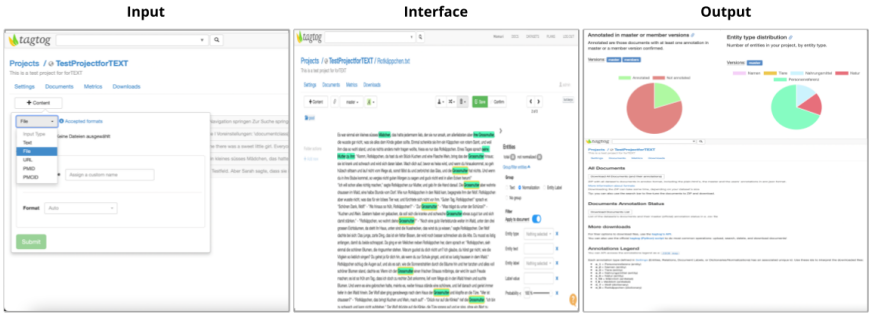
Toolbeitrag: Tagtog

Mareike Schumacher [2]
Mari Akazawa [1]
1. Technische Universität Darmstadt
2. Universität Regensburg

Thema:	Manuelle Annotation	DOI:	10.48694/fortext.3763
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2022-01-10 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.



Workflow: Im Webinterface werden Daten durch das Hochladen von Textdateien in verschiedenen Formaten, der direkten Eingabe eines Textes oder durch die Angabe einer URL importiert. Diese können im nächsten Schritt umfangreich mit neu erstellten oder hochgeladenen Tagsets, die auch Dictionaries beinhalten können, annotiert werden. Im Anschluss können die annotierten Versionen in der kostenfreien Version in Form einer ZIP-Datei im JSON-Format oder als TSV-Datei exportiert werden.

- **Systemanforderungen:** tagtog kostenfrei als cloudbasiertes (vgl. Cloudbasiert) Tool über einen Browser (z. B. Chrome, Firefox, Safari) oder über eine API genutzt werden. Kostenpflichtig kann tagtog auch lokal auf dem eigenen Server laufen. Die lokale Nutzung des Tools erfordert außerdem: Docker, Docker Compose, cURL und Commandline-Kenntnisse
- **Stand der Entwicklung**: Version 3.2021-W47.3 (Stand Dezember 2021); seit 2017 stetig weiterentwickelt
- **Herausgeber**: Dr. Juan Miguel Cejuela, Jorge Campos und weitere Entwickler*innen
- **Lizenz**: Creative Commons: Attribution 4.0 International (CC BY 4.0) für öffentliche Projekte
- **Weblink**: https://www.tagtog.com
- **Im- und Export**: Import von Formaten wie TXT (vgl. Reintext-Version), HTML, Bio XML-Format, Markdown; Import von CSV, TSV und PDF nur in kostenpflichtiger Version möglich; Export im JSON-Format und als TSV
- **Sprachen**: Sprachunabhängig (unterstützt Unicode)

1. Für welche Fragestellungen kann tagtog eingesetzt werden?

tagtog ist ein englischsprachiges Tool zur Annotation von Textdaten, das die Möglichkeit bietet, auf Grundlage manueller Annotationen, ein projektspezifisches Machine Learning durchzuführen, einen bereits vorhandenen ML-Algorithmus ins Projekt einzubinden, oder den tooleigenen ML-Classifier zur automatisierten Annotation zu nutzen. Neben der Annotation von Dokumenttentypen oder Entitäten mit eigenen Tagsets (vgl. Tagset), berechnet das Tool beispielsweise automatisch den Annotationsfortschritt bei kollaborativen Arbeiten oder die quantitative Verteilung der genutzten Tags und erstellt daraufhin Visualisierungen, die ebenfalls zum Download bereitstehen.
So bietet sich tagtog, durch die Erstellung bzw. das Hochladen von Tagsets zu Named Entities (vgl. Named Entities) oder Dokumenttypen und damit verknüpften Dictionaries, besonders dafür an, große Textmengen

automatisch oder halbautomatisch zu annotieren, und kann somit für eine große Vielfalt an Forschungsansätzen genutzt werden. Eine mögliche Fragestellung wäre: „Wie ist das Verhältnis von Sprecher- zu Sprecherinnen-Text in deutschsprachigen Dramen des 18. - 20. Jahrhunderts?“ oder „Welche realweltlichen Orte werden in Erzähltexten einer bestimmten Epoche erwähnt?“.

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

Funktionen:

- Erstellen oder Hochladen eigener Tagsets zur Annotation von NE oder Dokumenttypen
- Manuelle und automatische (nicht kostenfrei) Annotation von ganzen Paragraphen und Tabellen etc.
- Überlappende Annotationen
- Verknüpfung von NE durch Relationen oder Dictionaries
- Normalisierung von Tags
- Trainieren oder Hochladen eigener ML-Algorithmen
- Nutzung des tagtog-Machine-Learning-Algorithmus' (nicht kostenfrei)
- Kollaboratives Arbeiten mit automatischer Aufgabenverteilung
- Berechnung und Visualisierung von Statistiken zu annotierten Daten und zum Annotationsfortschritt eines Projektes
- Berechnung der Confidence Probability für alle Annotationen und Berechnung des IAA
- Query-Abfragen im Projekt zur Suche nach Dokumenten, Annotationsfortschritten oder bestimmten Tags
- Nutzung über eine API möglich

_Zuverlässigkeit: _tagtog wird kontinuierlich gepflegt und läuft zuverlässig.

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

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

tagtog bietet, im Vergleich zu anderen Tools, eine große Menge an Funktionalitäten und somit auch viele verschiedene Einsatzmöglichkeiten. Das Benutzerinterface in der Webversion ist übersichtlich, intuitiv aufgebaut und somit auch für DH-Einsteiger*innen geeignet. Die interaktive Benutzeroberfläche ist in vier Bereiche (Einstellungen, Projektübersicht & Annotationsbereich, Übersicht aller Statistiken, Downloadbereich) aufgeteilt. Um einen umfassenden Überblick aller Funktionen zu erlangen und die Vorteile der Funktionen zum kollaborativen Arbeiten in Gänze ausschöpfen zu können, ist es allerdings ratsam, sich zuvor intensiv mit der Dokumentation des Tools zu beschäftigen. Diese steht wie das Tool nur auf Englisch zur Verfügung.
Die Nutzung des Tools auf dem eigenen Server ist für DH-Einsteiger*innen aufgrund der aufwändigen Installation des Tools nicht zu empfehlen.

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

tagtog wurde ursprünglich als Text-Mining (vgl. Data Mining)-Tool für den Bereich der Biomedizin entwickelt (Cejuela u. a. 2014) und wird inzwischen in vielen weiteren wissenschaftlichen Disziplinen und auch im Finanz-, Gesundheits- und Rechtswesen eingesetzt (Goldberg u. a. 2015).
In den Literaturwissenschaften ist es bislang noch nicht sehr etabliert, wurde aber beispielsweise schon zum Trainieren von NLP-Modellen für Analysen historischer, lateinamerikanische Dokumente eingesetzt (Murrieta-Flores u. a. 2019). Die manuelle Annotation bietet die Möglichkeit traditionell-analoge Forschungsmethodik ins Digitale zu übertragen.

5. Unterstützt tagtog kollaboratives Arbeiten?

Ja, tagtog unterstützt kollaboratives Arbeiten. Das Tool ist darauf ausgelegt, kollaboratives Arbeiten zu erleichtern. Allen Teilnehmer*innen eines Projektes können verschiedene Rollen mit verschiedenen Berechtigungen zugewiesen werden. In einer Kopie vom Original, arbeiten die einzelnen Annotierenden an separaten Dokumenten, welche abschließend in einem Goldstandard zu einer Version zusammengesetzt werden können. Außerdem bietet tagtog die Möglichkeit, die zu annotierenden Dokumente zufällig auf die Annotierenden aufzuteilen. Durch die automatische Berechnung von Annotiationsfortschritten und der In-Text-Markierung von Annotationen nach Annotator*in, können die individuellen Annotationen besonders einfach miteinander abgeglichen werden. Außerdem berechnet tagtog bei kollaborativen Projekten automatisch das IAA und die Confidence Probability für jedes Dokument und jedes Projekt.

6. Sind meine Daten bei tagtog sicher?

Ja und Nein. In den kostenpflichtigen lokalen Versionen werden alle Daten auf dem eigenen Server/Rechner gespeichert. In den kostenpflichtigen Cloudversionen können die Projekte „privat“ gehalten werden. In der kostenfreien Cloudversion hingegen sind alle Projekte einschließlich aller Annotationen für andere Nutzer*innen frei zugänglich.

Personenbezogene Daten:
Zur Registrierung ist lediglich eine gültige Mailadresse nötig. Diese wird vertraulich behandelt. Verhaltensdaten werder von Besucher*innen sowie Nutzer*innen der Webseite gesammelt. Diese werden zur Interaktion mit Drittparteien verwendet. Weitere Informationen: https://docs.tagtog.com/projects.html#privacy

Urheberrechtlich geschützte Daten:
In der kostenfreien Cloudversion werden die Texte und Annotationen in die Cloud geladen und dort gespeichert. Somit sind die Daten von dort aus auch für andere Nutzer*innen einsehbar. Allerdings können die Textdaten nur in einem geschützten Login-Bereich verwaltet werden.

Externe und weiterführende Links

- tagtog: https://web.archive.org/save/https://www.tagtog.com (Letzter Zugriff: 03.07.2024)
- tagtog Dokumentation: https://web.archive.org/save/https://docs.tagtog.com (Letzter Zugriff: 03.07.2024)
- tagtog Datensicherheit: https://docs.tagtog.com/projects.html#privacy (Letzter Zugriff: 03.07.2024)

Bibliographie

Cejuela, Juan Miguel, Peter McQuilton, Laura Ponting, Steven J. Marygold, Raymund Stefancsik, Gillian H. Millburn, Burkhard Rost und FlyeBase Consortium. 2014. tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database* 2014, Nr. bau033. doi: https://doi.org/10.1093/database/bau033,.

Goldberg, Tatyana, Shrikant Vinchurkar, Juan Miguel Cejuela, Lars Juhl Jensen und Burkhard Rost. 2015. Linked annotations: a middle ground for manual curation of biomedical databases and text corpora. In: *BMC Proceedings 9*. Kashiwa, Japan. doi: 10.1186/1753-6561-9-S5-A4,.

Murrieta-Flores, Patricia, Raquel Liceras-Garrido, Katherine Bellamy, Mariana Favila-Vazquez, Jorge Campos, Juan Miguel Cejuela und Bruno Martins. 2019. Training NLP models for the analysis of 16th century Latin American historical documents:Tagtog and the Geographic Reports of New Spain. *Digital Humanities 2019*. doi: 10.6084/m9.figshare.11806185.v1,.

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.

Cloudbasiert Werden Ihnen Dienste, Speicherplatz oder Rechenleistung „cloudbasiert" angeboten, handelt es sich um die Bereitstellung dieser Ressource über das Internet. Eine Software, die nicht auf dem eigenen Server installiert ist, sondern auf den Servern des Herstellers, nennt man gehostete Software. Nutzt der/die Hersteller*in für die Bereitstellung selbst eine Cloud, so ist von cloudbasierter Software die Rede.

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.

Data Mining Data Mining gehört zum Fachbereich Information Retrieval und bezieht sich auf die systematische Anwendung computergestützter Methoden, die darauf abzielt, in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu erkennen. Textbasierte Formen des Data Minings sind u. a. Text Mining, Web Mining und Opinion Mining.

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.

Information Retrieval Die Teildisziplin der Informatik, das Information Retrieval, beschäftigt sich mit der computergestützten Suche und Erschließung komplexer Informationen in meist unstrukturierten Datensammlungen.

JSON JSON ist die englische Abkürzung für *JavaScript Object Notation* . Dabei handelt es sich um ein kompaktes Textformat, das insbesondere zum Datenaustausch entworfen wurde. Es ist für Menschen einfach zu lesen und zu schreiben und für Maschinen einfach zu analysieren und zu generieren. JSON ist ein Format, das unabhängig von Programmiersprachen ist.

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.

Opinion Mininig Unter Opinion Mining, oder Sentiment Analysis, versteht man die Analyse von Stimmungen oder Haltungen gegenüber einem Thema, durch die Analyse natürlicher Sprache. Das Opinion Mining gehört zu den Verfahren des Text Minings.

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.

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

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.

Tagset Ein Tagset definiert die Taxonomie, anhand derer Annotationen in einem Projekt erstellt werden. Ein Tagset beinhaltet immer mehrere Tags und ggf. auch Subtags. Ähnlich der Type/Token -Differenz in der Linguistik sind Tags deskriptive Kategorien, wohingegen Annotationen die einzelnen Vorkommnisse dieser Kategorien im Text sind.

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).

Text Mining Das Text Mining ist eine textbasierte Form des Data Minings. Prozesse & Methoden, computergestützt und automatisch Informationen bzw. Wissen aus unstrukturierten Textdaten zu extrahieren, werden als Text Mining zusammengefasst.

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.

Web Mining Unter Web Mining versteht man die Anwendung von Techniken des Data Mining zur Extraktion von Informationen aus dem World Wide Web. Das Web Mining ist ein Teilbereich des Data Minings und zählt zu einem der wichtigsten Anwendungsgebiete für das Text Mining.

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.