**Reporte selección de datos**

**Business understanding**

|  |  |
| --- | --- |
| 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.

*1.1.3 Business Success Criteria:*

* Describa los criterios para un resultado exitoso o útil del proyecto desde el punto de vista de negocio. Puede ser bastante específico y medible objetivamente, o puede ser general y subjetivo, (en este último caso, debe indicarse quién hace el juicio subjetivo).

La problemática a resolver mediante este proyecto, radica en aprovechar las funcionalidades que proveen tanto herramientas de software como lenguajes de programación como Python, para el análisis de diversos documentos de texto, enfocándose en la fase de preparación de datos.

Dichas funcionalidades se aplican a documentos que tratan sobre legislación ambiental en México, los cuales, aunque vienen en formato pdf, internamente presentan una estructura diversa, por lo que su análisis resulta más desafiante.

En este sentido, además de ejecutar tareas genéricas y específicas de preparación de los documentos, en este caso en los textos ambientales seleccionados como corpus, se ofrezcan datos cuantitativos acerca de la memoria y el tiempo transcurridos para las distintas configuraciones de tareas de preprocesamiento aplicadas, bajo una infraestructura de hardware determinadas.

El objetivo es 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.

**1.2 Evaluar la situación**

1.2.1 *Inventory of Resources Requirements*

Enumera todos los recursos disponibles para el proyecto:

* Identificar los stakeholders vinculados con los datos de interés y los stakeholders que toman decisiones.
* Determinar si hay perfiles especializados para la gobernanza de datos.
* Recursos tecnológicos (software y hardware)
* Suposiciones y restricciones: Enumerar todos los requisitos del proyecto, incluido el calendario de finalización, la comprensión y la calidad de resultados y seguridad, así como cuestiones legales. Como parte de este resultado, asegúrese de poder utilizar los datos.

La legislación ambiental mexicana presenta formatos y contenidos que no se apegan a una estructura determinada, por lo que su análisis resulta más desafiante. Si bien la minería de textos utiliza técnicas y algoritmos para preprocesar documentos, no siempre se está consciente del consumo de recursos computacionales requeridos para realizar una o más tareas específicas. En este sentido, resulta de interés que además de ejecutar tareas genéricas y específicas de preparación de los documentos, en este caso en los textos ambientales seleccionados como corpus, se ofrezcan datos cuantitativos acerca de la memoria y tiempo transcurridos para las distintas configuraciones de tareas de preprocesamiento aplicadas, bajo una infraestructura de *hardware* determinadas.

**Alcances**

* Se abordarán las tres primeras fases de CRISP-DM mediante las tareas de pre-procesamiento de datos las cuales se agrupan en genéricas como comprensión del negocio, comprensión de los datos y preparación de los datos; en especializadas como transformación pdf a texto, convertir texto plano a json, identificación de identidades y etiquetar los textos de la fase de preparación de datos que se va a utilizar.
* Utilizar Python con sus librerías, como herramienta de minería de textos y/o preprocesamiento.
* Analizar 9 leyes ambientales mexicanas de ámbito federal, las leyes a utilizar en el diseño experimental serán:
  + GEEPA: Ley General del Equilibrio Ecológico y la Protección al Ambiente
  + LAN: Ley de Aguas Nacionales
  + LDRS: Ley de Desarrollo Rural Sustentable
  + LGPAS: Ley General de Pesca y Acuacultura Sustentable
  + LGVS: Ley General de Vida Silvestre
  + LGDFS: Ley General de Desarrollo Forestal Sustentable
  + LGPGIR: Ley General Para la Prevención y Gestión Integral de Residuos
  + LFBOGB: Ley Federal de Bioseguridad de Organismos Genéticamente Modificados
  + LGCC: Ley General de Cambio Climático
* Los recursos computacionales de interés son el tiempo y memoria RAM.
* Las características técnicas de la computadora donde se monitorearán los recursos (tiempo y memoria) cuenta con (Memoria RAM de 3 GB. Disco duro de 500 GB. Procesador AMD 1.60 Ghz x64. Windows 8.1 Pro.). Las versiones utilizadas en la ejecución de la tarea de preprocesamiento (extracción de texto del archivo pdf) son Python 3.10, pdfplumber 0.1.2, pypdf2 2.10.3.

**Limitaciones**

* Limitación en la instancia EC2 de AWS pues que se trata de la capa gratuita únicamente.

**Recursos**

* Lista de recursos computacionales
  + Inventario (CC y Personal)
* Recursos humanos
  + Director del proyecto, codirector del proyecto con el expertise
* Software
  + Python y bibliotecas, Github (PyPDF2, pdfplumber)
* Listado de restricciones
  + 9 leyes
* Recursos computacionales
  + Tiempo y memoria

1.2.2 *Risks and Contingencies*

* Enumera los riesgos o eventos que podrían retrasar el proyecto o hacer que falle.
* Si existen contingencias correspondiente a planes, ¿qué acciones se tomarán si estos riesgos o eventos ocurren?

Tabla 2. Análisis FODA.

|  |  |
| --- | --- |
| **Fortalezas** | **Debilidades** |
| Existencia de gran variedad de normas, leyes para la experimentación del proyecto.  Gran cantidad de librerías para el proyecto | Librerías no adecuadas para el proyecto  Tiempo para la formación en el área de minería de textos y en el uso de sus herramientas y lenguajes.  No contar con un lugar adecuado de estancia para realizar pruebas |

|  |  |
| --- | --- |
| **Oportunidades** | **Amenazas** |
| Aprovechamiento del centro de cómputo de la MPSCO del ITSX | Número insuficiente de pruebas para producir resultados correctos, generales, informativos y útiles  Equipo insuficiente para realizar dichas pruebas |

1.2.3 *Terminology* (Glosario)

* Sobre el negocio.

Text Mining es el término en idioma inglés, que generalmente se refiere a la extracción automática de información interesante y no trivial de un texto no estructurado. Generalmente, su propósito no radica en comprender todo o parte de lo que dice un hablante/escritor en particular, sino más bien en extraer patrones de una gran cantidad de documentos. La minería de textos está relacionada con el Procesamiento del Lenguaje Natural (PNL), que incluye técnicas de inspiración lingüística, es decir, un texto se analiza típicamente desde un punto de vista léxico y sintáctico utilizando una gramática formal, la información resultante se interpreta semánticamente y se utiliza para extraer información sobre lo dicho (Kao & Poteet, 2007).

Tabla 3. Evolución de técnicas de minería de textos (Tandel et al., 2019).

|  |  |
| --- | --- |
| Extracción de información | La etapa principal para los ordenadores es reconocer la tipografía amorfa mediante el reconocimiento de frases y tratos importantes dentro del texto. Su objetivo es extraer información significativa de grandes trozos de texto. La información extraída se conserva en forma de registro para su posterior acceso o recuperación. |
| Recuperación de información | El campo de la recuperación de información ha estado en construcción con los sistemas de bases de datos durante más tiempo. El objetivo de la recuperación de información es obtener el documento con información precisa recuperada por el usuario. Por lo tanto, la recuperación del documento es seguida por una etapa llamada minería de texto que se centra principalmente en la solicitud enviada por el operador o es seguida por la etapa de extracción de información que utiliza técnicas de extracción de información. |
| Resumir | Se toma el texto en bruto y se realizan en él operaciones de preprocesamiento y procesamiento. En el preprocesamiento, se aplican tres métodos: la tokenización, el stemming y la eliminación de palabras. |
| *Clustering* | El proceso de seccionar un grupo de objetos o datos en una colección de subclases relevantes y comprensibles se denomina clustering. El clustering se utiliza principalmente para hacer un conjunto de documentos y archivos similares. |
| Categorización | En la categorización se reconocen los temas importantes de un documento. Esto se hace asignando los documentos a un conjunto de temas predefinidos. El documento categorizado puede tratarse como una "bolsa de palabras". La extracción de información trata de procesar la información real, mientras que la categorización no trata de procesar la información real. |

Preprocesamiento: Se refiere a la realización de operaciones o transformaciones sobre el texto en algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis. Esta fase se lleva a cabo sobre un conjunto de documentos objetos de estudio determinando el tipo de representación o patrones contenidos en los textos (Taeho, 2019).

Transformación pdf a texto: Los textos se transforman en representación estructurada o semiestructurada que facilite su posterior análisis. Se define el conjunto corpus de documentos. Estos documentos deben ser representativos y seleccionarse aleatoriamente o mediante algún método de muestreo probabilístico. Se debe evitar, en esta etapa, la duplicación de documentos dentro del corpus (Cortez, 2018).

Eliminación de ruido: La eliminación de palabras vacías se refiere al proceso de eliminar palabras vacías de la lista de tokens o palabras derivadas (Kowalski & Maybury, 2000). Las palabras de parada son las palabras gramaticales que son irrelevantes para el contenido del texto, por lo que deben eliminarse para una mayor eficiencia. La palabra clave de la lista se carga desde un archivo y, si están registrados en la lista, se eliminan.

Tokenización: La tokenización se define como el proceso de segmentar un texto o textos en tokens por el espacio en blanco o los signos de puntuación. Es capaz de aplicar la tokenización al códigos fuente en C, C ++ y Java (Aho et al., 2007) así como los textos que están escritos en un lenguaje natural. Sin embargo, el alcance está restringido al texto en este estudio, a pesar de la posibilidad.

Etiquetado de textos: Se conoce como derivación es una técnica de preprocesamiento que se ocupa de reducir una palabra a su forma básica, llamada raíz. En este contexto, no es importante que la palabra elegida sea una sola palabra, sino que las variantes morfológicas del mismo término, que en la mayoría de los casos tienen interpretaciones semánticas similares, se asignen todas a la misma raíz (Kowalski & Maybury, 2000).

CRISP-DM (Cross Industry Standard Process for Data Mining), el cual es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema (Chapman et al., 2000).

* Acerca de la tecnología utilizada o a utilizar.

Python está desarrollado bajo una licencia de código abierto aprobada por OSI (modelo de interconexión de sistemas abiertos) lo que lo hace de libre uso y distribución, incluso para uso comercial. La licencia es administrada por Python Software Foundation (Python, 2021).

NLTK (Natural Language Toolkit, de sus siglas en inglés) es una plataforma líder para crear programas de Python para trabajar con datos de lenguaje humano. Proporciona interfaces fáciles de usar para más de 50 corpus y recursos léxicos como WordNet, junto con un conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, lematización, etiquetado, análisis y razonamiento semántico, contenedores para bibliotecas NLP de potencia industrial, y un foro de discusión activo (NLTK, 2021).

RapidMiner lleva la inteligencia artificial a la empresa a través de una plataforma de ciencia de datos abierta y extensible. Creado para equipos de análisis, Unifica todo el ciclo de vida de la ciencia de datos, desde la preparación de datos hasta el aprendizaje automático y  la implementación de modelos predictivos. Más de 700,000 profesionales de análisis utilizan productos RapidMiner para generar ingresos, reducir costos y evitar riesgos (RapidMiner, 2021).

1.2.4 *Costos y beneficios*

* Construye un análisis de costo-beneficio para el proyecto, que compare los costos del proyecto con el potencial beneficio para la empresa si tiene éxito.

Aún no se tiene este punto

1.2.1 *Inventory of Resources Requirements*

Enumera todos los recursos disponibles para el proyecto:

Identificar los stakeholders vinculados con los datos de interés y los stakeholders

Aún no se tiene este punto

**2.1 Recoger los datos iniciales**

2.1.1 Adquirir los datos (o acceder a los datos) que figuran en los recursos del proyecto. Esta recopilación inicial incluye la carga de datos

Lista el conjunto o conjuntos de datos adquiridos, junto con su ubicación, los métodos utilizados para adquirirlos y cualquier problema encontrado junto con las soluciones alcanzadas.

Las leyes a analizar son 9 leyes ambientales mexicanas de ámbito federal, las cuales, “establecen las normas para la conservación, protección, mejoramiento y restauración del medio ambiente y los recursos naturales que lo integran” (Mejía, 2000).

Las leyes ambientales seleccionadas para el presente proyecto se encuentran en formato pdf todas disponibles en el sitio web: https://www.diputados.gob.mx

* + GEEPA: Ley General del Equilibrio Ecológico y la Protección al Ambiente
  + LAN: Ley de Aguas Nacionales
  + LDRS: Ley de Desarrollo Rural Sustentable
  + LGPAS: Ley General de Pesca y Acuacultura Sustentable
  + LGVS: Ley General de Vida Silvestre
  + LGDFS: Ley General de Desarrollo Forestal Sustentable
  + LGPGIR: Ley General Para la Prevención y Gestión Integral de Residuos
  + LFBOGB: Ley Federal de Bioseguridad de Organismos Genéticamente Modificados
  + LGCC: Ley General de Cambio Climático

**2.2 Descripción de los datos**

2.2.1 Describe los datos que se han adquirido (o generado), incluyendo el formato de los datos, la cantidad de datos (por ejemplo, el número de registros y campos de cada tabla), las identidades de los campos y cualquier otra característica que se haya descubierto.

Formato: Todas las leyes están en formato PDF

Cantidad de datos: 9 leyes ambientales mexicanas

Las leyes ambientales seleccionadas para el presente proyecto se encuentran en formato pdf todas disponibles en el sitio web: https://www.diputados.gob.mx

Tabla 1. Leyes ambientales con descripción y formato.

|  |  |
| --- | --- |
| **Ley Ambiental** | **Descripción** |
| GEEPA: Ley General del Equilibrio Ecológico y la Protección al Ambiente | Preservación y restauración del equilibrio ecológico, así como a la protección al ambiente, en el territorio nacional y las zonas sobre las que la nación ejerce su soberanía y jurisdicción (Congreso general de los estados Unidos Mexicanos, 2015). |
| LAN: Ley de Aguas Nacionales | En materia de aguas nacionales; es de observancia general en todo el territorio nacional, sus disposiciones son de orden público e interés social y tiene por objeto regular la explotación, uso o aprovechamiento de dichas aguas, su distribución y control, así como la preservación de su cantidad y calidad para lograr su desarrollo integral sustentable (Congreso general de los estados Unidos Mexicanos, 2020a). |
| LDRS: Ley de Desarrollo Rural Sustentable | Integrar una política de Estado para el desarrollo rural, por encima de las naturales diferencias entre las fuerzas políticas, capaz de construir acuerdos en puntos básicos que garanticen metas y programas en el largo plazo, creadora de seguridad, confianza y certidumbre; como una de las principales aspiraciones de los productores y sus organizaciones (Congreso general de los estados Unidos Mexicanos, 2007). |
| LGPAS: Ley General de Pesca y Acuacultura Sustentable | Ordenar, fomentar y regular el manejo integral y el aprovechamiento sustentable de la pesca y la acuacultura, considerando los aspectos sociales, tecnológicos, productivos, biológicos y ambientales (Congreso general de los estados Unidos Mexicanos, 2018a). |
| LGVS: Ley General de Vida Silvestre | El aprovechamiento sustentable de los recursos forestales maderables y no maderables y de las especies cuyo medio de vida total sea el agua, será regulado por las leyes forestales y de pesca, respectivamente, salvo que se trate de especies o poblaciones en riesgo (Congreso general de los estados Unidos Mexicanos, 2018b). |
| LGDFS: Ley General de Desarrollo Forestal Sustentable | Promover la legalidad en las actividades productivas, mejorar la capacidad de transformación e integración industrial, impulsar la comercialización y fortalecer la organización de redes locales de valor y cadenas productivas del sector forestal (Mexicanos, 2018). |
| LGPGIR: Ley General Para la Prevención y Gestión Integral de Residuos | Garantizar el derecho de toda persona al medio ambiente sano y propiciar el desarrollo sustentable a través de la prevención de la generación, la valorización y la gestión integral de los residuos peligrosos, de los residuos sólidos urbanos y de manejo especial; prevenir la contaminación de sitios con estos residuos (Mexicanos, 2021). |
| LFBOGB: Ley Federal de Bioseguridad de Organismos Genéticamente Modificados | Regular las actividades de utilización confinada, liberación experimental, liberación en programa piloto, liberación comercial, comercialización, importación y exportación de organismos genéticamente modificados, con el fin de prevenir, evitar o reducir los posibles riesgos que estas actividades pudieran ocasionar a la salud humana o al medio ambiente y a la diversidad biológica o a la sanidad animal, vegetal y acuícola (Mexicanos, 2020). |
| LGCC: Ley General De Cambio Climático | Garantizar el derecho a un medio ambiente sano y establecer la concurrencia de facultades de la federación, las entidades federativas y los municipios en la elaboración y aplicación de políticas públicas para la adaptación al cambio climático y la mitigación de emisiones de gases y compuestos de efecto invernadero (Congreso general de los estados Unidos Mexicanos, 2020b). |

Como se sabe cada ley tiene una estructura, la cual se verá en la siguiente figura 1:

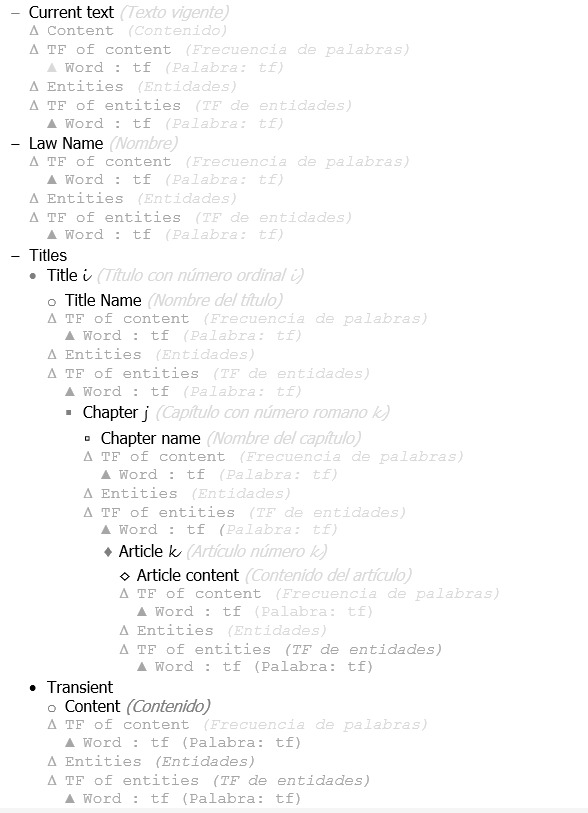
****

Figura 1. Estructura general de las leyes

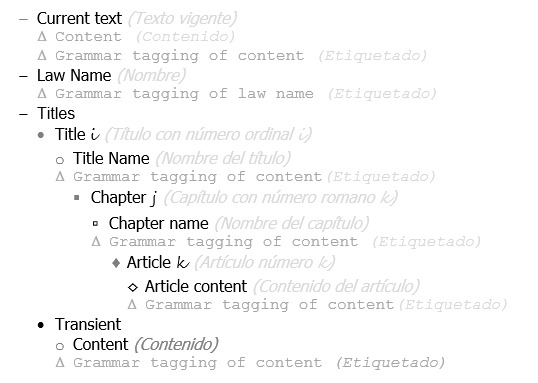
****

Figura 2. Estructura específica de las leyes

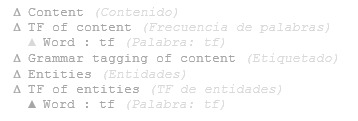
****

Figura 3. Contenido importante de las leyes

**2.3 Exploración de los datos**

2.3.1 Analiza mediante técnicas de consulta, visualización y elaboración de informes estadísticos. Estos incluyen la distribución de atributos clave (por ejemplo, el atributo objetivo de una tarea de predicción), las relaciones o correlaciones entre atributos.

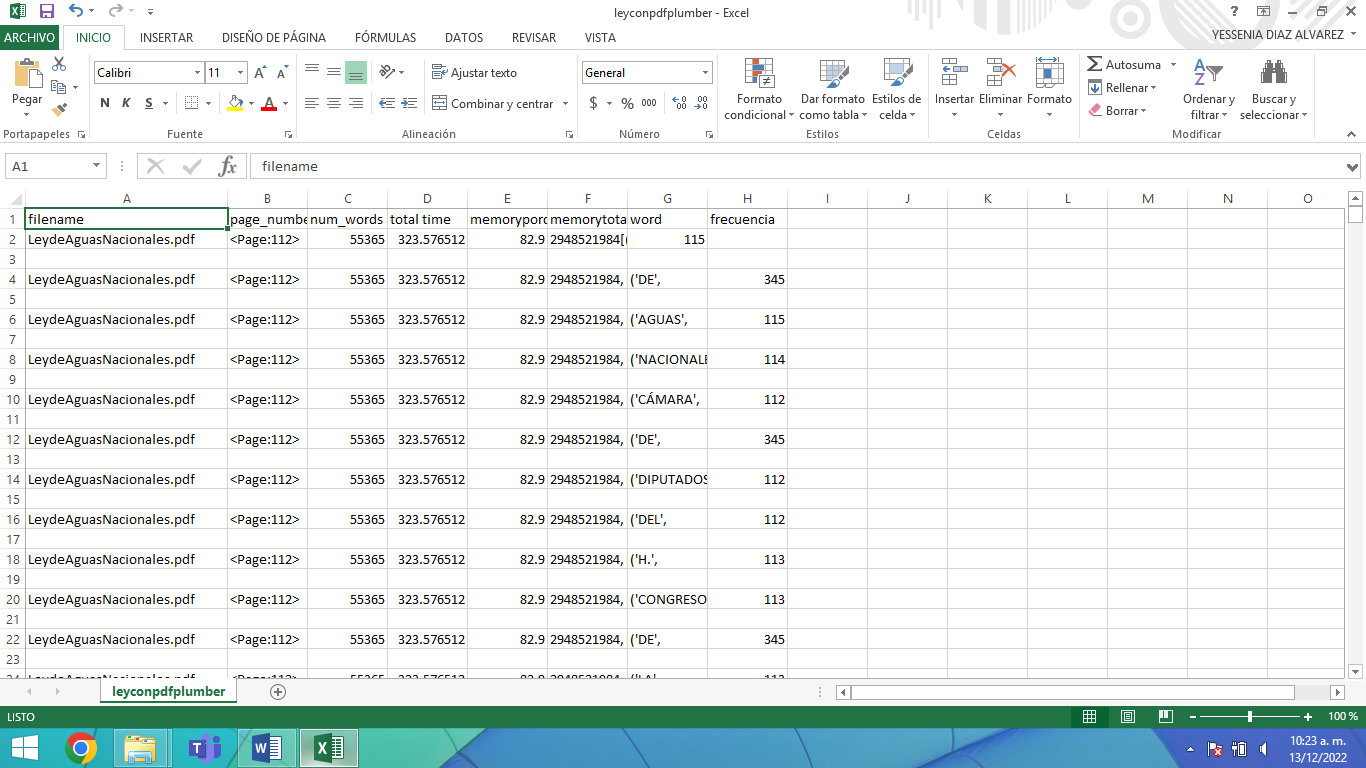
A continuación se muestra la siguiente tabla en pdfplumber

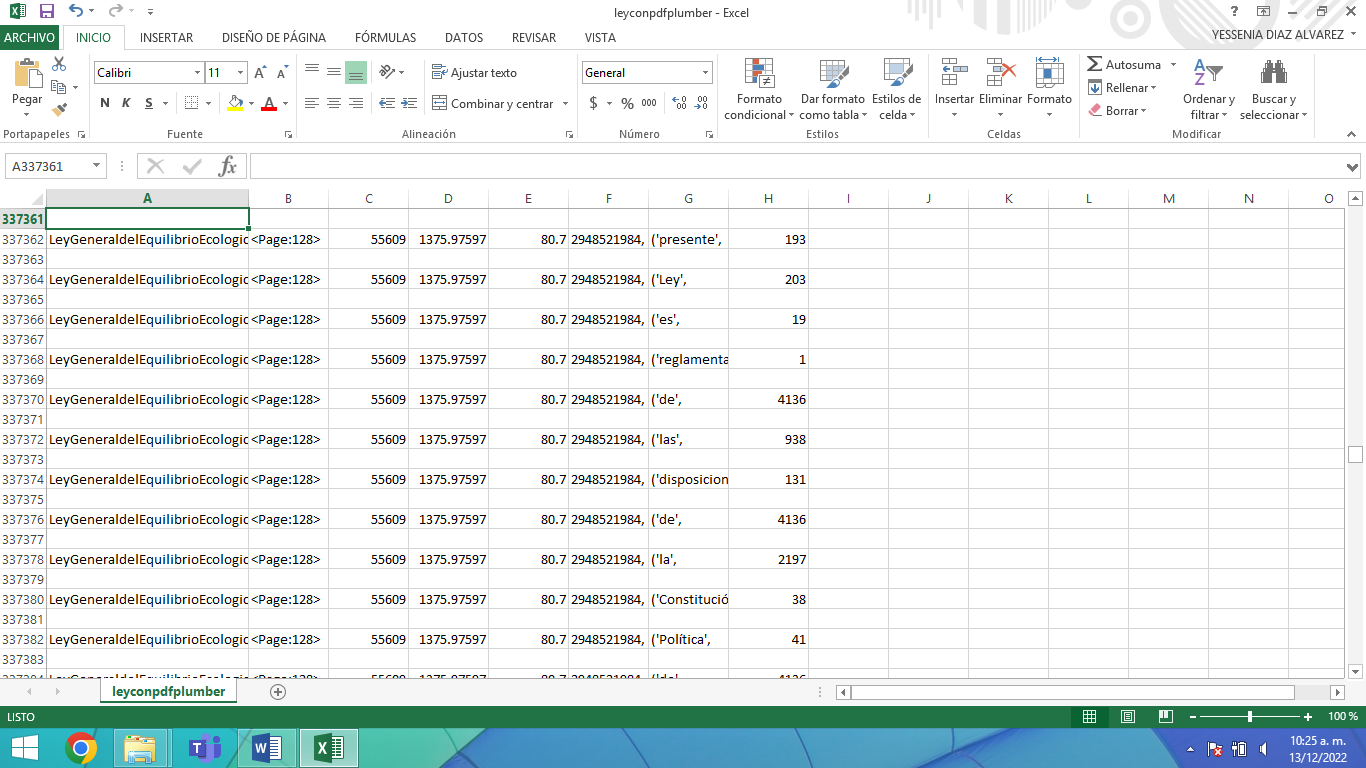
Se crea un csv con el nombre de la ley “Ley de Aguas Nacionales” y “Ley General del Equilibrio Ecológico y la Protección al Ambiente” en este caso

La primera ley consta de 112 páginas, tiene un total de 55365 palabras, el total del tiempo es de 323.576512 milisegundos y memoria virtual es de 82.9% bytes, la memoria RAM total de la computadora de 2.94 gb, se muestra el nombre de la palabra y la frecuencia en la que aparece cada una de ellas en esta ley.

La segunda ley consta de 128 páginas, tiene un total de 55609 palabras, el total del tiempo es de 1375.97597 milisegundos y memoria virtual es de 80.7% bytes, la memoria RAM total de la computadora de 2.94 gb, se muestra el nombre de la palabra y la frecuencia en la que aparece cada una de ellas en esta ley.

Tabla 5. Csv de leyes ambientales.





**2.4 Verificar la calidad de los datos**

2.4.1 ¿Están los datos completos (cubren todos los casos necesarios)?

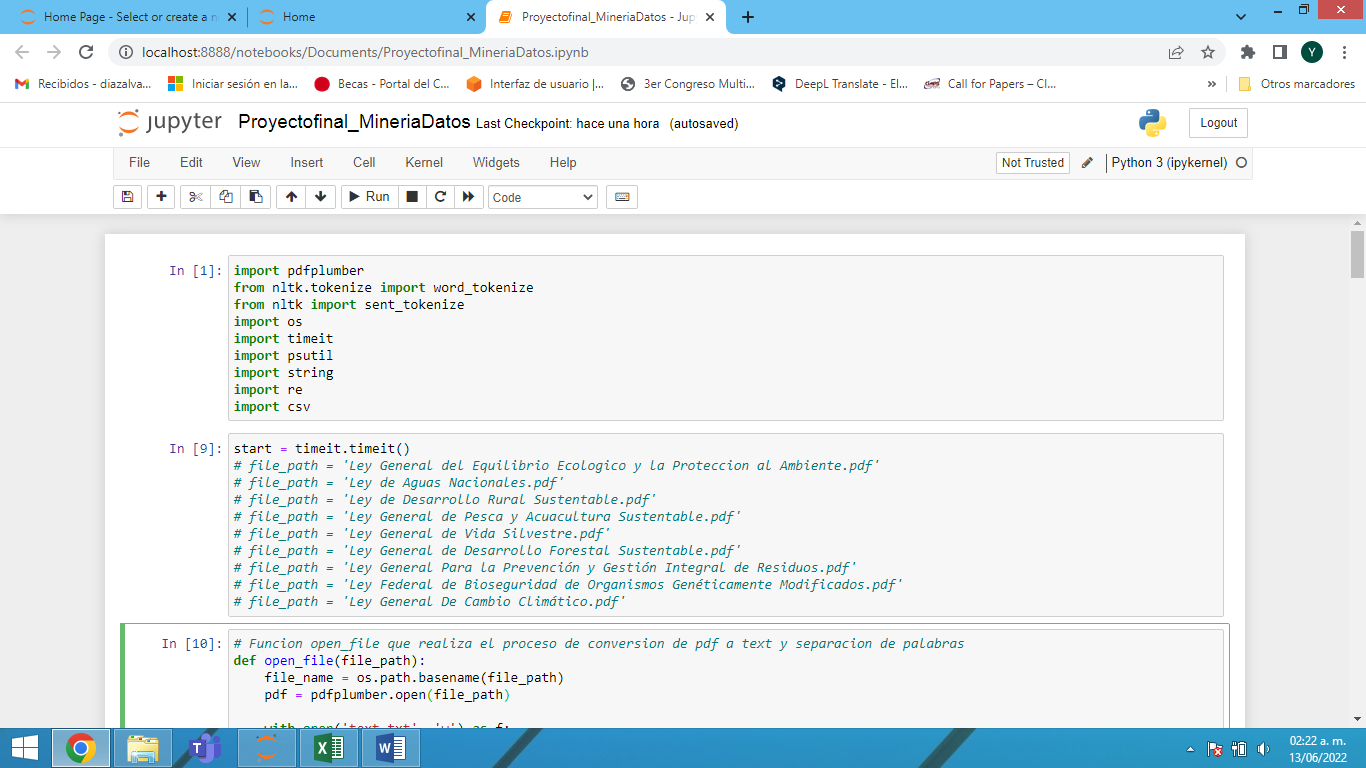
Si

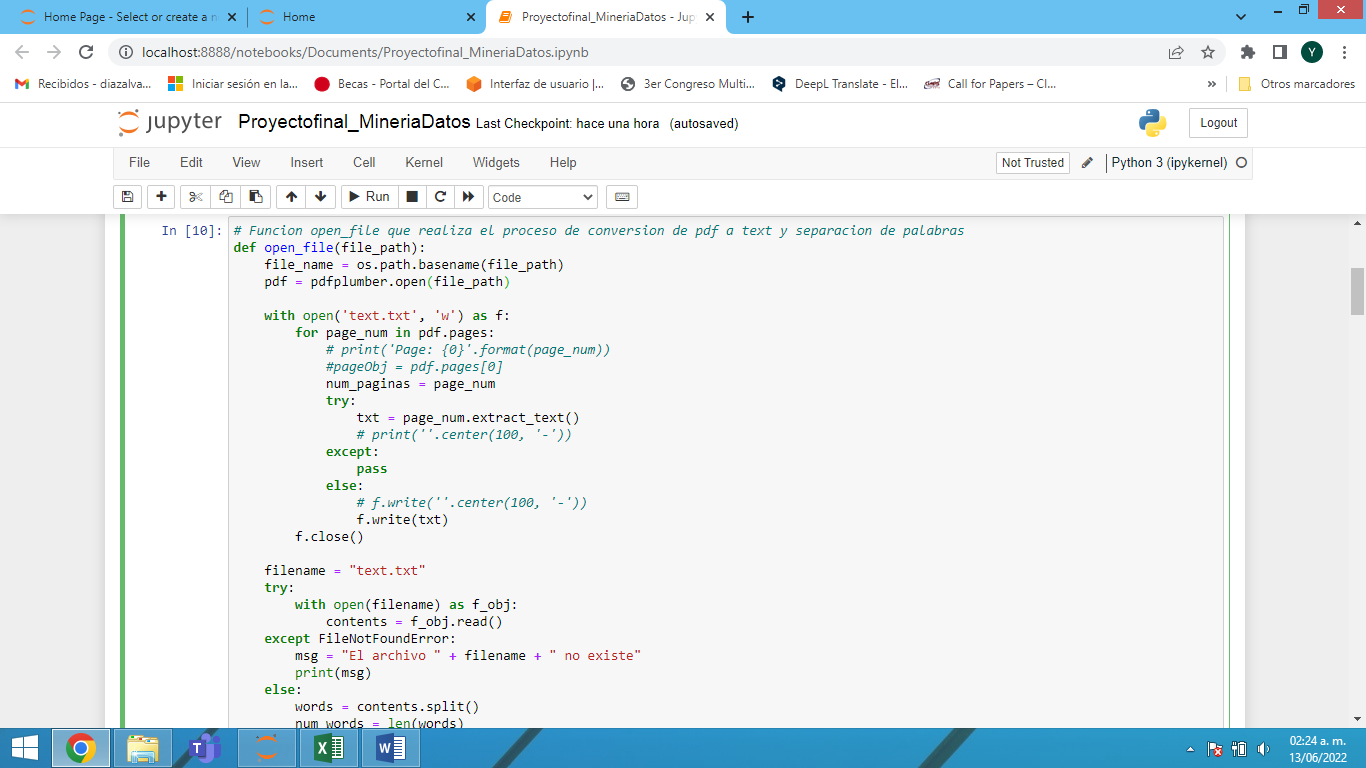
¿Son correctos o contienen errores y, si los hay, cuán comunes son?

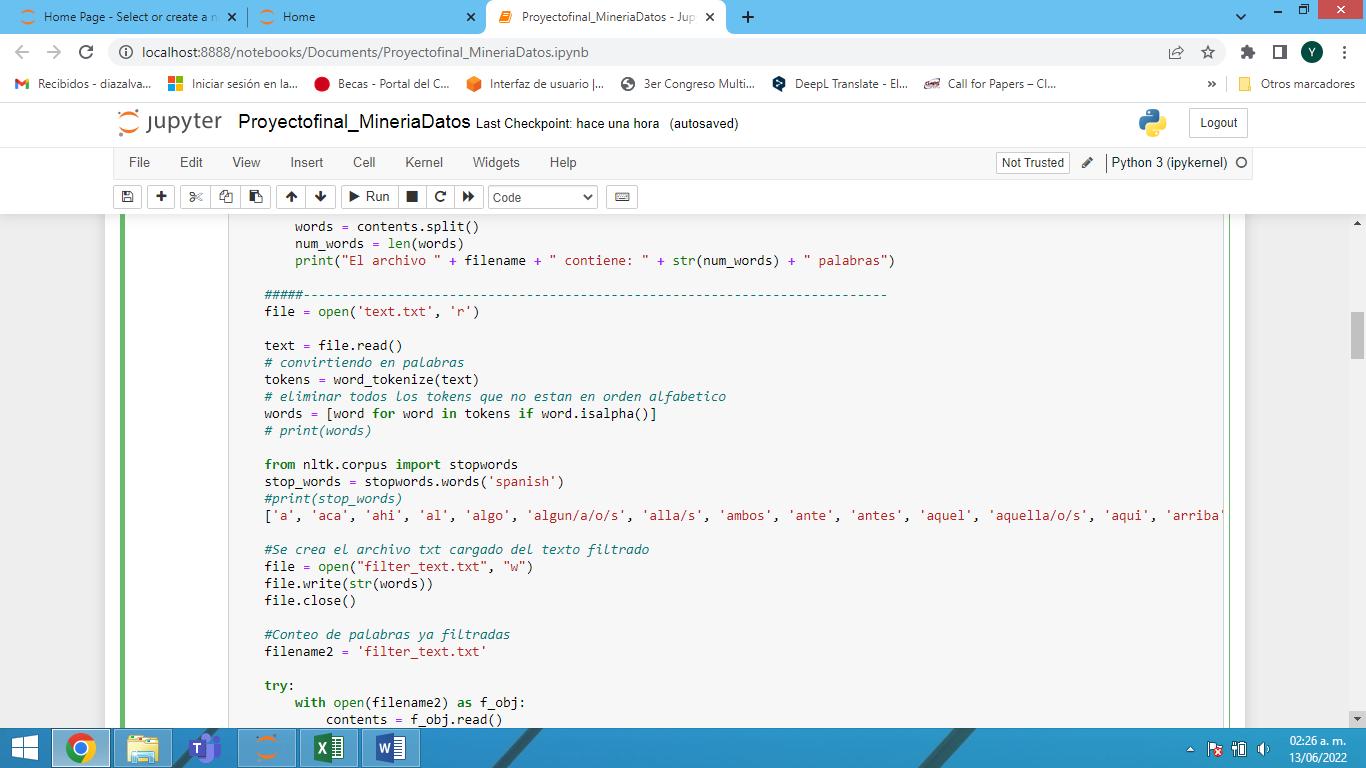
Estos datos contienen errores ya que cuentan con espacios y stopwords que hay que limpiar

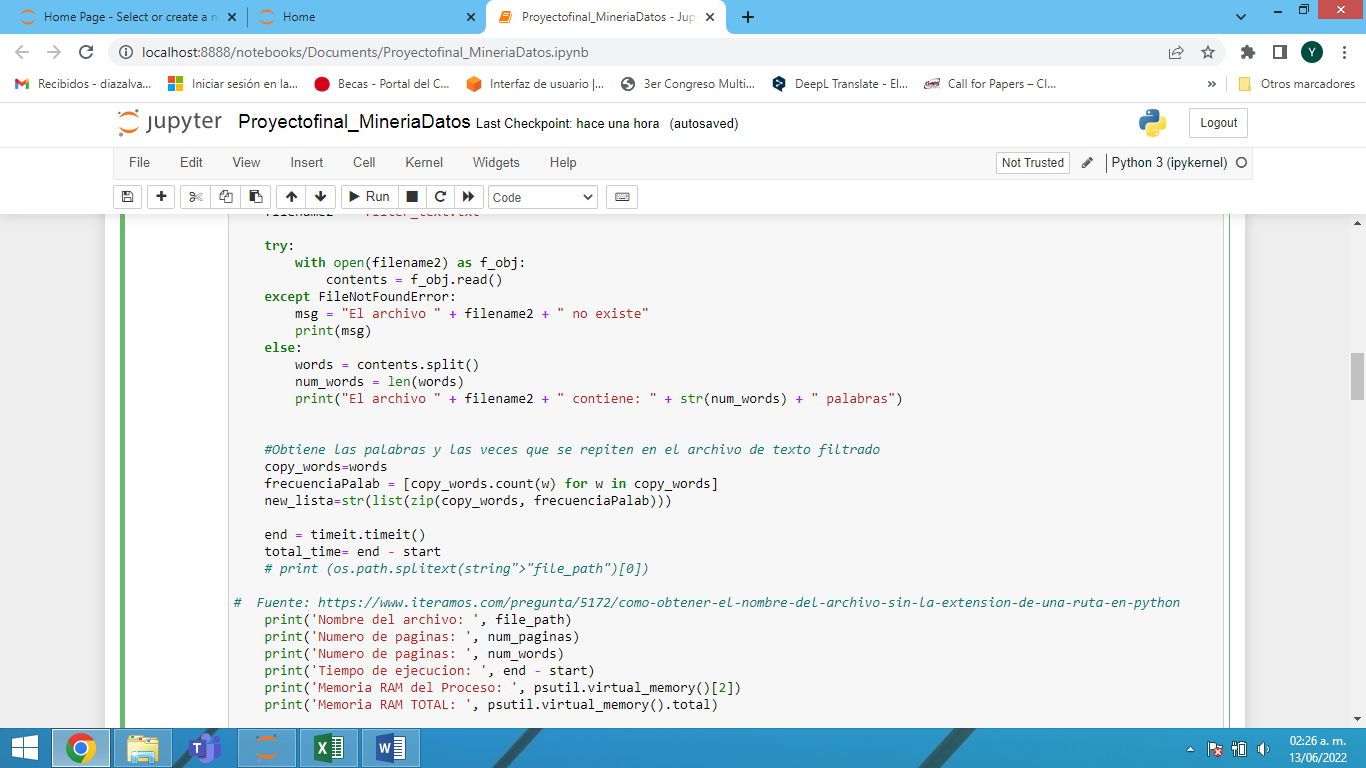
¿Hay valores que faltan en los datos? No

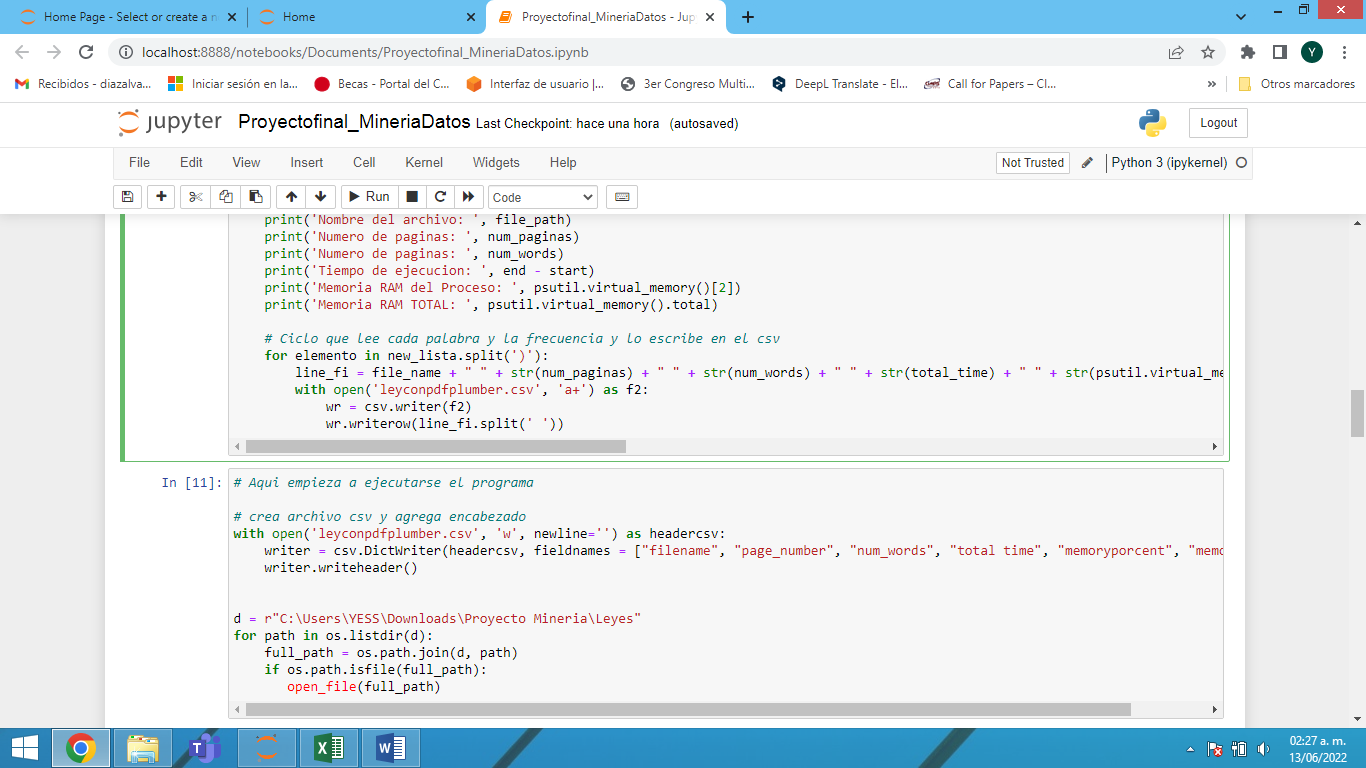
Si es así, ¿cómo se representan, dónde se producen y qué frecuencia tienen?

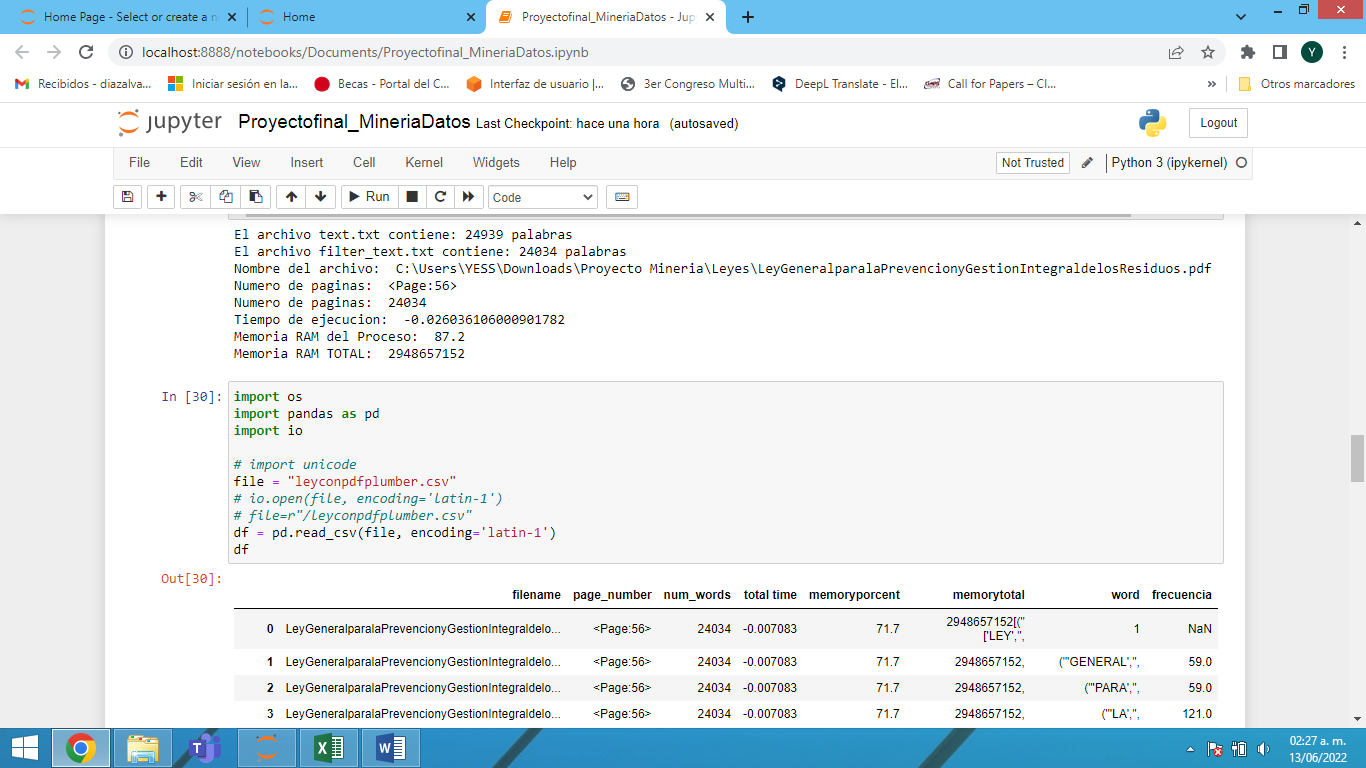












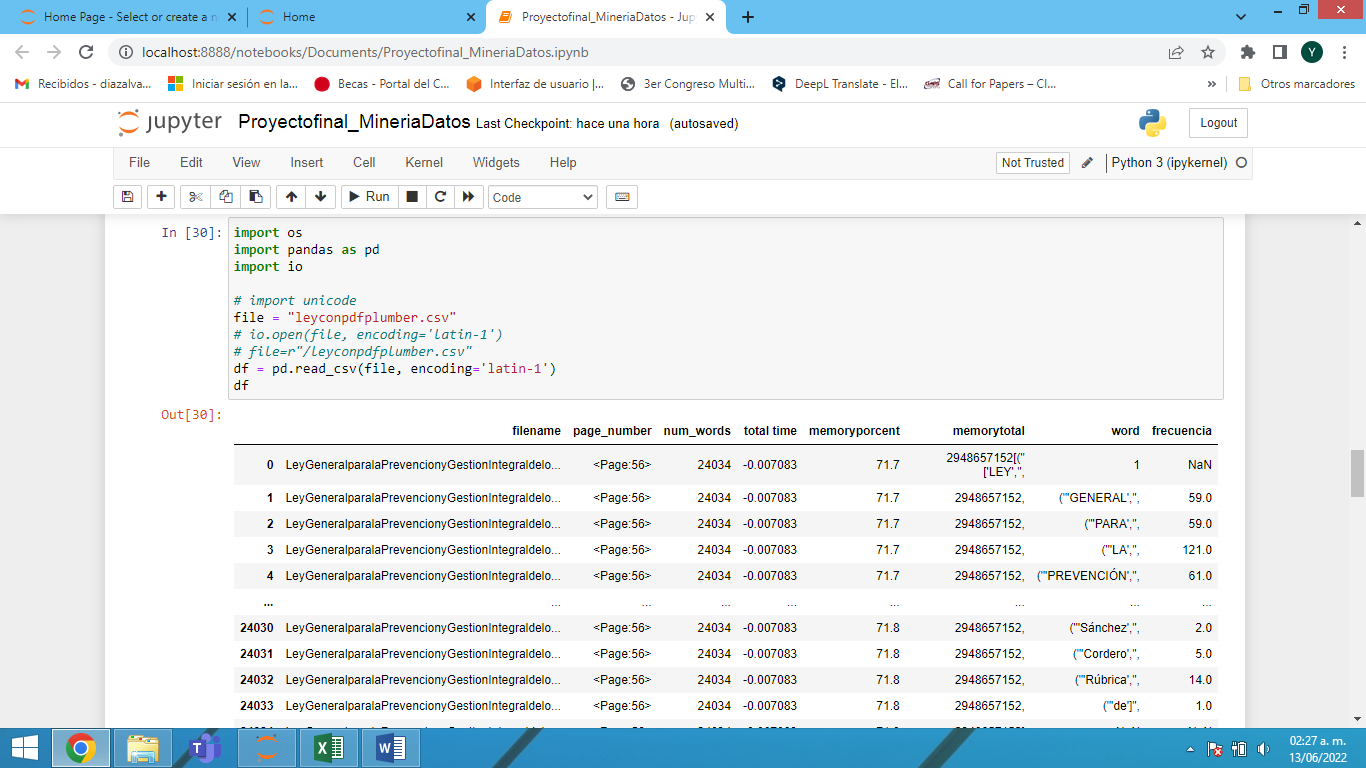
Como se muestra en el código se importa la librería pdfplumber, la cual se seleccionó ya que anteriormente se hicieron experimentaciones con las anteriores librerías y no daba los resultados que se esperaban. Se abre un archivo el cual llamado text.txt el cual se transforma de pdf de la ley a txt.

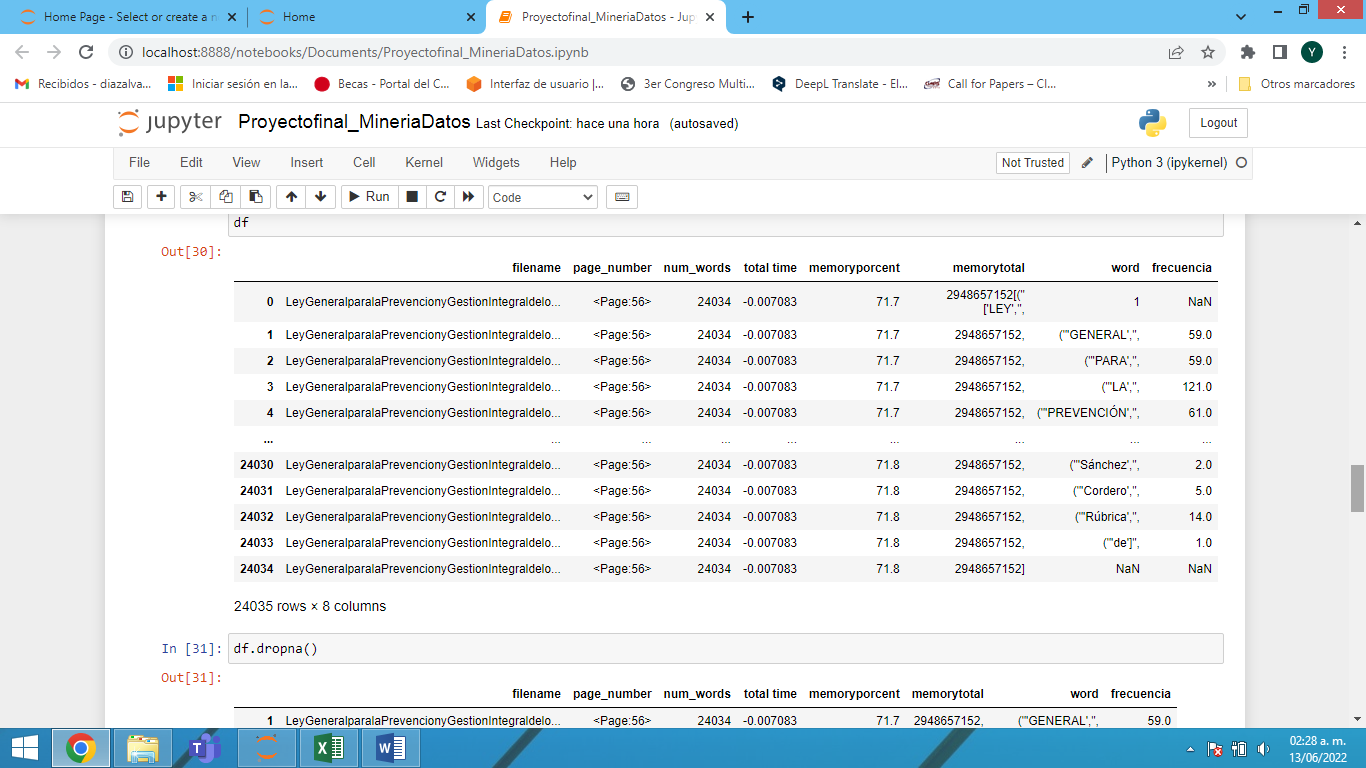
**3.1 Seleccionar los datos**

Decidir qué datos se van a utilizar para el análisis. Los criterios incluyen la relevancia para los objetivos de la minería de datos, la calidad y restricciones técnicas, como los límites del volumen o los tipos de datos.

(La selección de atributos (columnas), así como la selección de registros (filas) en una tabla).

Se abre el archivo csv que se tiene en el cual se muestra la tabla de los datos que se seleccionaran para tener una mayor visualización de los datos.



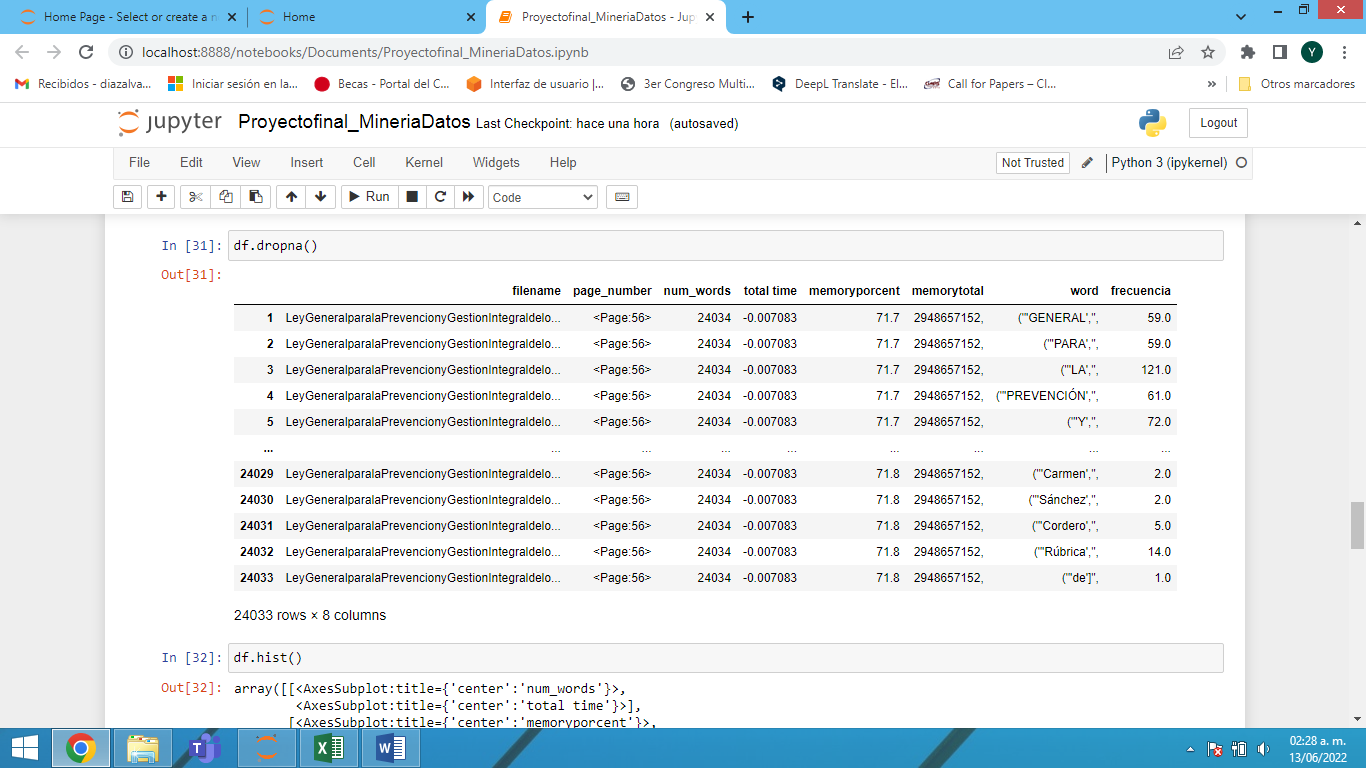


**3.2 Limpiar los datos**

Elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de subconjuntos limpios de los datos, la inserción de valores por defecto adecuados, o técnicas más ambiciosas como la

estimación de los datos que faltan.

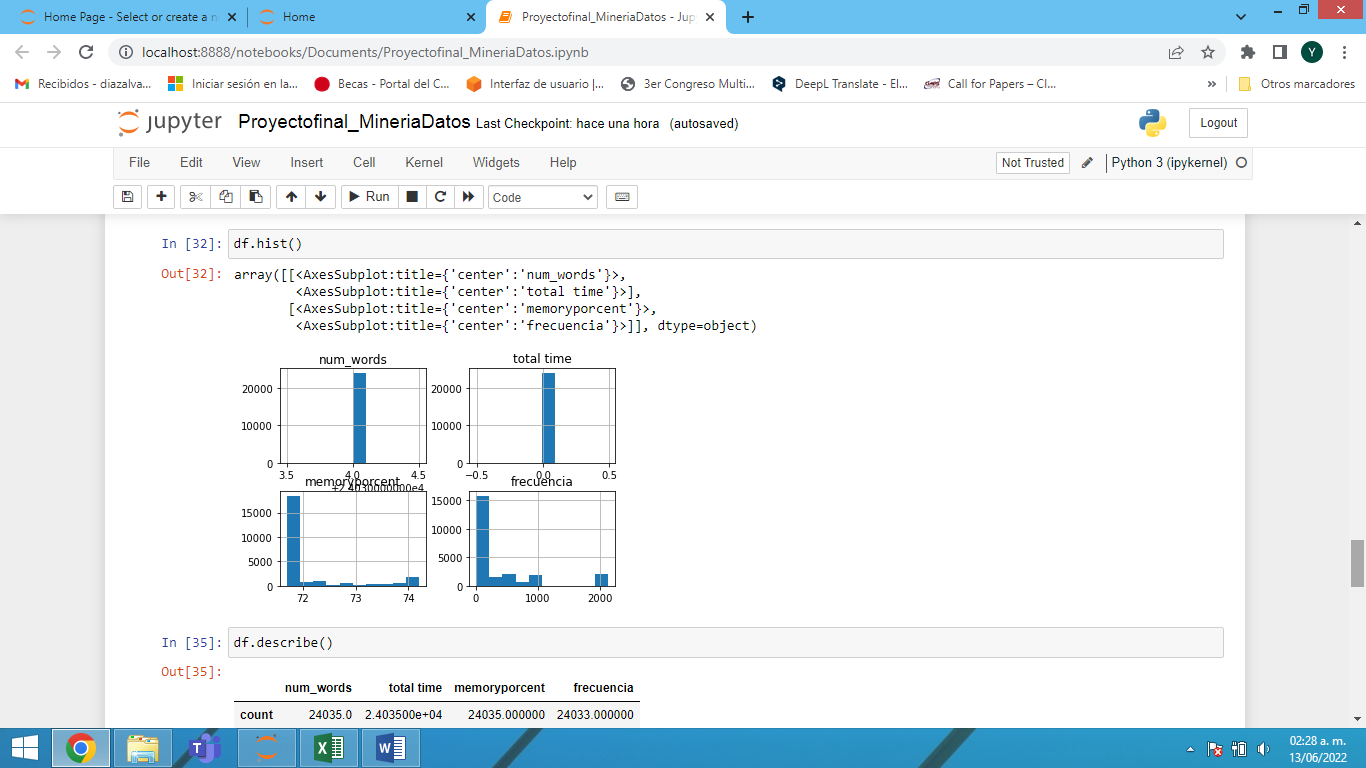
En la limpieza de los datos se muestran anteriormente 24035 filas las cuales con la limpieza de datos se eliminaron a 24033 filas.



**3.3 Construir los datos**

Esta tarea incluye operaciones que involucren derivar atributos derivados o nuevos registros completos, o valores transformados para los atributos existentes. (sólo si es necesario).

Se muestra el histograma de los nuevos registros anteriores, los cuales son número de palabras, total del tiempo, porcentaje de memoria y frecuencia de la palabra.



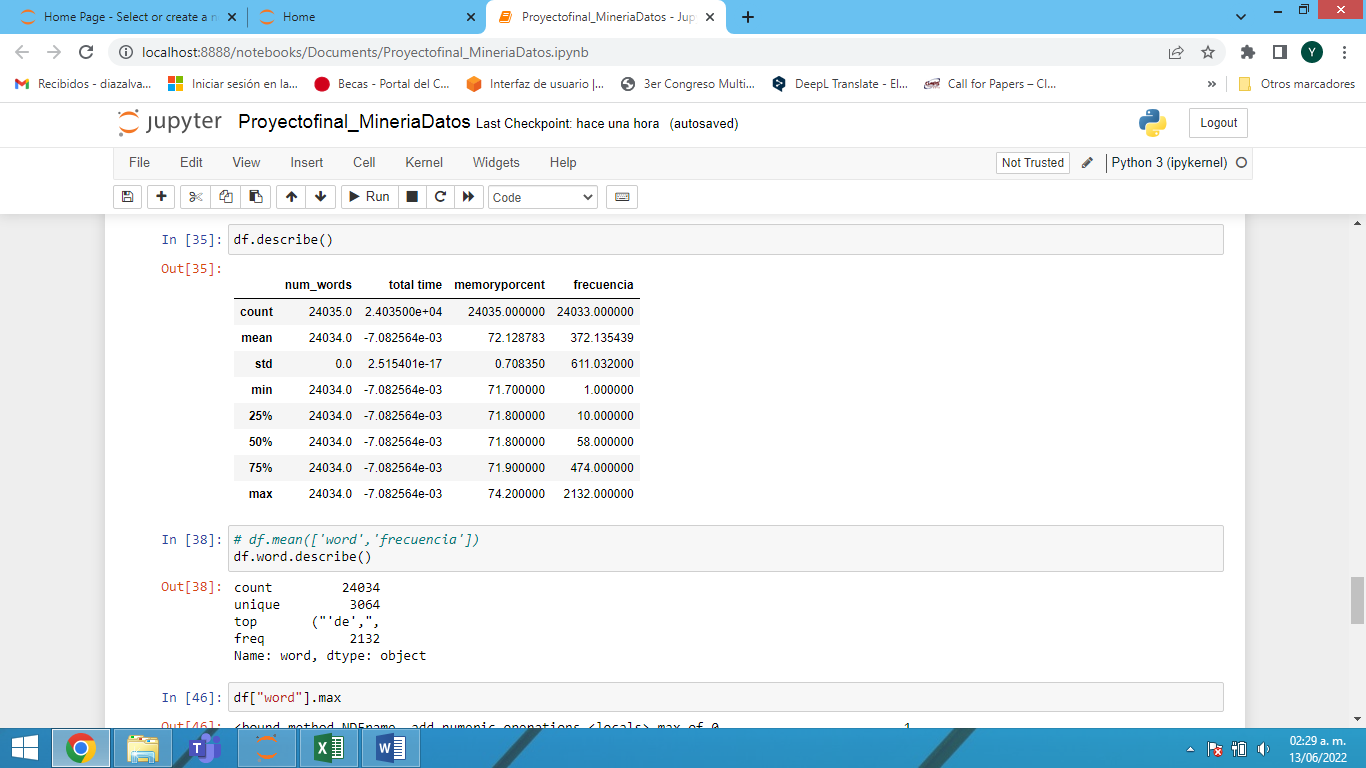
**3.4 Tarea Integrar datos**

Son métodos por los que se combina la información de varias tablas o registros para crear nuevos registros o valores (sólo si es necesario).

Ya que se tiene lo anterior, se describe de nuevo la tabla para verificar que datos quedaron en esta.

Se muestran los más importantes los cuales son número de palabras, total del tiempo, memoria y frecuencia de la palabra.

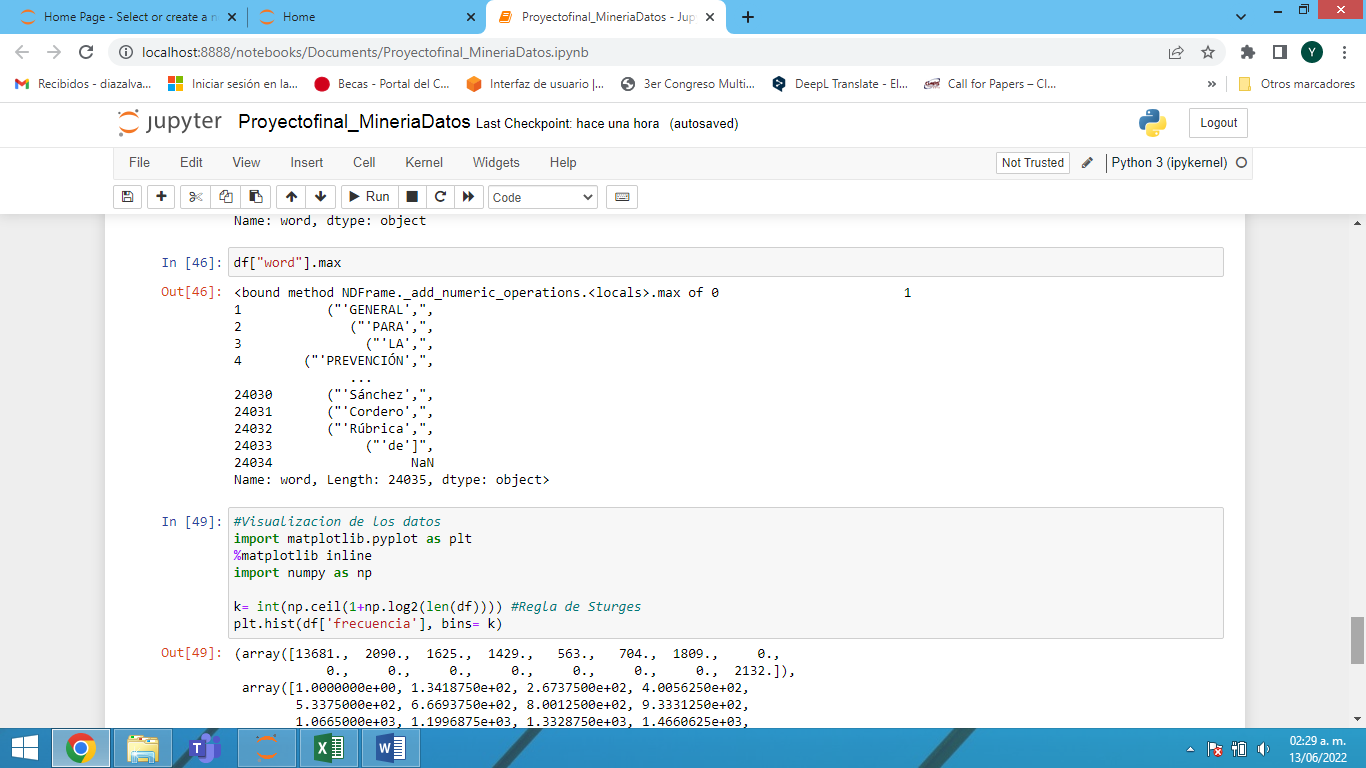
Se muestra la palabra con más frecuencia la cual fue “de”, las palabras totales fueron 24034, esta se contó 2132 veces repetida.

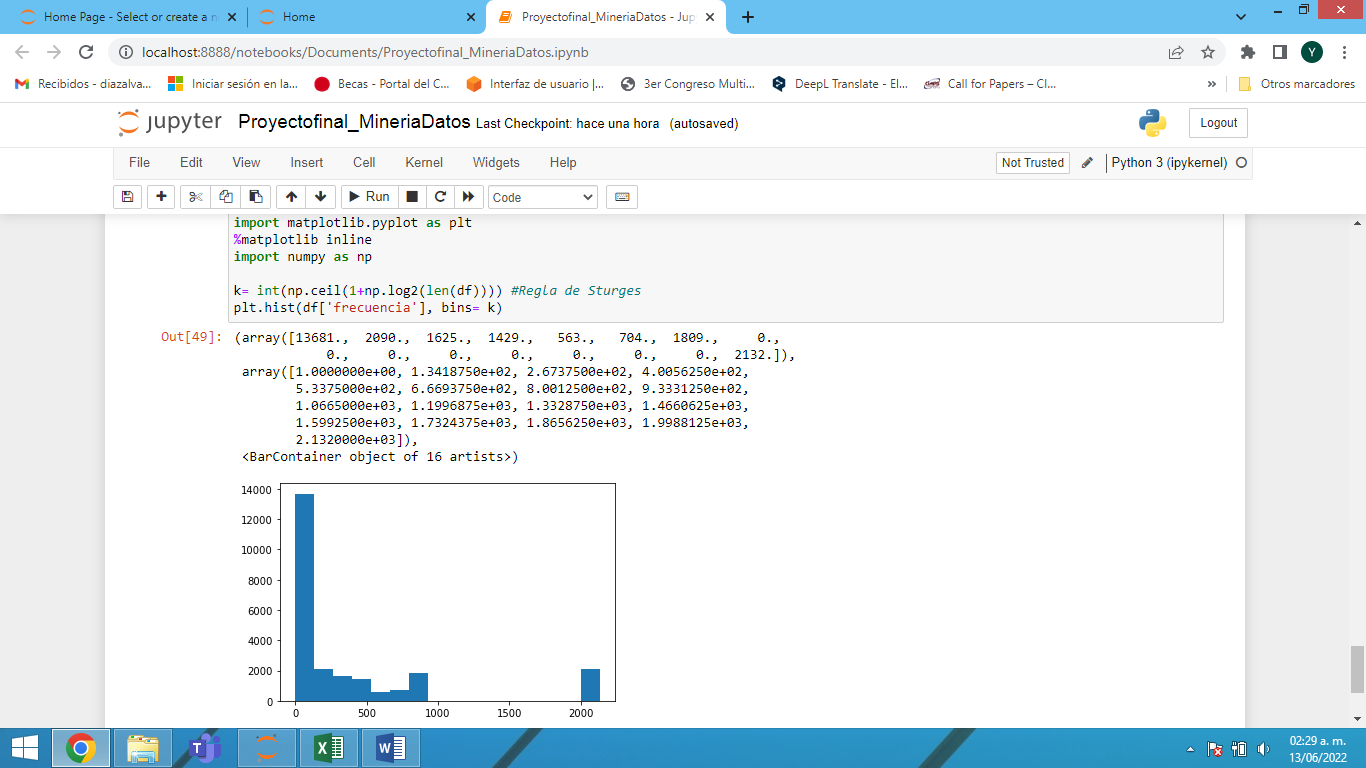


**3.5 Formatear datos**

Modificaciones sintácticas realizadas a los datos que no cambian su significado, pero que pueden ser requeridas por la herramienta de modelado. (sólo si es necesario).

Se muestra la palabra con menor frecuencia se tiene en el texto la cual fue “general” y la palabra “de” con mayor frecuencia.





Así como se muestra en el histograma el número de frecuencia de la palabra, la cual utiliza una formula logarítmica para obtener esta frecuencia.

**4.1 Seleccionar la técnica de modelado**

Seleccionar la técnica de modelización que se va a utilizar.

* Aprendizaje supervisado o no supervisado.
* Clasificación, regresión, clustering,…..
* Bayes, Prism, K-means, Random Forest,

Aplicarla y justificar por qué o cómo se utiliza y cuál es la utilidad, diseñando por lo menos 3 casos de prueba específicos.

**En caso de no ser factible a utilizar, explicar el por qué.**

**No se pudo hacer un modelo a este proyecto ya que contamos con palabras pero no tienen una clasificación aún.**