

**Course title: MCS7033** [**Collaborative Research Project 2**](https://lawrencetech.instructure.com/courses/15635)

[**Collaborative Research Project 2**](https://lawrencetech.instructure.com/courses/15635)**:   
 Multi-Source Data Analytics Chabot**

**Student Name: Sunil Kumar Kodi**

**Pooja Rani Mougli**

### **Abstract**

This project proposes a Chabot capable of analyzing and visualizing data from multiple sources while addressing specific healthcare needs. The system integrates web scraping, document processing, fine-tuned language models, and analytics. Additionally, it includes healthcare-specific capabilities such as prescription analysis, personalized health advice, and ethical compliance, making it a versatile and impactful tool.

### **Problem Statement**

Modern systems face challenges in efficiently processing diverse data formats and extracting actionable insights. This requires significant manual effort and specialized tools. In healthcare, real-time management of unstructured medical data is even more challenging, with ethical and compliance considerations adding complexity. A unified Chabot that automates data ingestion, analysis, and insight generation would address these issues across general and healthcare domains.

### **Project Objectives**

#### General Analytics Capabilities:

1. **Web Scraper**: Extract structured data from designated websites.
2. **Document Parsing**: Process Excel, Word, and PDF files to extract and unify relevant data.
3. **Language Model Interaction**: Fine-tune a GPT-like model to support natural language interaction.
4. **Analytics & Visualization**: Provide descriptive statistics and visualizations based on user queries.

#### Healthcare-Specific Features:

1. **Medical Query Handling**: Answer healthcare-related questions using fine-tuned domain-specific models like LLaMA and BioBERT.
2. **Prescription Analysis**: Extract and analyze data from prescriptions, focusing on medications, conditions, and recommendations.
3. **Ethical Compliance**: Ensure compliance with healthcare standards such as HIPAA and GDPR.

### **System Architecture**

#### **General Features:**

* **Web Scraping Module**: Automates data extraction from websites using tools like Beautiful Soup or Scrapy.
* **Document Handling Module**: Parses multiple formats (Excel, Word, PDF) with libraries like openpyxl, python-docx, and PyPDF2.
* **Language Model Integration**: Fine-tunes GPT or other language models using Hugging Face Transformers for customized interaction.
* **Visualization**: Generates charts (e.g., histograms, bar graphs) using Matplotlib or Seaborn.

#### **Healthcare Features:**

* **Medical Data Analysis**: Uses fine-tuned LLaMA and BioBERT models for extracting insights from medical queries and documents.
* **Custom NER Layers**: Adds Named Entity Recognition (NER) layers with CRF for precise medical term extraction.
* **Deployment**: Implements Flask for user interaction and scales using Docker and Kubernetes.

### **Methodology**

#### **Technologies**:

* **Programming Language**: Python
* **Libraries**: Beautiful Soup, openpyxl, python-docx, PyPDF2, pandas, Matplotlib, Seaborn, Hugging Face Transformers, PyTorch
* **Frameworks**: Flask for the chatbot interface
* **Deployment**: Docker for containerization, Kubernetes for scalability

#### **Steps to Implementation**:

1. **Python Foundations**: Establish a solid foundation in Python if needed.
2. **Web Scraping**: Target specific websites and extract structured data using web scraping libraries.
3. **Document Processing**: Load, process, and unify data from Excel, Word, and PDF files.
4. **Language Model Fine-Tuning**:
   * For general data: Fine-tune a GPT-like model.
   * For healthcare: Fine-tune LLaMA and BioBERT for domain-specific interaction.
5. **Analytics & Visualization**:
   * Implement descriptive statistics for general data analysis.
   * Generate visualizations like histograms and bar graphs based on user queries.
6. **Chatbot Interface**: Create a text-based or web-based interface using Flask for natural language interaction.
7. **Ethical Compliance**: Incorporate HIPAA and GDPR compliance measures, particularly for healthcare data handling.
8. **Scalable Deployment**: Use Docker for containerized deployment and Kubernetes for scalability.

### **Challenges and Mitigation**

1. **Data Quality**:
   * Issue: Real-world data can be messy.
   * Solution: Implement robust error handling and data-cleaning algorithms.
2. **Computational Limitations**:
   * Issue: Fine-tuning large language models requires significant resources.
   * Solution: Use cloud-based GPU/TPU resources like AWS or Google Cloud.
3. **Scope Management**:
   * Issue: Risk of expanding beyond achievable goals.
   * Solution: Prioritize core functionalities and evaluate progress regularly.

### **Evaluation Metrics**

#### General Data Analytics:

* **Functionality**: System’s ability to scrape, process, and analyze diverse data formats.
* **Query Handling**: Accuracy in understanding and responding to natural language queries.
* **Visualization**: Clarity and relevance of generated charts and graphs.

#### Healthcare-Specific:

* **Precision, Recall, and F1 Score**: Evaluate accuracy of medical term extraction and prescription analysis.
* **User Satisfaction**: Measure user satisfaction using a Likert scale.

### **Comparison of Features**

| **Aspect** | **General Data Chatbot** | **Healthcare Chatbot** |
| --- | --- | --- |
| **Scope** | General data analysis and visualization | Healthcare-specific insights and analysis |
| **Target Audience** | General users and organizations | Patients, healthcare professionals |
| **Core Technologies** | GPT-like models, document libraries | LLaMA, BioBERT, custom NER layers |
| **Data Types** | Web-scraped, Excel, Word, PDF | Medical Q&A, prescriptions, imaging |
| **Ethical Compliance** | Not explicitly mentioned | Strong focus on HIPAA and GDPR compliance |
| **Deployment** | Flask-based chatbot | Scalable Docker/K8s deployment |
| **Future Enhancements** | Better model support, expanded file processing | Integration with GPT-4, multimodal data |

### **Conclusion**

This unified Chabot combines general data analytics with healthcare-specific capabilities, providing a versatile solution for individuals and organizations. By integrating advanced technologies and maintaining ethical standards, the project offers a scalable and impactful tool ready for real-world deployment.