Capital One:
Letter Identification Using OCR

Team Members: Philip Foster, Andrew Mollenkamp, Austin Kirkpatrick, Jacob Camacho
Advisors: Arunachalam Saravanan (UTD)

Abstract: Capital One receives thousands requests via mail every month from customers requesting to change data listed on their credit report. Presently, employees must scan these letters into an electronic database, however they must still manually read these letters to extract important information such as the customer's name, address, SSN, account numbers, etc. 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 will save Capital One employees time and increase the number of letters Capital One can process without needing to hire additional staff.

Table of Contents

Executive Summary

Write the executive summary here…

Introduction

Each month, Capital One receives approximately 10,000 request letters via mail from customers to update information on their credit reports. Presently, these letters are read and processed manually by approximately 70 employees in the Credit Bureau Disputes Team. At present, the team's process for processing these letters involves scanning the paper letters and uploading them to a database. Employees then must read each letter and extract relevant fields such as customer name, address, SSN, etc. and enter them into a database by hand. This is time-consuming and prone to human errors. After entering the data, an employee must investigate the case to determine whether or not the request should be honored or rejected.

The Credit Bureau Disputes Team wants to automate part of this process by having an automated system read the letter to extract customer information and enter it into the database. This will save team members a significant amount of time by automating part of their current workflow. We are not aware of any alternative software solutions that exist to solve this problem.

Discussion

The core of the project will be written using Java and the Spring Framework. This gives us a good foundation which can easily be extended later on. This application will be placed in a Docker container, which will give team members, as well as Capital One, a consistent image that can be easily started and stopped, as well as replicated (using Kubernetes, or a similar tool) easily for scalability. We will be using Postgres for primary data storage, as well as Elasticsearch to store the converted text for fast searching.

To start, there will be an API where users can make requests to query the system and request new documents be ingested. This will be done with Spring Web - a part of the Spring Framework, which makes it easy to create RESTful APIs. The API paths will be divided by function:

The document ingest system will be the subsystem that is responsible for reading new letters and processing them to extract relevant information. The scanned images will be uploaded and converted to text via Tesseract OCR. The raw text will be stored in Elasticsearch for fast searching later on. The system will additionally process the text to extract important customer information.

Resources

- Personal computers of each of the members of the team
- Slack: https://slack.com/
- Cisco Webex Meetings (Provided by UTD)
- IntelliJ IDE: https://www.jetbrains.com/idea/
- GitHub: https://github.com/philipfoster/CapitalOne-OCR-Project
- Tesseract OCR (Optical Character Recognition) Open Source Project:
  https://opensource.google.com/projects/tesseract
- Java JNA wrapper for Tesseract OCR API: https://github.com/nguyenq/tess4j
- Fake letters (Provided by Capital One) for testing

Key Roles

Jacob Camacho	-	Role TBD
Philip Foster	-	Role TBD
Austin Kirkpatrick	-	Role TBD
Andrew Mellenkamp	-	Role TBD

Communication Plan

We plan to have weekly team meetings at minimum with biweekly meetings being preferred. These meetings will be each Tuesday and Thursday at noon, and will last as long as the team feels is necessary at that point in time. Our advisor will join in on these meetings occasionally as often as he feels is necessary. Additionally, we will meet with the corporate sponsor weekly on Friday at 3pm. Meetings will be held virtually via WebEx unless the team agrees that an in-person meeting is necessary. In this case, the team will work together to determine a suitable location to meet at. We have created a Slack group for general communication which does not require a meeting. Members can be in near-constant communication whenever it is necessary.

Risk Analysis / Contingency Plan

There are several risks present for this project.

Team Member is Unable to Work

There are several extreme circumstances, such as severe or prolonged illness, which may prevent a team member from being able to work on the project. If this happens, the affected member will communicate this with the rest of the group so that the group can plan to complete any work that the affected person cannot do. This will ensure that the project remains on-schedule.

Team is Unable to Complete the Project by Delivery Date

One of the goals of this project is to create a product that is well documented and have a high level of readability. In the event that the project is unfinished by the delivery date, our team hopes to give at least a partially finished product that another development team can easily continue the remaining development with as little transitionary difficulties as possible.

The Tesseract OCR Proves Unviable for the Project

From our research, Google’s Tesseract OCR API is the only reputable open-source product. In the event that Tesseract proves unviable for the project, other APIs will be further researched although the expectation is that other APIs will be subscription-based. If Capital One does not agree to purchasing/using another software, our goal is to deliver as much a finished product as possible, with functionality in all other parts of our project.

Costs

We do not anticipate any costs for this project. This is a purely software project, so there are no hardware requirements. Any software dependencies we use will be exclusively free and/or open source, per Capital One's request. Team members are free to use any developer tools that they already have to create the project. Any proprietary software that we use (i.e. WebEx) is either provided by the university, has a suitable free tier available, or has free student licenses available.

Timetable

Project Start

September 5th

Project Proposal Due

October 5th

Poster/Slides Due

December 10th

Project End

December 13th

Evaluation

TODO: Ask Faculty Advisor and Corporate Advisors what they prefer for measuring progress.

Performance Metrics

- Accuracy of using OCR to convert printed text to written text.
- Accuracy of recognizing useful fields within the letters.
- Speed of completion from receiving a letter to the output files of parsed data from the letter.

Conclusion

This project allows corporations, like Capital One, the ability to reduce mundane and labor intensive task by automating the recognition of important fields in customer complaint letters. While this project will not be able to fully automate the process, it will help to reduce the number of man-hours put into simply reading and parsing letters. With the use of our machine learning OCR software, the team which processes customer complaint letters could be reduced or their time better spent on other activities, better increasing the efficiency of the organization.

Contact Information

Fill in your information. Feel free to add additional relevant fields if you want.

Philip Foster:

Phone: (281) 740-9439
Email: philip@pfoster.me
Website: https://pfoster.me

Andrew Mollenkamp:

Phone: (469) 579-9524
Email: ajm151130@utdallas.edu

Austin Kirkpatrick:

Phone: (972) 983-8287
Email: abk150030@utdallas.edu

Jacob Camacho:

Phone: (214) 250-8946
Email: jbc140230@utdallas.edu

Deepesh Chaudhary:

Phone:
Email

Sources

Write any sources here… I don’t think we need formal citation, but we should probably ask Dr. Razo

Appendix

Approval

Signatures go here.