IDA Extraction

Intelligente Datenextraktion zur Reduzierung manueller Arbeit

IDA Extraction automatisiert die Datenerfassung aus strukturierten und unstrukturierten Dokumenten ohne komplexe Regelwerke. Ob Formulare, Verträge oder Freitextdokumente: Die Kombination aus Few-Shot Learning und LLM-Technologie extrahiert präzise die Daten, die Sie benötigen. Das Ergebnis: Drastisch reduzierter manueller Aufwand bei Validierung und Workflow-Einrichtung, signifikant höhere Straight-Through-Processing-Raten und qualitativ hochwertige Daten als Fundament für nachfolgende Prozesse.

ZWEI EXTRAKTIONSOPTIONEN

IDA Extraction Assistant (ExA)

- Für strukturierte und semistrukturierte Dokumente
- Key-Value-Pair-Extraktion

IDA LLM Entity Extraction

- Für unstrukturierte Dokumente ohne festes Layout
- Zero-Shot Named Entity Recognition

VORTEILE

Breites Anwendungsspektrum

IDA Extraction nutzt Technologien aus **Machine Learning** und **Natural Language Processing (NLP)**, wie z. B. intelligente zonale Extraktion, große Sprachmodelle (LLM) und LayoutLM. Die Anwendungen reichen von der Verarbeitung strukturierter Formulare bis hin zur Analyse von Dokumenten mit unstrukturiertem Text.

Das volle Potenzial unübertroffener OCR-Qualität

IDA Extraction basiert auf **IDA Recognition**, der OCR-Engine für hervorragende Ergebnisse in den schwierigsten Szenarien. Selbst bei verzerrten Scans, schlechter Bildqualität und schwer lesbarer Handschrift liefert IDA Recognition die hochwertige Textbasis, die für verlässliche Datenextraktion entscheidend ist. Warum das wichtig ist: Die Extraktionsqualität steht und fällt mit der Qualität der Eingabedaten.

IDA Recognition erfasst Maschinenschrift, Handschrift, Kontrollkästchen, Tabellen und historische Schriften als perfekte Grundlage für nachfolgende Extraktionsprozesse.

Vielseitige Ausgabeformate

Die extrahierten Daten werden in einem **JSON**-Format ausgegeben, das einen einfachen Zugriff für nachfolgende Aufgaben in der Weiterverarbeitung ermöglicht. Zusätzlich können die Daten in dem resultierenden **PDF** an ihrem ursprünglichen Ort markiert werden.

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 (alternativ: REST-API) ermöglicht eine schnelle Integration.

IDA EXTRACTION ASSISTANT

Lernen mit wenigen Trainingsdaten

Der IDA Extraction Assistant (ExA) nutzt fortschrittliches Few-Shot-Learning, das visuelle und textuelle Merkmale analysiert Mit wenigen Beispieldokumenten ist das System einsatzbereit und extrahiert Daten aus Dokumenten, die es beim Training nicht gesehen hat. Dies verkürzt nicht nur die Zeit bis zur Wertschöpfung, sondern minimiert auch den Aufwand für die Anpassung an wechselnde Layouts.

No-Code-Training

Benutzer:innen ohne Programmierkenntnisse können eigenständig Extraktionsmodelle erstellen, trainieren und anpassen. Die browserbasierte grafische Oberfläche macht Machine Learning einfach zugänglich.

VORAUSSETZUNGEN

Der **IDA Server** wird für die Verarbeitung von Inputdokumenten benötigt und bietet zudem eine Browser-Schnittstelle.

Für 64-Bit-Systeme
Linux: Ubuntu 18.04-25.10, Debian 11-13, CentOS 8-10, Red Hat 8.x-10.x, LEAP 15.4-15.6, 16.0; SLES 15 SP 4-7
Windows: 10, 11
Windows Server: 2016, 2019, 2022, 2025
Docker

Mind. **12 GB Festplattenspeicher**
Mind. **16 GB Arbeitsspeicher**

Zonale Datenextraktion

ExA extrahiert Schlüssel-Wert-Paaren (Key-Value Pair Extraction) aus strukturierten Dokumenten. Spezifizieren Sie einfach die Datenfelder, die Sie erfassen möchten – nicht nur Text, sondern auch Kontrollkästchen und numerische Werte. IDA liefert Positionsinformationen mit, die für nachfolgende Aufgaben wie Validierung wichtig sein können.

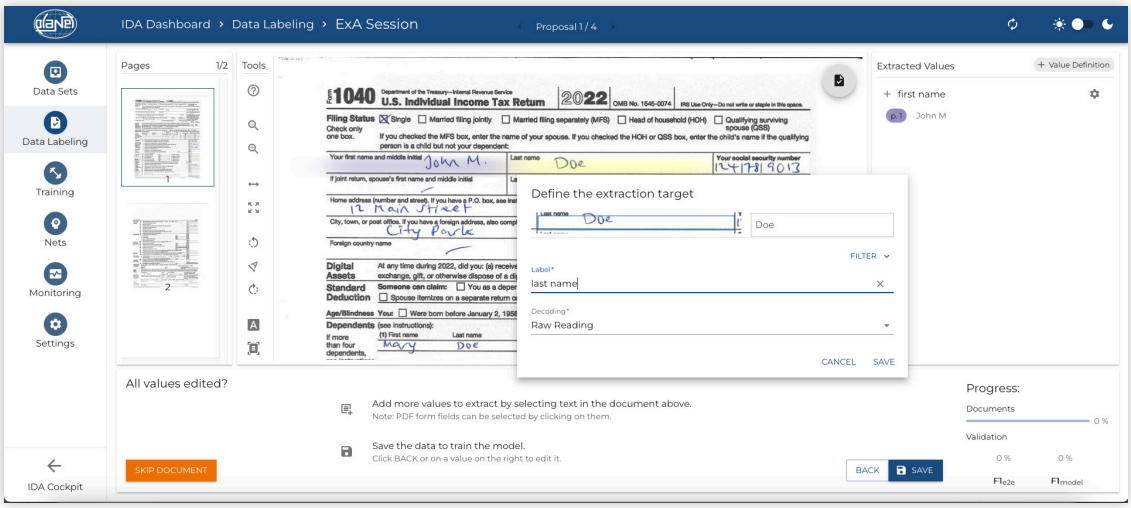
Modelltraining

Die grafische Oberfläche von ExA eliminiert die Notwendigkeit komplexer Datenaufbereitung. Trainieren Sie Modelle direkt im Browser – ohne technisches Setup. Derzeit funktioniert ExA am besten bei strukturierten und semistrukturierten Dokumenten wie z. B. Formularen.

Ein **automatisches Vortraining** beschleunigt den Prozess: ExA erkennt automatisch wiederkehrende Ankerpunkte in den Dokumenten und gruppiert ähnliche Layouts. So beschriften Sie nur ein Musterdokument pro Layout-Gruppe statt jedes einzelne Dokument.

So funktioniert's:

- Mindestens 3 Dokumente mit identischem Layout bilden ein Muster.
- ExA schlägt automatisch Extraktionsfelder vor.
- Sie validieren oder korrigieren die Vorschläge.
- AcroForm-Felder werden priorisiert und automatisch erkannt.



ExA-Oberfläche zur Kennzeichnung von Datenfeldern

Trainingsempfehlungen:

- Minimum: 5 Dokumente pro Dokumentenklasse
- Optimal: Je mehr Trainingsdokumente, desto besser das Modell
- Bei strukturierten Formularen: Wenige Beispiele genügen

Benutzer:innen können optional das **Layout Language Model (LayoutLM)** verwenden, um ihre Extraktionsergebnisse zu verbessern. Dieses Modell priorisiert visuelle Hinweise und kontextuelles Verständnis und ist damit besonders effektiv bei der Verarbeitung semistrukturierter Dokumente. Basierend auf der von IDA Classification durchgeführten Sortierung können die Dokumente an verschiedene Extraktionsmodelle weitergeleitet werden.

Seit IDA 5.3 dient ExA auch zur **Validierung von LLM Entity Extraction**. Erstellen Sie Ground-Truth-Sets und testen Sie verschiedene Prompt-Varianten objektiv für verlässliche LLM-Extraktion in Produktionsumgebungen.

IDA LLM ENTITY EXTRACTION

Keyword-basierte Abfragen

Automatisieren Sie die Datenextraktion aus unstrukturierten Dokumenten durch einfache Abfragen und ohne vorheriges Training (Zero-Shot Named Entity Recognition). Formulieren Sie, was Sie suchen: Ein Label kombiniert mit einem Schlüsselwort, einer Keyword-Liste oder einer kurzen Beschreibung.

Verifizierung der extrahierten Daten

KI-Halluzinationen sind ein bekanntes Problem bei LLM-basierten Systemen. IDA verhindert sie durch einen zusätzlichen Verifizierungsschritt: Die vom LLM extrahierten Daten werden mit der OCR-Transkription von IDA Recognition abgeglichen.

Prompt-Validierung

LLM-basierte Extraktion ist oft mit Unsicherheiten verbunden: Unvorhersehbare Modellleistung und fehlende Vergleichbarkeit zwischen Prompts machten manuelles Screening notwendig. Mit IDA können Sie LLM-Prompts gegen definierte Testdaten validieren und so verschiedene Prompt-Varianten vor dem Produktivbetrieb testen.

Rechnungsmodul[BETA]

IDA Extraction enthält ein vorkonfiguriertes Modul für die Rechnungsverarbeitung, das auf LLM Entity Extraction basiert. Es bietet voreingestellte Entitäten und ist damit sofort einsatzbereit.

VORAUSSETZUNGEN

Um große Sprachmodelle (LLMs) **on-premises** nutzen zu können, ist ein zusätzlicher **LLM-Server** erforderlich.

Für 64-Bit-Systeme

- **Docker (Ubuntu-basiert)**
- Mind. **40 GPU-Speicher** (aufteilbar auf mehrere GPUs)
- Mind. **6,5 GB Festplattenspeicher** + mind. 20 GB für LLM
- Mind. **64 GB Arbeitsspeicher (RAM)**

Wichtig: Festplatten- und Arbeitsspeicher hängen stark von den gewählten Modellen ab. Reiner CPU-Modus ist nicht möglich.

Der LLM Server kann zudem **OpenAI-Modelle** ansprechen und fungiert dann als Relais. Dadurch sind keine GPUs notwendig.

No-Code-Lösung

Benutzer:innen ohne Programmierkenntnisse können eigenständig Prompts erstellen. Die Anpassung des Moduls ist über das Browser-Interface einfach möglich.

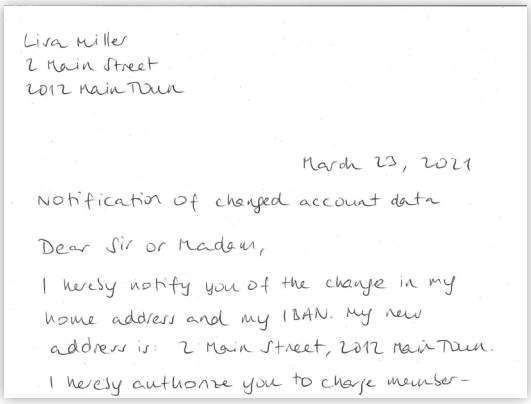
So funktioniert‘s

IDA LLM Entity Extraction eignet sich **am besten für die Verarbeitung unstrukturierter Dokumente**, die kein festes Layout oder Datenpunkte haben, wie z.B. Verträge oder Anschreiben. Basierend auf der von IDA Classification durchgeführten Kategorisierung können die Dokumente an verschiedene Extraktionsmodelle weitergeleitet werden.

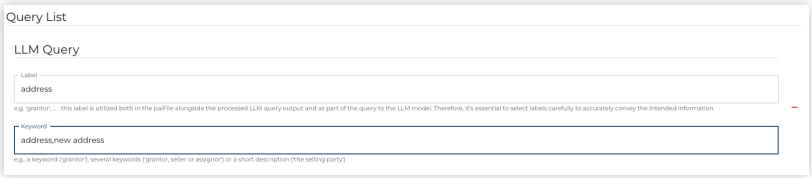
Eine **LLM-Anfrage (Query)** besteht aus:

- Label: Name des zu extrahierenden Datenfelds
- Definition: Schlüsselwort, Keyword-Liste oder Kurzbeschreibung
- Verifizierung: Abgleich mit Entity Finder und OCR-Transkription

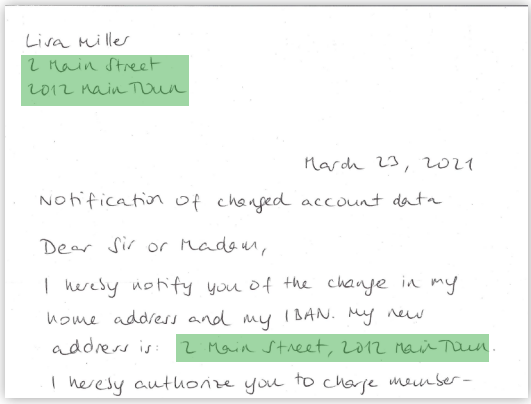
Inputdokument



+



Outputdokument



+

Erleben Sie IDA Extraction selbst und kontaktieren Sie uns für eine individuelle Demo.

Jetzt Demo buchen