TRABAJO DE IA SOBRE VÍCTIMAS DE DESAPARICIÓN FORZADA

1. En febrero de 2023 se celebró un convenio de colaboración entre la Subsecretaría de DDHH (SDH) y el Instituto Milenio Fundamento de los Datos (IMFD).
2. El objetivo de ese convenio era explorar formas en que la ciencia de datos podría colaborar con la búsqueda de personas víctimas de desaparición forzada.
3. Durante el primer periodo de la colaboración, se exploraron dos líneas de trabajo. La primera era el modelo de grafos inspirado en la plataforma Ángelus de la Comisión Nacional de Búsqueda (CNB) de México, que esta puso a disposición del Programa de Derechos Humanos; y la segunda era utilizar modelos de procesamiento de lenguaje natural para analizar los datos.
4. De parte de IMFD participaron de reuniones y desarrollaron tareas concretas dentro del marco del convenio, el entonces director del instituto, Marcelo Arenas, y un estudiante de postdoctorado francés. También apoyó en algunas tareas un pasante de pregrado.
5. Durante este periodo no hubo una planificación clara de una hoja de ruta de trabajo, productos esperados ni recursos a invertir. Se identificó desde ambas partes la dificultad que presentaba el convenio para disponer de recursos mínimos para ejecutar las tareas que nos proponíamos.
6. Por ejemplo, IMFD no contaba con la infraestructura (CPU, Unidades de procesamiento central; y GPU, unidades de procesamiento de gráficos) necesarios para el desarrollo de los modelos que se exploraban ni el procesamiento masivo de datos. Por esta razón, se solicitó a la SDH proveer dicha infraestructura, para lo cual se trató de contratar el Servicio por medio de convenio marco, sin éxito, estancando el avance del proyecto.
7. Otro ejemplo es que no había capacidad humana para hacer el trabajo de descifrar, traducir y comprender las tablas de datos del sistema de derechos humanos (plataforma interna del programa) para su utilización como base. Esto se trató de subsanar solicitando apoyo a la empresa que brinda soporte informático al sistema, pero lo que se requería era un contingente de cientistas de datos suficientes para desentrañar dicho sistema porque la base de datos del sistema originario no está bien estructurada y carece de documentación mínima para comprenderla. Se solicitó de parte del IMFD poder contratar estudiantes que pudieran colaborar en esta tarea, pero las restricciones presupuestarias del Programa de Derechos Humanos no permitieron contratar dicho apoyo.

8. Estos obstáculos financieros y la falta de personal no permitieron que se avanzara en productos concretos en la línea del objetivo del convenio marco. Así, llegando al último trimestre del año 2023, los únicos productos que se habían desarrollado fueron dos:

- La traducción de la base de datos del sistema de derechos humanos a una base de datos de grafos visualizables. Sin embargo, las relaciones eran la mayoría de las veces espurias, lejanas o incomprensible por las razones ya vertidas.
- El uso de algunos modelos de LLM[1] sobre la base de datos del sistema de DDHH y otras bases de datos de trayectorias elaboradas por el programa. Cabe señalar que nunca se hizo entrega formal de ningún producto en esta línea.

9. Desde el Programa no se tiene una mala evaluación de la colaboración del IMFD, todo lo contrario. Se les considera un aliado estratégico para el Plan Nacional de Búsqueda, pero el diagnóstico es claro: el convenio y la forma en que se estaba ejecutado no permitía llegar a productos concretos al ritmo que se requería.
10. Cuando se licitó el diseño del anteproyecto a fines de 2023, al cual IMFD no postuló, de parte del Programa se intentó que IMFD pudiera seguir acompañando el proceso, brindando su apoyo técnico y orientación para la definición de lo que debía ser la plataforma. En ese contexto, se sostuvieron reuniones entre la empresa adjudicataria, Unholster, y el IMFD, representado por su entonces director, Marcelo Arenas.
11. Sin embargo, en una reunión posterior, de parte del IMFD se planteó al Programa que como Instituto no podían brindarle asesoría a una empresa privada y que, si bien estaban 100% disponibles para seguir colaborando con la subsecretaría, no veían posible esa forma de trabajo, salvo que se pagara por su asesoría. Sobre esto no hubo acuerdo,
12. Finalmente, en el mes de enero de 2024 hubo reunión del PDH con el nuevo director del IMFD. Se conversó sobre nuevas formas de colaboración, específicamente sobre la forma en que serán analizados los datos una vez estos sean procesados por la plataforma que se había contratado con Unholster, señalando el IMFD que eso era lo que más les interesaba. Por su parte, desde el IMFD plantearon la posibilidad de postular a fondos ANID destinados al desarrollo de soluciones informáticas para problemas públicos complejos, para lo cual se requería el patrocinio de la Subsecretaría. En la reunión se planteó interés en explorar esa línea de colaboración, pero desde el IMFD señalaron que ya era muy tarde para la convocatoria 2024. (Sin

[1] Un modelo lingüístico grande (LLM) es un tipo de programa de inteligencia artificial (IA) que puede reconocer y generar texto, entre otras tareas.

perjuicio de ello, el Ministerio generalmente brinda patrocinio a los organismos académicos que lo solicitan en materias del conocimiento de este).

Participación de Unholster y trabajo en ejecución:

Efectivamente, durante el primer semestre de 2023, hubo una reunión con Unholster, en la que participó el Ministro Cordero (que no se registró como lobby, por cuanto no fue requerida para vender un producto en específico, sino para dar a conocer un trabajo que estaba realizando sobre el Informe Rettig, con motivo de los 50 años del Golpe Militar).

En atención a que los objetivos del Plan Nacional de Búsqueda era de una entidad superior a lo que estaba realizando el Instituto Milenio e incluso a lo que había expuesto Unholster respecto de 50 años, se resolvió llamar a licitación, a la que se presentaron varias empresas. Finalmente, se resolvió adjudicar a Unholster.

El trabajo de Unholster partió por sentar las bases: diseñar un sistema de procesamiento de capaz de extraer de la mejor forma la mayor cantidad de datos a partir de documentos digitalizados, y sobre esa base, montar modelos capaces de generar nuevo conocimiento, identificando nuevas relaciones, visualizando información de diferentes formas, y organizando el trabajo y la información orientada para la búsqueda.

Para eso, han tomado todo el repositorio digital del PDH, se han identificado todos los bloques de textos, cada bloque se ha descompuesto en secciones cortas, estas secciones cortas han sido indexadas y vinculadas geográficamente al documento, y luego procesadas por una cascada de modelos de OCR[2], generando secciones de texto transcrito.

Sobre estas secciones de texto, se han aplicado modelos de reconocimiento de entidades. Sobre las entidades identificadas, luego se aplica otro modelo capaz de identificar si corresponden a entidades conocidas. Para esto, se incorporó dentro del sistema todos los datos del antigua sistema de DDHH, utilizando las entidades ahí construidas como punto de partida.

Finalmente, sobre la base de todo lo anterior, se están aplicando modelos de LLM para el reconocimiento de eventos (por ejemplo, a X le pasó Z en un lugar Y a manso de M) que vinculen a las entidades del sistema, organizando esta información en bases de datos de grafos.

Todos estos desarrollos se traducen en interfaces de búsquedas de contenidos libres, de entidades, visualizaciones de grafos, geoespaciales, entre otras.

[2] Reconocimiento óptico de caracteres para convertir textos de documentos históricos en formato digital. (escaneo + algoritmo identifica y digitaliza para transformar en formato editable).

Finalmente, todo esto se traduce en los siguientes productos:

a) Entregados:

- Procesamiento de archivos para extracción de datos: sistema capaz de extraer de la mejor forma la mayor cantidad de datos a partir de documentos digitalizados.

- Migración de datos Sistema DDHH: se extrajeron, y procesaron todos los datos del sistema de DDHH, dándole estructura y sentido dentro de la nueva plataforma.

- Buscador de contenidos: herramienta de búsqueda de contenidos libres, que permite encontrar coincidencias tanto en los archivos procesados como en los datos provenientes del sistema DDHH.

- Repositorio de entidades: se extrajeron todas las entidades presentes en los contenidos, identificando cuales coinciden con las entidades provenientes del sistema DDHH, organizándola en repositorio.

- Repositorio digital: el repositorio digital se encuentra en la plataforma, con la descripción archivística recuperada de los sistemas del PDH.

b) Próximos productos:

- Identificación de eventos: modelo de LLM para la identificación de eventos en los bloques de textos, generando relaciones entre entidades.

- Visualización de grafos: estructuración de bases de datos de grafos, con visualización de las relaciones.

- Visualización georreferenciada: presentación georreferenciada de los datos del sistema, particularmente de los eventos.

- Módulo de investigación: organización de los distintos desarrollos para facilitar el trabajo de investigadores. Esto permitirá organizar el contenido en torno a la formulación de hipótesis interactuando con las entidades, relaciones y eventos.