Toolbeitrag: Gensim

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

Thema:	word2vec	DOI:	10.48694/fortext.3817
Jahrgang:	1	Ausgabe:	10
Erscheinungsdatum:	30-10-2024	Erstveröffentlichung:	2021-05-03 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.

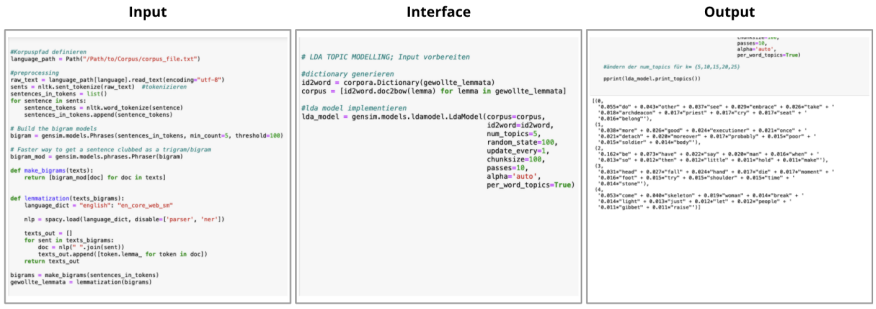

Abb. 1: Der Workflow von Gensim am Beispiel des LDA-Algorithmus' zum Topic Modeling: Vorab müssen alle Module und Packages installiert werden. Als erstes erfolgt die Definition des Korpuspfades & das Preprocessing. Nachdem das Korpus definiert und das Dictionary erstellt wurde, wird das Modell implementiert und Parametereinstellungen getroffen. Der Output ist beispielsweise eine Liste an Topicsets nach angegebener Topicanzahl.

- **Systemanforderungen:** Läuft auf Linux, Windows, MacOS und allen anderen Plattformen, die Python > 3.6 und NumPy unterstützen.
- **Stand der Entwicklung:** Wird seit 2008 entwickelt, letztes Release 01.April 2021 (Version 4.0.1.)
- **Herausgeber:** Radim Řehůřek und Petr Sojka
- **Lizenz:** GNU LGPL-Lizenz v2.1
- **Weblink:** https://radimrehurek.com/gensim/ 2
- **Im- und Export:** Import von Plain Text (vgl. Reintext-Version); Export möglich und individuell im Code anpassbar (Ergebnis-Speicherung als JSON oder Speicherung von Visualisierungen möglich)
- **Sprachen:** Vortrainierte Modelle für Englisch, Chinesisch, Deutsch, Französisch, Spanisch etc. vorhanden

1. Für welche Fragestellungen kann Gensim eingesetzt werden?

Gensim ist eine Open-Source-Bibliothek für Python und beinhaltet verschiedene Algorithmen, weshalb es für unterschiedliche Fragestellungen eingesetzt werden kann. Dabei ermöglichen es alle Algorithmen, automatisiert semantische Strukturen in den Textdaten zu entdecken. Gensim bietet sich insbesondere für die Verarbeitung großer Textsammlungen an.
Abhängig vom gewählten Modell, kann mit Topic-Modeling-Algorithmen (vgl. Topic Modeling) beispielsweise das Auftreten bestimmter Topics über einen Textverlauf betrachtet werden. Außerdem können Zusammenhänge zwischen bestimmten Themen und Faktoren wie Geschlecht, Nationalität des Autors, dem Erscheinungsjahr der Werke oder der Gattung der Texte erkannt werden (Jockers und Mimno 2013). Anhand von Textsammlungen eines Genres kann beispielsweise auch erörtert werden, ob verschiedene Autoren, Untergattungen und Zeiträume durch unterschiedliche Topic-Verteilungen charakterisiert sind (Schöch 2015). Mit Word2Vec hingegen können, auf Grundlage von Worteinbettungen, Figurenanalysen in großen Textsammlungen durchgeführt werden, welche wiederum Vergleiche zwischen Romanen oder Autoren erlauben (Grayson u. a. 2016). Zudem kann auch die semantische Komplexität von beispielsweise Romanen durch die Berechnung von Distanzen zwischen Worteinbettungen ermittelt werden (van Cranenburgh, van Dalen-Oskam und van Zundert 2019).

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

Funktionen:

- Bereitstellung von bereits trainierten Modellen und diversen Korpora in verschiedenen Formaten
- Unüberwachter Lernprozess (vgl. Machine Learning), keine Annotationen notwendig
- Verarbeitung von sehr großen Textsammlungen (vgl. Korpus)
- Wortvektoren trainieren mit Word2Vec, FastText und Doc2Vec
- Topic Modeling mit Latent Semantic Indexing (LsiModel)
- Topic Modeling mit Latent Dirichlet Allocation (vgl. LDA)

Zuverlässigkeit: Die Ergebnisse werden je nach Größe der Daten und manuell vorgenommenen Voreinstellungen zügig generiert. Die Ausführung von Word2Vec benötigt, je nach Korpusgröße, allerdings relativ viel Arbeitsspeicher und kann gegebenenfalls einige Stunden in Anspruch nehmen. Ein Tool, welches Textdaten in ähnlicher Größenordnung verarbeiten kann, ist derzeit nicht verfügbar.

3. Ist Gensim 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	teilweise
Gibt es eine gute Nutzerbetreuung?	✓

Gensim wurde entwickelt, um unstrukturierte digitale Textsammlungen im Plain-Text-Format (vgl. Reintext-Version) durch unüberwachte, maschinelle Lernverfahren (vgl. Machine Learning) zu verarbeiten, ohne dass dafür manuelle Annotationen (Jacke 2024) durchgeführt werden müssen. Als eine Open-Source-Bibliothek für Python ist Gensim allerdings nur für Nutzer*innen geeignet, die erste Programmierkenntnisse mit Python und generelles Codeverständnis mitbringen.
Ausführliche, englischsprachige Dokumentationen und Tutorials bieten, auf Grundlage bereits trainierter Modelle und vorverarbeiteter Korpora (vgl. Preprocessing), die Möglichkeit sich mit Gensim vertraut zu machen. Das Trainieren von Modellen mit eigenen Textsammlungen erfordert allerdings auch Kenntnisse im Bereich der computationellen Vorverarbeitung (vgl. Preprocessing) von Korpora (Bläß 2024). Außerdem müssen Parametereinstellungen bei der Implementierung der Algorithmen individuell an die Forschungsfrage angepasst werden.

Bei Fragen oder Problemen gibt es zwar nicht die Möglichkeit direkt Kontakt mit einem Support-Team aufzunehmen, Sie können aber über das Google-Forum und GitHub Hilfe erhalten.

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

Gensim etabliert sich zunehmend im Bereich der digitalen Literaturwissenschaften und ist z.B. in dem Digital-Humanities-Tools-Verzeichnis TAPoR eingetragen.
Da Gensim allerdings Grundkenntnisse in der Programmierung voraussetzt, ist es insbesondere in den digitalen Literaturwissenschaften für das Topic Modeling weniger etabliert als Tools wie DARIAH Topics Explorer (Schumacher 2024). Trotzdem ermöglicht die Nutzung von Gensim die individuelle Anpassung von Parametern an die Forschungsfrage.
Die Generierung von Worteinbettungen durch Word2Vec wird in den letzten Jahren auch zunehmend in den digitalen Literaturwissenschaften eingesetzt und dient beispielsweise als Werkzeug zur Unterstützung von quantitativen Literaturanalysen im Bereich des Distant Readings und Close Readings (Grayson u. a. 2016). In der traditionellen, literaturwissenschaftlichen Forschung findet Gensim noch keine Anwendung.

5. Unterstützt Gensim kollaboratives Arbeiten?

Nein, mit Gensim kann nicht direkt kollaborativ gearbeitet werden.
Ein Gensim-Projekt und die dazugehörigen Ressourcen können allerdings auf JupyterHub mit anderen For-

schenden geteilt werden, sodass zwar nicht direkt aber über einen Workaround kollaborativ gearbeitet werden kann.

6. Sind meine Daten bei Gensim sicher?

Ja, Gensim läuft auf dem eigenen Rechner. Alle Daten werden lokal verarbeitet, Texte müssen nirgendwo hochgeladen werden. Es werden keine personenbezogenen Daten erhoben.

Externe und weiterführende Links

- Gensim: https://web.archive.org/save/https://radimrehurek.com/gensim/ (Letzer Zugriff: 06.10.2024)
- Gensim Tutorials: https://web.archive.org/save/https://radimrehurek.com/gensim/auto_examples/index.html (Letzer Zugriff: 06.10.2024)
- Google-Forum: https://web.archive.org/save/https://groups.google.com/g/gensim (Letzter Zugriff: 06.10.2024)
- Gensim Q&A auf GitHub: https://web.archive.org/save/https://github.com/RaRe-Technologies/gensim/wiki/Recipes-&-FAQ (Letzer Zugriff: 06.10.2024)
- TAPoR: https://web.archive.org/save/http://tapor.ca/tools/1606 (Letzter Zugriff: 06.10.2024)

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.

Grayson, Siobhán, Maria Mulvany, Karen Wade, Gerardine Meaney und Derek Greene. 2016. Novel2Vec: Characterising 19th Century Fiction via Word Embeddings. In: https://researchrepository.ucd.ie/handle/10197/8360 (zugegriffen: 22. April 2021).

Jacke, Janina. 2024. Methodenbeitrag: Manuelle Annotation. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3748, https://fortext.net/routinen/methoden/manuelle-annotation.

Jockers, Matthew L. und David Mimno. 2013. Significant themes in 19th-century literature. *Poetics* 41, Nr. 6: 750–769. doi: 10.1016/j.poetic.2013.08.005,.

Schöch, Christof. 2015. Topic Modeling French Crime Fiction. In: *Digital Humanities 2015: Book of Abstracts*. Sydney: UWS.

Schumacher, Mareike. 2024. Toolbeitrag: DARIAH Topics Explorer. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 8. Topic Modeling (7. Oktober). doi: 10.48694/fortext.3728, https://fortext.net/tools/tools/dariah-topics-explorer.

van Cranenburgh, Andreas, Karina van Dalen-Oskam und Joris van Zundert. 2019. Vector space explorations of literary language. *Lang Resources & Evaluation* 53: 625–650. doi: 10.1007/s10579-018-09442-4,.

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.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

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.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte

selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

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.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

LDA LDA steht für *Latent Dirichlet Allocation* und ist ein generatives, statistisches Wahrscheinlichkeitsmodell, welches zum Topic Modeling angewendet werden kann. Bei der LDA werden auf Grundlage eines Wahrscheinlichkeitsmodells Wortgruppen aus Textdokumenten erstellt. Dabei wird jedes Dokument als eine Mischung von verborgenen Themen betrachtet und jedes Wort einem Thema zugeordnet. Wortreihenfolgen und Satzzusammenhänge spielen dabei keine Rolle.

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.

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.

Topic Modeling Das Topic Modeling ist ein statistisches, auf Wahrscheinlichkeitsrechnung basierendes, Verfahren zur thematischen Exploration größerer Textsammlungen. Das Verfahren erzeugt „Topics“ zur Abbildung häufig gemeinsam vorkommender Wörter in einem Text. Für die Durchführung können verschiedene Algorithmen und Modelle wie das LDA verwendet werden.

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.