TUIA NLP 2025
Práctica Unidad 5 Almacenamiento y representación de conocimiento

Trabajar en la carpeta Introducción del juego asignado en el TP1P2.

1. A partir de los textos extraídos en ejercicios anteriores, seleccioná fragmentos representativos del corpus, dividir con un splitter (explicando la elección del slitter seleccionado).
Nota: Apoyarse en TextSplitters de LangChain

2. Seleccioná un modelo de vectorización (por ejemplo, Sentence Transformers, spaCy, etc.) y justificá brevemente tu elección.

3. Iterar sobre cada fragmento de texto generado:
- Asignar un identificador único (por ejemplo, un contador incremental).
- Vectorizar el contenido.
- Extraer metadatos relevantes: nombre del archivo, idioma, autor, fuente, u otros disponibles.
Nota: Almacenar resultados en estructuras de datos temporales (listas, diccionarios, DataFrames, etc.), que se usarán en el siguiente paso.

4. Inicializar una base de datos vectorial usando ChromaDB y almacenar los vectores junto con sus identificadores y metadatos en la base de datos.

5. Realizar al menos 5 consultas sobre la base de datos (pueden ser frases o palabras representativas del contenido del juego).
- Evaluar la relevancia de los resultados obtenidos para cada consulta:
- Responder
    - ¿Qué tan relacionados son con la consulta?
    - ¿El sistema recupera fragmentos esperados?
    - ¿Existen casos ambiguos o fuera de contexto?