**Entrega 1**

**Formulación inicial del proyecto de Ciencia de Datos**

## 1. Contexto y Problema

Explica brevemente el entorno (empresa, investigación, industria) y cuál es el problema por resolver.

Gracias a la Ley de Derecho de Acceso a la Información Pública (Ley 27.275), cualquier persona puede consultar todos los movimientos y creaciones de las Sociedades en Argentina. Actualmente, no existen herramientas que nos permitan analizar las Sociedades Argentinas de manera eficiente. Al ser información no estructurada, se necesitan herramientas sofisticadas, que permitan entender el contexto de los negocios en Argentina.

## 2. Stakeholders

Lista los principales interesados (usuarios finales, áreas de negocio, tomadores de decisión).

Los principales interesados son:

* Analistas Políticos, Económicos y Sociológicos.
* Empresas de servicios en búsqueda de Leads.
* Fundadores y Emprendores analizando el mercado

## 3. Pregunta de Datos

Reformula el problema en términos de datos: ¿qué queremos medir, explicar o predecir?

* ¿Cómo podemos estructurar y normalizar los movimientos de las Sociedades Argentinas?
* ¿Qué tendencias están tomando las Sociedades en el pais?
* ¿Cuántas empresas han tenido que anunciarse en liquidación?
* ¿Cuántas sociedades existen por cada rubro?

## 4. Objetivo General

Enuncia de forma clara y medible qué busca el proyecto.

Construir un proceso que permita extraer, normalizar y analizar las Sociedades Argentinas.

## 5. Objetivos Específicos

Lista entre 3 y 5 objetivos concretos.

* Normalizar y estructurar todos los movimientos y creaciones de Sociedades en Argentina.
* Automatizar y definir un pipeline de ETL robusto que permita replicar el dataset para cualquier investigador.
* Investigar y analizar las tendencias de las Sociedades Argentinas.

## 6. Alcance

Define qué datos, periodo temporal y variables incluirás. Menciona exclusiones.

La meta del proyecto es que el proyecto se mantenga al día. Los primeros boletines digitalizados se realizaron en el año 2010, a partir de esa fecha, podemos empezar a analizar los movimientos hasta el presente.

Los datos a obtener del proceso de extracción son:

* Nombre de la sociedad
* Tipo de aviso (Constitución, Cesión, cese de actividad, aumento de capital, disolución)
* Fecha de publicación
* ID de Aviso
* Cuerpo del aviso

Con esta información, no es posible hacer una segmentación ni un análisis en profundidad del rubro societario. Para poder habilitar estos análisis, es necesario utilizar NLP, en particular LMs, que extraigan los rubros, objetivos, capitales y socios de la Sociedad.

Esto permite extraer más información:

* Accionistas
* Sede Social
* Actividades
* Rubro
* Capital Social
* Directivos y posiciones

Una Sociedad no ocurre únicamente en su concepción, esta es ente Jurídico que puede modificarse con el tiempo. Por esto, las modificaciones también deben ser almacenadas, para permitir comprender las tendencias de las Sociedades, su crecimiento, organigrama, y las acciones que tomaron.

## 7. Métricas de Éxito

Define cómo sabrás que el proyecto fue exitoso. Usa métricas cuantitativas.

* +10.000 Sociedades Argentinas normalizadas.
* +1.000 registros de alta fidelidad para utilizar en Finetuning.

## 8. Fuentes de Datos

Lista las fuentes, variables clave y observaciones sobre su calidad.

La fuente de datos es el Boletin Oficial. [https://www.boletinoficial.gob.ar](https://www.boletinoficial.gob.ar/). Los datos tienen una estructura básica para la extracción, pero es fundamental el procesamiento de información no estructurada, para poder hacer un análisis más holístico de las Sociedades.

## 9. Plan de Trabajo

Divide en etapas: Serán dependientes de la metodología que emplees. Por ejemplo: limpieza, EDA, modelado, evaluación, comunicación.

**Fase 1: Extracción de Datos Crudos**

El objetivo es extraer todos los boletines con un scraper en Python, y almacenarlos en una tabla temporal dentro de una base de datos.

1. Iterar por cada aviso del Boletin. No es necesario filtrar por fecha. Se puede acceder a cada aviso mediante la URL **detalleAviso/segunda/A1426199**, siendo el valor **A1426199** un identificador autoincremental.
2. Por cada URL, extraer el cuerpo del aviso, sociedad, fecha de publicación y el tipo de movimiento.
3. Almacenar la información extraida en una tabla en una Base de Datos MySQL.

**Fase 2: Finetuning de Gemma 3**

Para poder hacer un procesamiento óptimo de los avisos, se va a hacer un Finetuning del modelo Gemma 3 4B, utilizando la metodología QLoRA (Quantized Low Rank Adapters).

1. Construir un dataset de entrenamiento con la librería LangExtract, utilizando de input los avisos, y construyendo un JSON de respuesta con los campos clave. Utilizar Gemini 2.5 Flash para la extracción.
2. Obtener 1.000 registros de alta calidad, dividir el conjunto en 80/10/10 (test/train/validation).
3. Hacer un experimento de finetuning con Gemma 4B, 4bit quantization, y medir la accuracy.
4. Si la accuracy es menor al 60%, construir más registros e investigar los casos donde la extracción fue incorrecta.
5. Guardar el checkpoint final del modelo para utilizar en inferencia local.

**Fase 3: Modelado, Transformación y Enriquecimiento**

En esta fase convertiremos el texto no estructurado en una base de datos normalizada.

1. Diseñar el modelo de la base de datos, con Sociedades, Socios, Movimientos de capital y Movimientos de organigrama.
2. Construir un script para enviar cada aviso al motor de inferencia de Gemma 3 especializado en extracción de información societaria.
3. Almacenar las entidades extraidas de cada aviso en la BD normalizada.

**Fase 4: Análisis de las Sociedades Argentinas**

1. Hacer una investigación exhaustiva del conjunto de datos.
2. Identificar los rubros con más sociedades creadas, tendencias de sociedades constituidas por mes/año.

## 10. Cronograma

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mes | Semana | Fechas | Fase | Tareas Principales |
| Septiembre | Semana 2 | 8-14 sep | Inicio y Desarrollo del Scraper | - Configuración del entorno y repositorio. - Desarrollo y prueba del scraper (lote de 1,000 avisos). - Diseño del esquema de la base de datos. |
|  | Semana 3 | 15-21 sep | Ejecución de la Extracción Masiva | - Lanzamiento del scraper para todos los avisos. - Monitoreo del proceso. - Población de la tabla de datos crudos en MySQL. |
|  | Semana 4 | 22-28 sep | Inicio de Creación del Dataset de Entrenamiento | - Limpieza inicial de textos. - Generación de los primeros 250 registros de alta calidad. |
| Octubre | Semana 1 | 29 sep - 5 oct | Finalización del Dataset de Entrenamiento | - Generación y validación de los 750 registros restantes. - Revisión de la consistencia de las etiquetas JSON. |
|  | Semana 2 | 6-12 oct | Primer Experimento de Finetuning | - División del dataset (80/10/10). - Configuración del entorno de finetuning (QLoRA) - Ejecución del finetuning con Gemma 4B y evaluación. |
|  | Semana 3 | 13-19 oct | Iteración y Finetuning del Modelo Principal | - Si la precisión es < 60%, analizar errores y re-entrenar. |
|  | Semana 4 | 20-26 oct | Finalización y Validación del Modelo | - Conclusión del entrenamiento de Gemma 4B. - Evaluación con el conjunto de test. - Guardado del checkpoint del modelo. |
| Noviembre | Semana 1 | 27 oct - 2 nov | Transformación y Carga (Proceso ETL) | - Ejecución del script de inferencia sobre los datos crudos. - Población de las tablas normalizadas. |
|  | Semana 2 | 3-9 nov | Análisis de Datos y Cierre | - Análisis exploratorio. - Generación de métricas y visualizaciones. - Redacción del informe final y documentación. |
|  | Deadline | 10-14 nov | Entrega Final | - Margen para ajustes finales. - Entrega del proyecto. |

## 11. Riesgos y Limitaciones

Menciona problemas potenciales: calidad de datos, falta de recursos, privacidad.

* Tiempos necesarios para lograr buen rendimiento en Finetuning de Gemma 3. Gracias al uso de QLoRA, se puede reducir significativamente el consumo y los tiempos, al no ser necesario reentrenar todos los pesos del modelo, sino que únicamente se crea una nueva capa adaptativa, no se reentrenan los pesos.
* Los datos son de información pública por la Ley 27.275, pero debido a la Ley 25.326, se debe garantizar el derecho al honor y la intimidad de las personas. Por esto, se deben excluir los nombres de Socios específicos, ni información personal al momento de realizar el informe.

## 12. Entregables

Lista productos finales: dataset limpio, modelo, informe, dashboard, presentación.

* Una base de datos MySQL con tablas normalizadas que contengan la información estructurada de las sociedades.
* El código fuente del scraper, los scripts de ETL y el pipeline de inferencia.
* El checkpoint final del modelo Gemma 3 4B fine-tuneado para la extracción de entidades.
* Un informe final o dashboard con los hallazgos y visualizaciones que respondan a las preguntas de datos planteadas.

## 13. Comunicación

Define cómo y a quién presentarás resultados (formato y frecuencia).

(Completar aquí)

## 14. Recursos y Roles

Indica los recursos necesarios (humanos, tecnológicos, financieros) y los roles de cada integrante del equipo.  
Ejemplo: Data Scientist (modelado), Analista de Datos (limpieza y EDA), Ingeniero de Datos (ETL), Líder de Proyecto (coordinación), Recursos: servidores, software, presupuesto estimado

(Completar aquí)