Internship Diary

Introduction

During my two months of internship period, I got the opportunity to work on two significant projects that greatly enhanced my technical skills and provided practical experience in real-world applications. These projects involved advanced data extraction techniques and the implementation of an offline private ChatGPT. Through these experiences, I gained valuable insights into various technologies and tools, which have been instrumental in my professional development.

The detailed information about the projects is below:

1. PDF Data Extraction

Introduction

I worked on a project focused on extracting data from PDF documents. The data was present in both structured and unstructured formats, which required the use of various Python libraries to effectively extract the information.

Objective

- Extract data from PDF documents.
- Handle both structured and unstructured data formats.
- Utilize different Python libraries for data extraction.

Tools and Technologies used

- **Python Libraries**: PyMuPDF, PDFPlumber, Tabula, Camelot

Methodology

1. **Structured Data Extraction**:

- **PyMuPDF**: Used for extracting text and images from PDFs. It provided a straightforward way to access the content of the PDF files.
- **PDFPlumber**: Utilized for extracting tables and other structured data. It allowed for precise extraction of tabular data, making it easier to handle structured information.

2. **Unstructured Data Extraction**:

- **Tabula**: Employed for extracting tables from PDFs. It was particularly useful for dealing with complex table structures.
- **Camelot**: Used for extracting tables from PDFs with a focus on accuracy and handling various table formats.

Learnings

- Gained a deep understanding of the data extraction process from PDFs.
- Learned to use advanced Python libraries for handling both structured and unstructured data.
- Enhanced my Python coding skills, particularly in the context of data extraction and manipulation.

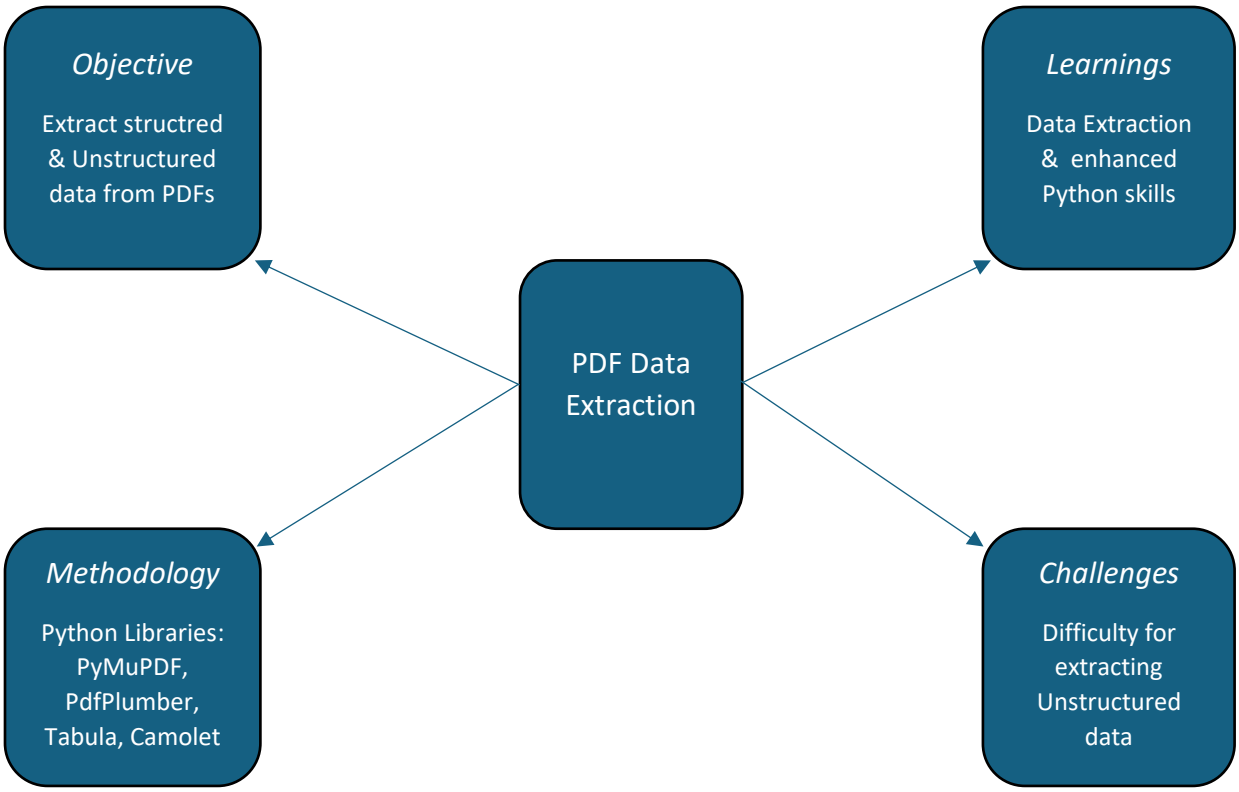
Challenges and Solutions

- **Challenge**: Handling PDFs with complex unstructured data.
- **Solution**: Experimented with different libraries for the extraction process to improve accuracy.

Conclusion

This project provided valuable insights into the data extraction process and significantly improved my Python programming skills. Using this project I also got the importance of choosing the right tools for specific tasks.

Flow Diagram



2. Offline Private ChatGPT

Introduction

The second project involved implementing an offline private ChatGPT. This project was particularly challenging as it required knowledge beyond traditional machine learning, exploring Generative AI (GenAI) and the LlamaIndex model of Ollama.

Objectives

- Implement an offline private ChatGPT.
- Ensure data privacy using Docker and PostgreSQL.
- Integrate GPU for enhanced performance.

Tools and Technologies Used

- **Generative AI**: LlamaIndex model of Ollama
- **Database**: PostgreSQL
- **Version Control**: GitHub

Methodology

- Cloned the repository from GitHub.
- Integrated the LlamaIndex model of Ollama for the ChatGPT implementation.
- Utilized GPU integration to enhance the performance of the model.
- Stored data securely in a PostgreSQL database.
- Ensured data privacy by using Docker containers.

Learnings

- Gained knowledge in Generative AI and the LlamaIndex model.
- Understood the integration of PostgreSQL for secure data storage.
- Enhanced my skills in using GitHub for version control and collaboration.

Challenges and Solutions

- **Challenge**: Implementing a new AI model with limited prior knowledge.
- **Solution**: Conducted extensive research and utilized online resources and YouTube videos to understand the model and its implementation.

Conclusion

This project was a significant learning experience, exposing me to new technologies and concepts. It helped me understand the practical applications of Generative AI and the importance of data privacy in AI implementations.