Technical Document

Text extraction from store bills

1. Introduction

Background

Extraction of text information from store bills is a crucial task for automating data entry and analysis. This document provides a comprehensive guide on the technical aspects of text extraction from store bills.

Purpose

The purpose of this document is to outline the methods, tools, and technologies involved in extracting text from store bills efficiently.

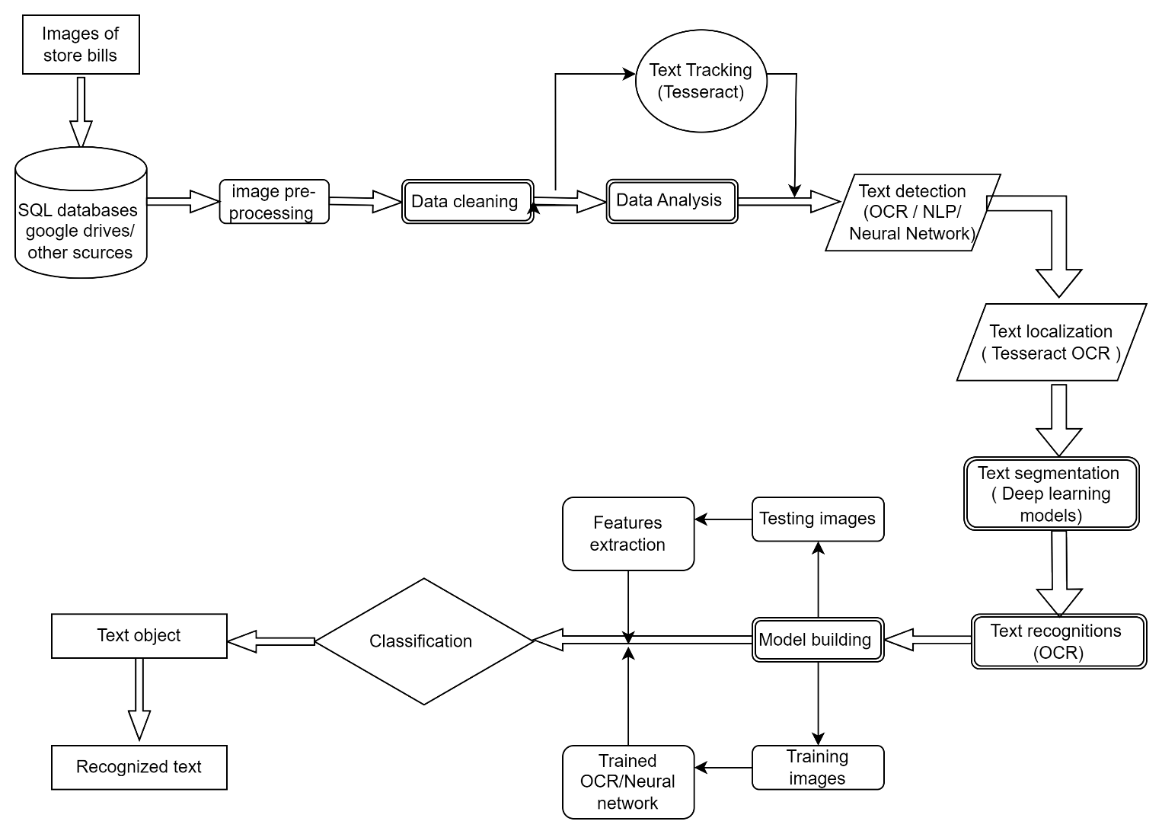
2. System Overview

Architecture

Provide an architectural overview of the system, highlighting the key components and their interactions.

Workflow

Detail the workflow of the text extraction process, from capturing the store bill image to obtaining structured text data.



3. Text Extraction Techniques

Optical Character Recognition (OCR)

Explain the use of OCR technologies for recognizing text within images. Detail the integration and configuration of Tesseract OCR for accurate text extraction.

* Tesseract OCR
* Keras OCR
* Wand OCR
* Paddle OCR
* Iron OCR

Regular Expressions

Demonstrate the use of regular expressions for pattern matching to extract specific information like product name, bill number etc.

Technology Stack

Programming Language

Python programming language used for development and MYSQL have used to database.

Libraries and Frameworks

List and describe the libraries and frameworks utilized, including Wand OCR and streamlit libraries are mainly used to deploy.

Data Storage and Management

Database

I have used MYSQL database to store and manage data as a structured data frame.

Data Cleaning

For data cleaning I have used opencv or cv2 libraries to convert the image into grayscale, resize, white balance, black balance to process and extract strings or words in the image

Deployment

Cloud Deployment

Deploying a text extraction system for store bills with OCR using Streamlit involves leveraging its simplicity and flexibility to create a web-based interface for users to interact with the application.

On-Premises Deployment

On-Premises Deployment for text extraction from store bills and OCR involves installing and running the text extraction system within the organization's own infrastructure, typically within its data centers or dedicated servers. This deployment model offers more control over the hardware, software, and security configurations, making it suitable for organizations with specific compliance requirements, data privacy concerns, or those that prefer to keep sensitive data within their premises.