Inteligencia artificial como apoyo al análisis de redes criminales

Luis Gabriel Martínez Abello

Asesor

Andres Felipe Hernandez Giraldo

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analitíca

2026

Resumen

La Fiscalía General de la Nación cuenta con la Sección de Análisis Criminal (SAC), el Centro Estratégico de Análisis Criminal (CEAC) y otros grupos dedicados a realizar análisis criminal a nivel nacional, adscritos al Cuerpo Técnico de Investigación. Estos grupos realizan análisis operativos y/o estratégicos de casos discriminados por delitos o temáticas. Para realizar estos análisis, se deben seguir ciertos pasos que están relacionados con el ciclo de la información: recolección, evaluación, tratamiento, cotejo, análisis y difusión.

El analista para cumplir con estos pasos, al momento del tratamiento y del cotejo de la información debe leer los expedientes (digitales o en físico), extraer la información relevante en una matriz (generalmente Excel) en la que se debe incluir la información de las entidades (personas naturales o jurídicas, contratos, contratista, contratante) y el vínculo entre esas entidades y, por último, utilizando un software para análisis (i2 Analyst´s Notebook) procesar la matriz para gráficamente visualizar las relaciones e identificar redes o asociaciones criminales.

Estos pasos de tratamiento y cotejo demandan mucho tiempo para el analista, tiempo que se debería usar para el análisis; en algunas ocasiones la extracción de la información y posterior construcción de la matriz puede ser un proceso de hasta seis (6) meses.

Otro aspecto, es que el análisis desarrollado por cada analista se utiliza para esclarecer un caso en particular, y en muy pocas ocasiones esos análisis se unen o se relacionan con otros informes de análisis, dentro o fuera de cada grupo.

El desarrollo del presente proyecto está enfocado en ayudar a esclarecer delitos contra la corrupción. Con apoyo del sistema, se pretende de manera ágil y eficiente determinar redes de personas naturales y jurídicas, y como se unen con la contratación estatal; sin importar la territorialidad y la escala de tiempo.

Teniendo en cuenta que los datos son No estructurados, se emplearán técnicas de Procesamiento de Lenguaje Natural para la extracción de la información y la identificación de entidades (personas, contratos, contratantes, contratistas).

Palabras clave**:** corrupción, análisis, criminal, redes.

Abstract

The Fiscalía General de la Nación has the Sección de Análisis Criminal (SAC), the Centro Estratégico de Análisis Criminal (CEAC), and other specialized units dedicated to criminal analysis at the national level, all attached to the Cuerpo Técnico de Investigación. These groups carry out operational and/or strategic analyses of cases categorized by crimes or themes. To perform these analyses, certain steps must be followed, which are related to the information cycle: collection, evaluation, processing, comparison, analysis, and dissemination.

In order to follow these steps, during the processing and comparison of information, the analyst must read the case files (digital or physical), extract the relevant information into a matrix (usually an Excel spreadsheet), which should include information on the entities involved (individuals or legal entities, contracts, contractors, contracting parties), and the links between those entities. Finally, using analysis software (i2 Analyst's Notebook), the matrix is processed to visually explore relationships and identify criminal networks.

These processing and comparison steps are time-consuming for the analyst—time that should ideally be used for analysis. In some cases, the information extraction and matrix construction process can take up to six (6) months.

Another issue is that the analysis conducted by each analyst is generally used to clarify a specific case, and very rarely are these analyses connected or related to other analytical reports, whether within the same group or externally.

The development of this project is focused on helping to clarify crimes related to corruption. With the support of the system, the aim is to efficiently and swiftly identify networks of individuals and legal entities and how they are connected to government contracting, regardless of territorial boundaries or time scale.

Considering that the data is unstructured, Natural Language Processing (NLP) techniques will be employed to extract information and identify entities (individuals, contracts, contracting parties, contractors).

Keywords**:** corruption, analysis, criminal, networks

Tabla de Contenido

Lista de Figuras

Introducción

En Colombia, la corrupción representa uno de los principales desafíos estructurales que afectan la confianza ciudadana, la eficacia institucional y el desarrollo económico. A pesar de los esfuerzos de entidades como la Fiscalía General de la Nación, los procesos investigativos en casos de corrupción siguen enfrentando limitaciones operativas, entre ellas, el tiempo y los recursos requeridos para procesar grandes volúmenes de información no estructurada contenida en expedientes judiciales.

La Sección de Análisis Criminal (SAC) y el Centro Estratégico de Análisis Criminal (CEAC) desempeñan un papel fundamental en la identificación y estudio de redes delictivas, aplicando metodologías que, aunque efectivas, dependen en gran medida del trabajo manual para la lectura, extracción, cotejo y análisis de información. Este enfoque, además de ser lento, dificulta la integración de datos entre casos, lo que limita la posibilidad de descubrir patrones complejos y conexiones relevantes entre actores involucrados en actos de corrupción.

Frente a este panorama, la presente propuesta de trabajo de grado plantea el desarrollo de un sistema basado en técnicas de Inteligencia Artificial (IA), Aprendizaje Automático (Machine Learning) y Procesamiento de Lenguaje Natural (NLP), que permita automatizar el análisis de expedientes judiciales digitales. El objetivo es identificar, de manera eficiente, relaciones entre personas naturales y jurídicas, contratos estatales y dinámicas delictivas asociadas a casos de corrupción.

Este enfoque permitirá no solo optimizar los tiempos de procesamiento y análisis, sino también fortalecer la capacidad investigativa de la Fiscalía, aportando a la lucha contra la corrupción desde una perspectiva tecnológica, analítica y estratégica, alineada con los principios de justicia, transparencia y modernización del Estado.

Descripción del Problema

Planteamiento del Problema

La corrupción es una manifestación social y económica porque se presenta en las relaciones humanas, y a su vez es promovida a favor de los intereses de dos o más particulares, donde por lo menos uno de estos es funcionario público (Zuleta, 2015). Por otro lado, según la Organización de las Naciones Unidas (ONU), la corrupción interviene directamente en el crecimiento económico, violando principios como la transparencia y la legitimidad, y entorpeciendo las administraciones.

En Colombia, la corrupción se ha manifestado de diferentes maneras con la compra de funcionarios, apropiación de bienes y del gasto público, favorecimiento en contratación, etc.

Con referencia al sistema de contratación pública, el Estado Colombiano ha establecido esquemas y requisitos para el acceso a las licitaciones públicas, siendo una de ellas el SECOP (Sistema Electrónico de Contratación Pública), que es el medio oficial de información del Estado sobre las contrataciones realizadas con fondos públicos. Permite realizar todo el proceso de contratación en línea.

De acuerdo con el ranking del Índice de Percepción de la Corrupción de Transparencia Internacional, para el año 2023 Colombia ocupa el puesto 81 entre 180 países evaluados. La corrupción en Colombia ha llevado a que se pierda la confianza en las instituciones y en quienes hacen parte de ellas.

Algunos de los casos más sonados en los últimos años han sido:

- Escándalo de Invercolsa (año 2004)
- Parapolítica (año 2006)
- Yidispolítica (año 2008)

- Farcpolitícia (año 2008)
- Agro Ingreso Seguro (año 2009)
- ChuzaDAS (año 2009)
- Carrusel de la contratación Bogotá (año 2010)
- Escándalo de Interbolsa (año 2012)
- Carrusel de la contratación Bucaramanga (año 2014)
- Caso Andrómeda y hacker (año 2014)
- Cartel de la hemofilia (año 2016)
- Odebrecht (año 2017)
- Caso Hyundai - Carlos Mattos (Año 2018)
- Ñeñepolítica (año 2020)
- Caso centros poblados (año 2021)
- Escándalo de las Marionetas (año 2022)
- Caso Nicolas Petro (año 2023)

Estos son algunos de los casos con mayor connotación nacional, algunos hasta de carácter internacional como es el Caso Odebrecht, pero vale la pena aclarar que son muchísimos más casos de corrupción a gran o menor escala los que azotan a Colombia y que, en el proceso penal se encuentran estancados.

Solo por mencionar un ejemplo, en el caso Odebrecht se pagaron coimas por un valor cercano a los 80 mil millones de pesos colombianos.

La Fiscalía General de la Nación desempeña un papel crucial en la lucha contra el crimen en Colombia, apoyándose en grupos especializados como la Sección de Análisis Criminal (SAC) y el Centro Estratégico de Análisis Criminal (CEAC). Estos grupos realizan labores esenciales

para esclarecer delitos y analizar dinámicas delictivas, basándose en un ciclo de información que incluye la recolección, evaluación, tratamiento, análisis y difusión de datos.

La Dirección Especializada contra la Corrupción, solo logró éxito en el 2% de los procesos que adelanto entre 2014 y 2021. Los principales obstáculos son:

- Falta de voluntad política al interior de la entidad y porque los casos implican riesgos y demandan mucho tiempo.
- La politización de la entidad: ocasionada por la forma de selección y la provisionalidad de algunos funcionarios.
- Falta de ambición, es habitual que los esfuerzos de funcionarios en la entidad se encaminen a superar un hecho puntual y no el contexto que los abarca; es decir, una red de corrupción más amplia.

Sin embargo, el proceso actual presenta varios desafíos significativos. Los pasos de tratamiento y cotejo de información, que incluyen la lectura de expedientes digitales o físicos, la extracción de datos relevantes y la construcción manual de matrices en herramientas como Excel, son altamente demandantes en tiempo y esfuerzo. Estos procesos pueden extenderse hasta seis meses, reduciendo drásticamente el tiempo disponible para realizar análisis estratégicos y operativos.

Además, los análisis suelen limitarse a casos individuales, con escasa integración entre informes desarrollados por diferentes analistas o grupos. Esta fragmentación limita la capacidad de identificar patrones y conexiones entre personas, grupos delictivos y casos judiciales, lo que es esencial para abordar fenómenos criminales complejos y transversales.

Ante este panorama, surge la necesidad de responder a la pregunta: ¿Cómo puede un sistema basado en Inteligencia Artificial y Machine Learning analizar expedientes digitales de la

Fiscalía General de Colombia para identificar conexiones entre personas, grupos delictivos y casos judiciales, facilitando el apoyo a investigaciones criminales?

Justificación

La corrupción, como fenómeno transversal, ha permeado múltiples niveles del sector público colombiano, debilitando la confianza ciudadana y afectando el desarrollo institucional. A pesar de los esfuerzos de la Fiscalía General de la Nación, la capacidad operativa para esclarecer este tipo de delitos se ve limitada por los tiempos extensos requeridos para procesar expedientes judiciales, muchos de ellos en formatos no estructurados y con volúmenes significativos de información.

Actualmente, los analistas deben invertir meses en tareas repetitivas de lectura, extracción manual y organización de datos, lo cual reduce su capacidad de concentrarse en el análisis profundo de las redes delictivas y la identificación de patrones de comportamiento. Además, los análisis se desarrollan de forma aislada, sin una articulación efectiva entre casos, lo que impide descubrir vínculos clave entre actores recurrentes o contratos sospechosos.

Ante este panorama, el desarrollo de un sistema automatizado basado en Inteligencia Artificial (IA), Machine Learning y Procesamiento de Lenguaje Natural (NLP) se justifica plenamente, al ofrecer una alternativa tecnológica capaz de transformar el modelo actual de análisis criminal. Esta herramienta permitirá extraer información relevante de manera rápida y precisa, estructurarla en grafos de relaciones y visualizarla en formatos comprensibles para los investigadores.

El impacto del proyecto se proyecta en varios niveles:

- Institucional, al fortalecer la eficiencia y efectividad de los procesos investigativos;
- Tecnológico, al incorporar soluciones modernas en el tratamiento de datos no estructurados;

- Social, al mejorar la capacidad de respuesta del Estado frente a la corrupción y restaurar la confianza ciudadana;
- Legal, al generar insumos procesables que puedan ser utilizados como apoyo en juicios orales.

En suma, el sistema propuesto no solo optimiza recursos y tiempos, sino que contribuye a una lucha más articulada y estratégica contra la corrupción en Colombia, con un enfoque preventivo, reactivo y sostenible.

Objetivos

Objetivo General

Desarrollar un sistema basado en Inteligencia Artificial y Machine Learning que permita implementar procesos de análisis más eficientes de expedientes digitales al interior de la Fiscalía General de la Nación, para identificar conexiones entre personas naturales y jurídicas, y casos enfocados a delitos relacionados con corrupción.

Objetivos Específicos

Diseñar un sistema de extracción automatizada de documentos relevantes contenidos en los expedientes digitales de la Fiscalía General de la Nación, como Contratos, Certificados de Matrícula y Certificados de Existencia y Representación Legal.

Establecer vínculos entre entidades clave dentro del texto digitalizado, tales como personas naturales, personas jurídicas, contratos, contratante, contratista con el fin de representar sus relaciones en el contexto de los expedientes judiciales.

Desarrollar un sistema de visualización que represente de forma gráfica o matricial los vínculos identificados entre entidades, permitiendo analizar las conexiones en los expedientes judiciales relacionados con casos de corrupción.

Marco de Referencia

Estado del Arte

En los últimos años, la aplicación de la Inteligencia Artificial (IA) en la lucha contra la criminalidad ha evolucionado significativamente, pasando de modelos analíticos descriptivos a sistemas predictivos y prescriptivos que permiten analizar grandes volúmenes de datos no estructurados. La investigación científica y tecnológica ha evidenciado que el Procesamiento de Lenguaje Natural (PLN), combinado con aprendizaje automático, puede extraer conocimiento útil desde documentos judiciales, transcripciones, contratos y fuentes abiertas.

Estudios como el de Caldwell et al. (2020) destacan cómo la IA, especialmente en áreas como la detección de relaciones y patrones en redes delictivas, puede generar ventajas operativas significativas para las agencias judiciales. Por su parte, Luna (2022) explora el uso de análisis de redes complejas como una metodología eficaz para mapear y entender la dinámica de grupos delictivos, incluidas las estructuras de corrupción administrativa.

La literatura especializada también ha abordado la automatización del análisis documental mediante técnicas como NER (Reconocimiento de Entidades Nombradas), extracción de relaciones (Relation Extraction), y modelos de lenguaje como BERT, RoBERTa o GPT, los cuales se utilizan para tareas como clasificación de expedientes, resumen automático y detección de similitud semántica entre casos.

En cuanto al análisis visual de redes, herramientas como i2 Analyst’s Notebook, Gephi y bibliotecas en Python como NetworkX y Graph-tool, han permitido representar gráficamente relaciones entre personas, entidades y hechos delictivos, facilitando la interpretación de fenómenos complejos.

En América Latina, algunos esfuerzos incipientes se han orientado al análisis de datos contractuales para prevenir corrupción. El trabajo de Odilla (2023) sobre el uso de "bots contra la corrupción" en Brasil, por ejemplo, identifica beneficios y limitaciones del uso de IA en tiempo real para monitorear licitaciones. En Colombia, sin embargo, las iniciativas institucionales aún se encuentran en fases exploratorias o experimentales, lo que evidencia la pertinencia e innovación del proyecto propuesto.

En resumen, el estado del arte demuestra que la IA aplicada al análisis criminal no solo es viable, sino altamente prometedora. Sin embargo, también plantea retos éticos, técnicos y legales que deben abordarse para garantizar su implementación efectiva en contextos judiciales.

Marco Contextual

Colombia enfrenta una crisis persistente de corrupción, que ha minado la credibilidad institucional y obstaculizado el desarrollo económico y social. Según el Índice de Percepción de la Corrupción de Transparencia Internacional, Colombia ocupó el puesto 81 de 180 países en 2023, con una calificación que evidencia altos niveles de desconfianza hacia el manejo público de recursos.

Casos emblemáticos como Odebrecht, el carrusel de la contratación en Bogotá, o el escándalo de Centros Poblados, muestran estructuras de corrupción que funcionan como redes criminales organizadas, donde convergen intereses políticos, empresariales y estatales. Estas redes no operan de forma aislada, sino mediante actores repetitivos, contratos amañados, empresas fachada y funcionarios cooptados, lo que exige herramientas avanzadas para su análisis.

La Fiscalía General de la Nación, a través de dependencias como la SAC (Sección de Análisis Criminal) y el CEAC (Centro Estratégico de Análisis Criminal), realiza esfuerzos

constantes por desentrañar estas redes. Sin embargo, los procesos siguen dependiendo en gran medida de la revisión manual de documentos judiciales —en muchos casos digitalizados pero no estructurados— lo que genera cuellos de botella en la fase de tratamiento y análisis.

Actualmente, el procesamiento de expedientes digitales implica la lectura individualizada de documentos PDF, la extracción manual a matrices Excel, y la visualización posterior en herramientas como i2 Analyst's Notebook. Este enfoque puede tomar entre 4 y 6 meses por caso, lo cual reduce el tiempo real para investigar, analizar, formular hipótesis y judicializar redes completas.

Además, la fragmentación del conocimiento es una problemática central: los análisis realizados por un grupo de analistas rara vez se integran con otros informes o bases de datos, limitando la posibilidad de identificar conexiones transversales entre actores que participan en múltiples casos.

En este contexto, el proyecto propone el desarrollo de un sistema inteligente de análisis criminal, que aproveche modelos de IA, PLN y aprendizaje automático para automatizar la extracción, categorización y visualización de datos en expedientes judiciales, especialmente en los delitos relacionados con la corrupción. Esta solución, además de reducir drásticamente los tiempos de procesamiento, permitiría integrar información entre casos, promover investigaciones más articuladas y robustecer la estrategia estatal contra la corrupción.

Marco Teórico

El marco teórico se basa en tres pilares fundamentales: análisis criminal, Inteligencia Artificial y Procesamiento de Lenguaje Natural.

El análisis criminal es definido como un proceso interdisciplinario para estudiar el delito, sus actores, patrones y contextos, con el fin de apoyar investigaciones y prevenir futuras

conductas delictivas (Tudela, 2015). En el caso de la Fiscalía General de la Nación, se articula dentro del ciclo de información criminal: recolección, evaluación, tratamiento, cotejo, análisis y difusión.

La Inteligencia Artificial (IA), entendida como la capacidad de las máquinas para imitar procesos cognitivos humanos, permite desarrollar sistemas que aprenden de los datos, detectan patrones y toman decisiones (Russell & Norvig, 2016). Dentro de la IA, el aprendizaje automático (Machine Learning) facilita la clasificación de expedientes, detección de anomalías y predicción de vínculos.

El Procesamiento de Lenguaje Natural (NLP) permite extraer conocimiento de textos en lenguaje humano. Herramientas como NER (Named Entity Recognition), análisis de relaciones, resumen automático o clasificación temática, permiten estructurar la información de expedientes judiciales, identificar entidades clave (personas, empresas, contratos) y visualizar las relaciones entre ellas (Bird et al., 2009; Jurafsky & Martin, 2023).

Marco Normativo

El proyecto se enmarca dentro de la normativa colombiana en materia de justicia, contratación pública, tratamiento de datos personales y uso ético de tecnologías.

A nivel judicial, la Ley 906 de 2004 establece el Código de Procedimiento Penal, esta norma regula las actuaciones penales en Colombia, bajo el sistema penal acusatorio. Define las funciones de la Fiscalía General de la Nación en la recolección, análisis y valoración de la prueba, así como la participación de los órganos de policía judicial. Establece también la legalidad y cadena de custodia de los elementos probatorios. La Resolución 0-0241 de 2022 de la Fiscalía General de la Nación establece los lineamientos para el análisis criminal y manejo de

expedientes judiciales, incluyendo las fases del ciclo de la información (recolección, evaluación, tratamiento, cotejo, análisis y difusión).

En cuanto a la contratación estatal, la Ley 80 de 1993 Estatuto General de Contratación de la Administración Pública establece el marco jurídico aplicable a los contratos estatales, sus modalidades, requisitos y principios; y la Ley 1150 de 2007 complementa y actualiza el Estatuto de Contratación, haciendo énfasis en mecanismos de control y uso de tecnologías de la información para mayor transparencia. Con el Decreto 1082 de 2015 se formaliza el SECOP como sistema oficial de publicación y seguimiento de la contratación pública en Colombia.

En lo relacionado con el manejo de datos, la Ley 1581 de 2012 regula la protección de datos personales por parte de entidades públicas y privadas, aspecto clave cuando se trabaja con expedientes judiciales que contienen información sensible. Finalmente, desde una perspectiva tecnológica, el CONPES 3975 de 2019 define la hoja de ruta del Gobierno colombiano para el uso responsable y ético de la IA recomendando el uso de IA en el sector público para mejorar eficiencia y toma de decisiones, siempre con respeto a los derechos humanos, ética y transparencia algorítmica. promueve el uso ético de la inteligencia artificial en el sector público colombiano, brindando lineamientos para su desarrollo e implementación responsable.

El sistema propuesto, por tanto, se alinea con las exigencias legales y éticas, al tiempo que responde a las necesidades institucionales de la Fiscalía para mejorar la eficacia en la lucha contra la corrupción.

Análisis Criminal: Estudio sistemático e interdisciplinario del delito y de los factores problemáticos que alteran la convivencia social e interesan a la investigación penal (sociodemográficos, espaciales, y temporales, entre otros), para apoyar la función constitucional

asignada a la Fiscalía General de la Nación y propender por la garantía de los derechos fundamentales de las víctimas a la verdad, la justicia, la reparación y la no repetición.

El análisis de información policial de naturaleza operativa tiene por objeto lograr resultados policiales concretos como una detención, un decomiso, la confiscación de activos o productos de origen delictivo, o el desmantelamiento de una empresa delictiva. En concreto, tiene por finalidad descubrir: vínculos entre sospechosos; el papel específico que una persona o personas de interés desempeñan en actividades delictivas; pistas de investigación; lagunas en la información.

Expediente Digital: Integridad completa del expediente penal que se verá reflejada y garantizada en el expediente digital SPOA.

i2 Analyst's Notebook: Software con capacidades de análisis visual avanzadas que convierten rápidamente conjuntos complejos de información dispar en información procesable de alta calidad para ayudarlos a ellos y a quienes participan en el análisis de inteligencia a identificar, predecir y contrarrestar actividades delictivas, terroristas y fraudulentas.

Recolección de Información: La Fiscalía General de la Nación, dentro del sistema que se rige por el principio acusatorio, desarrolla actividades de investigación a través de los órganos de policía judicial, tendientes a la recolección de elementos materiales probatorios, información legalmente obtenida y evidencia física con la dirección y control del fiscal del caso tendientes a establecer la veracidad de los hechos noticiados, la comisión de una conducta penal y el presunto responsable.

Evaluación de Información: Procedimiento dirigido a determinar la confiablidad de la fuente y la credibilidad de la información, con el fin de medir la exactitud de la misma.

Tratamiento de la Información: Serie de pasos como registro, almacenamiento y ordenación de la información.

Cotejo de la Información: Búsqueda y cotejo de datos registrados en bases mecánicas, magnéticas u otras similares de información de acceso público. Se recomienda que la policía judicial obtenga de la entidad una certificación en la que se indique la naturaleza pública de la base de datos sobre la cual se hace el cotejo.

Inteligencia Artificial: Es la ciencia e ingeniería de hacer máquinas inteligentes, especialmente programas informáticos inteligentes. Se relaciona con la tarea similar de usar equipos para comprender la inteligencia humana, pero la IA no tiene que ajustarse a los métodos biológicos observables.

Es la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano.

Machine Learning: Es un área de la inteligencia artificial que engloba un conjunto de técnicas que hacen posible el aprendizaje automático a través del entrenamiento con grandes volúmenes de datos.

Procesamiento de Lenguaje Natural (NLP): Es el campo de estudio que busca entender cómo funciona el lenguaje, su construcción, la generación de nuevo lenguaje, así como todas las tareas que tienen relación con el tratamiento del lenguaje. Entre estas tareas se tiene la generación de nuevo texto, traducciones de un idioma a otro, preguntas y respuestas, generar resumen, chatbots entre otros.

Para el Procesamiento de Lenguaje Natural, se usarán las siguientes técnicas:

- NER Named Entity Recognition. Extracción de información para identificar y clasificar entidades como personas, empresas, lugares, fechas, montos. Para ello existen herramientas como spaCy, Stanford Ner, Hugging Face Transformers.
- Análisis de relaciones entre entidades, utilizando spaCy con Dependency Parsing, BERT con Relation Extraction Models.
- Encontrar casos similares en los expedientes, utilizando herramientas como: TF-IDF + Cosine Similarity para comparar documentos; BERT embeddings para mejorar la precisión del análisis.
- Resumen automático de expedientes. Utilizando TF-IDF, TextRank.
- Clasificación de expedientes según el tipo de delito. Utilizando herramientas como Naive Bayes Random Forest.

Para identificar conexiones entre personas, grupos y casos, se pueden aplicar varas técnicas de Machine Learning, asi:

Modelos de aprendizaje supervisado para clasificación de expedientes, según el tipo de delito.

- Naive Bayes
- Random Forest XGBoost. Mayor precisión en grandes volúmenes de datos.
- BERT RoBERTa. Modelos de lenguaje para analizar documentos legales.

Modelos no supervisados para detección de patrones. Técnicas:

- Clustering (K-means, DBSCAN, HDBSCAN). Para agrupar casos similares.
- Isolation Forest One-Class SVM. Para detectar transacciones sospechosas.
- Modelos de redes para identificación de conexiones entre personas, empresas y contratos.

- Graph Neural Networks (GNNs). Modelos avanzados que identifican patrones en redes.
- Algoritmos de centralidad. PageRank, Betweenness Centrality.

Detección de relaciones con modelos de embeddings, para encontrar casos similares o relaciones ocultas. Técnicas:

- Word2Vec FastText. Para capturar relación entre las palabras en los documentos.
- BERT + Sentence Transformers. Para comparar documentos completos.

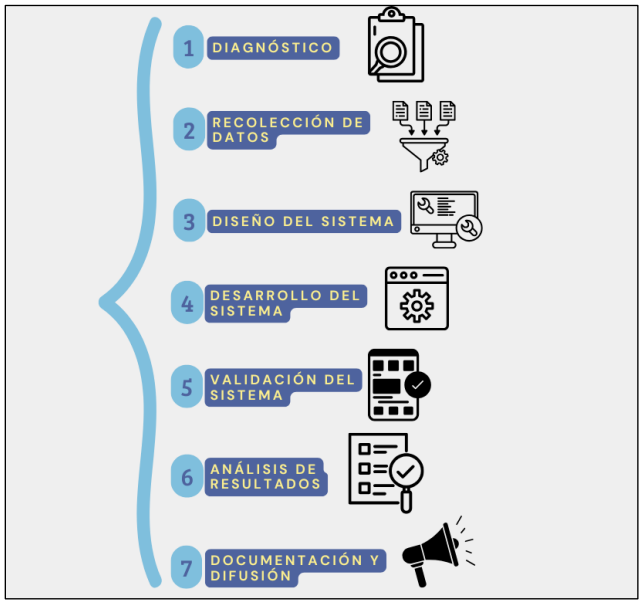
Metodologia

Esta investigación será de tipo aplicada, enfocada en desarrollar un sistema basado en Inteligencia Artificial, Machine Learning y Procesamiento de Lenguaje Natural (NLP) para mejorar los tiempos y optimizar el análisis criminal en la Fiscalía General de la Nación.

El diseño metodológico se estructura en 7 fases:

Figura 1

Fases del Diseño Metodológico



Diagnóstico

Objetivo: comprender el flujo actual de trabajo del análisis criminal y las necesidades específicas de los analistas.

Actividades:

- Revisión documental de los procedimientos operativos normalizados (PON) de análisis criminal de la Fiscalía.
- Entrevistas semiestructuradas a analistas del SAC y CEAC.

- Identificación de cuellos de botella en las fases de extracción, tratamiento y análisis.
- Detección de oportunidades para la automatización.

Recolección de Datos

Objetivo: obtener expedientes judiciales reales que sirvan como base de entrenamiento y prueba.

Actividades:

- Solicitud formal a la Dirección Especializada contra la Corrupción de casos ya cerrados.
- Selección de expedientes representativos que incluyan contratos, certificaciones, y actores involucrados.
- Clasificación de documentos según tipo, estructura, complejidad y formato (PDF, Word, imágenes escaneadas).
- Aseguramiento de la legalidad y privacidad de los datos.

Diseño del Sistema

Objetivo: estructurar el sistema en módulos funcionales para su desarrollo posterior.

Los modulos propuestos son:

- OCR (Reconocimiento Óptico de Caracteres): para transformar imágenes o PDFs escaneados en texto.
- Extracción de entidades y relaciones: mediante técnicas de NLP (spaCy, BERT, NER, dependency parsing).
- Análisis de relaciones: construcción de grafos con NetworkX, algoritmos de centralidad (PageRank, Betweenness).

- Visualización: representación gráfica o matricial de las redes con Graphviz o integración con i2.

Desarrollo del Sistema

Objetivo: implementar el sistema de análisis automatizado de expedientes.

- Limpieza y preprocesamiento de datos textuales.
- Entrenamiento de modelos de PLN para reconocimiento de entidades (Nombres, Cédulas, Empresas, Valores).
- Desarrollo de reglas o modelos de extracción de relaciones (contratista–contratante, vínculos legales).
- Construcción de grafos y paneles de visualización.
- Programación en Python (usando bibliotecas como spaCy, Hugging Face Transformers, Pandas, NetworkX).

Validación del Sistema

Objetivo: comprobar que el sistema funciona correctamente y supera al método manual tradicional.

- Procesamiento paralelo de expedientes: manual vs. automatizado.
- Evaluación de métricas: precisión, exhaustividad y tiempo.
- Validación de relaciones extraídas con analistas humanos.
- Encuestas de usabilidad y percepción a usuarios finales (analistas judiciales).

Análisis de Resultados

Objetivo: medir el impacto del sistema sobre la labor investigativa.

- Evaluación de reducción de tiempos (esperada > 50%).
- Ejemplos de casos donde se identificaron vínculos no evidentes manualmente.

- Revisión del sistema como apoyo en la formulación de hipótesis investigativas.
- Identificación de limitaciones técnicas, operativas o legales.

Documentación y Difusión

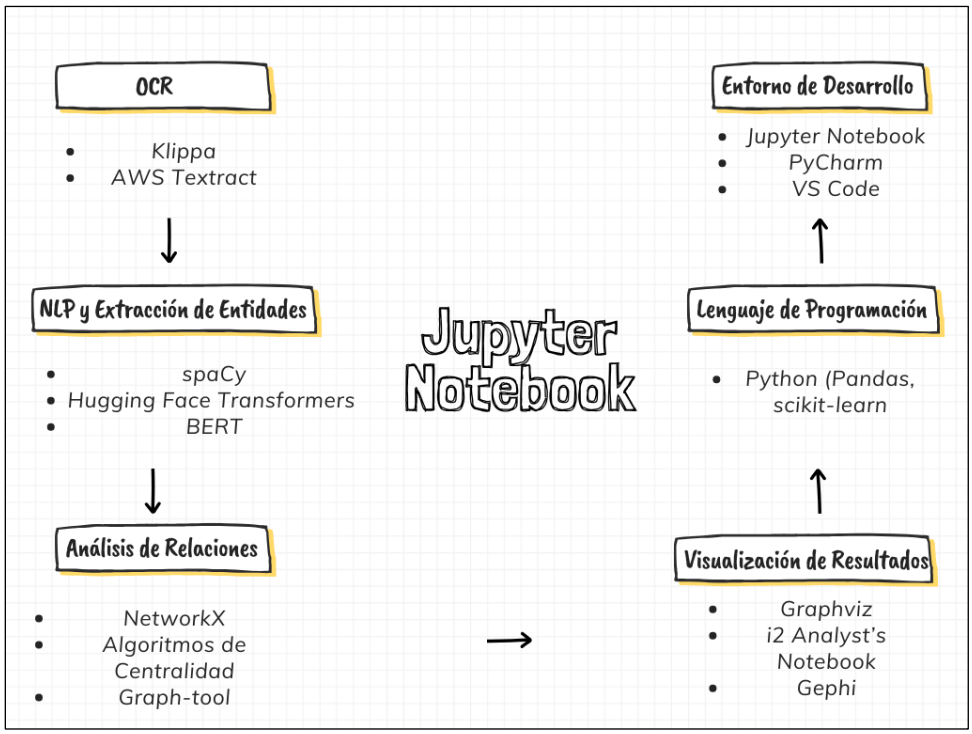
Objetivo: facilitar la adopción institucional del sistema y garantizar su uso responsable.

- Elaboración de manuales de usuario y guías de buenas prácticas.
- Preparación de presentaciones para tomadores de decisión en la Fiscalía.
- Diseño de protocolo de integración del sistema con los flujos institucionales existentes.
- Socialización académica y profesional de los resultados del proyecto.

Técnicas y Herramientas

Figura 2

Técnicas y Herramientas



Herramientas de OCR: Klippa, AWS Textract, UbiAI.

- Klippa: Plataforma SaaS de OCR avanzada que extrae datos estructurados de documentos escaneados o fotos. Soporta múltiples idiomas y formatos. Ideal para documentos legales.
- AWS Textrac: Servicio de Amazon Web Services que analiza automáticamente documentos escaneados, identifica texto, tablas y campos clave usando IA. Se integra fácilmente con otros servicios en la nube.

Procesamiento de Lenguaje Natural (NLP): SpaCy, Hugging Face Transformers.

- Spacy: Librería de procesamiento de lenguaje natural en Python muy eficiente. Realiza segmentación de oraciones, tokenización, lematización, y especialmente *Named Entity Recognition (NER)* para identificar personas, organizaciones, fechas, etc.
- Hugging Face Transformers: Repositorio de modelos preentrenados de PLN (como BERT, RoBERTa, DistilBERT). Proporciona herramientas para tareas como clasificación, respuesta automática, resumen, y más.

Análisis de relaciones: Algoritmos de centralidad, NetworkX

- Algoritmos de centralidad: PageRank, evalúa la importancia de un nodo en función de los enlaces que recibe. Betweenness Centrality, mide cuántas veces un nodo actúa como puente en la red.
- NetworkX: Librería de Python para la creación, manipulación y estudio de grafos y redes complejas. Permite modelar relaciones entre entidades como nodos y vínculos como aristas.

Visualización de datos: Graphviz, NetworkX, o integración con i2 Analyst′s Notebook.

- Graphviz: Software de código abierto para generar gráficos de relaciones a partir de descripciones en texto. Muy útil para crear visualizaciones claras y jerárquicas.
- i2 Analysts Notebook: Herramienta comercial avanzada de IBM para análisis visual e inteligencia. Usada por cuerpos policiales y agencias judiciales para detectar patrones, vínculos y secuencias en redes criminales.

Lenguaje de programación: Python

- Python: Lenguaje versátil, de código abierto, ideal para ciencia de datos, análisis de texto, y construcción de prototipos rápidos. Con librerías como Pandas para manipular estructuras de datos como DataFrames (muy similar a Excel), ideal para limpiar, transformar y analizar datos tabulados; y scikit-learn que es una librería de aprendizaje automático que permite aplicar algoritmos de clasificación, agrupamiento (clustering), reducción de dimensionalidad y validación cruzada.

Entorno de desarrollo: Jupyter Notebook o IDEs como PyCharm.

- Jupyter Notebook: Entorno interactivo que permite escribir código, visualizar resultados y documentar el proceso en una sola interfaz. Ideal para prototipado, pruebas y documentación de análisis.
- Pycharm: IDE profesional enfocado en Python, muy robusto para desarrollo de proyectos complejos. Soporta debugging, testing y control de versiones.

Análisis de Resultados

El sistema propuesto fue desarrollado e implementado en Python como una herramienta orientada al procesamiento automatizado de documentos digitales y análisis relacional de entidades asociadas a expedientes investigativos. La solución integra técnicas de procesamiento de texto, reconocimiento de entidades mediante modelos de lenguaje, OCR y análisis de redes, permitiendo transformar información no estructurada en representaciones relacionales útiles para el análisis criminal.

El análisis de resultados se presenta de acuerdo con los objetivos específicos planteados en la investigación:

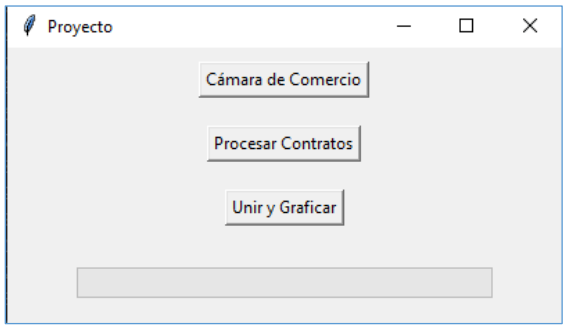
Diseñar un Sistema de Extracción Automatizada

Relacionado con documentos relevantes contenidos en los expedientes digitales de la Fiscalía General de la Nación, como Contratos, Certificados de Matrícula y Certificados de Existencia y Representación Legal.

Se desarrolló un sistema capaz de procesar documentos PDF correspondientes a contratos y certificados de Cámara de Comercio, extrayendo información relevante de manera automatizada.

Figura 3

Menú Principal

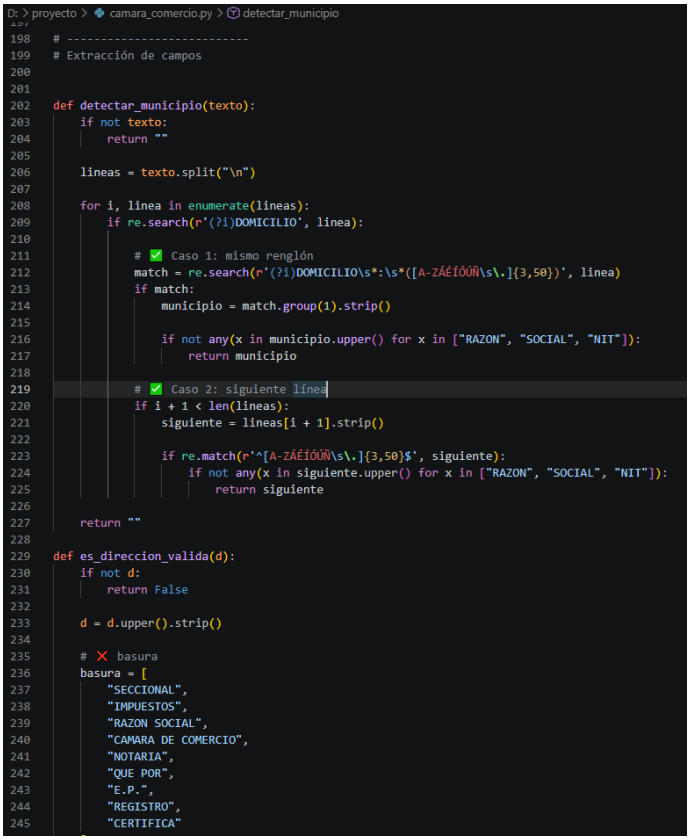


El sistema desarrollado tiene un menú principal, el cual da al usuario tres (03) opciones que son: Cámara de Comercio, Procesar Contratos, Unir y Graficar.

Se implementaron mecanismos de extracción de texto mediante herramientas especializadas para lectura de PDF y OCR en casos donde los documentos correspondían a imágenes escaneadas. Posteriormente, se aplicaron técnicas de procesamiento de texto y modelos de lenguaje para identificar entidades relevantes dentro del contenido documental.

Figura 4

Código Python Módulo camara_comercio.py

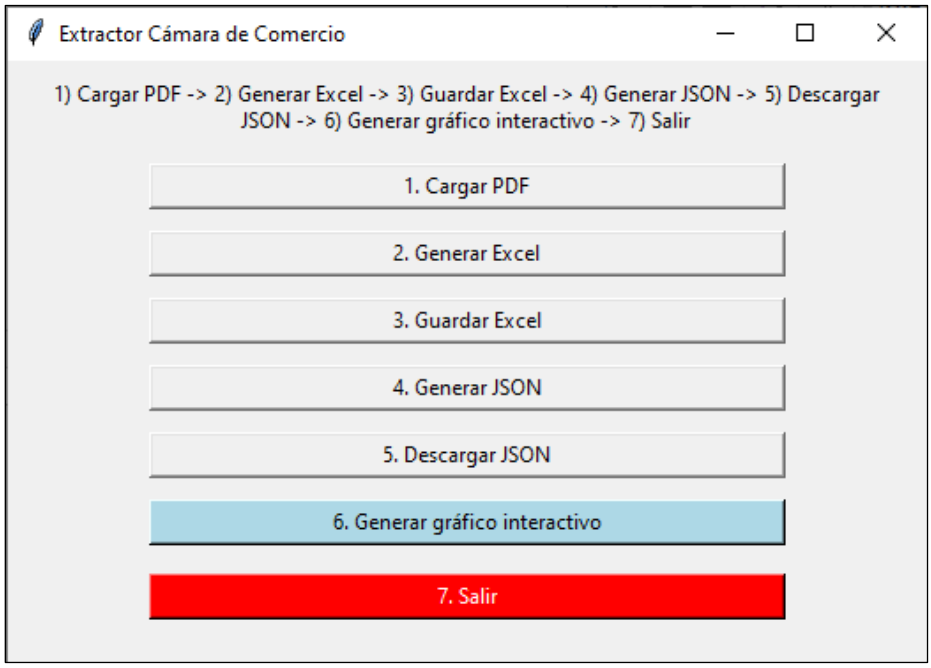


```
D: > proyecto > camara_comercio.py > detectar_municipio
198  # ---------------------------
199  # Extracción de campos
200
201
202  def detectar_municipio(texto):
203      if not texto:
204          return ""
205
206      lineas = texto.split("\n")
207
208      for i, linea in enumerate(lineas):
209          if re.search(r'(?i)DOMICILIO', linea):
210
211              # ✅ Caso 1: mismo renglón
212              match = re.search(r'(?i)DOMICILIO\s*:\s*([A-ZÁÉÍÓÚÑ\s\.]{3,50})', linea)
213              if match:
214                  municipio = match.group(1).strip()
215
216                  if not any(x in municipio.upper() for x in ["RAZON", "SOCIAL", "NIT"]):
217                      return municipio
218
219              # ✅ Caso 2: siguiente línea
220              if i + 1 < len(lineas):
221                  siguiente = lineas[i + 1].strip()
222
223                  if re.match(r'^[A-ZÁÉÍÓÚÑ\s\.]{3,50}$', siguiente):
224                      if not any(x in siguiente.upper() for x in ["RAZON", "SOCIAL", "NIT"]):
225                          return siguiente
226
227      return ""
228
229  def es_direccion_valida(d):
230      if not d:
231          return False
232
233      d = d.upper().strip()
234
235      # ❌ basura
236      basura = [
237          "SECCIONAL",
238          "IMPUESTOS",
239          "RAZON SOCIAL",
240          "CAMARA DE COMERCIO",
241          "NOTARIA",
242          "QUE POR",
243          "E.P.",
244          "REGISTRO",
245          "CERTIFICA"
```

La figura 4 muestra una parte del código desarrollado en Python en donde se extraen campos del Certificado de Cámara de Comercio; en este caso particular se extrae la información del municipio y dirección de la Persona Jurídica.

Figura 5

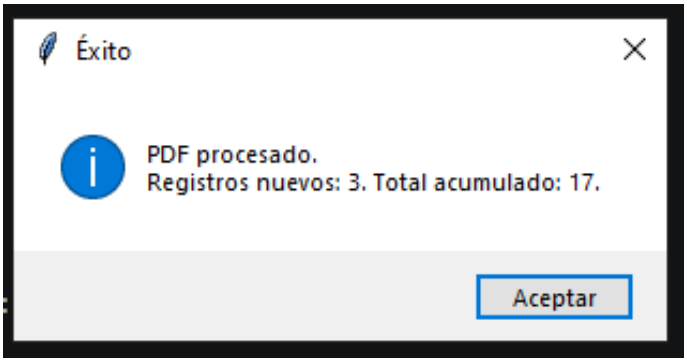
Módulo Cámara de Comercio



El módulo *Cámara de Comercio* permite al usuario cargar uno a uno archivos pdf para posteriormente generar un archivo Excel, se debe proceder a descargar un único archivo de Excel.

Figura 6

Mensaje Total de Registros Procesados por el Software Desarrollado



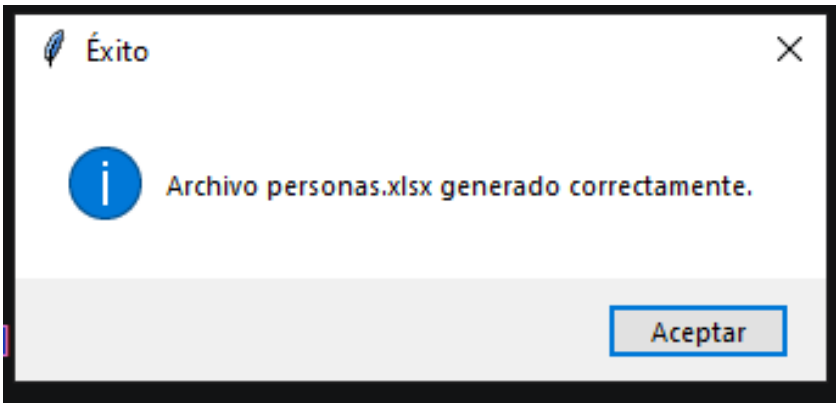
Cada vez que se carga un archivo pdf, el sistema muestra al usuario un mensaje en el que se indica que el archivo ha sido cargado y procesado satisfactoriamente, mostrando así mismo la cantidad de registros cargados y el total de registros acumulados.

Durante las pruebas realizadas, el sistema logró identificar correctamente información como:

- Nombres de empresas
- Nombres de personas
- Roles y/o cargos de las personas
- Contratantes y contratistas
- Números de identificación tributaria (NIT)
- Valores contractuales
- Fechas

Figura 7

Mensaje de Creación del Archivo personas.xlsx

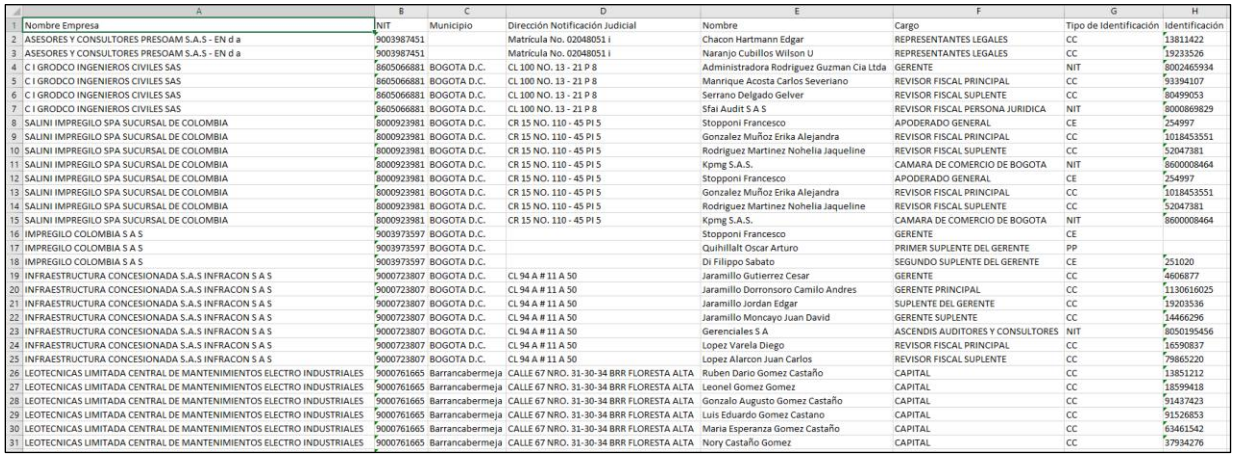


Al finalizar el proceso de carga de documentos de Cámara de Comercio se genera un archivo Excel, *personas.*xlsx, que contiene la totalidad de registros, con los siguientes campos:

Nombre empresa, NIT, Municipio, Dirección Notificación Judicial, Nombre, Cargo, Tipo Identificación e Identificación. Ver Figura 8.

Figura 8

Archivo personas.xlsx



	A	B	C	D	E	F	G	H
1	Nombre Empresa	NIT	Municipio	Dirección Notificación Judicial	Nombre	Cargo	Tipo de Identificación	Identificación
2	ASESORES Y CONSULTORES PRESOAM S.A.S - EN d a	9003987451		Matricula No. 02048051 i	Chacon Hartmann Edgar	REPRESENTANTES LEGALES	CC	13811422
3	ASESORES Y CONSULTORES PRESOAM S.A.S - EN d a	9003987451		Matricula No. 02048051 i	Naranjo Cubillos Wilson U	REPRESENTANTES LEGALES	CC	19233526
4	C I GRODCO INGENIEROS CIVILES SAS	8605066881	BOGOTA D.C.	CL 100 NO. 13 - 21 P 8	Administradora Rodriguez Guzman Cia Ltda	GERENTE	NIT	8002465934
5	C I GRODCO INGENIEROS CIVILES SAS	8605066881	BOGOTA D.C.	CL 100 NO. 13 - 21 P 8	Manrique Acosta Carlos Severiano	REVISOR FISCAL PRINCIPAL	CC	93394107
6	C I GRODCO INGENIEROS CIVILES SAS	8605066881	BOGOTA D.C.	CL 100 NO. 13 - 21 P 8	Serrano Delgado Gelver	REVISOR FISCAL SUPLENTE	CC	80499053
7	C I GRODCO INGENIEROS CIVILES SAS	8605066881	BOGOTA D.C.	CL 100 NO. 13 - 21 P 8	Sfai Audit S A S	REVISOR FISCAL PERSONA JURIDICA	NIT	8000869829
8	SALINI IMPREGILO SPA SUCURSAL DE COLOMBIA	8000923981	BOGOTA D.C.	CR 15 NO. 110 - 45 PI 5	Stopponi Francesco	APODERADO GENERAL	CE	254997
9	SALINI IMPREGILO SPA SUCURSAL DE COLOMBIA	8000923981	BOGOTA D.C.	CR 15 NO. 110 - 45 PI 5	Gonzalez Muñoz Erika Alejandra	REVISOR FISCAL PRINCIPAL	CC	1018453551
10	SALINI IMPREGILO SPA SUCURSAL DE COLOMBIA	8000923981	BOGOTA D.C.	CR 15 NO. 110 - 45 PI 5	Rodriguez Martinez Nohelia Jaqueline	REVISOR FISCAL SUPLENTE	CC	52047381
11	SALINI IMPREGILO SPA SUCURSAL DE COLOMBIA	8000923981	BOGOTA D.C.	CR 15 NO. 110 - 45 PI 5	Kpmg S.A.S.	CAMARA DE COMERCIO DE BOGOTA	NIT	8600008464
12	SALINI IMPREGILO SPA SUCURSAL DE COLOMBIA	8000923981	BOGOTA D.C.	CR 15 NO. 110 - 45 PI 5	Stopponi Francesco	APODERADO GENERAL	CE	254997
13	SALINI IMPREGILO SPA SUCURSAL DE COLOMBIA	8000923981	BOGOTA D.C.	CR 15 NO. 110 - 45 PI 5	Gonzalez Muñoz Erika Alejandra	REVISOR FISCAL PRINCIPAL	CC	1018453551
14	SALINI IMPREGILO SPA SUCURSAL DE COLOMBIA	8000923981	BOGOTA D.C.	CR 15 NO. 110 - 45 PI 5	Rodriguez Martinez Nohelia Jaqueline	REVISOR FISCAL SUPLENTE	CC	52047381
15	SALINI IMPREGILO SPA SUCURSAL DE COLOMBIA	8000923981	BOGOTA D.C.	CR 15 NO. 110 - 45 PI 5	Kpmg S.A.S.	CAMARA DE COMERCIO DE BOGOTA	NIT	8600008464
16	IMPREGILO COLOMBIA S A S	9003973597	BOGOTA D.C.		Stopponi Francesco	GERENTE	CE	
17	IMPREGILO COLOMBIA S A S	9003973597	BOGOTA D.C.		Quihillalt Oscar Arturo	PRIMER SUPLENTE DEL GERENTE	PP	
18	IMPREGILO COLOMBIA S A S	9003973597	BOGOTA D.C.		Di Filippo Sabato	SEGUNDO SUPLENTE DEL GERENTE	CE	251020
19	INFRAESTRUCTURA CONCESIONADA S.A.S INFRACON S A S	9000723807	BOGOTA D.C.	CL 94 A # 11 A 50	Jaramillo Gutierrez Cesar	GERENTE	CC	4606877
20	INFRAESTRUCTURA CONCESIONADA S.A.S INFRACON S A S	9000723807	BOGOTA D.C.	CL 94 A # 11 A 50	Jaramillo Dorronsoro Camilo Andres	GERENTE PRINCIPAL	CC	1130616025
21	INFRAESTRUCTURA CONCESIONADA S.A.S INFRACON S A S	9000723807	BOGOTA D.C.	CL 94 A # 11 A 50	Jaramillo Jordan Edgar	SUPLENTE DEL GERENTE	CC	19203536
22	INFRAESTRUCTURA CONCESIONADA S.A.S INFRACON S A S	9000723807	BOGOTA D.C.	CL 94 A # 11 A 50	Jaramillo Moncayo Juan David	GERENTE SUPLENTE	CC	14466296
23	INFRAESTRUCTURA CONCESIONADA S.A.S INFRACON S A S	9000723807	BOGOTA D.C.	CL 94 A # 11 A 50	Gerenciales S A	ASCENDIS AUDITORES Y CONSULTORES	NIT	8050195456
24	INFRAESTRUCTURA CONCESIONADA S.A.S INFRACON S A S	9000723807	BOGOTA D.C.	CL 94 A # 11 A 50	Lopez Varela Diego	REVISOR FISCAL PRINCIPAL	CC	16590837
25	INFRAESTRUCTURA CONCESIONADA S.A.S INFRACON S A S	9000723807	BOGOTA D.C.	CL 94 A # 11 A 50	Lopez Alarcon Juan Carlos	REVISOR FISCAL SUPLENTE	CC	79865220
26	LEOTECNICAS LIMITADA CENTRAL DE MANTENIMIENTOS ELECTRO INDUSTRIALES	9000761665	Barrancabermeja	CALLE 67 NRO. 31-30-34 BRR FLORESTA ALTA	Ruben Dario Gomez Castaño	CAPITAL	CC	13851212
27	LEOTECNICAS LIMITADA CENTRAL DE MANTENIMIENTOS ELECTRO INDUSTRIALES	9000761665	Barrancabermeja	CALLE 67 NRO. 31-30-34 BRR FLORESTA ALTA	Leonel Gomez Gomez	CAPITAL	CC	18599418
28	LEOTECNICAS LIMITADA CENTRAL DE MANTENIMIENTOS ELECTRO INDUSTRIALES	9000761665	Barrancabermeja	CALLE 67 NRO. 31-30-34 BRR FLORESTA ALTA	Gonzalo Augusto Gomez Castaño	CAPITAL	CC	91437423
29	LEOTECNICAS LIMITADA CENTRAL DE MANTENIMIENTOS ELECTRO INDUSTRIALES	9000761665	Barrancabermeja	CALLE 67 NRO. 31-30-34 BRR FLORESTA ALTA	Luis Eduardo Gomez Castano	CAPITAL	CC	91526853
30	LEOTECNICAS LIMITADA CENTRAL DE MANTENIMIENTOS ELECTRO INDUSTRIALES	9000761665	Barrancabermeja	CALLE 67 NRO. 31-30-34 BRR FLORESTA ALTA	Maria Esperanza Gomez Castaño	CAPITAL	CC	63461542
31	LEOTECNICAS LIMITADA CENTRAL DE MANTENIMIENTOS ELECTRO INDUSTRIALES	9000761665	Barrancabermeja	CALLE 67 NRO. 31-30-34 BRR FLORESTA ALTA	Nory Castaño Gomez	CAPITAL	CC	37934276

Para el módulo *Contratos*, el sistema automáticamente ejecuta todos los archivos pdf que se encuentren guardados en la carpeta *CONTRATOS,* extrayendo la información relevante y al finalizar este proceso genera y guarda el archivo Excel correspondiente con los siguientes campos: Archivo, número_contrato, fecha_contrato, alor_contrato, contratante, nit_contratante, representante_legal_contratante, contratista, nit_contratista, representante_legal_contratista, id_representante_legal_contratista; así como la gráfica de asociación. Es importante aclarar que a medida que se procesan los contratos, el sistema muestra una barra de progreso. Ver Figuras 9 y 10.

Figura 9

Proceso de Extracción de Datos del Módulo Contratos

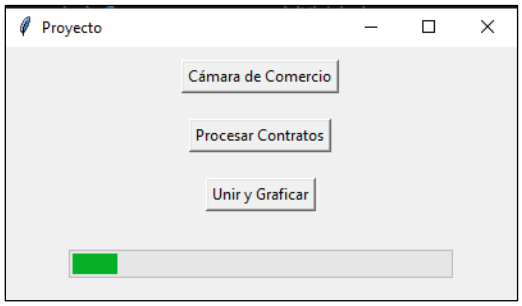
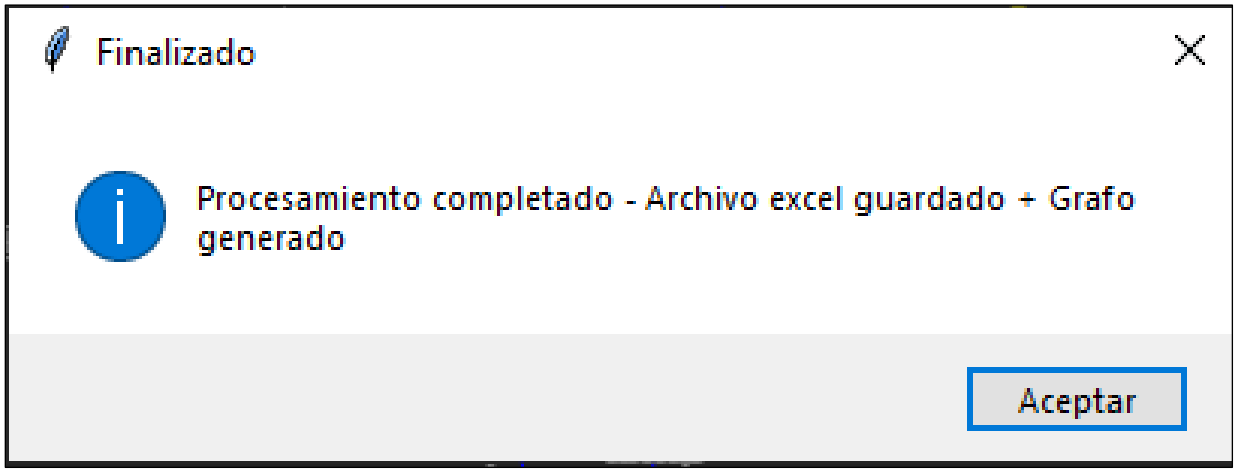


Figura 10

Finalización del Proceso de Extracción de Información



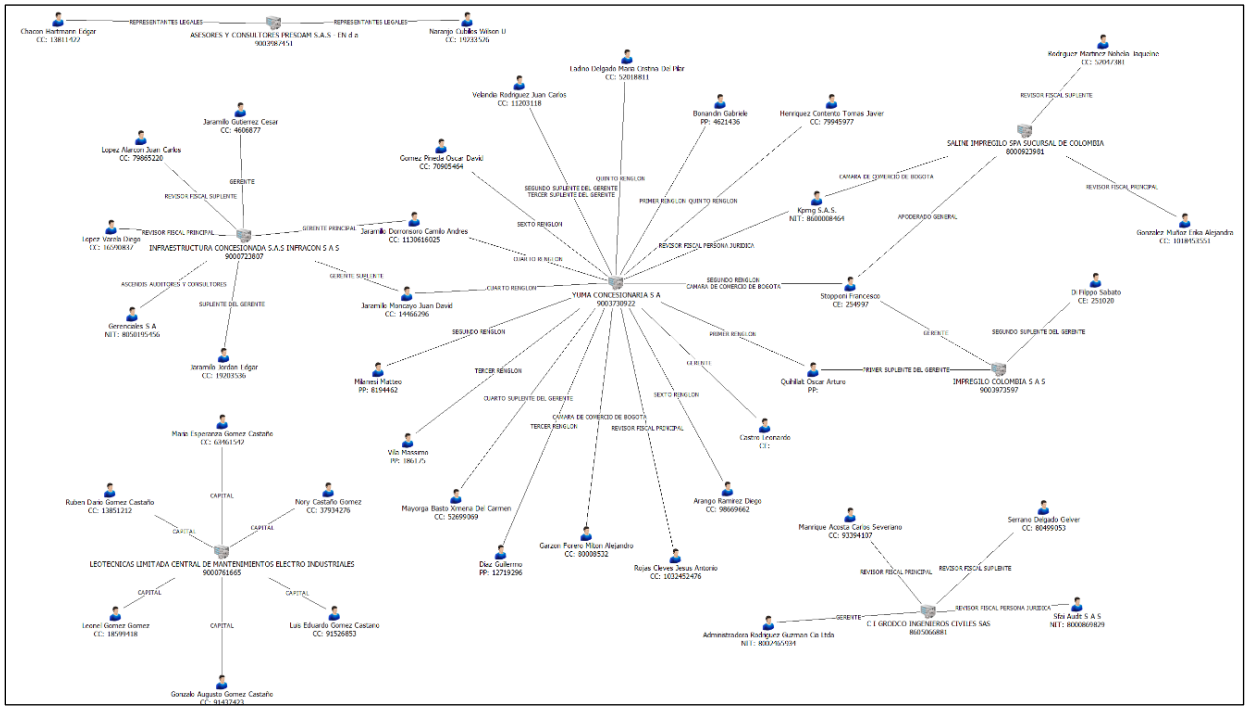
La automatización del proceso permitió reducir significativamente el tiempo requerido para la revisión manual de grandes volúmenes documentales, facilitando la estructuración de información que inicialmente se encontraba dispersa en documentos no estructurados. A modo de ejemplo, se procesaron ocho (08) documentos que tienen un total de ciento cuarenta (140) folios relacionados con contratos, un analista criminal con experiencia puede tardar 2 horas en leer y digitar la matriz de Excel, con el sistema desarrollado el procesamiento de estos ocho (08)

documentos fue de 2 a 3 minutos; lo que demuestra una reducción de aproximadamente el 95% en los tiempos de extracción y tratamiento de datos.

Adicionalmente, los resultados obtenidos fueron exportados automáticamente a archivos Excel, permitiendo su posterior verificación, validación, análisis y reutilización. Estos archivos de Excel están estructurados de tal manera que pueden ser utilizados en diferentes software de análisis como i2 Analyst´s Notebook.

Figura 11

Gráfico de Relaciones de Personas Naturales y Jurídicas

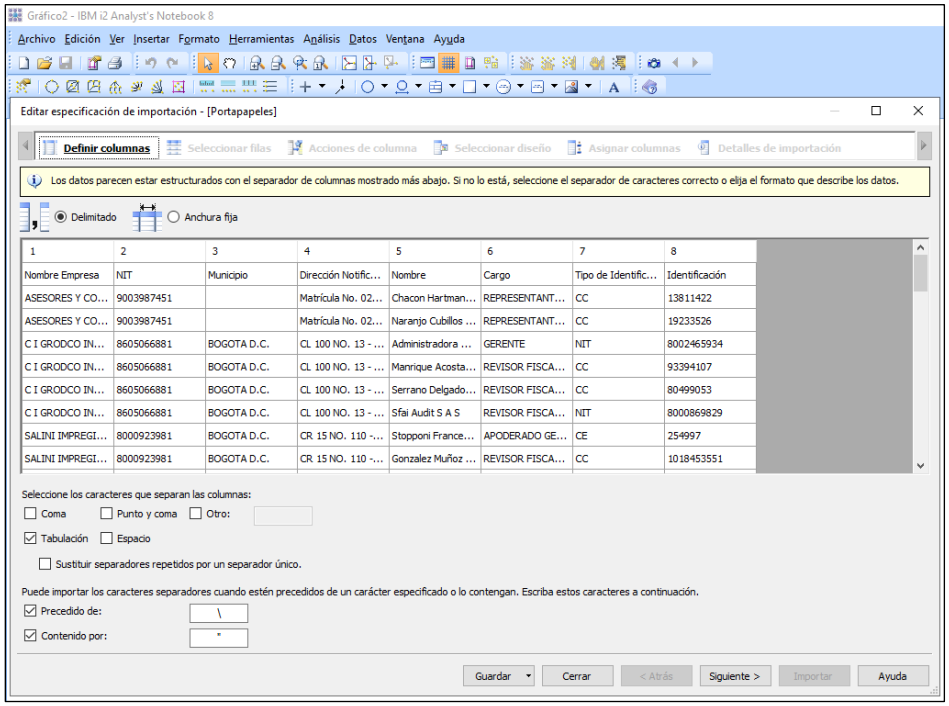


La figura 11 muestra el gráfico de relaciones de Cámara de comercio generado con el software i2 Analyst´s Notebook aprovechando el archivo Excel *personas.xlsx* generado con el sistema desarrollado. El proceso de importación requiere que el archivo Excel o de texto tenga una estructura de columnas y filas en donde cada fila es un registro que representa un vínculo

entre dos entidades (nodos), para este caso especifico las columnas que se cruzan son la de NIT de empresa e Identificación de cada persona registrada en cámara de comercio.

Figura 12

Proceso de Importación de Archivo personas.xlsx



1	2	3	4	5	6	7	8
Nombre Empresa	NIT	Municipio	Dirección Notific...	Nombre	Cargo	Tipo de Identific...	Identificación
ASESORES Y CO...	9003987451		Matrícula No. 02...	Chacon Hartman...	REPRESENTANT...	CC	13811422
ASESORES Y CO...	9003987451		Matrícula No. 02...	Naranjo Cubillos ...	REPRESENTANT...	CC	19233526
C I GRODCO IN...	8605066881	BOGOTA D.C.	CL 100 NO. 13 - ...	Administradora ...	GERENTE	NIT	8002465934
C I GRODCO IN...	8605066881	BOGOTA D.C.	CL 100 NO. 13 - ...	Manrique Acosta...	REVISOR FISCA...	CC	93394107
C I GRODCO IN...	8605066881	BOGOTA D.C.	CL 100 NO. 13 - ...	Serrano Delgado...	REVISOR FISCA...	CC	80499053
C I GRODCO IN...	8605066881	BOGOTA D.C.	CL 100 NO. 13 - ...	Sfai Audit S A S	REVISOR FISCA...	NIT	8000869829
SALINI IMPREGI...	8000923981	BOGOTA D.C.	CR 15 NO. 110 -...	Stopponi France...	APODERADO GE...	CE	254997
SALINI IMPREGI...	8000923981	BOGOTA D.C.	CR 15 NO. 110 -...	Gonzalez Muñoz ...	REVISOR FISCA...	CC	1018453551

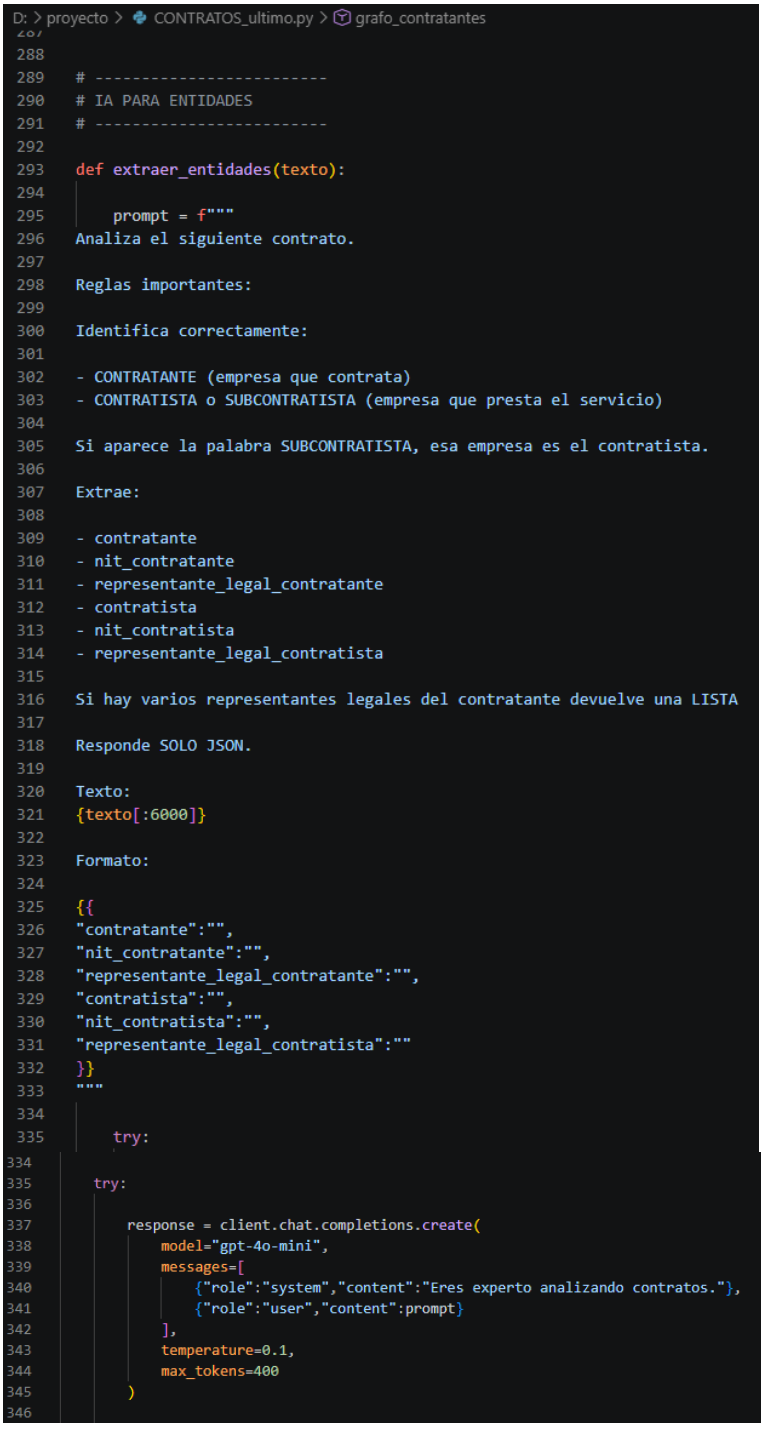
Establecer Vínculos entre Entidades Clave dentro del Texto Digitalizado

Tales como personas naturales, personas jurídicas, contratos, contratante, contratista con el fin de representar sus relaciones en el contexto de los expedientes judiciales.

El sistema implementó técnicas de Procesamiento de Lenguaje Natural mediante modelos de lenguaje preentrenados para realizar reconocimiento de entidades y extracción de relaciones dentro de los documentos procesados.

Figura 13

Script para Extraer Entidades en el Módulo Contratos_ultimo.py



```
D: > proyecto > CONTRATOS_ultimo.py > grafo_contratantes
288
289  # -------------------------
290  # IA PARA ENTIDADES
291  # -------------------------
292
293  def extraer_entidades(texto):
294
295      prompt = f"""
296  Analiza el siguiente contrato.
297
298  Reglas importantes:
299
300  Identifica correctamente:
301
302  - CONTRATANTE (empresa que contrata)
303  - CONTRATISTA o SUBCONTRATISTA (empresa que presta el servicio)
304
305  Si aparece la palabra SUBCONTRATISTA, esa empresa es el contratista.
306
307  Extrae:
308
309  - contratante
310  - nit_contratante
311  - representante_legal_contratante
312  - contratista
313  - nit_contratista
314  - representante_legal_contratista
315
316  Si hay varios representantes legales del contratante devuelve una LISTA
317
318  Responde SOLO JSON.
319
320  Texto:
321  {texto[:6000]}
322
323  Formato:
324
325  {{
326  "contratante":"",
327  "nit_contratante":"",
328  "representante_legal_contratante":"",
329  "contratista":"",
330  "nit_contratista":"",
331  "representante_legal_contratista":""
332  }}
333  """
334
335      try:
334
335    try:
336
337        response = client.chat.completions.create(
338            model="gpt-4o-mini",
339            messages=[
340                {"role":"system","content":"Eres experto analizando contratos."},
341                {"role":"user","content":prompt}
342            ],
343            temperature=0.1,
344            max_tokens=400
345        )
346
```

Como se muestra en la Figura 13 se utilizó un modelo de lenguaje preentrenado, específicamente GPT-4o-mini, consumido mediante API para realizar tareas de procesamiento de lenguaje natural sobre contratos. Para esto se ha dado la instrucción (prompt) necesaria para que el modelo realice a extracción de la información relacionada con *contratante, nit_contratante, representación_legal_contratante, contratista, nit_contratista, representante_legal_contratista*.

El modelo de lenguaje se utilizó únicamente para la extracción de entidades cuya identificación depende del contexto semántico del documento, como contratantes, contratistas y representantes legales, ya que estos elementos pueden presentarse con estructuras variables y redacción no uniforme. En cambio, información como fechas, valores y números de contrato presenta patrones estructurados y repetitivos, por lo que se extrajo de manera más eficiente mediante expresiones regulares en Python. De igual forma, los documentos de Cámara de Comercio poseen una estructura más estable, permitiendo realizar la extracción sin necesidad de modelos de lenguaje.

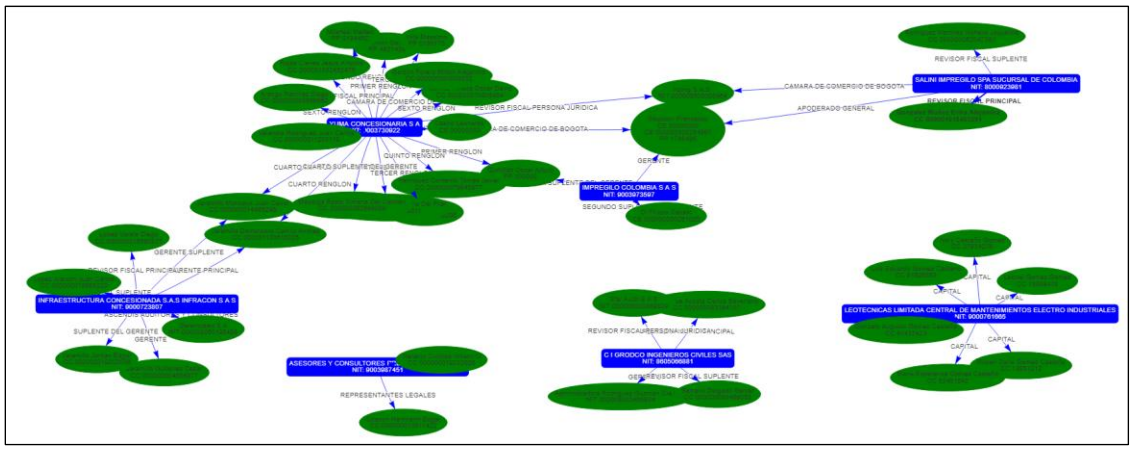
A partir del análisis de contratos y certificados empresariales, fue posible identificar relaciones entre:

- Personas naturales
- Personas jurídicas
- Contratos
- Representantes legales
- Empresas contratistas
- Empresas contratantes

Estas relaciones fueron estructuradas mediante grafos utilizando librerías de análisis de redes, permitiendo representar de manera explícita las conexiones existentes entre dichas entidades.

Figura 14

Gráfico de Relaciones Personas Naturales vs Personas Jurídicas

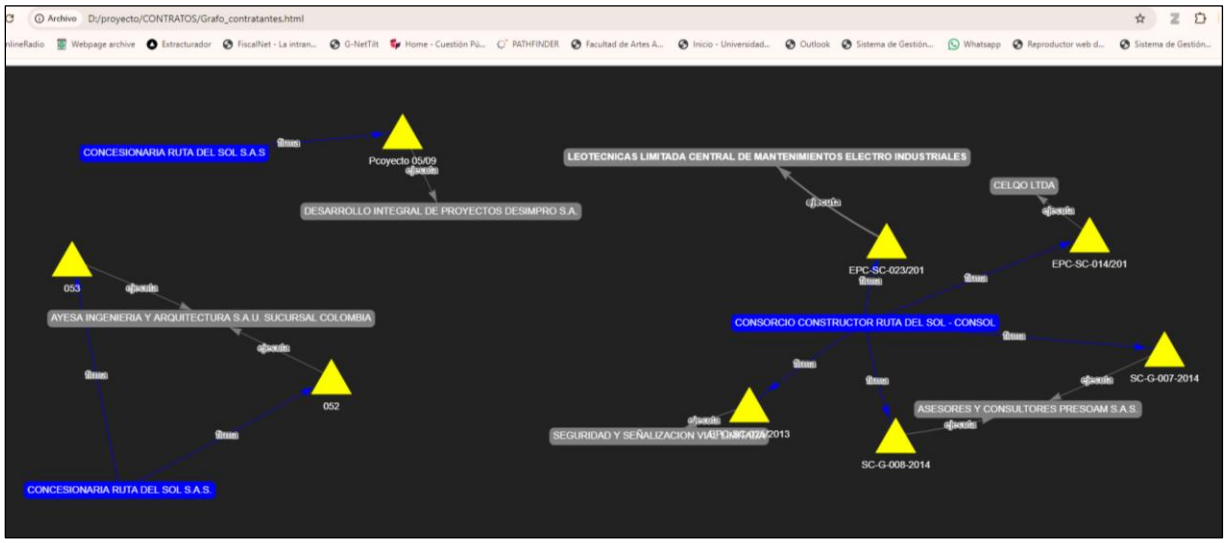


La figura 14 representa el gráfico de relaciones de Cámara de comercio generado por la aplicación desarrollada en el presente trabajo. Como se puede observar, existen siete (07) nodos que corresponden a las empresas o personas jurídicas, representadas en color azul, con los respectivos vínculos con las personas naturales que están registradas en cada una de los certificados de existencia y representacion legal. Para el analista criminal este tipo de gráficos es de suma importancia por que puede identificar a simple vista que personas pertenecen a diferentes empresas, como por ejemplo Francesco STOPPONI quien tiene relación con las empresas: YUMA CONCESIONARIA S.A, IMPREGILO COLOMBIA S.A. y SALINI IMPREGILO SUCURSAL COLOMBIA; con la sola lectura de documentos estas relaciones que no se ven a simple vista son difíciles de hallar para un analista. Si comparamos la Figura 15 (gráfico generado por la sistema) versus la Figura 12 (gráfico generado en i2) podemos observar

que las dos nos están mostrando la misma información, lo que indica que la generación de gráficos desarrollados en este proyecto está funcionando en un 100%. Es importante aclarar que ambos gráficos fueron realizados con la información extráida por el sistema de este proyecto aplicado.

Figura 15

Gráfico de Relaciones de Contratos



La figura 15 muestra el gráfico de Contratos generado por el sistema, donde cada contrato se representa con un triangulo amarillo, el contratante en azul y el contratista en gris. Con este gráfico el analista puede observar a simple vista que el Contratante CONSORCIO CONSTRUCTOR RUTA DEL SOL – CONSOL es el que más contratos presenta y que el contratista ASESORES Y CONSULTORES PRESOAM SAS ha firmado dos (02) contratos siendo estos el SC-G-007-2014 y SC-G-008-2014.

Los resultados evidenciaron la capacidad del sistema para consolidar información proveniente de múltiples fuentes y construir estructuras relacionales que facilitan la

identificación de patrones de interacción entre actores involucrados en los documentos analizados.

Desarrollar un Sistema de Visualización

Sistema que representa de forma gráfica o matricial los vínculos identificados entre entidades, permitiendo analizar las conexiones en los expedientes judiciales relacionados con casos de corrupción.

Como resultado final del procesamiento, el sistema generó grafos interactivos de relaciones mediante herramientas como NetworkX y PyVis.

Figura 16

Código Python del Módulo análisis.py

```
D: > proyecto > analisis.py > grafo_relaciones
  def grafo_relaciones(ruta_contratos, ruta_camara):
      # -------------------------
      # MERGE
      # -------------------------
      df = df_contratos.merge(
          df_camara,
          left_on="NIT_CONTRATISTA",
          right_on="NIT",
          how="left"
      )

      # -------------------------
      # CREAR GRAFO
      # -------------------------
      G = nx.Graph()

      for _, row in df.iterrows():

          contrato = str(row.get("NUMERO_CONTRATO", "")).strip()
          if not contrato:
              contrato = "SIN_CONTRATO"

          # -------------------------
          # EMPRESA (usar NIT como ID)
          # -------------------------
          nit_empresa = str(row.get("NIT_CONTRATISTA", "")).strip()
          nombre_empresa = str(row.get("CONTRATISTA", "")).strip()

          if not nit_empresa:
              continue  # evitar nodos basura

          G.add_node(
              nit_empresa,
              label=nombre_empresa,
              title=f"Empresa:\n{nombre_empresa}\nNIT: {nit_empresa}",
              color="blue",
              font={"color": "white"},
              physics=False,
              shape="box"
          )

          # -------------------------
          # CONTRATO
          # -------------------------
          G.add_node(
              contrato,
              label=contrato,
              title=f"Contrato: {contrato}",
```

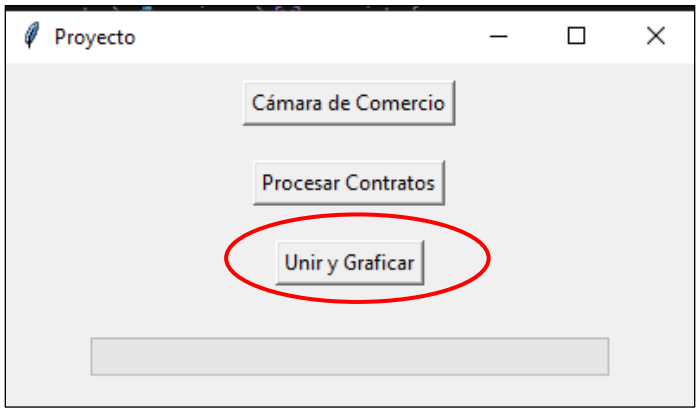
La figura 16 muestra parte del código Python con el que se leen los dos archivos de Excel, realizando una integración de datos mediante un join por identificador único, permitiendo enriquecer la información contractual con datos empresariales.

Está parte del proyecto implementado es de gran importancia, porque acá es donde se hace el análisis relacional uniendo dos fuentes distintas de información, Cámara de comercio y Contratos.

Este módulo integra dos fuentes de datos previamente procesadas mediante un join por NIT, permitiendo enriquecer la información contractual con datos de cámara de comercio. A partir de esta integración, se construye un grafo de relaciones donde se modelan empresas, contratos y personas, identificando vínculos clave como representantes legales y firmantes. Esta estructura permite realizar análisis relacional, facilitando la identificación de conexiones relevantes en contextos de investigación.

Figura 17

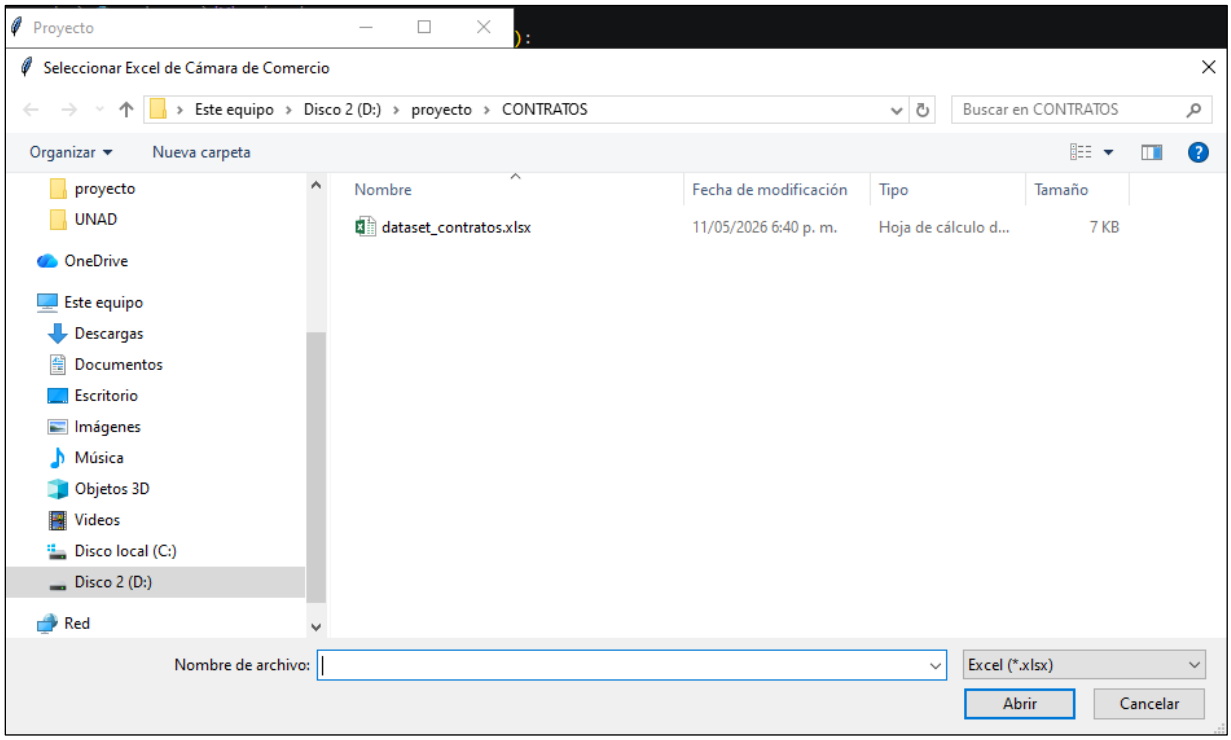
Ventana Principal



Al hacer clic en la opción de *Unir y Graficar* el sistema le solicita al usuario que seleccione el archivo Excel de Contratos y el archivo Excel de Cámara de Comercio.

Figura 18

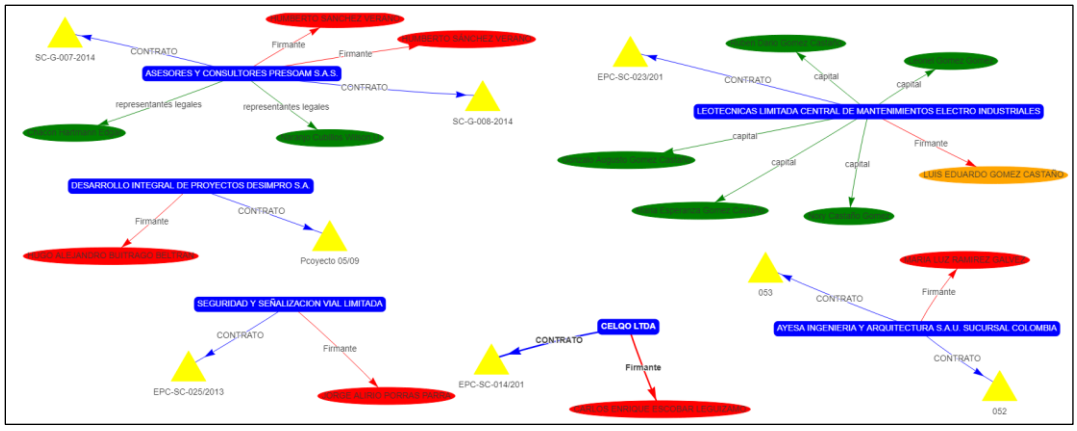
Ventana Selección Archivo Excel Cámara de Comercio



Paso seguido, el sistema genera un gráfico de relaciones entre la información contenida en estos dos archivos de Excel. Ver figura 19.

Figura 19

Gráfico de Relaciones Cámara de Comercio vs Contratos



La gráfica resultante, mostrada en la Figura 19, evidencia las relaciones existentes entre los contratos, las empresas contratistas y las personas naturales asociadas a estos documentos. En la visualización, la empresa contratista se representa mediante nodos de color azul, mientras que cada contrato se identifica con un nodo en forma de triángulo amarillo.

Por su parte, las personas naturales extraídas de los documentos de Cámara de Comercio se representan en color verde. Adicionalmente, las personas que firman los contratos en representación de la empresa contratista, pero que no aparecen registradas en los documentos de Cámara de Comercio, se muestran en color rojo. Finalmente, aquellas personas que firman el contrato y que sí coinciden con la información registrada en los documentos de Cámara de Comercio se representan mediante nodos de color naranja.

Esta representación gráfica permite identificar de manera visual las relaciones entre empresas, contratos y personas naturales, así como detectar coincidencias o posibles inconsistencias entre la información contractual y la información empresarial registrada.

Estas visualizaciones permitieron representar gráficamente nodos correspondientes a:

- Personas
- Empresas
- Contratos
- Representantes legales,

así como las relaciones o aristas existentes entre ellos. El tipo de gráfico aplicado al análisis criminal corresponde a la técnica conocida como diagrama de eslabones o relacional.

La representación gráfica facilitó el análisis visual de conexiones directas e indirectas entre actores, permitiendo identificar entidades con alta cantidad de vínculos y posibles estructuras relacionales relevantes para el análisis investigativo.

El uso de grafos permitió transformar grandes volúmenes de información textual en estructuras visuales más comprensibles para el analista, favoreciendo la generación de hipótesis investigativas y el análisis de relaciones complejas.

Asimismo, el sistema permitió detectar coincidencias entre personas registradas en documentos de Cámara de Comercio y participantes dentro de contratos, generando una visión integrada de las relaciones organizacionales y contractuales.

Finalmente, la integración de información contractual y empresarial en un único modelo relacional permitió enriquecer el contexto analítico de los expedientes procesados.

Todo el modelo desarrollado permite reducir hasta en un 90% el tiempo invertido por los analistas en las tareas de lectura, extracción, organización y estructuración de la información en matrices de análisis.

Conclusiones

El desarrollo del sistema permitió demostrar la viabilidad de aplicar técnicas de procesamiento automatizado de documentos, procesamiento de lenguaje natural y análisis de redes en contextos asociados al análisis investigativo y documental.

La solución implementada logró transformar información no estructurada contenida en documentos PDF en datos organizados y reutilizables, reduciendo significativamente la dependencia de procesos manuales de lectura y consolidación de información.

Asimismo, se evidenció que la integración de modelos de lenguaje con técnicas tradicionales de procesamiento de texto permite mejorar la extracción de entidades relevantes dentro de documentos jurídicos y contractuales, especialmente en escenarios donde la información presenta formatos variables o ambigüedad semántica.

El uso de grafos como mecanismo de representación permitió modelar relaciones entre personas, empresas y contratos, facilitando el análisis de conexiones que difícilmente pueden identificarse de manera manual cuando se manejan grandes volúmenes documentales.

Otro aspecto relevante del proyecto corresponde al enfoque de seguridad y confidencialidad de la información. Debido a que los documentos analizados pueden contener información sensible asociada a contextos investigativos, el sistema fue concebido como una solución orientada a entornos controlados, evitando la exposición innecesaria de datos a plataformas externas sin mecanismos adecuados de gobernanza y protección de la información.

Finalmente, el proyecto establece una base sólida para futuras investigaciones orientadas a incorporar técnicas de Machine Learning y análisis predictivo, tales como clasificación de riesgos, detección de anomalías o identificación automática de patrones relacionales complejos dentro de redes de entidades.

El modelo desarrollado permite reducir significativamente el tiempo invertido por los analistas criminales en las tareas de lectura, extracción, organización y estructuración de la información en matrices de análisis, alcanzando una disminución aproximada del 90% en comparación con el procesamiento manual tradicional. Este resultado adquiere mayor relevancia considerando que el sistema integra y correlaciona información proveniente de dos fuentes documentales distintas: contratos y documentos de Cámara de Comercio.

Adicionalmente, los archivos Excel generados por el sistema pueden ser utilizados e integrados fácilmente en otras herramientas tradicionales de análisis criminal e inteligencia ampliamente reconocidas a nivel internacional, como i2 Analyst's Notebook, software que ha sido utilizado durante más de 30 años en actividades de análisis investigativo, inteligencia y representación de redes relacionales. Esto permite garantizar interoperabilidad con metodologías y plataformas ya implementadas en entornos institucionales, facilitando la continuidad de los procesos de análisis y aprovechamiento de la información estructurada generada por el sistema.

Recomendaciones

A partir de los resultados obtenidos durante el desarrollo e implementación del sistema, se plantean las siguientes recomendaciones orientadas a fortalecer y ampliar el alcance de la solución propuesta.

1. Extender el sistema a otros contextos investigativos

Se recomienda ampliar el sistema para el análisis de otros tipos de delitos y estructuras investigativas, tales como lavado de activos, delitos financieros, enriquecimiento ilícito, contratación irregular y redes de tráfico ilegal.

La arquitectura modular desarrollada permite adaptar fácilmente los procesos de extracción y análisis a nuevos tipos documentales y escenarios investigativos, incorporando reglas específicas, nuevas entidades y relaciones adicionales según las necesidades institucionales.

Asimismo, la integración futura de modelos de Machine Learning permitiría evolucionar el sistema desde un enfoque descriptivo hacia uno predictivo, facilitando la detección automática de patrones de riesgo y comportamientos atípicos.

2. Integración con fuentes de información externas

Se recomienda integrar el sistema con bases de datos y plataformas externas de carácter público o institucional, tales como: SECOP, RUES, registros notariales y sistemas de contratación pública.

La integración de múltiples fuentes permitiría enriquecer el contexto analítico de las relaciones detectadas, aumentando la capacidad del sistema para identificar coincidencias, vínculos indirectos y patrones complejos entre personas, empresas y contratos.

3. Fortalecer la interfaz y experiencia de usuario

Aunque el sistema cuenta actualmente con una interfaz gráfica funcional desarrollada en Python, se recomienda evolucionar hacia una plataforma más robusta y orientada a usuarios no técnicos.

Entre las funcionalidades futuras podrían incluirse: búsqueda avanzada de entidades, filtros dinámicos sobre grafos, generación automática de reportes, paneles visuales interactivos, exportación de resultados, alertas automáticas de relaciones relevantes.

Estas mejoras permitirían optimizar la experiencia del analista y facilitar el uso del sistema dentro de entornos operativos reales.

4. Fortalecer los mecanismos de seguridad y gobernanza de datos

Debido a que el sistema puede operar sobre información sensible asociada a procesos investigativos y judiciales, se recomienda establecer lineamientos técnicos, éticos y legales claros para el uso de Inteligencia Artificial en este tipo de contextos.

En particular, se considera importante implementar controles de acceso y auditoría, garantizar trazabilidad sobre los procesos automatizados, definir políticas de manejo y almacenamiento seguro de datos, minimizar la exposición de información sensible a servicios externos, asegurar el cumplimiento de principios de confidencialidad y protección de datos.

En este sentido, se recomienda priorizar arquitecturas de procesamiento en entornos controlados o institucionales.

5. Incorporar técnicas avanzadas de análisis y Machine Learning

Como línea futura de investigación, se recomienda incorporar técnicas de Machine Learning y análisis avanzado de redes, tales como: clasificación automática de expedientes, detección de anomalías, clustering de casos similares, análisis predictivo, algoritmos de centralidad y modelos de similitud documental.

Estas técnicas permitirían incrementar las capacidades analíticas del sistema y apoyar la generación automática de alertas investigativas y priorización de casos.

Referencias

Bird, S., Klein, E., Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Caldwell, M., Andrews, J.T.A., Tanay, T. et al. AI-enabled future crime. Crime Sci 9, 14 (2020).

Cardoso, C. A., Pérez Abelleira, M. A. (2010). Minería de texto para la categorización automática de documentos.

Castro-Toledo, F.J., Miró-Llinares, F. Aguerri, J.C. Data-Driven Criminal Justice in the age of algorithms: epistemic challenges and practical implications. Crim Law Forum 34, 295 - 316 (2023).

Cloudera. (2023). Cloudera Manager Documentation.

Fiscalía General de la Nación (Colombia). (2022). Manual de análisis criminal y manejo de expedientes judiciales.

Gelbukh, A. (2010). Procesamiento de lenguaje natural y sus aplicaciones. Komputer Sapiens.

Gómez-Moreno, Pedro Ureña. La lucha contra el terrorismo y la delincuencia organizada: Una visión desde la lingüística y la ingeniería del conocimiento. Miscelánea: A Journal of English and American Studies (2016).

Goyal, P., Pandey, S., Jain, K. (2018). Deep learning for natural language processing. New York: Apress.

Hastie, T., Tibshirani, R., Friedman, J. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hidalgo Tapia, Pedro. Algoritmos para el análisis de redes y grafos. Universidad de Sevilla (2024).

Jurafsky, D., Martin, J. H. (2023). Speech and Language Processing. Pearson.

LUNA, Issa: El análisis de redes complejas aplicado a grupos de crimen y corrupción: introducción y perspectiva. Polít. Crim. Vol. 17 Nº 34 (Diciembre 2022), Art. 7, pp. 611-634. [http:politcrim.comwp-contentuploads202211Vol17N34A7.pdf]

Odilla, F. Bots against corruption: Exploring the benefits and limitations of AI-based anti-corruption technology. Crime Law Soc Change 80, 353 -396 (2023).

Procuraduría General de la Nación. Informe final: Protocolo metodológico de Análisis de Redes Criminales en la Procuraduría General de la Nación. 2020.

Quispe Angulo, C. A. (2018). El expediente digital y su incidencia en la administración de justicia en el Perú.

Santos, R. B. (2016). Crime analysis with crime mapping. Sage publications.

Scikit-learn Developers. (2023). Scikit-learn: Machine Learning in Python.

Solano-Avella, J. (2021). Desarrollo e implementación del Expediente Digital en la administración pública en Colombia: Futuras concepciones para llegar a la virtualidad de los procesos judiciales y administrativas. Universidad Católica de Colombia. Disponible en: https:hdl.handle.net1098326103

Torres Berru, Y., López Batista, V.F., Torres-Carrión, P., Jimenez, M.G. (2020). Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review. In: Botto-Tobar, M., Zambrano Vizuete, M., Torres-Carrión, P., Montes León, S., Pizarro Vásquez, G., Durakovic, B. (eds) Applied Technologies. ICAT 2019. Communications in Computer and Information Science, vol 1194. Springer, Cham.

Torres, M. M. E., Manjarrés-Betancur, R. (2020). Asistente virtual académico utilizando tecnologías cognitivas de procesamiento de lenguaje natural. Revista Politécnica.

Tudela Poblete, Patricio. (2015). Análisis criminal, proactividad y desarrollo de estrategias policiales basadas en la evidencia. Revista Criminalidad, 57(1), 137-152. Retrieved November 24, 2024, from http:www.scielo.org.coscielo.php?script=sci_arttextpid=S1794-31082015000100010lng=entlng=es.

Tunstall, L., Von Werra, L., Wolf, T. (2022). Natural language processing with transformers. " O'Reilly Media, Inc.".

https:revistas.udistrital.edu.coindex.phptiaarticleview1732317210

https:rinfi.fi.mdp.edu.arbitstreamhandle123456789354GAlias-RCassanelli-TFG-II-2019.pdf?sequence=1isAllowed=y

Manual de Procedimientos de la Fiscalía General de la Nación