IDA Extraction

Intelligente Datenextraktion mit Few-Shot Learning

IDA Extraction ist ein Feature zur intelligenten Datenextraktion aus Dokumenten mit **wenigen notwendigen Lernschritten** (Few-Shot Learning). Dank **fortschrittlicher Machine-Learning-Fähigkeiten** wird die Einrichtung und Wartung von Workflows im Vergleich zu regelbasierten oder manuellen Ansätzen drastisch beschleunigt. In Kombination mit der **außergewöhnlichen OCR und ICR** von PLANET AI sinkt der Bedarf an manuellen Korrekturen erheblich, was zu einer verbesserten durchgehenden Datenverarbeitung (Straight Through Processing) führt.

HAUPTMERKMALE

Verfeinerte zonale Datenextraktion

IDA Extraction verwendet eine fortschrittliche Methode zur Extraktion von Schlüssel-Wert-Paaren (Key-Value Pair Extraction), die es Benutzer:innen ermöglicht, die Datenfelder, die sie erfassen möchten, einfach zu spezifizieren. Diese Felder sind nicht auf Text beschränkt, sondern können auch Barcodes, Kontrollkästchen und numerische Werte umfassen. IDA liefert Positionsinformationen, die sich für nachfolgende Aufgaben, wie z. B. Validierung, als nützlich erweisen.

Lernen mit wenigen Trainingsdaten

IDA Extraction verfügt über fortschrittliches Few-Shot-Learning, das sowohl visuelle als auch textuelle Merkmale berücksichtigt. Indem man dem System nur einige wenige Trainingsdokumente vorlegt und die gewünschten Datenfelder angibt, kann es diese aus Seiten extrahieren, die es beim Training nicht gesehen hat.

SYSTEMVORAUSSETZUNGEN

Für 64-Bit-Systeme

Linux: Ubuntu 18.04 - 23.10, Debian 11, CentOS 8, Red Hat 8.x, LEAP 15.x, SLES 15 SP 4-5
Windows: 10, 11
Windows Server: 2012, 2012 R2, 2016, 2019, 2022
Docker

Mind. **9 GB Festplattenspeicher**
+ ca. 0,9 GB für volle Funktionalität
+ bis zu 0,5 GB zum Speichern von Protokolldaten

Dies verkürzt nicht nur die Zeit bis zur Wertschöpfung, sondern minimiert auch den Aufwand für die Anpassung an wechselnde Dokumentenlayouts.

Das volle Potenzial unübertroffener OCR-Qualität

IDA Extraction basiert auf **IDA Recognition**, einer optischen (OCR) und intelligenten (ICR) Zeichenerkennungs-Engine, die selbst in den schwierigsten Szenarien hervorragende Ergebnisse liefert. IDA Recognition erfasst Maschinen- und Handschrift, Kontrollkästchen, Tabellen und historische Schriften, selbst bei schlechter Scanqualität mit gedrehtem oder schiefem Druck. Qualitativ hochwertige Eingabedaten sind für Aufgaben der Datenextraktion entscheidend, da sie sich auf die Qualität der Extraktionsausgabe auswirken.

Vielseitiges JSON-Ausgabeformat

Die extrahierten Datenfelder warden in einem JSON-Format ausgegeben, das einen einfachen Zugriff für nachfolgende Aufgaben in der Weiterverarbeitung ermöglicht.

No-Code-Training

Der IDA Extraction Assistant (ExA) ermöglicht es Benutzer:innen ohne technisches Fachwissen, Datenextraktionsmodelle über eine browserbasierte grafische Oberfläche zu erstellen, zu trainieren und anzupassen.

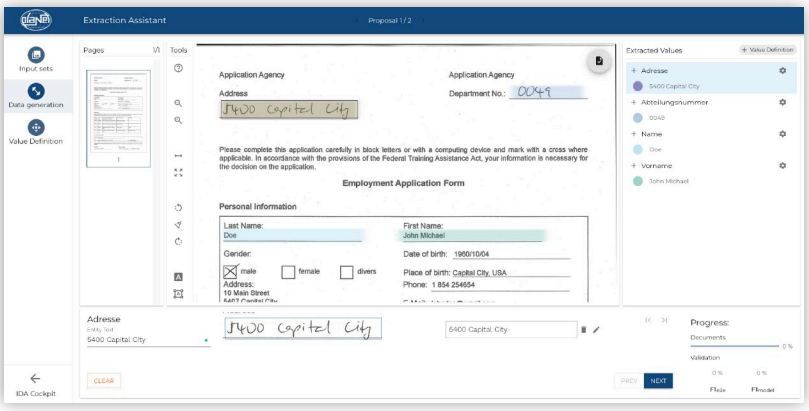
Einfache Bereitstellung und Integration

IDA wird entweder vor Ort (on-premises) oder in der Cloud als Java-Anwendung oder als Containerisierung mit Docker bereitgestellt. Die gRPC-API ermöglicht eine nahtlose und schnelle Integration.

MODELLTRAINING

IDA bietet den **Extraction Assistant (ExA)**, eine grafische Schnittstelle, die es Benutzer:innen ermöglicht, Modelle zu trainieren, ohne dass sie Programmierkenntnisse benötigen oder komplexe Datensätze vorbereiten müssen. Derzeit funktioniert ExA am besten bei **strukturierten und halbstrukturierten Dokumenten** wie Formularen oder Rechnungen. Auf der Grundlage der von IDA Classification durchgeführten Kategorisierung können die Dokumente an verschiedene Extraktionsmodelle weitergeleitet werden.

Als allgemeine Richtlinie wird ein **Minimum von fünf Dokumenten pro Dokumentenklasse** empfohlen. Es ist jedoch wichtig zu beachten, dass eine größere Anzahl von Trainingsdokumenten in der Regel zu einem besseren Modell führt.



ExA-Oberfläche zur Kennzeichnung von Datenfeldern

Benutzer:innen können eine unbegrenzte Anzahl von zu extrahierenden Datenfeldern definieren.

Um den manuellen Beschriftungsprozess zu beschleunigen, führt der Assistent ein **automatisches Vortraining** durch, um wiederkehrende Ankerpunkte innerhalb der Dokumente zu identifizieren. Als Ergebnis gruppiert ExA die Trainingsdokumente zur Verbesserung der Effizienz. Nach der Beschriftung eines dieser Beispieldokumente müssen Benutzer:innen nur noch die vorgeschlagenen extrahierten Felder validieren oder korrigieren. Für ein automatisches Pre-Training sind mindestens **drei Dokumente mit dem gleichen Layout** erforderlich. AcroForm-Felder werden priorisiert und automatisch zur Verwendung als Datenfeld vorgeschlagen.

Weitere Informationen sind in der Software-Dokumentation zu finden.