Carátula de versión pública		
PODER JUDICIAL DE LA FEDERACION SUPREMA CORTE DE JUSTICIA DE LA NACION	Fecha de clasificación	31 de agosto de 2021
	Área	Unidad General de Administración del Conocimiento Jurídico
	Documento	Informes mensuales de actividades del Contrato número SCJN/OM/DGRH-UGACJ-001/2020, prestación de servicios por honorarios asimilados a salarios de Omar Januario Castellanos Santa Cruz.
	Confidencial Reservado	• Firma de prestador de servicios • Direcciones IP de servidores internos
	Fundamento Legal	En términos de los previsto en los Artículos: 116 primer párrafo de la Ley General de Transparencia y Acceso a la Información Pública, 6 de la Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados, 113, fracción I de la Ley Federal de Transparencia y Acceso a la Información Pública, 110, fracción VII, de la Ley Federal de Transparencia y Acceso a la Información Pública, así como las resoluciones del Comité de Transparencia de la Suprema Corte de Justicia de la Nación: CT-CUM/A-36-2018, CT-CI/A-7-2021, CT-CUM/A-29-2018-III, CT-CUM/A-29-2018-II y CT-CI/A-7-2021.
	Firma del titular	Lic. Otilio Esteban Hernández Pérez Titular de la Unidad General de Administración del Conocimiento Jurídico

+IVFZo3qv+wIMgc1IH20F6AbGb/FwaaLzYea7LHtBlo=

México a 31 de julio de 2020

ING. AURELIO PEDRO VÁZQUEZ SÁNCHEZ
DIRECTOR DE ARQUITECTURA DE DATOS
UNIDAD GENERAL DE ADMINISTRACIÓN DEL CONOCIMIENTO JURIDICO
SUPREMA CORTE DE JUSTICIA DE LA NACIÓN
P R E S E N T E

Dando cumplimiento a las cláusulas del contrato celebrado con este Alto Tribunal, se presenta el siguiente reporte de productos generados en el periodo comprendido del 1 al 31 de julio de 2020, que forman parte de los proyectos en los que participo, consistentes en:

Proyecto: **Creación de modelos de administración de conocimiento y ciencia de datos para integración de líneas jurisprudenciales**

NLP en la construcción de líneas jurisprudenciales

- Revisión de la literatura sobre el tema de construcción de líneas jurisprudenciales, sentencias, precedentes y conceptos relacionados para la contextualización del problema.
- Se realizaron sesiones con el equipo de trabajo para definir el avance del proyecto, el enfoque y el alcance final del proyecto construcción de líneas jurisprudenciales.
- Análisis de los códigos Python que tiene relación con la construcción de modelos de procesamiento de lenguaje natural asociados con el proyecto construcción de líneas jurisprudenciales.
- Revisión, análisis y visto bueno de la arquitectura con la que se cuenta, así como el enfoque de ciencia de datos con el que se planteó el proyecto NLP para la construcción de líneas jurisprudenciales

Atentamente

[redacted]

MCC. Omar Januario Castellanos Santa Cruz

México, a 31 de agosto de 2020

ING. AURELIO PEDRO VÁZQUEZ SÁNCHEZ
DIRECTOR DE ARQUITECTURA DE DATOS
UNIDAD GENERAL DE ADMINISTRACIÓN DEL CONOCIMIENTO JURIDICO
SUPREMA CORTE DE JUSTICIA DE LA NACIÓN
P R E S E N T E

Dando cumplimiento a las cláusulas del contrato celebrado con este Alto Tribunal, se presenta el siguiente reporte de productos generados en el periodo comprendido del 1 al 31 de agosto de 2020, que forman parte de los proyectos en los que participo, consistentes en:

Proyecto: **Creación de modelos de administración de conocimiento y ciencia de datos para integración de líneas jurisprudenciales**

NLP en la construcción de líneas jurisprudenciales

- Creación de un modelo de *computer vision* utilizando *deep learning* para la detección y eliminación de notas al pie en documentos
- Obtención de los artículos citados dentro de los casos del cuadernillo 26 de la CIDH de forma automática utilizando NLP
- Configuración de ambiente GPU para Anaconda y CUDA en la maquina con la *ip* [redacted]
- Detección de notas al pie dentro de los párrafos de engrose y separación de los mismo para una mejor experiencia al utilizar el buscador jurídico para casos relacionados con líneas de jurisprudencia
- POC de viabilidad de conexión de Neo4j , Spark y el buscador jurídico para poder mostrar redes de vínculos de forma gráfica jurídico para casos relacionados con líneas de jurisprudencia

La evidencia de todos los cambios y desarrollos se listan a continuación:

Códigos

Ambiente	Servidor	Notebook
Desarrollo	[redacted]	Detección y etiquetado de notas al pie en engroses
Desarrollo	[redacted]	Análisis de tamaño de texto de notas el pie
Desarrollo	[redacted]	Eliminación de notas al pie en imágenes
Desarrollo	[redacted]	Detección de párrafos y segmentación
Desarrollo	[redacted]	Extracción de patrones de artículos citados en el cuadernillos 26 de la CIDH
Desarrollo	[redacted]	Python OCR en imágenes
Desarrollo	[redacted]	Pruebas de viabilidad de detección de párrafos por *computer vision* y *Deep learning*
Desarrollo	[redacted]	Modelo de detección y segmentación de párrafos creado en *pytorch*

Atentamente

[redacted]

MCC. Omar Januario Castellanos Santa Cruz

México, a 30 de septiembre de 2020

ING. AURELIO PEDRO VÁZQUEZ SÁNCHEZ
DIRECTOR DE ARQUITECTURA DE DATOS
UNIDAD GENERAL DE ADMINISTRACIÓN DEL CONOCIMIENTO JURIDICO
SUPREMA CORTE DE JUSTICIA DE LA NACIÓN
P R E S E N T E

Dando cumplimiento a las cláusulas del contrato celebrado con este Alto Tribunal, se presenta el siguiente reporte de productos generados en el periodo comprendido del 1 al 30 de septiembre de 2020, que forman parte de los proyectos en los que participo, consistentes en:

Proyecto: **Creación de modelos de administración de conocimiento y ciencia de datos para integración de líneas jurisprudenciales**

NLP en la construcción de líneas jurisprudenciales

- Para la especialización de un modelo NER sobre el contexto jurídico se realizó un análisis de las herramientas disponibles, una investigación de mejores prácticas, análisis del estado del arte y la evaluación de capacidades de datos internos. Posteriormente se implementaron modelos NER con data sets públicos con el objetivo de medir la complejidad que conlleva realizar dichos modelos desde cero.
- Se configuro la tarjeta gráfica NVIDIA y las bibliotecas de CUDA para la explotación de los *GPUs* en la máquina con el IP : [redacted]

- Se configuró el ecosistema *Hadoop* con *PySpark* sobre un sistema *CentOS* 8 y se instaló un base de datos basada en grafos (Neo4j) en la máquina con el IP : [redacted]

- Se creó una clase para solucionar el problema de similaridad entre números de expediente de engroses

La evidencia de todos los cambios y desarrollos se listan a continuación:

Códigos

Ambiente	Servidor	Notebook
Desarrollo	[redacted]	Funciones de similaridad y desimilaridad entre cadenas de texto
Desarrollo	[redacted]	Modelo NER en Sckitlearn utilizando CRF
Desarrollo	[redacted]	Pruebas de Entity Rules en Spark NLP
Desarrollo	[redacted]	Modelo NER utilizando Deep leanrning
Desarrollo	[redacted]	Modelo NER utilizando Spark y CRF
Desarrollo	[redacted]	Prueba de NER y Congnitive services con Spark
Desarrollo	[redacted]	Prueba de NER y análisis de sentimientos en Saprk
Desarrollo	[redacted]	Test de carga de archivos a Azure

[redacted]

MCC. Omar Januario Castellanos Santa Cruz

México a 31 de octubre de 2020

ING. AURELIO PEDRO VÁZQUEZ SÁNCHEZ
DIRECTOR DE ARQUITECTURA DE DATOS
UNIDAD GENERAL DE ADMINISTRACIÓN DEL CONOCIMIENTO JURIDICO
SUPREMA CORTE DE JUSTICIA DE LA NACIÓN
P R E S E N T E

Dando cumplimiento a las cláusulas del contrato celebrado con este Alto Tribunal, se presenta el siguiente reporte de productos generados en el periodo comprendido del 1 al 31 de octubre de 2020, que forman parte de los proyectos en los que participo, consistentes en:

Proyecto: **Creación de modelos de administración de conocimiento y ciencia de datos para integración de líneas jurisprudenciales**

Ingeniería de datos para el proyecto construcción de líneas jurisprudenciales

- Se inicia la instalación y despliegue de una herramienta de *data labeling (Prodigy)* para el etiquetado generalmente de un conjunto de datos, el cual está formado por párrafos de documentos jurídicos. Los párrafos fueron seleccionados de los diferentes tipos de documentos (engroses, tesis, etc.)
- Para el etiquetado inicial (exploratorio) del conjunto de datos utilizando *Prodigy* se generan con conjuntos de reglas provenientes de técnicas de área de análisis de lenguaje natural.
- Para general un posterior aprendizaje basado en *hypregrafos* se instala y despliega *Neo4J* utilizando las ventajas que ofrece las tecnología de contenedores (*Docker*). Posteriormente se plantea, analiza y definen las relaciones y reglas para la construcción de grafos. Dichos grafos representa las relaciones de citas entre los diferentes documentos jurídicos.

La evidencia de todos los cambios y desarrollos se listan a continuación:

Códigos

Ambiente	Servidor	Instancia
Desarrollo	[redacted]	Prodigy
Desarrollo	[redacted]	Neo4J

[redacted]

MCC. Omar Januario Castellanos Santa Cruz

México, a 30 de noviembre de 2020

ING. AURELIO PEDRO VÁZQUEZ SÁNCHEZ
DIRECTOR DE ARQUITECTURA DE DATOS
UNIDAD GENERAL DE ADMINISTRACIÓN DEL CONOCIMIENTO JURIDICO
SUPREMA CORTE DE JUSTICIA DE LA NACIÓN
P R E S E N T E

Dando cumplimiento a las cláusulas del contrato celebrado con este Alto Tribunal, se presenta el siguiente reporte de productos generados en el periodo comprendido del 1 al 30 de noviembre de 2020, que forman parte de los proyectos en los que participo, consistentes en:

Proyecto: **Creación de modelos de administración de conocimiento y ciencia de datos para integración de líneas jurisprudenciales**

Ingeniería de datos para el proyecto construcción de líneas jurisprudenciales

- Se inicia el análisis exploratorio de los párrafos que serán seleccionados para generar el *dataset* que será expuesto para su etiquetado en Prodigy.
- Posteriormente al análisis exploratorio, se generar los filtros y se crea el *dataset* (en formato *JSONL)* inicial para el etiquetado por parte de los expertos de negocio. Se realzan pruebas lógicas y pruebas de estrés para *Prodigy* con el *dataset* generado.
- Se generan más reglas para el etiquetado de las entidades y se refinan algunas existentes. Adicionalmente se restructuran algunos códigos de el *NER* basado en reglas existente para adaptarlo a la forma en que *Prodigy* lo requiere.

La evidencia de todos los cambios y desarrollos se listan a continuación:

Códigos

Ambiente	**Servidor**	**Instancia**
Desarrollo	[redacted]	Prodigy
Desarrollo	[redacted]	PySpark/Jupyter

[redacted]

MCC. Omar Januario Castellanos Santa Cruz

México, a 31 de diciembre de 2020

ING. AURELIO PEDRO VÁZQUEZ SÁNCHEZ
DIRECTOR DE ARQUITECTURA DE DATOS
UNIDAD GENERAL DE ADMINISTRACIÓN DEL CONOCIMIENTO JURIDICO
SUPREMA CORTE DE JUSTICIA DE LA NACIÓN
P R E S E N T E

Dando cumplimiento a las cláusulas del contrato celebrado con este Alto Tribunal, se presenta el siguiente reporte de productos generados en el periodo comprendido del 1 al 31 de diciembre de 2020, que forman parte de los proyectos en los que participo, consistentes en:

Proyecto: **Creación de modelos de administración de conocimiento y ciencia de datos para integración de líneas jurisprudenciales**

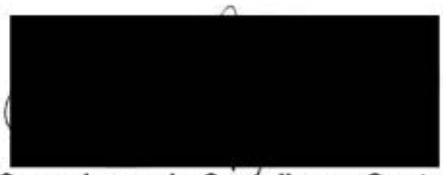
Etiquetado de párrafos para la generación de un NER de contexto jurídico

- De acuerdo al análisis de las reglas creadas para el etiquetado NER y los resultados de las pruebas del primer *dataset* jurídico generado utilizando *Prodigy* se decidió restructurar la forma en la que las diferentes clases interactuaban entre sí. Tomando en cuenta la escalabilidad del modelo NER se decidió unificar las diferentes clases utilizando el estándar de etiquetado de la biblioteca Spacy.
- Posteriormente de la restructuración y unificación de las clases se integró al *recipe* que utiliza Prodigy para la interface web de etiquetado.

La evidencia de todos los cambios y desarrollos se listan a continuación:

Códigos

Ambiente	**Servidor**	**Notebook**
Desarrollo	[redacted]	Prodigy



MCC. Omar Januario Castellanos Santa Cruz