RadiXplore Candidate Coding Challenge: Mining Project Intelligence System

Background

At RadiXplore, we specialize in extracting critical insights from vast geological and mining data. Accurate and automated identification of mining projects, along with their precise locations, is fundamental to our operations. This challenge simulates a real-world scenario where you'll build a system to achieve this.

You will be provided with:

- A collection of PDF files containing diverse geological and mining reports.
- A JSON annotation dataset, specifically highlighting "PROJECT" names within text (a sample of this format will be provided).

The Challenge: Building a Project Intelligence Pipeline

Your mission is to develop an automated pipeline that can identify mining project names from unstructured PDF text and infer their approximate geographical coordinates.

Part 1: Intelligent Information Extraction (Named Entity Recognition)

Objective: Develop a robust Named Entity Recognition (NER) model to pinpoint mining project names within the given pdf documents.

Steps:

1. **Text Extraction:** Implement a mechanism to accurately extract readable text content from the provided PDF files.
2. **NER Model Development/Fine-tuning:**
   - Utilize the provided JSON annotations as your training data.
   - Build or fine-tune an NER model specifically tailored to identify entities labeled as "PROJECT".
   - Your model should prioritize high accuracy in identifying these project names, even within noisy or ambiguously phrased text.
3. **Structured Output Generation:** For each identified mining project, generate a structured JSONL (JSON Lines) record. This record should contain:

```
{
  "pdf_file": "filename.pdf",
  "page_number": 3,
  "project_name": "Minyari Dome Project",
```

"context_sentence": "Minyari Dome Project is located in the Paterson region of WA and offers significant exploration upside.",
"coordinates": null
}

- pdf_file: The name of the PDF document where the project was found.
- page_number: The page number within the PDF where the project mention occurred.
- project_name: The exact text of the identified mining project.
- context_sentence: A sentence or short textual snippet that provides immediate context around the project name.
- coordinates: Initially null (this will be populated in Part 2).

Part 2: Geolocation Inference with Contextual Intelligence

Objective: For each extracted project name, infer its geographic coordinates (latitude, longitude)

Steps:

1. **Location Inference (LLM/Agentic Approach):**
   - Employ a Large Language Model (LLM) or construct an agentic pipeline to infer the most plausible geographic coordinates.
   - Your solution should demonstrate an intelligent approach to converting textual location cues into latitude and longitude. This could involve direct LLM querying, chain-of-thought prompting, integrating with external geospatial databases (gazetteers), or heuristic rules.
   - Populate the coordinates field in the JSONL record (e.g., [latitude, longitude]). If the location is highly ambiguous or cannot be reliably inferred, the coordinates field should remain null.

Deliverables

Please submit the following:

1. **Runnable Pipeline Code:** The complete, runnable code for your PDF text extraction, NER inference, and geolocation inference, leading to the JSONL output.
2. **JSONL Output File:** A single JSONL file containing all extracted projects and their inferred coordinates for all provided PDFs.
3. **README.md:** A comprehensive README.md file that includes:

    - Clear instructions on how to set up and run your entire pipeline to generate the output.
    - A list of all tools, libraries, and external APIs used.
    - Any key assumptions made during development.
    - A brief explanation of your model choices and geolocation strategy.

Evaluation Criteria

Your submission will be evaluated based on:

- **Accuracy & Robustness of NER:** How well your model correctly identifies mining projects across varied document structures and text quality.
- **Quality of Geolocation:** The precision and reasoning behind the inferred coordinates. We will assess the intelligence of your LLM/agentic approach in deriving locations from context.
- **Code Quality:** Clarity, modularity, readability, and maintainability of your code, along with comprehensive documentation.
- **Efficiency & Speed:** The performance of your pipeline in terms of processing time and resource utilization.
- **Error Handling & Real-world Complexity:** Your approach to handling noisy PDF text, ambiguous project mentions, and edge cases.
- **Reproducibility:** The ease with which we can set up and run your solution.

Bonus Points for:

- Implementing confidence scoring for extracted projects or inferred coordinates.
- Creative solutions for multi-modal fusion (if applicable, e.g., using visual cues from PDFs).

Notes for Candidates

- You are encouraged to use any open-source tools, AI Coding tools, pre-trained models, or publicly available APIs. **This includes the free tier of the Gemini API via Google AI Studio, which provides access to powerful LLMs.**
- The provided JSON annotation sample will clearly illustrate how "PROJECT" entities are labeled. Your NER model should consistently replicate this labeling style.
- For geolocation, creativity and innovative approaches are highly valued. Think about how you can leverage LLMs for "chain-of-thought" prompting.
- Aim for an efficient and reproducible pipeline. Containerization (e.g., Docker) is a plus but not strictly required.