

Was liest und schreibt man über Inklusion?

Web-Scraping und Text-Mining mit R am Beispiel einer Online-Nachrichten- und Diskussionsseite für Lehrkräfte

Pawel R. Kulawiak

2023-09-19

Preprint in Progress

Inhaltsverzeichnis

Einleitung	1
Ziele	4
Allgemeine Zielsetzung	4
Zielsetzung mit R: Web-Scraping und Text-Mining	4
News4teachers: Online-Nachrichten- und Diskussionsseite für Lehrkräfte	4
Inhalte von News4teachers und potenzielle Leserschaft aus Lehrkräften	5
Explorative Forschungsfragen	6
Web-Scraping	7
R-Zusatzpakete	7
R-Zusatzpaket <i>rvest</i>	7
R-Zusatzpaket <i>tidyverse</i>	7
Struktur und Inhalte der Webseite	8
Erster Web-Scraping-Versuch	12
Datenstruktur	15
Weitere Web-Scraping-Schritte	15
Literatur	18

Einleitung

Mein wertgeschätzter Kollege Timo Lüke (<https://timolueke.de/>) hat einst im Rahmen einer Medieninhaltsanalyse deutschsprachiger Printmedien (Lüke u. a. 2014) folgende Forschungsfragen aufgeworfen:

- Welches Verständnis von Inklusion wird in den deutschen meinungsführenden Medien kommuniziert?
- Welche Argumente für und gegen die Umsetzung von Inklusion werden genannt?
- Welche Fallbeispiele werden als Belege angeführt?

“Im Rahmen einer systematischen Inhaltsanalyse (Rössler, 2010) deutscher Printmedien untersuchen wir die öffentliche Berichterstattung zum Thema „Inklusion“. Dabei wollen wir verbreitete Definitionen, Argumente und Fallbeispiele systematisch erfassen. So sollen langfristig die Analyse des medialen Diskurses und in der Folge eine Versachlichung der kontroversen Debatte über Inklusion ermöglicht werden.” (Lüke u. a. 2014)

Erste Ergebnisse der Medieninhaltsanalyse sind in Form einer Posterpräsentation verfügbar (Lüke u. a. 2014) und ich erlaube mir die Darstellung des interessanten Posters (Abbildung 1).

Potsdamer Morgenpost

Freitag, 28. November 2014

Sonderausgabe zur AESF-Herbsttagung

10,50 €

Was liest man über Inklusion?

Konzeption einer Medieninhaltsanalyse deutschsprachiger Printmedien

Timo Lüke¹, Matthias R. Hastall², Christian Marschler³ & Michael Grosche¹

¹Universität Potsdam, ²Technische Universität Dortmund, ³Filmuniversität Babelsberg

Das Thema „Inklusion“ wird von den Medien zunehmend als relevant erkannt und entsprechend berücksichtigt. Inwiefern ihre Berichterstattung die Meinungen zur Inklusion in der Bevölkerung (und somit auch bei pädagogischen Fachkräften) beeinflussen, ist noch nicht untersucht worden.

Im Rahmen einer systematischen Inhaltsanalyse (Rössler, 2010) deutscher Printmedien untersuchen wir die öffentliche Berichterstattung zum Thema „Inklusion“. Dabei wollen wir verbreitete Definitionen, Argumente und Fallbeispiele systematisch erfassen. So sollen langfristig die Analyse des medialen Diskurses und in der Folge eine Versachlichung der kontroversen Debatte über Inklusion ermöglicht werden.

Forschungsfragen

- Welches Verständnis von Inklusion wird in den deutschen Meinungsführermedien kommuniziert?
- Welche Argumente für und gegen ihre Umsetzung werden dort genannt?
- Welche Fallbeispiele werden als Belege angeführt?

Material

- Aufgreifkriterium: Stichwort „Inklusion“ in einem der Textbestandteile (Überschrift, Text,...)
- Publikationszeitraum: Januar-Juni 2014
- Publikationen: alle überregionalen, deutschen Tages- und die reichweitenstärksten Wochenzeitungen und -magazine: Bild, Süddeutsche Zeitung, Frankfurter Allgemeine Zeitung (FAZ), Die Welt, Handelsblatt, Die Tageszeitung (taz), Neues Deutschland, Bild am Sonntag, Die Zeit, Welt am Sonntag, Frankfurter Allgemeine Sonntagszeitung (FAS), Der Spiegel, Stern, Bunte, Focus.

Formale Codierung

- Publikation
- Publikationsdatum
- Platzierung (Ressort, Seite, Buch)
- Beitragstyp
- Länge
- Bebildung

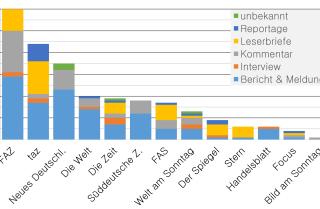
Zahlen des Tages

Im ersten Erhebungszeitraum erschienen insgesamt 252 Beiträge in 13 der 15 indizierten Zeitungen und Magazine. Die Bild (überregional) und Bunte druckten keine Beiträge zum Thema. Der „Fall“ Henri scheint ein wichtiger Auslöser der Berichterstattung

zu sein. Kommentare und Leserbriefe haben einen unerwartet hohen Anteil an den erschienenen Beiträgen. Der Diskurs (auch in den Onlineforen) wird in einem Tochterprojekt analysiert.



Artikel im Zeitverlauf



In Anlehnung an entsprechende Strukturierungsversuche von Göransson & Nilholm (2014) sowie Grosche (angenommen):

Definition / Verständnis von Inklusion

- Teilhabe in allen Lebensbereichen der Gesellschaft
- Entkategorisierung / Dekategorisierung
 - Behinderung, sonderpäd., Förderbedarf, etc.
 - andere Heterogenitätsdimension / Kategorie
- Veränderung der Zuordnung zu Schulen
 - Kinder mit Behinderungen gehen in allgemeine Schulen
 - Alle Kinder gehen in die wohnortnächste Schule
- Veränderung des Unterrichts / der Lehr-Lern-Prozesse
- Veränderung der Bewertungsmaßstäbe
- Veränderung der äußeren Strukturen (Schulsystem & lokal)
 - Abschaffung der Förderschulen
 - Abschaffung der Mehrgliedrigkeit
- Veränderung der inneren Strukturen (Schule & Schulleben)
 - Multiprofessionelle Teams
 - Demokratische Schule / Mitbestimmung der SchülerInnen
- Veränderung der pädagogischen Profession(en) / Berufsbilder
 - Sonderpädagogik geht in die allgemeine Pädagogik auf
 - Allgemein- & Sonderpädagogen unterrichten gemeinsam
 - SonderpädagogInnen als ExpertInnen zur Unterstützung
- politischer Begriff der etwas mit den Interessen bzw. Rechten von Menschen mit Behinderung zu tun hat
- unklares Konstrukt / noch nicht klar

Unterschied zwischen Inklusion und Integration

- „Inklusion“ ist gleichbedeutend mit „Integration“
- Unterscheidung undifferenziert / implizit
- Unterscheidung differenziert / explizit

Literatur

Göransson, K., & Nilholm, C. (2014). Conceptual Diversities and Empirical Shortcomings – A Critical Analysis of Research on Inclusive Education. *EJSNE*, 29, 265–280. doi:10.1080/0856257.2014.933545

Grosche, M. (In Druck). Was ist Inklusion?. In P. Kuhl, P. Stanat, B. Lütje-Klose, C. Griesch, M. Prenzel, & H. A. Pant (Hsg.), *Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen: Grundlagen und Befunde*. Wiesbaden: Verlag für Sozialwissenschaften.

Rössler, P. (2010). *Inhaltsanalyse*. Konstanz: UTB.
Originalgrafik (Mitte): Robert Aehnelt (CC BY-SA 3.0). commons.wikimedia.org/wiki/File:Schrifte_zur_Inklusion.svg

Impressum

Timo Lüke
Universität Potsdam
Humanwissenschaften
Inklusionspädagogik
timo.lueke@uni-potsdam.de

Download



Abbildung 1: Posterpräsentation von Lüke u. a. (2014): Was liest man über Inklusion?

Ziele

Allgemeine Zielsetzung

Ich möchte die Medieninhaltsanalyse von Lüke u. a. (2014) replizieren sowie erweitern und mich dabei auf die Textinhalte einer Online-Nachrichten- und Diskussionsseite für Lehrkräfte fokussieren, nämlich News4teachers (News4teachers 2022).

Zielsetzung mit R: Web-Scraping und Text-Mining

Ich möchte exemplarisch aufzeigen, wie die einzelnen Projektphasen der Medieninhaltsanalyse mit der Programmiersprache R umgesetzt werden können. Hierfür werden wir uns auf zwei wichtige Arbeitsschritte fokussieren:

- **Web-Scraping**, also eine automatisierte Methode zum Extrahieren der Textinformationen von der Webseite News4teachers.
- **Text-Mining**: Die mittels Web-Scraping gesammelten Textdaten sollen mit Methoden des Text-Minings analysiert werden. Methoden des Text-Minings fokussieren sich auf die Extraktion von nützlichen Informationen aus unstrukturierten Textdaten. Unstrukturierte Textdaten sind Texte, die nicht in einer festen Datenbankstruktur vorliegen, also z.B. Textinhalte von Webseiten. Mit Methoden des Text-Minings kann auch der sentimentale Ton oder die subjektive Meinung eines Textinhalts ermittelt werden. Das Hauptziel der sogenannten Sentimentanalyse besteht also darin, die in einem Textdokument geäußerten Emotionen und Ansichten bezüglich eines bestimmten Themas zu identifizieren, in unserem Fall also z.B. geäußerte Meinungen zum Thema Inklusion.

News4teachers: Online-Nachrichten- und Diskussionsseite für Lehrkräfte

Bevor wir mit dem Web-Scraping und Text-Mining beginnen, betrachten wir zunächst das Arbeitsmaterial, also die Webinhalte der Webseite News4teachers, und die entsprechende Selbstbeschreibung der Webseite (News4teachers 2022):

“Wer steckt hinter News4teachers?

News4teachers wird von einer Redaktion aus Lehrern und Journalisten betrieben. Die Seite ist ein gemeinsames Projekt von [4teachers](#), der Service-Plattform von Lehrern für Lehrer, sowie [der Agentur für Bildungsjournalismus](#).

Was ist News4teachers?

News4teachers ist eine Nachrichten- und Diskussionsseite, die sich mit seriösen Berichten, Analysen und Kommentaren an pädagogische Profis und die an

Bildungsthemen interessierte Öffentlichkeit richtet. Die Redaktion sichtet täglich die Meldungen aus Politik, Forschung und Gesellschaft. Auf die Seite gelangt alles, was für die Bildung wichtig ist. News4teachers bietet also einen aktuellen Überblick über die relevanten Informationen für Lehrer, Erzieher, Schüler und Eltern. Und zwar: unabhängig und überparteilich.

Was ist die Idee hinter News4teachers?

News4teachers fühlt sich dem klassischen Journalismus verpflichtet. Das heißt konkret: Wir unterwerfen uns den publizistischen Grundsätzen des Deutschen Presserats, dem [Pressekodex](#). Informationen, die auf die Seite gelangen, wurden zuvor von der Redaktion mit der gebotenen Sorgfalt geprüft. Quellen werden stets genannt, Meinung und Bericht voneinander getrennt. News4teachers unterliegt zudem einer Chronistenpflicht: Alles, was für die Bildungsdebatte in Deutschland von Bedeutung ist, wird aktuell berichtet. Regelmäßige Nutzer von News4teachers sind also immer im Bild.” (News4teachers 2022)

Inhalte von News4teachers und potenzielle Leserschaft aus Lehrkräften

News4teachers verspricht eine unabhängige und überparteiliche Berichterstattung zu Bildungsthemen, wahrscheinlich auch zum Thema Inklusion. Die Inhalte sind für die Leserschaft kostenfrei (werbefinanziertes Angebot). Die Inhalte von News4teachers sind außerdem speziell auf Lehrkräfte ausgerichtet. Somit kann angenommen werden, dass ein großer Teil der Leserschaft aus Lehrkräften besteht. Die Internetseite News4teachers hatte folgende Besucherzahlen (Jahr 2023): Mai (54000 Personen), Juni (60000 Personen) und Juli und August jeweils 55000 Personen ([Zahlen ermittelt mit: https://neilpatel.com/website-traffic-checker/](https://neilpatel.com/website-traffic-checker/)). Nehmen wir an, dass die Leserschaft von News4teachers zu 75% aus Lehrkräften aus Deutschland bestünde, dann hätten wir bei einer monatlichen Besucherzahl von 55000 Personen eine monatliche Leserschaft von ca. 41250 Lehrkräften (55000 * 0,75 = 41250). In Deutschland gibt es aber laut Mikrozensus 2022 rund 975000 Lehrkräfte an allgemeinbildenden Schulen. Die potenzielle News4teachers-Leserschaft aus Lehrkräften (41250 Personen) entspräche dann einem Anteil von ca. 5.64% aller Lehrkräfte an allgemeinbildenden Schulen (55000 / 975000 * 100 = 5.64%). Im dargestellten Szenario würden die Inhalte von News4teachers also pro Monat ca. 5.64% der Lehrkräfte an allgemeinbildenden Schulen in Deutschland erreichen (5 von 100 Lehrkräften lesen News4teachers). Dies sind aber nur vage Vermutungen zur Reichweite von News4teachers unter Lehrkräften an allgemeinbildenden Schulen in Deutschland, unter der Annahme, dass 75% der Leserschaft von News4teachers aus Lehrkräften bestünde.

Die Webseite News4teachers bieten der Leserschaft die Möglichkeit die Inhalte zu kommentieren und zu diskutieren (Abbildung 2 und Abbildung 6). Hierfür formuliert die Redaktion spezifische Richtlinien (News4teachers 2022):

“Gibt’s Regeln für die Leserzuschriften in den Foren?

Grundsätzlich gilt: Niemand hat einen Anspruch darauf, in den Foren zu den einzelnen Artikeln eine eigene Wortmeldung zu veröffentlichen. Die Redaktion

legt Wert darauf, nur Leserzuschriften zu veröffentlichen, die erkennbar darauf abzielen, einen inhaltlichen Beitrag zur Diskussion des darüberstehenden Artikels zu leisten. Das bedeutet konkret: Auch für Leserzuschriften gelten die publizistischen Grundsätze des Deutschen Presserats, gilt also [der Pressekodex](#).

Kurzgefasst:

- Wir veröffentlichen keine Leserbeiträge, in denen ungeprüfte, unbelegte oder falsche Tatsachenbehauptungen verbreitet werden.
- Wir veröffentlichen keine Hetze gegen Menschen oder Menschengruppen.
- Wir veröffentlichen keine Werbung, ob nun für Produkte oder Parteien.
- Und wir veröffentlichen keine Links auf un seriöse Quellen.

Wir behalten uns darüber hinaus vor, Leserbriefe, die lediglich der Stimmungsmache dienen, zu löschen. Oder Leserbriefe sinnwährend zu kürzen.“ (News4teachers 2022)

[Hier weitere Erläuterungen einfügen]



„Schämt Euch!“ – Deutschland steht vor den Vereinten Nationen am Pranger, weil es die Inklusion an Schulen praktisch verweigert

29. August 2023

152

GENF. „Schämt Euch!“ – so heißt es auf einem Transparent, das Aktivistinnen und Aktivisten des Berliner Bündnisses für schulische Inklusion vor dem Palais der...

Abbildung 2: Beitrag zum Thema Inklusion mit 152 Leserkommentaren auf der Internetseite News4teachers (Quelle: <https://www.news4teachers.de/2023/08/schaeamt-euch-deutschland-steht-vor-den-vereinten-nationen-am-pranger-weil-es-die-inklusion-an-schulen-verweigert/>)

Explorative Forschungsfragen

Die Inhalte von der Webseite News4teachers und die Kommentare und Diskussionen der Leserschaft eignen sich eventuell zur Beantwortung folgender Forschungsfragen:

- Auf welche Art und Weise wird das Thema Inklusion auf der Online-Nachrichten- und Diskussionsseite für Lehrkräfte dargestellt?
- Auf welche Art und Weise werden die Inhalte zum Thema Inklusion von der Leserschaft kommentiert und diskutiert?

Web-Scraping

Der erste Arbeitsschritt, hin zum Text-Mining, also hin zur Medieninhaltsanalyse, wird nun das Web-Scraping sein, also die automatisierte Extraktion der Webinhalte (z.B. Textinformationen) von der Webseite News4teachers. Traditionellerweise bzw. almodischerweise würde man Webinhalte mit der Methode “*copy-and-paste*” in einen Datensatz übertragen, also z.B. Text von einer Webseite kopieren und anschließend die kopierte Textinformation in einen Datensatz einfügen (z.B. bei Excel). Dieses Verfahren ist aber fehleranfällig, da z.B. die Gefahr besteht, dass aufgrund mangelnder Konzentration falsche oder unvollständige Textinhalte übertragen werden. Web-Scraping ist daher als automatisierte Methode der Extraktion von Webinhalten weniger anfällig für Fehler und somit die Methode der Wahl.

R-Zusatzpakete

R-Zusatzpaket *rvest*

Für das Web-Scraping nutzen wir nun das R-Zusatzpaket *rvest*. Der Name des R-Zusatzpaketes ist eine gelungene Anspielung auf das englische Wort *harvest* (ernten, sammeln), denn wir wollen ja Informationen aus dem Internet sammeln (mit R). Das kreative Wortspiel ist auch im Logo des R-Zusatzpaketes visualisiert (Abbildung 3). Zunächst müssen wir das R-Zusatzpaket installieren und laden.

```
install.packages("rvest")
library(rvest)
```



Abbildung 3: Logo des R-Zusatzpaketes *rvest*

Das R-Zusatzpaket *rvest* verfügt über eine umfassende und hilfreiche Online-Dokumentation:

<https://rvest.tidyverse.org/>

R-Zusatzpaket *tidyverse*

Das R-Zusatzpaket *tidyverse* ist eine Zusammenstellung unterschiedlicher R-Zusatzpakete. Wir werden an diversen Stellen die herausragende Funktionalität des R-Zusatzpaketes

tidyverse nutzen. An den entsprechenden Stellen wird ein Verweis auf die R-Zusatzpakete erfolgen. Informationen zum R-Zusatzpaket *tidyverse* findet man hier:

<https://www.tidyverse.org/>

Wir installieren und laden das R-Zusatzpaket:

```
install.packages("tidyverse")
library(tidyverse)
```

Struktur und Inhalte der Webseite

Ziel des Web-Scrapings wird es sein, die relevanten Webinhalte von News4teachers automatisiert zu extrahieren. Hierfür müssen wir uns erstmal einen Überblick über die Struktur und Inhalte der Webseite verschaffen. Die Beiträge auf den Internetseiten von News4teachers haben eine spezifische Struktur mit spezifischen Webinhalten. Wir betrachten den Beitrag mit dem Titel „*Schämt Euch!*“ – *Deutschland steht vor den Vereinten Nationen am Pranger, weil es die Inklusion an Schulen praktisch verweigert*“.¹ Für unsere Forschungsfragen mehr oder weniger interessante Webinhalte sind in den Abbildungen kenntlich gemacht (Abbildung 4, Abbildung 5 und Abbildung 6).

¹<https://www.news4teachers.de/2023/08/schaemt-euch-deutschland-steht-vor-den-vereinten-nationen-am-pranger-weil-es-die-inklusion-an-schulen-verweigert/>

Titel

Erscheinungsdatum

Gefällt-Mir-Anzahl

„Schämt Euch!“ – Deutschland steht vor den Vereinten Nationen am Pranger, weil es die Inklusion an Schulen praktisch verweigert

29. August 2023

Gefällt mir 608

Teilen

Anzahl der Kommentare

GENF. „Schämt Euch!“ – so heißt es auf einem Transparent, das Aktivistinnen und Aktivisten des Berliner Bündnisses für schulische Inklusion vor dem Palais der Vereinten Nationen in Genf platziert haben. Und: „Deutschland verweigert das Menschenrecht auf inklusive Bildung.“ Der Ort des Protests, zu dem Dutzende von Initiativen aufgerufen haben, ist kein Zufall: Heute und morgen findet hier eine sogenannte Staatenprüfung statt, in der Deutschland im Mittelpunkt steht – genauer: das Engagement, das die Bundesrepublik zeigt, um die UN-Behindertenrechtskonvention umzusetzen. Die Kritik daran ist scharf.

Berliner Bündnis für schulische Inklusion
@bbsinklusion · Folgen

#InklusiveBildungJETZT
Wir sind auf dem Place de Nations angekommen! Die #Staatenprüfung für Deutschland beginnt in Kürze.

Abbildung 4: (a) Beitrag auf der Internetseite News4teachers und Struktur der Webinhalte
 (Quelle: <https://www.news4teachers.de/2023/08/schaeamt-euch-deutschland-steht-vor-den-vereinten-nationen-am-pranger-weil-es-die-inklusion-an-schulen-verweigert/>)

Externe Inhalte und/oder Bilder

#WirFahrenNachGenf



9:40 vorm. · 29. Aug. 2023 aus Genf, Schweiz

(i)

21 Antworten Teilen

1 Antwort lesen

Text

Der offizielle Beitrag Deutschlands fällt dünn aus – bezeichnend für das, was sich in den vergangenen Jahren in Sachen Inklusion in der Schule getan hat: Gerade mal eine halbe Seite bringt die Bundesregierung zusammen, um die Maßnahmen seit 2019 zu beschreiben, um den Anspruch der Behindertenrechtskonvention auf ein "integratives Schulsystem auf allen Ebenen" (Artikel 24) zu realisieren. Konkret wird angeführt: ein Zwischenbericht der KMK zur Lehrerbildung von 2020, eine Empfehlung der KMK zur individuellen Förderung an Berufsschulen sowie eine Förderrichtlinie "Unterstützende Diagnostik in der inklusiven Bildung". Zahlen? Daten? Fakten? Fehlanzeige.

Abbildung 5: (b) Beitrag auf der Internetseite News4teachers und Struktur der Webinhalte
(Quelle: <https://www.news4teachers.de/2023/08/schaemt-euch-deutschland-steht-vor-den-vereinten-nationen-am-pranger-weil-es-die-inklusion-an-schulen-verweigert/>)

Anzahl Kommentare **157 KOMMENTARE**

User-Name **DerechteNorden** ⏱ 7 Tage zuvor

Kommentar
Welches Deutschland? Das bayerische?
In SH wird Inklusion nicht verweigert. Trotz der z.T. widrigen Umstände.
Nachdem, was ich hier gelesen habe, ist das in NRW z.B. auch nicht der Fall.

Anzahl Likes für Kommentar **9** Antworten

User-Name **Redaktion** ⏱ 7 Tage zuvor

Relation zwischen Antwort und User-Name

| ↗ Antwortet *DerechteNorden*

In Schleswig-Holstein lag die Exklusionsquote 2008 bei 3,1 Prozent, 2020/21 lag sie bei 2,3 Prozent. Das heißt: Es werden 0,8 Prozentpunkte weniger Schülerinnen und Schüler an Sonderschulen unterrichtet als vor 15 Jahren. Damit liegt Schleswig-Holstein tatsächlich noch relativ gut im Vergleich zu anderen Bundesländern.

In Nordrhein-Westfalen gingen 2020/2021 4,8 Prozent der Schülerinnen und Schüler auf Sonderschulen – 2008 waren es 5,2 Prozent gewesen, 2018/2019 4,7. Das heißt: Vor 15 Jahren wurden gerade mal 0,4 Prozentpunkte mehr Schülerinnen und Schüler an Sonderschulen unterrichtet. Zuletzt ist ihr Anteil sogar wieder gestiegen.

Für ganz Deutschland sieht die Bilanz nach 14 Jahren UN-Behindertenrechtskonvention so aus: 4,4 Prozent der Schülerinnen und Schüler besuchen Sonderschulen – 2008 waren es 4,9 Prozent gewesen, also nur 0,5 Prozentpunkte mehr.

Gerne hier nachlesen: <https://www.aktion-mensch.de/inklusion/bildung/hintergrund/zahlen-daten-und-fakten/inklusionsquoten-in-deutschland>

Herzliche Grüße
Die Redaktion

Anzahl Likes für Antwort **2** Antworten

User-Name **Alex** ⏱ 7 Tage zuvor

Relation zwischen Antwort und User-Name

| ↗ Antwortet *Redaktion*

Kommentar als Antwort auf Antwort
Na ja, die Eltern haben halt die Wahl. Für manche ist die Förderschule eben die bessere Entscheidung. So, wie Kinder nicht per se auf

Abbildung 6: (c) Beitrag auf der Internetseite News4teachers und Struktur der Webinhalte
 (Quelle: <https://www.news4teachers.de/2023/08/schaemt-euch-deutschland-steht-vor-den-vereinten-nationen-am-pranger-weil-es-die-inklusion-an-schulen-verweigert/>)

[Erläuterungen zu den Abbildungen und Inhalten hinzufügen]

Erster Web-Scraping-Versuch

Zuvor haben wir uns einen Überblick über die zu extrahierenden Webinhalte verschafft. Für den ersten Web-Scraping-Versuch nutzen wir weiterhin den Beitrag mit dem Titel „*Schämt Euch!*“ – *Deutschland steht vor den Vereinten Nationen am Pranger, weil es die Inklusion an Schulen praktisch verweigert*“ (Abbildung 4). Dies ist der Link zum Beitrag:

<https://www.news4teachers.de/2023/08/schaemt-euch-deutschland-steht-vor-den-vereinten-nationen-am-pranger-weil-es-die-inklusion-an-schulen-verweigert/>

Wir nutzen den Befehl `read_html()` und den entsprechenden Link, um sämtliche Informationen von der Webseite zu extrahieren.

```
html <- read_html("https://www.news4teachers.de/2023/08/schaemt-euch-deutschland-steht-vor-den-vereinten-nationen-am-pranger-weil-es-die-inklusion-an-schulen-verweigert/")
```

Alle Webinhalte sind nun im Objekt `html` hinterlegt. Wir sind allerdings nur an spezifischen Webinhalten interessiert und möchten daher im nächsten Schritt einen spezifischen Textinhalt aus dem Objekt `html` auslesen. Beginnen wir mit einem Textinhalt, welcher sich relativ leicht extrahieren lässt. Wir wollen den Titel des Beitrages extrahieren: „*Schämt Euch!*“ – *Deutschland steht vor den Vereinten Nationen am Pranger, weil es die Inklusion an Schulen praktisch verweigert*“. Dabei ist es gar nicht so leicht, einen spezifischen Inhalt wie den Titel zu lokalisieren und auszulesen. Hierfür ist HTML-² und CSS-Selector-Grundlagenwissen³ hilfreich. Die eigentlichen Textinhalte sind nämlich im HTML-Dokument der Webseite hinterlegt. Ist eine Internetseite im Browser geöffnet, so gelangen wir mit einem Rechtsklick i.d.R. zur Option „*Seitenquelltext anzeigen*“ (Abbildung 7). Dies führt uns zum HTML-Dokument der Webseite (Abbildung 8).

²<https://developer.mozilla.org/en-US/docs/Web/HTML>

³https://developer.mozilla.org/en-US/docs/Learn/CSS/Building_blocks/Selectors; „CSS includes a miniature language for selecting elements on a page called CSS selectors. CSS selectors define patterns for locating HTML elements, and are useful for scraping because they provide a concise way of describing which elements you want to extract.“, Quelle: <https://rvest.tidyverse.org/articles/rvest.html>

„Schämt Euch!“ vor den Vereinten Nationen am Pranger, weil es Schulen praktisch

29. August 2023

 Gefällt mir 609



GENF. „Schämt Euch!“ – so heißt es

Aktivisten des Berliner Bündnisses „... Schändliche Inkriminierung vor dem Palais der Vereinten Nationen in Genf platziert haben. Und: „Deutschland verweigert das Menschenrecht auf inklusive Bildung.“ Der Ort des Protests, zu dem Dutzende von

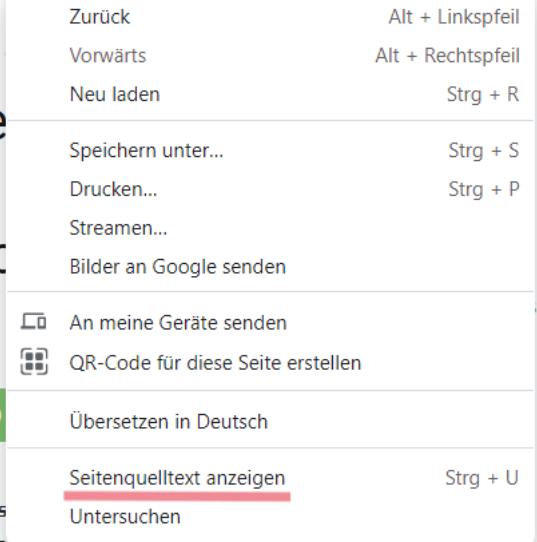


Abbildung 7: Seitenquelltext (HTML) anzeigen

```

1292 <article id="post-132285" class="post-132285 post type-post status-publish format-standard has-post-thumbnail category-leben ca
1293 <div class="td-post-header">
1294
1295 <!-- category --><ul class="td-category"><li class="entry-category"><a href="https://www.news4teachers.de/bildung/lebe
1296 <header class="td-post-title">
1297 <h1 class="entry-title">Schämt Euch!#8220; #8211; Deutschland steht vor den Vereinten Nationen am Pranger,
1298
1299
1300 <div class="td-module-meta-info">
1301 <!-- author --> <!-- date --><span class="td-post-date"><time class="entry-date updated td-m

```

Abbildung 8: HTML-Dokument/Seitenquelltext (Ausschnitt)

Der HTML-Code aus Abbildung 8 ist zwecks besserer Lesbarkeit auch nachfolgend dargestellt:

```

<article id="post-132285" class="post-132285 post type-post status-publish format-standard has-post-thumbnail category-leben ca
<div class="td-post-header">
<!-- category -->
<ul class="td-category">
<li class="entry-category"><a href="https://www.news4teachers.de/bildung/lebe
<li class="entry-category"><a href="https://www.news4teachers.de/bildung/titel
<li class="entry-category"><a href="https://www.news4teachers.de/bildung/wisse
</ul>
<header class="td-post-title">
<h1 class="entry-title">Schämt Euch!#8220; #8211; Deutschland steht v
<div class="td-module-meta-info">

```

```

<!-- author -->
<!-- date -->
<span class="td-post-date">
    <time class="entry-date updated td-module-date" datetime="2023-08-29T1
</span>
<!-- comments -->
<div class="td-post-comments">
    <a href="https://www.news4teachers.de/2023/08/schaemt-euch-deutschland
        <i class="td-icon-comments"></i>150
    </a>
</div>
<!-- views -->
</div>
</header>
</div>
</article>
```

Das HTML-Dokument (Abbildung 8) ist riesig (mehr als 10000 Zeilen) und wir müssen etwas stöbern, um den passenden Webinhalt zu lokalisieren. Wir sehen z.B. in der Zeile 1297, dass der Titel des Beitrages ein h1-HTML-Element⁴ ist (header 1: Überschrift erster Ebene):

```
<h1 class="entry-title">„Schämt Euch!“ - Deutschland steht vor den Verein ...
```

Diese Information benötigen wir, um den Titel des Beitrags gezielt auszulesen. Hierfür nutzen wir den Befehl `html_elements("h1")` und übergeben das Objekt `html` an diesen Befehl.

```
html |> html_elements("h1")

{xml_nodeset (1)}
[1] <h1 class="entry-title">„Schämt Euch!“ - Deutschland steht vor den Verein ...
```

Die Information `<h1 class="entry-title">` ist überflüssig, da wir nur am HTML-Textinhalt interessiert sind. Daher extrahieren wir den reinen Textinhalt, also den Titel, mit dem Befehl `html_text()`

```
html |>
    html_elements("h1") |>
    html_text()

[1] "„Schämt Euch!“ - Deutschland steht vor den Vereinten Nationen am Pranger, weil es die I...
```

Herzlichen Glückwunsch! Somit haben wir erfolgreich alle Informationen von der Webseite extrahiert und eine relevante Textstelle (den Titel) ausgelesen.

⁴https://developer.mozilla.org/en-US/docs/Web/HTML/Element/Heading_Elements

Datenstruktur

Im Seitenquelltext (Abbildung 8) sehen wir, dass anscheinend jeder Beitrag über eine ID verfügt (`id="post-132285"`). Wenn wir in unserem zukünftigen Datensatz mehrere Beiträge abspeichern wollen, dann wird eine ID-Variable zwecks Unterscheidung der Beiträge eine hilfreiche Sache sein. Tabelle 1 ist eine erste Idee bezüglich einer möglichen/sinnvollen Datenstruktur. Bei dieser Datenstruktur ignorieren wir der Einfachheit halber ein paar relevante Webinhalte (z.B. Kommentare und Anzahl der Likes).

Tabelle 1: Erste Idee b

id	datum	titel
132285	29. August 2023	'Schämt Euch!' – Deutschland steht vor den Vereinten Nationen am Pranger...
...
...

Weitere Web-Scraping-Schritte

Um die Datenstruktur aus Tabelle 1 zu realisieren, müssen wir nun die ID des Beitrags, das Erscheinungsdatum, die Zusammenfassung und den eigentlichen Haupttext des Beitrages auslesen (den Titel haben wir ja bereits erfolgreich extrahiert). Beginnen wir mit der ID.

* ID

In Abbildung 8 sehen wir, dass die ID (`id="post-132285"`) ein Attribut⁵ eines HTML-Elements ist (HTML-Element: `article`⁶):

```
<article id="post-132285" class="post-132285 post type-post status-publish format-standard">
```

Daher übergeben wir das Objekt `html` zunächst an den Befehl `html_elements("article")` und dann an den Befehl `html_attr("id")` zwecks Auslesung der ID.

```
html |>  
  html_elements("article") |>  
  html_attr("id")
```

```
[1] "post-132285"
```

⁵https://developer.mozilla.org/en-US/docs/Web/HTML/Global_attributes/id

⁶<https://developer.mozilla.org/en-US/docs/Web/HTML/Element/article>

Die ID des Beitrags erscheint mit dem Präfix "post-", eine nicht notwendigerweise nützliche Information. Das Präfix entfernen wir daher mit dem Befehl `str_remove("post-")` und überführen die ID mit dem Befehl `as.numeric()` in ein nummerisches Format. Somit erhalten wir die nummerische ID 132285.

```
html |>
  html_elements("article") |>
  html_attr("id") |>
  str_remove("post-") |> # R-Zusatzpaket stringr (tidyverse)
  as.numeric()
```

```
[1] 132285
```

* Erscheinungsdatum

Fahren wir fort mit dem Auslesen des Erscheinungsdatums des Beitrages. Im Quelltext (Abbildung 8, Zeile 1301) erscheint folgende Information:

```
<span class="td-post-date"><time class="entry-date updated td-module-date" datetime="2023-
```

Wir sehen, dass das Datum ein HTML-Element ist, nämlich ein `time`-Element⁷. Dieses `time`-Element ist innerhalb eines `span`-Elements⁸ geschachtelt. Wir können hier also von einer hierarchischen Schachtelung der HTML-Elemente sprechen (`span, time`). Entsprechend erfolgt die Extraktion des Datums mit der Übergabe des Objektes `html`, zunächst an den Befehl `html_elements("span")`, und anschließend an den Befehl `html_elements("time")`.

```
html |>
  html_elements("span") |>
  html_elements("time")

{xml_nodeset (7)}
[1] <time class="entry-date updated td-module-date" datetime="2023-08-29T12:4 ...
[2] <time class="entry-date updated td-module-date" datetime="2023-09-18T14:0 ...
[3] <time class="entry-date updated td-module-date" datetime="2023-09-17T17:3 ...
[4] <time class="entry-date updated td-module-date" datetime="2023-09-18T15:1 ...
[5] <time class="entry-date updated td-module-date" datetime="2023-09-19T12:0 ...
[6] <time class="entry-date updated td-module-date" datetime="2023-09-19T10:5 ...
[7] <time class="entry-date updated td-module-date" datetime="2023-09-18T19:2 ...
```

Das Ergebnis ist aber nicht ganz befriedigend, da mehrere Datumsangaben extrahiert worden sind, unter anderem das gewünschte Erscheinungsdatum des Beitrages (2023-08-29), aber auch

⁷<https://developer.mozilla.org/en-US/docs/Web/HTML/Element/time>

⁸<https://developer.mozilla.org/en-US/docs/Web/HTML/Element/span>

andere, nicht relevante Datumsangaben (z.B. 2023-09-17), welche ebenfalls auf der Webseite erscheinen (Abbildung 9).

Oft gelesen



Vereinte Nationen rügen Deutschland wegen Stillstand bei der Inklusion – Land...

17. September 2023

30

Abbildung 9: Verweis auf einen anderen Beitrag mit nicht relevanter Datumsangabe

Wir müssen daher beim Auslesen noch genauer die hierarchische Position des Erscheinungsdatums definieren. Ein Blick auf Abbildung 8 offenbart, dass die beiden HTML-Elemente `span` und `time` innerhalb des bereits bekannten HTML-Elements `article` geschachtelt sind. Diese hierarchische Schachtelung (`article`, `span`, `time`) muss daher beim Auslesen des Erscheinungsdatums beachtet werden:

```
html |>
  html_elements("article") |>
  html_elements("span") |>
  html_elements("time")

{xml_nodeset (1)}
[1] <time class="entry-date updated td-module-date" datetime="2023-08-29T12:4 ...
```

Das Erscheinungsdatum ist in diesem Falle das einzige `time`-Element innerhalb des `article`-Elements. Daher führt auch das Weglassen des `span`-Elements und somit die Anwendung einer reduzierten hierarchische Schachtelung der HTML-Elemente (`article`, `time`) zum gewünschten Erfolg:

```

html |>
  html_elements("article") |>
  html_elements("time")

{xml_nodeset (1)}
[1] <time class="entry-date updated td-module-date" datetime="2023-08-29T12:4 ...

```

Auch bei der Datumsangabe wollen wir uns auf die wesentliche Information fokussieren und extrahieren daher die reine Datumsangabe, die dem Attribut "datetime" zugeordnet ist. Die Befehlskette wird daher um den Befehl "html_attr("datetime")" ergänzt:

```

html |>
  html_elements("article") |>
  html_elements("time") |>
  html_attr("datetime")

[1] "2023-08-29T12:46:06+02:00"

```

Die Datumsangabe ("2023-08-29T12:46:06+02:00") beinhaltet auch eine für uns nicht relevante Zeitangabe, also die genaue Uhrzeit der Beitragserscheinung (T12:46:06+02:00). Die ersten 10 Zeichen (inkl. Bindestriche: JJJJ-MM-TT/2023-08-29) beibehalten die relevante Datumsangabe. Die nicht relevante Zeitangabe entfernen wir, indem wir lediglich die ersten 10 Zeichen der Datumsangabe beibehalten. Hierfür ergänzen wir die Befehlskette um den Befehl `str_sub(end = 10)`.

```

html |>
  html_elements("article") |>
  html_elements("time") |>
  html_attr("datetime") |>
  str_sub(end = 10) # R-Zusatzpaket stringr (tidyverse)

[1] "2023-08-29"

```

- * Zusammenfassung

Literatur

Lüke, Timo, Matthias R. Hastall, Christian Marschler, und Michael Grosche. 2014. „Was liest man über Inklusion?“ <https://doi.org/10.6084/M9.FIGSHARE.1252227>.
 News4teachers. 2022. „Über uns“. *News4teachers*. <https://www.news4teachers.de/uberuns/>.