Analyse und Visualisierung der OCR-Konfidenz für Dokumente aus den Digitalen Beständen der Staats- und Universitätsbibliothek Hamburg

Exposé zur Abschlussarbeit im Qualifikationskurs „Data Librarian“ an der TH Köln

Michael Kubina – 15.07.2022

Zielsetzung

Die geplante Abschlussarbeit hat zum Ziel am Beispiel von retrodigitalisierten Dokumenten der Staats- und Universitätsbibliothek Hamburg die vermutete Qualität von OCR-Ergebnissen statistisch auszuwerten und die Ergebnisse zu visualisieren. Dies ist möglich für Dokumente, bei denen die OCR-Engine Konfidenzwerte ausgegeben hat. Die OCR-Konfidenz bemisst den Grad der Sicherheit einer OCR-Engine die jeweiligen Textkomponenten richtig erkannt zu haben (IMPACT, 2022). Die OCR-Konfidenz gibt somit nicht zwangsläufig Auskunft darüber, wie hoch die tatsächlichen Erkennungsraten waren und ob die Volltexte korrekt sind, sondern ist höchstens ein Indikator dafür. Dennoch kann die Erkennungssicherheit wichtige Rückschlüsse dafür liefern, wo potentiell schlechte OCR-Ergebnisse vorliegen könnten.

Vorgehen

Für die Analyse und Visualisierung der Ergebnisse wird ein mehrstufiges Python-Script geschrieben, welches auf Bibliotheken für Datenanalyse und Datenvisualisierung zurückgreift. Im ersten Schritt muss die METS-Datei, die das digitale Objekt zusammenhält und beschreibt, heruntergeladen werden. Daraus werden die URLs für die OCR-Volltexte (ALTO-Dateien) extrahiert und anschließend auch diese heruntergeladen. Im zweiten Schritt werden alle Wortkonfidenzen aus den ALTO-Dateien extrahiert. Dies erfolgt für jeweils jedes Wort jeder Zeile einer jeder Seite eines Dokuments. Im dritten Schritt wird aus den extrahierten Wortkonfidenzen eine Liste aus Dataframes erzeugt, bei der jeder Dataframe einer Seite mit allen Textzeilen und deren Wortkonfidenzen entspricht. Diese Dataframes können sodann bereinigt und statistisch ausgewertet werden. Im vierten Schritt werden die Werte in Anlehnung an die Klimastreifen von Ed Hawkings visualisiert (HAWKINS, 2018), bei der die globale Temperaturentwicklung in Form von aneinandergereihten vertikalen Streifen dargestellt wird. Die Farbe eines jeden Streifen hängt davon ab, inwieweit dieser von einem gegebenen Mittelwert abweicht. In diesem Projekt wird daher das Gesamtdokument so visualisiert, dass jeder Streifen einer Textseite entspricht und die Abweichung der mittleren OCR-Konfidenz der gesamten Seite von einem Mittelwert farblich hervorgehoben wird. Zudem wird für jede Textzeile einer Textseite ebenfalls eine solche Visualisierung vorgenommen, bei der ein Streifen der OCR-Konfidenz eines Wortes dieser Textzeile entspricht. Die so visualisierten Textzeilen werden anschließend zusammengefügt um eine Heatmap für die jeweilige Seite zu erzeugen. Im fünften Schritt wird ein Report erstellt, der die Ergebnisse der Auswertung auf einer übersichtlichen Seite aggregiert. Hierfür wird ein HTML-Dokument erzeugt und es werden alle generierten Bilder und statistischen Daten eingebunden. Im sechsten und letzten Schritt soll das Python-Script auf mehrere echte Dokumente angewandt werden, so dass eine Betrachtung und kurze Diskussion der Ergebnisse erfolgen kann.

Alle Dokumente sollen auf einem öffentlichen Repository und soweit möglich unter freier Lizenz veröffentlich werden.

Literaturverzeichnis

IMPACT Centre of Competence in Digitisation (2022) *Confidence Level (OCR)* [online]. San Vicente del Raspeig: Fundación General de la Universidad de Alicante. https://www.digitisation.eu/glossary/confidence-level-ocr/ [Zugriff am: 15.07.2022]

HAKINS, ED (2018) *Warming stripes* [online.]. *Ed Hawkins*: Climate Lab Book. https://www.climate-lab-book.ac.uk/2018/warming-stripes/ [Zugriff am: 15.07.2022]