शोध पत्र / Research Paper

प्राकृतिक भाषा प्रसंस्करण और प्रकाशिक संप्रतीक अभिज्ञान प्रक्रियाओं का उपयोग कर एंकर निर्धारण द्वारा विभिन्न इनवॉइस छवियों से स्वचालित डेटा अभिग्रहण तथा डेटा निष्कर्षण

Automatic Data Capture and Data Extraction from Different Invoice Images by Anchor Determination using NLP and OCR Processes

डॉ. विनायक डी. शिंदे[1] एवं स्वरांजली संदीप जगताप[2]

Dr. Vinayak D. Shinde[1] and Swaranjaly Sandeep Jagtap[2]

[1]Associate Professor, Shree L.R. Tiwari College of Engineering, Thane, India -401107

[2]Shree L. R. Tiwari College of Engineering, Thane, India

vdshinde@gmail.com, jagtap41swara@gmail.com

https://doie.org/10.10427/VP.2025527105

सारांश

आज के गतिशील कारोबारी माहौल में, इनवॉइस के भीतर डेटा का प्रभावी प्रसंस्करण और हैंडलिंग किसी भी संबंधित संचालन के अनुकूलन के लिए आवश्यक है। यह शोध पत्र एक ऐसे दृष्टिकोण को प्रस्तुत करता है जिसमें हम प्रकाशिक संप्रतीक अभिज्ञान (Optical Character Recognition (OCR)) और प्राकृतिक भाषा प्रसंस्करण (Natural Language Processing (NLP))) तकनीकों को नियोजित करके इनवॉइस छवियों से निकालने के बाद डेटा कैप्चरिंग को स्वचालित करते हैं। प्रारंभ में, OCR का उपयोग इनवॉइस छवि को TSV (टैब–सेपरेटेड वैल्यू) फ़ाइल में परिवर्तित करने के लिए किया जाता है, आगे इसे व्यवस्थित विश्लेषण के उद्देश्य से JSON ऑब्जेक्ट में परिवर्तित किया जाता है। इनवॉइस के भीतर प्रासंगिक डेटा फ़ील्ड की सटीक पहचान करने के लिए हम रेगेक्स (नियमित अभिव्यक्ति) के माध्यम से एंकर निर्धारण का उपयोग करते हैं। संगठनों, स्थानों, लोगों जैसे पूर्वनिर्धारित संस्थाओं के साथ निकाले गए डेटा को मान्य करने के लिए compromise.js प्राकृतिक भाषा प्रसंस्करण लाइब्रेरी का उपयोग किया जाता है। इसके अलावा हम इस डेटा को एक इकाई सेट में निकालते हैं और फ़िल्टर करते हैं। इस डेटा को CSV फ़ाइल में सहेजा जा सकता है ताकि निकट भविष्य में इसका उपयोग कुछ डेटाबेस के साथ एकीकृत करने के लिए किया जा सके। यह पद्धति दो इनवॉइस लेआउट के लिए लागू की जाती है, ताकि न म्निलिखित मापदंडों के आधार पर इसकी सटीकता की जांच की जा सके। जब हम बड़ी संख्या में इनवॉइस का उपयोग करते हैं तो ये प्रतिशत भिन्न हो सकते हैं। चरित्र स्तर सटीकता: 99%, शब्द–स्तर सटीकता: 98.5%, लाइन–स्तर सटीकता: 97%। हालांकि, सीमा रेगेक्स पैटर्न में निहित है जो विशिष्ट है और सामान्य नहीं है। इसलिए, केवल उन इनवॉइस के लिए प्रभावी है जिनका एक विशेष प्रारूप है। इससे पता चलता है कि भविष्य में मशीन लर्निंग और प्राकृतिक भाषा प्रसंस्करण तकनीकों के साथ अधिक चर या चौतरफा रेगेक्स पैटर्न लागू किए जा सकते हैं ताकि विभिन्न प्रारूपों और लेआउट से इनवॉइस डेटा के लिए सटीक परिणाम प्राप्त किए जा सकें।

Abstract

In today's dynamic business environment, effective processing and handling of data within invoices is essential for optimizing any related operations. This research paper presents an approach where we automate data capturing followed by extracting from invoice images by employing Optical Character Recognition (OCR) and Natural Language Processing (NLP) techniques. Initially, OCR is utilized for converting the invoice image into a TSV (Tab-Separated Values) file, further it is

converted to a JSON object for the purpose of systematic analysis. For accurately identifying relevant data fields within the invoice, we use anchor determination through Regex (regular expressions). In order to validate the data extracted with predefined entities like organizations, places and people ensuring the compromise.js NLP library is used. Further, we extract and filter this data into an entity set. This data can be saved in a CSV file so that it can be used in the near future to be integrated with some databases. This methodology is applied for two invoice layouts, to examine its accuracy based on the following parameters. These percentages may vary when we use large number of invoices. Character level accuracy: 99%, Word-level accuracy: 98.5%, Line-level accuracy: 97%.However, the limitation lays in the Regex pattern, which is specific and not general. Hence, only effective for invoices which have a particular format. This suggests that in future, more variable or all-round regex patterns can be implemented along with Machine learning and Natural language processing techniques so that accurate results can be achieved for invoice data from varied formats and layouts.

मुख्य शब्दः इनवॉइस छवियाँ, प्रकाशिक संप्रतीक अभिज्ञान (OCR), डेटा अभिग्रहण, डेटा निष्कर्षण, प्राकृतिक भाषा प्रसंस्करण (NLP), डेटा रूपांतरण, एंकर निर्धारण, नियमित अभिव्यक्ति आकृति।

Keywords : Invoice Images, Optical Character Recognition (OCR), Data capture, Data extraction, Natural Language Processing (NLP), Data transformation, Anchor determination, Regular Expression Pattern.

परिचय

एक इनवॉइस ऐसा दस्तावेज है जो एक विक्रेता द्वारा खरीदार को जारी किया जाता है जिसमें उसी के लिए भुगतान का अनुरोध करने के दृष्टिकोण के साथ प्रदान की गई सेवाओं या सामानों के बारे में विशिष्टताएं होती हैं। यह भुगतान के संबंध में एक औपचारिक अनुरोध की तरह काम करता है, यह रिकॉर्ड रखने और अन्य व्यवसाय प्रबंधन कार्यों के लिए उपयोगी है। इनवॉइस, कंपनियों और संगठनों के लिए महत्वपूर्ण दस्तावेज हैं क्योंकि वे खरीद के बारे में सबूत/प्रमाण के रूप में कार्य करते हैं, तथा कर और लेखांकन के उद्देश्य के लिए आवश्यक हैं। इन दिनों इनवॉइस अक्सर अधिकांश संगठनों में होते हैं, जिससे उत्पन्न होने वाले कई इनवॉइस को कुशलतापूर्वक संभालने की आवश्यकता होती है [3]। एक इनवॉइस प्रसंस्करण की मैनुअल हैंडलिंग स्पष्ट रूप से समय लेने वाली है और मानवीय त्रुटि के लिए प्रवण है। इसलिए, इनवॉइस से डेटा को पकड़ने और निकालने और भविष्य में उपयोग के लिए इस उपयोगी जानकारी को उपलब्ध कराने के लिए एक स्वचालित प्रणाली की आवश्यकता उत्पन्न होती है। डिजिटल दस्तावेजों के आगमन से प्रकाशिक संप्रतीक अभिज्ञान (OCR) जैसी विभिन्न महत्वपूर्ण तकनीकों का उदय हुआ। प्रकाशिक संप्रतीक अभिज्ञान (OCR) तकनीक का उपयोग करके डिजिटल दस्तावेजों को मशीन पठनीय प्रारूप में बदलना संभव है। प्रकाशिक संप्रतीक अभिज्ञान (OCR) पाठ्य डेटा में अनुक्रम की भविष्यवाणी के लिए चरित्र दीर्घ अल्पकालिक स्मृति (Long Short Term Memory (LSTM)) और आवर्तक तंत्रिका जालतंत्र (Recurrent Neural Network (RNN)) के उद्देश्य से संवलयी तंत्रिका जालतंत्र (Convulsive Neural Network (CNN)) जैसे गहन शिक्षण मॉडल का उपयोग करता है [8][10]। प्राकृतिक भाषा प्रसंस्करण (NLP) मशीनों को प्राकृतिक या मानवीय भाषा में डेटा का विश्लेषण और प्रसंस्करण करने में सक्षम बनाता है। यह शोध एक ऐसे दृष्टिकोण को प्रस्तावित करता है जो इनवॉइस डेटा को पकड़ने और निकालने के लिए एंकर निर्धारण के साथ–साथ प्रकाशिक संप्रतीक अभिज्ञान (OCR) और प्राकृतिक भाषा प्रसंस्करण (Natural Language Processing) का उपयोग करता है और इस डेटा को एक्सेल शीट में संग्रहीत करता है जहां इसे विश्लेषण या रिकॉर्ड बनाए रखने के उद्देश्य के लिए आगे इस्तेमाल किया जा सकता है। इसके अलावा, इसका उपयोग डेटाबेस और अन्य व्यावसायिक प्रक्रियाओं के साथ एकीकरण के

लिए किया जा सकता है। एंकर उस डेटा से पहले के मुख्य शब्द या वे शब्द हैं जिन्हें हम निकालना चाहते हैं। वे दस्तावेज़ के भीतर आवश्यक डेटा का पता लगाने में मदद करते हैं। पैटर्न मिलान कार्यों के लिए नियमित अभिव्यक्ति (रेगेक्स) का प्रभावी ढंग से उपयोग किया जा सकता है। रेगेक्स, नियमित अभिव्यक्तियों के लिए छोटा, स्ट्रिंग मिलान और स्ट्रिंग हेरफेर के लिए एक व्यापक उपकरण है। इस शोध पत्र का संगठन निम्नानुसार है: परिचय अनुभाग, साहित्य समीक्षा, प्रस्तावित सिस्टम एल्गोरिथ्म तथा अंतिम खंड निष्कर्ष और भविष्य के दायरे के साथ कार्यान्वयन और परिणाम हैं।

संबंधित कार्य

बेयर एवं अन्य [1] ने एक ऐसी प्रणाली प्रस्तुत की जो विभिन्न डोमेन में अलग–अलग लेआउट के साथ इनवॉइस से डेटा के निष्कर्षण को स्वचालित करती है। सिस्टम में दस्तावेज़ों को डिजिटाइज़ करने के लिए एक OCR टूल और FRESCO शामिल है, जो डोमेन–विशिष्ट नियमों के आधार पर मॉडलिंग और सामग्री निकालने की भाषा है। यह प्रणाली स्वास्थ्य बीमा इनवॉइस के लिए 1% से कम त्रुटि के साथ 50% से अधिक स्वचालन प्राप्त करती है, उच्च दक्षता और अनुकूलन क्षमता का प्रदर्शन करती है। रेड्डी एवं अन्य [2] ने गहरी शिक्षण तकनीकों का उपयोग करके हस्तलिखित पाठ पहचान की आवश्यकता को संबोधित किया। दृष्टिकोण शब्द वर्गीकरण के लिए सीएनएन और चरित्र विभाजन के लिए लॉन्ग शॉर्ट टर्म मेमोरी (LSTM) नेटवर्क का उपयोग करता है, ईएमएनआईएसटी (EMNIST) डेटासेट के साथ उच्च सटीकता प्राप्त करता है। प्रस्तावित प्रणाली डेटा सुरक्षा और पहुंच में सुधार के लिए, विशेष रूप से ऐतिहासिक दस्तावेज़ संरक्षण और बड़े पैमाने पर डिजिटलीकरण प्रयासों में महत्वपूर्ण क्षमता प्रदर्शित करती है। शि एवं अन्य [3] ने अधिक स्पष्ट चित्र प्राप्त करने के लिए बैचों में स्कैन किए गए इनवॉइस को, स्थान जानकारी टेम्पलेट मिलान एल्गोरिथ्म के लिए उपयोग किया था, चरित्र पहचान सीएनएन का उपयोग करके किया गया था जिसने मान्यता दर में सुधार किया जो 99.8% तक पहुंच सकता है। स्वयं निर्णय और हस्तनिर्देश सुधार की एक प्रक्रिया, मान्यता सटीकता के निर्धारण के लिए वैट (VAT) इनवॉइस के अंदर जानकारी के सहसंबंध का उपयोग करके जोड़ा जाता है। औदिच्य एवं अन्य [4] ने डिजिटल रूप से टाइप की गई मुद्रित छवियों से वर्णों को पहचानने के लिए टेसरैक्ट द्वारा प्रदान किए गए अंतर्निहित प्रशिक्षित परीक्षण डेटा के साथ टेसरैक्ट प्रकाशिक संप्रतीक अभिज्ञान (OCR) इंजन का उपयोग करके गुजराती वर्ण पहचाना। विभिन्न स्थितियों जैसे कि विभिन्न फ़ॉन्ट शैलियों, फ़ॉन्ट आकार आदि का उपयोग करना, के लिए अवलोकन किया गया था। यह कुछ जटिल पात्रों और समान दिखने वाले पात्रों की अस्पष्टता के साथ मामलों के अलावा अच्छा प्रदर्शन करने वाला पाया गया था। औंग एवं अन्य [5] ने छवि पहचान प्रणाली में फीचर निष्कर्षण और वर्गीकरण कार्यों के लिए दृढ़ तंत्रिका नेटवर्क (CNN) के साथ एक प्रणाली का प्रस्ताव दिया। अनुक्रम–से–अनुक्रम कार्यों के लिए आवर्तक तंत्रिका नेटवर्क (RNN), जैसे इनवॉइस छवियों से पाठ्य जानकारी को पहचानना और निकालना। डेटासेट के रूप में वास्तविक दुनिया के इनवॉइस की एक विस्तृत विविधता का उपयोग किया गया था, सिस्टम ने इनवॉइस छवियों से प्रासंगिक जानकारी निकालने में 95% की उच्च स्तर की सटीकता का प्रदर्शन किया। हालांकि, इनवॉइस के लिए मामूली प्रदर्शन गिरावट देखी गई थी जिसमें अपरंपरागत लेआउट और खराब छवि गुणवत्ता थी। सातव एवं अन्य [6] ने एक वेब आधारित अनुप्रयोग विकसित किया। यह विकसित अनुप्रयोग एक तीन चरण प्रक्रिया का पालन करता है जहां विभिन्न ओपनसीवी तकनीकों का उपयोग करके पूर्व प्रसंस्करण किया जाता है, फिर प्रकाशिक संप्रतीक अभिज्ञान (OCR) तकनीक का उपयोग करके डेटा निष्कर्षण किया जाता है और प्रसंस्करण के बाद नियमित अभिव्यक्ति आकृति का उपयोग किया जाता है। सिस्टम खुदरा विक्रेताओं और थोक विक्रेताओं के लिए विकसित किया गया है ताकि वे अपने व्यावसायिक डेटा को अधिक कुशलता से प्रबंधित कर सकें।

सिधवा एवं अन्य [7] ने पता लगाने के उद्देश्य से पायथन में एक वेबएप विकसित किया। एज डिटेक्शन कैनी एज डिटेक्शन एल्गोरिथ्म का उपयोग करके किया जाता है, कंटूर डिटेक्शन के लिए ओपनसीवी के कंटूर () फ़ंक्शन का उपयोग किया जाता है, इसके बाद चार पॉइंट पर्सपेक्टिव ट्रांसफॉर्म का उपयोग किया जाता है। छवि को सीमाओं के साथ क्रॉप किया गया है और प्रकाशिक संप्रतीक अभिज्ञान (OCR) के माध्यम से पारित किया गया है। मल्लाधी एवं अन्य [8] ने एआई–प्रकाशिक संप्रतीक अभिज्ञान (OCR) के साथ आरएल के समामेलन का सुझाव दिया है | इनवॉइस के विभिन्न बदलते प्रारूपों के लिए आरएल की अनुकूलन क्षमता के विभिन्न लाभों को इंगित करके इनवॉइस प्रसंस्करण में आरएल के उपयोग पर जोर देता है, लगातार पुरस्कार या दंड के आधार पर अपनी रणनीतियों को परिष्कृत करता है इसलिए डेटा के साथ विकसित करने में सक्षम होने के कारण प्रक्रिया, स्केलेबिलिटी के रूप में कंपनी बड़ी संख्या में इनवॉइस संसाधित करने के लिए बढ़ेगी। यह प्रणाली प्रभावी रूप से स्केल कर सकती है और बड़े डेटासेट और इससे भी अधिक विविध स्वरूपों के अनुकूल है। दोषी एवं अन्य [9] ने UiPath का उपयोग करके एक प्रणाली विकसित की है। प्रारंभिक चरण डेटा स्क्रैपिंग है | इनवॉइस फ़ाइलों को ईमेल से पुर्नप्राप्त किया जाता है और एक फ़ोल्डर में संग्रहीत किया जाता है। इनवॉइस डेटा फ़ील्ड UiPath से पुर्नप्राप्त किए जाते हैं और डेटाबेस में एप्लिकेशन स्टोर में इनपुट होते हैं। वाक् पहचान वाला चैटबॉट वेबसाइट में एकीकृत है। लीं एवं अन्य [10] ने हस्तलिखित रसीदों और मुद्रित रसीदों दोनों के लिए एक प्रणाली विकसित की। हस्तलिखित रसीदों के लिए, रसीदों और सीएनएन की टेम्पलेट मिलान और निश्चित सुविधाओं का उपयोग किया जाता है। हस्तलिखित रसीद वर्णों के लिए सिस्टम की पहचान सटीकता दर 80.93% थी जबकि मुद्रित रसीद के लिए 99.39% थी।

प्रस्तावित एल्गोरिथ्म

इस अध्ययन का उद्देश्य एक स्वचालित प्रणाली को लागू करना है जो एक इनवॉइस छवि से डेटा को प्राप्त करती है और इसे एक्सेल शीट में सटीक रूप से संग्रहीत करती है, इस प्रकार इसे डेटाबेस के साथ आगे एकीकरण के लिए उपयोग करने योग्य बनाती है। हम एंकर निर्धारण और प्राकृतिक भाषा प्रसंस्करण (NLP) और प्रकाशिक संप्रतीक अभिज्ञान (OCR) को नियोजित करते हैं। प्रकाशिक संप्रतीक अभिज्ञान (OCR) के साथ, इनवॉइस से मुद्रित या हस्तलिखित पाठ मशीन–पठनीय डेटा में परिवर्तित हो जाता है, एनएलपी निकाली गई जानकारी को सही ढंग से समझने और वर्गीकृत करने में मदद करता है। इस प्रणाली के विकास को बढ़ाने के लिए, हमने जावास्क्रिप्ट पुस्तकालयों का उपयोग किया है। जावास्क्रिप्ट और Node.js रनटाइम वातावरण के रूप में उपयोग किए जाते हैं। सिस्टम एक इंटेल i3 प्रोसेसर, 7 वीं पीढ़ी, 8GB रैम के साथ संचालित है। चित्र 1 प्रस्तावित सिस्टम फ्रेमवर्क दिखाता है। इसमें निम्नलिखित चरण शामिल हैं: i) इनवॉइस छवि, ii) प्रकाशिक संप्रतीक अभिज्ञान (OCR) प्रणाली, iii) डेटा परिवर्तन, iv) फ़ाइल सफाई, v) प्रासंगिक डेटा फ़ील्ड की पहचान, vi) डेटा सत्यापन, vii) निकाली गई जानकारी को सहेजना, viii) एक्सेल शीट में स्टोर करना। निम्नलिखित चरण प्रस्तावित सिस्टम फ्रेमवर्क को दर्शाते हैं:

(1) छवि (इनपुट):

एक इनवॉइस एक विक्रेता द्वारा एक खरीदार को जारी किया गया एक विस्तृत दस्तावेज है, जो प्रदान की गई वस्तुओं या सेवाओं, उनकी मात्रा, कीमतों, करों और कुल देय राशि को रेखांकित करता है। इस प्रणाली में, इनवॉइस छवि को इनपुट के रूप में पारित किया जाता है जिस पर आगे की कार्रवाई की जाती है।

(2) प्रकाशिक संप्रतीक अभिज्ञान (OCR) प्रणाली (runOcr()):

tesseract.js पैकेज का उपयोग प्रकाशिक संप्रतीक अभिज्ञान (OCR) के लिए किया जाता है। फ़ंक्शन runOcr() इनवॉइस छवि को संसाधित करता है और संगत TSV (टैब–सेपरेटेड वैल्यू) फ़ाइल जनरेट करता है। यह रूपांतरण शब्द और चरित्र विभाजन के

बाद शोर में कमी, बिनाराइजेशन, डेस्कविंग जैसे चरणों की एक श्रृंखला को शामिल करता है। यह TSV प्रारूप सारणीबद्ध डेटा फ़ाइल संग्रहीत करने के लिए पाठ–आधारित फ़ाइल स्वरूप है। यह सारणीबद्ध डेटा को स्पष्ट रूप से प्रस्तुत करने में मदद करता है।

(3) डेटा परिवर्तन (tsvtotext(), fs मॉड्यूल):

फ़ंक्शन tsvtotext() TSV फ़ाइल को संभालता है, जिसमें एक संरचित, सारणीबद्ध प्रारूप में छवि से निकाला गया डेटा होता है और इस डेटा को एक पाठ फ़ाइल में परिवर्तित करता है, यह सुनिश्चित करता है कि यह पठनीय बना रहे और मूल छवि डेटा के समान प्रारूप बनाए रखे। यह फ़ंक्शन TSV डेटा को सादे पाठ प्रारूप में परिवर्तित करने के लिए Node.js fs मॉड्यूल का उपयोग करता है।

(4) डेटा परिवर्तन (tsvtojsonobject(), fs मॉड्यूल):

TSV फ़ाइल को JSON (JavaScript Object Notation) फॉर्मेट में बदला जाता है। tsvtojsonobject() फंक्शन के अंदर, आवश्यक जानकारी छवि–पाठ फाइल से निकाली जाती है। इसमें प्रमुख विवरण जैसे इनवॉइस संख्या, खरीद आदेश संख्या, बिलिंग पता, शिपिंग पता और कुल राशि शामिल है।

(5) प्रासंगिक डेटा फ़ील्ड (एंकर निर्धारण) की पहचान करना:

एंकर निर्धारण की मदद से इनवॉइस छवि के भीतर प्रासंगिक डेटा फ़ील्ड की पहचान करता है। यह स्ट्रिंग मिलान का उपयोग करता है। स्ट्रिंग मिलान के लिए नियमित अभिव्यक्तियों का उपयोग एक प्रभावी उपकरण के रूप में किया जाता है।

(6) फाइल सफाई (रेगेक्स):

इस प्रक्रिया में अनावश्यक वर्णों को हटाना, प्रारूपों को मानकीकृत करना और सटीक विश्लेषण और प्रसंस्करण के लिए डेटा अखंडता सुनिश्चित करना शामिल है। रेगेक्स अवांछित वर्णों को हटाने में मदद करता है, जैसे गैर–अल्फान्यूमेरिक प्रतीक और पाठ स्थिरता सुनिश्चित करते हुए अतिरिक्त रिक्त स्थान को साफ करते हैं। यह त्रुटियों को कम करना और दक्षता बढ़ाना सुनिश्चित करता है।

(7) डेटा सत्यापन (एनएलपी प्रक्रियाएं):

डेटा के प्रत्येक टुकड़े को क्रमिक रूप से निकालने पर, यह सफाई से गुजरता है और मुख्य इकाई सेट में एकीकृत होता है। डेटा निष्कर्षण प्रक्रिया के दौरान, compromise.js एनएलपी लाइब्रेरी का उपयोग विशिष्ट डेटा श्रेणियों जैसे व्यक्तियों, स्थानों और संगठनों को मान्य करने के लिए किया जाता है। Compromise.js एक Node.js और ब्राउज़र–आधारित प्राकृतिक भाषा प्रसंस्करण लाइब्रेरी है। यह अंग्रेजी पाठ को पार्स करने, समझने और हेरफेर करने के लिए उपकरणों का एक सेट प्रदान करता है। यह सत्यापन निकाली गई जानकारी की सटीकता सुनिश्चित करता है। निकाले गए और फ़िल्टर किए गए डेटा को एक निकाय सेट में समेकित किया जाता है |

(8) निकाली गई जानकारी सहेजना (CSV फाइल में सम्मिलन):

निकाले गए डेटा को संसाधित और मान्य करने के बाद, इसे एक एकल इकाई सेट में संकलित किया जाता है, जो इनवॉइस छवि से निकाली गई सभी महत्वपूर्ण जानकारियों को एक संरचित प्रारूप में समेकित करता है। यह संरचित डेटा एक CSV (अल्पविराम से अलग मान) में संग्रहीत किया जाता है, यह सुनिश्चित करता है कि संसाधित इनवॉइस डेटा भविष्य के संदर्भ और विश्लेषण के लिए सुलभ और उपयोग करने योग्य बना रहे।

(9) आउटपुट (एक्सेल शीट में सहेजें):

मान्य और संरचित इनवॉइस डेटा एक एक्सेल शीट में संग्रहीत किया जाता है जो विभिन्न व्यावसायिक कार्यों के साथ व्यापक विश्लेषण, रिपोर्टिंग और सहज एकीकरण को सक्षम करता है। यह क्षमता सुनिश्चित करती है कि निकाली गई जानकारी वित्तीय ऑडिट, पूर्वानुमान और अनुपालन उद्देश्यों के लिए आसानी से सुलभ है।

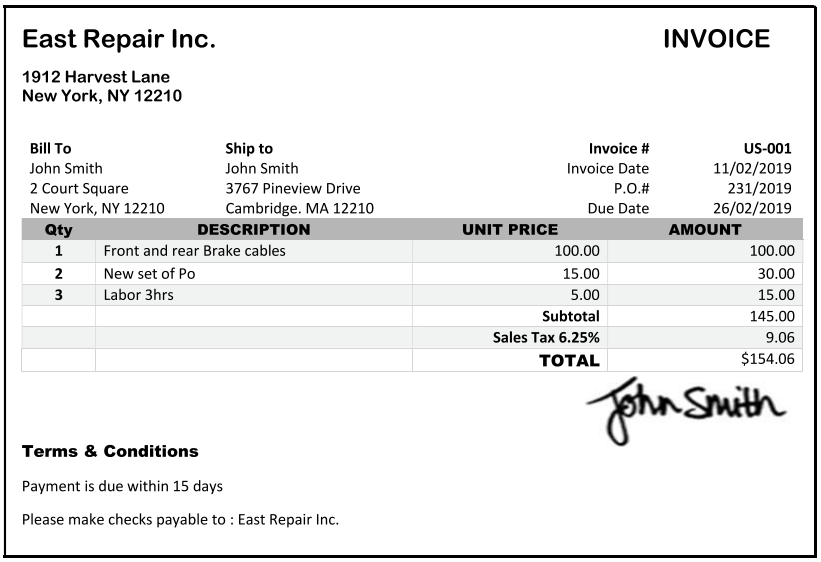
चित्र 1: प्रस्तावित प्रणाली पद्धति

एल्गोरिथ्म		
निवेश	:	इनवॉइस छवि
आउटपुट	:	इनवॉइस डेटा को Excel पत्रक में संग्रहीत करें
चरण 1	:	छवि को TSV फ़ाइल में बदलें
चरण 2	:	TSV फ़ाइल स्वरूप को पाठ फ़ाइल स्वरूप में बदलें
चरण 3	:	पिछले चरण में प्राप्त टेक्स्ट फ़ाइल को JSON ऑब्जेक्ट में बदलें।
चरण 4	:	एंकर निर्धारण का उपयोग करके प्रासंगिक डेटा फ़ील्ड की पहचान करें।
चरण 5	:	नियमित अभिव्यक्ति आकृति (रेगेक्स) का उपयोग करके डेटा सफाई करें।
चरण 6	:	एनएलपी प्रक्रिया का उपयोग करके डेटा सत्यापन करें।
चरण 7	:	इस निकाय सेट डेटा को CSV फ़ाइल में डालें।
चरण 8	:	आगे उपयोग के लिए इस डेटा को एक्सेल शीट में स्टोर करें।

कार्यान्वयन और परिणाम

प्रस्तावित प्रणाली में, डेटा को इनवॉइस से निकाला जाता है और निकालने के बाद इस डेटा को आगे उपयोग और प्रसंस्करण के लिए एक्सेल शीट में संग्रहीत किया जाता है। चित्र 2 एक नमूना इनवॉइस छवि दिखाता है। इनवॉइस इमेज इनपुट होने के बाद, पहला कदम OCRing का उपयोग करके इनवॉइस इमेज को TSV फ़ाइल में बदलना है। इसके लिए tesseract.js पैकेज का उपयोग किया जाता है। runOcr() विधि एक TSV (टैब–सेपरेटेड वैल्यू) फ़ाइल स्वरूप बनाने के लिए इनवॉइस छवि को संसाधित करती है, जो सारणीबद्ध डेटा को संग्रहीत करने के लिए उपयोग किया जाने वाला एक साधारण पाठ फ़ाइल स्वरूप है। चित्र 3 इनपुट इनवॉइस के द्वारा उत्पन्न संबंधित TSV फ़ाइल दिखाता है। इस फ़ाइल स्वरूप में पाठ की स्थिति और पदानुक्रमित लेआउट के बारे में जानकारी होती है। इसके अलावा आत्मविश्वास स्कोर की गणना की जाती है। जिसका मान 0 और 100 के बीच होता है,जो मान्यता प्राप्त छवि की सटीकता के बारे में निश्चितता की सीमा को इंगित करता है। उदाहरण, कुछ छवि 'ईस्ट रिपेयर इंक' के लिए आत्मविश्वास स्कोर 97 निकलता है, तो इसका मतलब है कि प्रकाशिक संप्रतीक अभिज्ञान (OCR) इंजन इस पाठ की मान्यता की सटीकता के बारे में 97% आश्वस्त है। इनवॉइस में पाठ का स्थानिक लेआउट बनाए रखा जाता है जो तालिकाओं और सूचियों का सटीक प्रतिनिधित्व करने के लिए महत्वपूर्ण है। छवि फ़ाइल से TSV फ़ाइल में रूपांतरण को पाटना चरणों की एक श्रृंखला है जैसे कि छवि प्रीप्रोसेसिंग जिसमें शोर में कमी जैसे विभिन्न चरण शामिल हैं, जो पाठ पहचान प्रक्रिया में बाधा डाल सकते हैं। बाइनरीकरण, जो छवि को काले और सफेद में परिवर्तित कर रहा है। डेस्क्यूइंग

जो छवि में किसी भी झुकाव को ठीक कर रहा है। अगले चरण में, यह TSV फ़ाइल एक टेक्स्ट फ़ाइल में कनवर्ट हो जाती है। फ़ंक्शन tsvtotex(), TSV फ़ाइल का उपयोग करता है। जिसमें पिछले चरण से सारणीबद्ध प्रारूप में छवि से निकाला गया डेटा होता है और इसे संरचित प्रारूप के साथ टेक्स्ट फ़ाइल में परिवर्तित करती है। यहां fs (फाइल सिस्टम) लाइब्रेरी का उपयोग फाइलों को पढ़ने और लिखने के लिए किया जाता है। ऐसा इसलिए किया गया ताकि TSV फ़ाइल की सामग्री को अधिक पठनीय टेक्स्ट प्रारूप में परिवर्तित किया जा सके। चित्र 4 उत्पन्न संबंधित पाठ फ़ाइल दिखाता है।



East Repair Inc. **INVOICE**

1912 Harvest Lane
New York, NY 12210

Bill To	Ship to	Invoice #	US-001
John Smith	John Smith	Invoice Date	11/02/2019
2 Court Square	3767 Pineview Drive	P.O.#	231/2019
New York, NY 12210	Cambridge. MA 12210	Due Date	26/02/2019

Qty	DESCRIPTION	UNIT PRICE	AMOUNT
1	Front and rear Brake cables	100.00	100.00
2	New set of Po	15.00	30.00
3	Labor 3hrs	5.00	15.00
		Subtotal	145.00
		Sales Tax 6.25%	9.06
		TOTAL	$154.06

John Smith

Terms & Conditions

Payment is due within 15 days

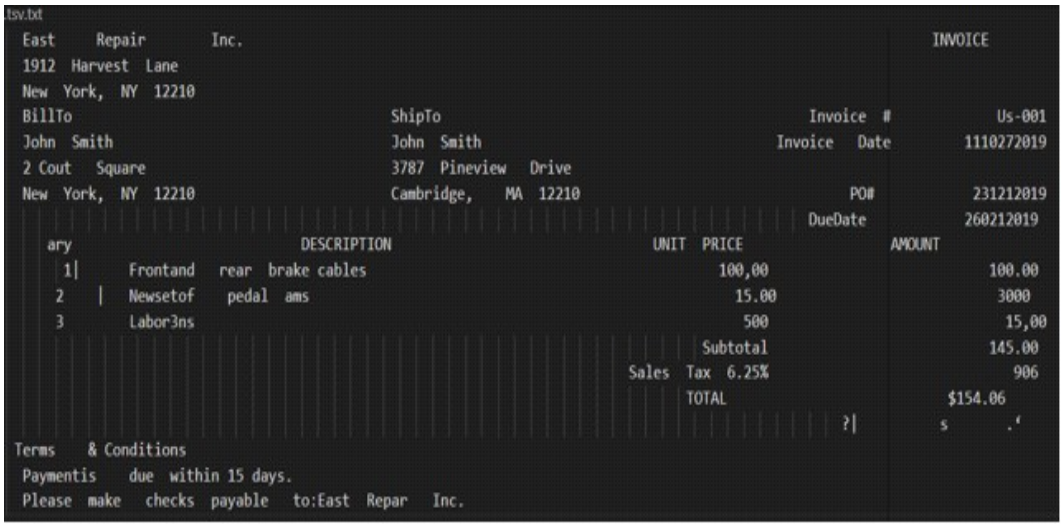
Please make checks payable to : East Repair Inc.

चित्र 2: नमूना इनवॉइस छवि (इनपुट)

files > invoice.png.tsv

```
1   1   1   0   0   0   0   0   0   750 1061    -1
2   2   1   1   0   0   0   54  72  642 919 -1
3   3   1   1   1   0   0   54  72  642 919 -1
4   4   1   1   1   1   0   56  72  638 20  -1
5   5   1   1   1   1   1   56  73  41  15  92.319328   East
6   5   1   1   1   1   2   104 73  62  19  91.892830   Repair
7   5   1   1   1   1   3   172 73  33  15  88.638100   Inc.
8   5   1   1   1   1   4   613 72  81  16  92.179382   INVOICE
9   4   1   1   1   2   0   56  115 105 10  -1
10  5   1   1   1   2   1   56  115 26  10  93.212112   1912
11  5   1   1   1   2   2   87  115 42  10  82.004639   Harvest
12  5   1   1   1   2   3   134 115 27  10  91.880775   Lane
13  4   1   1   1   3   0   55  133 117 11  -1
14  5   1   1   1   3   1   55  133 25  10  92.515610   New
15  5   1   1   1   3   2   83  133 29  11  72.773911   York,
16  5   1   1   1   3   3   117 133 17  10  90.714806   NY
17  5   1   1   1   3   4   138 133 34  10  90.122070   12210
18  4   1   1   1   4   0   55  243 639 13  -1
19  5   1   1   1   4   1   55  243 38  10  53.891293   BillTo
20  5   1   1   1   4   2   282 243 46  13  33.965225   ShipTo
21  5   1   1   1   4   3   538 243 43  10  54.111404   Invoice
22  5   1   1   1   4   4   584 243 8   10  92.367500   #
23  5   1   1   1   4   5   654 243 40  10  78.898384   Us-001
24  4   1   1   1   5   0   55  265 641 14  -1
25  5   1   1   1   5   1   55  265 26  10  82.970634   John
26  5   1   1   1   5   2   86  265 31  10  84.745361   Smith
```

चित्र 3: इनपुट इनवॉइस के द्वारा उत्पन्न संबंधित TSV फ़ाइल


tsv.txt

East Repair Inc. INVOICE
1912 Harvest Lane
New York, NY 12210

BillTo	ShipTo	Invoice #	Us-001
John Smith	John Smith	Invoice Date	1110272019
2 Cout Square	3787 Pineview Drive		
New York, NY 12210	Cambridge, MA 12210	PO#	231212019
		DueDate	260212019

ary	DESCRIPTION	UNIT PRICE	AMOUNT
1	Frontand rear brake cables	100,00	100.00
2	Newsetof pedal ams	15.00	3000
3	Labor3ns	500	15,00
		Subtotal	145.00
		Sales Tax 6.25%	906
		TOTAL	$154.06

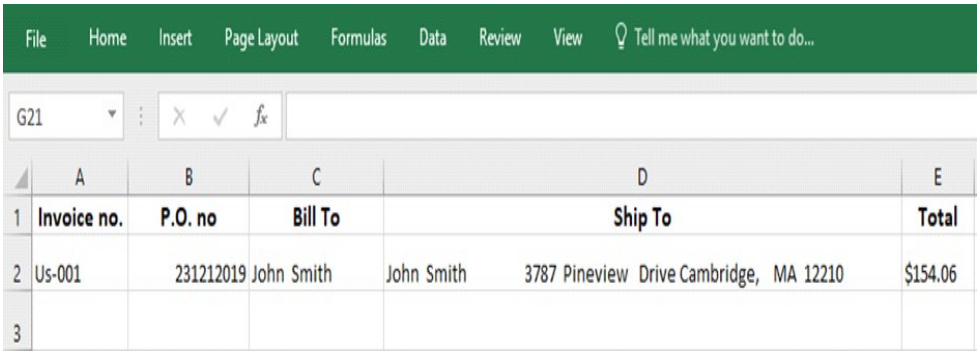
?| s .'

Terms & Conditions
Paymentis due within 15 days.
Please make checks payable to:East Repar Inc.

चित्र 4: उत्पन्न संबंधित पाठ फ़ाइल

इसके बाद, tsvtojsonobject() फ़ंक्शन छवि–पाठ फ़ाइल का उपयोग करता है और उपयोगी डेटा जैसे इनवॉइस संख्या, क्रय आदेश संख्या, बिल का पता, शिप का पता, एंकर टेक्स्ट पर आधारित कुल राशि निकालता है। एंकर पाठ के भीतर भरोसेमंद मार्कर के रूप में काम करता है। एंकर निर्धारण का उपयोग महत्वपूर्ण डेटा फ़ील्ड जैसे इनवॉइस नंबर, खरीद ऑर्डर नंबर, बिलिंग टू, शिपिंग एड्रेस और कुल राशि की पहचान करने के लिए किया जाता है। यह JSON ऑब्जेक्ट पंक्ति पाठ और रेगेक्स पैटर्न के बीच स्ट्रिंग मिलान के साथ किया जाता है। पैटर्न मैच होते ही एंकर की पहचान कर ली गई है। इस एंकर के बाद इस एंकर से संबंधित वास्तविक मूल्य होगा, यानी इनवॉइस डेटा फ़ील्ड का वास्तविक मूल्य उदाहरण के लिए इनवॉइस में वास्तविक इनवॉइस संख्या यूएस–001। इनवॉइस में अन्य प्रासंगिक क्षेत्रों को स्थितीय तर्क का उपयोग करके पहचाना जाता है, जो एंकर के संबंध में इनवॉइस में अन्य आस–पास के डेटा फ़ील्ड का पता लगा रहा है। अगली फ़ाइल सफाई, नियमित अभिव्यक्ति आकृति का उपयोग करके की जाती है। इस चरण में अवांछित वर्णों को हटा दिया जाता है, प्रारूपों को मानकीकृत किया जाता है, किसी भी अतिरिक्त सफेद रिक्त स्थान को ठीक किया जाता है। यह सुनिश्चित करने के लिए किया जाता है कि डेटा साफ और प्रयोग करने योग्य है। अगला चरण डेटा सत्यापन है। एनएलपी लाइब्रेरी विशिष्ट डेटा श्रेणियों, जैसे व्यक्तियों, स्थानों और संगठनों को मान्य करने के लिए compromise.js है। यह चरण निकाली गई जानकारी की सटीकता और विश्वसनीयता सुनिश्चित करता है। एक बार मान्य होने के बाद, साफ और फ़िल्टर किए गए डेटा को एक एकल इकाई सेट में समेकित किया जाता है, जिसका उपयोग अन्य व्यावसायिक प्रणालियों के साथ आगे की प्रक्रिया और एकीकरण के लिए किया जा सकता है। यह डेटा CSV (अल्पविराम से अलग मान) फ़ाइल में संग्रहीत किया जाता है। आगे हम इस डेटा को एक्सेल शीट में स्टोर करते हैं। CSV–लेखक लाइब्रेरी का उपयोग निकाले गए और संसाधित डेटा को CSV फाइलों में लिखने के लिए किया जाता है। यह निकाले गए इनवॉइस डेटा को आगे के विश्लेषण और रिकॉर्ड–कीपिंग के लिए संरचित प्रारूप में संग्रहीत करने के लिए उपयोगी है (चित्र 5)। हमने अपने रेगेक्स पैटर्न को दो इनवॉइस लेआउट पर भी लागू किया और निम्नलिखित सटीकता हासिल की: चरित्र–स्तर सटीकता: 99%, शब्द–स्तर सटीकता: 98.5%, लाइन–स्तरीय सटीकता: 97%। जब हम बड़ी संख्या में इनवॉइस का उपयोग करते हैं, तो ये प्रतिशत भिन्न हो सकते हैं। हमारा रेगेक्स पैटर्न विशिष्ट संरचना या लेआउट के प्रारूप के लिए सिलवाया गया है। ऐसा इसलिए है क्योंकि प्रत्येक इनवॉइस दस्तावेज़ में पाठ, प्रारूपों और समग्र विभिन्न लेआउट की विभिन्न व्यवस्थाएं हो सकती हैं। इनवॉइस सामग्री की नियुक्ति जैसे, पीओ नंबर, कुल राशि,

इनवॉइस संख्या, आदि अलग–अलग इनवॉइस के लिये भिन्न हो सकते हैं विभिन्न व्यवसायों या विक्रेताओं, खुदरा विक्रेताओं के पास आमतौर पर अपनी व्यक्तिगत इनवॉइस संरचना और लेआउट होते हैं। [8] प्रत्येक अलग इनवॉइस संरचना और लेआउट के लिए हमें अलग–अलग रेगेक्स पैटर्न बनाने की आवश्यकता होगी ताकि यह उन लेआउट के लिए अनुकूलित हो।



	A	B	C	D	E
1	**Invoice no.**	**P.O. no**	**Bill To**	**Ship To**	**Total**
2	Us-001	231212019	John Smith	John Smith 3787 Pineview Drive Cambridge, MA 12210	$154.06
3					

चित्र 5: आउटपुटः इनवॉइस डेटा एक्सेल शीट में संग्रहीत

निष्कर्ष और भविष्य की संभावनाएं

प्रस्तावित प्रणाली प्रकाशिक संप्रतीक अभिज्ञान (OCR), प्राकृतिक भाषा प्रसंस्करण (NLP) और एंकर निर्धारण का उपयोग करके इनवॉइस डेटा को कब्जा करती हैं और निकालती है, निकाले गए डेटा को रिकॉर्ड रखने के विश्लेषण और उद्देश्य के लिए एक्सेल शीट में संग्रहीत किया जाता है। यह इनवॉइस डेटा की मैनुअल डेटा प्रविष्टि की आवश्यकता को समाप्त करता है, जो अन्यथा एक समय लेने वाली प्रक्रिया होगी। समग्र इनवॉइस प्रसंस्करण के लिए प्रकाशिक संप्रतीक अभिज्ञान, प्राकृतिक भाषा प्रसंस्करण, एंकर निर्धारण इष्टतम दृष्टिकोण हैं। हमने अपने स्ट्रिंग मैचर को दो इनवॉइस प्रारूप पर लागू किया और मैट्रिक्स पर निम्नलिखित सटीकता हासिल की : चरित्र स्तर सटीकता: 99, शब्द–स्तर सटीकता: 98.5%, लाइन–स्तरीय सटीकता: 97%। भविष्य में, प्रस्तावित प्रणाली को अधिक रेगेक्स पैटर्न को लागू करके सुधार किया जा सकता है, इनवॉइस प्रारूप की विविध श्रेणी पर काम करने में सक्षम होने के लिए लचीलेपन में सुधार किया जा सकता है। इसके अलावा अधिक उन्नत मशीन लर्निंग एल्गोरिथ्म और प्राकृतिक भाषा प्रसंस्करण तकनीकों को लागू किया जा सकता है।

शोध पत्र में प्रयुक्त अंग्रेज़ी शब्दों की समानार्थक हिंदी शब्दावली

Alphabetically sorted terminology in English	**वर्णमाला अनुक्रमित हिंदी शब्दावली**
Convulsive neural network	संवलयी तंत्रिका जालतंत्र
Data capture	डेटा अभिग्रहण
Data extraction	डेटा निष्कर्षण
Natural language processing	प्राकृतिक भाषा प्रसंस्करण
Optical character recognition	प्राकृतिक संप्रतीक अभिज्ञान
Recurrent neural network	आवर्तक तंत्रिका जालतंत्र

संदर्भ

1. Bayer, T. and Mogg-Schneider, H. (1997). A Generic System for Processing Invoices. Published in *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, Ulm, Germany.

2. Reddy, S. and Rao, M. (2020). A Survey on Deep Learning For Hand Written Recognition by Digit & Character. *International Journal of Creative Research Thoughts (IJCRT), V*ol. 8, no. 6, pp. 3310-3314.

3. Shi, S., Cui, C. and Xiao, Y. (2020). An Invoice Recognition System Using Deep Learning. Published in *International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, Chongqing, China.

4. Audichya, M. and Saini, J. (2017). A Study to Recognize Printed Gujarati Characters Using Tesseract OCR. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), V*ol. 5, no. IX, pp. 1505-1510.

5. Aung, T. , Paw, S. and Tin, H.H.K. (2022). Automated Invoice Processing Using Image Recognition in Business Information Systems. *Economics, Commerce and Trade Management: An International Journal (ECTIJ), V*ol. 3, pp. 73-79.

6. Satav, M. S., Varade, T., Kothavale, D., Thombare, S. and Lokhande, P. (2020). Data Extraction From Invoices Using Computer Vision. Presented in *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS).*

7. Sidhwa, H., Kulshrestha, S. and Malhotra, S. (2018). Text Extraction from Bills and Invoices. Presented in *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India.

8. Malladhi, A. (2021). Adaptive Reinforcement Learning in AI-OCR Invoice Processing for the Digital Age. *International Journal of Computer Techniques ,* Vol. 8, no. 6, pp. 1-22.

9. Doshi, P., Kotak, Y. and Sahitya, A. (2020). Automated Invoice Processing System Along with Chatbot. *International Journal of Research in Engineering, Science and Management, V*ol. 3, no. 5, pp. 29-31.

10. Lin, C.-J., Liu, Y.-C and Lee, C.-L. (2022). Automatic Receipt Recognition System Based on Artificial Intelligence Technology. *Applied Sciences ,* vol. 12, no. 2, p. 853, pp.1-22.