



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

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

2023-09-29

Preprint in Progress

Inhaltsverzeichnis

1 Vorwort	4
2 Einleitung	4
3 Ziele	6
3.1 Allgemeine Zielsetzung	6
3.2 Zielsetzung mit R: Web-Scraping und Text-Mining	6
4 News4teachers: Online-Nachrichten- und Diskussionsseite für Lehrkräfte	6
4.1 Inhalte von News4teachers und potenzielle Leserschaft aus Lehrkräften	7
4.1.1 Kommentare und Diskussionen	8
5 Explorative Forschungsfragen	9
6 Web-Scraping	9
6.1 R-Zusatzpakete	9
6.1.1 R-Zusatzpaket <i>rvest</i>	9
6.1.2 R-Zusatzpaket <i>tidyverse</i>	10
6.2 Struktur und Inhalte der Webseite	10
6.3 Erster Web-Scraping-Versuch	14
6.3.1 Datenstruktur	17
6.3.2 Weitere Web-Scraping-Schritte	17
6.3.2.1 ID	17
6.3.2.2 Link	18
6.3.2.3 Erscheinungsdatum	19
6.3.2.4 Ort der Berichterstattung	22
6.3.2.5 Zusammenfassung des Beitrages	23
6.3.2.6 Haupttext	24
6.3.2.7 Verfasser	31
7 Weiteres	31

1 Vorwort

TBA

2 Einleitung

Mein wertgeschätzter Kollege Timo Lüke¹ 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).

¹<https://timolueke.de/>

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 verbriefte 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 nach Beitragstyp und Publikation



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 (Hrsg.), *Inklusion von Schülerrinnen 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?

3 Ziele

3.1 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).

3.2 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. Eine Einführung in das Thema Web-Scraping mit R bieten Wickham, Cetinkaya-Rundel, und Grolemund (2023, Kap. 25)².
- **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 eines Textinhalts bzw. die im Text vermittelte subjektive Meinung analysiert 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. Eine Einführung in das Thema Text-Mining mit R bieten Silge und Robinson (2017).

4 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](#).

²<https://r4ds.hadley.nz/webscraping>

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)

Die Redaktion besteht aus folgenden Personen (News4teachers 2023a): Anna Hückelheim, Sonja Mankowsky, Laura Millmann, Nina Odenius, Thomas Zab und Milla Priboschek (Podcast-Redaktion).

4.1 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/>). 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 (Bundesagentur für Arbeit 2022). 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.

AUFGREIFEN: [<https://www.news4teachers.de/2021/12/liebe-leserin-lieber-leser-ein-wort-zum-jahreswechsel-in-eigener-sache/>]

4.1.1 Kommentare und Diskussionen

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 unseriöse 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 (News4teachers 2023b)

5 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?

6 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. altmodischerweise 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. Eine Einführung in das Thema Web-Scraping mit R bieten Wickham, Cetinkaya-Rundel, und Grolemund (2023, Kap. 25)³.

6.1 R-Zusatzpakete

6.1.1 R-Zusatzpaket *rvest*

Für das Web-Scraping nutzen wir nun das R-Zusatzpaket *rvest* (Wickham 2022). 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)
```

³<https://r4ds.hadley.nz/webscraping>



Abbildung 3: Logo des R-Zusatzpaketes *rvest*

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

- <https://rvest.tidyverse.org/>
- <https://r4ds.hadley.nz/webscraping> (Wickham, Cetinkaya-Rundel, und Grolemund 2023, Kap. 25)

6.1.2 R-Zusatzpaket *tidyverse*

Das R-Zusatzpaket *tidyverse* (Wickham u. a. 2019) ist eine Zusammenstellung unterschiedlicher R-Zusatzpakete. Auch das R-Zusatzpaket *rvest* ist Bestandteil des R-Zusatzpakets *tidyverse*. 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)
```

6.2 Struktur und Inhalte der Webseite

Ziel des Web-Scapings 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“ (News4teachers 2023b). Für unsere Forschungsfragen mehr oder weniger interessante Webinhalte sind in den Abbildungen kenntlich gemacht (Abbildung 4, Abbildung 5 und Abbildung 6).

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

157

Anzahl der Kommentare

Zusammenfassung

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.

Externe Inhalte und/oder Bilder

Berliner Bündnis für schulische Inklusion
@bbsinklusion · [Folgen](#)

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

X

Abbildung 4: (a) Beitrag auf der Internetseite News4teachers und Struktur der Webinhalte (News4teachers 2023b)

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 (News4teachers 2023b)

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 (News4teachers 2023b)

[Erläuterungen zu den Abbildungen und Inhalten hinzufügen]

6.3 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-  
  
#html <- read_html("https://www.news4teachers.de/2012/02/im-kern-sind-wir-uns-einig-kein-s  
  
#html <- read_html("https://www.news4teachers.de/2019/03/mobbing-ritual-unter-grundschuele
```

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 HTML⁴ und CSS-Selector-Grundlagenwissen⁵ hilfreich. Die eigentlichen Textinhalte sind nämlich im HTML-Dokument der Webseite hinterlegt (HTML-Quelltext). 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).

⁴ <https://developer.mozilla.org/en-US/docs/Web/HTML>

⁵ https://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 Pranger der Vereinten Nationen in Genf platziert haben. Und: „Deutschland verweigert das Menschenrecht auf inklusive Bildung.“ Der Ort des Protests, zu dem Dutzende von

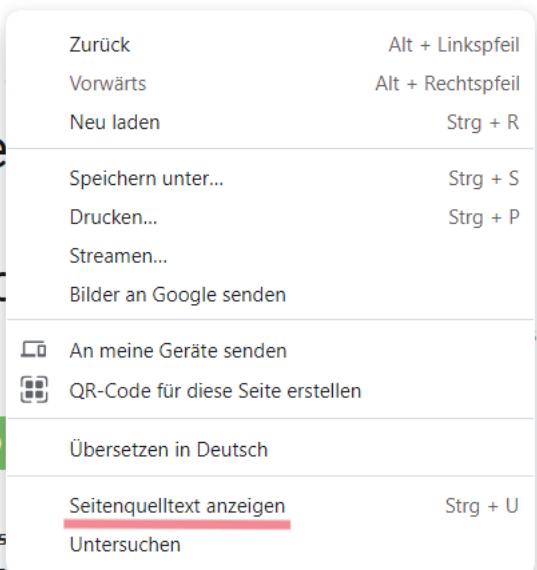


Abbildung 7: Seitenquelltext (HTML) anzeigen

```

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   <!-- category --><ul class="td-category"><li class="entry-category"><a href="https://www.news4teachers.de/bildung/lebe
1295   <header class="td-post-title">
1296     <h1 class="entry-title">Schämt Euch! Deutschland steht vor den Vereinten Nationen am Pranger,
1297
1298
1299
1300
1301   <div class="td-module-meta-info">
      <!-- author -->           <!-- date --><span class="td-post-date"><time class="entry-date updated td-m

```

Abbildung 8: HTML-Dokument/Seitenquelltext (Ausschnitt)

Das HTML-Dokument (Abbildung 8) ist riesig (mehr als 10000 Zeilen) und wir müssen etwas stöbern, um den passenden Webinhalt zu lokalisieren. 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 cat
  <div class="td-post-header">
    <!-- category -->
    <ul class="td-category">
      <li class="entry-category"><a href="https://www.news4teachers.de/bildung/leben">Leben</a></li>
      <li class="entry-category"><a href="https://www.news4teachers.de/bildung/titelthema">Titelthema</a></li>
      <li class="entry-category"><a href="https://www.news4teachers.de/bildung/wissen">Wissen</a></li>
    </ul>
    <header class="td-post-title">
      <h1 class="entry-title">Schämt Euch! Deutschland steht vor den Vereinten Nationen am Pranger</h1>
    </header>
    <div class="td-post-content">
      <p>GENF. „Schämt Euch!“ – so heißt es Aktivisten des Berliner Bündnisses „... Schändliche Inkriminierung vor dem Pranger der Vereinten Nationen in Genf platziert haben. Und: „Deutschland verweigert das Menschenrecht auf inklusive Bildung.“ Der Ort des Protests, zu dem Dutzende von</p>

```

```

<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/schämt-euch-deutschland
            <i class="td-icon-comments"></i>150
        </a>
    </div>
    <!-- views -->
</div>
</header>
</div>
</article>
```

Wir sehen z.B. in der Zeile 1297 (Abbildung 8), dass der Titel des Beitrages ein `h1`-HTML-Element⁶ 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()`. Die Befehlskette wird entsprechend erweitert:

```
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.

⁶https://developer.mozilla.org/en-US/docs/Web/HTML/Element/Heading_Elements

6.3.1 Datenstruktur

Im HTML-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 vorerst ein paar relevante Webinhalte (z.B. Kommentare und Anzahl der Likes).

Tabelle 1: Erste Datenstruktur

id	link	datum	ort	titel
132285	https://...	29. August 2023	GENF	'Schämt Euch!' – Deutschland steht vor den Vereinten Nationen
...
...

HINZUFÜGEN [LINK, LIKES]

6.3.2 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.

[REIHENFOLGE und Vollständigkeit prüfen]

6.3.2.1 ID

In Abbildung 8 sehen wir, dass die ID des Beitrages (`id="post-132285"`) ein Attribut⁷ eines HTML-Elements ist (HTML-Element: `article`⁸):

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

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

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

⁷ https://developer.mozilla.org/en-US/docs/Web/HTML/Global_attributes/id

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

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

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
```

6.3.2.2 Link

Der Beitrag verfügt über einen langen Link:

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

Im HTML-Quelltext ist allerdings auch ein kurzer Link, also ein `shortlink`, ausgewiesen:

```
<link rel='shortlink' href='https://www.news4teachers.de/?p=132285' />
```

Die ID des Beitrags (132285) ist Bestandteil des kurzen Links. Wir können also den ersten Teil des kurzen Links ("`https://www.news4teachers.de/?p=`") mit der ID (132285) verbinden, um den gewünschten Kurzlink zu generieren. Hierfür nutzen wir nach der Auslesung der ID den Befehl `paste0("https://www.news4teachers.de/?p=", .)`. Mit dem magrittr-Pipe-Operator (`%>%`⁹) wird die ID an das zweite Argument des Befehls `paste0("https://www.news4teachers.de/?p=", .)` übergeben, also an die Stelle mit dem Punkt (.). Eine Übergabe an das zweite Argument wäre mit der sogenannten base-Pipe (`|>`) nicht möglich, daher nutzen wir die magrittr-Pipe (`%>%`). Die Befehlskette zur Erstellung des Links gestaltet sich somit folgendermaßen:

```
html |>
  html_elements("article") |>
  html_attr("id") |>
  str_remove("post-") |> # R-Zusatzpaket stringr (tidyverse)
  as.numeric() %>% # Pipe-Operator, R-Zusatzpaket magrittr (tidyverse)
  paste0("https://www.news4teachers.de/?p=", .)
```

```
[1] "https://www.news4teachers.de/?p=132285"
```

⁹<https://magrittr.tidyverse.org/>

6.3.2.3 Erscheinungsdatum

Fahren wir fort mit dem Auslesen des Erscheinungsdatums des Beitrages. Im HTML-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`-Element¹⁰. Dieses `time`-Element ist innerhalb eines `span`-Elements¹¹ geschachtelt. Wir können hier also von einer hierarchischen Schachtelung der HTML-Elemente sprechen (`span` → `time`, Abbildung 9).

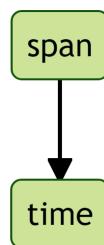


Abbildung 9: Hierarchische Schachtelung der HTML-Elemente `span` und `time`

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="2023-09-26T10:5 ...
[3] <time class="entry-date updated td-module-date" datetime="2023-09-28T12:5 ...
[4] <time class="entry-date updated td-module-date" datetime="2023-09-26T16:2 ...
[5] <time class="entry-date updated td-module-date" datetime="2023-09-29T16:3 ...
[6] <time class="entry-date updated td-module-date" datetime="2023-09-29T13:4 ...
[7] <time class="entry-date updated td-module-date" datetime="2023-09-29T12:5 ...
```

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

¹⁰<https://developer.mozilla.org/en-US/docs/Web/HTML/Element/time>

¹¹<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 10).

Oft gelesen



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

17. September 2023

30

Abbildung 10: 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 (`article -> span -> time`, Abbildung 11).

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

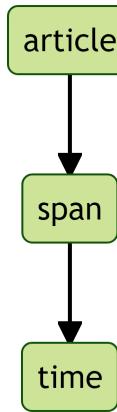


Abbildung 11: Hierarchische Schachtelung der HTML-Elemente `article`, `span` und `time`

einer reduzierten hierarchischen Schachtelung der HTML-Elemente (`article -> time`, Abbildung 12) zum gewünschten Erfolg:

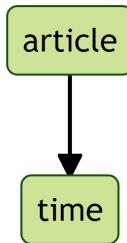


Abbildung 12: Reduzierte hierarchische Schachtelung der HTML-Elemente `article` und `time`

```

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 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"
```

6.3.2.4 Ort der Berichterstattung

Die Ortsangabe ist Bestandteil der Zusammenfassung (siehe Abbildung 4) und die Zusammenfassung ist ein Absatz, i.d.R. der erste Absatz des Beitrages. Für die Extraktion der Ortsangabe ist daher das HTML-Element für Absätze notwendig (`p`¹²; `p` steht für "paragraph"). Dieses HTML-Element (`p`) ist wie gewohnt innerhalb des HTML-Elements `article` geschachtelt. Zur Extraktion des ersten Absatzes wird diesmal der Befehl `html_element("p")` anstatt `html_elements("p")` genutzt. Der Befehl `html_elements("p")` würde alle Absätze des Beitrages extrahieren. Wir benötigen aber nur den ersten Absatz mit der Ortsangabe und daher nutzen wir diesmal den Befehl `html_element("p")` anstatt `html_elements("p")`. Die Befehlskette gestaltet sich daher wie folgt:

```
html |>
  html_elements("article") |>
  html_element("p") |> # html_element anstatt html_elements
  html_text()
```

```
[1] "GENF. „Schämt Euch!“ – so heißt es auf einem Transparent, das Aktivistinnen und Aktivis-
```

genauer: das Engagement, das die Bundesrepublik zeigt, um die UN-Behindertenrechtskonvention

¹²<https://developer.mozilla.org/en-US/docs/Web/HTML/Element/p>

Somit sehen wir den Absatz mit der Ortsangabe. Wir benötigen allerdings nur die Ortsangabe, also das erste Wort des Absatzes. Hinter der gewünschten Ortsangabe steht ein Punkt (GENF.). Mit dem Befehl `str_extract("[^.]+")`¹³ extrahieren wir alle Zeichen vor dem ersten Punkt, also die Ortsangabe GENF. Die Befehlskette gestaltet sich daher wie folgt:

```
html |>
  html_elements("article") |>
  html_element("p") |> # html_element anstatt html_elements
  html_text() |>
  str_extract("[^\\.]+") # R-Zusatzpaket stringr (tidyverse)
```

```
[1] "GENF"
```

6.3.2.5 Zusammenfassung des Beitrages

Wie soeben bei der Extraktion der Ortsangabe erwähnt, ist die Zusammenfassung des Beitrages der erste Absatz des Textes (siehe Abbildung 4). Der erste Absatz wurde soeben folgendermaßen extrahiert:

```
html |>
  html_elements("article") |>
  html_element("p") |> # html_element anstatt html_elements
  html_text()
```

```
[1] "GENF. „Schämt Euch!“ - so heißt es auf einem Transparent, das Aktivistinnen und Aktivis  
genauer: das Engagement, das die Bundesrepublik zeigt, um die UN-Behindertenrechtskonvention
```

Somit erhalten wir die Zusammenfassung mit der Ortsangabe inkl. Punkt (GENF.). Nun wollen wir die überflüssige Ortsangabe entfernen und nur die eigentliche Zusammenfassung beibehalten. Dies erreichen wir mit dem Befehl `str_extract("\\.\\. [\\s](.*)")`. Die regex-Formel "`\\.\\. [\\s](.*)`" hat folgende Bedeutung:

- \\. Suche und extrahiere Zeichen nach dem ersten Punkt (einschließlich des ersten Punktes)
- [\s] Die extrahierten Zeichen können Leerzeichen sein ("s" steht für "space")
- (.*) Extrahiere außerdem alle weiteren Zeichen

Die Befehlskette gestaltet sich daher wie folgt:

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

¹³Bei so einer kryptischen Formel ("[^.]+") handelt es sich um eine sogenannte "regular expression" (regex). Eine Einführung in diese Thematik findet man hier: <https://r4ds.hadley.nz/regexp.html> (Wickham, Cetinkaya-Rundel, und Grolemund 2023, Kap. 16).

```

html_element("p") |> # html_element anstatt html_elements
html_text() |>
str_extract("\\.\\.\\s](.*))" # R-Zusatzpaket stringr (tidyverse)

```

[1] ". „Schämt Euch!" - so heißt es auf einem Transparent, das Aktivistinnen und Aktivisten genauer: das Engagement, das die Bundesrepublik zeigt, um die UN-Behindertenrechtskonvention

Die Ortsangabe (GENF) wurde erfolgreich entfernt. Der Punkt hinter der Ortsangabe (GENF.) wurde allerdings nicht entfernt und bleibt bestehen. Die Zusammenfassung beginnt daher nun mit einem Punkt(.) gefolgt von einem Leerzeichen. Wir entfernen den Punkt und das Leerzeichen ("." ") mit dem Befehl `str_remove(". ")`. Die Befehlskette zur Extraktion der Zusammenfassung gestaltet sich daher folgendermaßen:

```

html |>
html_elements("article") |>
html_element("p") |> # html_element anstatt html_elements
html_text() |>
str_extract("\\.\\.\\s](.*))" |> # R-Zusatzpaket stringr (tidyverse)
str_remove(". ") # R-Zusatzpaket stringr (tidyverse)

```

[1] ". „Schämt Euch!" - so heißt es auf einem Transparent, das Aktivistinnen und Aktivisten genauer: das Engagement, das die Bundesrepublik zeigt, um die UN-Behindertenrechtskonvention

6.3.2.6 Haupttext

Kommen wir nun zum Filetstück, also zum eigentlichen Haupttext des Beitrages. Und wie das so ist beim Filetieren: Es ist gar nicht so trivial! Der Beitragstext besteht aus Absätzen. Also können wir, wie bereits gewohnt, das HTML-Element `p` berücksichtigen. Und dieses HTML-Element `p` ist bekannterweise innerhalb des HTML-Elements `article` geschachtelt:

```

# Seitenquelltext (HTML)
html |>
html_elements("article") |>
html_elements("p")

{xml_nodeset (25)}
[1] <p>GENF. <strong>„Schämt Euch!" - so heißt es auf einem Transparent, das ...
[2] <p dir="ltr" lang="de"><a href="https://twitter.com/hashtag/InklusiveBil ...
[3] <p>- Berliner Bündnis für schulische Inklusion (@bbsinklusion) <a href=" ...
[4] <p><script charset="utf-8" async consent-original-src_="https://platfor ...
[5] <p>Es ist fast schon dreist, wie Deutschland auf die offizielle Staatenp ...
[6] <p>Das Deutsche Institut für Menschenrechte, das vom Bundestag mit dem M ...

```

[7] <p>Auch in einer gemeinsamen Stellungnahme von einem Bündnis deutscher N ...
[8] <p>„In keinem Bildungsbereich – von der Kita über Schule, Ausbildung und ...
[9] <p>Die Ausführungen der Bundesregierung im Staatenbericht, so heißt es w ...
[10] <p>Sonderpädagoginnen und -pädagogen würden immer noch weitestgehend für ...
[11] <p>Die Einführung inklusiver Bildung in Regelschulen sei von erheblichem ...
[12] <p dir="ltr" lang="de"><a href="https://twitter.com/hashtag/WirFahrenNac ...
[13] <p>- mittendrin e.V. (@mittendrinev) <a href="https://twitter.com/mitten ...
[14] <p><script charset="utf-8" async consent-original-src_="https://platfor ...
[15] <p dir="ltr" lang="de">Noch 1 Tag bis zur <a href="https://twitter.com/h ...
[16] <p>Janine aus Berlin ist dabei!
<a href="https://twitter.com/hashtag/ ...
[17] <p>- Berliner Bündnis für schulische Inklusion (@bbsinklusion) <a href=" ...
[18] <p><script charset="utf-8" async consent-original-src_="https://platfor ...
[19] <p>Der Rechtsanspruch auf inklusive Schulbildung sei in den meisten Bund ...
[20] <p><a href="https://www.vdk.de/deutscher-behindertenrat/mime/00134312D16 ...
...
...

```
# Text
html |>
  html_elements("article") |>
  html_elements("p") |>
  html_text()
```

[1] "GENF. „Schämt Euch!" – so heißt es auf einem Transparent, das Aktivistinnen und Aktivier genauer: das Engagement, das die Bundesrepublik zeigt, um die UN-Behindertenrechtskonvention
[2] "#InklusiveBildungJETZT\nWir sind auf dem Place de Nations angekommen! Die #Staatenprüfung
[3] "- Berliner Bündnis für schulische Inklusion (@bbsinklusion) August 29, 2023"
[4] " Der offizielle Beitrag Deutschlands fällt dünn aus – bezeichnend für das, was sich in
[5] "Es ist fast schon dreist, wie Deutschland auf die offizielle Staatenprüfung, mit der d
[6] "Das Deutsche Institut für Menschenrechte, das vom Bundestag mit dem Mandat ausgestatte
[7] "Auch in einer gemeinsamen Stellungnahme von einem Bündnis deutscher Nichtregierungsorg
[8] „In keinem Bildungsbereich – von der Kita über Schule, Ausbildung und Hochschule bis z
liegt eine verbindliche Gesamtstrategie (Ziele, Zeitplan, Qualitätskriterien, Ressourcen) vo
[9] "Die Ausführungen der Bundesregierung im Staatenbericht, so heißt es weiter, „betrachte
[10] "Sonderpädagoginnen und -pädagogen würden immer noch weitestgehend für die Arbeit in Fö
[11] "Die Einführung inklusiver Bildung in Regelschulen sei von erheblichem Personalmangel g
[12] "#WirFahrenNachGenf und die Vorhut ist schon da. Morgen beginnt das Protestcamp vor der
[13] "- mittendrin e.V. (@mittendrinev) August 28, 2023"
[14] " Alles in allem erscheint die Inklusion als Mogelpackung. „Hohe ‚Inklusionsraten‘ der
[15] "Noch 1 Tag bis zur #Staatenprüfung #UNBRK in Genf!"
[16] "Janine aus Berlin ist dabei!#WirFahrenNachGenf - Kommt am 29. & 30.8. mit auf den Pla
[17] "- Berliner Bündnis für schulische Inklusion (@bbsinklusion) August 28, 2023"
[18] " Kein Wunder: „‘Inklusive Schulen‘ sind nicht flächendeckend vorhanden und beschränken
[19] "Der Rechtsanspruch auf inklusive Schulbildung sei in den meisten Bundesländern mit (Re
[20] "Hier geht es zum vollständigen „Gemeinsamen Bericht der Zivilgesellschaft zum 2. und 3
[21] "Der Beitrag wird auch auf der Facebook-Seite von News4teachers heiß diskutiert."

```
[22] ""
[23] "Eltern protestieren vor den Vereinten Nationen gegen Deutschland –
weil es die Inklusion in Schulen schleifen lässt"
[24] ""
[25] " "
```

Mit der obigen Befehlskette werden 25 Absätze extrahiert. Aber nur 11 Absätze gehören zum eigentlichen Haupttext (4 bis 11, 14, 18 und 19). Es ist also auch eine Menge Gedöns dabei, also primär nicht relevante Informationen, verstreut innerhalb des Beitrages, z.B. Verweise auf externe Quellen:

```
[2] "#InklusiveBildungJETZT\nWir sind auf dem Place de Nations angekommen!
Die #Staatenpruefung für Deutschland beginnt in Kürze.#WirFahrenNachGenf
pic.twitter.com/XoBPh69iU0"

[20] "Hier geht es zum vollständigen „Gemeinsamen Bericht der Zivilgesellschaft
zum 2. und 3. Bericht der Bundesregierung zur Umsetzung der UN-Behindertenrechtskonvention
durch Deutschland"."

[21] "Der Beitrag wird auch auf der Facebook-Seite von News4teachers heiß
diskutiert."
```

Wir müssen diese nicht relevanten Absätze also mittels einer geeigneten Systematik entfernen. Woran erkennen wir die nicht relevanten Absätze? Es sind Verweise auf externe Quellen und diese Verweise enthalten Links auf externe Quellen. Im nächsten Schritt wollen wir daher Absätze mit Links entfernen. Ob in einem Absatz ein Link vorhanden ist, erkennen wir am HTML-Element `a`¹⁴ (`a` steht für “*anchor*”), z.B. beim Verweis auf externe Inhalte bei Twitter (Abbildung 4 und Abbildung 5):

```
"<p dir=\"ltr\" lang=\"de\"><a href=\"https://twitter.com/hashtag/InklusiveBildungJETZT?src=
sind auf dem Place de Nations angekommen! Die <a href=\"https://twitter.com/hashtag/Staatenp
für Deutschland beginnt in Kürze.<a href=\"https://twitter.com/hashtag/WirFahrenNachGenf?src
<a href=\"https://t.co/XoBPh69iU0\">pic.twitter.com/XoBPh69iU0</a></p>"
```

Das Ende eines Links wird also immer mit dem HTML-Element `` gekennzeichnet sein. Wir wollen nun alle Absätze entfernen, die das HTML-Element `` enthalten¹⁵. Hierfür speichern wir vorübergehend alle Absätze im HTML-Format als Objekt `all_p` ab. Wir überführen alle Absätze (`all_p`) mit dem Befehl `as.character()` ins reine Textformat und überprüfen anschließend mit dem Befehl `str_detect("")`, ob Verlinkungen in den Absätzen vorhanden sind. Das Ergebnis ist ein TRUE-FALSE-Aussage für jeden Absatz (TRUE: Link vorhanden; FALSE: Link nicht vorhanden). Diese TRUE-FALSE-Aussage speichern wir als Objekt `link` ab und lassen uns abschließend anzeigen, welche Absätze keine Links enthalten (`all_p[!link] |> html_text()`):

¹⁴<https://developer.mozilla.org/en-US/docs/Web/HTML/Element/a>

¹⁵Bei diesem Arbeitsschritt habe ich die R-Community um Unterstützung gebeten: <https://stackoverflow.com/questions/77152943/>

```

all_p <-
  html |>
  html_elements("article") |>
  html_elements("p")

all_p |>
  as.character() |>
  str_detect("</a>") # R-Zusatzpaket stringr (tidyverse)

```

```

[1] FALSE TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE
[13] TRUE FALSE TRUE TRUE TRUE FALSE FALSE TRUE TRUE TRUE TRUE FALSE
[25] FALSE

```

```

link <-
  all_p |>
  as.character() |>
  str_detect("</a>") # R-Zusatzpaket stringr (tidyverse)

all_p[!link] |>
  html_text()

```

```

[1] "GENF. „Schämt Euch!" - so heißt es auf einem Transparent, das Aktivistinnen und Aktivisten genauer: das Engagement, das die Bundesrepublik zeigt, um die UN-Behindertenrechtskonvention zu respektieren. „Es ist fast schon dreist, wie Deutschland auf die offizielle Staatenprüfung, mit der die Welt überzeugt werden soll, reagiert.“ Auch in einer gemeinsamen Stellungnahme von einem Bündnis deutscher Nichtregierungsorganisationen („In keinem Bildungsbereich - von der Kita über Schule, Ausbildung und Hochschule bis zur Arbeitswelt - liegt eine verbindliche Gesamtstrategie (Ziele, Zeitplan, Qualitätskriterien, Ressourcen) vor.“) wird kritisiert, dass die Ausführungen der Bundesregierung im Staatenbericht, so heißt es weiter, „betrachtet werden müssen.“ Sonderpädagoginnen und -pädagogen würden immer noch weitestgehend für die Arbeit in Förderzentren eingesetzt. „Die Einführung inklusiver Bildung in Regelschulen sei von erheblichem Personalmangel geprägt.“ Alles in allem erscheint die Inklusion als Mogelpackung. „Hohe ‚Inklusionsraten‘ der Schulen sind nicht flächendeckend vorhanden und beschränken sich auf wenige Bundesländer.“ Kein Wunder: „Inklusive Schulen“ sind nicht flächendeckend vorhanden und beschränken sich auf wenige Bundesländer.“ Der Rechtsanspruch auf inklusive Schulbildung sei in den meisten Bundesländern mit (Rechts-)Vorwürfen beladen.“
[12] ""
[13] ""

```

Das Ergebnis ist schon ziemlich befriedigend: Absätze mit Links wurden entfernt. Aber wir haben auch einen inhaltlich relevanten Absatz entfernt. Ein Absatz enthält nämlich einen Link und ist zugleich inhaltlich relevant (Abbildung 13).

Das Deutsche Institut für Menschenrechte, das vom Bundestag mit dem Mandat ausgestattet wurde, den Vereinten Nationen offiziell zu berichten, fällt in seinem Gutachten ein vernichtendes Urteil ([News4teachers berichtete](#)): „In Deutschland herrscht in der Politik und auch in weiten Teilen der Gesellschaft ein verfehltes Inklusionsverständnis. So wird die Mehrheit der Kinder mit Behinderungen weiterhin nicht inklusiv beschult und wächst ohne schulischen Kontakt zu nichtbehinderten Kindern auf. Das Ziel einer inklusiven Gesellschaft ist so nicht zu erfüllen.“

Abbildung 13: Inhaltlich relevanter Absatz mit Link (“News4teachers berichtete”)

Wir müssen unsere Systematik daher erweitern und entsprechende Absätze mit “*News4teachers berichtete*”-Links beibehalten, da diese Absätze inhaltlich relevant erscheinen. Dies erreichen wir, in dem wir alle Absätze (`all_p`) dahingehend überprüfen, ob die Zeichenabfolge “*News4teachers berichtete*” in den Absätzen vorhanden ist (`str_detect("News4teachers berichtete")`). Das Ergebnis dieser Überprüfung ist wieder eine TRUE-FALSE-Aussage für jeden Absatz (TRUE: “*News4teachers berichtete*”-Link vorhanden; FALSE: “*News4teachers berichtete*”-Link nicht vorhanden). Diese TRUE-FALSE-Aussage speichern wir als Objekt `n4t_link` ab und lassen uns abschließend Absätze anzeigen, welche “*News4teachers berichtete*”-Links enthalten, aber keine anderen Links enthalten (`all_p[n4t_link | !link] |> html_text()`):

```
all_p |>
  as.character() |>
  str_detect("News4teachers berichtete") # R-Zusatzpaket stringr (tidyverse)
```

```
[1] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[13] FALSE FALSE
[25] FALSE
```

```
n4t_link <-
  all_p |>
  as.character() |>
  str_detect("News4teachers berichtete") # R-Zusatzpaket stringr (tidyverse)

all_p[!link | n4t_link] |>
  html_text()
```

```
[1] "GENF. „Schämt Euch!“ - so heißt es auf einem Transparent, das Aktivistinnen und Aktivisten genauer: das Engagement, das die Bundesrepublik zeigt, um die UN-Behindertenrechtskonvention
[2] " Der offizielle Beitrag Deutschlands fällt dünn aus - bezeichnend für das, was sich in
[3] "Es ist fast schon dreist, wie Deutschland auf die offizielle Staatenprüfung, mit der d
[4] "Das Deutsche Institut für Menschenrechte, das vom Bundestag mit dem Mandat ausgestatte
```

```

[5] "Auch in einer gemeinsamen Stellungnahme von einem Bündnis deutscher Nichtregierungsorga
[6] „In keinem Bildungsbereich - von der Kita über Schule, Ausbildung und Hochschule bis zu
liegt eine verbindliche Gesamtstrategie (Ziele, Zeitplan, Qualitätskriterien, Ressourcen) vor
[7] „Die Ausführungen der Bundesregierung im Staatenbericht, so heißt es weiter, „betrachte
[8] „Sonderpädagoginnen und -pädagogen würden immer noch weitestgehend für die Arbeit in Fö
[9] „Die Einführung inklusiver Bildung in Regelschulen sei von erheblichem Personalmangel ge
[10] „ Alles in allem erscheint die Inklusion als Mogelpackung. „Hohe ‚Inklusionsraten‘ der
[11] „ Kein Wunder: „Inklusive Schulen“ sind nicht flächendeckend vorhanden und beschränken
[12] „Der Rechtsanspruch auf inklusive Schulbildung sei in den meisten Bundesländern mit (Re
[13] ""
[14] ""

```

Das sieht schon ziemlich gut aus. Der inhaltlich relevante Absatz mit „*News4teachers berichtete*“-Link (Abbildung 13) wurde nun beibehalten ([4] „Das Deutsche Institut für Menschenrechte ...“). Aber wir haben zwei Absätze ohne sinnvollen Inhalt ([13] „“ und [14] „“), da diese Absätze keine Zeichen oder ausschließlich Leerzeichen enthalten. Wir wollen daher Absätze, die keine Zeichen oder ausschließlich Leerzeichen enthalten, entfernen. Wir Überprüfen daher, ob der Textinhalt der Absätze (`all_p[!link | n4t_link] |> html_text()`) keine Zeichen oder ausschließlich Leerzeichen enthält (`str_detect("^\\s*$")`). Die regex-Formel `"^\\s*$"` hat folgende Bedeutung:

- ^ Beginne die Suche am Anfang des Absatzes
- \s* Suche Leerzeichen bzw. keine Zeichen ("s" steht für "space")
- \$ Führe diese Suche bis zum Ende des Absatzes durch

Das Ergebnis der Überprüfung ist erneut eine TRUE-FALSE-Aussage für jeden Absatz (TRUE: keine Zeichen oder ausschließlich Leerzeichen). Diese TRUE-FALSE-Aussage speichern wir als Objekt `spaces` ab und lassen uns abschließend Absätze anzeigen, welche nicht ausschließlich aus Leerzeichen oder keinen Zeichen bestehen (`.[!spaces]`). Der Punkt(.) symbolisiert die Absätze, welche zuvor mittels `all_p[!link | n4t_link] |> html_text()` erzeugt worden sind. Bei der Übergabe dieser Absätze nutzen wir, wie zuvor beim Auslesen des Links (Kapitel 6.3.2.1), die magrittr-Pipe (%>%), da nur diese Pipe, und nicht die base-Pipe (|>), eine Übergabe an eine Bedingung in eckigen Klammern ermöglicht (`.[!spaces]`):

```

all_p[!link | n4t_link] |>
  html_text() |>
  str_detect("^\\s*$") # R-Zusatzpaket stringr (tidyverse)

```

```

[1] FALSE FALSE
[13] TRUE  TRUE

```

```

spaces <-
  all_p[!link | n4t_link] |>
  html_text() |>

```

```

str_detect("^\s*$") # R-Zusatzpaket stringr (tidyverse)

all_p[!link | n4t_link] |>
  html_text() %>% # Pipe-Operator, R-Zusatzpaket magrittr (tidyverse)
  .[!spaces]

```

[1] "GENF. „Schämt Euch!" - so heißt es auf einem Transparent, das Aktivistinnen und Aktivisten genauer: das Engagement, das die Bundesrepublik zeigt, um die UN-Behindertenrechtskonvention zu unterstützen.
[2] "Der offizielle Beitrag Deutschlands fällt dünn aus - bezeichnend für das, was sich in den Berichten abzeichnet.
[3] "Es ist fast schon dreist, wie Deutschland auf die offizielle Staatenprüfung, mit der die Organisationen der Menschenrechte (OHR) reagieren.
[4] "Das Deutsche Institut für Menschenrechte, das vom Bundestag mit dem Mandat ausgestattet wurde, hat eine verbindliche Gesamtstrategie (Ziele, Zeitplan, Qualitätskriterien, Ressourcen) vorgelegt.
[5] "Auch in einer gemeinsamen Stellungnahme von einem Bündnis deutscher Nichtregierungsorganisationen (NRO) wird die Inklusion als zentrale Prinzipien der Bildungspolitik hervorgehoben.
[6] „In keinem Bildungsbereich - von der Kita über Schule, Ausbildung und Hochschule bis zur beruflichen Bildung - liegt eine verbindliche Gesamtstrategie (Ziele, Zeitplan, Qualitätskriterien, Ressourcen) vor.
[7] "Die Ausführungen der Bundesregierung im Staatenbericht, so heißt es weiter, „betrachten die Inklusion als zentrale Prinzipien der Bildungspolitik.“
[8] "Sonderpädagoginnen und -pädagogen würden immer noch weitestgehend für die Arbeit in Förderzentren und -einrichtungen eingesetzt.
[9] "Die Einführung inklusiver Bildung in Regelschulen sei von erheblichem Personalmangel geprägt.
[10] "Alles in allem erscheint die Inklusion als Mogelpackung. „Hohe ‚Inklusionsraten‘ der Schulen sind nicht flächendeckend vorhanden und beschränken sich auf wenige Bundesländer.“
[11] "Kein Wunder: „‘Inklusive Schulen‘ sind nicht flächendeckend vorhanden und beschränken sich auf wenige Bundesländer.“
[12] "Der Rechtsanspruch auf inklusive Schulpflicht ist in den meisten Bundesländern mit (Rechts-)streitigkeiten verbunden."

Fast geschafft! Wir müssen nur noch den ersten Absatz entfernen (.[-1]), da der erste Absatz die Zusammenfassung mit Ortsangabe darstellt ([1] "GENF. „Schämt Euch!" - so heißt es auf einem Transparent...") und daher nicht zum Haupttext gezählt werden kann. Dies ist daher die schlussendliche Befehlskette:

```

all_p[!link | n4t_link] |>
  html_text() %>% # Pipe-Operator, R-Zusatzpaket magrittr (tidyverse)
  .[!spaces] %>% # Pipe-Operator, R-Zusatzpaket magrittr (tidyverse)
  .[-1]

```

[1] "Der offizielle Beitrag Deutschlands fällt dünn aus - bezeichnend für das, was sich in den Berichten abzeichnet.
[2] "Es ist fast schon dreist, wie Deutschland auf die offizielle Staatenprüfung, mit der die Organisationen der Menschenrechte (OHR) reagieren.
[3] "Das Deutsche Institut für Menschenrechte, das vom Bundestag mit dem Mandat ausgestattet wurde, hat eine verbindliche Gesamtstrategie (Ziele, Zeitplan, Qualitätskriterien, Ressourcen) vorgelegt.
[4] "Auch in einer gemeinsamen Stellungnahme von einem Bündnis deutscher Nichtregierungsorganisationen (NRO) wird die Inklusion als zentrale Prinzipien der Bildungspolitik hervorgehoben.
[5] „In keinem Bildungsbereich - von der Kita über Schule, Ausbildung und Hochschule bis zur beruflichen Bildung - liegt eine verbindliche Gesamtstrategie (Ziele, Zeitplan, Qualitätskriterien, Ressourcen) vor.
[6] "Die Ausführungen der Bundesregierung im Staatenbericht, so heißt es weiter, „betrachten die Inklusion als zentrale Prinzipien der Bildungspolitik.“
[7] "Sonderpädagoginnen und -pädagogen würden immer noch weitestgehend für die Arbeit in Förderzentren und -einrichtungen eingesetzt.
[8] "Die Einführung inklusiver Bildung in Regelschulen sei von erheblichem Personalmangel geprägt.
[9] "Alles in allem erscheint die Inklusion als Mogelpackung. „Hohe ‚Inklusionsraten‘ der Schulen sind nicht flächendeckend vorhanden und beschränken sich auf wenige Bundesländer.“
[10] "Kein Wunder: „‘Inklusive Schulen‘ sind nicht flächendeckend vorhanden und beschränken sich auf wenige Bundesländer.“
[11] "Der Rechtsanspruch auf inklusive Schulpflicht ist in den meisten Bundesländern mit (Rechts-)streitigkeiten verbunden."

6.3.2.7 Verfasser

TBA

7 Weiteres

Evtl. interessant für das Auslesen der Kommentare:

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

Literatur

- Bundesagentur für Arbeit. 2022. *Statistik der Bundesagentur für Arbeit. Berichte: Blickpunkt Arbeitsmarkt (Online-Bericht) – Akademiker/-innen.* <https://statistik.arbeitsagentur.de/DE/Statischer-Content/Statistiken/Themen-im-Fokus/Berufe/AkademikerInnen/Berufsgruppen/Generische-Publikationen/2-8-Lehrkraefte.pdf?blob=publicationFile&v=3>.
- 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/uber-uns/>.
- . 2023a. „Impressum“. *News4teachers.* <https://www.news4teachers.de/impressum/>.
- . 2023b. „„Schämt euch!“ – Deutschland steht vor den Vereinten Nationen am Pranger, weil es die inklusion an Schulen Praktisch Verweigert“. *News4teachers.* <https://www.news4teachers.de/2023/08/schaemt-euch-deutschland-steht-vor-den-vereinten-nationen-am-pranger-weil-es-die-inklusion-an-schulen-verweigert/>.
- Silge, Julia, und David Robinson. 2017. *Text Mining with R*. Sebastopol, CA: O'Reilly Media. <https://www.tidytextmining.com/>.
- Wickham, Hadley. 2022. „rvest: Easily Harvest (Scrape) Web Pages“. <https://CRAN.R-project.org/package=rvest>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, u. a. 2019. „Welcome to the {tidyverse}“ 4: 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, Mine Cetinkaya-Rundel, und Garrett Grolemund. 2023. *R for data science*. 2. Aufl. Sebastopol, CA: O'Reilly Media. <https://r4ds.hadley.nz/>.