Capstone Project (Optional)

Big Data Course

Dr. Tegawendé F. Bissyandé | Dr. Iyiola Emmanuel Olatunji

Project Overview:

The modern enterprise is drowning in unstructured Big Data including petabytes of text logs, customer feedback, medical records, financial reports, and internal documentation. While traditional Big Data tools excel at managing structured, relational information, they struggle to extract contextual, semantic, and actionable insights from this textual deluge. Large Language Models (LLMs) represent a paradigm shift, offering the capability to read, comprehend, and synthesize vast quantities of human-generated text at scale. This capstone project challenges you to explore the critical intersection: How can advanced LLM techniques be leveraged to solve a real-world Big Data problem, driving quantifiable business value?

Objectives:

Each student will select a unique Big Data domain and develop a solution that addresses at least three of the following objectives:

1. Technical Implementation: Develop a robust, end-to-end LLM-based solution (e.g., using a Retrieval-Augmented Generation (RAG) system, fine-tuning a smaller model, or deploying a multi-agent system).

2. Scalability and Performance: Demonstrate the solution's ability to handle Big Data volumes (e.g., millions of documents, terabytes of log files). This must include a discussion and implementation of Big Data tools (e.g., Apache Spark, Vector Databases, distributed processing) to manage the data pipeline.

3. Advanced Technique/Model Evaluation: Research and implement a complex LLM technique (e.g., Parameter-Efficient Fine-Tuning (PEFT), Quantization for compression, advanced prompt chaining, or utilizing a specialized open-source model like LLaMA/Mistral).

4. Ethical/Bias Analysis: Perform a rigorous analysis of potential biases, ethical concerns, or data privacy challenges inherent in the selected data and model, offering concrete mitigation strategies. This is dependent if the chosen project domain e.g. privacy-sensitive application.

5. Quantifiable Impact: Define and measure the success of the LLM solution against a clear baseline (e.g., accuracy improvements over a traditional NLP model, time-to-insight reduction, or cost savings from model compression).

Recommended Project Domains:

- Real-Time Customer Sentiment and Topic Clustering

- Question and Answering System using RAG

- Financial Report Analyzer

- Simple language translator

Suggestion on LLMs and CPU-friendly frameworks:

- LLMs: Mistral 7B Instruct v0.2/v0.3 or Llama 3 8B Instruct or Phi-3 Mini (3.8B) or DeepSeek R1 1.5B

- RAG Framework: LlamaIndex or Haystack

- Data Processing: Hugging Face datasets

- Big Data Processing: Polars or Pandas + Dask

- Embedding Model: all-MiniLM-L6-v2 or bge-small-en-v1.5

- Evaluation: ROUGE (for summarization) and Sequence Labeling Metrics (F1 for entity extraction)

- Agent Framework: LangChain or LlamaIndex

Deliverables & Evaluation:

The solution of the project should be hosted on GitHub. **Please send the link to** **emmauel.olatunji@uni.lu** with the subject "Big Data Captone Project-studentID". Replace with your studentID. The final grade will be based on the following:

- Proposal (50%): A detailed document outlining the problem, dataset, chosen methodology, and expected outcomes.

- Technical Implementation (50%): A fully documented, reproducible codebase (e.g., GitHub repository) demonstrating the data pipeline, model training/fine-tuning, and deployment strategy (even if mock-deployed).