Text Analysis in R

Christian Overgaard
CME, Journalism Vertical

Big thanks to Jo Lukito!!

The University of Texas at Austin
Center for Media Engagement
Moody College of Communication

2023-04-25

Workshop Goals

Provide you with the skills needed to do basic text analysis.

1. How to look for messages containing a word? (e.g., 🇺🇦🇺🇦)
2. How to find the most common words in a dataset?
    - ... and how to visualize top words/frequencies?

This **beginner** workshop is meant for both qualitative and quantitative researchers. We will **not** go into depth about advanced NLP models.

Examples of what you'll learn today

A	B
id	text
1	@TheDemCoalition @BarackObama @realDonald
2	Trump Laid the Groundwork for a Coronavirus Mes
3	In the face of a potential infectious disease outbrea
4	Fact check: McConnell claims Obama didn't leave T
5	Coronavirus quarantine hits MORE of Trump's inne
6	Government Accountability Office To Investigate T
7	@andweknow in his newest video on youtube high
8	The Trump Administration's COVID-19 Non-Guidar
9	@DeniseShearin @BGrueskin He's also the incomp
10	President Trump is fast tracking these antiviral dru
11	Trump is now relying for his future on the very exp
12	@jcbudllc And his answer has been the same, there
13	Replace Trump And Bolster The CDC, A Leading Me
14	They plan on leaving all of their cars running becau
15	@SusanBr86829147 @KayBurley @MattHancock
16	We are likely to cross 10k deaths today or tomorro
17	We are facing two deadly illnesses. The Corona viru
18	White House hopes to wind down COVID-19 task fo
19	@GregMolidor Trump says his administration will
20	Rural communities without a hospital struggle to f
21	Americans slam Trump for bragging about ratings c
22	Alarm, Denial, Blame: The Pro-Trump Media's Co
23	What's the point having a #SecretService to protec
24	@liamstone_19 Japanese Americans post Pearl Har
25	Why does the liberal media not want to show @r
26	Coronavirus Live Updates: As Governors Look to Re
27	@HuffPostPol She's a stronger woman than me. Be
28	What the FUCK. Arrerregggusdfghvcujkl Russia to ac
29	@DavidAgStone @ProudResister @BernieSanders
30	@gmbutts #Covid_19 take a read Gerald. Maybe yo
31	Vice President Mike Pence and His Wife Test Negati
32	@RepGalonski "The real push [against] hydroxychl
33	trump could even send us all a million bucks during
34	@marcorubio So the coronavirus is a democrat? H
35	Irrelevant Never-Trump Loser Steve Schmidt: Trum
36	As I get ready to pass out courtesy of a few glasses o
37	Biden Surges Past Trump As Just 34% Of Independe
38	3 patterns of dishonesty to look out for in Trump's
39	Phase 205/04/2020Hon'ble Chief Minister Smt. Ma

Extract just the messages that

a) where written by a **specific account**, or

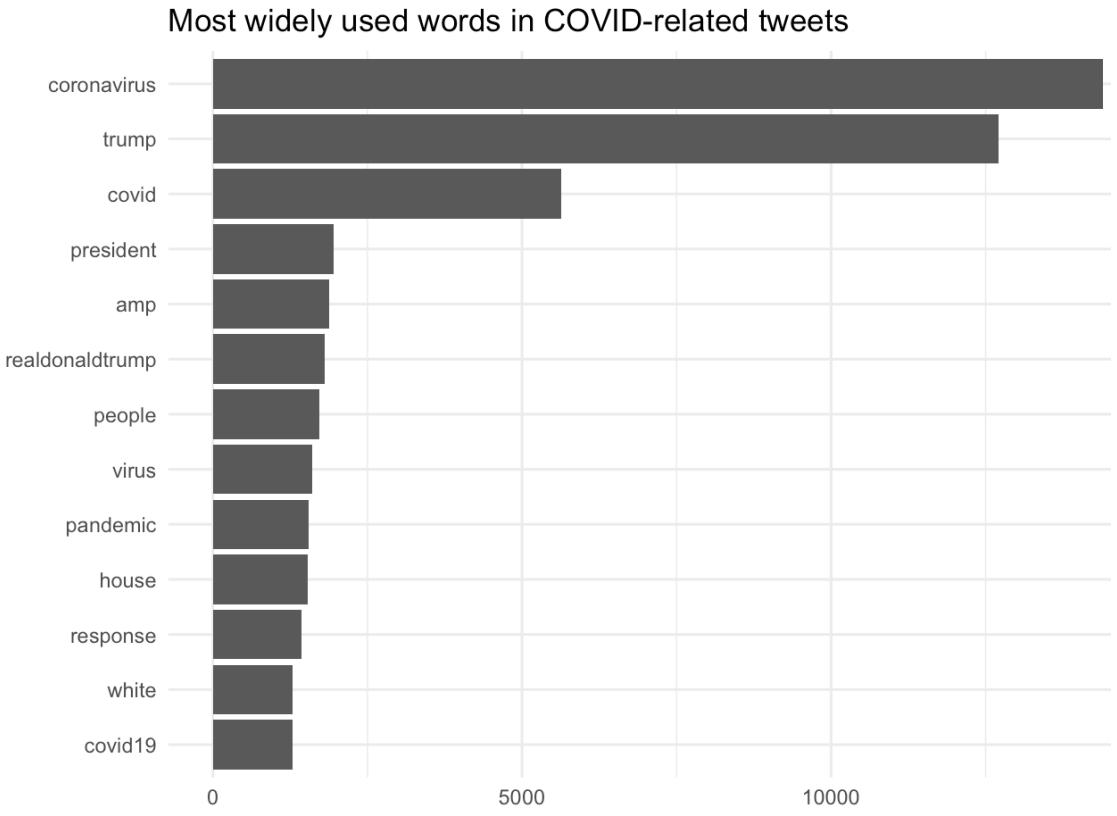
b) that mention a **specific word** or set of words

A	B
id	text
1	@andweknow in his newest video on youtube highlights info on UV
15	@SusanBr86829147 @KayBurley @MattHancock Sympathise with
22	Alarm, Denial, Blame: The Pro-Trump Media's Coronavirus Disto
26	Coronavirus Live Updates: As Governors Look to Reopen, Trump Enc
33	trump could even send us all a million bucks during the pandemic a
38	3 patterns of dishonesty to look out for in Trump's remarks https://
40	If #corona a conspiracy its most likely a group effort led by ambitio
46	"The vice-president, Mike Pence, was speaking at Fema headquarte
...	"Based on JCT analysis, millionaire tax filers benefiting from one of
...	Trump supporters attacking new COVID-19 drug because the presid
...	This cause is close to my heart - please sign: https://t.co/arSV5i8z1
...	Coronavirus: Trump Bans Travel From Europe To Us https://t.co/CV
...	Alright no matter what you think of Trump, begrudging the guy a da
...	#TrumpMeltdown 2020:DO NOT #Retweet This To Enrage #Trump
...	Economists Rip #Trump For Getting #CoronavirusStimulus Totally V
...	@WhiteHouse @realDonaldTrump The corona virus will make the r
...	But he did though....https://t.co/SzPwP7l5HL https://t.co/BLJzzW4
...	@ToddLlewellyn "This is their new hoax." - President Trumphttps:/
...	@BBCWorld has been following 3 Trump voters (from #NorthCar
n	It just keeps getting better and better.https://t.co/GrAivOh9SR

Examples of what you'll learn today

A	B
id	text
1	@TheDemCoalition @BarackObama @realDonald
2	Trump Laid the Groundwork for a Coronavirus Mes
3	In the face of a potential infectious disease outbrea
4	Fact check: McConnell claims Obama didn't leave T
5	Coronavirus quarantine hits MORE of Trump's inne
6	Government Accountability Office To Investigate T
7	@andweknow in his newest video on youtube high
8	The Trump Administration's COVID-19 Non-Guidar
9	@DeniseShearin @BGrueskin He's also the incomp
10	President Trump is fast tracking these antiviral dru
11	Trump is now relying for his future on the very exp
12	@jcbudllc And his answer has been the same, there
13	Replace Trump And Bolster The CDC, A Leading Me
14	They plan on leaving all of their cars running becau
15	@SusanBr86829147 @KayBurley @MattHancock
16	We are likely to cross 10k deaths today or tomorro
17	We are facing two deadly illnesses. The Corona viru
18	White House hopes to wind down COVID-19 task fo
19	@GregMolidor Trump says his administration will
20	Rural communities without a hospital struggle to f
21	Americans slam Trump for bragging about ratings c
22	Alarm, Denial, Blame: The Pro-Trump Media's Co
23	What's the point having a #SecretService to protec
24	@liamstone_19 Japanese Americans post Pearl Har
25	Why does the liberal media not want to show @r
26	Coronavirus Live Updates: As Governors Look to Re
27	@HuffPostPol She's a stronger woman than me. Be
28	What the FUCK. Arrerregggusdfghvcujkl Russia to ac
29	@DavidAgStone @ProudResister @BernieSanders
30	@gmbutts #Covid_19 take a read Gerald. Maybe yo
31	Vice President Mike Pence and His Wife Test Negati
32	@RepGalonski "The real push [against] hydroxychl
33	trump could even send us all a million bucks durin
34	@marcorubio So the coronavirus is a democrat? H
35	Irrelevant Never-Trump Loser Steve Schmidt: Trum
36	As I get ready to pass out courtesy of a few glasses o
37	Biden Surges Past Trump As Just 34% Of Independe
38	3 patterns of dishonesty to look out for in Trump's
39	Phase 205/04/2020Hon'ble Chief Minister Smt. Ma

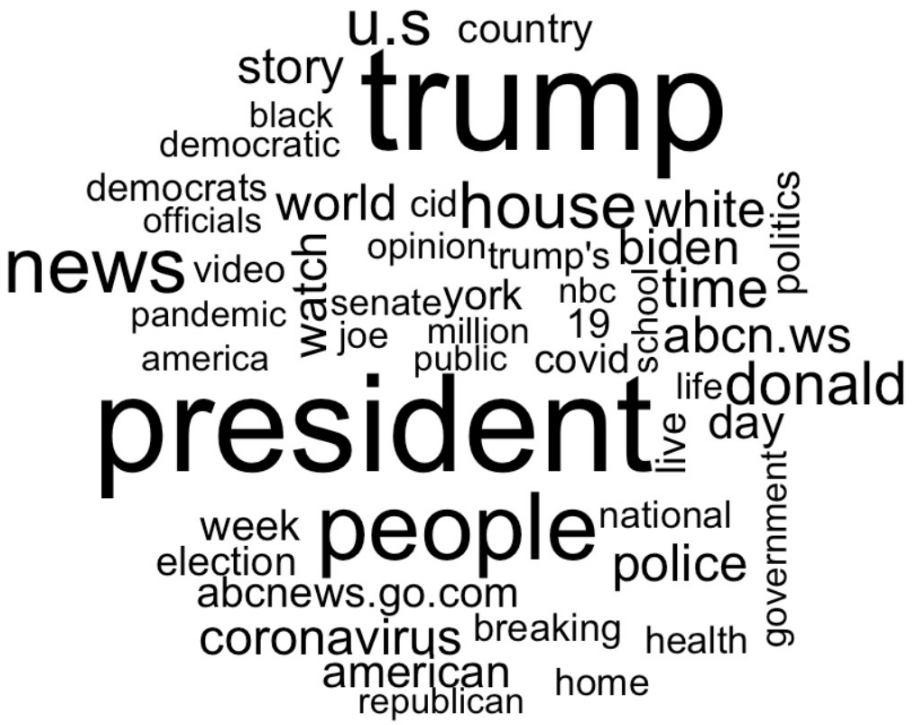
Bar chart



Examples of what you'll learn today

A	B
id	text
1	@TheDemCoalition @BarackObama @realDonald
2	Trump Laid the Groundwork for a Coronavirus Mes
3	In the face of a potential infectious disease outbrea
4	Fact check: McConnell claims Obama didn't leave T
5	Coronavirus quarantine hits MORE of Trump's inne
6	Government Accountability Office To Investigate T
7	@andweknow in his newest video on youtube high
8	The Trump Administration's COVID-19 Non-Guidar
9	@DeniseShearin @BGrueskin He's also the incomp
10	President Trump is fast tracking these antiviral dru
11	Trump is now relying for his future on the very exp
12	@jcbudllc And his answer has been the same, there
13	Replace Trump And Bolster The CDC, A Leading Me
14	They plan on leaving all of their cars running becau
15	@SusanBr86829147 @KayBurley @MattHancock
16	We are likely to cross 10k deaths today or tomorro
17	We are facing two deadly illnesses. The Corona viru
18	White House hopes to wind down COVID-19 task fo
19	@GregMolidor Trump says his administration will
20	Rural communities without a hospital struggle to f
21	Americans slam Trump for bragging about ratings c
22	Alarm, Denial, Blame: The Pro-Trump Media's Co
23	What's the point having a #SecretService to protec
24	@liamstone_19 Japanese Americans post Pearl Har
25	Why does the liberal media not want to show @r
26	Coronavirus Live Updates: As Governors Look to Re
27	@HuffPostPol She's a stronger woman than me. Be
28	What the FUCK. Arrerregggusdfghvcujkl Russia to ac
29	@DavidAgStone @ProudResister @BernieSanders
30	@gmbutts #Covid_19 take a read Gerald. Maybe yo
31	Vice President Mike Pence and His Wife Test Negati
32	@RepGalonski "The real push [against] hydroxychl
33	trump could even send us all a million bucks during
34	@marcorubio So the coronavirus is a democrat? H
35	Irrelevant Never-Trump Loser Steve Schmidt: Trum
36	As I get ready to pass out courtesy of a few glasses o
37	Biden Surges Past Trump As Just 34% Of Independe
38	3 patterns of dishonesty to look out for in Trump's
39	Phase 205/04/2020Hon'ble Chief Minister Smt. Ma

Word cloud



“Subfields” Analyzing Text

- Natural Language Processing (computer science)
- Computational Linguistics (linguistics)
- Text as Data (political science, political communication)
- Computer-assisted content analysis (communication)

Natural language: language generated by humans.
Text: written natural language.

How computers “read” text

Computers are very dumb and cannot read text.

But they *can* read when two ***characters*** are different (“a” != “A”). And they can tell how many characters are in a ***string***.

A character: a unit of information equivalent to one alphabetic letter or symbol.
A string: a bunch of characters.

Characters and Strings

How many characters are in this string?
(Don’t count the quotation marks)

“apple”

Characters and Strings

How many characters are in this string?
(Don’t count the quotation marks)

“apples”

Characters and Strings

How many characters are in this string?
(Don’t count the quotation marks)

“Apples suck.”

Note: “Apple” is a substring in “Apples suck.”

Characters, Strings, and Patterns

Computers can't understand text, but they can recognize patterns.

*This is where **regular expressions** come in!*

"a sequence of symbols and characters expressing a string or pattern to be searched for within a longer piece of text."

Using Regular Expressions to look for things

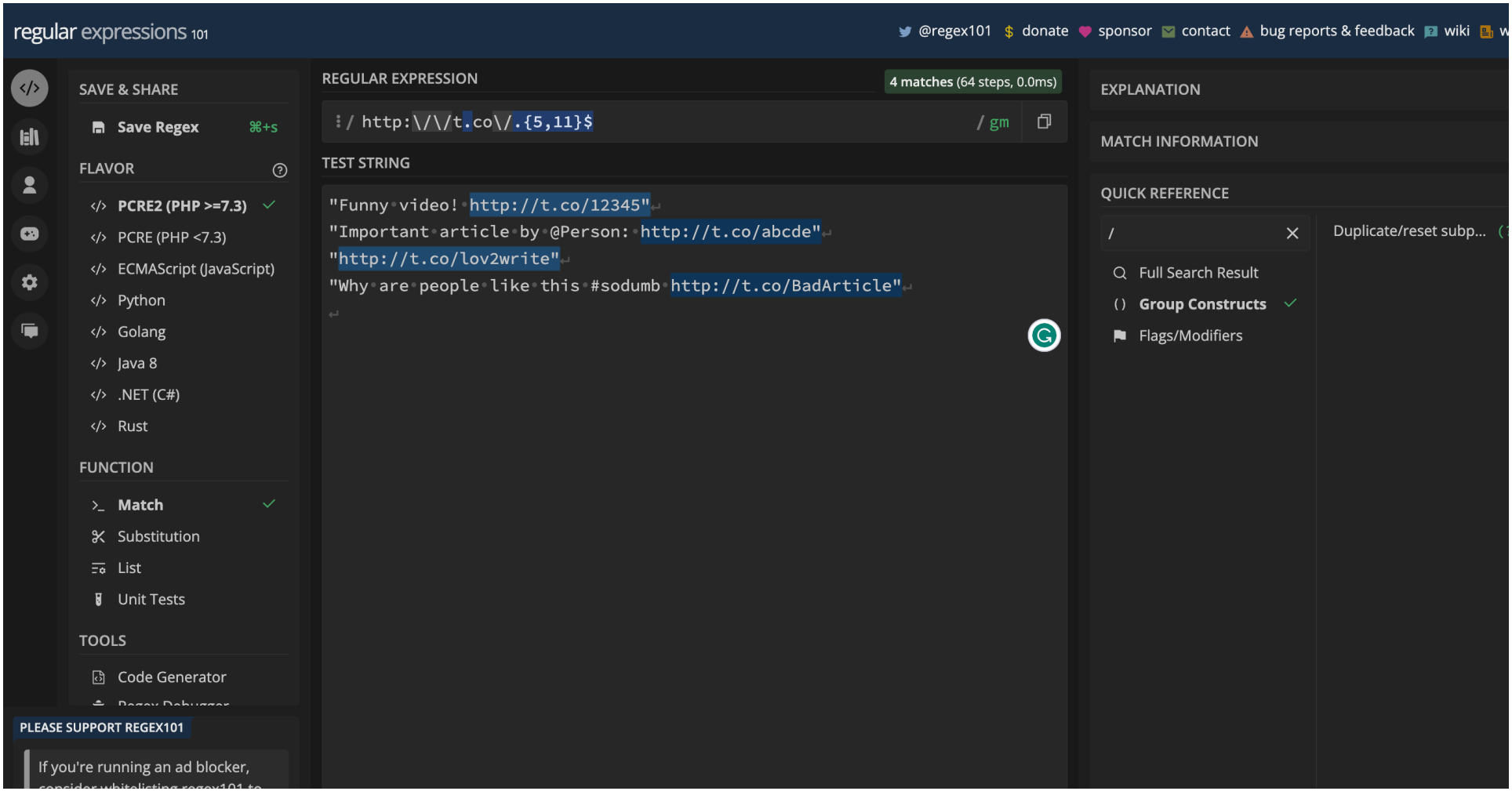
"Funny video! http://t.co/12345"
"Important article by @Person: http://t.co/abcde"
"http://t.co/lov2write"
"Why are people like this #sodumb http://t.co/BadArticle"

We can use regular expressions to identify the urls:

"http://t.co/.{5,11}$"

Two resources:
(a) https://regex101.com/
(b) Cheatsheet: https://paulvanderlaken.files.wordpress.com/2017/08/r-regular-expression-cheetsheat.pdf

https://regex101.com/



Basic Regular Expressions in R

Cheat Sheet

Character Classes

[[:digit:]] or \\d	Digits; `[0-9]`
\\D	Non-digits; `[^0-9]`
[[:lower:]]	Lower-case letters; `[a-z]`
[[:upper:]]	Upper-case letters; `[A-Z]`
[[:alpha:]]	Alphabetic characters; `[A-z]`
[[:alnum:]]	Alphanumeric characters `[A-z0-9]`
\\w	Word characters; `[A-z0-9_]`
\\W	Non-word characters
[[:xdigit:]] or \\x	Hexadec. digits; `[0-9A-Fa-f]`
[[:blank:]]	Space and tab
[[:space:]] or \\s	Space, tab, vertical tab, newline, form feed, carriage return
\\S	Not space; `[^[:space:]]`
[[:punct:]]	Punctuation characters; `!"#$%&'()*+,-./:;<=>?@[]^_`{|}~`
[[:graph:]]	Graphical char.; `[[:alnum:][:punct:]]`
[[:print:]]	Printable characters; `[[:alnum:][:punct:]\\s]`
[[:cntrl:]] or \\c	Control characters; `\n, \r` etc.

Special Metacharacters

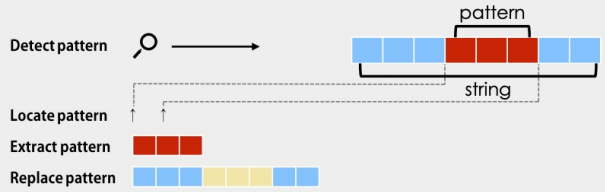
\n	New line
\r	Carriage return
\t	Tab
\v	Vertical tab
\f	Form feed

Lookaraounds and Conditionals*

(?=)	Lookahead (requires PERL = TRUE), e.g. `(?=yx)`: position followed by `'xy'`
(?!)	Negative lookahead (PERL = TRUE); position NOT followed by pattern
(?<=)	Lookbehind (PERL = TRUE), e.g. `(?<=yx)`: position following `'xy'`
(?<!)	Negative lookbehind (PERL = TRUE); position NOT following pattern
?(if)then	If-then-condition (PERL = TRUE); use lookaheads, optional char. etc in if-clause
?(if)then|else	If-then-else-condition (PERL = TRUE)

*see, e.g. http://www.regular-expressions.info/lookaround.html
http://www.regular-expressions.info/conditional.html

Functions for Pattern Matching




```
> string <- c("Hiphopopotamus", "Rhymenoceros", "time for bottomless lyrics")
> pattern <- "t.m"
```

Detect Patterns

grep(pattern, string)
`[1] 1 3`

grep(pattern, string, value = TRUE)
`[1] "Hiphopopotamus"`
`[2] "time for bottomless lyrics"`

grepl(pattern, string)
`[1] TRUE FALSE TRUE`

stringr::str_detect(string, pattern)
`[1] TRUE FALSE TRUE`

Locate Patterns

regexpr(pattern, string)
find starting position and length of first match

gregexpr(pattern, string)
find starting position and length of all matches

stringr::**str_locate(string, pattern)**
find starting and end position of first match

stringr::**str_locate_all(string, pattern)**
find starting and end position of all matches

Split a String using a Pattern

strsplit(string, pattern) or stringr::**str_split(string, pattern)**

Extract Patterns

regmatches(string, regexpr(pattern, string))
extract first match `[1] "tam" "tim"`

regmatches(string, gregexpr(pattern, string))
extracts all matches, outputs a list
`[[1]] "tam" [[2]] character(0) [[3]] "tim" "tom"`

stringr::**str_extract(string, pattern)**
extract first match `[1] "tam" NA "tim"`

stringr::**str_extract_all(string, pattern)**
extract all matches, outputs a list

stringr::**str_extract_all(string, pattern, simplify = TRUE)**
extract all matches, outputs a matrix

stringr::**str_match(string, pattern)**
extract first match + individual character groups

stringr::**str_match_all(string, pattern)**
extract all matches + individual character groups

Replace Patterns

sub(pattern, replacement, string)
replace first match

gsub(pattern, replacement, string)
replace all matches

stringr::**str_replace(string, pattern, replacement)**
replace first match

stringr::**str_replace_all(string, pattern, replacement)**
replace all matches

Character Classes and Groups

.	Any character except \n
|	Or, e.g. `(a|b)`
[...]	List permitted characters, e.g. `[abc]`
[a-z]	Specify character ranges
[^...]	List excluded characters
(...)	Grouping, enables back referencing using \\N where N is an integer

Anchors

^	Start of the string
$	End of the string
\\b	Empty string at either edge of a word
\\B	NOT the edge of a word
\\<	Beginning of a word
\\>	End of a word

Quantifiers

*	Matches at least 0 times
+	Matches at least 1 time
?	Matches at most 1 time; optional string
{n}	Matches exactly n times
{n,}	Matches at least n times
{,n}	Matches at most n times
{n,m}	Matches between n and m times

General Modes

By default R uses *POSIX extended regular expressions*. You can switch to *PCRE regular expressions* using `PERL = TRUE` for base or by wrapping patterns with `perl()` for stringr.

All functions can be used with literal searches using `fixed = TRUE` for base or by wrapping patterns with `fixed()` for stringr.

All base functions can be made case insensitive by specifying `ignore.cases = TRUE`.

Escaping Characters

Metacharacters (`. * +` etc.) can be used as literal characters by escaping them. Characters can be escaped using `\\` or by enclosing them in `\\Q...\\E`.

Case Conversions

Regular expressions can be made case insensitive using `(?i)`. In backreferences, the strings can be converted to lower or upper case using `\\L` or `\\U` (e.g. `\\L\\1`). This requires `PERL = TRUE`.

Greedy Matching

By default the asterisk * is greedy, i.e. it always matches the longest possible string. It can be used in lazy mode by adding `?`, i.e. `*?`.

Greedy mode can be turned off using `(?U)`. This switches the syntax, so that `(?U)a*` is lazy and `(?U)a*?` is greedy.

Note

Regular expressions can conveniently be created using rex::**rex()**.

Updated: 09/16

Beyond Regular Expressions: Tokenizing

Searching for messages containing a word may be useful, but sometimes, you want some aggregate descriptives, too.

How do we go about processing data so our (stupid) computer can understand what we mean?

Tokenizing (v.)

“to break (text) into individual linguistic units.”

(Source: Oxford Languages)

We often tokenize sentences into individual words (called “tokens”). But you can also tokenize an article into its sentences, or a sentence into “n-grams”

Tokenizing (v.)

“to break (text) into individual linguistic units.”

(Source: Oxford Languages)

We often tokenize sentences into individual words (called “tokens”). But you can also tokenize an article into its sentences, or a sentence into “n-grams”

“Apples suck.”

Tokenizing (v.)

“to break (text) into individual linguistic units.”

(Source: Oxford Languages)

We often tokenize sentences into individual words (called “tokens”). But you can also tokenize an article into its sentences, or a sentence into “n-grams”

“Apples suck.”

“Bananas are so so much better”

Tokenized Structure

When you tokenize a sentence, or message into word-tokens, each "word" is treated as its own thing:

One string: "Bananas are so so much better"

Tokenized: c("Bananas", "are", "so", "so", "much", "better")

N-grams

Unigram: One word (e.g., "banana")
Bigram: Two words (e.g., "good banana")
Trigram: Three words (e.g., "very good banana")
N-gram: N words

“Bananas are so much better”

R Workshop

- I encourage you to follow along on your own computer (but it's not required). To do so:
  - Open this file: TextAnalysisWorkshopApril2023.Rproj
  - Open this file: 01-Text Analysis in R - Activity 1.Rmd
  - Check that you have the required R packages