Mittelstand 4.0
Kompetenzzentrum
eStandards

Mittelstand-
Digital

Gefördert durch:
Bundesministerium
für Wirtschaft
und Klimaschutz
aufgrund eines Beschlusses
des Deutschen Bundestages

Innovint Aircraft Interior GmbH: Mit neuronalen Netzwerken zur automatisierten Texterkennung

Best Practice

Foto: Unsplash / Alexander Schimmeck

Im Fokus:

Eine effiziente Digitalisierung ist im Büroalltag kaum noch wegzudenken. Immer wieder stellt sich Unternehmenden die Frage, welche Möglichkeiten neue, smarte Technologien, wie zum Beispiel die künstliche Intelligenz, mit sich bringen und wie diese in ihr Unternehmen integriert werden können. Um zeitsparende und zukunftsfähige Strukturen in der Datenverarbeitung aufzubauen, hat das mittelständische Unternehmen Innovint Aircraft Interior GmbH zusammen mit dem Mittelstand 4.0-Kompetenzzentrum eStandards in einem Praxisprojekt nach einer nachhaltigen Lösung gesucht. Durch die Kooperation konnten verschiedene Ansätze mit Hilfe von neuronalen Netzwerken ausprobiert und ein maschinelles Ausleseverfahren von Text erzielt werden.

Die Innovint Aircraft Interior GmbH ist Innenausstatter für Flugzeuge und stellt unter anderem eine Vielfalt an individuell anpassbaren Rollstühlen, Korbwiegen, Trennwänden und Schrankfächern aus einer Hand her. Um ihren Kund:innen einen reibungslosen Ablauf in der Auftragsbearbeitung zu gewährleisten, müssen Wünsche und Bedürfnisse in den Auftragsformularen pünktlich ausgewertet und umgesetzt werden. Da das manuelle Auslesen der Dateien jedoch oftmals mühsam und zeitintensiv ist, hat das Unternehmen gemeinsam mit dem Kompetenzzentrum eStandards ein praktisches Verfahren entwickelt, das diesem Aufwand entgegenwirkt.

KI ist nicht gleich KI

Das Problem: Die Auftragsdateien hatten diverse unterschiedliche Dokumentlayouts. Da die ursprüngliche KI jedoch auf ein einzelnes Dokumentlayout trainiert wurde, musste sie bei jedem abweichenden Layout eines Auftraggebers neu trainiert werden. Dies war nicht nur mit Mehraufwand und Kosten für Innovint verbunden, sondern hätte alternativ nur das händische Auslesen der Dateien zugelassen.

Info: Die Basis von KI

Künstliche neuronale Netze sind die Grundlage für eine künstliche Intelligenz. Sie sind dazu in der Lage, große Mengen an unstrukturierten Daten auszuwerten und Muster in ihnen zu finden. Zu den unstrukturierten Daten gehören unter anderem auch Texte und Bilder. Es wird hier also nach Mustern in den unterschiedlichen Dokumenttexten gesucht. Da das neuronale Netzwerk teilweise fehlerhafte Ergebnisse ausgibt, benötigt es ein ausgiebiges Training mit einem klassifizierten Datensatz – im Fall von Innovint mit dem des Dokumentlayouts.

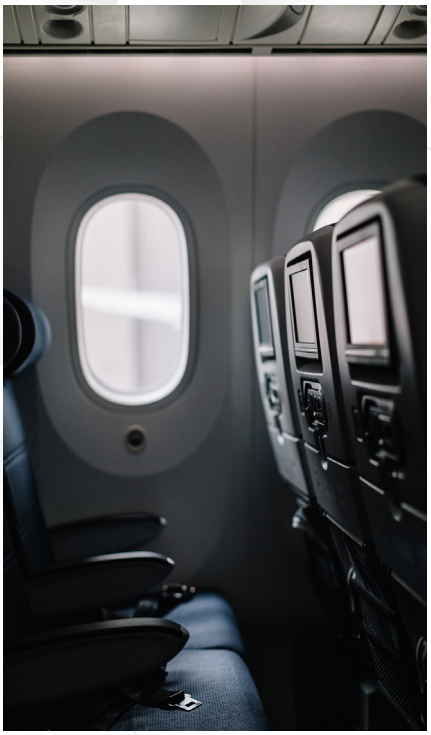

Foto: Pixabay / Stuart Bailey

Mit neuem Ansatz flugbereit

Während des Praxisprojekts wurde ein Ansatz entwickelt, der das Layout übergreifende Auslesen der Daten in den Vordergrund stellte. Die zu lesenden Auftragsdateien liegen dem KMU in der Regel im Portable Document Format (PDF) vor und mussten zunächst zwischen computergenerierten PDFs und direkten Scans unterschieden werden. Während computergenerierte PDFs kopierbaren – und somit maschinell auslesbaren Text wie in einer einfachen Textdatei – beinhalten, handelt es sich bei den Scans um Bilder.

Erkennung textbasierter PDFs

Für textbasierte PDFs gibt es mehrere Auslesemöglichkeiten, da die gängigsten Programmiersprachen darauf ausgelegt sind. Die Expert:innen des Kompetenzzentrums benutzten für die Programmierung der Künstlichen Intelligenz Python. Innerhalb dieser Programmiersprache gibt es Bausteine, sogenannte Module, welche die Software Poppler bereitstellen. Poppler erlaubt es, den Text aus PDF-Dateien auf zweierlei Weisen zu lesen, wie das untenstehende Beispielbild zeigt.

Strukturerhaltend:

Liegt ein PDF mit dem in Abbildung 1 gezeigten Inhalt vor, würde die strukturerhaltende Methode den Text auslesen als „Hier steht eine Kopfzeile. Ein Block voller Text steht auf dieser Seite hier. Ein weiterer Block wurde hier drüben geschrieben“

Roh:

Die rohe Methode ignoriert die Struktur und erkennt „Hier steht eine Kopfzeile. Ein Block voller Ein weiterer Block Text steht auf wurde hier drüben dieser Seite hier. geschrieben.“

Aus beiden Ausgaben kann der gesuchte Inhalt nun durch reguläre Ausdrücke ausgelesen werden. Mit Hilfe von regulären Ausdrücken werden in der Regel Übereinstimmungen, falls vorhanden, an einer beliebigen Stelle innerhalb einer Zeichenkette gefunden. Ein praktisches Beispiel dafür ist die Textsuchfunktion in einem Dokument. Vorteilhaft ist vor allem, dass für Python bereits ein weitentwickeltes Modul namens Invoice2Data existiert. Damit können Vorlagen angelegt werden, welche die einzelnen, auszulesen-

```
Hier steht eine Kopfzeile.


Ein Block voller                                    Ein weiterer Block
Text steht auf                                      wurde hier drüben
dieser Seite hier.                                  geschrieben.
```

Abbildung 1: Beispieltext für die Erkennung textbasierter PDFs

den Felder definieren.
Im Laufe des Praxisprojekts wurde das Modul mit einigen Anpassungen und Erweiterungen übernommen und die endgültige Ausgabe generiert. Dabei ergab sich jedoch eine Schwierigkeit: Die Endnutzer:innen mussten für jeden Auftraggeber bzw. für jedes Format von Auftragsdokumenten eine Vorlage anlegen, welche auch die Beschreibung von regulären Ausdrücken beinhalten musste. Zum Glück ist dafür nicht unbedingt eine IT-Fachkraft von Nöten! Online gibt es eine Reihe von kostenfreien Werkzeugen, die Laien bei der Beschreibung von regulären Ausdrücken behilflich sein können, etwa auf regex101.com oder regex-generator.olaf-neumann.org.

Erkennung bildbasierter PDFs

Die Erkennung bildbasierter PDFs stellte dagegen eine größere Herausforderung dar. Wie oben beschrieben, kann der Text hier nicht einfach aus dem Dokument herauskopiert, sondern muss mit einer Software für Texterkennung generiert werden. Dafür nutzten die Expert:innen die OCR-Software (Optical Character Recognition) Tesseract und steißen auf einige Probleme:

- Tesseract hat in seinem aktuellen Entwicklungsstand Probleme mit unterschiedlichen Schriftarten: Text kann verloren gehen.
- Die Texterkennung ist nicht fehlerfrei: Ein „l“ kann schlecht von einem „I“ oder eine „1“ von einem „t“ unterschieden werden.
- Tesseract gibt zwar an, wie sicher die Richtigkeit der Ergebnisse ist, allerdings kann diese trotzdem bei über 80 Prozent liegen, auch wenn ein falsches Zeichen ausgelesen oder Text nicht erkannt wurde.
- Die Ausgabe muss unter Umständen automatisch oder händisch gegengeprüft werden, um Fehler in Lieferadressen oder Auftragsnummern zu vermeiden.
- Tesseract kann nicht direkt auf PDFs angewendet werden: PDFs müssen konvertiert werden, um ausgelesen zu werden, z. B. in PNG-Dateien mit einer Bildauflösung von mindestens 200 dpi (dots per inch). Bei sehr großen PDFs kann die Konvertierung an die Grenzen der Endnutzerhardware stoßen und sehr lange dauern.
- Es gibt keine Lösung bei einer schlechten Scanqualität: Hier muss der Text per Hand ausgelesen werden.

All diese Komplikationen führten im Fall von Innovint dazu, dass sich die finale Lösung auf die Text-PDFs konzentriert. Da bildbasierte PDFs im KMU nur einen Bruchteil der Auftragsdateien ausmachen, wurden sie zunächst außen vor gelassen.

In Zukunft kann Innovint hier ansetzen und mit neuen, KI-basierten Methoden auch eine möglichst fehlerfreie Erkennung der bildbasierten PDFs erreichen. Da sich smarte Technologien aktuell so schnell wie noch nie weiterentwickeln, ist eine baldige Lösung nicht ausgeschlossen.

Wertschöpfende Übertragbarkeit

Die Herangehensweise im Praxisprojekt mit Innovint ist auf jedes andere KMU übertragbar, das Teil einer Lieferkette oder eines Wertschöpfungsnetzwerks ist und mit Auftragsdateien, Lieferscheinen oder anderen textbasierten Dokumenten zu tun hat. Das Projekt zeigt, dass der Einsatz von KI auf vielfältige Weise möglich ist und dafür nicht unbedingt IT-Fachkräfte im Unternehmen gebraucht werden.

Wir unterstützen Sie gerne beim Einsatz von Künstlicher Intelligenz in diesem Bereich der Digitalisierung. Sprechen Sie uns einfach an!

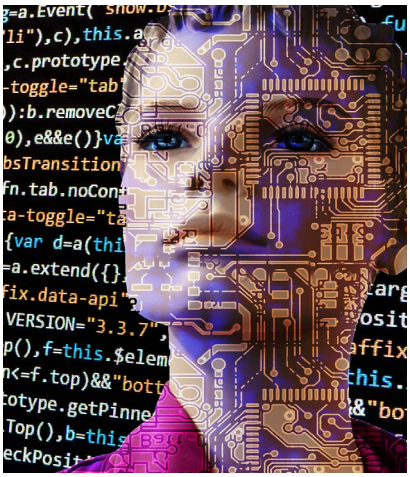

Foto: Pixabay / Gerd Altmann

Impressum:

Autoren: An Pham, Moritz Busch
Redaktion: Lena Köppen
Fotos: Unsplash, Pixabay

Mittelstand 4.0-Kompetenzzentrum eStandards
Co-Working-Space Sankt Augustin
Fraunhofer-Institut für Angewandte Informationstechnik FIT

Kontakt:

Tel: +49 2241 14-3712
koeppen@kompetenzzentrum-estandards.digital
www.kompetenzzentrum-estandards.digital

Das Mittelstand 4.0-Kompetenzzentrum eStandards gehört zu Mittelstand-Digital. Mit Mittelstand-Digital unterstützt das Bundesministerium für Wirtschaft und Klimaschutz die Digitalisierung in kleinen und mittleren Unternehmen und dem Handwerk.

Was ist Mittelstand-Digital?

Mittelstand-Digital informiert kleine und mittlere Unternehmen über die Chancen und Herausforderungen der Digitalisierung. Die geförderten Kompetenzzentren helfen mit Expertenwissen, Demonstrationszentren, Best-Practice-Beispielen sowie Netzwerken, die dem Erfahrungsaustausch dienen. Das Bundesministerium für Wirtschaft und Klimaschutz ermöglicht die kostenfreie Nutzung aller Angebote von Mittelstand-Digital.

Weitere Informationen finden Sie unter
www.mittelstand-digital.de