**Objetivos del negocio**

|  |  |
| --- | --- |
| Título del proyecto de tesis: | Caracterización de recursos computacionales para la fase de preprocesamiento de minería de textos |
| Objetivo: | Medir los recursos computacionales como tiempo y memoria empleados para realizar tareas de preprocesamiento de documentos de texto, específicamente en la fase de preparación de datos mediante técnicas de minería de textos con la finalidad de caracterizar una dupla (una tarea y un recurso computacional) y de este modo, crear una instancia EC2 para aprovechar el cómputo en la nube. |
| Objetivos específicos: | ● Identificar y ejecutar tareas de pre-procesamiento de textos genéricas y especializadas con el fin de homogeneizar los documentos de leyes ambientales mexicanas (generación de un corpus legislativo) utilizando operadores de minería de textos proporcionados por las herramientas y/o lenguajes de programación específicos.  ● Comprender los objetivos del negocio (business undestanding) y de los datos (data understanding).  ● Monitorear los recursos (tiempo y memoria) a partir de diseños experimentales factoriales, utilizando scripts de Python para el preprocesamiento de los datos, es decir, se los 9 documentos de ley ambiental seleccionados.  ● Realizar pruebas escalables con el número de documentos y con las diferentes tareas de preprocesamiento de texto a realizar, con el fin de obtener la configuración personalizada del paquete de tareas a ejecutar en la capa gratuita de la máquina virtual de AWS. |

**1.1 Determinar los objetivos del negocio**

*1.1.1 Background*

* Registra la información que se conoce sobre la situación comercial de la organización al comienzo del proyecto.

**Caracterización de recursos computacionales para la fase de preprocesamiento de minería de textos**

Alumna

Yessenia Díaz Álvarez

Directora

Dra. Virginia Lagunes Barradas

Co-Director

Dr. Miguel Ángel Hidalgo Reyes

La Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas es una escuela de nivel superior y posgrado del Instituto Politécnico Nacional localizada en la Ciudad de México de México. Fue fundada en 1996.

**UPIITA-IPN**

Dra. Obdulia Pichardo

Dra. Bella Martínez

Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas IPN

Uno de los proyectos que han hecho es el proyecto Integralidad Gamma (Cuántico, 2020) que como lo mencionan (Pichardo et al., 2020) concibe el procesamiento de legislación ambiental como interrogar los datos. Estos datos giran en torno a diversas áreas: finanzas, educación, recreación, salud, e incluso ecología.

Los datos, por sí solos, no proporcionan la información requerida para la adecuada toma de decisiones, por lo que la mayoría de las veces tenemos que “interrogarlos para que confiesen”, pero ¿cómo interrogarlos para que confiesen? La respuesta a esta pregunta es la Ciencia de Datos.

Los datos tienen que sufrir todo un proceso para ser transformados en información que sea útil para el usuario final. La Ciencia de Datos (CD) es la disciplina que extrae conocimientos apoyándose de Estadística, Minería de Datos y Aprendizaje Automático a partir de una cantidad de datos. La CD tiene diversas subáreas de especialidad. Cuando los datos están almacenados en forma de texto se utilizan algoritmos, métodos y herramientas vinculadas con la Minería de Texto. Por ejemplo, la identificación automática de la temática de una nota periodística o un documento legislativo es una tarea de Minería de Texto.

Actualmente, el equipo de Ciencia de Datos, de la Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas del Instituto Politécnico Nacional (UPIITA-IPN) en vinculación con el Instituto de Ecología (Inecol), hace un análisis de documentos legislativos en materia ambiental. Se busca identificar automáticamente el nivel de coherencia de la jurisprudencia nacional implementando técnicas de minería automatizada de textos para el análisis de instrumentos legales aplicables.

Para llegar a este fin, se enfrentan varios retos. El primero de ellos se refiere a la representación de la información, ya que no todas las leyes, reglamentos, normas, planes, programas, estrategias, acuerdos y lineamientos siguen la misma estructura. Por lo que, los documentos deben ser procesados para obtener un estándar de representación legible por la computadora; por ello es recomendable que el sector gubernamental trabaje en un estándar estructural en la creación de documentos legislativos, como sucede actualmente en la generación de facturas, las cuales son legibles fácilmente por las computadoras.

En la estructura a evaluar para la marcación de coherencias legislativas, también se debe analizar la relación jerárquica que existe entre los diferentes documentos y dentro de ellos. Es decir, debe ser considerada la pirámide legislativa, además de la relación de los artículos en su respectivo capítulo.

Una vez que se logra la representación normalizada, se puede obtener la temática y términos asociados, apoyándose de algoritmos de clasificación, que hacen uso de técnicas de minería de textos y algoritmos de inteligencia artificial. La identificación de términos relevantes en un documento nos permite establecer la relación temática entre pares, con el fin de compararlos y establecer el nivel de coherencia entre las fuentes.

La identificación de la coherencia en el texto legislativo orientado al ámbito ecológico, no solo se basa en un análisis estructural del texto y obtención de términos, también se involucra la semántica.

Por tal motivo, se destacan algunos de los trabajos previos a esta investigación en los que se analizan textos orientados a dicho ámbito.

En primera instancia, el equipo de UPIITA-IPN (UPIITA-IPN, 2019), actualmente evalúa diversas técnicas empleadas para la detección de incoherencias, esto significa que busca la relación entre las definiciones de los términos en el contexto en que son usados. Por ejemplo, si hay dos documentos que regulan la contaminación ambiental, pero uno la define en el ámbito de aguas y otro en el de suelo, entonces las definiciones podrían ser diferentes y no mantener coherencia entre ellas en un contexto general de contaminación.

Por el momento, el trabajo se enfoca en el ámbito ambiental, pero el análisis que se lleva acabo podría ampliarse a diferentes campos. Cabe destacar que este trabajo se realiza bajo la supervisión y en colaboración de politólogos de la Universidad de Luxemburgo y especialistas en ecología del Instituto de Ecología del Estado de Veracruz, México (Inecol). Al finalizar, se espera que el analisis obtenido de los datos sirva a los expertos para identificar las oportunidades de mejora en la representación de documentos legislativos en el ámbito ambiental.

*1.1.2 Business Objectives:*

* Describe los objetivos primarios y secundarios de la empresa/institución (Indicar todos los que tenga la empresa).

**Identifica los objetivos primarios**

● Identificar y ejecutar tareas de pre-procesamiento de textos genéricas y especializadas con el fin de homogeneizar los documentos de leyes ambientales mexicanas (generación de un corpus legislativo) utilizando operadores de minería de textos proporcionados por las herramientas y/o lenguajes de programación específicos.

● Comprender los objetivos del negocio (business undestanding) y de los datos (data understanding).

● Monitorear los recursos (tiempo y memoria) a partir de diseños experimentales factoriales, utilizando scripts de Python para el preprocesamiento de los datos, es decir, se los 9 documentos de ley ambiental seleccionados.

● Realizar pruebas escalables con el número de documentos y con las diferentes tareas de preprocesamiento de texto a realizar, con el fin de obtener la configuración personalizada del paquete de tareas a ejecutar en la capa gratuita de la máquina virtual de AWS.

**Luego los objetivos secundarios**

* Identificación y análisis de bibliotecas que permitan medir recursos.
* Evaluación de tiempo y recursos en documentos utilizados para el preprocesamiento de documentos seleccionados.
* Descripción de las herramientas de monitoreo utilizadas.
* Análisis de los resultados obtenidos en cada uno de los casos de pruebas.

De este modo, un usuario tendrá mayor control o conocimiento sobre las horas-*core* que demanda una tarea y así evitaría una experimentación incierta en una máquina virtual que ocasione costos no deseados.