SMJE 4263 COMPUTER INTEGRATED MANUFACTURING

EXTRACTING INFORMATION FROM INVOICES AND RECEIPT USING PYTHON

Name: Voon Kim Vui

Matrix: A19MJ0230

Lecturer: Prof. Madya.Ir. Dr. Zool Hilmi Bin Ismail

Abstract

This report presents a solution for extracting information from receipts and invoices using Optical Character Recognition (OCR) with Tesseract. The code utilizes regular expressions to extract relevant data such as reference numbers, dates, and amounts from the OCR text. The extracted information is organized into separate data frames for invoices and receipts, which are then combined into a single data frame. The report provides an overview of the code implementation, discusses the problem statement, outlines the objectives, describes the methodology, presents the results, and concludes with the significance of the solution.

Introduction

The management and organization of financial documents, such as receipts and invoices, can be a time-consuming and error-prone task. Automating the extraction of key information from these documents can greatly enhance efficiency and accuracy. This report introduces a solution that employs OCR techniques to extract relevant data from receipts and invoices. The extracted information can then be used for various purposes, such as financial analysis, record keeping, and auditing.

Problem Statement

The manual extraction of information from receipts and invoices is a laborious and error-prone process. It requires individuals to manually read and transcribe data, which can result in mistakes and inconsistencies. Moreover, the sheer volume of documents makes manual processing inefficient and time-consuming. The problem statement revolves around automating this process by developing a system that can accurately extract key information from receipts and invoices.

Objectives

The main objectives of the solution are as follows:

1. Develop a system to automatically extract information from receipts and invoices.
2. Extract key data points such as reference numbers, dates, and amounts.
3. Handle variations in document formats and layouts.
4. Organize the extracted information into separate data frames for invoices and receipts.

Provide a consolidated view of the extracted data in a single data frame.

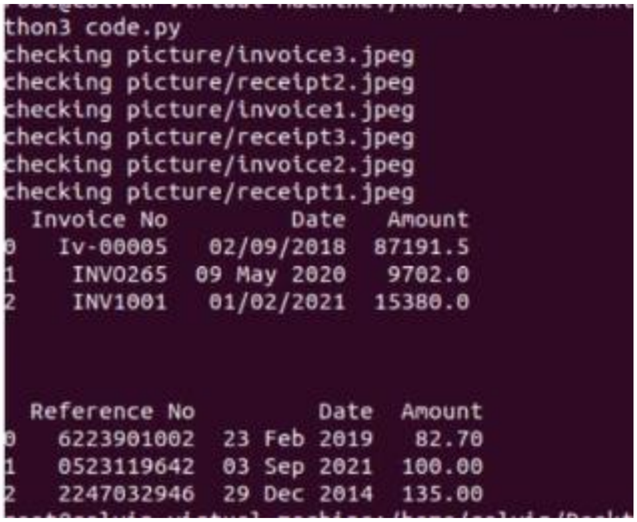
Methodology

The solution utilizes the following methodology:

The code employs the Tesseract OCR engine to perform optical character recognition on receipt and invoice images. Regular expressions are used to extract relevant information from the OCR text, such as reference numbers, dates, and amounts. Two separate functions, extract_information_from_receipt and extract_information_from_invoice, are implemented to handle receipts and invoices, respectively. The extracted information is stored in separate lists for invoices and receipts. The lists are then used to create separate data frames, df1 and df2, for invoices and receipts, respectively. Finally, the data frames are concatenated along the columns axis to create a single data frame, combined, containing all the extracted information.

Results

The code successfully extracts information from receipts and invoices using OCR with Tesseract. The extracted data is organized into separate data frames for invoices and receipts. The invoice data frame, df1, contains columns for invoice number, date, and amount. The receipt data frame, df2, includes columns for reference number, date, and amount. These data frames provide a structured representation of the extracted information, making it easier to analyze and utilize for further processing.



```
thon3 code.py
checking picture/invoice3.jpeg
checking picture/receipt2.jpeg
checking picture/invoice1.jpeg
checking picture/receipt3.jpeg
checking picture/invoice2.jpeg
checking picture/receipt1.jpeg
  Invoice No        Date   Amount
0   Iv-00005   02/09/2018  87191.5
1    INVO265  09 May 2020   9702.0
2    INV1001   01/02/2021  15380.0



  Reference No          Date  Amount
0   6223901002  23 Feb 2019   82.70
1   0523119642  03 Sep 2021  100.00
2   2247032946  29 Dec 2014  135.00
```

Conclusion

The developed solution provides an efficient and accurate method for extracting information from receipts and invoices. By leveraging OCR techniques and regular expressions, the code can effectively locate and extract key data points from various document formats and layouts. The separate data frames for invoices and receipts allow for easy access and analysis of the extracted information. This solution significantly reduces the manual effort involved in processing financial documents, leading to improved efficiency and accuracy in financial management tasks.