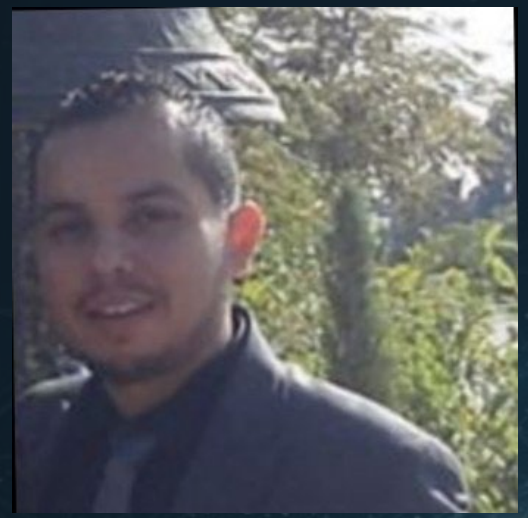
21 de Marzo 2025

Investigación

EXTRACCIÓN Y VALIDACION DE CONOCIMIENTO DE DOCUMENTOS DE TEXTOS EN UN AREA DE DOMINIO.

Alder López Cerda

ACTUALMENTE EN DOCTORADO EN CIENCIAS EN CIENCIA DE DATOS



8110503147
alder.lopz@gmail.com

Alder López con más de 20 años de experiencia en desarrollo, gestión y liderazgo tecnológico en la industria. Actualmente, en NEORIS, liderando iniciativas estratégicas en IA Generativa, Modelos LLM y Modelos de Difusión, enfocándose en soluciones avanzadas en la nube, aprendizaje diferencial y federado, seguridad en LLM y automatización del ciclo de desarrollo de software.

Como Maestro en Ciencia de Datos y doctorante en el Doctorado en Ciencias en Ciencia de Datos en INFOTEC, investigador sobre un enfoque de intersección entre Procesamiento de Lenguaje Natural (NLP), geometría no euclidiana y optimización de arquitecturas RAG , LLMs. La investigación doctoral se centra en técnicas como cuantización, destilación, LoRA, pruning y circuitos de características dispersas para mejorar el rendimiento en entornos con recursos limitados, así como en métodos con redes neuronales, interpretación causal y geometría hiperbólica para obtener métricas de coherencia textual, extracción de conocimiento y la generación de respuestas de mayor valor.

Ingeniero en Sistemas Computacionales, posterior Especialidad en Métodos Estadísticos por CIMAT, continuando con Maestría en Ciencia de Datos por la UANL y actualmente Doctorado en Ciencias en Ciencia de Datos en INFOTEC enfocado a NLP, LLM.

Problemas

La extracción y generación de conocimiento son fundamentales en dominios técnicos y especializados como el desarrollo de software, medicina, entre otros.

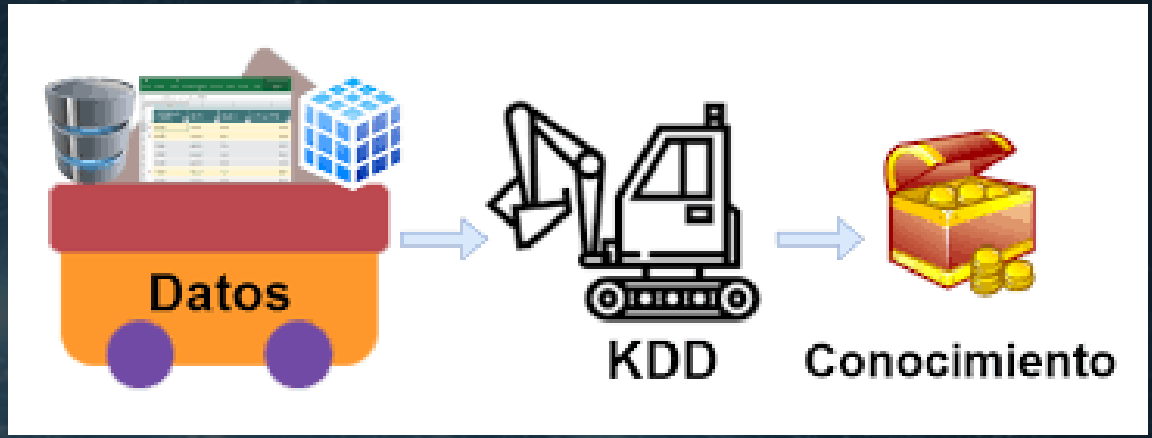
El Procesamiento de Lenguaje Natural (NLP) enfrenta desafíos clave que afectan su uso en el mundo real. Desde documentos con errores que confunden decisiones hasta chatbots que responden fuera de contexto. Los LLM como BERT o GPT son poderosos, pero a menudo son lentos, caros y poco precisos en tareas específicas. Mi investigación aborda estos problemas para hacer el NLP más útil y eficiente

Ejemplos :

- Incoherencias: Imaginen un manual médico con instrucciones contradictorias: 'dar 5 ml' en una página y 'dar 10 ml' en otra. Esto puede causar errores graves si no se detecta.
- Recuperación imprecisa: Un chatbot de soporte técnico busca en una base de datos enorme, pero trae un artículo irrelevante porque no entiende bien la pregunta. El usuario se frustra y pierde tiempo.
- Huecos en la especificación: El sistema de gestión de biblioteca permitirá a los usuarios gestionar libros y préstamos de forma eficiente. Debe ser fácil de usar y rápido

Líneas de investigación

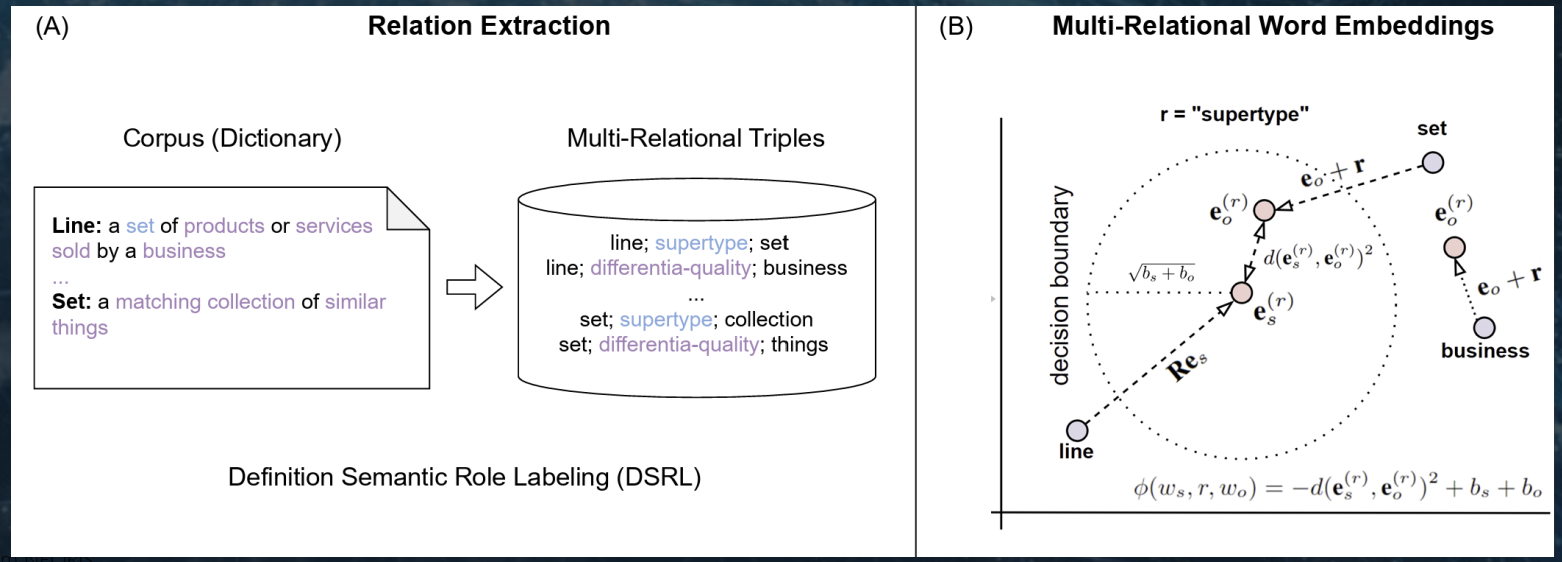
- Métricas de calidad de texto de entrada en un contexto.
- Extracción de conocimiento, identificación de entidad y sus relaciones basadas en RAG y otros métodos para entender , interpretar y reescribir contenido.
- Generación de contenido en base al contexto de entrada, conocimiento entidades, dominio de industria y sus relaciones.
- Fine-Tunning, LoRA de un modelo para especializarlo a un dominio en particular.



KDD (Knowledge Discovery in Databases)

Enfoques de exploración

- **Redes neuronales y modelos probabilísticos** para encontrar errores semánticos y cohesión en los textos.
- Uso de **geometría hiperbólica** para organizar datos y responder mejor. Usando embeddings hiperbólicos para capturar relaciones semánticas jerárquicas y detectar lagunas en la estructura lógica
- **Ecuaciones estructurales** para **coherencia** textual, para modelar relaciones causales entre componentes textuales o entre variables derivadas de textos.
- **Modelo de difusión puede "rellenar" huecos al intentar reconstruir un texto completo a partir de uno parcial o ruidoso. Los puntos donde el modelo introduce información significativa (no presente en el original) señalan huecos.**



Aplicación al Desarrollo de Software

Las especificaciones técnicas son fundamentales en el desarrollo de software, pero su generación manual es costosa y propensa a errores. Esta investigación propone automatizar la identificación de documentos, estructurar conocimiento técnico, generar contenido alineado con estándares y validar su calidad con métricas objetivas.

Importancia del dominio:

- Las especificaciones técnicas son esenciales para definir requisitos, dependencias y estándares en proyectos de software.
- Actualmente, estas especificaciones se generan manualmente, lo que resulta en procesos costosos y propensos a errores.

Contribuciones de la investigación:

- Automatización de la identificación de documentos relevantes para especificaciones.
- Estructuración de conocimiento técnico (entidades, relaciones) para proyectos complejos.
- Generación de contenido técnico alineado con estándares del desarrollo de software.
- Validación de la calidad del contenido generado con métricas objetivas de acuerdo al dominio.

Resultados esperados:

- Reducción de tiempo y esfuerzo en la creación de especificaciones técnicas.
- Incremento en la calidad y precisión de los documentos generados.
- Mejora en la adaptabilidad y eficiencia de modelos de lenguaje para tareas específicas.