not opaque flow – Workflows zur Aufbereitung und Auswertung historischer Dokumente

Weber, Dominic
dominic.weber@unibe.ch
Digital Humanities, Walter Benjamin Kolleg, Universität Bern, Schweiz
ORCID: 0000-0002-9265-3388

Schwandt, Silke
silke.schwandt@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland
ORCID: 0000-0001-8303-4668

Huang, Angela
alhuang@fgho.eu
Forschungsstelle für die Geschichte der Hanse und des Ostseeraums, Europäisches Hansemuseum Lübeck, Deutschland
ORCID: 0000-0002-5321-9888

Hodel, Tobias
tobias.hodel@unibe.ch
Digital Humanities, Walter Benjamin Kolleg, Universität Bern, Schweiz
ORCID: 0000-0002-2071-6407

Tolino, Serena
serena.tolino@unibe.ch
Institut für Studien zum Nahen Osten und zu muslimischen Gesellschaften, Universität Bern, Schweiz
ORCID: 0000-0001-7740-5805

Kuhlmann, Christopher
christopher.kuhlmann@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Meyer, Dana
dameyer@techfak.uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Wilde, Melvin
melvin.wilde@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Kirschnick, Inga
inga.kirschnick@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Jentsch, Patrick
p.jentsch@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Hostettler, Myrjam
myrjam.hostettler@unibe.ch
Digital Humanities, Walter Benjamin Kolleg, Universität Bern, Schweiz
ORCID: 0000-0001-9316-6330

Widmer, Jonas
jonas.widmer@unibe.ch
Digital Humanities, Walter Benjamin Kolleg, Universität Bern, Schweiz

Lange, Inga
ilange@fgho.eu
Forschungsstelle für die Geschichte der Hanse und des Ostseeraums, Europäisches Hansemuseum Lübeck, Deutschland

Popken, Vivien
vpopken@fgho.eu
Forschungsstelle für die Geschichte der Hanse und des Ostseeraums, Europäisches Hansemuseum Lübeck, Deutschland

Ausgangslage

Jüngste Entwicklungen im Bereich des *Deep Learning* haben große Fortschritte in vielen Anwendungsbereichen mit sich gebracht. Automatisches Erkennen von (Hand-)Schriften, das Identifizieren von Entitäten, Events und Beziehungen sowie *Topic Modeling* sind dabei für die *Digital Humanities* von besonderem Interesse. Während für moderne Sprachformen und Sprachen des globalen Nordens bereits eine Vielzahl von Lösungen existieren, müssen sie für vormoderne Sprachen, Sprachstufen und Sprachen des globalen Südens, noch ausgiebig getestet und angepasst werden. Auch den ihnen zugrunde liegenden Algorithmen und Datensätze müssen erweitert und kritisch untersucht werden.

Das Projekt ‘The Flow – From Deep Learning to Digital Analysis and the Role in the Humanities’ nimmt sich diesen Chancen und Herausforderungen aus einer dezidiert historischen Perspektive anhand vier sehr unterschiedlicher Korpora aus verschiedenen Zeiten und Räumen an und versucht generalisierbare und bewegliche Lösungen zu

entwickeln. Ausgehend von der an der Universität Bielefeld lancierten Applikation *nopaque* wird der gesamte Weg vom digitalisierten Quellenkorpus zu den für die Forschung nutzbaren Daten in niederschwellig, wiederverwendbare Workflows übertragen (Jentsch und Porada 2022). Das Ziel ist es, die Vorteile maschineller Lernverfahren für Geisteswissenschaftler:innen zugänglicher zu machen und ihre Herausforderungen und Grenzen in den historisch arbeitenden Geisteswissenschaften hervorzuheben. Gleichzeitig sollen die Ansätze transparent und kritisch diskutiert und somit das methodologische Instrumentarium der Fachwissenschaften geschärft werden.

Alle bereits erwähnten Schritte von der Quelle zu den Daten sind mit einigem Anpassungsaufwand und den entsprechenden Kenntnissen auf historische Korpora anwendbar (siehe Bspw. für Latein und Französisch Torres Aguilar und Stutzmann, 2021; Cafiero u. a., 2021). Layoutanalyse und Handschriftenerkennung werden von Transkribus, eScriptorium und ähnlichen Anwendungen abgedeckt (Muehlberger u. a., 2019; Kiessling u. a., 2019). Für (*Named) Entity Recognition* und *Event Extraction* existieren Programmbibliotheken und bereits trainierte Modelle, auf denen aufgebaut werden kann (Akbik u. a. 2019; Vasiliev 2020; Brunner u. a. 2020).

Unser Ausgangspunkt bildet nopaque, eine Anwendung der Universität Bielefeld zur Prozessierung von Textkorpora. Derzeit kann die Webanwendung nopaque aus einer Reihe von Digitalisaten ein Korpus erstellen, mit Tesseract OCR (Metzger und Weil, 2019) oder Transkribus, basierend auf TrHTR und pyLaia (Puigcerver [2017] 2022) den Text erkennen sowie mit spacy tokenisieren, lemmatisieren, Part-of-Speech-Tagging und Named Entity Recognition durchführen. Dies alles wird ermöglicht durch eine Applikation mit intuitiver graphischer Benutzeroberfläche, die über einen Browser aufgerufen werden kann.

Im Rahmen des Projekts “The Flow” werden die genutzten Algorithmen erweitert und zusätzliche Formen der Aufbereitung in nopaque integriert. Aktuell sind in unserem Projekt in der Quellenerschließung und -analyse drei Aufbereitungsschritte vorgesehen, die alle jeweils unabhängig voneinander stehen und als *open source* Pakete publiziert werden. Über die nopaque-Oberfläche werden alle Teile verbunden und über ein Jobmanagement zu einem Workflow kombiniert.

Prozess 1: OCR/HTR

Die neuesten Entwicklungen in *Computer Vision* und Texterkennung mit Transformers versprechen auch für vormoderne und nicht-lateinische Texte höhere Präzision und bessere Wiederverwendbarkeit von Modellen (Ströbel u. a. 2022). Dieses Versprechen gilt es im Rahmen des “The Flow”-Projekts zu überprüfen und große Modelle zu integrieren bzw. die Integration über HuggingFace zu erleichtern.

In diesem Arbeitsschritt soll insbesondere die Publikation von Ground Truth Daten mit dem Training (im Sinne von *Fine-Tuning* großer Modelle) verknüpft werden. Der Schritt resultiert daher in neuen HTR-Modellen für individuelle Handschriften und der Generierung einer Datenbasis für spezifische Sprachformen.

Prozess 2: Entity, Event und Relation Extraction

Automatisiertes Identifizieren von Entitäten, Ereignissen und Relationen sind klassische Aufgaben des *Natural Language Processing*. Für diese Aufgaben, wie auch für das Part-of-speech-Tagging sind stabile Language Models, die entweder enorm groß sind (im Sinne von Large Language Models) oder die spezifisch für eine bestimmte Domäne oder Zeit entwickelt wurden, erforderlich sind. Wiederum geht es um die Generierung großer Modelle und die Nachnutzung bestehender Modelle, die danach durch *Tagger* oder anderer Sequenz-zu-Sequenz Annotationsformen ausgezeichnet werden.

Für vormoderne Sprachen haben sich, wie für moderne Sprachstufen, insbesondere große oder domänenspezifische Sprachmodelle als hilfreich herausgestellt (Hodel, Prada Ziegler, und Schneider, 2023), sodass keine Normalisierungen mehr notwendig sind. Dies bedeutet gleichzeitig, dass Grundlagen (Sprachmodelle) erarbeitet und best-practices (beispielsweise für Annotationen) definiert werden müssen.

Prozess 3: Topic Modeling, Clustering und Vergleichsstudien

In diesem Schritt sollen auch die in den vorherigen Schritten trainierten Modelle für das thematische Clustern von Dokumenten innerhalb der jeweiligen Korpora verwendet werden. Auf Basis der Sprachmodelle können Abschnitte oder ganze Dokumente vektorisiert und mittels Clusteringverfahren miteinander verglichen werden. Die Resultate daraus können mit bereits intensiv genutzten Verfahren der Themenextraktion, insbesondere klassische Topic Modeling Verfahren mit LDA, abgeglichen werden (Graham, Weingart, und Milligan, 2012; Hodel, Möbus, und Serif, 2022; Schöch, 2017), was Vergleiche und die Kombination der Ergebnisse ermöglicht.

Dadurch wird ein Fundament gelegt, um die Quellenkritik, *close-reading* sowie Methoden und Standards der Geschichtswissenschaft wieder einzubringen und die methodologische Brücke zwischen *Digital* und *Humanities* zu schlagen.

Workflow Prozesse: Implementierung in *nopaque*

Aus den oben eingeführten Prozessen der Datenaufbereitung und -modellierung lassen sich wiederverwendbare Workflows entwickeln. Diese werden modular in *nopaque* implementiert, sodass sie auch von Forschenden eingesetzt werden können, die vorwiegend an der Anwendung interessiert sind.

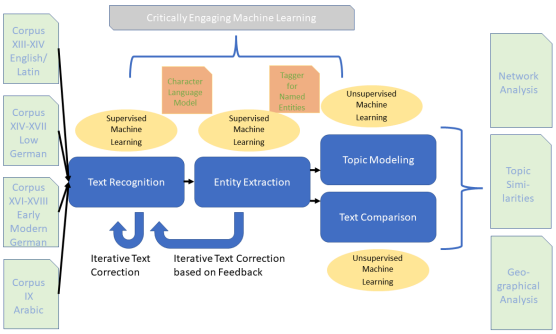

Abbildung 1: Schema der Workflows in einzelne Schritte zerlegt.

Die einzelnen Stationen dieser Workflows sind als Mikroprozesse zu verstehen, die jeweils eine bestimmte (nicht unbedingt kleine) Aufgabe erfüllen. Dazu gehört neben den oben eingeführten Prozessen beispielsweise eine Indizierung der Texte. Diese Technologie ermöglicht schnelleren und zuverlässigen Zugriff auf die Texte, wodurch das Trainieren und Anwenden von Machine-Learning-Modellen effizienter gestaltet werden kann.

Die Abläufe und die Interaktion zwischen den einzelnen Services werden über APIs (Programmierschnittstellen) abgewickelt, welche von einem von außen zugänglichen API-Gateway zusammengehalten und koordiniert werden. So werden unter anderem domänenspezifische Anpassungen und die Nachnutzbarkeit des Outputs in beliebige Analyse- und Auswertungsumgebungen vereinfacht. Das Zusammenspiel von Microservices, internen und externen APIs stellt die Skalierbarkeit sicher und erhöht die Flexibilität der Workflows.

Außerdem können so nach Bedarf zusätzliche Microservices zwischengeschaltet werden. Die intuitive graphische Benutzeroberfläche von nopaque muss dadurch nicht die einzige mögliche Anwendungsform bleiben. Die einzelnen Module sind zwar für die Anwendung in nopaque intendiert und optimiert, sind davon aber grundsätzlich unabhängig und können einzeln angesteuert werden.

Die modulare Struktur schafft großes Entwicklungspotenzial und bereitet *nopaque* auf längerfristige Weiterentwicklung und Anpassung an künftige technische Innovationen vor. Gleichzeitig kann auch das Projekt ‘The Flow’ derzeit noch offene Fragen in Zukunft einfacher in die Workflows einbinden. Dies betrifft beispielsweise die Anbindung an HPC-Clusters (High Performance Computing), was vor allem für aufwendige Prozesse wie das Training von HTR- und Sprachmodellen interessant ist.

Ziel des Workshops

Der Workshop verfolgt zwei Ziele: Erstens wird der aktuelle Stand von *nopaque* und seine Module der Community vorgestellt. Zweitens wird in einem geführten Brainstorming überlegt, welche Ansätze und Methoden in den Prozessen mitgedacht werden können. Das heißt, die Teilnehmenden haben die Möglichkeit sich in die weitere Entwicklung einzubringen und eigene Erfahrungen zu tauschen. Darum wird für den Workshop ein Call for Participation veröffentlicht (siehe Call anbei).

Ziel des Brainstorming-Teils ist es, die Diversität der historisch arbeitenden digitalen Geisteswissenschaften ernst zu nehmen und gemeinsam mit Critical Friends nach Ansätzen und Auswertungsformen zu suchen, die im Projektteam nicht abgedeckt sind und blinde Flecken darstellen. Aus diesem Grund sind Teilnehmende eingeladen, ihre Arbeitsweisen mit Auswertungs- und Analysetools zu demonstrieren.

Auf dieser Grundlage sollen in einem abschließenden Teil, die jeweiligen Implikationen für wiederverwendbare Workflows und mögliche Implementierungen und Integrationen in nopaque diskutiert werden. Ein besonderer Fokus soll dabei auf aufgrund ungewöhnlicher Layouts, Schriften oder Sprachen auftretender Bedürfnisse an Tools – und wo diese nicht bedient werden – liegen.

Format des Workshops

Der Workshop wird in vier Teilen plus Kaffeepause (3h30min) durchgeführt

00:00-00:30 Einführung und Vorstellungsrunde
00:30-01:30 Einführung in *nopaque*
01:30-01:50 Kaffeepause
01:50-02:20 Pitches der eingereichten Algorithmen und Methoden (aus CFP)
02:20-03:00 Brainstorming-Session
03:00-03:30 Schlussdiskussion

Zielpublikum und notwendiges Vorwissen:

Der Workshop richtet sich an Personen, die mit historischen Dokumenten arbeiten oder aus technischer Warte Prozessierung von Dokumenten anstreben oder verantworten. Vorwissen ist keines notwendig und die Einreichung eines eigenen Beitrags wird auch nicht vorausgesetzt.

Bibliographie

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, und Roland Vollgraf. 2019. „FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP“. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4010.

Brunner, Annelen, Ngoc Duyen Tanja Tu, Lukas Weimer, und Fotis Jannidis. 2020. „To BERT or Not to BERT – Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of Four Types of Speech, Thought and Writing Representation“. In . CEUR-WS. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/11561.

Cafiero, Florian, Thibault Clérice, Paul Fièvre, Simon Gabay, und Jean-Baptiste Camps. 2021. „Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre“. *Journal of Data Mining & Digital Humanities* 2021 (Februar). https://doi.org/10.46298/jdmdh.6485.

Graham, Shawn, Scott Weingart, und Ian Milligan. 2012. „Getting Started with Topic Modeling and MALLET“. *Programming Historian*. https://doi.org/10.46430/phen0017.

Hodel, Tobias, Dennis Möbus, und Ina Serif. 2022. „Von Inferenzen und Differenzen. Ein Vergleich von Topic-Modeling-Engines auf Grundlage historischer Korpora“. In *Von Menschen und Maschinen: Mensch-Maschine-Interaktionen in digitalen Kulturen*, herausgegeben von Selin Gerlek, Sarah Kissler, Thorben Mämecke, Dennis Möbus, Jennifer Eickelmann, Katrin Köppert, Peter Risthaus, und Florian Sprenger, 1. Auflage, 1:181–205. Digitale Kultur. Hagen: Hagen University Press. https://doi.org/10.57813/20220623-153139-0.

Hodel, Tobias, Ismail Prada Ziegler, und Christa Schneider. 2023. „Pre-Modern Data: Applying Language Modeling and Named Entity Recognition on Criminal Records in the City of Bern“. Graz, Austria, Juni 30. https://doi.org/10.5281/zenodo.8107616.

Jentsch, Patrick, und Stefan Porada. 2022. „nopaque“. nopaque | from text > to data > to analysis. 2022. https://nopaque.sfb1288.uni-bielefeld.de/.

Kiessling, Benjamin, Robin Tissot, Peter Stokes, und Daniel Stökl Ben Ezra. 2019. „eScriptorium: An Open Source Platform for Historical Document Analysis“. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2:19–19. https://doi.org/10.1109/ICDARW.2019.10032.

Metzger, Noah, und Stefan Weil. 2019. „Optimierter Einsatz von OCR-Verfahren – Tesseract als Komponente im OCR-D-Workflow“. Workshop gehalten auf der MAD HD, Heidelberg, Juli 30.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, u. a. 2019. „Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study“. *Journal of Documentation* 75 (5): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

Puigcerver, Joan. (2017) 2022. „PyLaia“. Python. https://github.com/jpuigcerver/PyLaia.

Schöch, Christof. 2017. „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama“. *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Ströbel, Phillip Benjamin, Simon Clematide, Martin Volk, und Tobias Hodel. 2022. „Transformer-based HTR for Historical Documents“. Preprint. ArXiv: https://doi.org/10.48550/arXiv.2203.11008.

Torres Aguilar, Sergio, und Dominique Stutzmann. 2021. „Named Entity Recognition for French medieval charters“. In *Workshop on Natural Language Processing for Digital Humanities*. Workshop on Natural Language Processing for Digital Humanities Proceedings of the Workshop. Helsinki, Finland. https://hal.archives-ouvertes.fr/hal-03503055.

Vasiliev, Yuli. 2020. *Natural language processing with Python and spaCy: a practical introduction*. San Francisco: No Starch Press.