Kompetenzzentrum OCR

Automatische Texterkennung als Serviceangebot

Will, Larissa
larissa.will[at]uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID-iD: 0009-0004-6220-8939

Huff, Dorothee
dorothee.huff[at]uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID-iD: 0000-0003-0866-9967

Zusammenfassung: Die Möglichkeiten, die verschiedenen Programme im Bereich automatisierter Texterkennung heutzutage bieten, sind vielfältig. Deren Anwendung sowie die Vor- und Nachverarbeitung der Digitalisate ist jedoch nicht immer intuitiv. Im Projekt OCR-BW haben die Universitätsbibliotheken Mannheim und Tübingen seit 2019 das „Kompetenzzentrum Volltexterkennung von handschriftlichen und gedruckten Werken“ aufgebaut und beraten seitdem Informationseinrichtungen und wissenschaftliche Projekte in Baden-Württemberg zu diesem Thema. Das umfangreiche Know-how im Bereich automatisierte Texterkennung und die verschiedenen Serviceangebote des Kompetenzzentrums sollen hier erläutert werden und Wissenschaftler*innen hinsichtlich der Einsatzmöglichkeiten von Texterkennungssoftware informiert werden.

Abstract

Durchsuchbare Volltexte für historische Drucke und Handschriften bieten einen zeitgemäßen, umfassenden Zugang zum Kulturgut und können als Grundlage für Anwendungen im Bereich der Data Science dienen.[1]
Die Möglichkeiten, die die verschiedenen Programme in diesem Bereich mittlerweile bieten, sind breit, jedoch ist die Anwendung sowie die Vor- und Nachverarbeitung nicht immer intuitiv. Im Projekt OCR-BW haben die Universitätsbibliotheken Mannheim und Tübingen seit 2019 das „Kompetenzzentrum Volltexterkennung von handschriftlichen und gedruckten Werken“ aufgebaut und beraten seitdem

[1] Weil und Kamlah, 2019.

Informationseinrichtungen und wissenschaftliche Projekte in Baden-Württemberg und darüber hinaus zu diesem Thema.[2]

Das Kompetenzzentrum kann ein breites Know-how für unterschiedliche Programme wie z. B. Tesseract[3], Transkribus[4] und eScriptorium[5] vorweisen. Die UB Mannheim ist aktuell mit zwei Teilprojekten an OCR-D[6] beteiligt, wodurch auch hier Synergien entstehen.[7] Nach Auslaufen des Projekts OCR-BW 2022 werden die Services im Sinne der Nachhaltigkeit als Teil des bibliothekarischen Portfolios fortgeführt. Durch die Volltexterkennung von Handschriften und historischen Drucken werden sowohl Forschenden neue Möglichkeiten im Umgang mit Quellen in der wissenschaftlichen Arbeit ermöglicht als auch Bibliotheken ein doppeltes Tätigkeitsfeld eröffnet.[8] Neben dem Einsatz für die Bereitstellung von Volltexten zum Zweck der weiteren Erschließung von eigenen Beständen ist das Thema auch für den wissenschaftsunterstützenden Dienst einer Bibliothek relevant.[9] Bedarf für die Verwendung von Texterkennungsprogrammen besteht nicht nur in den Geisteswissenschaften, sondern – wie sich gezeigt hat – auch für konkrete Forschungsfragen aus anderen Disziplinen. Zum einen können mithilfe von automatischer Texterkennung große Textkorpora bearbeitet werden, zum anderen wird der Zugriff auf Originalquellen auch ohne paläographische Kenntnisse erleichtert. So werden Kurrent-, Sütterlin- oder Frakturschrift in vielen geisteswissenschaftlichen Studiengängen nur rudimentär behandelt, Naturwissenschaftler*innen fehlt die paläographische Grundausbildung oftmals gänzlich.

Die Anwendung der Texterkennungssoftware und das Lesen des Quellenmaterials stellen jedoch nicht die einzigen Hürden dar, sondern auch zahlreiche andere Fragestellungen müssen im Vorfeld geklärt werden: Welchen rechtlichen Beschränkungen unterliegen die Werke? Ist die Bereitstellung von durchsuchbaren Volltexten im Einzelfall kritisch zu bewerten? Nach welchen Richtlinien werden die Texte transkribiert und Trainingsmaterial erzeugt? Wie wird mit Fehlerraten (sog. Character Error Rate oder Word Error Rate) umgegangen? Und ist die Nachnutzung

---

[2] Weil und Kamlah, 2020; Projektübersicht OCR-BW, 2023.
[3] Tesseract OCR, 2023.
[4] READ-COOP, 2023.
[5] Scripta/escriptorium, 2023.
[6] OCR-D, 2023.
[7] Projekte der UB Mannheim, 2023.
[8] Gehrlein et. Al, 2020.
[9] Weil, 2018.

des Trainingsmaterials oder sogar der Modelle möglich und wie können diese bereitgestellt werden? Wenn ja, unter welchen Einschränkungen?

Das Angebot des Kompetenzzentrums umfasst ein breites Portfolio. Neben individueller Beratung und Unterstützung werden verschiedene Dokumentationen zu Texterkennungsprogrammen sowie auch Infrastruktur für Forschende z. B. in Form einer Instanz der Texterkennungs- und Transkriptionsplattform eScriptorium zur Verfügung gestellt.[10]
Seit November 2022 bietet das Kompetenzzentrum zudem das niedrigschwellige Angebot einer offenen Online-Sprechstunde an.[11] Hier können sich Interessierte aus allen Bereichen mit Fragen rund um das Thema automatische Texterkennung an das Team des Kompetenzzentrums wenden. Diese Sprechstunde wird ergänzt durch eine stetig aktualisierte FAQ-Sektion auf der Projekthomepage.[12]

Auf diesem Poster soll das Serviceangebot der Universitätsbibliotheken Mannheim und Tübingen im Bereich der automatischen Texterkennung für Forschende vorgestellt und Wissenschaftler*innen über die Einsatzmöglichkeiten von Texterkennungssoftware informiert werden.

Bibliografie

Gehrlein, Sabine, Jan Kamlah, Matthias Pintsch, Irene Schumm und Stefan Weil. 2020. “Vom Papier zur Datenanalyse. ‘Neue’ historische Forschungsdaten für die Wirtschaftswissenschaften.” In E-Science-Tage 2019: Data to Knowledge, herausgegeben von Vincent Heuveline, 598:140–52. Heidelberg: heiBOOKS. https://doi.org/10.11588/heibooks.598.c8423.

„Home - READ-COOP“. READ-COOP. Abgerufen am 27.04.2023. https://readcoop.eu/.

„OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen“. OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen. Abgerufen am 13.07.2023. https://ocr-bw.bib.uni-mannheim.de/.

„ OCR-D“. OCR-D. Abgerufen am 12.07.2023. https://ocr-d.de/.

[10] eScriptorium/Universitätsbibliothek Mannheim, 2023.
[11] Will, 2022.
[12] OCR-BW, 2023.

„Projekte der UB | Universität Mannheim“. Universitätsbibliothek | Universität Mannheim. Abgerufen am 12.07.2023. https://www.bib.uni-mannheim.de/ihre-ub/projekte-der-ub/.

„Projektübersicht | OCR-BW.“ OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen. https://ocr-bw.bib.uni-mannheim.de/projektuebersicht/

„Scripta / escriptorium - GitLab“. GitLab, 27.04.2023. https://gitlab.com/scripta/escriptorium/.

„GitHub - tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository)“. GitHub. Abgerufen am 12.07.2023. https://github.com/tesseract-ocr/tesseract.

Universitätsbibliothek Mannheim, „eScriptorium - Homepage“, OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen, abgerufen am 12.05.2023, https://ocr-bw.bib.uni-mannheim.de/escriptorium/.

Weil, Stefan. 2018. “126 Jahre Zeitung online - Fundgrube für historisch Interessierte und Motor für die Bibliotheks-IT: 126 years of the newspaper online.”, präsentiert bei 107. Deutscher Bibliothekartag, Berlin, Deutschland.

Weil, Stefan und Jan Kamlah. 2019. “Forschungsdaten aus Digitalisaten.” In E-Science-Tage 2019: Data to Knowledge, herausgegeben von Vincent Heuveline, 598:189. Heidelberg: heiBOOKS.

Weil, Stefan und Jan Kamlah. 2020. “OCR-BW – Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen: Texterkennung von historischen Drucken mit OCR-D und Tesseract.”, präsentiert bei Dokumentenerbe digital - Digitalisierung historischer Bestände baden-württembergischer Bibliotheken, Online.

Will, Larissa (2022, 28. Oktober). Projektende OCR-BW und 1. offene OCR-Sprechstunde | OCR-BW. OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen. https://ocr-bw.bib.uni-mannheim.de/2022/10/28/projektende-ocr-bw-und-1-offene-ocr-sprechstunde/