Data Engineer Intern at AIDF-CRI

Background

The Asian Institute of Digital Finance (AIDF) is a university-level institute in NUS, jointly founded by the Monetary Authority of Singapore (MAS), the National Research Foundation (NRF) and NUS. AIDF aspires to be a thought leader, a Fintech knowledge hub, and an experimental site for developing digital financial technologies as well as for nurturing current and future Fintech researchers and practitioners in Asia. The Credit Research Initiative (CRI) is a non-profit undertaking under the AIDF. Pioneering the "public good" credit risk measures, the initiative is committed to advancing big data analytics and providing directly useful credit intelligence to academic and professional communities.

AIDF-CRI is dedicated to staying updated with the latest trends and technologies, especially for AI and LLMs. Notably, efficient and seamless data collection and preprocessing are often overlooked but are critical components for supporting high-performing models. We are continually developing and enhancing our data pipelines by integrating new technologies to meet both project and market demands.

Responsibilities

AIDF-CRI is actively seeking candidates with a strong interest in data management and pipeline optimization. The selected candidate will play a key role in enhancing the performance of our advanced models and applications by integrating the latest technologies to build data pipelines that are not only robust but also tailored to meet evolving needs.

Particularly, the responsibilities will include:

- Alternative Data Pipeline Development
    - Build and maintain scalable data pipelines to automatically scrape and process alternative data sources such as news websites, financial filings, and social media platforms.
    - Utilize modern scraping tools and automation technologies to ensure efficient and reliable data extraction.
- Advanced NLP Integration
    - Apply advanced NLP techniques, such as tokenization, named entity recognition (NER), and sentiment analysis, to extract meaningful insights from unstructured data.
    - Use cutting-edge NLP models and frameworks to analyze large datasets and produce high-quality outputs.
- Database Management
    - Work with both NoSQL databases & Vector Database (e.g., MongoDB, Milvus) to store, query, and manage the processed data.
    - Ensure that the database design supports efficient data retrieval and integration with machine learning models.
- Automation & Continuous Improvement
    - Automate data processing workflows, ensuring the scalability and reliability of the pipelines.

    - Identify opportunities for process improvements and enhance the robustness of existing systems.
- Documentation
    - Maintain comprehensive documentation of backend processes, system architecture, and database schema to promote knowledge sharing and ensure smooth onboarding for new team members.

Minimum Requirements

- Experience in web scraping such as news, social media, and financial reports.
- Strong understanding of NLP techniques, including tokenization, named entity recognition, and text classification.
- Proficiency in at least one programming language, such as Python, and experience working with NoSQL databases.
- Familiarity with web scraping tools and frameworks (e.g., Scrapy, Selenium).
- Strong problem-solving skills and the ability to work independently in a fast-paced environment.
- Excellent communication and documentation skills, both written and verbal, to collaborate effectively with cross-functional teams and stakeholders.

Bonus Skills

- Knowledge of vector databases or retrieval-based AI systems.
- Experience integrating LLMs into data pipelines and retrieval-based systems.
- Experience working with large datasets, especially alternative data sources like news and 10-K filings.
- Familiarity with machine learning techniques for text analysis, especially in data classification, clustering, and prediction.