Text Mining

Lang, Sarah; sarah.lang@uni-graz.at

Der Begriff des *Text Mining* (TM) wurde 1995 durch Ronen Feldman und Ido Dagan unter dem Titel *Knowledge Discovery from Text* (KDT) eingeführt; er bleibt jedoch bis heute wenig klar abgegrenzt. Allgemein wird TM aus dem Blickwinkel der Informatik als "a subfield devoted to the extraction of knowledge from unstructured text" angesehen (Jockers/Underwood 2016, S. 292). Zugehörig zur Domäne von *Data Mining* und *Data Science*, wird es als '*Data Mining* unter Benutzung von Textdaten' definiert und mitunter auch *Text Data Mining* genannt. Ziele dabei sind *Information Extraction* (IE), *Information Retrieval* (IR) und *Knowledge Discovery*.

Zur Datenverarbeitung wird *Natural Language Processing (NLP)* verwendet, wodurch eine Nähe zum Feld der Computerlinguistik entsteht. Im Gegensatz zum *Distant Reading* wird der Begriff *Text Mining* eher im Kontext der Informatik verwendet. Distant Reading findet sich in den Digital Humanities zumeist in Form von Computerphilologie (*Computational Literary Studies*), wobei digitale Analysemethoden für die Interpretation von Text fruchtbar gemacht werden sollen. *Text Mining* dagegen verfolgt Ziele aus dem Bereich der *Information Extraction*. *Text Mining* versteht Text als reines Datenbündel (vgl. *bag-of-words*) oder als Datenlieferanten, der selbst in der Analyse keine weitere Bedeutung mehr haben muss. Die Resultate werden dem Text nur 'entnommen'. Also ist *Text Mining* nicht primär ein hermeneutisches Tool zur Textinterpretation, sondern eher ein Werkzeug zur Textauswertung.

Die Aufgabe des *Text Mining* besteht in statistischer Pattern-Erkennung, die in Anwendungen wie *Text Clustering*, *Text Categorization*, *Entity Extraction*, *Document Summarization* oder auch *Sentiment Analysis* vorkommt. Aber auch TF-IDF (*term frequency-inverse document frequency*), Intertextualitäts- oder Plagiatserkennung (*Intertextuality / Text Reuse / Plagiarism Detection*) gehören dazu sowie das Pre-Processing von Inputtext durch Parsen und *Natural Language Processing (NLP)*, um eine gewisse Strukturierung der ansonsten als unstrukturiert bezeichneten Datengattung 'Text' zu erzielen. Die Bezeichnung des 'Mining' verweist auch besonders auf die Analyse der Big Data des Internet (*Web Mining*). Mitunter wird *Text Mining* auch mithilfe von *Machine Learning*-Algorithmen betrieben. Ressourcen, die speziell zum *Text Mining* erarbeitet wurden, sind außerdem zumeist nicht primär für die Anwendung auf historische Texte und Sprachen beziehungsweise überhaupt auf geisteswissenschaftliche Anwendungsszenarien ausgelegt. Im Fall von *Text Mining* wird außerdem tendenziell eher von Big Data-Anwendungen ausgegangen, wohingegen *Distant Reading*-Methoden zur quantitativen Textanalyse auch schon mit verhältnismäßig kleineren Textkorpora durchgeführt werden.

Literatur:

- Jockers, Matthew L.; Underwood, Ted: Text-Mining the Humanities. In: A New Companion to Digital Humanities. Chichester: 2016, S. 291–306.

Software:

Natural Language Toolkit (nltk), R

Verweise:

Data Mining, Distant Reading, NLP, Analysemethoden, NER, Datenvisualisierung, Interpretation

Themen:

Datenanalyse

Zitiervorschlag:

Lang, Sarah. 2021. Text Mining. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: https://gams.uni-graz.at/o:konde.194