Letter Identification
Using OCR

Abstract
Capital One receives thousands requests via mail every month from customers requesting to change data listed on their credit report. The proposed system will ingest scanned letters and process them using OCR and Natural Language Processing technology to extract relevant customer details and add them to the database, as well as attempt to detect if the request is frivolous. This automation will save Capital One employees time and increase the number of letters Capital One can process without needing to hire additional staff.

Architecture
- Tesseract OCR
- Docker
- OpenNLP
- Maven
- Spring Framework
- Natty
- PostgreSQL

Performance Metric
- Good (90% accuracy) extraction average for most fields
- Good OCR performance (80% accuracy)
- Average speed for 1 page document (30-35 seconds)

Impact
Capital One is on a mission to change banking for good and to help people live their best lives as an information-based technology company. Our team’s work helped to reach those goals.

Before
- Employees manually extracted key customer information from letters

After
- Automated data extraction
- Multi-instance and queue support
- Template Detection

Summary
- Implemented text extraction using Tesseract OCR library with Java wrapper
- Parse relevant user information from extracted data using regex
- Used SimHash algorithm to detect similarity to other letters and decide if letter is “template” mail
- Store data and results in PostgreSQL database