**Inventario de Recursos**

|  |  |
| --- | --- |
| Título del proyecto de tesis: | Caracterización de recursos computacionales para la fase de preprocesamiento de minería de textos |
| Datos de interés: | 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. |
| 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 (Tika, NLTK, PyPDF2 + 2 más) * Listado de restricciones   + 9 leyes * Recursos computacionales   + Tiempo y memoria |

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

## 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 8 GB. Disco duro de 500 GB. Procesador Intel Celeron Processor 1.60 Ghz x64. Windows 10 Home.

## Limitaciones

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