Toolbeitrag: OCR4all

Mareike Schumacher [1]

1. Universität Regensburg

Thema:	Textdigitalisierung und Edition	DOI:	10.48694/fortext.3743
Jahrgang:	1	Ausgabe:	3
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2019-09-09 auf fortext.net
Lizenz:	CC BY SA		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

Der Workflow von OCR4all: Die Bilddateien werden in der vorbereiteten Ordnerstruktur abgelegt und können dann auf der Benutzeroberfläche ausgewählt und bearbeitet werden; der erkannte und verbesserte Text wird schließlich als TXT- oder XML-Datei wieder in den Ordnern abgespeichert

- **Systemanforderungen:** Nutzbar mit Linux (empfohlen), Windows und MacOS
- **Stand der Entwicklung:** +OCR4all läuft in der ersten Produktivversion, die kontinuierlich verbessert wird
- **Herausgeber:** Universität Würzburg
- **Lizenz:** Kostenfrei zugänglich
- **Weblink:** https://www.ocr4all.org/
- **Im- und Export:** Als Importformate eignen sich sowohl Bildformate (z. B. PNG, JPG) als auch das PDF-Format (vgl. PDF), Texte können im TXT- (vgl. Reintext-Version) oder XML-Format (vgl. XML) gespeichert werden
- **Sprachen:** Erkennung von über 200 Sprachen (u.a. Latein, Deutsch, Französisch, Niederländisch)

1. Für welche Fragestellungen kann OCR4all eingesetzt werden?

OCR4all erleichtert die Digitalisierung (Horstmann 2024) historischer Drucke und ermöglicht die Texterkennung unterschiedlicher Schrifttypen. Das OCR4all-Texterkennungstool könnte besonders gut für ein Forschungsprojekt eingesetzt werden, in dem historische Faksimiles die Untersuchungsgegenstände darstellen. Mit Hilfe des Tools können beispielsweise Texte auf Handzetteln von Theatervorführungen aus dem 19. Jahrhundert erkannt werden, um diese dann miteinander abzugleichen. Mit Hilfe eigener Verbesserungen, durch die das Tool lernt (vgl. Machine Learning), können auch weitaus ältere Texte computerlesbar gemacht werden. Eine weitere mögliche Fragestellung wäre z. B.: Welche Persönlichkeiten wurden besonders häufig in Schriften des frühen Protestantismus erwähnt?

2. Welche Funktionalitäten bietet OCR4all und wie zuverlässig ist das Tool?

Funktionen (Auswahl):

- Integration gängiger und sehr mächtiger Texterkennungsprogramme in eine einheitliche Benutzeroberfläche (vgl. GUI)
- Automatische Texterkennung in Scans von Fraktur- und Antiquaschriften aus dem 19. Jahrhundert

- Semi-automatische Erkennung frühneuzeitlicher gedruckter Texte
- Preprocessing der Scans (z. B. Erkennung der Schriftbereiche in binären und graustufigen Bildern, Umrechnen schief eingescannter Textbereiche in gerade Textblöcke)
- Segmentierung
- Erkennung von Layout und Textregionen sowie Textzeilen, dazu eine Korrekturoberfläche zur Verbesserung der Ergebnisse
- Zeichenerkennung auf Grundlage von Zeilenbildern, die im Layout erkannt bzw. festgelegt wurden
- Abschließende Textkorrektur
- Training (vgl. Machine Learning) eigener, projektspezifischer OCR-Modelle
- Evaluation der eigenen Korrektur- und Trainingsarbeit

Zuverlässigkeit: OCR4all ist ein schnell und zuverlässig laufendes Texterkennungstool. Die Qualität der gespeicherten Textdokumente hängt stark von der Bildqualität der eingescannten Faksimiles ab. Mit einem iterativen Ansatz, bei dem die Ergebnisse in mehreren Durchläufen verbessert werden und das Tool dabei für die projektspezifischen Materialien optimiert wird, erreicht OCR4all Erkennungsquoten von bis zu 99,5%.

3. Ist OCR4all für DH-Einsteiger*innen geeignet?

Checkliste	✓ / teilweise / -
Methodische Nähe zur traditionellen Literaturwissenschaft	✓
Grafische Benutzeroberfläche	✓
Intuitive Bedienbarkeit	teilweise
Leichter Einstieg	teilweise
Handbuch vorhanden	✓
Handbuch aktuell	✓
Tutorials vorhanden	teilweise
Erklärung von Fachbegriffen	✓
Gibt es eine gute Nutzerbetreuung?	✓

Die Entwickler von OCR4all haben es sich zur Aufgabe gemacht, einen Einstieg in die Texterkennung zu bieten, der besonders für Nutzer*innen ohne Vorkenntnisse geeignet ist. Dieses Ziel erreichen sie mit einer gut strukturierten grafischen Benutzeroberfläche, Handbüchern in Deutsch und Englisch und einer sehr zugewandten Nutzerbetreuung. Allerdings läuft OCR4all in einer Docker-Umgebung, die vor der Nutzung auf dem eigenen PC eingerichtet werden muss. Für diese Einrichtung werden Commandline-Programme (vgl. Commandline) benötigt, deren Verwendung für wenig technikaffine Nutzer*innen ungewohnt sein kann. Gleiches gilt für die Vorgänge des Daten-Up- und -Downloads, die über die computerinterne Ordnerstruktur geregelt werden. Die von OCR4all bereitgestellten Tutorials zur Einrichtung des Programmes decken derzeit Linux- und Windows-, nicht aber Mac-Umgebungen ab.

4. Wie etabliert ist OCR4all in den (Literatur-)Wissenschaften?

OCR4all wurde im Jahr 2019 herausgebracht. Es handelt sich folglich um eine Neuerscheinung, bei der sich zeigen wird, wie kompatibel das Werkzeug mit den Ansprüchen der Fachgemeinschaft ist. Es arbeiten bereits zahlreiche, auch eher traditionell arbeitende Geisteswissenschaftler*innen, mit OCR4all. In wissenschaftlichen Veröffentlichungen wird das Tool bisher allerdings noch nicht erwähnt. Eine Publikation zur Entwicklung von OCR4all wird derzeit vorbereitet.

5. Unterstützt OCR4all kollaboratives Arbeiten?

OCR4all wurde für die Benutzung durch einzelne Nutzer*innen entwickelt, kann allerdings auch gemeinsam genutzt werden, wenn das Tool auf einem durch ein Passwort geschützten Server installiert wird.

6. Sind meine Daten bei OCR4all sicher?

Ja, wenn Sie die von OCR4all genutzte Container-Software Docker lokal auf Ihrem Computer installieren und keinen Web-Zugang wählen. Während der Installation von OCR4all installieren Sie einen Docker-Container auf Ihrem PC. Dabei handelt es sich um einen Teil ihres Arbeitsspeichers, der eingekapselt wird und über den die Installationsdaten über ein mit Hilfe von Docker geteiltes Laufwerk ausgetauscht werden. Um die Installation durchführen zu können, ist ein Admin-Zugriff auf Ihren Computer notwendig. Theoretisch birgt dieses Verfahren

die Gefahr, dass auch schadhafte Dateien auf Ihrem System eingerichtet werden könnten. Da OCR4all an der Universität Würzburg entwickelt wurde und somit nur Daten (bzw. Updates) von regionalen Servern zu ihnen gelangen, die hohen Sicherheitsanforderungen genügen, ist diese Gefahr allerdings sehr gering.

Ihre Bild-und Textdaten bleiben auf Ihrem eigenen System. Ausschließlich Nutzer*innen Ihres Computers können darauf zugreifen. Selbst wenn Sie OCR4all auf einem Server installieren und kollaborativ nutzen, haben nur diejenigen Zugang zu den hier abgelegten Daten, die auf den Server zugreifen können. Aus urheberrechtlicher Sicht ist die Nutzung von OCR4all also vollkommen unbedenklich.

Externe und weiterführende Links

- OCR4all: https://web.archive.org/save/https://www.ocr4all.org/ (Letzter Zugriff: 04.06.2024)

Bibliographie

Horstmann, Jan. 2024. Methodenbeitrag: Möglichkeiten der Textdigitalisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3741, https://fortext.net/routinen/methoden/moeglichkeiten-der-textdigitalisierung.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

CSV CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv`, sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML

implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

PDF PDF steht für *Portable Document Format* . Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein OCRter Scan oder ein am Computer erstellter Text.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

Server Ein Server kann sowohl hard- als auch softwarebasiert sein. Ein hardwarebasierter Server ist ein Computer, der in ein Rechnernetz eingebunden ist und der so Ressourcen über ein Netzwerk zur Verfügung stellt. Ein softwarebasierter Server hingegen ist ein Programm, das einen spezifischen Service bietet, welcher von anderen Programmen (Clients) lokal oder über ein Netzwerk in Anspruch genommen wird.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.