

**D.Y. PATIL COLLEGE OF ENGINEERING &
TECHNOLOGY, KOLHAPUR**

(An Autonomous Institute)



DEPARTMENT OF CSE (DATA SCIENCE)

A

**Project Report
on**

“ AUTOMATING INVOICE TEXT EXTRACTION ”

Submitted by

Name

Roll No.

Samruddhi Sanjay Kale

05

Shubha Muralidhar Desai

06

Satej Santosh Kulkarni

07

Akhilesh Shilbuddha Damke

08

KASABA BAWADA, KOLHAPUR

Under the guidance of

Prof.Poonam.Patil

Third Year B. Tech. CSE (Data Science)

Academic Year 2023-24

**D. Y. PATIL COLLEGE OF ENGINEERING &
TECHNOLOGY, KOLHAPUR**

(An Autonomous Institute)



DEPARTMENT OF CSE (DATA SCIENCE)

CERTIFICATE

This is to certify that,

Roll No.	Unique ID	Student Name	Exam Seat No.
05	EN21122724	Samruddhi Sanjay Kale	16745
06	EN21191939	Shubha Muralidhar Desai	16772
07	EN21165406	Satej Santosh Kulkarni	16755
08	EN21148233	Akhilesh Shilbuddha Damke	16668

have successfully completed the project work entitled,

“Automating Invoice Text Extraction”

In partial fulfilment for the curriculum of **T. Y. B. Tech. CSE (Data Science)**. This is the record of their work carried out during academic year 2023-2024.

Date:

Place: Kolhapur

Prof.Poonam.Patil
Project Guide

Prof. DR. G. V. Patil
HoD

Prof. DR. S. D. Chede
Principal

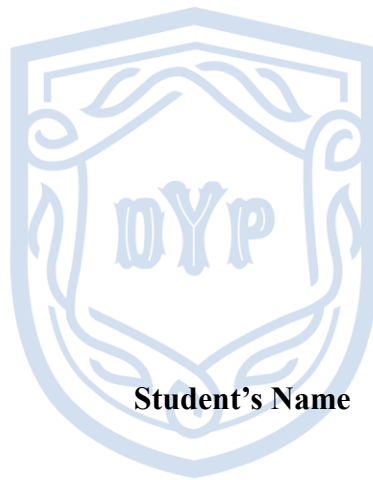
External Examiner

DECLARATION

We the undersigned students of **T. Y. B. Tech. CSE (Data Science)** declare that the project work report entitled “**AUTOMATING INVOICE TEXT EXTRACTION** ” written and submitted by us, under the guidance of **Prof.Poonam Patil** is our original work. The empirical findings in this report are based on the data collected by us. The matter assimilated in this report is not the reproduction of any readymade report. We have not violated any of the provisions under the Copyright and Piracy / Cyber / IPR Act amended from time to time.

Date:

Place: Kolhapur



Roll No.	Unique ID	Student's Name	Signature
05	EN21122724	Samruddhi Sanjay Kale	
06	EN21191939	Shubha Muralidhar Desai	
07	EN21165406	Satej Santosh Kulkarni	
08	EN21148233	Akhilesh Shilbhuddha Damke	

KASABA BAWADA, KOLHAPUR

ACKNOWLEDGMENT

We extend our heartfelt gratitude to Prof. DR. A. K. Gupta Sir, Executive Director DYPCET, whose visionary leadership and unwavering support provided the foundation for this endeavor. We also express our sincere gratitude to Prof. DR. S. D. Chede Sir, Principal DYPCET, for their guidance and encouragement throughout this journey. Special thanks to DR. G. V. Patil Sir, Head of Department Data Science, for their invaluable insights and mentorship, shaping our understanding and approach within the field of data science.

We are immensely grateful to our project guide ---Guide Name--- whose expertise and dedication empowered us to navigate challenges and achieve milestones with confidence.

To our esteemed colleagues in the project group, your collaboration and dedication have been instrumental in realizing the goals of this project. Together, we have embraced innovation and teamwork, driving the project towards success. This project report stands as a testament to the collective efforts and support extended by each individual mentioned above. We acknowledge and appreciate your contributions, which have enriched our learning experience and propelled us towards excellence.

Date: (AN AUTONOMOUS INSTITUTE)

KASABA BAWADA, KOLHAPUR

Place: Kolhapur

INDEX

Sr. No.	Topic	Page Number
1.	Abstract	1
2.	Introduction	2
3.	Problem Statement & Objectives	3
4.	Requirement Specification	4
5.	System Architecture	5
6.	Modules	6
7.	Software & Hardware Requirements	8
8.	Result Analysis	9
9.	Conclusion & Future Scope	11
10.	References	12

(AN AUTONOMOUS INSTITUTE)

KASABA BAWADA, KOLHAPUR

Abstract

In today's dynamic business landscape, efficient financial management is paramount for organizational success. However, traditional invoice processing methods often pose significant challenges, including manual data entry errors, time-consuming workflows, and difficulties in adapting to diverse invoice formats. To address these issues, this project endeavors to develop a comprehensive software solution leveraging advanced image processing techniques.

At its core, the project aims to automate and streamline the extraction of essential information from invoices, such as dates, total amounts, and item quantities. Through the implementation of a sophisticated image processing algorithm tailored specifically for Tally invoices, the system seeks to achieve unparalleled accuracy in data extraction.

Furthermore, the project includes the integration of an Optical Character Recognition (OCR) system, enabling the conversion of image-based text into editable format. This capability not only facilitates further processing and analysis of invoice data but also enhances the system's flexibility and utility across a wide range of invoice layouts and fonts.

Central to the project's success is the development of a user-friendly interface for configuring and monitoring the invoice processing pipeline. By providing intuitive controls and real-time insights, this interface empowers users to customize and manage the system effortlessly, thereby maximizing efficiency and productivity.

Through meticulous research, experimentation, and collaboration, the project endeavors to deliver a robust, adaptable, and user-centric solution that revolutionizes invoice processing and financial management practices. By achieving these objectives, the project aims to empower chartered accountants and organizations alike, enabling them to navigate the complexities of modern finance with confidence and precision.

(AN AUTONOMOUS INSTITUTE)

KASABA BAWADA, KOLHAPUR

INTRODUCTION

In today's fast-paced business environment, characterized by rapid technological advancements and ever-evolving market dynamics, the efficient management of financial transactions stands as a critical pillar for the success and sustainability of enterprises. Among the myriad tools and platforms available to businesses, Tally emerges as a stalwart in the realm of accounting software, offering a comprehensive suite of features for streamlined invoicing, meticulous accounting, and insightful financial reporting. Yet, amidst its array of capabilities, a persistent challenge looms large: the reliance on manual data entry in the invoicing process.

Manual data entry, though a longstanding practice, presents a significant bottleneck in the seamless flow of financial operations within organizations. The arduous task of transcribing data from physical invoices into digital formats not only consumes valuable time and resources but also introduces the risk of human error, potentially compromising the accuracy and integrity of financial records. Recognizing this impediment as a prime opportunity for innovation and improvement, our project sets out to harness the transformative potential of image processing techniques to revolutionize Tally invoice processing.

At the heart of our endeavor lies the fusion of cutting-edge computer vision and machine learning technologies, aimed at automating character generation for Tally invoices. By meticulously analyzing and decoding invoice images through sophisticated algorithms and neural networks, our system endeavors to extract pertinent information seamlessly and accurately, transforming visual data into digital text format with unparalleled precision. This automated approach not only obviates the need for manual intervention but also augments the efficiency and reliability of the invoicing process, empowering businesses to navigate financial complexities with confidence and clarity.

In this comprehensive synopsis, we delineate the overarching objectives, methodological framework, and potential benefits of our project. Through a synthesis of innovation, expertise, and dedication, we aspire to deliver a transformative solution that not only streamlines invoicing processes but also catalyzes operational efficiency, reduces overhead costs, and fosters optimal resource allocation. By embracing the vanguard of technological advancement, we endeavor to empower businesses with the tools they need to thrive and flourish in an increasingly competitive landscape.

KASABA BAWADA, KOLHAPUR

PROBLEM STATEMENT

To Automate Invoice Text Extraction using OCR.

OBJECTIVES

- To develop image processing algorithm for accurate extraction of date, amount, and item quantity from invoices.
- To implement OCR system to convert scanned image text into editable format for further data processing.
- To explore techniques for handling variations in invoice layouts and fonts to enhance system reliability.
- To provide user-friendly interface for easy configuration and monitoring of invoice processing pipeline.
- To seamlessly integrate solution with Tally software to automate invoice generation and streamline accounting workflows.

D Y P A T I L
COLLEGE *Of*
ENGINEERING & TECHNOLOGY
(AN AUTONOMOUS INSTITUTE)

KASABA BAWADA, KOLHAPUR

Requirement Specification

The Invoice Data Extraction project aims to develop a robust and user-friendly application for automating the extraction of essential information from invoice images. The primary objective is to streamline the process of retrieving key data elements, including invoice total, tax amounts (e.g., SGST, CGST), invoice date, and party total, from scanned or photographed invoices. The application targets businesses and organizations that deal with a high volume of invoices, seeking to improve operational efficiency, reduce manual effort, and minimize errors associated with manual data entry.

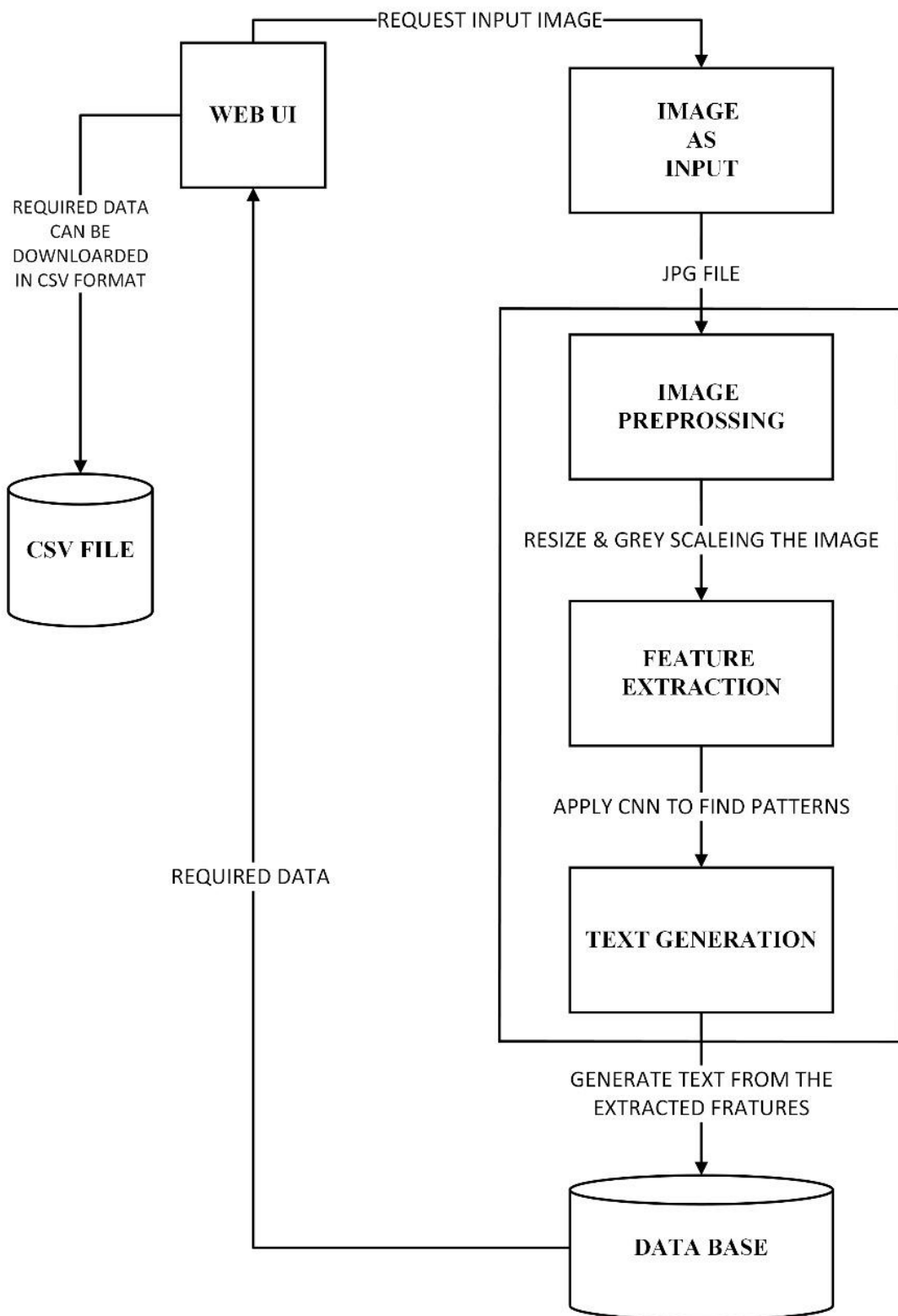
The system shall provide a simple and intuitive user interface for uploading invoice images in common formats such as JPEG, PNG, or PDF. Upon uploading an image, the application will leverage Optical Character Recognition (OCR) technology, implemented using the pytesseract library, to extract text content from the invoice. This extracted text will then undergo parsing and pattern matching using regular expressions to identify relevant information such as tax amounts, invoice total, date, and party total. The application shall handle various invoice formats and layouts, adapting its extraction algorithms dynamically to accommodate different templates and styles.

To ensure accuracy and reliability, the system will incorporate error handling mechanisms to address OCR inaccuracies, noisy images, and ambiguous text patterns. Additionally, the application will leverage the word2number library to convert textual representations of monetary values (e.g., "two hundred dollars") into numerical equivalents, facilitating arithmetic calculations and data analysis.

The requirement specification includes the following key features:

- 1. Image Upload:** Users can upload invoice images via the application's web-based interface.
- 2. OCR-based Text Extraction:** The system will utilize pytesseract for extracting text content from uploaded images.
- 3. Data Parsing and Extraction:** Extracted text will undergo parsing and pattern matching to identify relevant data elements such as tax amounts, invoice total, date, and party total.
- 4. Error Handling:** The application will incorporate error handling mechanisms to mitigate OCR inaccuracies and handle exceptions gracefully.
- 5. Dynamic Adaptation:** The system will dynamically adapt its extraction algorithms to accommodate different invoice formats and layouts.
- 6. Monetary Value Conversion:** Textual representations of monetary values will be converted into numerical equivalents using the word2number library.

System Architecture



Modules

Image Processing Module:

- **Functionality:** Handles image upload, format validation, and preprocessing tasks.
- **Components:** Utilizes Pillow (PIL) library for image manipulation.
- **Dependencies:** Requires PIL library for image processing.

Text Extraction Module:

- **Functionality:** Implements Optical Character Recognition (OCR) for text extraction.
- **Components:** Utilizes pytesseract library for OCR tasks.
- **Dependencies:** Requires pytesseract and Tesseract OCR engine.

Data Parsing Module:

- **Functionality:** Parses extracted text to identify relevant data elements.
- **Components:** Implements regex patterns for data extraction.
- **Dependencies:** Relies on Python's re module for pattern matching.

Error Handling Module:

- **Functionality:** Manages errors encountered during image processing and text extraction.
- **Components:** Implements robust error handling mechanisms.
- **Dependencies:** Utilizes Python's exception handling and logging.

Data Conversion Module:

- **Functionality:** Converts textual monetary values to numerical equivalents.
- **Components:** Utilizes word2number library for text-to-number conversion.
- **Dependencies:** Requires word2number library.

User Interface Module:

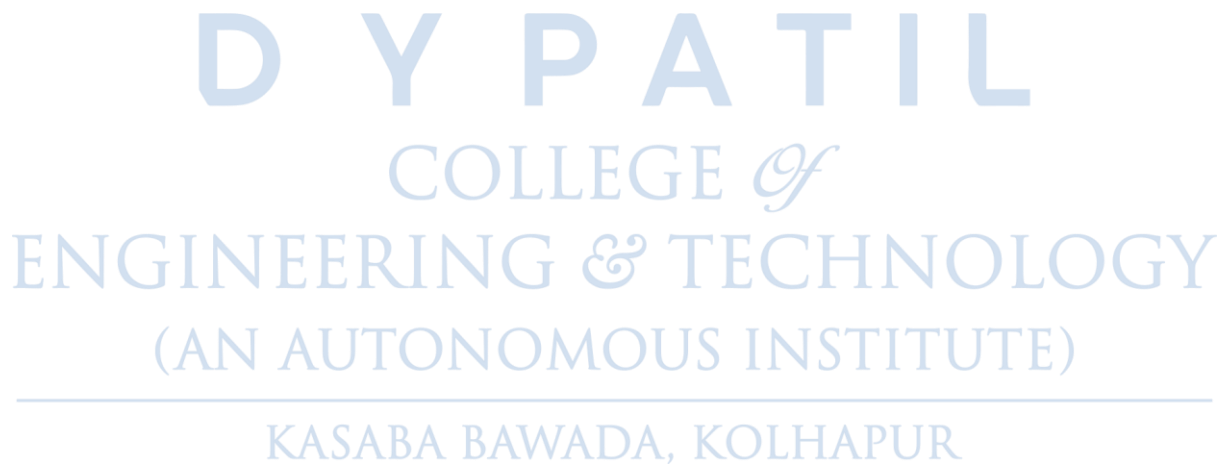
- **Functionality:** Provides user-friendly interface for image upload and data visualization.
- **Components:** Implements web-based UI using Streamlit library.
- **Dependencies:** Requires Streamlit library for building interactive interfaces.

Reporting and Export Module:

- **Functionality:** Supports generating reports and exporting data in CSV or Excel format.
- **Components:** Implements functionality for data export.
- **Dependencies:** Relies on pandas library for data manipulation.

Integration and Testing Module:

- **Functionality:** Integrates modules, conducts unit tests, and performs system testing.
- **Components:** Implements integration tests and system testing procedures.
- **Dependencies:** Utilizes testing frameworks such as pytest and Selenium.



Software & Hardware Requirements

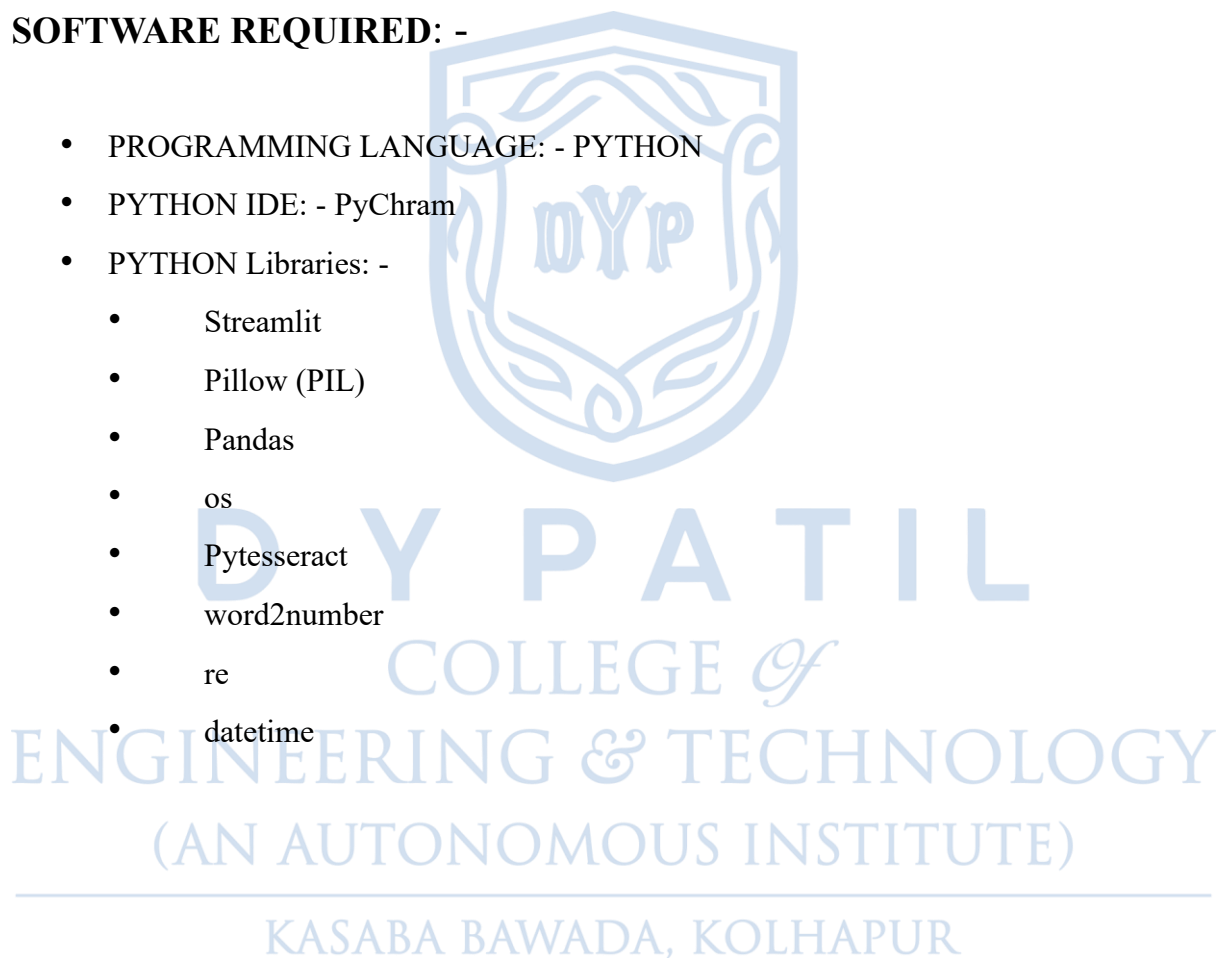
HARDWARE REQUIRED: -

- Processor: Intel Pentium or higher
- RAM: 4 GB or more
- Storage: 256 GB or more
- Internet connection: Broadband connection with good speed
- GPU: Not mandatory but can be helpful for faster training of machine learning models.

SOFTWARE REQUIRED: -

- PROGRAMMING LANGUAGE: - PYTHON
- PYTHON IDE: - PyCharm
- PYTHON Libraries: -

- Streamlit
- Pillow (PIL)
- Pandas
- os
- pytesseract
- word2number
- re
- datetime




Result Analysis

Deploy

Invoice Data Extraction

Upload Single Invoice Image



Drag and drop file here

Limit 200MB per file • JPG, PNG, JPEG

Browse files

Upload Folder Containing Invoice Images

Process

[illegible]

Invoice Data Extraction

Upload Single Invoice Image



Drag and drop file here
Limit 200MB per file • JPG, PNG, JPEG

Browse files

Upload Folder Containing Invoice Images

/Users/samruddhikale/Desktop/New folder/d

Process

	Invoice Name	Date	SGST Payable	CGST Payable	Total
0	ELL-010	10-05-2023	81.00	81.00	509
1	ACL-050	26-04-2024	78.75	78.75	2,563
2	HWL-004	22-01-2024	54.00	54.00	1,908
3	HWL-002	78-41-2023	72.00	72.00	1,372
4	HWL-001	25-03-2024	78.75	78.75	2,563



	Invoice Name	Date	SGST Payable	CGST Payable	Total
0	ELL-010	10-05-2023	81.00	81.00	509
1	ACL-050	26-04-2024	78.75	78.75	2,563
2	HWL-004	22-01-2024	54.00	54.00	1,908
3	HWL-002	78-41-2023	72.00	72.00	1,372
4	HWL-001	25-03-2024	78.75	78.75	2,563

Sales Line Chart



Conclusion

In conclusion, the development of a software solution utilizing image processing techniques to automate invoice data extraction presents a significant advancement in the realm of financial management. By successfully achieving the outlined objectives, our project aims to revolutionize the way businesses handle invoice processing.

Through the implementation of sophisticated algorithms and machine learning models, we aspire to deliver a robust system capable of accurately extracting essential information such as dates, amounts, and item quantities from invoice images. The integration of character recognition (OCR) technology further enhances the system's capabilities by converting textual content into editable digital format, thereby facilitating seamless data processing.

Moreover, our project acknowledges the diverse landscape of invoice layouts and fonts, and endeavors to address this challenge by exploring techniques for adaptability and reliability. By providing a user-friendly interface for configuration and monitoring, we ensure that businesses can easily customize and manage the invoice processing pipeline according to their specific requirements.

Future Scope

While our project endeavours to fulfil current needs in invoice automation, there exist several avenues for future enhancement and expansion. Some potential areas of focus include:

- **Enhanced Accuracy and Robustness:** Continual refinement of the image processing algorithms and OCR system to improve accuracy, particularly in handling complex invoice formats and low-quality images.
- **Integration with Cloud Services:** Exploration of cloud-based solutions for scalability and accessibility, enabling seamless integration with Tally software and other accounting platforms.
- **Invoice Verification and Validation:** Extension of the system capabilities to include verification and validation of extracted data against predefined rules or databases, ensuring accuracy and compliance.
- **Natural Language Processing (NLP) Integration:** Incorporation of NLP techniques to extract additional context from invoice text, such as vendor information, invoice purpose, and payment terms.
- **Mobile Application Development:** Development of a mobile application for capturing and processing invoice images directly from mobile devices, offering convenience and flexibility for users on the go.
- **Advanced Reporting and Analytics:** Implementation of advanced reporting and analytics features to derive actionable insights from invoice data, facilitating strategic decision-making and financial planning.

References

- BODAPATI SOHAN CHIDVILAS and VENKATA NAGA SAI RAKESH KAMISETTY, “DIGITIZATION OF DATA FROM INVOICE USING OCR “
- T.M.Rath and R. Manmatha, “Word spotting for historical documents”, International Journal on Document Analysis and Recognition (IJDAR), Vol.9, No 2 – 4, pp. 139 – 152 , 2006.
- V. Lavrenko, T. M. Rath, R. Manmatha: “Holistic Word Recognition for Handwritten Historical Documents”, Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04),pp 278-287, 2004.
- T. Adamek, N. E. O'Connor, A. F. Smeaton, “Word Matching Using Single-Closed Contours for Indexing Handwritten Historical Documents”, International Journal on Document Analysis and Recognition (IJDAR), special Issue on Analysis of Historical Documents, 2006.
- T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis and S. J. Perantonis, “ Keyword - Guided Word Spotting in Historical Printed Documents Using Synthetic Data and User feedback ”, International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents, Vol. 9, No. 2-4, pp. 167-177, 2007.
- V.G.Gezerlis and S.Theodoridis, “Optical Character Recognition for the Orthodox Hellenic Byzantine music notation”, Pattern Recognition, Vol.35, pp. 895 – 914, 2002.
- L. Laskov, “Classification and Recognition of Neume Note Notation in Historical Documents”, International Conference of Computer Systems and Technologies (CompSysTech), 2006.
- K. Ntzios, B. Gatos, I. Pratikakis, T. Konidaris and S.J. Perantonis, “An Old Greek Handwritten OCR System based on an Efficient Segmentation-free Approach”, International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents, Vol. 9, No. 2-4, pp. 179-192, 2007