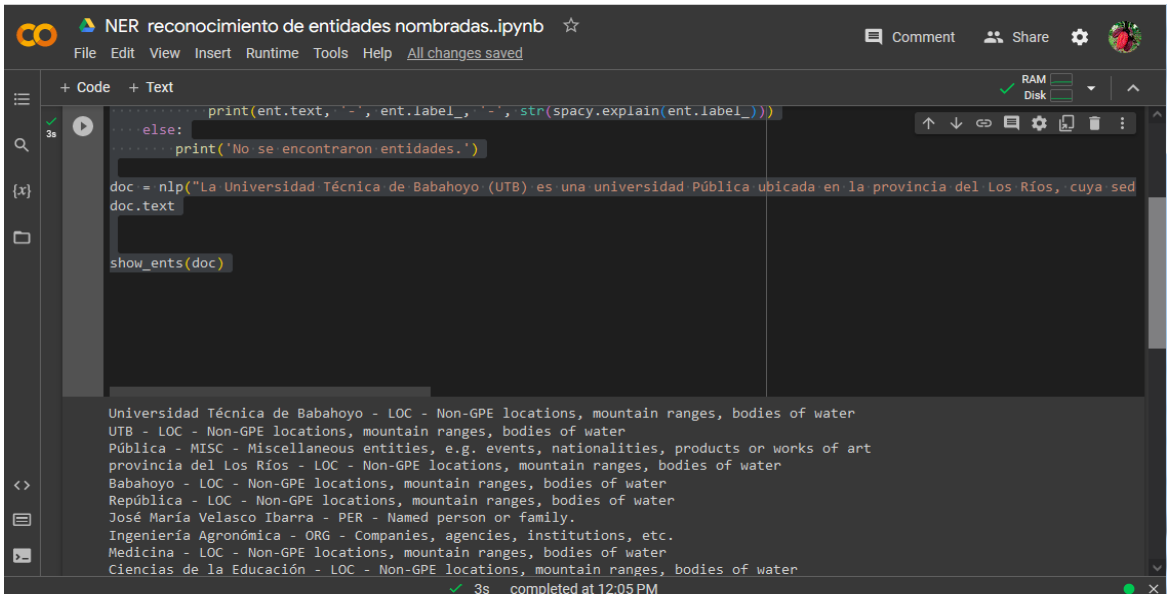

Fig. Análisis de reconocimiento de entidades

Análisis de dependencia en Spacy

from spacy import displacy

nlp = spacy.load('es_core_news_lg')

doc = nlp ("Por su parte Emilia Santillán estudiante de la FAFI señaló: antes no se veía mucho interés de los jóvenes, pero ahora a nosotros nos interesa dar un voto inteligente y somos más críticos y eso es un reto para los candidatos.")

doc.text

displacy.render(doc, style='dep', jupyter=True)

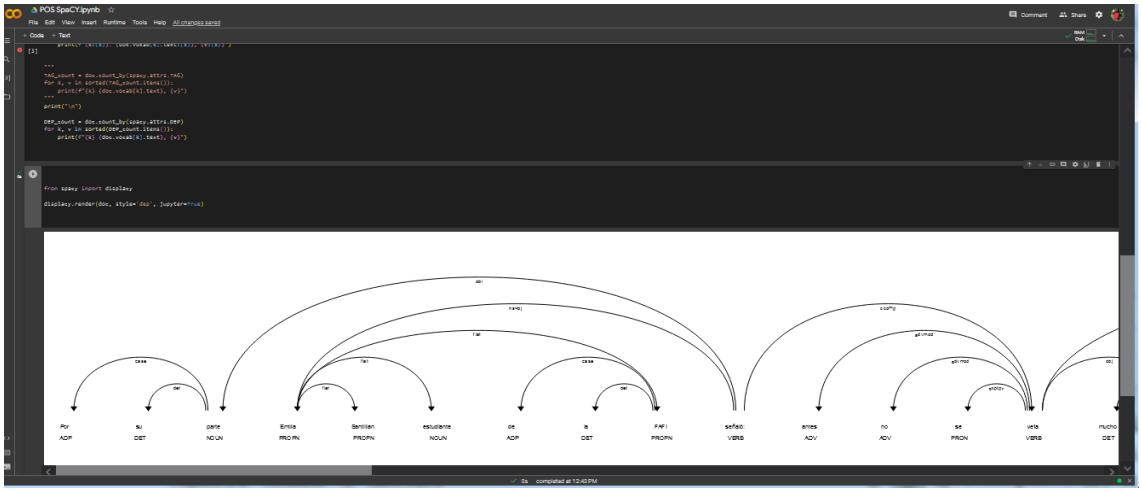

Fig. Análisis de dependencia

Tokenizacion de texto en Spacy

from spacy.tokenizer import Tokenizer

py_nlp= spacy.load("es_core_news_lg")

py_tokenizer = Tokenizer (py_nlp.vocab)

py_tokens = py_tokenizer ("La Universidad Técnica de Babahoyo (UTB) es una universidad Pública ubicada en la provincia del Los Ríos, cuya sede se encuentra en la ciudad de Babahoyo. El 5 de octubre de 1971 fue creada oficialmente por decreto del entonces presidente de la República, José María Velasco Ibarra con las facultades de Ingeniería Agronómica, Medicina, Veterinaria y Ciencias de la Educación ")

print ("Blank tokenizer", end=" : ")

for py_tokens in py_tokens:

print (py_tokens, end=', ')

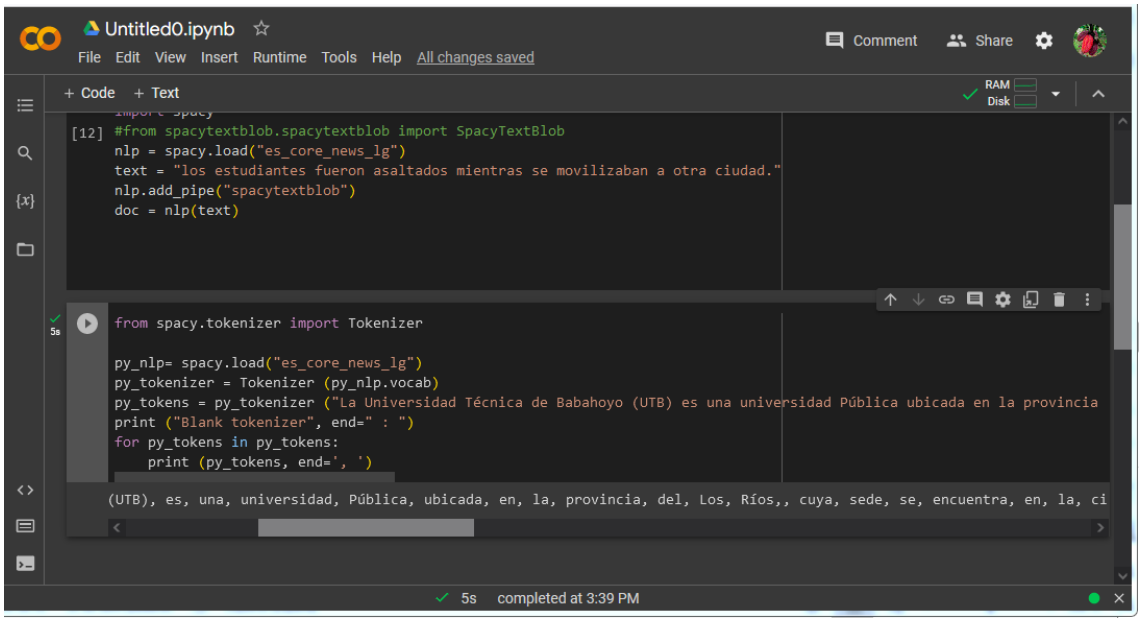

Fig. Tokenizacion

CERTIFICADO DE ANÁLISIS ANTIPLAGIO

CERTIFICADO DE ANÁLISIS
magister

INTRODUCCIÓN

8% Similitudes

< 1% Texto entre comillas
0% similitudes entre comillas

2% Idioma no reconocido

Nombre del documento: INTRODUCCIÓN.pdf
ID del documento: 2fc15000e67f17f19031461dd63ada0cd30fa9d7
Tamaño del documento original: 466,54 kB
Autor: Klever Angamarca

Depositante: Klever Angamarca
Fecha de depósito: 27/9/2023
Tipo de carga: url_submission
fecha de fin de análisis: 27/9/2023

Número de palabras: 3811
Número de caracteres: 25.480

Ubicación de las similitudes en el documento:

Fuentes principales detectadas

N°	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	docs.aws.amazon.com | ¿Qué es Amazon Comprehend? - Amazon Comprehend https://docs.aws.amazon.com/es_es/comprehend/latest/dg/what-is.html	2%		Palabras idénticas: 2% (76 palabras)
2	www.oracle.com | ¿Qué es el procesamiento de lenguaje natural (NLP)? | Oracle C... https://www.oracle.com/co/artificial-intelligence/what-is-natural-language-processing/ 3 fuentes similares	2%		Palabras idénticas: 2% (69 palabras)
3	Documento de otro usuario #267a82 El documento proviene de otro grupo	< 1%		Palabras idénticas: < 1% (32 palabras)

Fuentes con similitudes fortuitas

N°	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	learn.microsoft.com | Procesamiento de lenguaje natural personalizado a gran es... https://learn.microsoft.com/es-ES/azure/architecture/solution-ideas/articles/large-scale-custom-nat...	< 1%		Palabras idénticas: < 1% (20 palabras)
2	www.doi.org | IEEE 2019 International Conference on Information Systems and C... https://www.doi.org/10.1109/INCISCOS49368.2019.00041	< 1%		Palabras idénticas: < 1% (19 palabras)