International Journal for Modern Trends in Science and Technology, 6(12): 520-523, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061299
Available online at: http://www.ijmtst.com/vol6issue12.html

Electronic Invoicing using Image Processing

Harsh Arora

B. Tech Scholar, Information Technology Department, Maharaja Agrasen Institute of Technology, New Delhi, India

To Cite this Article

Harsh Arora, "Electronic Invoicing using Image Processing", *International Journal for Modern Trends in Science and Technology*, 6(12): 520-523, 2020.

Article Info

Received on 18-November-2020, Revised on 16-December-2020, Accepted on 19-December-2020, Published on 23-December-2020.

ABSTRACT

Most ecommerce companies have their receive to pay process as predominantly manual, leading to non-reliability of payments & delayed visibility for sellers and requirement of additional manpower for scaling up for buyers. With the correct image and pdf processing tools, it is possible to automate this process for more efficient and cost-effective results. The research paper focuses on automating the task of invoice processing which is predominantly done manually. The idea is to save time, effort, and costs while eliminating human errors from the process. There are several existing image and pdf processing tools of which we will discuss pdftotext, tesseract and tesseract4.

KEYWORDS: *Image Processing, Electronic invoicing, pdftotext, tesseract, tesseract4.*

INTRODUCTION

Most ecommerce companies have their receive to pay process as predominantly manual, leading to non-reliability of payments & delayed visibility for sellers and requirement of additional manpower for scaling up for buyers. For a seller this involves high turn around time to know the money that the buyer is going to acknowledge. This in turn leads to reduction in time to respond to any deductions without impact to payment for the invoice and unpredictability in working capital planning.

Manual invoice processing forces a buyer to scale up in operations which in turn require scale up in manpower. Manual processing ensures that the cost of processing per invoice is high while being error prone.

The system will determine the viability of different image processing techniques to obtain all relevant information from the document and determine their accuracy.

Most ecommerce companies have their receive to pay process as predominantly manual, leading to non-reliability of payments & delayed visibility for sellers and requirement of additional manpower for scaling up for buyers.

STRUCTURE OF PAPER

The paper is organized as follows: In Section 1, the introduction of the paper is provided along with the structure, important terms, objectives and overall description. In Section 2 we discuss related work. In Section 3 we have the complete information about image processing tools. Section 4 shares information about the flexible YAML templating system created for it, its advantages and disadvantages. Section 5 tells us about the methodology and the process description. Section 6 tells us about the future scope and concludes the paper with acknowledgement and references.

OBJECTIVES

The predominant invoice processing systems are either entirely manual or they follow a rigid single template system. Whether an individual is a buyer or a seller, this leads to a lot of inefficiencies and high costs.

This project aims to address some of the problems in current systems by greatly minimizing the human intervention in the process and thus reducing costs and errors. The aim is to ease the task of both the buyer and the seller.

II. RELATED WORK

There are numerous works that have been done related to image processing machine learning algorithms.

J.E. Cross[1] has investigated methods to recover the maximum amount of available information from an image. Some radio frequency and optical sensors collect large-scale sets of spatial imagery data whose content is often obscured by fog, clouds, foliage and other intervening structures. Often, the obstruction is such as to render unreliable the definition of underlying images. Various mathematical operations used in image processing to remove obstructions from images and to recover reliable information were investigated, to include Spatial Domain Processing, Frequency Domain Processing, and non-Abelian group operations.

John C. Russ [2] has investigated techniques of image processing. These are operations that start with a grey scale (or color) image and return another grey scale image. The next chapter will deal with some additional techniques that operate on grey-scale images for purposes of locating feature edges in the context of isolating features for measurement.

J. M. White [3] and G. D. Rohrer, "Image Thresholding for Optical Character Recognition and Other Applications Requiring Character Image Extraction," in IBM Journal of Research and Development have researched on Two new, cost-effective thresholding algorithms for use in extracting binary images of characters from machine- or hand-printed documents.

However there has been little to no work put into the viability of image processing to achieve electronic automated invoicing.

III. IMAGE PROCESSING

There are numerous image and pdf processing libraries that we can use to extract the raw text of our invoice from. We will discuss pdftotext, tesseract and tesseract4.

pdftotext [4]

pdftotext is an open source command-line utility for changing PDF files to plain text files—i.e. extracting text information from PDF files.

It's freely offered and enclosed by default with several LINUX distributions, and is additionally offered for Windows as an element of the Xpdf Windows port. Such text extraction is sophisticated as PDF files are internally designed on page drawing primitives, which means the boundaries between words and paragraphs are often inferred to support their position on the page.

pdftotext is part of the Xpdf software suite. Poppler, which is derived from Xpdf, also includes an implementation of pdftotext. On most Linux distributions, pdftotext is included as part of the poppler-utils package.

Tesseract [5]

Tesseract is an optical character recognition engine, released under the apache licence. In 2006, it was considered to be one of the most accurate open source ocr engines.

The tesseract engine was released as open source by HP labs and the University of Nevada. Google has been sponsoring the development of tesseract since 2006. Tesseract can be executed via a command line interface.

It also has support for opening images using the leptonica [6] library.

Tesseract4 [7]

Tesseract4 uses a completely overhauled internal OCR engine. It uses a new neural network based recognition engine which enables it to deliver a significantly superior accuracy on document images than its predecessors.

However an increased accuracy comes at a tradeoff with a significant increase in the required compute power. The new engine helps deliver

impressively accurate results even without manual enhancements of the images.

It has also shown capability in recognizing handwritten text.

IV. YAML TEMPLATING

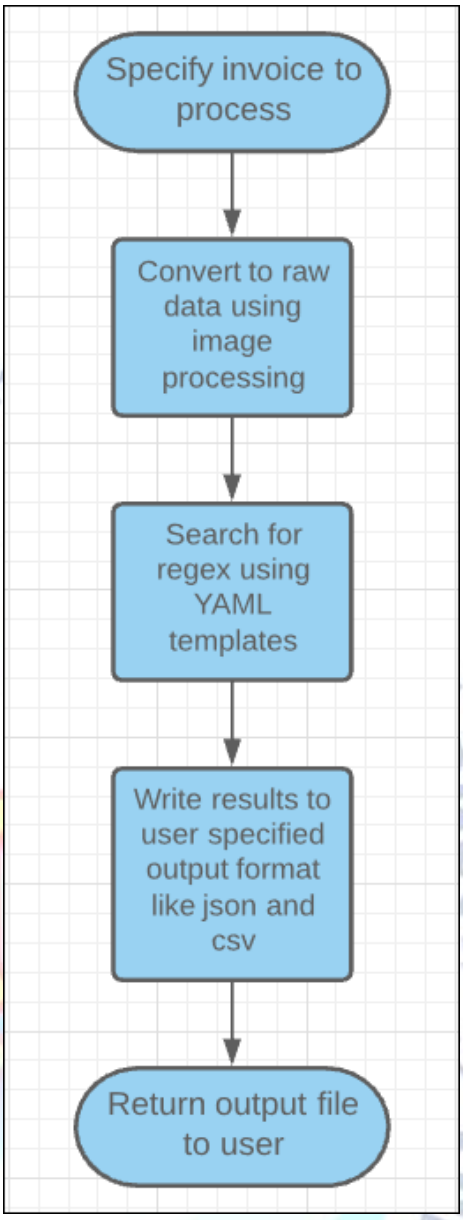
The system uses a YAML based templating engine to search for regex in the raw text extracted from the image processing libraries. A flexible YAML based templating system allows us to match content pdf files precisely. It also enables us to define static fields that are identical for every invoice. It provides enough flexibility to an organization to define custom fields or to have multiple regex per field based on requirements. The templating system has plugins available to match line items and tables in an invoice. Using YAML templating allows flexibility for users to quickly add new invoice templates to the system by quickly creating a template for it and the system can process invoices based on the new templates immediately. This templating system results searches for regex in the raw text and helps return accurate results from an invoice.

V. METHODOLOGY

The objective of this project is to allow an efficient, cost effective and a convenient invoicing system which can adapt to different types of invoice templates. The idea is to get the path to the invoice from the user via the command line. With the help of different image and pdf processing libraries, raw text from the invoice can be extracted. The system automatically recognizes the format of the invoice and uses the appropriate YAML template to search for regex in the raw data. Once the data has been obtained, the data is then converted to an output format which is specified by the user via the command line. The project is aimed at building a flexible invoicing system which can precisely match content PDF files, easily match line items and tables and automate the entire invoicing process for any major organization. The proposed system will have the ability to obtain all relevant information from the document with 100% accuracy while ensuring speed and reliability.

Process Description

The following diagram makes it easier to understand how we proceed.



- The user is required to input the path to invoice via the command line.
- Additionally, they can also specify the image processing library, output folder and output file format via the command line itself.
- The invoice is then processed and the raw text received is searched for regex based on provided YAML templates.
- In case no YAML template is found, an error is shown to the user and they can add the new template to it.
- The obtained result is written to an output file as specified by the user or in default configuration.
- The output file is saved to a specified folder for the user to access.
- This flexible invoicing system would help minimize human interaction, which in turn will increase efficiency and reduce costs.

VI. FUTURE SCOPE AND CONCLUSION

The project is aimed at building a flexible invoicing system which can precisely match content PDF files, easily match line items and tables and automate the entire invoicing process for any major organization. The system has the ability to obtain all relevant information from the document with 100% accuracy while ensuring speed and reliability. This project has a huge potential for further development. While the problem focuses on digitization of invoices, this could be extended to digitizing any document for processing, thereby removing any manual efforts, errors and management of document processing within companies.

REFERENCES

[1] J. Luo and J. Cross, "Advanced Image Processing Techniques for Maximum Information Recovery," 2007 Thirty-Ninth Southeastern Symposium on System Theory, Macon, GA, 2007, pp. 58-62, doi: 10.1109/SSST.2007.352317.

[2] Russ J.C. (1990) Image Processing. In: Computer-Assisted Microscopy. Springer, Boston, MA. https://doi.org/10.1007/978-1-4613-0563-7_3

[3] J. M. White and G. D. Rohrer, "Image Thresholding for Optical Character Recognition and Other Applications Requiring Character Image Extraction," in IBM Journal of Research and Development, vol. 27, no. 4, pp. 400-411, July 1983, doi: 10.1147/rd.274.0400.

[4] pdf2text library https://github.com/syllabs/pdf2text

[5] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2007, pp. 629-633, doi: 10.1109/ICDAR.2007.4376991.

[6] Leptonica library for image support https://github.com/DanBloomberg/leptonica

[7] Tesseract4 https://github.com/tesseract-ocr/tesseract