BFSI - OCR of Bank Statements

Author: Bhavya Taneja

Project Overview:

The project will consist of developing several key modules to facilitate the automation and analysis of financial data. These modules will ensure seamless API integration, accurate OCR data extraction, comprehensive salary and expenses analysis, and full deployment and integration with existing financial systems.

Technologies and Libraries Used

Web Scraping (Week 1)

- **Languages:** Python
- **Libraries:**
    - requests: For sending HTTP requests and fetching HTML content.
    - BeautifulSoup (from bs4): For parsing and extracting image URLs from HTML pages.
- **Concepts:** Bing Image Search, Directory Management, Data Downloading.

Key Features:

- Automated image scraping from Bing Image Search
- Organized directory structure for different document types

- Error handling for failed downloads
- Progress tracking and reporting

Cloud Storage and Retrieval (Week 2)

- **Platform**: Cloudinary
    - Used for securely storing and managing images in the cloud.
- **Languages**: Python
- **Libraries**:
    - cloudinary and cloudinary.api: For interacting with the Cloudinary API.

Concepts:

- Cloudinary API integration.
- Pagination for retrieving large datasets.
- Efficient handling of cloud resources.

Key Features:

- Cloud-based image storage
- Pagination support for large datasets
- API-based image retrieval
- Error handling for API operations

Install required packages:

pip install requests beautifulsoup4 cloudinary

Configuration: Web Scraping Setup

No additional configuration required. The script automatically creates necessary directories.

Cloudinary Setup:

Configure your Cloudinary credentials:

```
cloudinary.config(
  cloud_name="your_cloud_name",
  api_key="your_api_key",
  api_secret="your_api_secret"
)
```

USAGE:

Web Scraping :

python web_scraper.py

- Automatically creates 'downloaded_images' directory
- Organizes images by document type
- Displays download progress

Cloud Integration :

python cloud_retrieval.py

- Retrieves all images from Cloudinary
- Displays image URLs and public IDs
- Shows total count of stored images

Directory Structure

```
bfsi-ocr-project/
├── downloaded_images/
│   ├── bank_statement/
│   ├── profit_and_loss_statement/
│   ├── balance_sheet/
│   ├── payslip/
│   ├── passbook/
│   └── affidavit_document/
├── web_scraper.py
├── cloud_retrieval.py
└── README.md
```

Future Scope

- Implementation of OCR functionality
- Enhanced document classification
- Automated data extraction
- Document verification system

Contributing