PDF OCR Metadata Extractor

A Python tool that extracts company and contact information from PDF documents using OCR (Optical Character Recognition). The tool processes PDFs by scanning pages before the Table of Contents or first chapter to identify key business information.

Features

- **Smart Page Detection**: Automatically finds Table of Contents or Chapter 1 to limit scanning scope
- **OCR Integration**: Uses Tesseract OCR for scanned documents and direct text extraction for text-based PDFs
- **Structured Data Extraction**: Extracts emails, phone numbers, addresses, URLs, and company information
- **CSV Export**: Outputs structured data in CSV format with predefined company and contact schemas
- **Batch Processing**: Processes multiple PDFs automatically (configurable limit)
- **Logging**: Comprehensive logging for debugging and monitoring

Requirements

System Requirements

- Python 3.8 or higher
- Windows OS (batch files provided for Windows)
- Tesseract OCR engine

Python Dependencies

See `requirements.txt` for complete list:

- PyMuPDF (fitz) - PDF processing
- pdf2image - PDF to image conversion
- pytesseract - OCR interface
- Pillow - Image processing
- pandas - Data manipulation

Installation

1. Clone or Download

```bash
git clone <your-repo-url>
cd pdf-ocr-extractor
```

2. Install Tesseract OCR

Run the setup script:

```cmd
setup_tesseract.bat
```

Or manually download from: https://github.com/UB-Mannheim/tesseract/wiki

3. Install Python Dependencies

```cmd
pip install -r requirements.txt
```

Configuration

Edit the folder paths in `pdf_ocr_extractor.py`:

```python
FOLDER_IN = r"C:\Users\brenn\n8n-docker\upload-to-n8n-Waiting"
FOLDER_OUT = r"C:\Users\brenn\n8n-docker\upload-to-n8n-Waiting"
```

Usage

Quick Start

Run the batch file:

```cmd
run_extractor.bat
```

Manual Execution

```
cmd

python pdf_ocr_extractor.py
```

Output Files

The tool generates timestamped CSV files:

- `companies_YYYYMMDD_HHMMSS.csv` - Company information
- `contacts_YYYYMMDD_HHMMSS.csv` - Contact information
- `pdf_extractor.log` - Processing log

Data Schema

Company Data Fields

Column	Description
Company_Name_Location	Combined company name and location
Company_Name	Official company name
Company_MAIN_Phone	Primary phone number
Company_Job_Type	Company role (Owner/Developer, etc.)
Company_Street	Street address
Company_City	City
Company_State	State
Company_Zip/Postal	Zip/postal code
Company_Country	Country
Company_URL	Website URL
Industry	Industry classification
ALL_Company_Contact_Emails	All associated email addresses
ALL_Contact_Names	All associated contact names
Date_Created	Record creation timestamp

Contact Data Fields

Column	Description
Contact_Email	Primary email address
Contact_Name (First Last)	Full contact name
Contact_Phone_Direct	Direct phone number
Company_Name_Location	Associated company
Company_ID (LU)	Company lookup ID
Contact_Job_Type	Job role/title
Contact_ALL_Phones_JSON	All phone numbers (JSON format)
LinkedIn_URL	LinkedIn profile
Contact_Notes	General notes
Date_Created	Record creation timestamp

How It Works

1. **File Discovery**: Scans input folder for PDF files (processes first 5 by default)
2. **Page Analysis**: Identifies Table of Contents or Chapter 1 to limit scanning scope
3. **Text Extraction**: Uses direct PDF text extraction or OCR for scanned documents
4. **Pattern Matching**: Applies regex patterns to extract structured data:
   - Email addresses
   - Phone numbers
   - URLs
   - Street addresses
   - City/State/ZIP combinations
5. **Data Structuring**: Creates company and contact records following predefined schemas
6. **CSV Export**: Saves results with timestamps for easy tracking

Troubleshooting

Common Issues

Tesseract not found

- Ensure Tesseract is installed and in PATH
- Or specify path in script: `pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract.exe'`

No text extracted

- Check if PDFs are image-based (OCR required)
- Verify PDF files are not corrupted
- Check log file for detailed error messages

Poor extraction quality

- OCR works best on high-quality scanned documents
- Consider preprocessing images for better OCR results
- Adjust DPI settings in `pdf2image.convert_from_path()`

Missing dependencies

- Run: `pip install -r requirements.txt`
- Ensure all packages install successfully

Customization

Modify Data Fields

Edit the `company_headers` and `contact_headers` lists in the `PDFMetadataExtractor` class.

Adjust Regex Patterns

Modify the `patterns` dictionary to change extraction rules:

```python
self.patterns = {
    'email': r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b',
    'phone': r'(?:\+?1[-.\s]?)?\(?([0-9]{3})\)?[-.\s]?([0-9]{3})[-.\s]?([0-9]{4})',
    # Add custom patterns here
}
```

Change Processing Limits

Modify the `limit` parameter in the `main()` function or `process_pdfs()` method.

Contributing

1. Fork the repository
2. Create a feature branch

3. Make your changes
4. Add tests if applicable
5. Submit a pull request

License

Support

For issues and questions:

1. Check the log file (`pdf_extractor.log`) for detailed error information
2. Review the troubleshooting section above
3. Create an issue in the GitHub repository with:
   - Error messages
   - Sample PDF (if possible)
   - System information
   - Steps to reproduce

Version History

- **v1.0.0** - Initial release with basic OCR and metadata extraction