QBurst

Data Extraction for

Faster Document Processing

Overview

A leading medical diagnostics company sought to enhance its data processing capabilities by automating information extraction from lab reports. The objective was to streamline the extraction of personal details, test information, and handwritten data present in various formats within these reports. We implemented an AI-based data extraction solution that leverages optical character recognition (OCR), computer vision, and Natural Language Processing (NLP) to identify and extract relevant data points from different types of documents. The extracted data can be downloaded or pushed to third-party applications, thus speeding up the processing of those documents.

Client Profile

A premier medical diagnostics company.

Business Challenges

The client faced challenges with inefficient manual data entry from diverse lab reports, resulting in inaccuracies and resource-intensive processes. Handling variability in report formats, accurately extracting handwritten data, and customizing solutions across domains were major pain points.

- Extract personal details (name, registration identifier, gender, age) from PDF lab reports
- Identify specific test information mentioned in the reports
- Implement image processing and OCR techniques for accurate data extraction
- Adapt to different report formats and support scalability

- Customize domain-specific layers to accommodate various healthcare domains and report types
- Recognize and extract handwritten data, even if it overlaps with printed text

QBurst Solution

The solution leveraged image processing and optical character recognition (OCR) to efficiently extract data from scanned PDF lab reports, overcoming limitations encountered by traditional parsing mechanisms. Its scalable architecture provided the flexibility to fine-tune operations, ensuring compatibility with diverse report formats. Integration of a pluggable domain-specific layer, enabled seamless customization across different domains, enabling the identification and extraction of relevant information based on specific use cases, thereby enhancing precision and relevance.

The solution employed image-based models to accurately extract handwritten data. Through sophisticated algorithms, it discerned handwritten text, even in varied orientations, overlapping with printed details, or existing within diverse document layouts. The solution features logic mechanisms to reject less accurate outputs, thereby significantly improving the precision and reliability of data extracted from handwritten sections.

The comprehensive approach not only streamlined the extraction process from scanned documents but also provided a robust framework for accommodating various forms, domains, and data types. Its adaptability and accuracy enhancements served as crucial components in automating data entry, reducing errors, and ultimately empowering the client with more efficient and accurate data analysis capabilities.

Project Highlights

- **Multi-source data extraction:** Capable of extracting personal details, test information, and handwritten data from scanned PDF lab reports.
- **Robust image processing:** Utilizes image-based models to adjust for skewed images and minor orientation issues, ensuring accurate data extraction.
- **Domain-specific customization:** Offers a pluggable domain-specific layer allowing customization for different domains, enhancing accuracy and relevancy.
- **Handwritten text recognition:** Pecialized algorithms to extract handwritten data from scanned documents, handling varying angles, formats, and overlaps with printed text.
- **Predefined data coordinates support:** Ability to define and extract data from specific areas within documents, ensuring higher accuracy in extracting targeted information.
- **Adaptability and customization:** Customizable to handle multiple forms and domains, ensuring flexibility in data extraction processes.
- **Error rejection logic:** Incorporates logic to reject inaccurate or less reliable outputs, improving overall data accuracy and integrity.

Technologies Used

- Open AI
- OCR
- Python
- Java
- Computer vision
- Google Cloud Platform

Benefits

- Automated data extraction reduced the manual effort required for analyzing lab reports, improving overall operational efficiency.
- Implementation of domain-specific logic and rejection of inaccurate outputs increased the reliability of extracted data, reducing errors.
- By automating repetitive data entry tasks, the company saved significant time and resources, leading to cost savings.
- The customizable nature of the solution allowed it to adapt to various forms and domains within the healthcare industry, enhancing its versatility.