Richtlinie zum „Software-Sprint“

Listen OCR – Listen OCR fü r die Arolsen Archive

Schlussbericht

Zuwendungsempfänger:

Thomas Werkmeister

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS21S34 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Die Arolsen Archive (https://arolsen-archives.org/) sind ein archivarisches Denkmal für die Verfolgten des Nationalsozialismus. In ihrem Bestand befinden sich Millionen von Dokumenten aus den Arbeits- und Konzentrationslagern der NS- Zeit sowie Dokumente über Familienzusammenführungen und Übergangslagern für die befreiten Menschen nach dem Ende des Zweiten Weltkriegs.

Für viele dieser Dokumente bestanden bereits Prozesse und technische Verfahren, die die gewünschten Ergebnisse mit vertretbarem Aufwand erbringen konnten. Bei Listendokumenten, wie zum Beispiel Transportlisten oder Häftlingsbüchern stießen die Archive jedoch immer wieder an ihre Grenzen. In den Archiven existieren in etwa 700.000 dieser Dokumente, die ganz zentral für die Nachverfolgung der Opfer sind. Meine Motivation war hier einen Beitrag zu leisten, um die wichtige, geschichtliche Arbeit der Archive technisch zu unterstützen.

Die Vorgehensweise habe ich zu Anfang grob in die folgenden drei Phasen unterteilt:

- Vorbereitung: tiefgehende Sichtung der Daten. Typen, Charakteristika, Cluster finden
- Hauptteil: Entwicklung von verlässlichen Methoden zum automatischen Erfassen der (Teil)Inhalte oder zum Zerteilen der Dokumente fürs Crowd-sourcing. Iteration der Methoden
- Nachbereitung: Veröffentlichung der Dokumente durch die Arolsen Archives, Veröffentlichung der Methoden, Erkenntnisse und des Source Codes

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Meine direkte Zielgruppe sind die Arolsen Archives. Zu ihrem Projekt der Aufarbeitung der Dokumente würde habe ich direkt beitragen. Durch das Lösen der Probleme mit Listendokumenten wird das offene Online Archiv vollständiger, und weiterhin werden weiterführende Themen wie das Nachvollziehbarmachen von Einzelschicksalen möglich werden.

Darüber hinaus sind bei diesem Projekt interessante Softwarekomponenten entstanden, die sich für ähnliche Probleme wiederverwenden lassen. So musste ich im Projektverlauf zum Beispiel feststellen, dass ich auch nach einiger Suche keine funktionierende Software für die Annotation von Tabellen in Bildern finden konnte. Die Annotation von Tabellen ist besonders aufwendig, da man bei Naiven Ansätzen die gesamte quadratische Komplexität der Tabelle abarbeiten muss. Eine Tabelle mit 5 Spalten und 50 Zeilen hat 250 Zellen. Diese Zellen einzeln über Rechtecke zu annotieren ist unwirtschaftlich. Für die Dokumente der Arolsen Archive kann mein Annotator eine derartige Tabelle im besten Fall mit lediglich 8 Klicks annotieren. Zwei für den Tabellenkörper, einer für das automatische Setzen der Zeilenlinien, Fünf für die Spaltenlinien.

Im Laufe der weiteren Entwicklung, werde ich diese Funktionalitäten über den Anwendungsfall der Arolsen Archive hinaus generalisieren und auch für andere Archive anbieten. Damit kann der Tabellenannotator für Archive und Forscher im Bereich Tabellenstrukturextraktion Nutzen schaffen.

Ausführliche Darstellung der Ergebnisse

Die wichtigsten Ergebnisse sind die Entwicklung einer praktikablen Methode und eines funktionierenden Prototyps für die Problemstellung der Arolsen Archive. Mit dem Prototypen, der bereits von den Archiven genutzt wird, lässt sich die Zeit, die es braucht die Daten einer langen Tabelle zu extrahieren, von 20 Minuten auf sechs Minuten reduzieren. Für die gegebene Herausforderung, ist das ein großer Erfolg. Mir wurde seitens der Arolsen Archive mehrfach erwähnt, dass sich an den Listendokumente in der Vergangenheit schon Entwickler mehrerer namhafter Unternehmen ohne Erfolg versucht haben. Ein großer Anteil des Erfolgsrezeptes war die Erkenntnis, dass die Dokumente zu heterogen für einen vollständig automatisierten Ansatz sind. Dazu kam die Bereitschaft auf einen semiautomatisierten Ansatz zu setzen, der nicht nur auf der Daten- und KI-Ebene stattfindet, sondern auch die Annotatoren der Arolsen Archive und deren Freiwillige miteinbezieht.

Gleichzeitig war es durch diese Umstände unmöglich die gesamten Dokumente innerhalb der Projektphase zu bearbeiten. Die Annahme am Anfang des Projektes, dass ein automatischer Ansatz möglich wäre, hat sich wie beschrieben als falsch herausgestellt. Dies ist jedoch kein zu großer Rückschlag für das Projekt, da ein praktikabler Ansatz geschaffen wurde, der bei den Arolsen Archiven auf viel Resonanz getroffen ist und jetzt auch in Zusammenarbeit nach der Projektphase weiterentwickelt wird.

Über den Softwareprototypen hinaus, ist außerdem bisher ein Datensatz von rund 250 Tabellenstrukturannotationen und 18000 Zellbildern mit Text entstanden, der sich verwenden lässt um ähnliche Systeme zu trainieren.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Für die Arolsen Archive habe ich einen Ansatz geschaffen, der vielversprechend aussieht, um die Personalien aus den Listendokumenten zu extrahieren. Seit Anfang März 2022 haben wir daher mit den Archiven eine weitere Projektphase eingeleitet, um die Software in Arbeitsabläufe zu integrieren und die Nutzerfreundlichkeit zu verbessern. Es gibt Pläne für weitere Projektphasen zur Erweiterung der Funktionalität, wie zum Beispiel die bessere Erkennung von Handschrift, oder die Annotation von Datentypen auf Text- statt auf Spaltenebene.

Diese Verbesserungen werden auch weiterhin in das Projekt einfließen. Gleichzeitig muss immer abgewägt werden, welchen Anteile Nutzerspezifisch sind, und welche generalisierbar für eine breitere Nutzergruppe. Die Open-Source-Stellung wird vorallem andere Nutzergruppen mit eigenen Entwicklungskapazitäten und größeren Dokumentenbeständen ansprechen, da diese dann selbst Anpassungen für ihre Dokumententypen vornehmen können. Das können sowohl andere Archive als auch Personen aus dem Forschungsbereich sein.

Gleichzeitig sehe ich Potential, die Software auch als fertige, gehostete Lösung anzubieten. Ich denke es gibt hier durchaus eine Möglichkeit auch einer breiteren Nutzergruppe, die nur ab und an mit Tabellen zu tun haben eine Lösung zu bieten. Das könnte auch für Archive ohne eigene Entwickler interessant sein.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

In der Anfangsphase waren sieben verschiedene Teilansätze im Gespräch. Am meisten Zeit ist dabei in den Ansatz geflossen, der zuerst Text auf Dokumenten erkennt und basierend auf Mustern der Textpositionen versucht Tabellen ausfindig zu machen. Dieser Ansatz wurde auch bereits in der Forschungsliteratur verfolgt und mir in einigen Gesprächen mit Forschern als sinnvoll bestätigt. Allerdings habe ich diesen Ansatz schon im ersten Monat wieder verworfen, da die Korrektur der Texterkennung sehr aufwendig ist, nicht alle Tabellenzellen Text beinhalten, und die Dokumente oft irrelevantes Beiwerk haben. Im Vergleich der Vor- und Nachteile der Ansätze hat sich der Ansatz behauptet, der die Tabellen an als wichtigste und erste Einheit sieht und daraufhin die Texterkennung innerhalb der Zellen durchführt.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Demovideo:

https://demoweek.prototypefund.de/projects/17-listen-ocr.html

Code:

https://github.com/twerkmeister/table-annotator

https://github.com/twerkmeister/table-segmenter

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Insgesamt hat sich das Projekt als schwieriger herausgestellt als Anfangs gedacht. So war zum Beispiel eine vollständig automatisierte Lösung des Problems nicht möglich. Gleichzeitig haben sich dadurch auch viele andere methodisch wertvolle Lösungswege eröffnet, die innerhalb des geplanten Budgets zu einem gut funktionerendem Prototypen geführt haben, dessen Weiterentwicklung jetzt auch über die Arolsen Archive finanziert wird.

Während der Projektphase wurde stets gewissenhaft gearbeitet und es wurden permanent Fortschritte erzielt. Der iterative Ansatz der Entwicklung hat sich hier bewährt. Es wäre unmöglich gewesen von vornerein die konkreten Herausforderungen und Reibungspunkte der späteren Entwicklung vorherzusagen. Man braucht echtes Feedback der Nutzung und von objektiven Messwerten, zum jeweiligen aktuellen Stand, um abzuschätzen welche Weiterentwicklungen derzeit das beste Verhältnis zwischen Kosten und Nutzen haben.

Ein enorm wichtiger Punkt bei der Kosteneinhaltung war die regelmäßige Einbeziehung externer Fachleute. So habe ich im zweiwöchentlichen Takt die derzeitigen Entwicklungen mit einem KI-Forscher in diesem Bereich und mit einem sehr kompetenten Softwareentwickler/Produktmanager diskutiert. Diese Gespräche haben mir oft frühzeitig Sackgassen aufgezeigt und viele Fehlinvestitionen meiner Zeit erspart.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Während der Projektphase habe ich im wöchentlichen Austausch mit den Arolsen Archiven gestanden, und mit Ihnen den aktuellen Fortschritt diskutiert, und mehr über die Hintergründe und Eigenheiten der Dokumente erfahren. Mir ist nicht bekannt, dass es während der Projektphase weitere Entwciklungen bei den Arolsen Archiven oder anderen Firmen zu diesem Thema gegeben hat.