CYBLY

Datenerfassung, Organisation und Management

ReMeP 2024 – 18.11.2024

Bedeutung von Daten für AI

AI-Modelle sind nur so gut wie die Daten, die ihnen zugrunde liegen

Datenqualität ist entscheidend

AI-Modelle „lernen“ aus Beispielen. Schlechte Daten führen zu unzuverlässigen Ergebnissen

Beispiele für negative Auswirkungen

Fehlerhafte Prognosen: Ungenaue Daten können zu verzerrten Ergebnissen führen.
Bias und Diskriminierung: Verzerrte Trainingsdaten führen zu voreingenommenen Entscheidungen.

Merkmale hochwertiger Daten

Vollständigkeit: Alle relevanten Informationen müssen enthalten sein.
Konsistenz: Daten sollten präzise und einheitlich sein.
Aktualität: Veraltete Daten führen zu ungenauen Aussagen.

Datenquellen

Wichtige Grundlagen für effektive AI-Modelle

Interne Datenquellen

Alte Fallakten, Vertragsarchive und juristische Gutachten

Externe Datenquellen

- Öffentliche juristische Datenbanken (z. B. RIS, EurLex).
- Regierungsportale, und offene Datenquellen

Datenanreicherung

Named Entity Recognition (NER) zur Extraktion relevanter Entitäten wie Gerichte, Parteien und Paragraphen.
Automatische Klassifikation von Dokumenten nach Themen (z. B. Arbeitsrecht, Vertragsrecht, Strafrecht).

Datentypen

Der Unterscheidung hängt davon ab, wie systematisch die Inhalte organisiert sind und wie leicht sie automatisch extrahiert und analysiert werden können

Strukturierte Daten

Liegen in festen Formaten vor, die leicht analysiert und in Datenbanken gespeichert werden können (z. B. Tabellen mit definierten Spalten in MySQL).

Halbstrukturierte Daten

Haben eine teilweise Struktur, sind aber nicht vollständig in Datenbankformaten organisiert. Sie enthalten oft Metadaten, die die Struktur beschreiben (z. B. JSON-Dateien, XML, E-Mails).

Unstrukturierte Daten

Haben keine festgelegte Struktur. Inhalte sind in freiem Text oder beliebigen Formaten vorliegend, was die automatische Analyse schwieriger macht (z. B. Fließtexte in Berichten oder juristische Dokumente)

Juristische Dokumente

- **Sichtbare Struktur ≠ Maschinenlesbare Struktur**
  Juristische Dokumente haben Abschnitte und Klauseln, sind aber oft schwer automatisiert zu verarbeiten (PDFs, Scans).
- **Hohe Kontextabhängigkeit**
  Rechtliche Formulierungen sind stark kontextgebunden und erfordern tiefergehende Analyse. Die Interpretation hängt stark vom Kontext und der spezifischen Bedeutung der Formulierungen ab
- **Fehlende Standardisierung**
  Dokumente wie Verträge, Vereinbarungen und rechtliche Gutachten nutzen unterschiedliche Formate und Strukturen.

Ohne eine umfangreiche NLP-Verarbeitung (z. B. Named Entity Recognition, Relation Extraction) sind die Inhalte schwer in strukturierte Daten umzuwandeln. Ein einfaches Parsing oder die Extraktion von Absätzen reicht oft nicht aus, um die Bedeutung vollständig zu erfassen

Daten-anreicherung

Optimierung juristischer Dokumente für AI-gestützte Lösungen

Was bedeutet Datenanreicherung im juristischen Bereich

- Ergänzung bestehender Dokumente mit zusätzlichen Informationen zur Verbesserung der **Datenqualität**.
- Ziel: **Präzisere Analysen, bessere Entscheidungsunterstützung und fundierte Vorhersagen** für juristische AI-Modelle.

Warum ist Datenanreicherung im juristischen Kontext wichtig

- **Vermeidung von Fehlinterpretationen**: Vollständigere und genauere Daten verringern die Wahrscheinlichkeit von **Fehlprognosen** in Rechtsanalysen.
- **Bessere Dokumentensuche und -klassifikation**: AI-Modelle können relevantere Urteile oder Präzedenzfälle identifizieren, wenn die zugrunde liegenden Daten angereichert sind.

Datenablage für AI-Anwendungen

Optimale Datenablage und Anwendungsfälle

- **Vektordatenbanken**
  Vektordatenbanken wie OpenSearch, werden genutzt, um semantisch ähnliche Dokumente aus großen Textkorpora abzurufen, bevor ein Sprachmodell darauf basierend Antworten generiert.
- **Object Stores**
  In Kombination mit Vektordatenbanken, um Rohdaten wie PDFs, Bilder oder gescannte Dokumente zu speichern und dann bei Bedarf darauf zuzugreifen.
- **Graphdatenbanken**
  AI-Modelle nutzen Graphdatenbanken, um Wissensgraphen zu erstellen, die Beziehungen zwischen Entitäten (z. B. Personen, Orte, Ereignisse) abbilden. Dies verbessert die Fragebeantwortung und die Erklärung von Ergebnissen (Explainable AI).

Herausforderungen und Lösungen

Datenqualität und Konsistenz

Oft mehrdeutige Formulierungen, Verwiese, veraltete Informationen.

Lösung:

- Einsatz von NLP zur Bereinigung und Vereinheitlichung von Texten.
- Automatisierte Validierungstools zur Erkennung von Fehlern und Inkonsistenzen in Dokumenten.

Datenschutz

Vertrauliche, sensible Informationen, DSGVO

Lösung:

- Anonymisierungstools
- Zugriffskontrollen und Verschlüsselung zur Sicherstellung der Datensicherheit

Unstrukturiert und komplexe

komplexe Formulierungen, lange Paragraphen und verschachtelte Klauseln.

Lösung:

- Nutzung von NER und Text Parsing zur Extraktion relevanter Informationen.
- Vektordatenbanken zur semantischen Suche und effizienten Abfrage

Heterogene Datenquellen

Daten stammen aus verschiedenen Quellen mit unterschiedlichen Formaten

Lösung:

- ETL-Pipelines
- APIs zur Echtzeit-Datenintegration

Best Practices

- ✓ Datenbereinigung frühzeitig durchführen.
- ✓ Implementieren Sie automatisierte ETL-Pipelines, um Daten aus verschiedenen Quellen zu konsolidieren und eine konsistente Datenbasis zu schaffen.
- ✓ Regelmäßige Überwachung der Datenqualität.
- ✓ Dokumentation und Versionierung aller Datenquellen.
- ✓ Interdisziplinäre Teams zur besseren Dateninterpretation.

LawThek LegalNetics

Die Rechtsinformations-Plattform und das Tool-set zur Datenextraktion und Speicherung der Cybly GmbH

Automatisierte Datenerfassung

Strukturierte Wissensaufbereitung vor allem rechtlicher und administrativer Informationen. Grundlage für Automatisation von Geschäftsprozessen und KI-Anwendungen zur Effizienzsteigerung

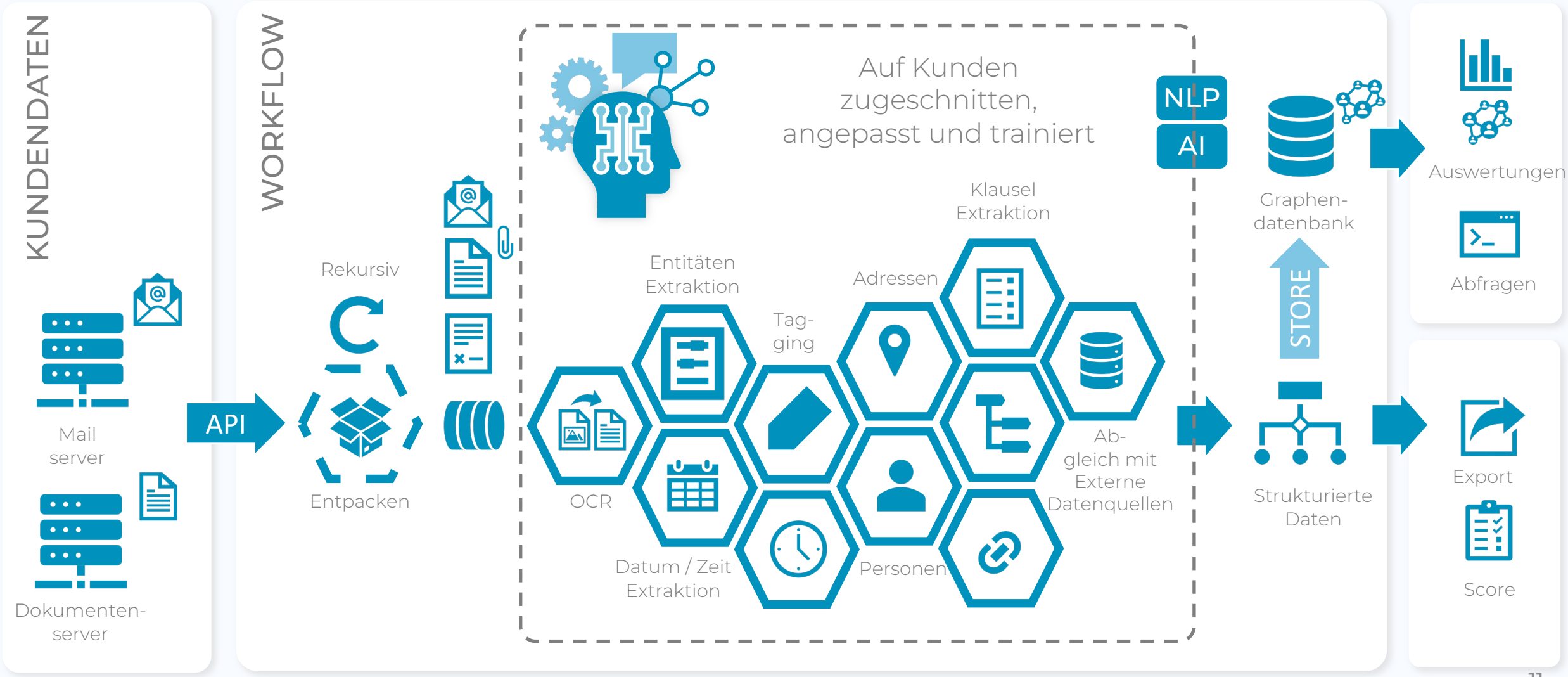
Speicherung in einem Knowledge Graphen

Ergänzung mit unternehmensinternem Know-How, Verlinkung von Inhalten mit internen Quellen und externen Gesetzen, Normen und Urteilen

Schnittstellen zu anderen Anwendungen

Zu unternehmensinternen Anwendungen
Verknüpfung mit dem LawThek Rechtsinformationssystem der Cybly

Beispiel Wissensextraktion



Ausblick

Der Weg zu effektiven AI-Lösungen im juristischen Bereich

Lassen Sie uns darüber sprechen, wie wir Ihre juristischen Prozesse durch datengetriebene AI optimieren können.

Daten sind der Schlüssel

Der Erfolg von AI-Modellen hängt maßgeblich von der Qualität der zugrunde liegenden juristischen Daten ab.

Ganzheitlicher Ansatz

Kombinieren Sie Datenanreicherung, strenge Datenverwaltung und moderne Speichersysteme, um das Potenzial Ihrer AI-Lösungen auszuschöpfen

Potenzial

Der Einsatz von AI in der Rechtsbranche steht erst am Anfang – mit der richtigen Datenstrategie können Sie innovative Lösungen schaffen, die den juristischen Alltag revolutionieren.

Gemeinsam gestalten

Nutzen Sie die Gelegenheit, Juristen, Datenexperten und Technologen zusammenzubringen, um die nächste Generation juristischer AI-Lösungen zu entwickeln.