Individual Plan & Reflection

Adam Holt

Reflection on Skill Auditing

I have foundational technical skills in statistical analysis and Python programming, which are essential for data science projects. I also have familiarity with natural language processing (NLP) using deep learning. Specifically, transformer-based Large Language Models (LLMs), which will play a principal role in the project.

Regarding soft skills, I am well-suited for a team-based project because I am naturally collaborative. Additionally, I enjoy teaching others who may not possess the same knowledge. These characteristics will be essential in actively contributing to the team; facilitating effective communication; and mentoring team members.

Self-directed Learning Plan

My primary technical shortages regard the specific datatypes, techniques, and models which apply to the project. Namely, the data preprocessing of scanned PDF files with Optical Character Recognition (OCR); the fundamentals of RAG architecture and implementation; and how to integrate OCR, LLMs and RAG for the data processing pipeline.

In order to address these shortages I have researched OCR models and learned about RAG architecture and its components and have presented my findings to the team. Additionally, I have been working to understand how to preprocess and format data for OCR/LLM/RAG integration.

Individual Tasks and Timeline

My tasks will be spread across the entire project timeline and will consist of:

- Research (1 week - complete):
    - Learning about OCR and RAG with LLMs.
- Data Preparation and Preprocessing (1-2 weeks - in progress):
    - Collect and preprocess coroner's report data.
    - Implement OCR model.
- RAG Model Development (2-4 weeks):
    - Implement the RAG model which integrates with the OCR preprocessor and LLM.

- Testing (continuous):
    - Test and validate the integrated data processing pipeline and individual components.
- Review and Fine Tuning (continuous):
    - Address bottlenecks and improve overall application quality and stability.
- Teamwork (continuous):
    - Collaborate with team members to ensure effective communication.
    - Regularly review and update the project plan to ensure progress and address issues.
- Stakeholder Engagement (continuous):
    - Liaise with the client to ensure project is within scope and meeting expectations.

I will manage my commitments with an iterative evaluation of the requirements to complete them.

Computing Environment and Teamwork Tools

We will set up a self-contained computing environment using Python and the Anaconda environment manager. This, in addition to a security audit, will ensure the data processing application uses open source software and executes locally without network access or calls to external APIs.

Python is well-suited for this project due to its extensive libraries and frameworks for data processing, NLP, and deep learning. Furthermore, Anaconda is a distribution of Python that includes a package manager (Conda) and a collection of popular data science and machine learning libraries which enables a convenient and reproducible way to manage dependencies and environments. This setup ensures that the project environment is self-contained, reducing the risk of conflicts or dependencies with external libraries.

For teamwork in developing the data application we will use git - a version control system that enables tracking changes, collaboration, and management of different versions of the codebase.

Risk Identification and Mitigation

Several important risks have been identified. The first relates to the data, namely the possibility of having too little data or poor data quality, both of which can lead to inaccurate or incomplete results. To ameliorate these risks we will introduce data preprocessing techniques (OCR) and generate synthetic data.

The second risk involves technical issues with the integration of the OCR, LLM, and RAG models. I have identified specific software (Docling) that will mitigate some of these risks, but due to the nature of the project many such risks remain and will need to be addressed on an ad hoc basis.

Another risk involves poor team collaboration and communication which will lead to delays and misunderstandings. To address this type of risk we shall establish clear communication channels and project management processes and regularly schedule team meetings and reviews to make progress and prioritise important tasks, and to ensure requirements and client expectations are met.