IDA Extraction

Intelligente Datenextraktion zur Reduzierung manueller Arbeit

IDA Extraction bietet **modernste Datenextraktion** sowohl für strukturierte als auch für unstrukturierte Dokumente. Die Nutzung von Technologien wie **Few-Shot Learning** und **großen Sprachmodellen (LLM)** ermöglicht die präzise Erfassung von Daten in Dokumenten. Dies reduziert den Bedarf an manueller Arbeit erheblich, sowohl bei der Validierung von Ergebnissen als auch bei der Einrichtung von regelbasierten Workflows, was zu einer **verbesserten durchgehenden Datenverarbeitung** (Straight Through Processing) führt.

PRODUKTKONFIGURATIONEN

IDA Extraction Assistant (ExA)

- Am besten geeignet für strukturierte und semistrukturierte Dokumente
- z. B. für Formularverarbeitung

IDA LLM Entity Extraction

- Am besten geeignet für unstrukturierte Dokumente
- z. B. für Volltext-Dokumentenindexierung

HAUPTMERKMALE

Breites Anwendungsspektrum

IDA Extraction nutzt eine Reihe von Technologien aus dem Bereich des **Machine Learning** und des **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, 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.

Hochwertige Eingabedaten sind für Aufgaben der Datenextraktion entscheidend, da sie sich direkt auf die Qualität der Extraktionsergebnisse auswirken.

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

No-Code-Training

ExA ermöglicht es Benutzer:innen **ohne technisches Fachwissen**, Datenextraktionsmodelle über eine browserbasierte grafische Oberfläche zu erstellen, zu trainieren und anzupassen.

SYSTEMVORAUSSETZUNGEN

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, 12; CentOS 8, Red Hat 8.x, 9; LEAP 15.x, SLES 15 SP 4-6
Windows: 10, 11
Windows Server: 2016, 2019, 2022
Docker

Mind. **12 GB Festplattenspeicher**

Mind. **16 GB Arbeitsspeicher**

Zonale Datenextraktion

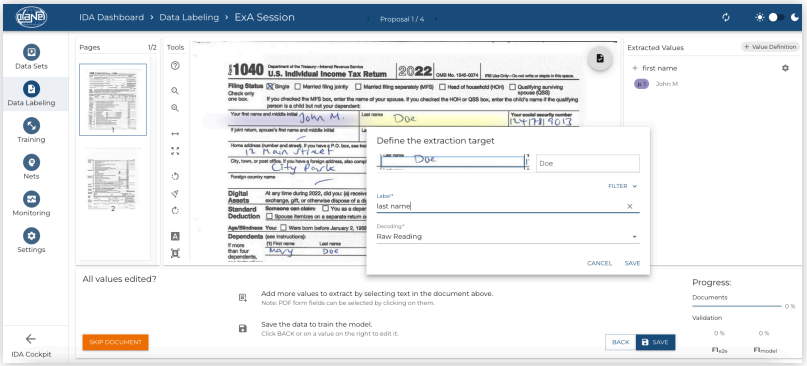
ExA verwendet eine fortschrittliche Methode zur Extraktion von **Schlüssel-Wert-Paaren (Key-Value Pair Extraction)**, die es Benutzer:innen ermöglicht, die Datenfelder, die sie erfassen möchten, einfach zu spezifizieren. Diese Felder sind nicht auf Text beschränkt, sondern können auch Kontrollkästchen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

Modelltraining

Mit dem **Extraction Assistant** bietet IDA eine grafische Schnittstelle, die es Benutzer:innen ermöglicht, Modelle zu trainieren, ohne dass sie komplexe Datensätze vorbereiten müssen. Derzeit funktioniert ExA am besten bei **strukturierten und semistrukturierten Dokumenten** wie z. B. Formularen.

Benutzer:innen können optional das **Layout Language Model (LayoutLM)** verwenden, um ihre Extraktionsergebnisse zu verbessern. Dieses Modell priorisiert visuelle Hinweise und kontextuelles Verständnis und erweist sich als besonders effektiv bei der Verarbeitung semistrukturierter Dokumente. Basierend auf 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

Anwender: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 für eine effizientere Bearbeitung. Nach der Beschriftung eines dieser Musterdokumente (Sample) müssen Benutzer:innen lediglich die vorgeschlagenen extrahierten Felder validieren oder korrigieren. Um ein Muster zu erstellen, sind mindestens **drei Dokumente mit identischem Layout** erforderlich.

AcroForm-Felder werden priorisiert und automatisch zur Verwendung als Datenfeld vorgeschlagen.

IDA LLM ENTITY EXTRACTION

Keyword-basierte Entitätserkennung

Mit IDA LLM Entity Extraction können Sie die Datenextraktion aus unstrukturierten Dokumenten automatisieren, indem Sie **einfache Abfragen (Queries)** formulieren. Dazu gehört die Beschriftung der zu extrahierenden Elemente sowie eine kurze Beschreibung oder eine Liste von Schlüsselwörtern. Ein vorheriges Training ist nicht erforderlich (Zero-Shot Named Entity Recognition).

Verifizierung der extrahierten Daten

Um KI-Halluzinationen zu vermeiden werden die extrahierten Daten mit der Transkription von IDA Recognition **abgeglichen**.

Rechnungsmodul[BETA]

IDA Extraction enthält ein Modul, das speziell für die Verarbeitungen von **Rechnungen** entwickelt wurde und auf LLM Entity Extraction basiert.

No-Code-Lösung

IDA LLM Entity Extraction befähigt Anwender:innen, denen es an technischen Kenntnissen oder an Erfahrung in der Formulierung von Prompts mangelt. Die Anpassung des Moduls ist über ein **Browser-Interface** einfach möglich.

SYSTEMVORAUSSETZUNGEN

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

Für 64-Bit-Systeme

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

Sowohl der benötigte Festplatten- als auch der Arbeitsspeicher sind stark von den jeweiligen Modellen abhängig, die auf dem LLM Server laufen sollen. Es ist kein reiner CPU-Modus möglich.

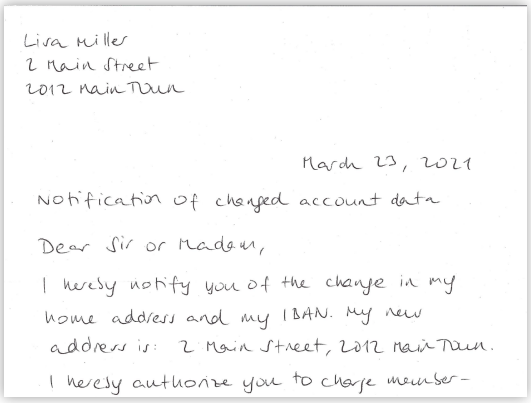
Der LLM Server kann zudem **OpenAI-Modelle** ansprechen, z. B. um Hardwareanforderungen zu reduzieren.

So funktioniert‘s

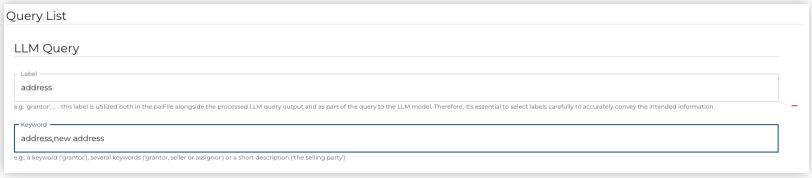
IDA LLM Entity Extraction eignet sich **am besten für die Verarbeitung unstrukturierter Dokumente**, die kein festes Layout oder Datenpunkte haben, wie z.B. Verträge oder Anschreiben. Basierend auf der von IDA Classification durchgeführten Kategorisierung können die Dokumente an verschiedene Extraktionsmodelle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

Eine LLM-Anfrage (Query) besteht aus einem Label, das mit einem Schlüsselwort, einer Liste von Schlüsselwörtern oder einer kurzen Beschreibung gepaart wird. Die Ergebnisse des LLMs werden **anschließend mit dem Entity Finder und der Transkription abgeglichen**. Dies gewährleistet eine hohe Genauigkeit und verhindert fehlerhafte Ergebnisse, die gemeinhin als Halluzinationen bekannt sind.

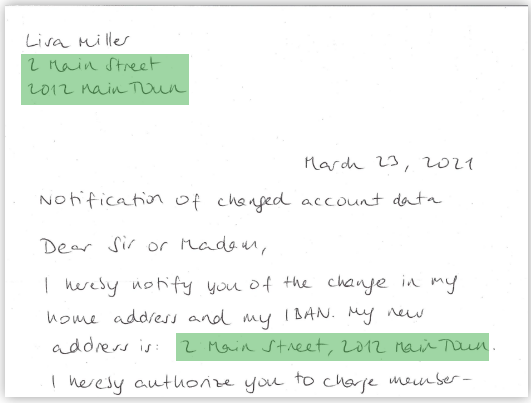
Inputdokument



+



Outputdokument



+

Weitere Informationen sind in der Software-Dokumentation zu finden.