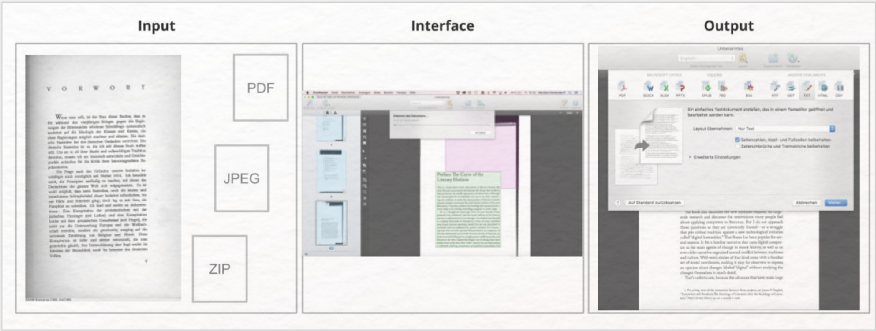
Toolbeitrag: Abbyy FineReader

Mareike Schumacher [1]

1. Universität Regensburg

Thema:	Textdigitalisierung und Edition	DOI:	10.48694/fortext.3742
Jahrgang:	1	Ausgabe:	3
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2019-07-15 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 des Abbyy FineReaders: Import einer eingescannten Datei hochladen, Analyse sowie Umwandlung durch das Tool und Export der computerlesbaren Datei

- **Systemanforderungen:** Abbyy FineReader ist für Linux, Mac OS und Windows erhältlich, die Versionen unterscheiden sich je nach Betriebssystem
- **Stand der Entwicklung:** Abbyy FineReader läuft in der 14. Version
- **Herausgeber:** Abbyy Europe
- **Lizenz:** kostenpflichtig
- **Weblink:** https://pdf.abbyy.com/de/
- **Im- und Export:** Importformate sind DOC-Formate, CSV, TIFF, JPEG, ZIP und PDF, Exportformate sind PDF, DOC-Formate, TXT (vgl. Reintext-Version), CSV, ODT, PPT, HTML, RTF, EPUB, FB2, DjVu
- **Sprachen:** Erkennungssprachen knapp 200, Sprachen der Benutzeroberfläche: 30

1. Für welche Fragestellungen kann der Abbyy FineReader eingesetzt werden?

Der Abbyy FineReader ist ein Tool zur Optical Character Recognition (OCR), die eingescannte Texte computerlesbar macht (vgl. Möglichkeiten der Textdigitalisierung (Horstmann 2024a)). Das Tool unterstützt v.a. die Vorbereitung von Forschungsprojekten, deren Fokus auf Werken liegt, die noch nicht gemeinfrei (vgl. Open Access) sind und die deshalb zunächst in ein Format umgewandelt werden müssen, mit dem Sie digital arbeiten können. Eine mögliche Fragestellung wäre: Wie wird die Stadt in Alfred Döblins *Berlin Alexanderplatz* dargestellt?

2. Welche Funktionalitäten bietet der Abbyy FineReader und wie zuverlässig ist das Tool?

Funktionen:

- Optical Character Recognition (OCR)
- Vorschläge für zu überprüfende Textstellen (Kontrollfunktion)
- Export von FineReader-Dokumenten als bearbeitbare Textdokumente (z. B.: TXT, DOCX)

Zuverlässigkeit: Der Abbyy FineReader läuft sehr zuverlässig und hochperformant. Die automatische Erkennung von Text in Scans ist gut, aber nicht fehlerfrei. Die Vorschlagsfunktion zur Überprüfung von Textstellen, die das Tool nicht eindeutig erkennen konnte, ist äußerst hilfreich. Mit Hilfe dieser Funktion können Sie die erkannten Texte zügig überprüfen und ggf. manuell korrigieren. Von einer Nutzung ohne anschließende Überprüfung wird

abgeraten, da die Fehlerquote dafür zu hoch ist. Leider ist die Funktionalität zur halbautomatischen Überprüfung der OCR-Ergebnisse nicht für die MacOS-Version implementiert. Nutzen Sie den Abbyy FineReader daher am besten mit einem Windows-Betriebssystem.

3. Ist der Abbyy FineReader für DH-Einsteiger*innen geeignet?

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

Die Digitalisierung von Textdokumenten, die der Abbyy FineReader erleichtert, ist nicht Teil des eigentlichen geisteswissenschaftlichen Forschungsprozesses. Die Aufbereitung von Texten mit Hilfe der automatischen Texterkennung stellt allerdings in vielen Fällen einen essentiellen Teil der literaturwissenschaftlichen Arbeit dar: Um überhaupt (digital) mit Texten arbeiten zu können, müssen diese in einem geeigneten Format vorliegen. In der Regel stellt die Digitalisierung von Textdokumenten also auch für weniger technikaffine Forscher*innen einen wichtigen Teil ihres Arbeitsprozesses dar. Der Abbyy FineReader bietet umfangreiche Tutorials für den Einstieg in die Nutzung und einen Support, der nach Versionen gestaffelt ist. Insbesondere für die aktuelle Version gibt es einen sehr umfangreichen technischen Support, die Vorgänger-Version verfügt über einen leicht eingeschränkten Support. Für Versionen, die noch älter sind, gibt es keinen Support mehr.

Die Nutzung des Abbyy FineReaders ist kostenpflichtig. Allerdings können einige der Funktionen in limitiertem Umfang kostenfrei im Handschriften-Digitalisierungs-Tool Transkribus (Horstmann 2024b) genutzt werden (vgl. Digitale Manuskriptanalyse (Horstmann 2024c)).

4. Wie etabliert ist der Abbyy FineReader in den (Literatur-)Wissenschaften?

Der Abbyy FineReader findet vor allem in computerlinguistischen und informationstechnologischen Publikationen Erwähnung und kann somit als gut etabliert eingestuft werden. Da die Digitalisierung von Texten für geisteswissenschaftliche Forschungsprozesse eher eine Vorbereitung der Forschungsarbeit als Teil des Analyseprozesses oder gar -gegenstandes ist, findet er hier wenig Erwähnung.

5. Unterstützt der Abbyy FineReader kollaboratives Arbeiten?

Nein, der Abbyy FineReader ist für die Einzelnutzung optimiert.

6. Sind meine Daten beim Abbyy FineReader sicher?

Ja, da der Abbyy FineReader auf Ihrem eigenen Computer installiert und ohne eine Verbindung zum Internet verwendet wird, ist die Nutzung aus datenschutzrechtlicher Perspektive unbedenklich.

Externe und weiterführende Links

- Abby FineReader: https://web.archive.org/save/https://pdf.abbyy.com/de/ (Letzter Zugriff: 04.06.2024)

Bibliographie

Horstmann, Jan. 2024c. Methodenbeitrag: Digitale Manuskriptanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3744, https://fortext.net/routinen/methoden/digitale-manuskriptanalyse.

———. 2024a. 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.

———. 2024b. Toolbeitrag: Transkribus. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3746, https://fortext.net/tools/tools/transkribus.

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.

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.

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 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.

Open Access Open Access bezeichnet den freien Zugang zu wissenschaftlicher Literatur und anderen Materialien im Internet.

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.

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.

ZIP ZIP steht für ein Dateiformat (zip = engl. Reißverschluss), in welchem mehrere Einzeldateien verlustfrei, komprimiert zusammengefasst werden. ZIP-Dateien werden beim Öffnen entweder automatisch entpackt oder lassen sich per Rechtsklick extrahieren.