HEINZ TEMPL
Der Autor ist Rechtsanwalt in Wien.

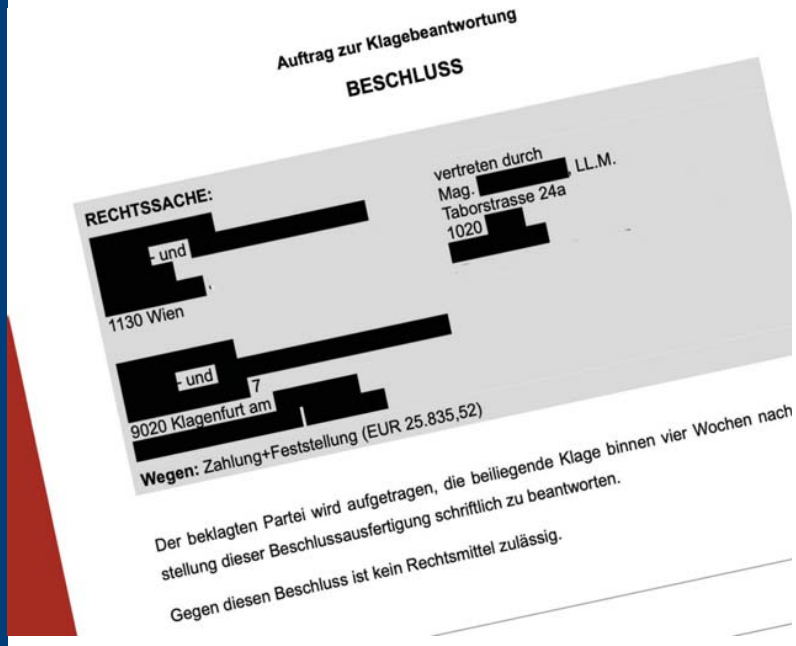
2025/93

Lokale Lösung zur Dokumentenanonymisierung: Ein Praxisbericht

#ki #business #schwaerzen

Eine kostenfreie Python-basierte Methode zur automatisierten Schwärzung von Dokumenten

Der Produktivitätsschub durch KI-gestützte Systeme in der juristischen Praxis ist enorm. Bei der täglichen Arbeit mit Legal-Tech-Tools stellt sich aber rasch die Frage nach einer schnellen, zuverlässigen Möglichkeit zur Anonymisierung von Dokumenten und damit zur Wahrung des Datenschutzes beim Durchsatz von Dateien in KI-Modellen von Drittanbietern. Von der daraufhin entwickelten Lösung möchte ich der juristischen Praxis hier berichten und sie zur Verfügung stellen. Eine schlanke Python-Anwendung, die lokal läuft und Dokumente recht gut effizient automatisch schwärzt.


Auftrag zur Klagebeantwortung

BESCHLUSS

RECHTSSACHE:

[redacted] und [redacted]

1130 Wien

vertreten durch [redacted] Mag. [redacted] LL.M.

Taborstrasse 24a

1020 [redacted]

[redacted] und [redacted] 7

9020 Klagenfurt am [redacted]

Wegen: Zahlung+Feststellung (EUR 25.835,52)

Der beklagten Partei wird aufgetragen, die beiliegende Klage binnen vier Wochen nach stellung dieser Beschlussausfertigung schriftlich zu beantworten.

Gegen diesen Beschluss ist kein Rechtsmittel zulässig.

Anonymisiertes Dokument Foto: RA Templ

Die technische Umsetzung

Nach verschiedenen Tests hat sich eine Kombination zweier Ansätze als besonders effektiv erwiesen: In einem ersten Schritt werden mittels regulärer Ausdrücke (Regex) standardisierte Datenformate wie E-Mail-Adressen und Telefonnummern identifiziert und durch Platzhalter ersetzt. Anschließend analysiert die Natural Language Processing-Bibliothek spaCy den vorbereinigten Text und erkennt kontextabhängig komplexere Entitäten wie Personennamen oder Firmennamen. Diese werden in Word-Dokumenten bei Erkennung automatisch in anonymisierte Bezeichner (zB „Person A“, „Person B“) umgewandelt.

Praktischer Einsatz

Die Implementierung erfolgt als lokales Python-Skript und lässt sich problemlos in bestehende Dokumentenmanagementsysteme integrieren. Der Workflow ist denkbar einfach:

1. Auswahl eines beliebigen Ordners (lokaler Pfad oder Netzwerkverzeichnis)
2. Optional: Automatische Konvertierung von DOCX- und MSG-Dateien in PDF
3. Durchführung der Anonymisierung
4. Export der geschwärzten Dokumente in eigenen Ordner

Für Anwendungsfälle, die eine direkte API-Anbindung erfordern, wurde eine DSGVO-konforme Integration der OpenAI-API implementiert, abgesichert durch das entsprechende Data Processing Agreement.

Technische Details und Verfügbarkeit

Das Skript kombiniert bewährte Python-Bibliotheken: Während Regex zuverlässig strukturierte Daten erkennt, übernimmt spaCy die kontextabhängige Analyse komplexerer Textbestandteile. Diese hybride Methode gewährleistet eine hohe Erkennungsrate bei gleichzeitig schneller Verarbeitung.

Die Lösung steht der juristischen Praxis kostenfrei unter der MIT-Lizenz zur Verfügung. Der Quellcode sowie detaillierte Implementierungshinweise sind über GitHub zugänglich und können von IT-Betreuern in lokalen Softwareumgebungen implementiert werden. Das Skript lässt sich leicht anpassen, wenn Sie etwa mit Azure AI Language oder Microsoft Presidio arbeiten möchten.

Der Vollständigkeit halber: Am Markt sind verschiedene kostenpflichtige Softwarelösungen für derartige Pre-KI-Schwärzungsworkflows verfügbar.

Ausblick

Die bisherigen Praxiserfahrungen zeigen, dass diese schlanke Lösung den Arbeitsalltag spürbar erleichtert. Feedback aus der juristischen Praxis und Weiterentwicklung sind ausdrücklich erwünscht, um die Funktionalität weiter an die Bedürfnisse von Anwendern anzupassen.

Videobeschreibung auf Youtube:

Github: Bibliothek
V1: schlank, nur docx, pdf und msg-Dateien

V2: docx & doc Files können verarbeitet werden, verbesserte Erkennung in Tabellen, Kopf- und Fußzeilen von Wordfiles, verbesserte Open-AI API Einbindung

INFOBOX

Beachten Sie, dass die Vorstellung der hier gezeigten Tools keine Anleitung für die Integration der Anwendungen in Ihre Datenschutz-Umgebung beinhaltet. Betreiben Sie derartige Modelle bzw Software in Ihrer Kanzlei, behalten Sie als Verantwortlicher stets den Überblick über die Art und Rechtmäßigkeit der Verarbeitung und integrieren Sie neue Software insbesondere in Ihr Verarbeitungsverzeichnis.