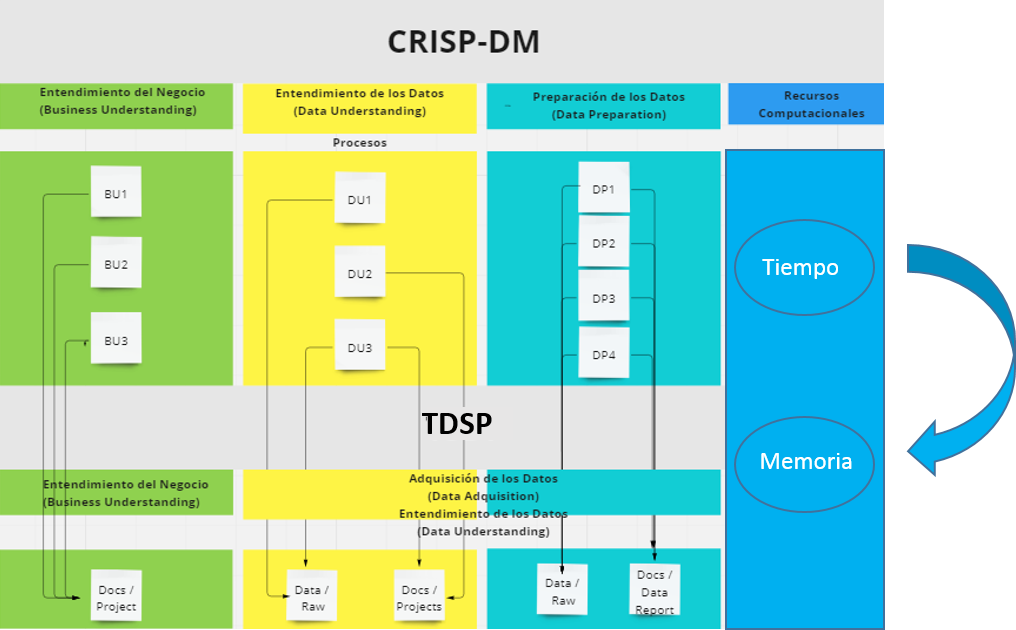
**Plan del Proyecto**

|  |  |  |  |
| --- | --- | --- | --- |
| **Fases/Tareas** | **Tipo** | **Acciones** | **Entregables** |
| **BU = Entendimiento del Negocio (Business Understanding).** | **Fase** |  |  |
| BU-G1 Determinar los objetivos del negocio | Tarea Genérica | Describir el objetivo principal del cliente desde el punto de vista del negocio. | 1. Objetivos del negocio 2. Criterios de éxito. |
| BU-G2 Evaluar la situación | Tarea Genérica | Hardware, software y recursos humanos. | 1. Inventario de recursos 2. Terminología. |
| BU-G3 Producir el plan del proyecto | Tarea Genérica | Enumerar los recursos necesarios, las funciones, las responsabilidades y los costes y beneficios. Además de especificar cómo se van a combinar las dos metodologías. | Plan del Proyecto |
| BU-E1 Creación del Repositorio de GitHub | Tarea Especializada | Crear un repositorio en GitHub | 1. Carpetas. 2. Descripción de las carpetas o repositorios. |
|  |  |  |  |
| **DU = Entendimiento de los Datos (Data Understanding).** | **Fase** |  |  |
| DU-G1 Recolección de Datos Iniciales | Tarea Genérica | Enumerar los documentos junto con su ubicación y el procedimiento para acceder a ellos. | Reporte inicial de recopilación de datos |
| DU-G2 Descripción de los datos (Leyes Ambientales en General) | Tarea Genérica | Describa los datos adquiridos, incluyendo su formato, cantidad e identidad de los campos (Estructura de leyes (Logo, nombre, artículos)) | Reporte de descripción de datos |
|  |  |  |  |
| **DP = Preparación de los Datos (Data Preparation).** | **Fase** | Preparar los datos | Datos listos para ser preparados |
| DP-G1 Limpiar datos | Tarea Genérica | Describir las transformaciones a realizar sobre los datos para su limpieza. | Reporte de procesos a aplicar de limpieza de datos |
| DP-E1 Transformación de pdf a texto plano | Tarea Especializada | Transformar las leyes en formato pdf a txt mediante las librerias de python | 1. Plantilla de Diseño experimental a realizar. 2. Algoritmo de transformación pdf a txt (dos librerías). 3. Texto de las leyes ambientales transformadas a txt. 4. Tablas de resultados. 5. Análisis e interpretación de resultados. |
| DP-E2. Eliminación de ruido (Quitar encabezados 1 vez) | Tarea Especializada | Quitar encabezados en todas las páginas, exceptuando en la primera. | Codigo mas elaborado |
| DP-E3. Quitar stopwords, paginas, logos | Tarea Especializada | Quitar stopwords o pies de página a algunas hojas | Codigo de limpieza |
| DP-G2 Formateo de datos | Tarea Genérica | Describa los cambios sintácticos realizados para satisfacer el objetivo empresarial. | Informe de datos reformulado |
| DP-E4. Etiquetado de textos. | Tarea Especializada | Etiquetar con html | Codigo etiquetado |
| DP-E5. Identificación de entidades. | Tarea Especializada | Identificar si las entidades son sinonimos, verbos, persona, objeto mediante vectores | Codigo vector |
| DP-E6. Transformar de TXT a estructura JSON | Tarea Especializada | Transformar txt a json | Codigo txt a json |
| DP-E7 Descripción del corpus | Tarea Especializada | Describir los textos de leyes ambientales con el fin de servir como muestra representativa para el análisis. | Conjunto cerrado de textos o de datos destinado a la investigación científica |

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

**Metodología Híbrida**

Derivado de la descripción de las dos metodologías utilizadas en proyectos de ciencia de datos y de la guía de diseño experimental, es decir, de la definición de CRISP-DM como modelo para el preprocesamiento de datos, de TDSP como guía para el manejo de la estructura del proyecto a través de plantillas, así como de la guía propuesta por McGeoch para el diseño de los experimentos, se plantea el uso de una metodología híbrida la cual se explica a través de la Figura.



Esquema de interacción entre las metodologías McGeoch, CRISP-DM y TDSP. Fuente: Elaboración propia

En la figura anterior, se muestra una metodología compuesta por las tres fases de CRISP-DM a utilizar en este proyecto, junto con las tareas genéricas y entregables de cada fase. TDSP añade un artefacto general representado por la estructura de directorios, así como un número variable de plantillas para cada tarea. Por último, la guía experimental mantiene un ciclo de vida activo de planeación y ejecución de las fases y sus respectivas tareas genéricas, controlando la repetición de las mismas o la adición o modificación de tareas especializadas.

La integración de las metodologías propuestas permite complementar las tareas sugeridas por CRISP-DM, tal y como mencionan los expertos que es necesario hacer en caso de contar con necesidades específicas del proyecto.

**Creación del repositorio**

|  |  |
| --- | --- |
| Repositorio | <https://github.com/YesseniaD/CRCPMT-Project> |