OAKSOL INTERSHIP ASSESSMENT

Ganesula Pradeep
Pradeepganesula6@gmail.com

AWS Textract OCR & MySQL Integration for Automated Text Extraction

Objective:

Implementing an automated text extraction pipeline with AWS Textract and storing structured data in a MySQL database are the goals of this assessment. The goal of this project is to rapidly extract text from scanned photographs, process the information, and store it for later research.

Project Requirements:

1. AWS Setup

- Create an AWS account.
- Configure an IAM user with the following permissions:
- AmazonTextractFullAccess
- AmazonAdministratorAccess
- Install and configure AWS CLI:
  AWS Access Key ID :
  AWS Secret Access Key:
  Default region name [us-east-1]:
  Default output format [json]:

2. MySQL Setup

- Install MySQL Workbench.
- Create a database named patients_data**.**
- **Create tables:**

```
CREATE TABLE patients (
    id INT AUTO_INCREMENT PRIMARY KEY,
    name VARCHAR(255),
    dob DATE
);
```

To store the json data:

```
CREATE TABLE forms_data (
    id INT AUTO_INCREMENT PRIMARY KEY,
    patient_id INT REFERENCES patients(id),
    form_json JSON,
    created_at TIMESTAMP DEFAULT NOW()
);
```

3. **Project Development**
- Extract text from an image using AWS Textract.
- Preprocess the extracted text and structure it in JSON.
- Insert data into MySQL with proper relationships.

4. **Write Python scripts for:**
- OCR extraction (aws_textract_ocr.py)
- Data insertion into MySQL (datainsertion.py)

5. **GitHub Submission**
- Initialize Git repository in the project directory.
- Commit all necessary files (excluding data.json and logs).
- Push the project to GitHub and provide a README with:

GITHUB REPOSITORY: https://github.com/Pradeep2625/Text_Extraction/