PROVINZIAL NordWest Konzern

OCR im großen Stil – Briefpost als Datenfundament für Machine-Learning-Projekte

Success Story Provinzial NordWest

Aufgabe

- Kostengünstige Texterkennung der gesamten Eingangspost mit Open Source Software
- Vorverarbeitung der Textdokumente zur Verbesserung der Ergebnisse
- Verarbeitung des täglichen Datenaufkommens spätestens bis zu Beginn des nächsten Arbeitstages
- Skalierbare Architektur, die mit dem Datenaufkommen mitwächst

Herausforderungen

- Zurzeit wird nur die erste Seite der Post erkannt und abgelegt.
- Die Abarbeitung ist über den Tag verteilt sehr ungleichmäßig, mit einigen wenigen Peaks.
- Neben der täglichen Volltextindizierung sollen auch große Altarchive verarbeitet werden können.

Vorgehen

- Lose gekoppelte Microservices in Docker-Containern
- Pre-Processing der Dokumente mit openCV
- Deep-Learning-Modelle klassifizieren Dokumente vor
- Für die Schrifterkennung wird Tesseract OCR verwendet
- Volltextsuche und Bewertung der Ergebnisse über Levenshtein-Distanz
- Alle Ergebnisse und Metriken werden in Elasticsearch abgelegt

Ergebnisse

- Docker wird erstmals produktiv im Unternehmen eingesetzt.
- Durch Open Source Tools beschränken sich die Kosten auf den Betrieb
- Hochskalierende und dynamische Lösung verarbeitet täglich über 160.000 Seiten
- Erkennungsrate bei über 70 % für mehr als 90 % der Dokumente
- Hohe Qualität der Ergebnisse bereitet Weg für den Einsatz weiterer KI-Projekte

Der Provinzial NordWest Versicherungskonzern (PNW) startete Ende 2017 das Projekt „Sherlock" zur Volltextindizierung der Eingangspost. Es bildet, neben der Optimierung von Geschäftsprozessen, das Datenfundament für Machine- und Deep-Learning-Modelle. Durch die Nutzung freier und moderner Technologien wie Docker, Python, TensorFlow und Elasticsearch wurde eine kostengünstige und hochskalierbare Lösung entwickelt, die täglich mehr als 200.000 Dokumente indiziert ablegt und durchsuchbar macht.

Ausgangssituation

Der Provinzial NordWest Versicherungskonzern ist in Schleswig-Holstein, Mecklenburg-Vorpommern, Hamburg und Westfalen für seine Kunden vor Ort: Von Westerland bis Rügen und von Viöl bis Hamburg-Harburg reicht das Netz der 220 Versicherungsfachgeschäfte der Provinzial Nord, in Westfalen ist die Westfälische Provinzial zwischen Bocholt und Höxter mit 438 Geschäftsstellen vertreten.

Die PNW digitalisiert schon seit Jahren die Eingangspost (Papier, Fax, Mail) und legt die eingescannten Dokumente digital im Bildformat „TIFF" ab. Aufgrund von Kosten und Rechenlast wurde bisher lediglich die erste Seite einer Dokumentenmappe durch OCR erkannt und zur Klassifikation herangezogen.

Ende 2018 wurde das Projekt „Sherlock" ins Leben gerufen. Zunächst als Proof of Concept entwickelt, hatte Sherlock das Ziel, die gesamte Eingangspost als Volltext durchsuchbar abzulegen.

Die PNW verarbeitet täglich weit über 100.000 Seiten aus Briefen und E-Mail-Dokumenten. Grundvoraussetzung von Sherlock war es, diese Last binnen 24 Stunden zu verarbeiten und damit eine tagesaktuelle Datenbasis über die gesamte Eingangspost für Volltextsuchen aus dem CRM bereitzustellen.

Lösung

Das Projekt stellt gleich eine ganze Reihe neuer Anforderungen an die interne Software-Entwicklung und den IT-Betrieb.

Der Einsatz moderner, aber heterogener Technologien, wie OpenCV, Tesseract, TensorFlow und Keras, erfordert ein hohes Maß an Flexibilität hinsichtlich Entwicklung, Build und Deployment. Um insbesondere in den letzteren Punkten einen gemeinsamen Standard zu schaffen, werden die einzelnen Services von Sherlock in Docker-Containern betrieben.

Zum aktuellen Zeitpunkt besteht Sherlock aus neun, lose durch Queues gekoppelte Services, die über die Anzahl ihrer Container individuell skaliert werden können. Das ist insbesondere aufgrund der hohen Last zu bestimmten Kernzeiten, wie am frühen Vormittag oder abends, wichtig.

Jeder Service führt Tagebuch über seine aktuellen Durchlaufzeiten. Ein Tesseract-Service benötigt zum Beispiel

im Durchschnitt zehn Sekunden pro Seite, während hingegen das Pre-processing, wie Säubern und Hochskalieren, in unter einer Sekunde erledigt ist. Durch die Microservice-Architektur kann Sherlock auf dieses Ungleichgewicht ausgerichtet werden.

Um die Texterkennung zu entlasten, werden die Seiten mithilfe eines trainierten, tiefen neuronalen Netzes in Text- und Bilddokumente unterteilt. Damit lassen sich bereits zu Beginn größere TIFF-Dateien herausfiltern, die ohnehin keinen Text enthalten.

Die Erkennungsrate wird zur Laufzeit anhand eines großen Wörterbuchs in Elasticsearch abgeglichen und gemessen. Elasticsearch stellt im selben Zug auch einen Mechanismus für Wortvorschläge bereit, mit dem Sherlock Fehler in der Erkennung noch einmal ausgleicht.

In Elasticsearch werden die Volltexte anschließend auch persistiert und bereitgestellt.

Ergebnis

Sherlock hat zu Peak-Zeiten 230.000 Seiten pro Tag abgearbeitet. Die Erkennungsraten liegen dabei bei über 70 Prozent für 90 Prozent der eingehenden Dokumente. Hinzu kommen richtig erkannte Eigennamen, die nicht im Wörterbuch enthalten sind.

Das System ist seit September produktiv und hat bereits über 12 Millionen Seiten persistiert, die dem CRM-System mit einer Volltextsuche zur Verfügung stehen.

Außerdem sind bereits neue Projekte auf dem Weg, die auf den Daten aufsetzen. Die Projekte reichen über neue Verfahren zur Dokumentenklassifikation mit Machine-Learning-Modellen bis hin zur Intentionserkennung im Schriftverkehr mit den Kunden.

Neben den Ergebnissen des Projekts und den Folgeprojekten im KI- und Data-Science Bereich wurden auch Erfahrungen im Betrieb von Docker und heterogenen Architekturen gemacht. Mithilfe der Container stellt der Betrieb der Anwendung keinen hohen Aufwand dar und ebnet den Weg für eine heterogene Anwendungslandschaft und damit auch für neue Tools und Möglichkeiten.

Mark Keinhörster
Data Engineer, Münster
mark.keinhoerster@codecentric.de

> *„Das Sherlock Verfahren integriert sich sehr flexibel und hoch skalierbar in unsere Anwendungslandschaft! Dabei unterstützt Docker hervorragend. Das Nutzenpotenzial der Volltextdatenbank ist sehr groß, was sich sowohl in der Bearbeitung einzelner Dokumente bemerkbar macht, als auch in der übergreifenden Analyse von Dokumenten."*
>
> ***Matthias Kortbus**, ITK-Aktivitäten/Dokumente*