Application Project: A Scalable Data Enrichment Pipeline for Scientific Document Analysis

Objective: The primary objective of this project is to design, implement, and validate a scalable data pipeline for the transformation of millions of unstructured scientific documents into a clean, structured, and annotated dataset suitable for advanced artificial intelligence analysis.

Application: Interested candidates are invited to submit their curriculum vitae, a current transcript of records, and a brief statement of motivation. Further details are provided below.

1. Introduction and Motivation

The effective analysis of large-scale scientific corpora is predicated on the initial transformation of unstructured data into a coherent, machine-readable format. A significant data engineering challenge lies in processing raw, heterogeneous documents, such as PDFs from research papers and patents, into a pristine state. This project addresses the critical need to develop a robust system capable of processing a corpus exceeding 20 million documents. The core task involves not only text extraction but also the identification and structuring of critical entities, concepts, and their intrinsic relationships. The successful completion of this foundational layer is indispensable for enabling all subsequent AI-driven intelligence and analysis.

2. Organizational Context.

Paper Pulse is an emerging technology, spun out of UTUM, developing a sophisticated Software-as-a-Service (SaaS) platform to innovate the process of technology scouting. The platform is designed to assist organizations in screening and evaluating scientific literature to identify its commercialization potential. A minimum viable product is currently operational at the Technical University of Munich (TUM), and the venture has garnered significant interest from other leading institutions.

3. Candidate's Responsibilities

We are seeking a technically-oriented candidate to construct the core data infrastructure for our platform. The key responsibilities will include:

- The design and implementation of a scalable Extract, Transform, Load (ETL) pipeline capable of ingesting and processing millions of documents from diverse sources.
- The development of an intelligent PDF parsing system that dynamically routes various document structures (e.g., text-intensive versus formula-intensive) to specialized parsing tools (such as PyMuPDF or Nougat) to ensure optimal data extraction.
- Leveraging Large Language Models (LLMs) for advanced Named Entity Recognition (NER) to identify and extract key semantic information, including technologies, methodologies, and equipment, from unstructured text.

- The creation of a "human-in-the-loop" smart annotation workflow designed to efficiently validate AI-generated labels, thereby constructing a high-quality dataset for subsequent machine learning model training.

4. Relevant Research Areas

This project is situated at the intersection of several key technical domains, including:

- **Data Engineering:** ETL Pipeline Design, Scalable Data Processing, Cloud Architecture (GCP/AWS).
- **Machine Learning:** Natural Language Processing (NLP), Named Entity Recognition (NER), LLM-based Data Extraction.
- **MLOps:** Data Versioning, Workflow Orchestration (e.g., Airflow), Smart Annotation Systems.

5. Required Candidate Profile

The ideal candidate will possess the following qualifications:

- Demonstrated proficiency in programming, particularly in Python and associated data processing libraries (e.g., Pandas).
- A demonstrable interest in constructing robust, scalable, and automated data systems.
- A foundational understanding of databases (SQL/NoSQL) and cloud-based workflows.
- Fluency in English.
- **Preferred qualifications include:** Experience with Docker, Kubernetes, or workflow orchestration tools.