Toolbeitrag: GROBID

Dominik Gerstorfer [1]

1. Technische Universität Darmstadt

Thema:	Bibliografie	DOI:	10.48694/fortext.3788
Jahrgang:	1	Ausgabe:	11
Erscheinungsdatum:	30-11-2024	Erstveröffentlichung:	2021-03-08 auf fortext.net
Lizenz:			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.

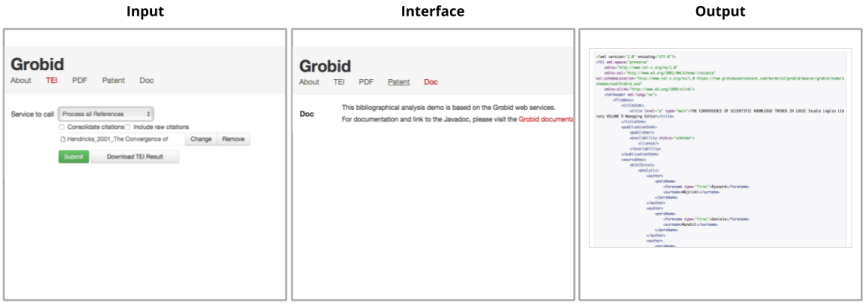

Abb. 1: Der GROBID-Workflow: Im Webinterface werden Dateien ausgewählt, die strukturierten Daten können als TEI heruntergeladen werden.

- **Systemanforderungen:** Läuft auf Linux und Mac, benötigt JDK 8.
- **Stand der Entwicklung:** Wird seit 2008 entwickelt, letztes Release August 2020.
- **Herausgeber:** Patrice Lopez
- **Lizenz:** 0.6.1
- **Weblink:** https://github.com/kermitt2/grobid/releases/
- **Im- und Export:** GROBID liest TXT- und PDF-Dateien und extrahiert strukturierte Daten als XML-TEI oder BibTeX.
- **Sprachen:** Keine Angabe

1. Für welche Fragestellungen kann GROBID eingesetzt werden?

GROBID liest TXT (vgl. Reintext-Version)- und PDF-Dateien und extrahiert strukturierte Daten als XML-TEI (vgl. TEI) oder BibTeX. Das Tool dient primär der Vorverarbeitung (vgl. Preprocessing) von Texten, etwa im Prozess der Korpusbildung (Bläß 2024).

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

Funktion: GROBID kann eingesetzt werden um bibliographische Informationen aus Texten zu extrahieren, dabei kann zwischen den Informationen des Textes oder der darin enthaltenen Bibliographie gewählt werden. Des Weiteren kann der Volltext einer PDF- als strukturierte TEI-Datei ausgelesen werden.

Folgende Funktionen sind Verfügbar:

- Header-Informationen eines Artikels (Titel, Autoren, Abstract, Keywords, etc.) extrahieren und parsen.
- Bibliographische Daten extrahieren.
- Zitate im Text erkennen und mit der Bibliographie verknüpfen
- Einzelne bibliographische Angaben parsen.
- Adressen und Institutszugehörigkeiten parsen.
- Volltext einer PDF strukturieren und als TEI ausgeben.

GROBID bietet auch die Möglichkeit eigene Modelle zu trainieren und Module zu schreiben, so dass fortgeschrittene Nutzer*innen das Tool sehr flexibel anpassen können.

Zuverlässigkeit: GROBID setzt ML-Modelle (vgl. Machine Learning) ein, die F-Scores (vgl. F-score) zwischen 0,76 und 0,89 erreichen, abhängig von der Qualität der Quelltexte und der eingesetzten Funktion. In der Regel müssen die extrahierten Daten noch manuell nachbearbeitet werden.

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

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

Eine direkte methodische Nähe zu den traditionellen Literaturwissenschaften ist nicht gegeben, da mit GROBID selbst keine Analysen möglich sind. GROBID ist vielmehr ein Hilfsprogramm, mit dem mühsame und arbeitsintensive Aufgaben automatisiert und erleichtert werden können.
Ein leichter Einstieg ist über das Webinterface möglich, welches für einfache Anwendungsfälle eine graphische Benutzeroberfläche bereitstellt. Der volle Funktionsumfang ist jedoch erst über die API zugänglich, hierfür ist es nötig auf der Kommandozeile (vgl. Commandline) mit cURL oder den GROBID-Client-Programmen entsprechende Anfragen zu stellen. Welche Optionen zur Verfügung stehen und wie auch größere Datenbestände automatisiert bearbeitet werden können, ist in der umfangreichen Dokumentation mit Beispielen beschrieben.

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

GROBID ist in den Naturwissenschaften und den Digitalen Geisteswissenschaften etabliert und wird bereits stabil in privaten und öffentlichen Projekten eingesetzt, u.a. von ResearchGate, dem Internet Archive und dem CERN (Invenio).

5. Unterstützt GROBID kollaboratives Arbeiten?

Nein, mit GROBID kann nicht kollaborativ gearbeitet werden.

6. Sind meine Daten bei GROBID sicher?

Ja, GROBID läuft als Server auf dem eigenen Rechner, alle Daten werden lokal verarbeitet.

Externe und weiterführende Links

- CERN (Invenio): https://web.archive.org/web/20241106112023/https://invenio-software.org/ (Letzter Zugriff: 06.11.2024)
- cURL: https://web.archive.org/web/20241106112021/https://de.wikipedia.org/wiki/CURL (Letzter Zugriff: 06.11.2024)
- Grobid Dokumentation: https://web.archive.org/web/20241106112305/https://grobid.readthedocs.io/en/latest/ (Letzter Zugriff: 06.11.2024)
- Grobid auf GitHub: https://web.archive.org/web/20241106112023/https://github.com/kermitt2/grobid/releases/ (Letzter Zugriff: 06.11.2024)
- Grobid Web Application: https://web.archive.org/web/20241106112452/https://kermitt2-grobid.hf.space/ (Letzter Zugriff: 06.11.2024)
- Internet Archive: https://archive.org (Letzter Zugriff: 06.11.2024)
- ResearchGate: https://www.researchgate.net/ (Letzter Zugriff: 06.11.2024)

Bibliographie

Bläß, Sandra. 2024. Methodenbeitrag: Korpusbildung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 2. Korpusbildung (12. Juni). doi: 10.48694/fortext.3708, https://fortext.net/routinen/methoden/korpusbildung.

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.

API API steht für *Application Programming Interface* und bezeichnet eine Programmierschnittstelle, die Soft- und Hardwarekomponenten wie Anwendungen, Festplatten oder Benutzeroberflächen verbindet. Sie vereinheitlicht die Datenübergabe zwischen Programmteilen, etwa Modulen, und Programmen.

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.

F-score Der F-Score steht für ein statistisches Maß, welches das Verhältnis von Genauigkeit (*Precision*) und Trefferquote (*Recall*) als gewichtetes harmonisches Mittel angibt, und deshalb als gerichtetes, harmonisches Mittel gilt.

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.

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.