KYC Identity Verification System - Setup & Usage Guide

Overview

The KYC Identity Verification System is an AI-powered solution for extracting structured data from identity documents (driver's licenses, passports, and ID cards) to support financial services compliance requirements. It leverages Fireworks.ai's vision model capabilities to automate the document processing workflow.

System Architecture

Core Components

1. **Document Loader**
   - Discovers and validates image files in the designated directory
   - Supports PNG, JPG, JPEG, BMP, and TIFF formats
   - Validates file integrity before processing
2. **Vision Processor**
   - Interfaces with Fireworks.ai's `llama4-maverick-instruct-basic` vision model
   - Extracts structured KYC data from document images
   - Implements retry logic for API resilience
3. **Batch Processor**
   - Manages sequential or parallel processing of multiple documents
   - Generates comprehensive analytics and reports
   - Saves results in multiple formats (JSON, CSV)
4. **CLI Interface**
   - Provides easy-to-use commands for all operations
   - Displays real-time progress and results
   - Offers various output options

Data Flow

```
Documents Directory → Document Validation → Vision AI Processing →
Data Extraction → Result Validation → Output Generation (JSON/CSV)
```

Installation & Setup

Prerequisites

- Python 3.8 or higher
- Fireworks.ai API key
- Identity document images (driver's licenses, passports, ID cards)

Quick Setup (5 minutes)

1. **Clone or create project directory**

```bash
mkdir kyc-verification
cd kyc-verification
```

2. **Create Python virtual environment**

```bash
python -m venv venv
source venv/bin/activate  # On Windows: venv\Scripts\activate
```

3. **Install required packages**

```bash
pip install openai click pydantic pillow
```

4. **Set Fireworks.ai API key**

```bash
export FIREWORKS_API_KEY="your_api_key_here"
# On Windows: set FIREWORKS_API_KEY=your_api_key_here
```

5. **Create documents directory and add images**

```bash
mkdir documents
# Copy your identity document images to the documents/ folder
```

6. **Place the kyc_processor.py script in the project directory**

7. **Test the setup**

```bash
python kyc_processor.py test-connection
python kyc_processor.py list-documents
```

Usage Commands

Basic Commands

List Available Documents

```bash
python kyc_processor.py list-documents
```

Shows all documents in the documents directory with validation status.

Test API Connection

```bash
python kyc_processor.py test-connection
```

Verifies Fireworks.ai API connectivity and model availability.

Processing Commands

Process Single Document

```bash
python kyc_processor.py process-single "License 1.png"
```

Processes a specific document and displays extracted information.

Process All Documents (Sequential)

```bash
python kyc_processor.py process-all
```

Processes all valid documents in the documents directory sequentially.

Process All Documents (Parallel)

```bash
python kyc_processor.py process-all --parallel
```

Processes documents in parallel (up to 3 concurrent) for faster batch processing.

Utility Commands

Clear Output Files

```bash
python kyc_processor.py clear-outputs
```

Removes all generated output files (with confirmation prompt).

Output Structure

The system generates multiple output files for different use cases:

```
outputs/
├── batch_summary_YYYY-MM-DD_HH-MM-SS.json    # Processing statistics
├── detailed_results_YYYY-MM-DD_HH-MM-SS.json # Complete extraction data
├── kyc_extractions_YYYY-MM-DD_HH-MM-SS.csv   # Spreadsheet-ready format
└── individual/
    ├── License-1_results.json                # Individual document results
    ├── License-2_results.json
    └── ...
```

Output File Descriptions

1. **batch_summary_*.json**
   - Processing statistics (success rate, timing)
   - Document type distribution
   - Field extraction rates
   - Quality metrics
2. **detailed_results_*.json**
   - Complete extraction data for all documents
   - Failed document error messages
   - Processing metadata
3. **kyc_extractions_*.csv**
   - Key fields in tabular format
   - Easy import into Excel/Google Sheets
   - Ideal for compliance reporting
4. **individual/*.json**
   - Per-document extraction results
   - Full field details
   - Processing confidence scores

Extracted Data Fields

The system extracts the following information when available:

Personal Information

- Full name, first name, last name, middle name
- Date of birth
- Sex/gender

Document Information

- Document type (driver's license, passport, ID card)
- Document number
- Issue date and expiration date
- Issuing authority

Address Information

- Street address
- City, state/province
- ZIP/postal code
- Country

Physical Characteristics

- Height and weight
- Eye color and hair color

Additional Features

- Photo presence detection
- License class (for driver's licenses)
- Restrictions and endorsements
- Security feature detection

Performance Expectations

- **Processing Time**: 3-8 seconds per document (depending on size and complexity)
- **Success Rate**: Typically 90%+ for clear, well-lit document images
- **Parallel Processing**: Up to 3x faster for batch operations
- **API Rate Limits**: Built-in retry logic handles temporary failures

Best Practices

Document Image Quality

1. Use high-resolution images (300+ DPI recommended)
2. Ensure good lighting without glare
3. Keep documents flat and fully visible
4. Avoid shadows or obstructions

File Organization

1. Use descriptive filenames (e.g., "john_doe_license.jpg")
2. Group similar document types if processing many
3. Remove non-document images from the directory

API Usage

1. Set appropriate retry parameters for your use case
2. Monitor API usage to stay within limits
3. Use parallel processing for large batches

Troubleshooting

Common Issues and Solutions

1. **"API connection failed"**
   - Verify FIREWORKS_API_KEY is set correctly
   - Check internet connectivity
   - Ensure API key has proper permissions
2. **"Cannot open as image"**
   - Verify file is not corrupted
   - Check file format is supported
   - Try re-saving the image in a standard format
3. **Low extraction rates**
   - Improve image quality
   - Ensure documents are clearly visible
   - Check for supported document types
4. **Processing failures**
   - Review error messages in logs
   - Check individual result files for details
   - Verify API quota hasn't been exceeded

Appendix: Improvement Roadmap

Option 1: Enhanced Data Validation & Security

- **Advanced Validation Rules**: Implement checksums for document numbers, date logic validation
- **Fraud Detection**: Add tamper detection using image analysis
- **Data Encryption**: Encrypt sensitive data at rest and in transit
- **Audit Logging**: Comprehensive compliance audit trail
- **PII Redaction**: Automatic redaction options for sharing

Option 2: Advanced Analytics & Integration

- **OCR Fallback**: Integrate traditional OCR for text extraction backup
- **Multi-Document Correlation**: Link related documents (e.g., front/back of license)
- **Database Integration**: Direct export to SQL/NoSQL databases
- **REST API**: Web service wrapper for integration with other systems
- **Real-time Monitoring**: Dashboard for processing statistics
- **Machine Learning**: Continuous improvement through result feedback
- **Document Templates**: Custom extraction templates for specific document types

Additional Enhancements Under Consideration

- Support for additional document types (military IDs, state IDs)
- Multi-language document support
- Facial recognition integration for photo matching
- Blockchain integration for verification records
- Mobile app for document capture and processing
- Cloud deployment options (AWS, Azure, GCP)

Support & Maintenance

Logging

The system maintains a detailed log file (`kyc_processing.log`) with:

- Processing timestamps
- Success/failure records
- Error details
- Performance metrics

Updates

- Check Fireworks.ai documentation for model updates
- Monitor the vision model performance
- Update extraction prompts based on results

Backup

Regular backup recommendations:

- Archive processed documents
- Backup extraction results
- Maintain API configuration