Automatisierte Inhaltsanalyse mit R

Einleitung

Cornelius Puschmann

Contents

inhaltsanalyse-mit-r.de

Diese Einführung gliedert sich in neun inhaltliche Kapitel, in denen wesentliche Ansätze der automatisierten Inhaltsanalyse in R anhand von Beispielen vorgestellt werden, und zwei Beispielstudien, in denen eine Kombination zuvor vorgestellter Methoden systematisch angewandt wird. Derzeit ist nur die Arbeit an den Kapiteln 1-6 abgeschlossen.

Inhalt

Downloads

Sämtliche in dieser Einführung verwendeten R-Notebooks, Korpora und Lexika und können hier heruntergeladen werden.

R-Pakete

Die technische Grundlage dieser Einführung liefert das Paket quanteda, entwickelt von Ken Benoit und Kollegen, welches eine umfangreiche Basisinfrastruktur für die Analyse von Textdaten in R darstellt. Mit quanteda liest man Textdaten ein, legt man Korpora an, zählt man Wörter und wendet Lexika auf Daten an. Damit ist quanteda deutlich umfangreicher als die meisten vergleichbaren Pakete und eher ein vollumfängliches Textmining-Framework. Äquivalent mit Blick auf die Funktionalität sind allenfalls die Pakete tm und (begrenzt) tidytext. Im Vergleich zu tm ist quanteda zwar jünger, zeichnet sich aber durch einen großen Funktionsumfang, sehr gute Performance, und eine exzellente Dokumentation aus. Tatsächlich sind zahlreiche hier vorgestellte Beispiele direkt der quanteda-Dokumentation entnommen, die allerdings häufig etwas gegenüber dem Paket selbst hinterherhinkt.

Weitere Pakete werden für spezialisierte Aufgaben verwendet, die quanteda nicht abdeckt. Dazu gehören Themenmodelle genauso wie das überwachte maschinelle Lernen. Für den ersten Punkt setzen wir die Pakete topicmodels und stm ein, für den zweiten Punkt das Paket RTextTools.

Schließlich arbeiten wir intensiv mit den Paketen des tidyverse, die gemeinsam so etwas wie den großangelegten Versuch des neuseeländischen Statistikers Hadley Wickham darstellen, R trotz zahlreicher nicht unwesentlicher syntaktischer Indiosynkratien und Leistungsprobleme zu einer führenenden Sprache im Bereich der Datenwissenschaft zu machen. Wer sich einmal an die Logik von tidyr, dplyr und ggplot gewöhnt hat, möchte sie für gewöhnlich nicht mehr missen, auch wenn der Weg dorthin mitunter beschwerlich sein kann, etwa weil die spezielle Syntax nicht jedermanns Sache ist. Eine essentielle Einführung in das tidyverse ist das Buch R for Data Science von Garrett Grolemund und Hadley Wickham.

Korpora

Wir verwenden in dieser Einführung insgesamt neun unterschiedlich Korpora, anhand derer die vorgestellten Methoden anschaulich gemacht werden sollen. Bewusst unterscheiden sich die Daten in Bezug auf Sprache, Genre, Medium, Struktur und Umfang sehr deutlich. Von Social Media-Daten aus Facebook und Twitter über Pressetexte (in Auszügen) aus Deutschland, der Schweiz und den USA, bis hin zu politischen Reden und Parlmamentsdebatten ist eine große Bandbreite an Textsorten und Kontexten vertreten. Einige Korpora, etwa das Sherlock-Holmes-Korpus, sind primär wegen ihrer Anschaulichkeit und (durch vergleichsweise geringe Größe) gute Handhabbarkeit ausgewählt worden, und weniger deshalb, weil ihnen unbedingt große sozialwissenschaftliche Relevanz unterstellt wird. Zudem sind die Daten gemeinfrei, d.h. das Urheberrecht ist entweder erloschen oder schützt den Inhalt nicht (im Fall von Tweets oder Kommentaren). Im Korpus der Facebook-Kommentare sind bewusst keinelei Metadaten enthalten, um die Privatsphäre der Autoren so gut möglich zu schützen. Vom Sherlock Holmes-Korpus abgesehen sind die Daten ausgesprochen aktuell.

Korpus	Beschreibung	Texte	Wörter	Genre	Sprache	Quelle
Sherlock Holmes	Detektiv-Erzählungen von Arthur Conan Doyle	12	126.804	Literatur	en	archive.org
Twitter	Tweets von Donald Trump und Hillary Clinton im US-Präsidentschaftswahlkampf 2016	18.826	458.764	Social Media	en	trumptwitterarchive.com eigene Sammlung
Finanzkrise	Artikel aus fünf Schweizer Tageszeitungen mit dem Schlagwort 'Finanzkrise'	21.280	3.989.262	Presse	de	COSMAS
Bundestag	Transkripte der Plenardebatte des 18. Deutschen Bundestags (2013-2017)	205.584	15.296.742	Politik	de	offenenesparlament.de
EU	EUSpeech-Korpus aus Reden europäischer Politiker (national/EU) zwischen 2007 und 2015	17.505	14.279.385	Poltik	en	Schumacher et al, 2016

Korpus	Beschreibung	Texte	Wörter	Genre	Sprache	Quelle
UN	United Nations General Debate Corpus aus Transkripten der jährlichen UN-Generaldebatte nach Land, 1970-2017	7.897	24.420.083	Politik	en	Mikhaylov et al, 2017
Facebook	Zufallssample aus Kommentaren von sechs öffentlichen deutschsprachigen Facebook-Seiten, geposted zwischen 2015-2016	20.000	1.054.477	Social Media	de	eigene Sammlung
Die Zeit	Zufallssample von zwischen 2011 und 2016 veröffentlichten Nachrichtenbeiträgen	377	195.734	Presse	de	eigene Sammlung
New York Times	Inhaltanalyse von Beiträgen aus der New York Times zu dem Projekt 'Making the News' (1996-2006)	30.862	215.275	Presse	en	Boydstun, 2013

Hintergrund

Warum eine deutschsprachige Einführung in die automatisierte inhaltsanalyse mit R? Die klassische (d.h. manuelle) Inhaltsanalyse ist eine der wichtigsten Methoden der empirischen Sozialwissenschaften, und es existieren zahlreiche auflagenstarke Standardwerke, an die dieser knappe Überblick ganz sicher nicht heranreicht, was seinen Detailreichtum, die Tiefe der methodischen Einordnung, und den Grad der praktischen Erprobtheit angeht. Allerdings ist das Angebot an Lehrbüchern bereits deutlich eingeschränkter, wenn man sich der (teil)automatisierten Inhaltsanalyse zuwendet, und eine hinreichend anwendungsnahe Beschreibung sucht, die vor Code nicht zurückschreckt, und die zudem noch frei verfügbar ist. In dieser zugegeben engen Sparte gibt es deutlich weniger Auswahl, und zumeist liegt der Fokus auf einzelnen proprietären Programmen mit grafischer Nutzeroberfläche, die i.d.R. nicht kostenlos verfügbar und wenig leistungsstark sind, und zudem zum Teil relativ schnell veralten.

Programmiersprachen für die Datenwissenschaft, vor allem R und Python, bieten seit einigen Jahren viele neue Möglichkeiten für die Anwendung innerhalb der sozialwissenschaftlichen Forschung – nicht nur im Bereich Statistik. Solche Sprachen sind häufig flexibler, vielseitiger und leistungsstärker als kommerzielle Standardwerkzeuge wie SPSS und MaxQDA, was nicht bedeutet, dass man nicht beides nebeneinander nutzen kann. Aber gerade in den letzten Jahren hat R gewaltig zugelegt, was sein Potenzial für Bereiche wie die Inhaltsanalyse angeht, in denen zuvor eine Kombination aus manuellen Ansätzen und unflexiblen

Standardprogrammen vorherrschte. Die Entwicklung von leistungsfähigen R-Paketen speziell für die sozialwissenschaftliche Forschung, wie etwa quanteda, stm und RTextTools, erleichtert die Arbeit mit Inhaltsdaten so stark, dass R nicht mehr hinter dem Hauptkonkurrenten Python zurückstehen muss, wenn es um die effiziente Analyse von Textdaten geht.

Aber was ist hier überhaupt mit Inhaltsanalyse gemeint? Um Enttäuschungen vorzubeugen: In dieser Einführung wird ausschließlich mit Text gearbeitet, auch wenn mit der klassischen Inhaltsanalyse natürlich auch Bilder und Videoinhalte umtersucht werden. Zwar tut sich in diesen Bereichen in den letzten Jahren sehr viel, Verfahren für die Analyse nicht-textueller Inhalte würden den Rahmen dieser Einführung aber klar sprengen. Ich habe dem Begriff *Inhaltsanalyse* dennoch bewusst der Vorzug gegenüber verwandten Begriffen wie *Textmining* gegeben, um klarzustellen, dass für uns das sozialwissenschaftliche Erkenntnisinteresse im Mittelpunkt steht, nicht die Feinheiten einzelner technischer Verfahren. Zugleich verwende ich den Begriff nicht in der relativ engen Lesart, die in der Kommunikations- und Medienwissenschaft häufig vorherrscht. Grund hierfür ist einerseits, dass diese Einführung idealerweise für Kommunikations- und Medienwissenschaftler genauso wie für Soziologen und Politikwissenschaftler nützlich ist (und natürlich auch gerne über diese Fachbereiche hinaus verwendet werden kann), und andererseits, dass es in den folgenden neun Kapiteln immer wieder Bezüge zu Ansätzen gibt, die klar aus der Computerlinguistik und Informatik kommen, und die das sozialwissenschaftliche Methodenrepertoire eindeutig bereichern. Zugleich existieren in diesen Disziplinen wichtige Techniken, die für Sozialwissenschaftler vergleichsweise wenig relevant sind, etwa die Wortartbestimmung (Tagging) oder die syntaktische Analyse (Parsing), die wir hier nur sehr am Rand behandeln. Auch werden gängige Verfahren in technischen Fächern so behandelt, dass Studierende diese idealerweise selbst verbessern oder erweitern können. Dieses Interesse steht in dieser Einführung eindeutig nicht im Vordergrund. Stattdessen geht es bei der computergestützten Analyse innerhalb der computational communication science um die kompetente Anwendung solcher Techniken, mit dem klaren Ziel, Erkenntnisse über gesellschaftliche Phänomene aus Texten zu gewinnen. Deshalb spreche ich ganz bewusst nicht von *Text- oder Datamining*, sondern von *Inhaltsanalyse*, auch wenn wir im Verlauf der folgenden neun Kapitel sehr viel mit Begiffen wie Korpus, Wortfrequenz und Textstatistik hantieren, die vermutlich in den meisten klassischen Einführung in die Inhaltsanalyse fehlen. Dass schließlich trotzdem der Weg über R gewählt, und nicht etwa ein Tool mit graphischer Bedienoberfläche herangezogen wird, ist kein Wiederspruch. Das Vorurteil, *Programmieren = Informatik* hält sich leider immer noch in den Sozialwissenschaften, gerade unter älteren Semestern, auch wenn sich mit dem Internet radikal verändert hat wie und was man programmiert, und im Zuge dessen auch, wie relevant das Programmieren für die Sozialwissenschaften ist.

Eine frei verfügbare deutschsprachige Einführung in die automatisierte Inhaltsanalyse mit R, die genau diesen anwendungsbezogenen und sozialwissenschaftliche Fokus hat, und die zugleich ganz konkrete Code-Beispiele liefert, statt die Inhaltsanalyse primär abstrakt zu erklären, fehlte in meinen Augen bislang, auch wenn fortgeschrittene Überblicke zu Themen wie der Verknüpfung von maschinellem Lernen und Inhaltsanalyse bereits existieren. Ob das Experiment nun gelungen ist oder nicht, entscheiden wie immer die Leserinnen und Leser. Dabei steht die konkrete Anwendung in den folgenden Kapiteln klar gegenüber der theoretischen Reflexion im Hintergrund. Standardwerke wie die von Rössler, Mayring, Früh oder Merten sollten hier dringend herangezogen werden, um die Inhaltsanalyse inklusive ihrer Entwicklungsgeschichte besser zu verstehen. Wem die zu Beginn vorgestellten Techniken zu sehr von der klassischen Inhaltsanalyse mit manueller Kodierung entfernt sind, lege ich das Kapitel 6 besonders ans Herz, in dem es um das Ableiten von Inhaltsanalyse-Kategorien aus strukturellen Merkmalen (in der Regel sind das Wörter) anhand von Verfahren des überwachten maschinellen Lernens geht. Dort dürfte der Bezug zwischen traditioneller und automatisierter Inhaltsanalyse am deutlichsten werden.

Last but not least: Diese Einführung setzte neben Wissen über die Inhaltsanalyse und etwas sozialwissenschaftlicher Vorbildung auch grundlegende R-Kenntnisse voraus. Einführungen in R gibt es zuhauf, etwa die von Zuckarelli und Luhmann. Über 'normales R' hinaus werden nahezu alle Pakete aus dem sogenannten tidyverse eingesetzt, allen voran ggplot2 und dplyr. Das Buch R for Data Science von Garrett Grolemund und Hadley Wickham ist hier besonders nützlich, um die Codebeispiele zu verstehen. Und natürlich handelt sich bei dieser Einführung um *work in progress* – Feedback und Kritik sind mir sehr willkommen!

CoRnelius Puschmann

puschmann@gmail.com / cbpuschmann

Hamburg, August 2018