Toolbeitrag: WebLicht

Mareike Schumacher [1]

1. Universität Regensburg

Thema:	Korpusbildung	DOI:	10.48694/fortext.3811
Jahrgang:	1	Ausgabe:	2
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2019-08-05 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 von WebLicht: Text hochladen, Beispieltext wählen oder Text per Copy/Paste einfügen, dann Tool-Pipeline auswählen (Easy Mode) oder Tools selbst zusammen stellen (Advanced Mode). Die Ergebnisse können als XML-Dateien einzeln oder alle zusammen heruntergeladen werden

- **Systemanforderungen:** Webbasiertes (vgl. Webanwendung) Tool, über den Browser (z. B. Chrome, Firefox, Safari) nutzbar
- **Stand der Entwicklung:** WebLicht wird seit 2008 im Rahmen des CLARIN-D-Projektes stetig weiterentwickelt
- **Herausgeber:** Universität Tübingen
- **Lizenz:** Kostenfrei zugänglich
- **Weblink:** https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page
- **Im- und Export:** Importformate: Text (per Copy/Paste oder Direkteingabe), CONLL-U, Microsoft-Word-Formate, PDF, RTF (vgl. Reintext-Version), DTA-TEI, ISO/TEI for spoken language; Downloadformat: XML
- **Sprachen:** Deutsch, Englisch

1. Für welche Fragestellungen kann WebLicht eingesetzt werden?

WebLicht ist ein browserbasiertes Tool, mit dem sich eine Reihe von vorbereitenden Routinen Preprocessing durchführen lassen. Dazu gehören Lemmatisierung (vgl. Lemmatisieren), Part-of-Speech-Tagging (vgl. POS), Named Entity Recognition (vgl. Named Entities) und Geotagging. Für viele der Funktionen (vgl. Feature) in WebLicht können Sie aus mehreren, unterschiedlichen Tools auswählen. Darum kann WebLicht für sehr unterschiedliche Forschungsprojekte eingesetzt werden. Eine mögliche Fragestellung wäre: Welche Orte kommen in Theodor Storms *Schimmelreiter* vor und mit welchen Figuren sind sie verknüpft?

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

Funktionen (Auswahl):

- Aufspaltung von Texten in einzelne Sätze (Sentence-Splitter)
- Aufspaltung von Texten in einzelne Wörter (Tokenizer (vgl. Type/Token))
- Zuordnung von Wörtern zu Wortarten (Part-of-Speech-Tagger)
- Lemmatisierung
- Morphologische Analyse
- Syntaxanalyse
- Automatische Annotation von Eigennamen (Named Entity Recognition (Schumacher 2024))

- Geolokalisierung und -visualisierung
- Nutzung von WebLicht im Einsteiger-Modus (*Easy Mode*) und im Fortgeschrittenen-Modus (*Advanced Mode*)

Zuverlässigkeit: WebLicht wird seit 2008 kontinuierlich weiterentwickelt. Das webbasierte Tool braucht nicht auf dem eigenen Rechner installiert zu werden und ist sehr zuverlässig.

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

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

WebLicht bietet die Nutzung in zwei unterschiedlichen Modi an, dem *Easy Mode* und dem *Advanced Mode.* Im Easy Mode können Sie vorbereitete Tool-Abfolgen nutzen. Im *Advanced Mode* können die sogenannten Pipelines selbst zusammen gestellt werden. Die Nutzung des *Easy Modes* ist sehr einsteigerfreundlich und bereitet auf die Nutzung des *Advanced Mode* vor. WebLicht ist in erster Linie ein computerlinguistisches Tool mit einer Reihe linguistisch relevanter Funktionen. Einige der Tools, die in WebLicht genutzt werden können, knüpfen aber auch an literaturwissenschaftliche Traditionen an, wie z. B. Named Entity Recognition (Schumacher 2024) oder Geolokalisierung. Die Benutzeroberfläche (vgl. GUI) ist leicht zu bedienen, doch computerlinguistisch wenig vorgebildeten Nutzer*innen kann es schwer fallen, im *Advanced Mode* die Kombinierbarkeit der unterschiedlichen Tools herauszufinden. Ein Einsteiger-Tutorial im Videoformat sowie Expert*innen-Interviews sind auf dem CLARIN-D-YouTube-Kanal zu finden, doch die Nutzung der unterschiedlichen Tools und ihrer Kombinationsmöglichkeiten im *Advanced Mode* ist damit nicht komplett abgedeckt. Es gibt aber ein ausführliches Manual, in dem alle Funktionen erläutert werden.

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

WebLicht ist ein in der Computerlinguistik gut etabliertes Tool. Auch in den digitalen Geisteswissenschaften wird WebLicht – vor allem als Tool-Suite für die Vorbereitung von Textanalyse - eingesetzt. Für traditionellere literaturwissenschaftliche Ansätze wird es derzeit noch eher selten genutzt.

5. Unterstützt WebLicht kollaboratives Arbeiten?

Nein, in WebLicht arbeiten Sie alleine an einem Text.

6. Sind meine Daten bei WebLicht sicher?

Ja. WebLicht hält sich an den Code of Conduct for Service Providers, einen Standard zur Datensicherheit, der im Rahmen der universitären Forschung und Lehre entwickelt wurde.

Externe und weiterführende Links

- Code of Conduct for Service Providers: https://web.archive.org/save/https://geant3plus.archive.geant.net/Pages/uri/V1.html (Letzter Zugriff: 04.06.2024)
- WebLicht: https://web.archive.org/save/https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page (Letzter Zugriff: 04.06.2024)

Bibliographie

Schumacher, Mareike. 2024. Methodenbeitrag: Named Entity Recognition (NER). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3765, https://fortext.net/routinen/methoden/named-entity-recognition-ner.

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.

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.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktionseinheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

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.

Property Property steht für „Eigenschaft“, „Komponente“ oder „Attribut“. In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt werden.

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.

Webanwendung Eine webbasierte Anwendung ist ein Anwendungsprogramm, welches eine Webseite als Schnittstelle oder Front-End verwendet. Im Gegensatz zu klassischen Desktopanwendungen werden diese nicht lokal auf dem Rechner der Nutzer*innen installiert, sondern können von jedem Computer über einen Webbrowser „online“ genutzt werden. Webanwendungen erfordern daher kein spezielles Betriebssystem.

Wordcloud Eine *Wordcloud* , oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.

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.