

Sistema informático para el análisis de entidades nombradas en el procesamiento de documentos digitales en la empresa DATYS

Autor

Luis Andrés Licea Berenguer

Tutores

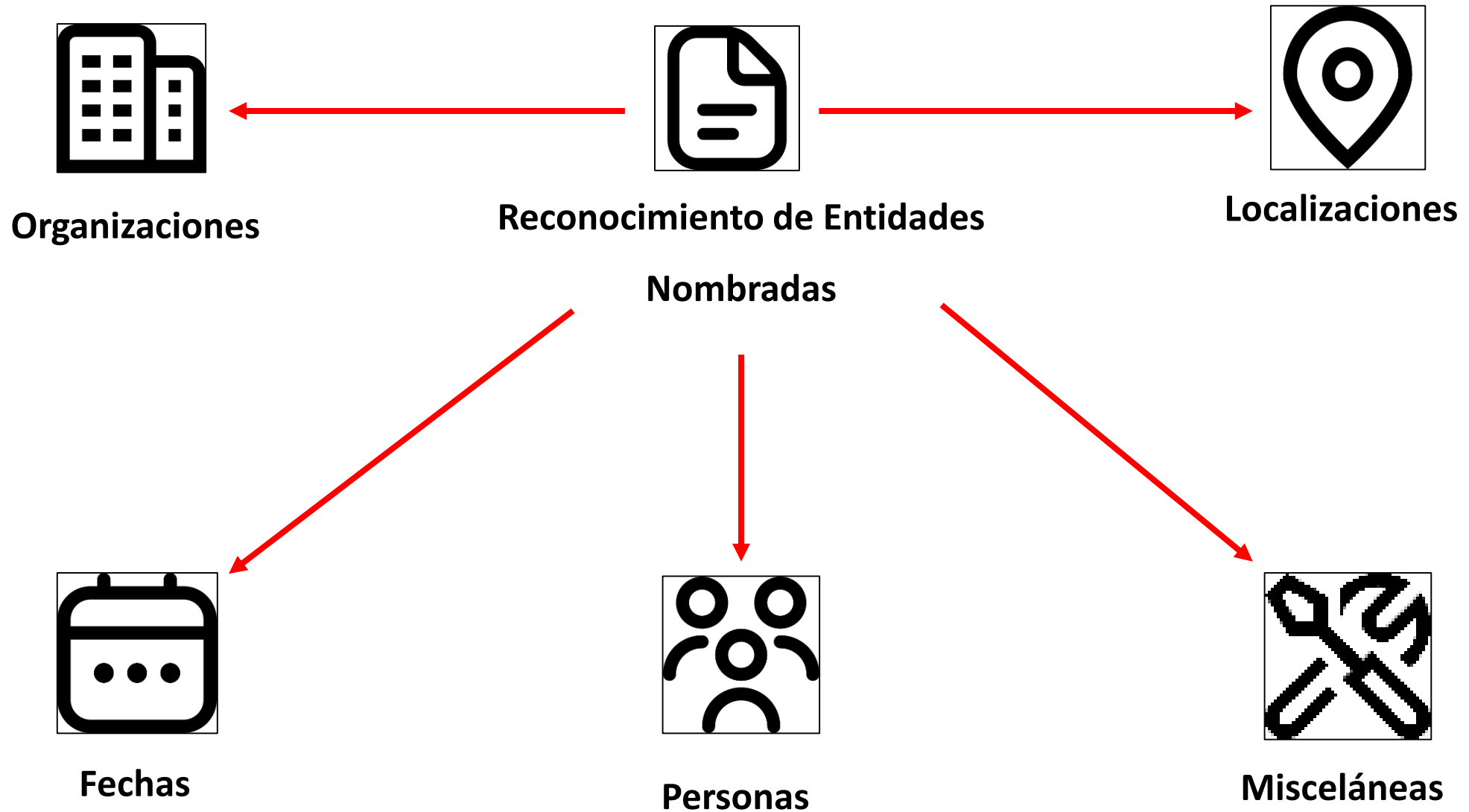
Dr.C. Dionis López Ramos

MSc. Jose Erinaldo Cruzata Ferrer

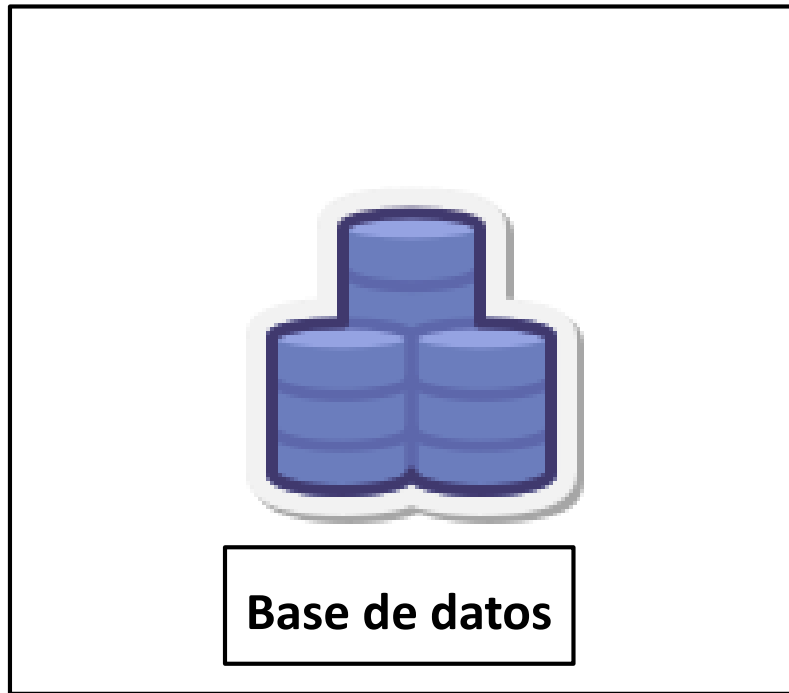
**Facultad de Ingeniería en Telecomunicaciones, Informática y Biomédica
Departamento de Ingeniería Informática**

Santiago de Cuba, abril, 2024

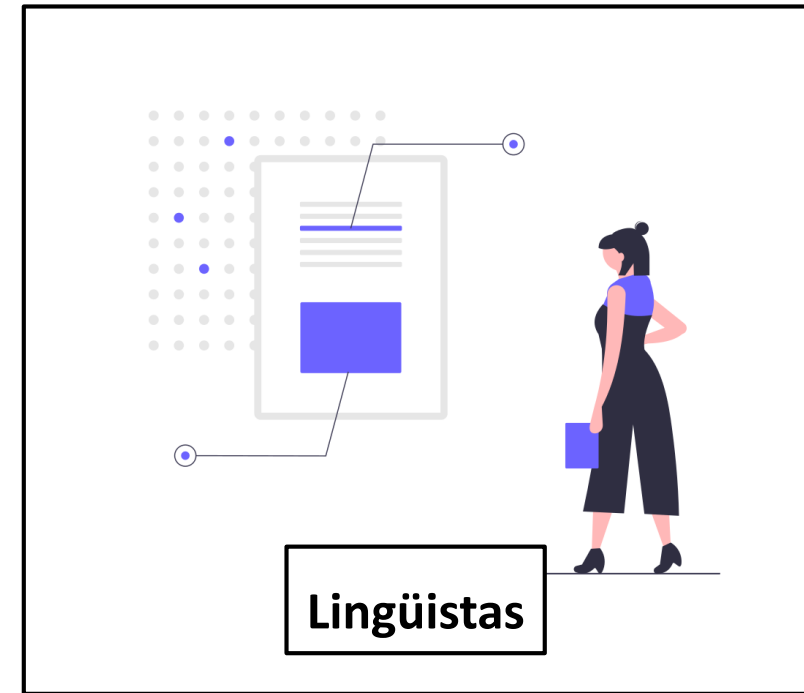
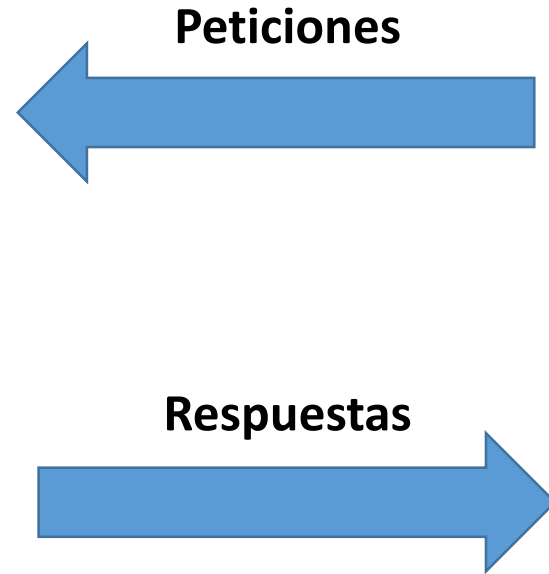
INTRODUCCIÓN



Situación Problemática



- Volumen grande de información



- Horario laboral
- Costoso
- Poco personal capacitado

Problema de Investigación

La empresa DATYS de Santiago de Cuba tiene limitaciones en el reconocimiento de entidades nombradas en español al procesar volúmenes masivos de datos diariamente en Elasticsearch.

Objeto de estudio

Reconocimiento de Entidades Nombradas.

Campo de estudio

Modelos de procesamiento del lenguaje natural

Objetivo general

Crear un sistema de procesamiento del lenguaje natural para el reconocimiento de entidades nombradas en volúmenes masivos de datos en español almacenados en Elasticsearch.

Objetivos Específicos

- Estudio del estado del arte de las técnicas de reconocimiento de entidades nombradas.
- Estudio de las herramientas de Procesamiento de Lenguaje Natural (Spacy)
- Estudio de las propuestas de grandes modelos de lenguajes y su uso en la generación de oraciones.
- Análisis del modelo de indexación de información en Elasticsearch de la empresa DATYS.
- Diseño de una propuesta para el reconocimiento de entidades en información indexado en Elasticsearch según los requerimientos de la empresa DATYS para el procesamiento de información.
- Implementación del modelo y sistemas diseñado para el reconocimiento de entidades nombradas y la generación de datos de entrenamiento.
- Pruebas de la propuesta del modelo y sistema implementado.
- Evaluación por la empresa Datys del sistema implementado.

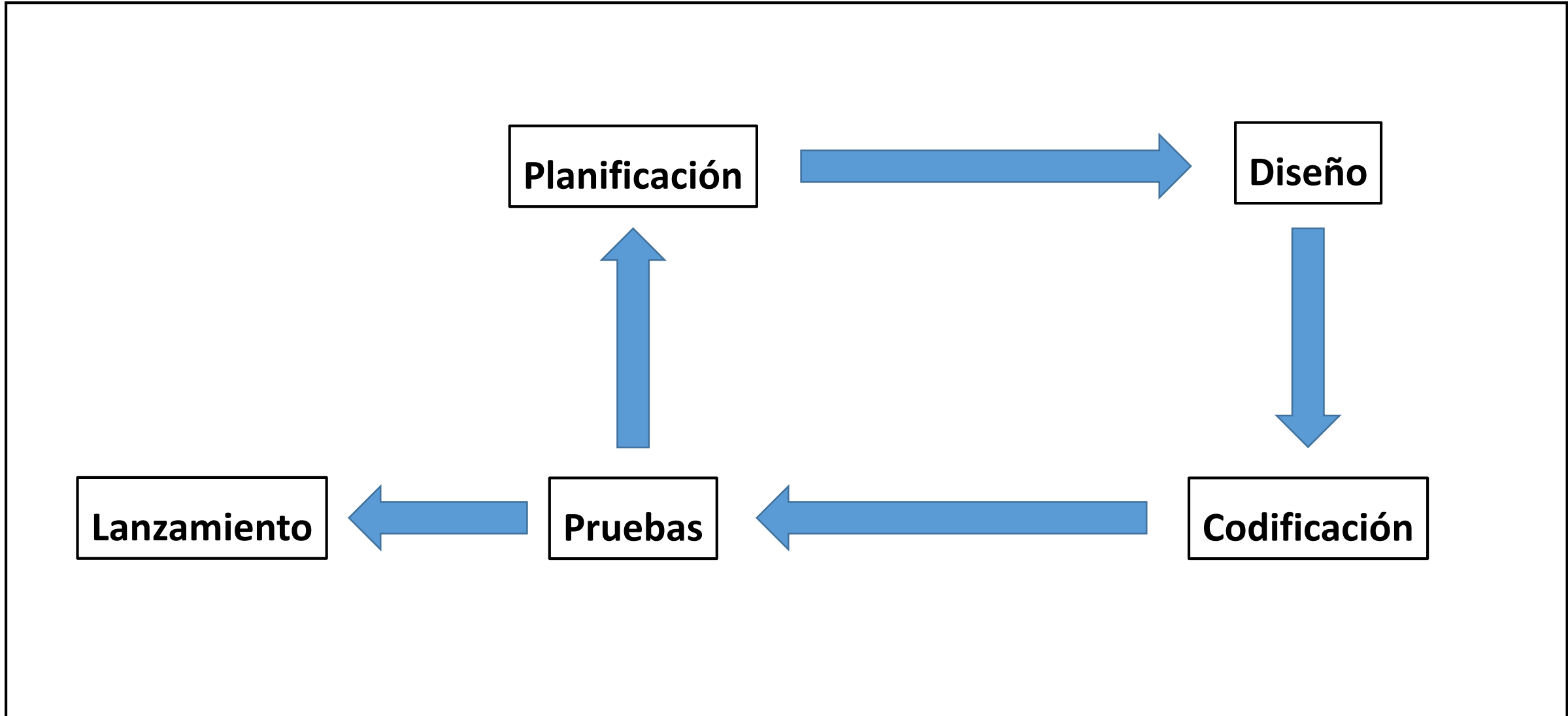
Hipótesis

Si se desarrolla un sistema de procesamiento de lenguaje natural para el reconocimiento de entidades nombradas, se podrá contribuir a disminuir las limitaciones actuales en el reconocimiento de las mismas en grandes volúmenes de datos en Elasticsearch. Este sistema buscará mejorar la eficiencia y efectividad en la identificación y clasificación de entidades, facilitando así el estudio lingüístico y la gestión de datos masivos para fines de análisis.

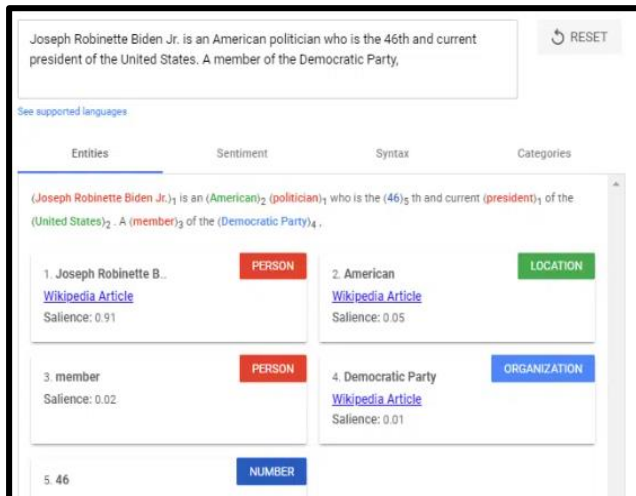
Aportes

- 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 en caso de surgir o no detectar una nueva entidad.
- Generación de datos de entrenamiento y testing de la entidad que no reconozca el modelo.
- Capacidad de retroceder en caso de el modelo perder conocimiento al realizar el reentrenamiento.

Metodología XP



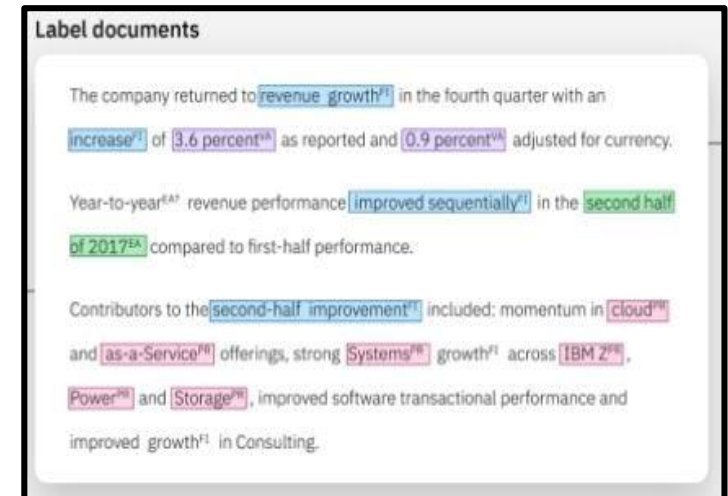
Sistemas con reconocimiento de entidades nombradas



Google NLP



ChatGPT



IBM Watson Discovery

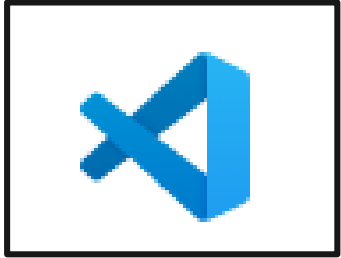
Modelos de reconocimiento de Entidades nombradas

MODELO	PESO	PRECISIÓN
Spacy	541 MB	90%
Bert	420 MB	89.86%
Flair	1.72 GB	86.65%
Stanford	635 MB	N/A

Modelos generadores de oraciones

Modelo	Llama2	GPT-3.5	GPT-4
Parametros	70 billones	154 – 175 billones	1 – 1.76 trillones
Modalidades	Texto solamente	Texto solamente	Texto e imagen
Precisión	68.9%	70%	86.4%
Complejidad	Baja	Alta	Alta
Velocidad	Rápido	Lento	Lento
Eficiencia	Más eficiente	Menos eficiente	Menos eficiente

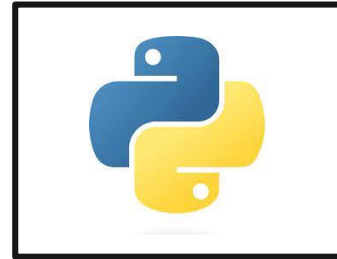
Herramientas y tecnologías de desarrollo



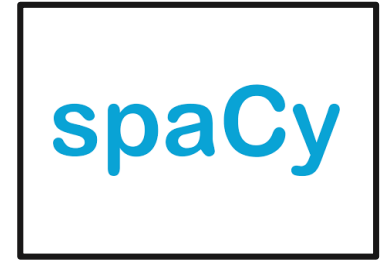
Visual Studio Code
(versión 1.82)



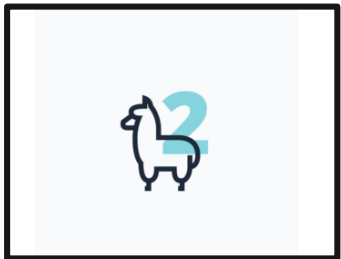
Elasticsearch
(versión 8.3.3)



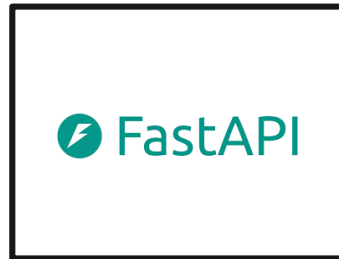
Python
(versión 3.10.8)



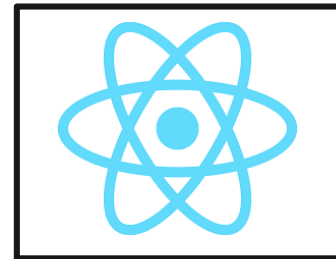
Spacy
(versión 3.6.1)



Llama2 7Billones



FastAPI v0.100.1



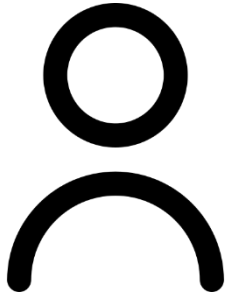
ReactJS v18



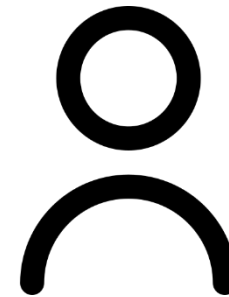
Material UI v5.15.15

Usuarios del Sistema y Responsabilidades

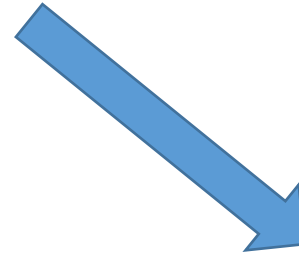
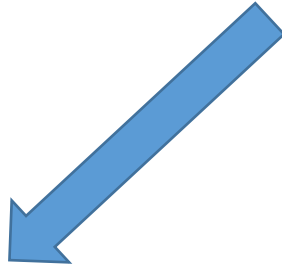
Usuario del Sistema



ADMINISTRADOR



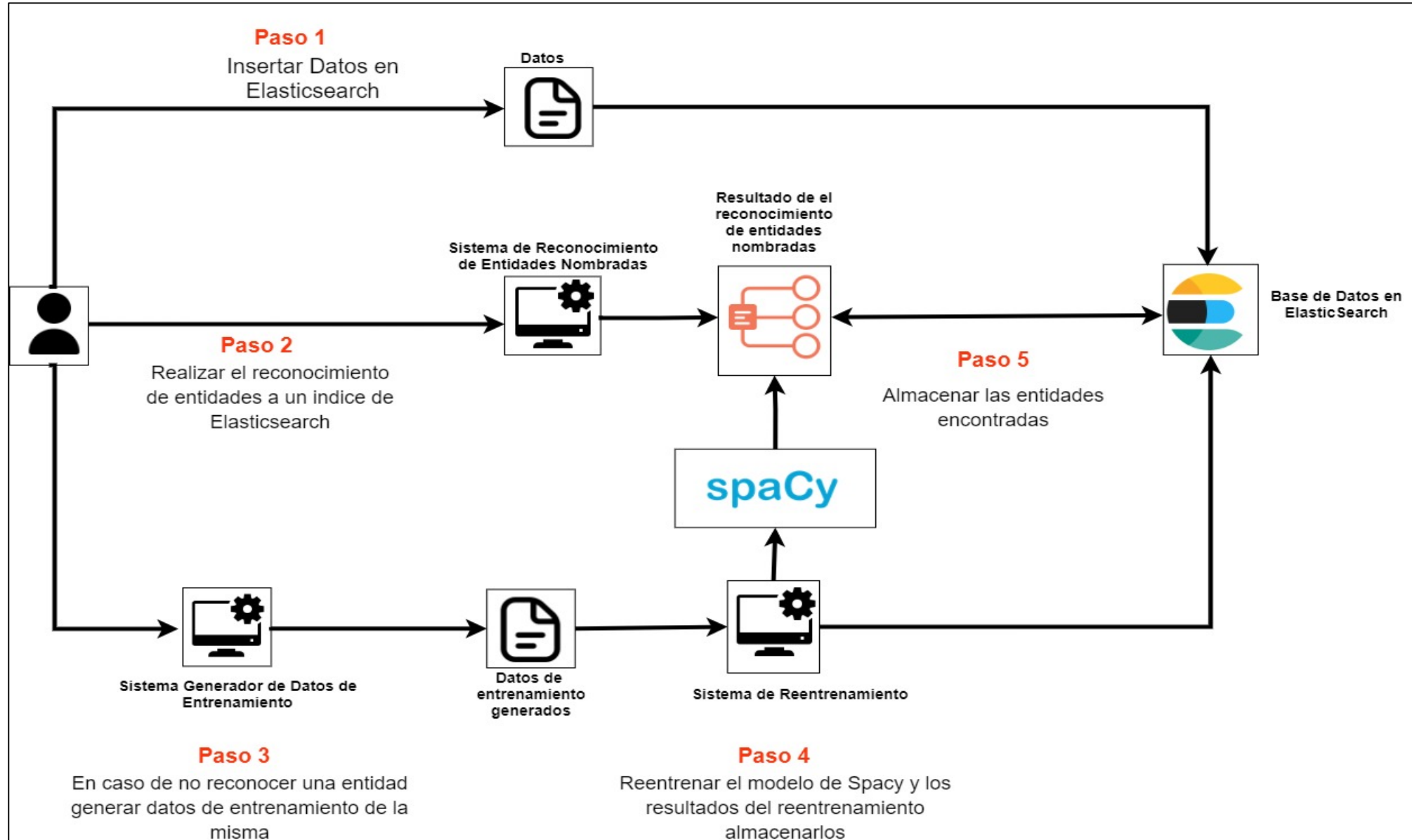
USUARIOS



Historias de Usuarios

- 1: Autenticar usuario en el sistema
- 2: Iniciar sistema con datos de prueba
- 3: Reconocer entidades nombradas en los índices de Elasticsearch
- 4: Reentrenar modelo
- 5: Administrador de usuarios
- 6: Soporte en español e inglés
- 7: Reentrenar por especialista

Diagrama del sistema



Elasticsearch base de datos no relacional

Índices de Elasticsearch

Usuario: id_usuario, nombre, contraseña, rol

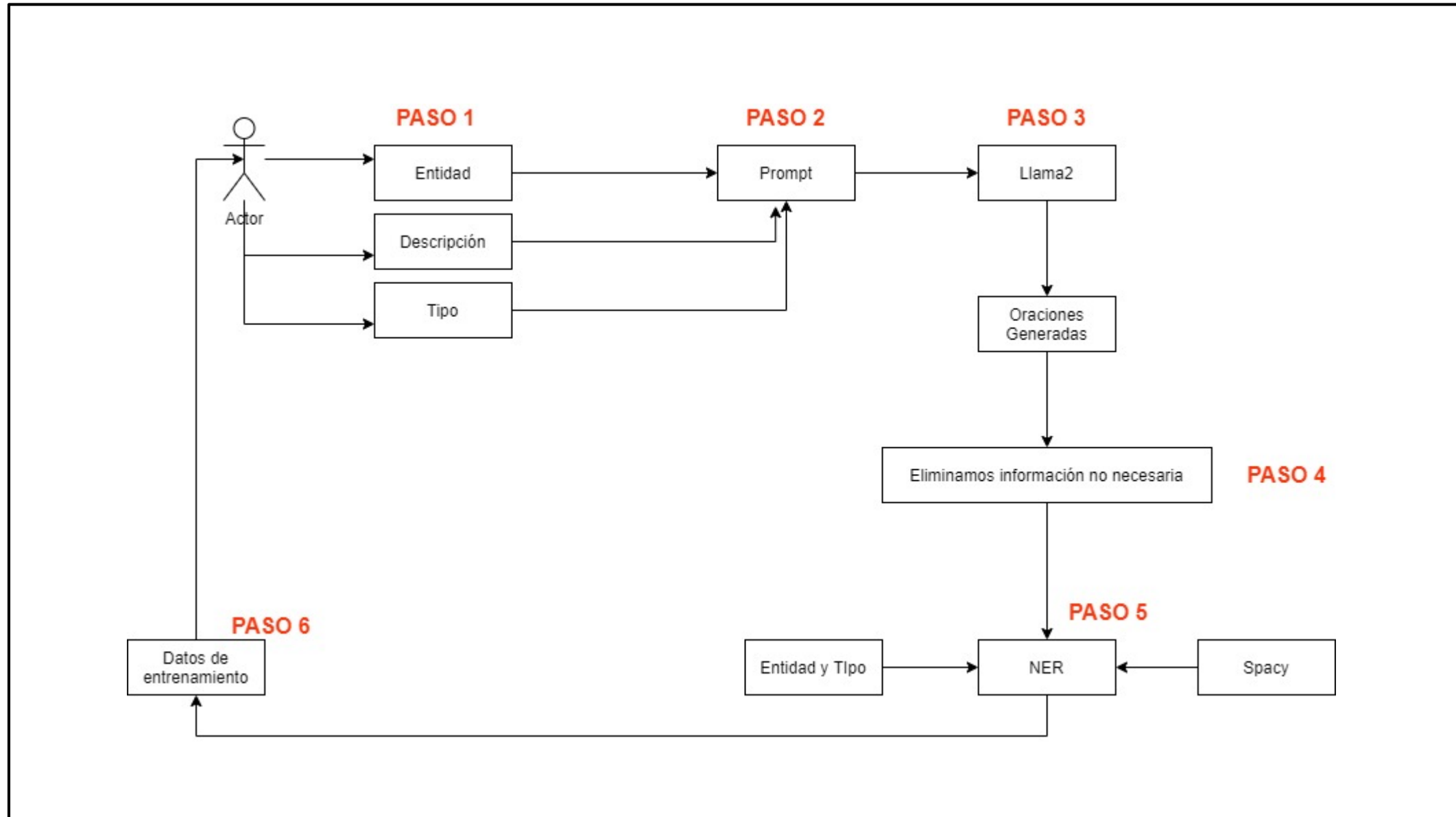
Datos: id_datos, texto, entidades

Modelo: id_modelo, precisión, exhaustividad,
puntuación F1, exactitud

Trazas: id_trazas, tema, fecha, id_usuario

- Estructura y Organización: Colección de documentos JSON
- Alto rendimiento y escalabilidad
- Open Source
- Maneja y procesa grandes volúmenes de datos
- Basado en la librería Lucene
- Índices Invertidos

Algoritmo generador de oraciones



Ejemplo de oración generada

```
{
  "text": "El equipo de pelota de Cuba, teamacere, está formado por un grupo de jugadores talentosos y  
entrenados por el experto técnico, Juan Carlos.",
  "entities": [
    {
      "name": "Cuba",
      "start": 23,
      "end": 27,
      "label": "LOC"
    },
    {
      "name": "teamacere",
      "start": 29, "end": 38,
      "label": "ORG"
    },
    {
      "name": "Juan Carlos",
      "start": 127,
      "end": 138,
      "label": "PERSON"
    }
  ]
}
```

Olvido Catastrófico

Reconocimiento de Entidades
Nombradas



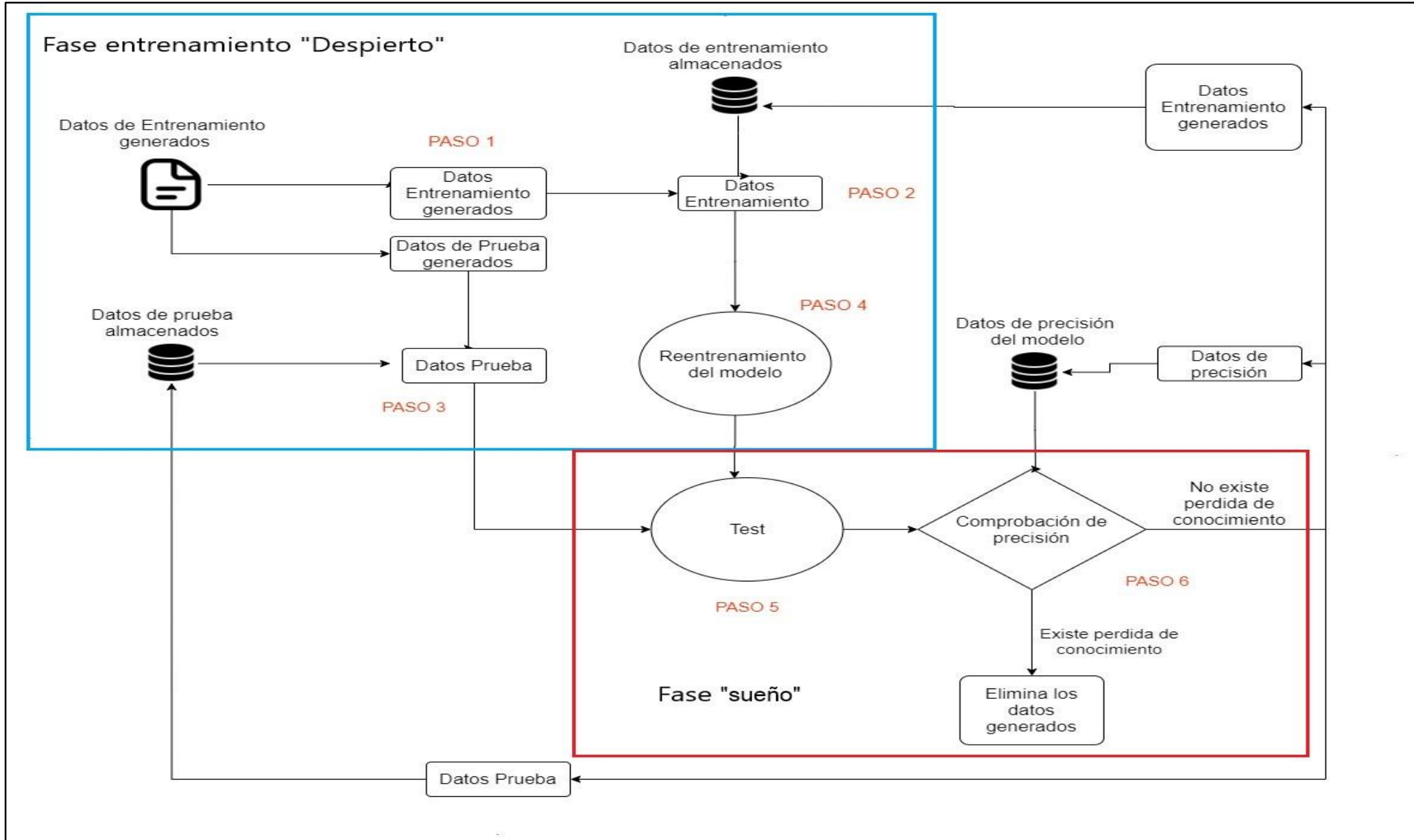
Pérdida de habilidades o conocimientos previamente adquiridos, por un modelo de IA cuando se le enseña nuevas tareas

Método de aprendizaje consolidado “despierto y sueño”

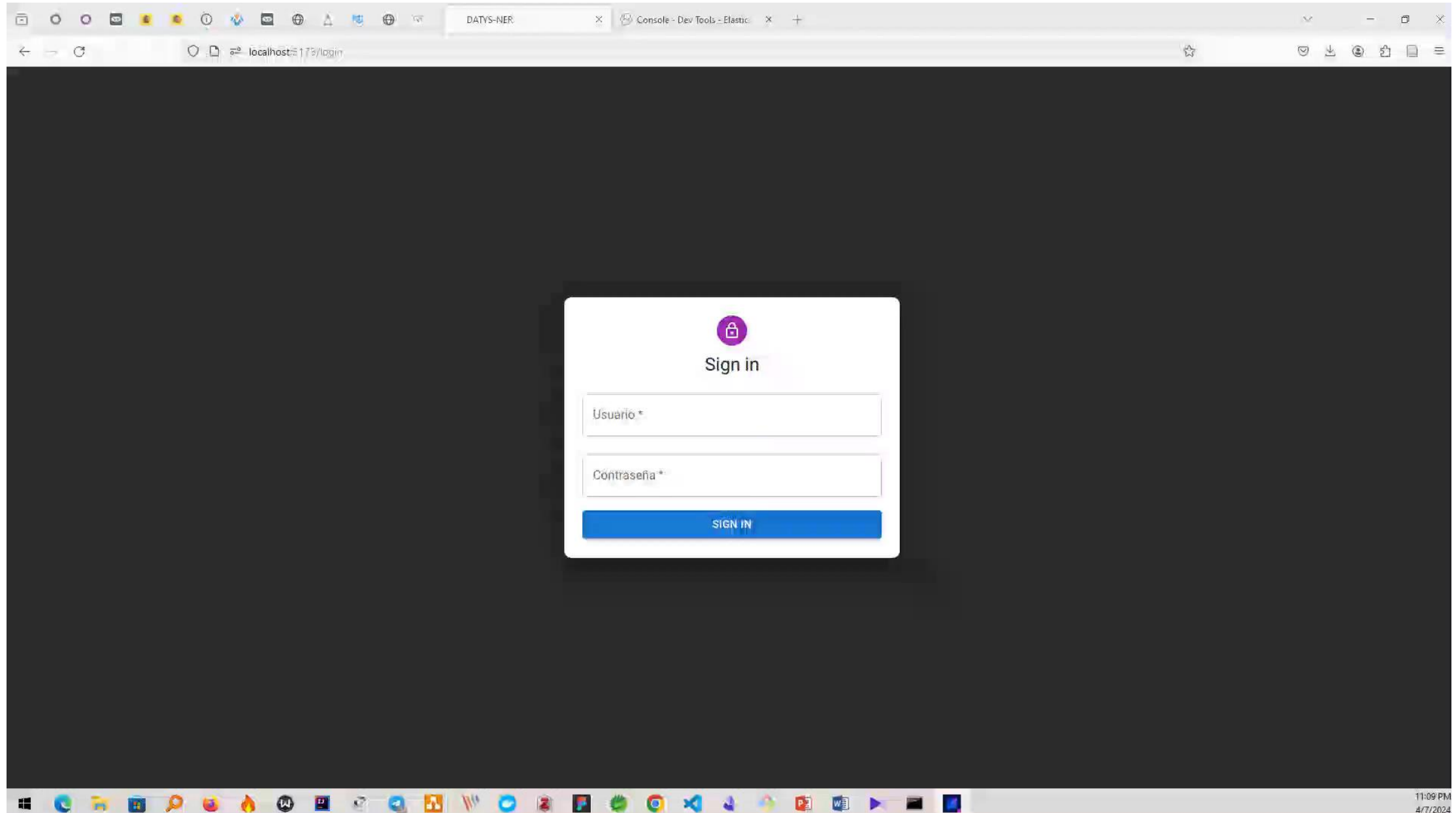
Despierto

Sueño

Algoritmo para evitar el olvido catastrófico



Funcionamiento del sistema



Conclusiones

- Se realizó un estudio del estado del arte de las técnicas de reconocimiento de entidades nombradas, las herramientas de Procesamiento de Lenguaje Natural (Spacy) y de las propuestas de grandes modelos de lenguajes y su uso en la generación de oraciones.
- Se analizó el modelo de indexación de información en Elasticsearch de la empresa DATYS.
- Se diseñó una propuesta para el reconocimiento de entidades en información indexada en Elasticsearch según los requerimientos de la empresa DATYS para el procesamiento de información.
- Se implementó el sistema diseñado para el reconocimiento de entidades nombradas y la generación de datos de entrenamiento.
- Se realizaron pruebas de las propuestas del modelo y sistema implementado.
- Se evaluó por la empresa Datys el sistema implementado.

Aval del Centro de Datys

22 de febrero de 2024

"Año del 66 aniversario de la Revolución"

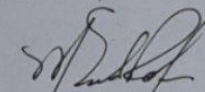
A quién pueda interesar:

Por medio de la presente, en calidad de Directora de DATYS Santiago de Cuba, perteneciente al MININT, certifico este aval a favor del estudiante de 4to año de Ingeniería Informática Luis Andrés Licea Berenguer. Luis ha desarrollado su tesis de pregrado en el tema "Herramienta digital para el reconocimiento de entidades nombradas en flujos de documentos en idioma español", el cual forma parte de un problema de investigación asociado al trabajo de la institución.

El estudiante ha demostrado una notable capacidad y dedicación en el desarrollo de su proyecto, el cual ha sido realizado con un alto nivel de competencia y rigor académico. Su trabajo ha contribuido a la investigación y desarrollo en el campo de la Minería de Datos y Textos, y su tesis representa un avance importante en el campo de la Inteligencia Artificial y el Procesamiento del Lenguaje Natural. Sus resultados serán analizados por la institución para su posterior uso en la misma.

Agradezco de antemano su atención y quedo a su disposición para cualquier información adicional que pueda requerir.

Atentamente,



Dra. Mónica Rubio Rojas

Directora DATYS Santiago de Cuba

Correo Electrónico: monica.rubio@datys.cu



Recomendaciones

- Integrar más funcionalidades al sistema como la posibilidad de tener varios modelos de procesamiento de lenguaje natural para realizar otras tareas como la traducción de texto, el análisis de sentimientos en las oraciones, similitud en textos.
- Integrar modelos de reconocimiento de entidades nombradas en otros idiomas como inglés, francés, portugués.

Sistema informático para el análisis de entidades nombradas en el procesamiento de documentos digitales en la empresa DATYS

Autor

Luis Andrés Licea Berenguer

Tutor

Dr.C. Dionis López Ramos

MSc. Jose Erinaldo Cruzata Ferrer

**Facultad de Ingeniería en Telecomunicaciones, Informática y Biomédica
Departamento de Ingeniería Informática
Proyecto de Investigación**

Santiago de Cuba, abril, 2024