Project Instructions

This project focuses on relevant knowledge extraction and integration to existing knowledge base. The goal is to extract structured knowledge (in the form of subject–predicate–object triplets) from documents, compare it with existing knowledge, and decide whether the information is new, partially new, or already known.

Project Workflow

1. Input Knowledge (Ground Truth)

   - Start with introductory knowledge provided as sentence structures.

   - Convert these sentences into structured triplets (subject–predicate–object) to form the initial knowledge base. Use NLP/LLM. Store sentence information as property of predicate.

2. Select Document Corpus

   - Use a document (e.g., a PDF or text corpus) relevant to your chosen domain (e.g., environmental reports, technical standards, scientific papers).

   - This document serves as the unstructured source of new knowledge.

3. Chunking Process
   - Segment the document into smaller chunks (e.g., paragraphs, fixed-size windows, semantic units, etc.).

   - Experiment with different chunking strategies and justify which works best for your corpus. – Take care of coreference resolution.

4. Triplet Extraction
   - Use either traditional NLP pipelines (dependency parsing, OpenIE, spaCy) or LLM-based extraction (prompting) to extract triplets.
   - Each chunk should yield candidate triplets.
5. Knowledge Comparison

   - Compare extracted triplets against the initial ground truth.

   - Categorize them into: Exists (already in the knowledge base), Partially new (some overlap with existing knowledge), New (completely new information). Keep log of decision makings.

   – Initially consider nodes for graph comparison.

6. Knowledge Integration

- Expand the knowledge base by integrating new or partially new triplets.
- Ensure consistency and avoid duplication.
-Take care of normalizing entity representation by considering singular, plural cases, etc.

Considerations for ML/AI/LLM Students

- Ground Truth: Chose introductory knowledge (sentence structures converted into triplets). This acts as your baseline knowledge graph. Example:

Data Preprocessing is subclass of Data Science Task.
Supervised Learning is subclass of Data Science Task.
Classification is subclass of Supervised Learning.
SVM is subclass of Classification.
sklearn.svm.SVC is an Operator.
sklearn.svm.SVC is a Sklearn Estimator.
sklearn.svm.SVC implements SVM.
SVM is implemented by sklearn.svm.SVC.

- Document Corpus: Choose a related document in your area of interest (e.g., AI, ML, LLM, etc). This will be chunked and processed for new knowledge extraction.

- Evaluation: Use qualitative analysis to judge relevancy.

Expand the knowledge base by starting from a limited but fundamental ground truth. Even if the initial base knowledge does not cover all possible information, use it as an anchor point to process related documents, extract structured triplets, and iteratively integrate new knowledge. This approach ensures that the knowledge graph grows systematically while maintaining alignment with the initial truth.