

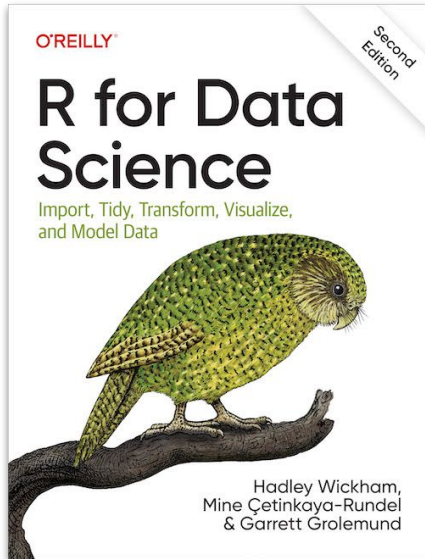


SEARCHING AND TRANSFORMING TEXT

R & stringr

- Why is text different?
- Working with text data using `{stringr}`
 - Searching
 - Extract matches
 - Replace matches
 - Splitting, substrings, and concatenating
 - Case conversion
 - Trimming

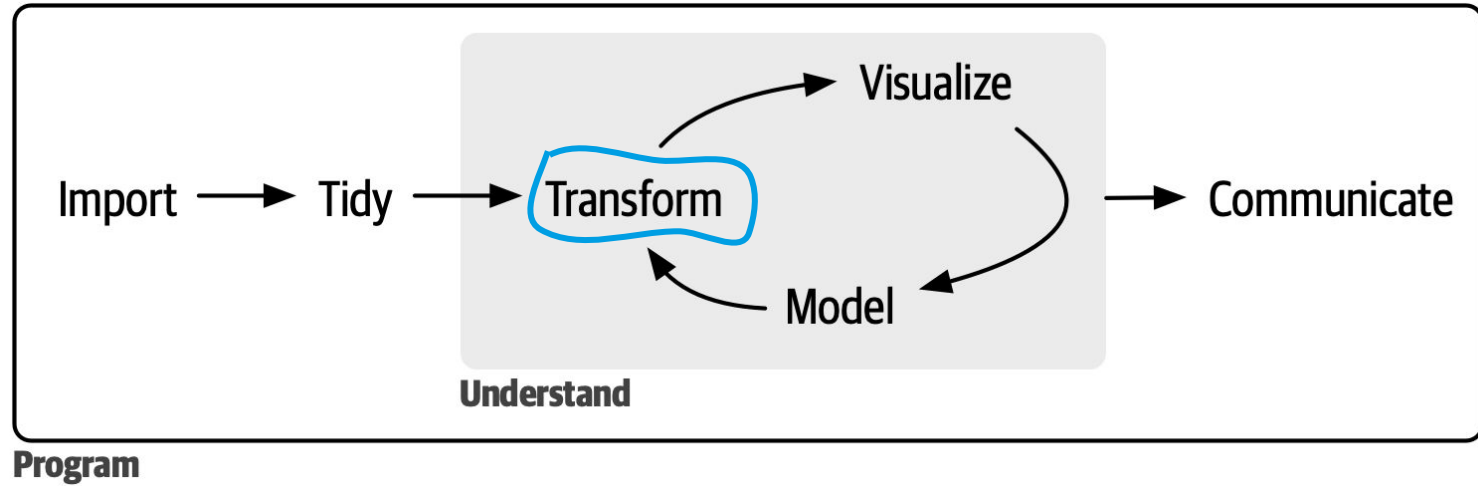
RECOMMENDED LITERATURE



Wickham, Hadley, and Garrett Golemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. 2nd edition, O'Reilly, 2023. Available online: <https://r4ds.hadley.nz/>

- Chapter 14 - Strings
- Chapter 15 - Regular Expressions

WHERE ARE WE?



Source: Wickham, Hadley, and Garrett Golemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. First edition, O'Reilly, 2016. URL: <https://r4ds.hadley.nz/diagrams/data-science/base.png>

WHY IS TEXT DIFFERENT?

WHY IS TEXT DIFFERENT?

STRUCTURED VS UNSTRUCTURED

Tweets in a Spreadsheet					
File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive					
screen_name					
A	B	C	D	E	
screen_name	lang	created_at	is_retweet	retweeted_user	text
berlinlieblich	en	2020-04-29T15:37:40.000+0000	TRUE	MayorofLondon	RT @MayorofLondon: Thank you @RegBerlin Mayor Müller for convening Mayors from across the world today to share experiences of managing this...
berlinlieblich	de	2020-04-29T14:35:31.000+0000	FALSE	null	@biko3467 @Danke_Tegel Aber hallo.
berlinlieblich	de	2020-04-29T20:14:24.000+0000	FALSE	null	Die Berliner sind mit dem Senat zufrieden und die Mehrheit für Mitte-Links steht. Was will man mehr ...https://t.co/Cld5BgUdc3
berlinlieblich	de	2020-04-29T16:00:29.000+0000	FALSE	null	Bei der @Linksfraktion gibt es gleich nebenan bei Facebook einen spannenden Livestream (u.a. mit mir) und natürlich mit Gebärdendolmetschung. (@EinAugensch
c_lindner	de	2020-04-29T16:05:38.000+0000	FALSE	null	Mit @FuestClemens, VWL-Professor, Präsident des #ifo und Mitglied der #Leopoldina spreche ich in einer neuen #COVID19-Sonderfolge über den Schaden an der #
SWagenknecht	de	2020-04-29T10:28:22.000+0000	FALSE	null	300 Infizierte in einem Schlachtbetrieb? Gab es dort genug #Arbeitsschutz, Abstands- und Hygieneregeln? Offenbar braucht es mehr Kontrollen – und Betriebe, die s
berlinlieblich	de	2020-04-29T11:09:54.000+0000	FALSE	null	„Das Mindeste, was die Bundesregierung machen sollte, ist ein weiteres Ausbluten der Unternehmen durch Gewinnausschüttungen zu verhindern. Sie muss die Not
c_lindner	de	2020-04-29T17:57:29.000+0000	TRUE	ZDFheute	RT @ZDFheute: "Wenn Hygienekonzepte vorliegen - Masken, Desinfektion, ausreichen Abstand - , dann müssen Lokale, Läden, Schulen das gesells...
c_lindner	de	2020-04-29T13:17:56.000+0000	TRUE	maybrittlin	RT @maybrittlin: Thema morgen bei #linke "Die Politik macht auf – die Unsicherheit bleibt" Mehr Infos zu allen 6 Gästen ➡ https://t.co/G...
SWagenknecht	de	2020-04-29T10:28:22.000+0000	FALSE	WDRinvestigativ	300 Infizierte in einem Schlachtbetrieb? Gab es dort genug #Arbeitsschutz, Abstands- und Hygieneregeln? Offenbar braucht es mehr Kontrollen – und Betriebe, die s
berlinlieblich	de	2020-04-29T10:28:22.000+0000	FALSE	WDRinvestigativ	RT @WDRinvestigativ: In der Sp... #ZDFheute: "Wenn Hygienekonzepte vorliegen - Masken, Desinfektion, ausreichen Abstand - , dann müssen Lokale, Läden, Schulen das gesells...
berlinlieblich	de	2020-04-29T10:28:22.000+0000	FALSE	WDRinvestigativ	Tegel schließt. Endlich. Die Umstände in... #ZDFheute: "Wenn Hygienekonzepte vorliegen - Masken, Desinfektion, ausreichen Abstand - , dann müssen Lokale, Läden, Schulen das gesells...
berlinlieblich	de	2020-04-29T10:28:22.000+0000	FALSE	WDRinvestigativ	RT @WDRinvestigativ: In der Sp... #ZDFheute: "Wenn Hygienekonzepte vorliegen - Masken, Desinfektion, ausreichen Abstand - , dann müssen Lokale, Läden, Schulen das gesells...
brantner	de	2020-04-28T17:34:18.000+0000	TRUE	ThomasZawalski	RT @ThomasZawalski: Dr. Franziska Brantner MdB europapolitische Sprecherin der Grünen Bundestagsfraktion und Thomas Zawalski, Wirtschaftsbe...
DoroBaer	de	2020-04-28T17:34:18.000+0000	FALSE	null	@LenaRogl Also um den Toaster zum Brennen zu bringen, brauchte ich meine Kinder nicht. Ich konnte das ganz alleine... :-)
brantner	und	2020-04-28T17:34:18.000+0000	FALSE	YoYo_Ma	RT @YoYo_Ma: V. Epilogue from "The Fence, the Rooftop and the... Closed and performed with my dear friend @KinanAzmeh كان...
Volker_Beck	de	2020-04-28T17:34:18.000+0000	TRUE	Volker_Beck	RT @Volker_Beck: Ich sage mal so: "Wir werden womöglich alle s... #ZDFheute: "Wenn Hygienekonzepte vorliegen - Masken, Desinfektion, ausreichen Abstand - , dann müssen Lokale, Läden, Schulen das gesells...
Volker_Beck	de	2020-04-28T18:00:00.000+0000	TRUE	TspWissenschaft	RT @TspWissenschaft: Die deutsche Erinnerungspolitik gibt sich häufig als vorbildlich aus. Der Politikwissenschaftler @ProfSalzborn erklärt...
kahrs	fr	2020-04-29T05:36:02.000+0000	FALSE	null	moins
DoroBaer	de	2020-04-28T17:34:18.000+0000	FALSE	null	@LenaRogl Also um den Toaster zum Brennen zu bringen, brauchte ich meine Kinder nicht. Ich konnte das ganz alleine... :-)
brantner	de	2020-04-28T17:34:18.000+0000	TRUE	ThomasZawalski	RT @ThomasZawalski: Dr. Franziska Brantner MdB europapolitische S... cherin der Grünen Bundestagsfraktion und Thomas Zawalski, Wirtschaftsbe...
Volker_Beck	de	2020-04-28T18:01:11.000+0000	FALSE	null	Ich kann das Buch von @ProfSalzborn nur zur Lektüre empfehlen. Mirling ein ganzer Krok... https://t.co/h6WwF18bCo
Volker_Beck	de	2020-04-28T18:01:11.000+0000	TRUE	profocur	RT @profocur: Homosexualität führt letztlich zu Seuchen und damit sind Homosexuelle auch indirekt für die Coronapandemie in der Türkei ver...
Volker_Beck	de	2020-04-28T18:00:00.000+0000	TRUE	TspWissenschaft	RT @TspWissenschaft: Die deutsche Erinnerungspolitik gibt sich häufig als vorbildlich aus. Der Politikwissenschaftler @ProfSalzborn erklärt...
DoroBaer	de	2020-04-29T06:17:54.000+0000	TRUE	MarkusBlume	RT @MarkusBlume: Wir gedenken heute der Befreiung des #Konzentrationslagers #Dachau. Das menschenverachtende Grauen fand dort vor 75 Jahren...
kahrs	fr	2020-04-29T05:36:02.000+0000	FALSE	null	moins
juergenhardt	de	2020-04-29T05:21:41.000+0000	FALSE	null	Gut, dass die @KASonline das Scheinverfälscht auf die Krise in ve lenkt. Als @cdcsut haben wir uns immer dafür stark gemacht, den demokratischen Übergang u
juergenhardt	de	2020-04-29T05:21:41.000+0000	FALSE	null	Gut, dass die @KASonline das Scheinverfälscht auf die Krise in ve lenkt. Als @cdcsut haben wir uns immer dafür stark gemacht, den demokratischen Übergang u
GoeringEckardt	de	2020-04-27T06:57:19.000+0000	FALSE	null	Das ist mal ein Montag Morgen! #happy #GRUENE 100000 das ist 100 Prozent #Zukunft #Mut #Klimaschutz #Zusammenhalt @Die_Gruenen https://t.co/Uw0xbDf67
ZakinNastic	de	2020-04-26T19:37:22.000+0000	TRUE	Amira_M_Ali	RT @Amira_M_Ali: Links wirkt! Keine #Staatshilfen für Unternehmen, die Dividenden ausschütten. Aber jetzt bitte auch konsequent umsetzen. n...

STRUCTURED
META DATA
FILTER GROUP BY
SUMMARIZE - ARRANGE

UNSTRUCTURED
DATA
???

WHY IS TEXT DIFFERENT?

WHAT OPTIONS DO WE HAVE?

What can we do with text [without changing the structure](#)?

Apply **filter**, but with different operators from `{stringr}`:

- Search for keywords with `str_detect`
- Search for patterns with regular expressions

Apply **mutate** and extract matches:

- Extract whether a keyword has been matched
- Extract the matched keyword(s)
- Extract multiple matches and explode to rows

SEARCHING IN TEXT WITH `{stringr}`

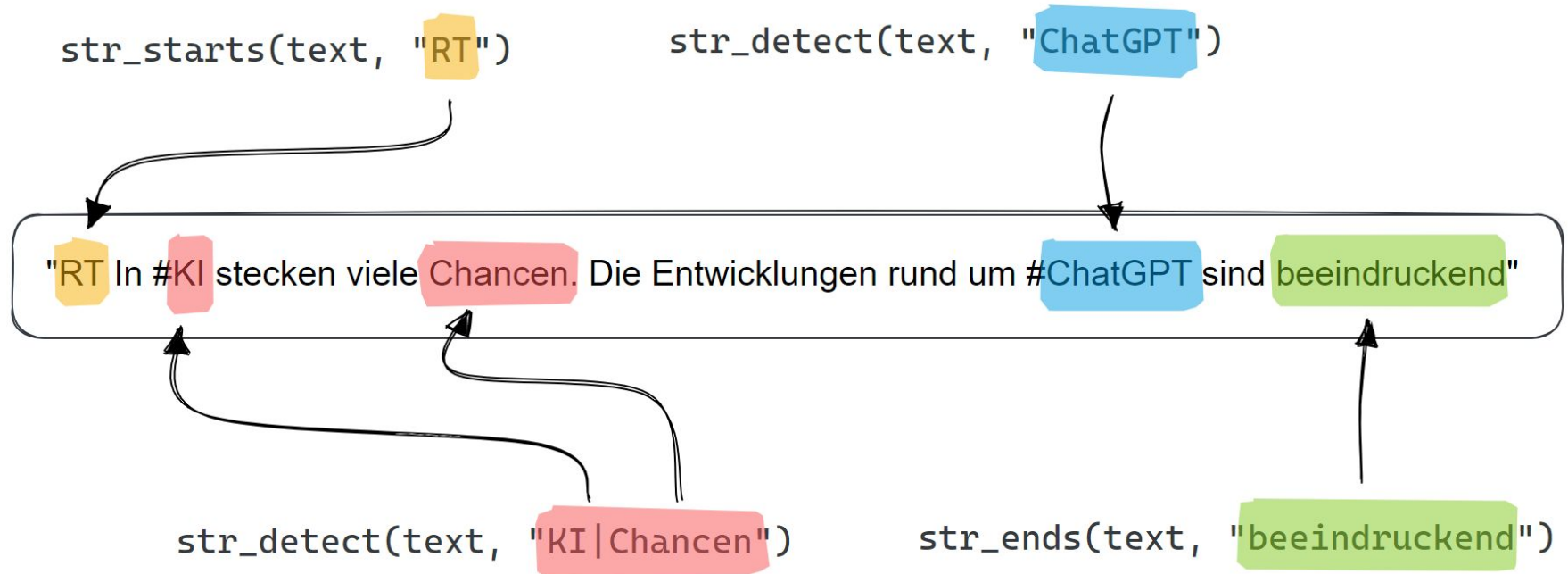
SEARCHES IN TEXT

SIMPLE KEYWORD MATCHES

- The `{stringr}` package contains functions for working with text
- We can use `{stringr}` to search in text:
 - `str_detect` for simple keyword matches
 - `str_starts` and `str_ends` as special cases
 - `str_to_lower` to improve matches
 - `str_count` to count occurrences of keywords and patterns

SEARCHES IN TEXT

SIMPLE KEYWORD MATCHES



SEARCHES IN TEXT

SIMPLE KEYWORD MATCHES

Simple keyword matches anywhere in the text:

```
tweets |>  
  filter(str_detect(text, "ChatGPT")) |>  
  select(screen_name, text)
```

SEARCHES IN TEXT

SIMPLE KEYWORD MATCHES

Searching at the start of a text:

```
tweets |>  
  filter(str_starts(text, "RT", negate = TRUE)) |>  
  select(screen_name, text)
```

SEARCHES IN TEXT

SIMPLE KEYWORD MATCHES

Searching at the end of a text:

```
tweets |>  
  filter(str_ends(text, "CL")) |>  
  select(screen_name, text)
```

SEARCHES IN TEXT WITH IMPROVED KEYWORD MATCHES

Improved keyword matches with `str_to_lower`:

```
tweets |>  
  mutate(text = str_to_lower(text)) |>  
  filter(str_detect(text, "chatgpt")) |>  
  select(screen_name, text)
```

SEARCHES IN TEXT WITH REGULAR EXPRESSIONS

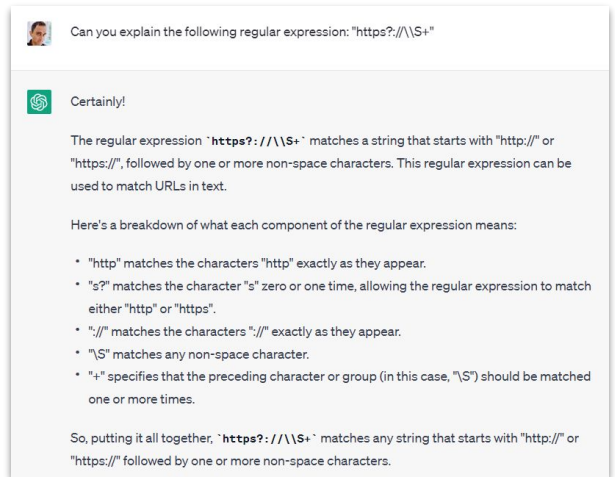
Regular expressions are powerful, yet can be complex to write.

Some examples:

- `str_detect(text, "chatgpt|gpt3|gpt4")`
- `str_detect(text, "@\\w+")`
- `str_detect(text, "https?:\\/\\/\\S+")`
- `str_detect(text, "\\s{2,}")`

All `{stringr}` functions actually take regular expressions!

ChatGPT to the rescue!



EXTRACT MATCHES WITH
`{stringr}`

It is good to know whether and which keyword matched in a text:

- `str_extract` to pull a match from the text into a new column
- `str_extract_all` to pull all matches into a new column as a list
- `str_extract_all` with `str_c` and `unlist` to concatenate all matches into a single string
- `str_extract_all` with `unnest_longer` from `{tidyr}` to extract all matches into separate rows

EXTRACT MATCHES

SINGLE MATCHES

Extract only the first match:

```
tweets |>  
  mutate(text = str_to_lower(text)) |>  
  mutate(first_match = str_extract(text, "chatgpt|gpt3|gpt4")) |>  
  filter(!is.na(first_match)) |>  
  select(first_match, text)
```

EXTRACT MATCHES

ALL MATCHES AS LIST

Extract all matches as a list:

```
tweets |>  
  mutate(text = str_to_lower(text)) |>  
  mutate(matches = str_extract_all(text, "chatgpt|gpt3|gpt4")) |>  
  filter(!is.na(matches)) |>  
  select(matches, text)
```

EXTRACT MATCHES

ALL MATCHES AS STRING

With `str_c` and `unlist`, we can transform a list of strings into a concatenated string:

```
tweets |>  
  mutate(text = str_to_lower(text)) |>  
  mutate(matches = str_extract_all(text, "chatgpt|gpt3|gpt4")) |>  
  mutate(matches_flat = str_c(unlist(matches), collapse = ",")) |>  
  select(matches_flat, text)
```

EXTRACT MATCHES

ALL MATCHES AS ROWS

With `unnest_longer`, we can explode a list of character strings into separate rows:

```
tweets |>
  mutate(extracted_urls = str_extract_all(text, "https?://\\S+")) |>
  unnest_longer(extracted_urls, keep_empty = TRUE) |>
  select(id, screen_name, extracted_urls, text) |>
  arrange(id)
```

REPLACE MATCHES WITH
`{stringr}`

REPLACE MATCHES

Occasionally, we want to remove things from text:

- `str_replace` to replace the first match with a new string
- `str_replace_all` to replace all matches with a new string
- `str_remove` and `str_remove_all` to delete occurrences
- `str_trim` to remove leading and trailing white spaces from text

All code examples are on GitHub: <https://github.com/winf-hsos/data-analytics-code>

SPLITTING, SUBSTRINGS, AND CONCATENATING TEXT WITH `{stringr}`

SPLIT TEXT

BREAK UP A TEXT

Splitting a string into multiple pieces can be helpful at times:

- `str_split` to break a long string into a list of strings using a separator character or pattern
- `str_split_i` to break a long string using a separator and keeping only the i-th result
- `str_sub` extracts a part from a string via start and end locations
- `str_c` to concatenate a list of strings into one strings

CASE CONVERSION WITH `{stringr}`

CASE CONVERSION

ALL CAPS OR LOWER

Splitting a string into multiple pieces can be helpful at times:

- `str_to_lower` to convert all letters to lowercase
- `str_to_upper` to convert all letter to uppercase
- `str_title` to make the first letter uppercase

TRIMMING TEXT WITH `{stringr}`

TRIMMING

Trimming a string is sometimes necessary, especially to correct for faulty user input:

- `str_trim` removes white spaces at the beginning and end of a string
- `str_squish` also replaces multiple white spaces within a string with a single one