Tabellen automatisiert analysieren

Künstliche Intelligenz als Performance-Treiber?

Editorial

Neuronale Netze, künstliche Intelligenz (KI), Robotik – in Zeiten der zunehmenden Digitalisierung gewinnen smarte Lösungen an Bedeutung. Und das über Branchen hinweg. Doch was steckt hinter dem mittlerweile schon zum Buzzword gewordenen Begriff? Und was bedeutet KI für die Finanzwelt und insbesondere für die Erfassung von Daten? Wir werfen einen Blick auf die automatisierte Extraktion von Fakten aus den Tabellen von Jahresabschlüssen nach dem HGB.

Viel Freude beim Lesen. Für individuelle Fragen kontaktieren Sie uns gerne.

Herzlichst
Felix Gratz
Leiter Automatisierung, PPA

Was ist künstliche Intelligenz?

Eine klare Definition, was mit Künstlicher Intelligenz bezeichnet werden kann, gibt es nicht. Im weitesten Sinne umfasst der Begriff intelligent erscheinende Systeme. Im engeren Sinne würde man hierunter selbstlernende, auf neuronalen Netzen aufbauende Systeme verstehen. Eine auf solchen Netzwerken aufbauende KI steht einem auf Logik und Arithmetik aufbauenden Algorithmus gegenüber. Setzen Unternehmen für die Analyse ihrer Daten auf eine künstliche Intelligenz: Wo liegen dann die Unterschiede bei Ressourceneinsatz und Ergebnissen?

KI-gestützte Extraktion aus HGB-Haupttabellen: Wo liegen die Herausforderungen?

Um die Fakten einer Tabelle erfolgreich zu extrahieren, stehen einige Bedingungen im Fokus:

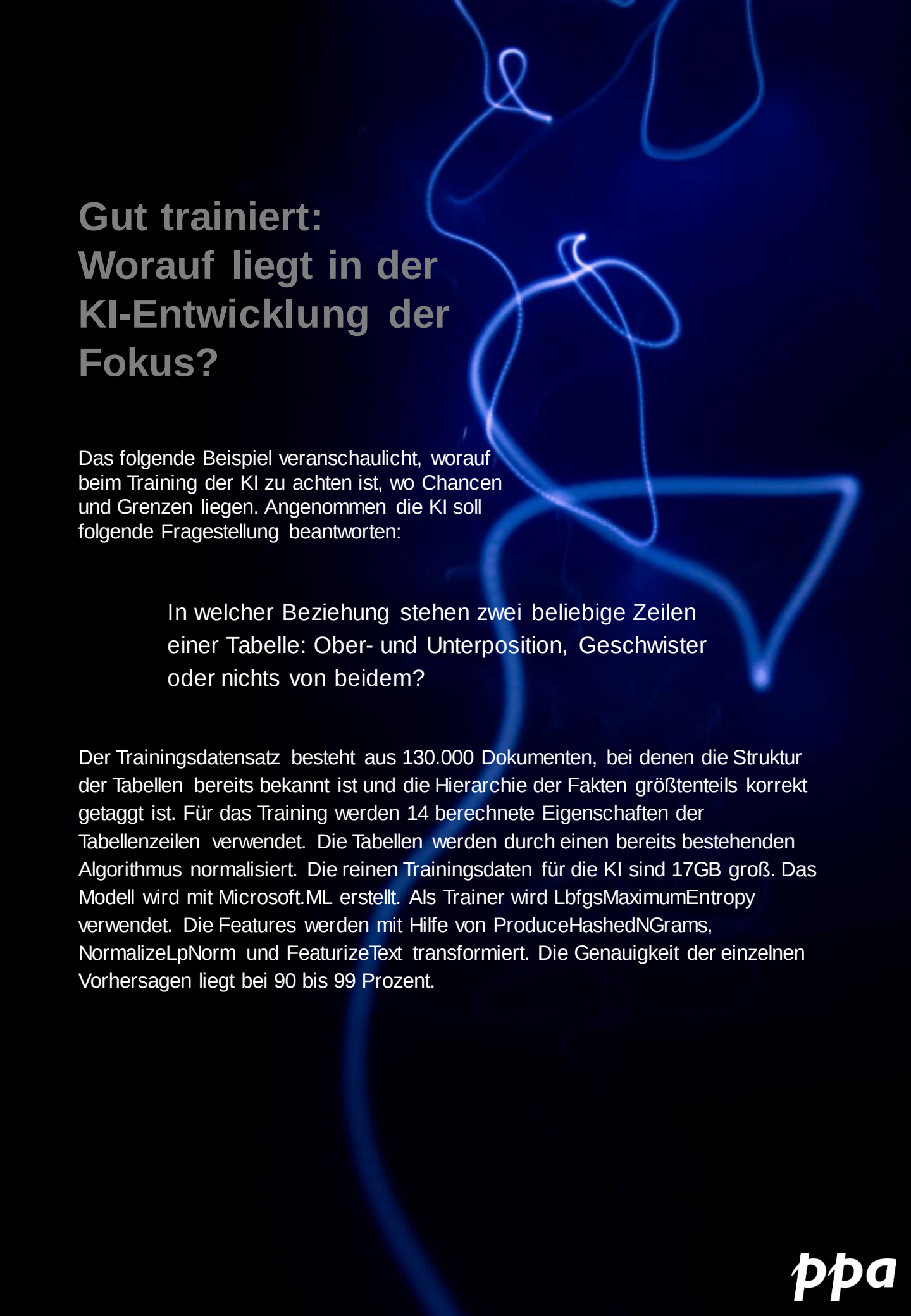
Der Wertetyp jeder Spalte muss bekannt sein (Geschäftsjahr, Vorjahr, Differenz, Prozentwerte etc.).

Jede Zuordnung von Label und Zahl muss bekannt sein.

Die Hierarchie der einzelnen Fakten muss bekannt sein. Ein Fakt ist die Kombination aus Zahl und Label.

Bei einer KI-gestützte Extraktion ist ein gutes Training ausschlaggebend für die Zuverlässigkeit der Daten. Beim Training der KI stehen die folgenden Herausforderungen im Fokus:

- Die Tabellen bestehen aus mehrzeiligen Zeilen. Wie die Zeilen zusammengefasst werden müssen, ist dem Dokument typischerweise nicht zu entnehmen.
- Das Label einer Zahl muss nicht unbedingt in derselben Zeile stehen. Typischerweise stehen die Unterelemente eines Fakts zwischen der „Zahlenzeile“ und der „Labelzeile“ des Fakts.
- Nicht jede Zahl hat ein zugehöriges Label. → Es gibt Fakten, die nur einen Wert aber kein Label haben.
- Nicht jedes Label hat eine zugehörige Zahl. → Es gibt Fakten, die nur ein Label aber keinen Wert haben.



Gut trainiert: Worauf liegt in der KI-Entwicklung der Fokus?

Das folgende Beispiel veranschaulicht, worauf beim Training der KI zu achten ist, wo Chancen und Grenzen liegen. Angenommen die KI soll folgende Fragestellung beantworten:

> In welcher Beziehung stehen zwei beliebige Zeilen einer Tabelle: Ober- und Unterposition, Geschwister oder nichts von beidem?

Der Trainingsdatensatz besteht aus 130.000 Dokumenten, bei denen die Struktur der Tabellen bereits bekannt ist und die Hierarchie der Fakten größtenteils korrekt getaggt ist. Für das Training werden 14 berechnete Eigenschaften der Tabellenzeilen verwendet. Die Tabellen werden durch einen bereits bestehenden Algorithmus normalisiert. Die reinen Trainingsdaten für die KI sind 17GB groß. Das Modell wird mit Microsoft.ML erstellt. Als Trainer wird LbfgsMaximumEntropy verwendet. Die Features werden mit Hilfe von ProduceHashedNGrams, NormalizeLpNorm und FeaturizeText transformiert. Die Genauigkeit der einzelnen Vorhersagen liegt bei 90 bis 99 Prozent.

Klassische Logik: Wie sieht der Aufbau des klassischen Algorithmus aus?

Der klassische Algorithmus versucht auf Basis der arithmetischen Beziehungen der Zahlen die Hierarchie der einzelnen Zeilen zu bestimmen: Wenn ein Satz von Zahlen eine Linearkombination einer anderen in der Tabelle existierenden Zahl ist, steigt die Wahrscheinlichkeit, dass diese Zahl die Oberposition ist. Hinzu kommt noch eine manuell gepflegte Datenbank, die beschreibt in welcher Relation sich bestimmte Labels befinden dürfen.

KI als Ergänzung zum klassischen Algorithmus

Die Entwicklung der KI-basierten Lösung benötigt nur 50 Prozent der Entwicklungsressourcen im Vergleich zum klassischen Ansatz. Allerdings lassen sich die Ergebnisse der KI nicht alleine verwenden. Um korrekte Ergebnisse zu erhalten, müssen weitere klassische Validierungsstufen integriert werden. Sowohl das KI-gestützte System als auch das klassische System können 70 Prozent der Tabellen korrekt erkennen. Allerdings lässt sich mit weiterem Aufwand das klassische System auf Erfolgsquoten von über 80 Prozent bringen. Das KI-basierte System ist hier bereits an seinen Grenzen.

Die Mischung macht´s: KI als Ergänzung zum klassischen Algorithmus

Fazit: Bei ausreichend vorhandenen Trainingsdaten, lässt sich ein KI-basiertes System sehr schnell entwickeln. Wenn dies bereits eine ausreichende Genauigkeit bietet, kann das der effizienteste Weg sein. Die besten Resultate erzielen Unternehmen allerdings, wenn sie beide Systeme intelligent kombiniert einsetzen.

Smart und solide: Integrierte Datenextraktion mit PPA

Wir arbeiten mit smarten Tools und verzichten nicht die auf die Genauigkeit durch tradierte Verfahren. Die PPA erfasst seit über 20 Jahren strukturierte Daten aus Jahresabschlüssen für internationale und deutsche Großbanken. Die maschinelle Unterstützung hierfür bauen die IT-Experten im Unternehmen seit 2010 kontinuierlich aus. Heute extrahiert PPA 45 bis 60 Prozent aller Haupttabellen in HGB-Berichten vollständig und korrekt automatisch. Durch neue Algorithmen und KI erzielte das Darmstädter Unternehmen im Q1 2020 ein deutlicher Sprung der Extraktionsraten. Das Team arbeitet fortlaufend an der weiteren Digitalisierung und Optimierung der Analyse- und Extraktionsprozesse, um Banken stets ein zuverlässiges Ergebnis zu liefern.

Wir freuen uns auf den Austausch

Kontakt:

FELIX GRATZ
Leitung Automatisierung
Felix.Gratz@ppaworld.com
06151 7804 698

ppa