Neuer Glanz für alte PDFs – Metadaten automatisch generieren

Dr. Jan Oevermann, plusmeta GmbH, Karlsruhe

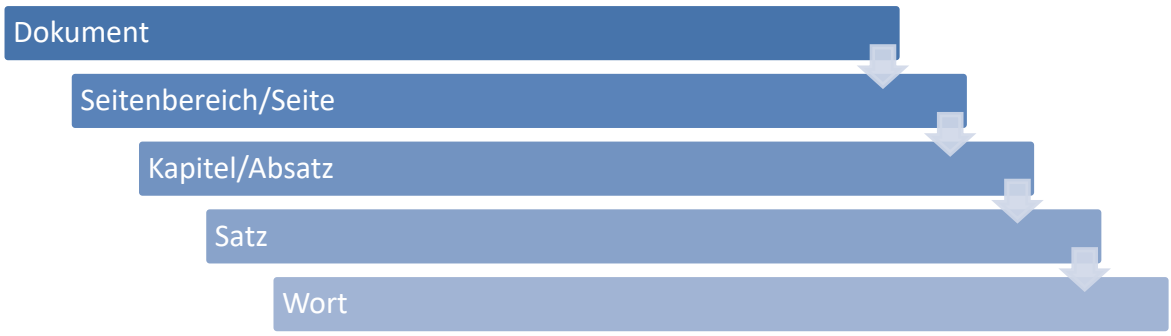
Um PDF-Dateien in moderne Nutzungsszenarien einzubinden, werden Metadaten und Struktur benötigt. Mit Hilfe regelbasierter Ansätze oder Methoden der Künstlichen Intelligenz können Informationen automatisiert extrahiert und ausgewertet werden. Doch Dokumente bieten verschiedene Ebenen zur Metadatenextraktion und -vergabe, die unterschiedliche Anforderungen an die verarbeitenden Systeme stellen und stark abhängig vom individuellen Anwendungsfall sind.

Ebenen für Metadaten

Bei dokumentbasierten PDF-Dateien können auf verschiedenen Ebenen Metadaten vergeben werden – von Metadaten auf Dokumentebene (z. B. dem Dokumenttyp) bis hin zu Annotationen auf Wortebene (z. B. ein erwähntes Produkt oder ein Ort). Je feingranularer die Ebene, desto komplexer wird die Vergabe und desto größer wird die Menge der extrahierten Metadaten.

Ob ein Metadatum, dass in einer Ebenen vergeben wurde (z. B. die Informationsart eines Absatzes) auch auf dieser Ebene vom verarbeitenden System (z. B. einem Content-Delivery-Portal) ausgewertet werden kann, hängt oft von den technischen Begebenheiten ab und ist nicht immer selbstverständlich. So können viele CDP-Systeme Metadaten nur auf der Dokumentebene verarbeiten. Jedoch können Metadaten in der Regel von unten nach oben aggregiert werden, so dass auf Basis einer erkannten Produkterwähnung auf Wortebene die entsprechenden Dokumentmetadaten abgeleitet werden können (z. B. das zum Dokument gehörige Produkt).

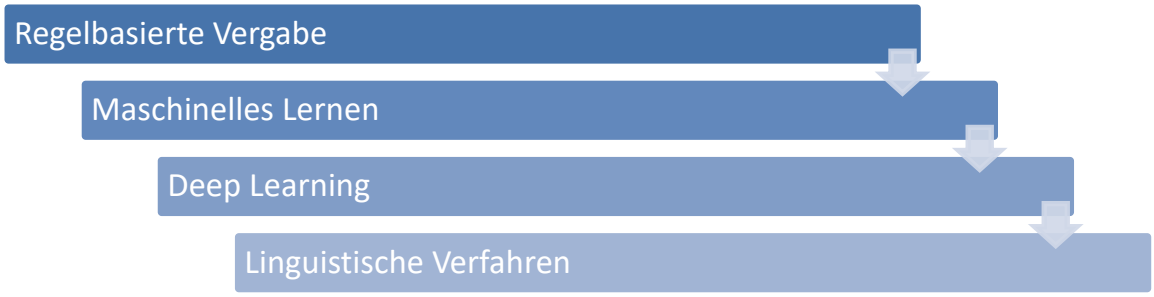
Welches Vorgehen bzw. welche Ebenen am besten geeignet ist, muss vom jeweiligen Anwendungsfall abhängig gemacht werden. Manchmal ist eine regelbasierte Strukturierung eines PDFs in Seitenbereiche auf Basis des Inhaltsverzeichnisses ausreichend; in anderen Fällen sollen auf semantischer Ebene einzelne Absätze gezielt verfügbar gemacht werden.



Ebenen für die Metadatenvergabe bei PDF-Dokumenten

Generierung von Metadaten

Die maschinelle Erzeugung von Metadaten kann mit verschiedenen Verfahren realisiert werden. Beginnend bei regelbasierten Verfahren („wenn Betriebsanleitung auf der Titelseite steht, ist der Dokumenttyp ‚Betriebsanleitung'") bis hin zu leistungsfähigen aber komplizierten Methoden des maschinellen Lernens („die Wortverteilung in diesem Abschnitt deutet auf einen anleitenden Text hin") können verschiedene Wege zum gewünschten Ergebnis führen. Diese können mit den eingangs erwähnten Ebenen oder untereinander kombiniert werden.



Beispielhafte Verfahren zur Generierung von Metadaten

Anwendung

In der realen Anwendung ergibt sich das Verfahren zur Metadatengenerierung oft aus der Ausgangssituation heraus. Sind keine ausreichenden Trainingsdaten (also genügend Dokumente mit manuell vergebenen Metadaten) vorhanden, wie das etwa bei Zulieferdokumentationen der Fall ist, können regelbasierte Verfahren helfen. Die Methoden des Machine und Deep Learning arbeiten sehr effizient und akkurat, sind jedoch auf die oben genannten Trainingsdaten angewiesen und dadurch nicht immer umsetzbar. Eine weitere Möglichkeit bieten linguistische Verfahren, die auf Basis von sprachlichen Merkmalen bestimmte Informationen extrahieren und damit Metadaten erzeugen können (etwa die Unterscheidung von instruktiven und deskriptiven Aussagen in Texten). Durch ihre enge Kopplung an die spezifische Grammatik der jeweiligen Sprache sind sie jedoch sprachabhängig und verhältnismäßig komplex.

Während manche Systeme nur auf eine der oben genannten Methoden zur Metadatengenerierung spezialisiert sind, existieren auch kombinierte Lösungen, die plattformbasiert mehrere Verfahren einsetzen, um einen Anwendungsfall abzudecken. Prinzipiell gilt die Regel, dass immer fallabhängig entschieden werden sollte, welches das beste Werkzeug ist, um die Problemstellung zu lösen.

Literatur

– Oevermann, Jan (2019): Optimierung des semantischen Informationszugriffs auf Technische Dokumentation. tekom Hochschulschriften (Band 25). Stuttgart : tcworld.

Kontakt:
jan@plusmeta.de