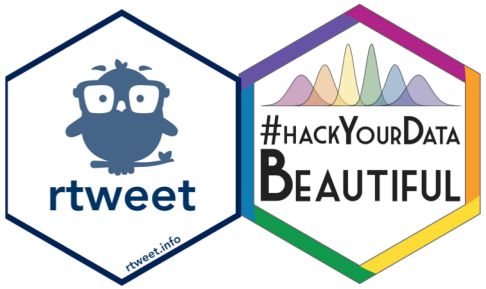


Scraping and visualizing Twitter data

@AnnaHenschel

Anna.Henschel@glasgow.ac.uk

A short introduction to Twitter (and rtweet).

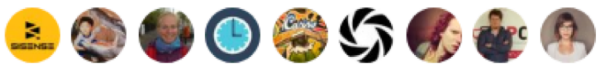
Mike Kearney
@kearneymw

Following

Social media + data science == super trendy, right? Well then install my R package, rtweet, and start figuring it out. #rstats

9:33 PM - 11 Jan 2017

7 Retweets **22** Likes



1 7 22

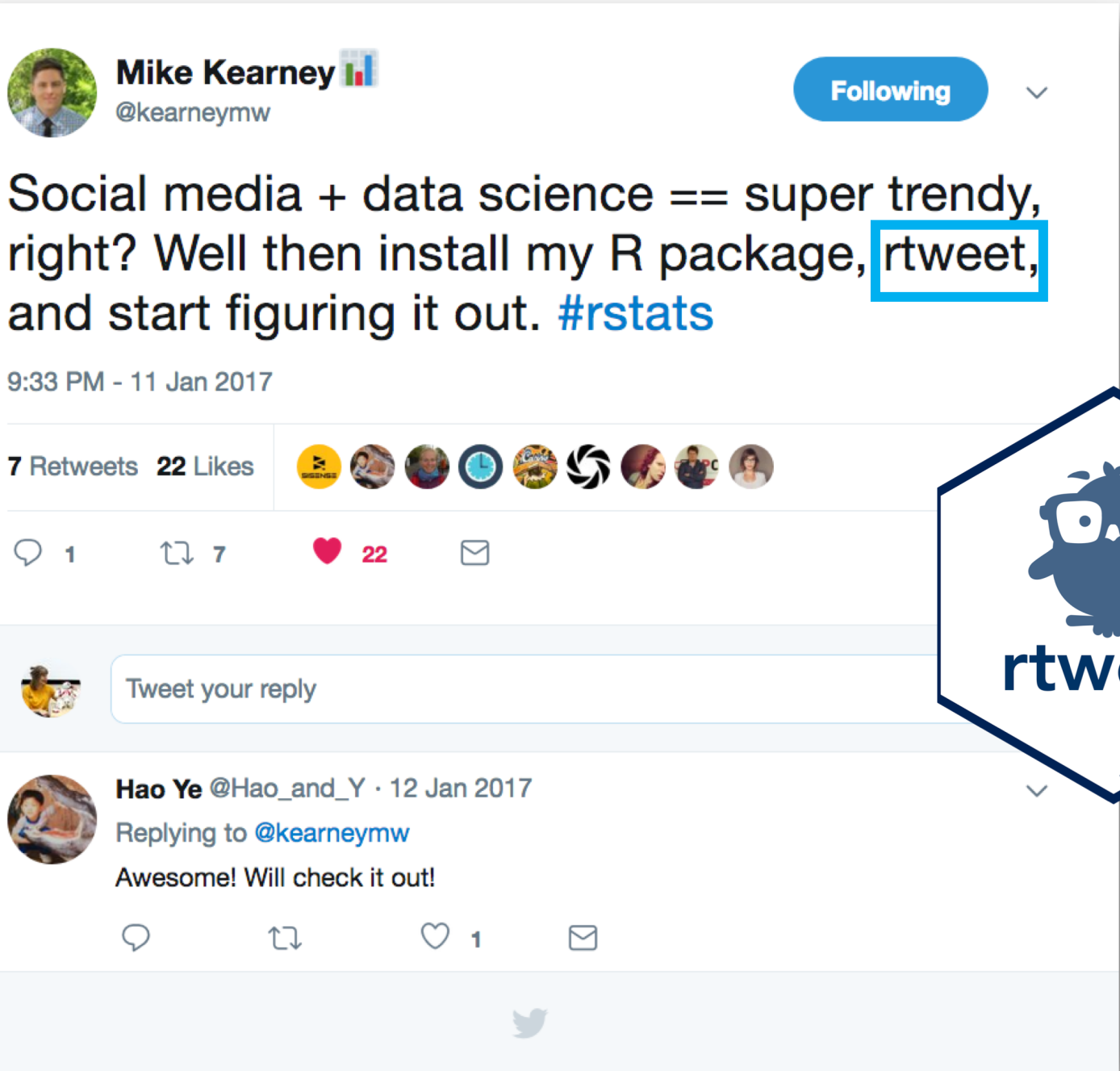
Tweet your reply

Hao Ye @Hao_and_Y · 12 Jan 2017

Replying to @kearneymw

Awesome! Will check it out!

1



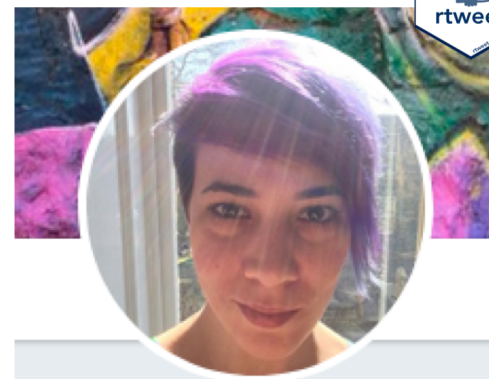
https://mikewk.com/

Many scientists are on Twitter!

Steph NicAllan
@eolasinntinn Follows you

Jack Taylor
@JackEdTaylor Follows you

Rebecca Lai
@_R_Lai_ Follows you



Lisa DeBruine 🏳️‍🌈
@LisaDeBruine Follows you

Lovisa Sundin
@menimagerie

Carolyn
@CarolynBot Follows you

Shannon McNee
@ShannonMcNee2 Follows you

Niamh Stack
@Eavanmac Follows you

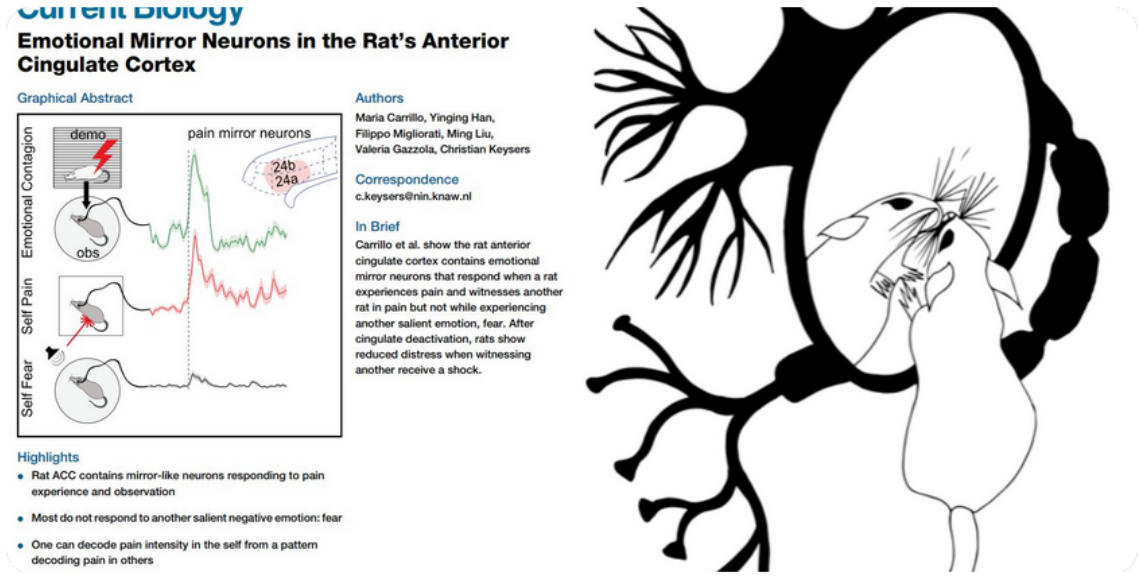
Why though ?

New papers / preprints

Socialbrainlab
@sbl_nin

Following

We have a new paper out that demonstrates the existence of emotional mirror neurons in the cingulate and shows that if you deactivate the region, you interfere with emotional contagion. Check it out here: cell.com /current-biolog ...



Current Biology

Emotional Mirror Neurons in the Rat's Anterior Cingulate Cortex

Graphical Abstract

Authors

Maria Carrillo, Yinging Han, Filippo Migliorati, Ming Liu, Valeria Gazzola, Christian Keysers

Correspondence

c.keysers@nin.knaw.nl

In Brief

Carrillo et al. show the rat anterior cingulate cortex contains emotional mirror neurons that respond when a rat experiences pain and witnesses another rat in pain but not while experiencing another salient emotion, fear. After cingulate deactivation, rats show reduced distress when witnessing another receive a shock.

Highlights

- Rat ACC contains mirror-like neurons responding to pain experience and observation
- Most do not respond to another salient negative emotion: fear
- One can decode pain intensity in the self from a pattern decoding pain in others

Anna Henschel
@AnnaHenschel

Hey #sciencetwitter, check out my preprint! Come for the robots and interpersonal synchrony, stay for #rstats pirateplots and multivariate assumption checks 🤓🤖 #phdchat #preprint

Preprint: psyarxiv.com/q9ku8

> **PsyArXiv-bot** @PsyArXivBot
> The effect of interpersonal synchrony with a robot on likeability and social motivation osf.io/q9ku8/

7:48 AM - 5 Nov 2018

12 Retweets **36** Likes

1 12 36

Academic journals with a presence on Twitter are more widely disseminated and receive a higher number of citations

Previous research has shown that researchers' active participation on Twitter can be a powerful way of promoting and disseminating academic outputs and improving the prospects of increased citations. But does the same hold true for the presence of academic journals on Twitter? ***José Luis Ortega*** *examined the role of 350 scholarly journals, analysing how their articles were tweeted and cited. Findings reveal that articles from those journals that have their own individual Twitter handle are more tweeted about than articles from journals whose only Twitter presence is through a scientific society or publisher account. Articles published in journals with any sort of Twitter presence also receive more citations than those published in journals with no Twitter presence.*

Email Address

Subscribe to the Impact Blog

This work is licensed under a Creative Commons Attribution 3.0 Unported License unless otherwise stated.

Dr Claudia Antolini #FBPE

@CA_AstroComm

Follow

Be your best.

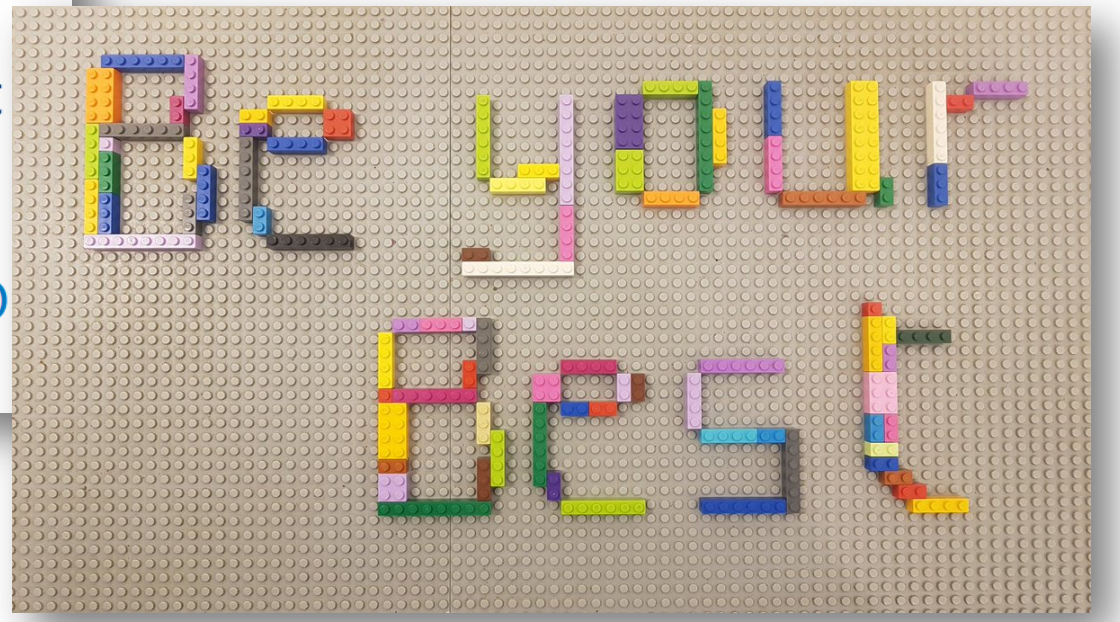
Not better than your colleague, not the best there ever was.

Be the best you can be today.

Sometimes being your best is getting out of bed. Sometimes it's forgiving yourself for not even being able to do that.

#Motivation #mentalhealth #PhDChat #LEGO #colours

A supportive community



And ... data!

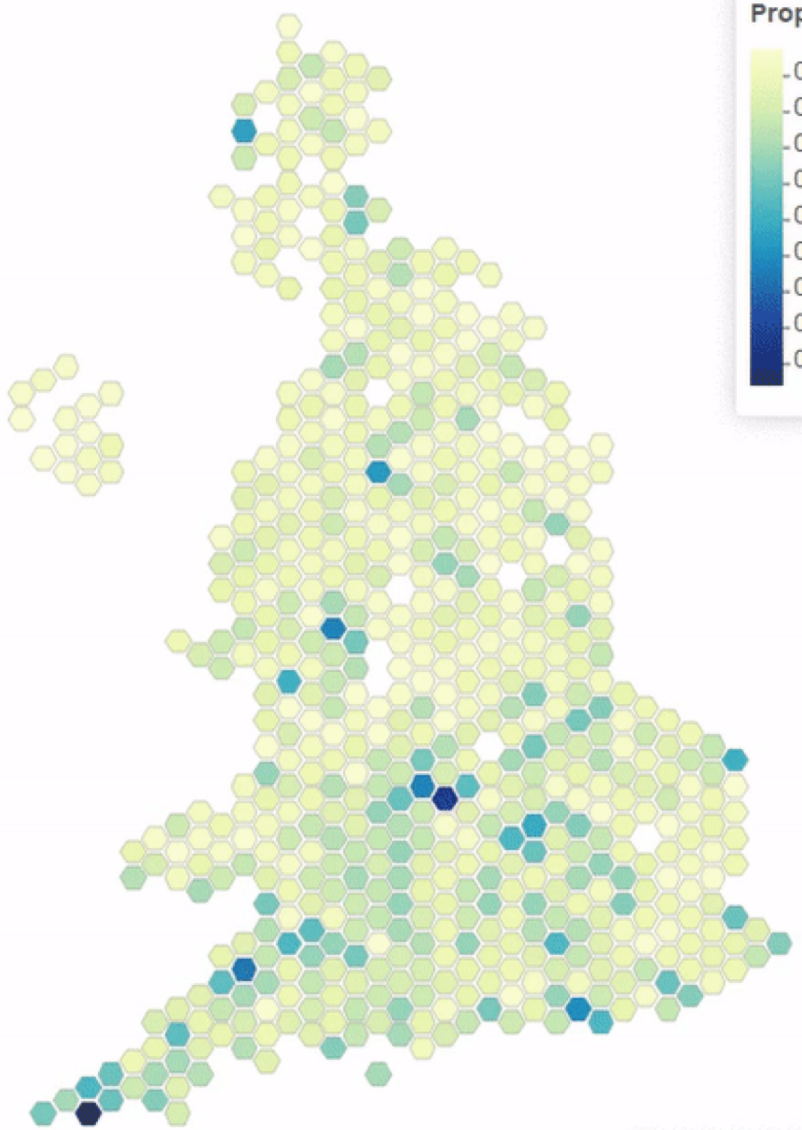
Jack Taylor
@JackEdTaylor

Following

It turns out you can use API calls to get more detailed data on E-petitions (explore.data.parliament.uk/?learnmore=e-P...), here's an animation I made last night showing how the proportion of #RevokeArticle50 signatures from each UK constituency changes over time.

#rstats #PeoplesVote

And ... data!



Day: 23/02/19

Signatures Today: +3,660

Total Signatures so far: 3,660

A word on ethics.

- *Twitter developer terms of service*
- *Don't derive or store sensitive information*
- *The role of consent?*

Taylor & Pagliari , 2018; Williams, Burnap, & Sloan, 2017

datenkraken

English [edit]

Etymology [edit]

Borrowed from German *Datenkraken*, plural of *Datenkrake*, itself a compound of *Daten* (“data”) and *Krake* (“octopus”), invoking an imagery of such a company being an octopus having its virtual "tentacles" deeply penetrating its users' online habits.

What are we going to do in this tutorial?

- *Get data from Twitter using rtweet*
- *Wrangle Twitter data with tidytext*
- *Sentiment analysis*
- *(Additional practice)*

Remember:

- *The red text does not always mean* 
- *If you fall behind, copy/paste from the web materials for this session*
- *Write the code in a **.rmd (R Markdown)** file – not in the console!*

Installing rtweet

```
# install rtweet from CRAN
install.packages("rtweet")

# load rtweet package
library(rtweet)
```

Other packages:

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

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

```
library(tidyverse)
```

Tipp:

Rtweet interacts with Twitter's API. In order to use the package you need to allow Rstudio to authenticate you as a user. When running the first function, a popup window in your browser will appear, confirming this.

[Lego Grad Student]

[Lego Grad Student]

[Lego Grad Student]

Getting (almost all) tweets of a user


```
lego<- get_timeline("@legogradstudent", n=3200)
```

Getting (almost all) tweets of a user

```
view(lego)

# Look at first few lines of the dataframe
head(lego)
```

```
head(lego)
A tibbl~: 6 x 88
user_id status_id created_at          screen_name text  source display_text_wi~ reply_to_status~
<chr>   <chr>     <dttm>              <chr>       <chr> <chr>              <dbl> <chr>
744046~ 11102414~ 2019-03-25 18:05:52 legogradst~ @Kea~ Twitt~               56 111024033186024~
744046~ 11102285~ 2019-03-25 17:14:49 legogradst~ @Chr~ Twitt~               23 111022809928587~
744046~ 11102246~ 2019-03-25 16:59:12 legogradst~ Jamm~ Twitt~              136 NA
744046~ 11096614~ 2019-03-24 03:41:32 legogradst~ "@em~ Twitt~              169 110964892965265~
744046~ 11092093~ 2019-03-22 21:44:51 legogradst~ We a~ Twitt~              145 NA
744046~ 11081976~ 2019-03-20 02:44:50 legogradst~ @mjb~ Twitt~               96 110819630522089~
```

Tidy tweets = *one word per row format*


```
tidy_tweets <- lego %>%
    filter(is_retweet==FALSE)
```

Tidy tweets = *one word per row format*


```
tidy_tweets <- lego %>%
     filter(is_retweet==FALSE) %>%
     select(status_id, text)
```

Tidy tweets = *one word per row format*


```
tidy_tweets <- lego %>%
     filter(is_retweet==FALSE) %>%
     select(status_id, text)
```

Run this code and have a look at the dataframe!

Tidy tweets = *one word per row format*


```
tidy_tweets <- lego %>%
     filter(is_retweet==FALSE) %>%
     select(status_id, text) %>%
     unnest_tokens(word, text)
```

Tidy tweets = *one word per row format*


```
tidy_tweets <- lego %>%
     filter(is_retweet==FALSE) %>%
     select(status_id, text) %>%
     unnest_tokens(word, text)
```

Did it work?

Tidy tweets *= one word per row format*


```
tidy_tweets <- lego %>%
     filter(is_retweet==FALSE) %>%
     select(status_id, text) %>%
     unnest_tokens(word, text)


# Look at the dataframe
view(tidy_tweets)
```

Stop words

= most common words in a language (e.g. "the" or "is")


```
stop_words
```

Stop words = *most common words in a language (e.g. "the" or "is")*

```
stop_words
```

```
A tibble: 1,149 x 2
   word        lexicon
   <chr>       <chr>
 1 a           SMART
 2 a's         SMART
 3 able        SMART
 4 about       SMART
 5 above       SMART
 6 according   SMART
 7 accordingly SMART
 8 across      SMART
 9 actually    SMART
10 after       SMART
... with 1,139 more rows
```

Custom stop words for Internet text data


```
my_stop_words <- tibble(
     word = c(
          "https",
          "t.co",
          "rt",
          "amp",
          "rstats",
          "gt"),
          lexicon = "twitter" )
```

Custom stop words for Internet text data


```
# Check if it worked
View(my_stop_words)
```

Custom stop words for Internet text data

```
# Check if it worked
View(my_stop_words)
```

	word	lexicon
1	https	twitter
2	t.co	twitter
3	rt	twitter
4	amp	twitter
5	rstats	twitter
6	gt	twitter

Adding custom stop words and removing numbers


```
# Connect all stop words
all_stop_words <- stop_words %>%
    bind_rows(my_stop_words)

# Remove numbers
no_numbers <- tidy_tweets %>%
    filter(is.na(as.numeric(word)))
```

Adding custom stop words and removing numbers


```
# Remove numbers
no_numbers <- tidy_tweets %>%
    filter(is.na(as.numeric(word)))
```

```
## Warning in rlang::eval_tidy(~is.na(as.numeric(word)), <e
nvironment>): NAs
## introduced by coercion
```

Removing stop words with anti_join()


```
# Get rid off all stop words
no_stop_words <- no_numbers %>%
     anti_join(all_stop_words, by = "word")
```

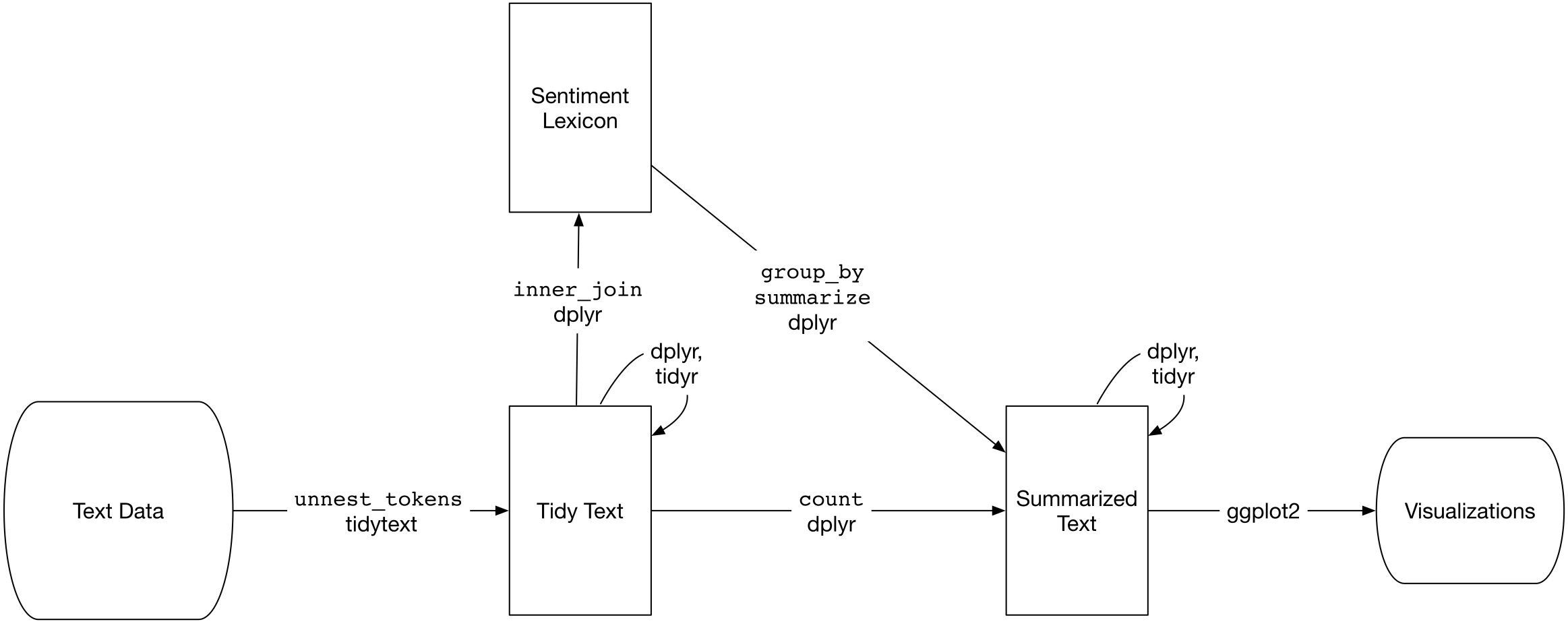
How many words are we left with?

Check in the environment (on the top right hand side).

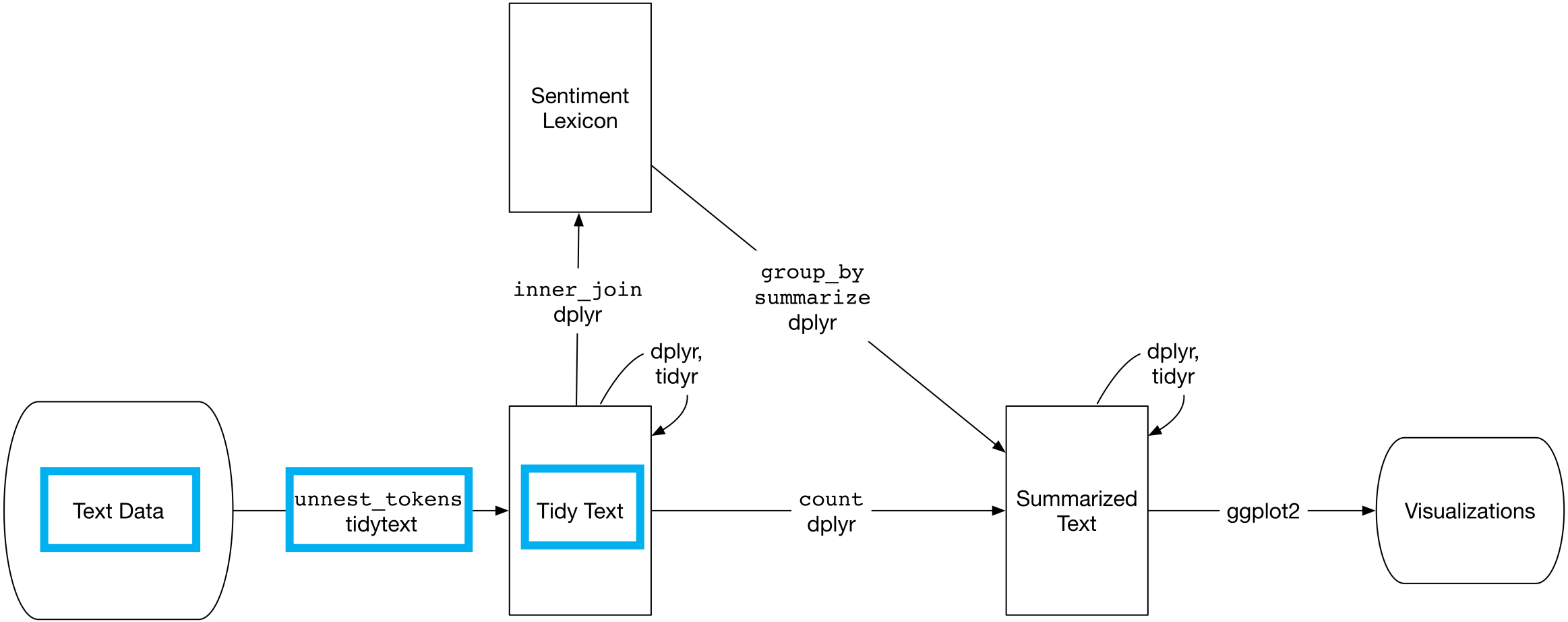
How many rows does `tidy_tweets` *have, how many rows for* `no_stop_words`*?*

Sentiment analysis

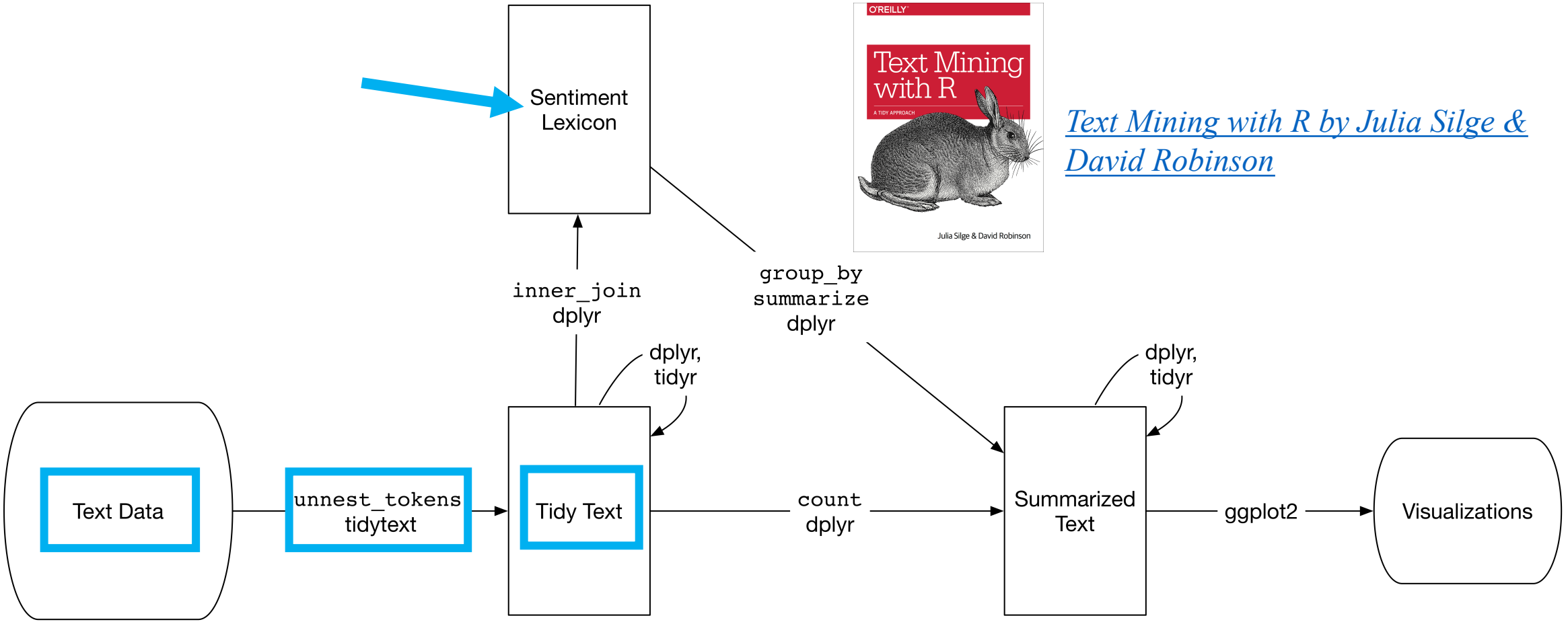
Sentiment analysis



Sentiment analysis



Sentiment analysis



Text Mining with R by Julia Silge & David Robinson

Sentiment analysis?


```
# Add sentiments by using a lexicon
nrc_words <- no_stop_words %>%
    inner_join(get_sentiments("nrc"),
    by="word")

view(nrc_words)
```

Sentiment analysis?


```
pie_words<- nrc_words %>%
    group_by(sentiment)
```

Sentiment analysis?


```
pie_words<- nrc_words %>%
    group_by(sentiment) %>%
    tally
```

Sentiment analysis?

```
pie_words<- nrc_words %>%
    group_by(sentiment) %>%
    tally %>%
    arrange(desc(n))
```

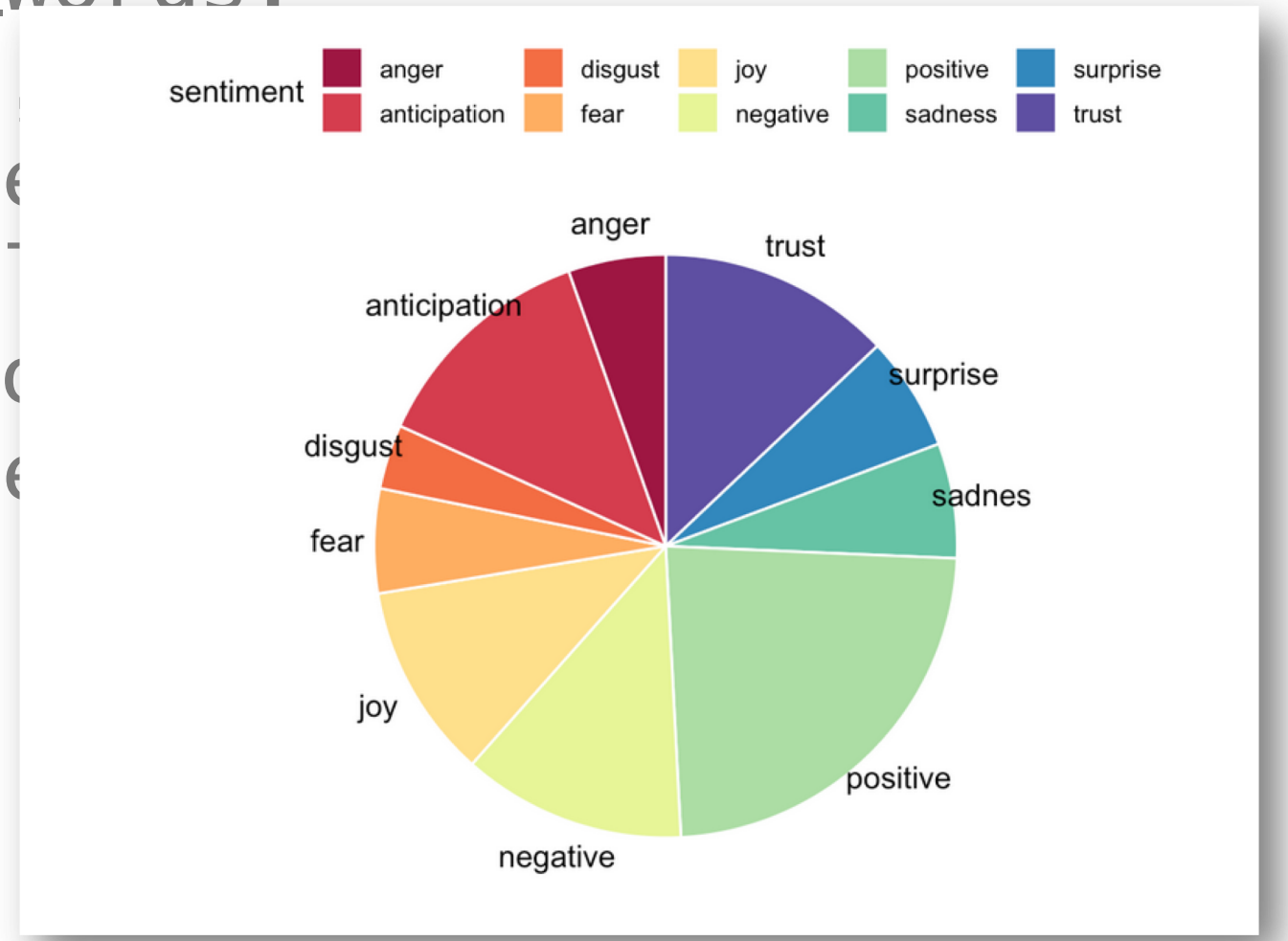
```
pie_words
A tibble: 10 x 2
 sentiment         n
 <chr>         <int>
 positive       2716
 trust          1497
 anticipation   1485
 negative       1432
 joy            1252
 sadness         736
 surprise        735
 fear            669
 anger           624
 disgust         414
```

Pie chart

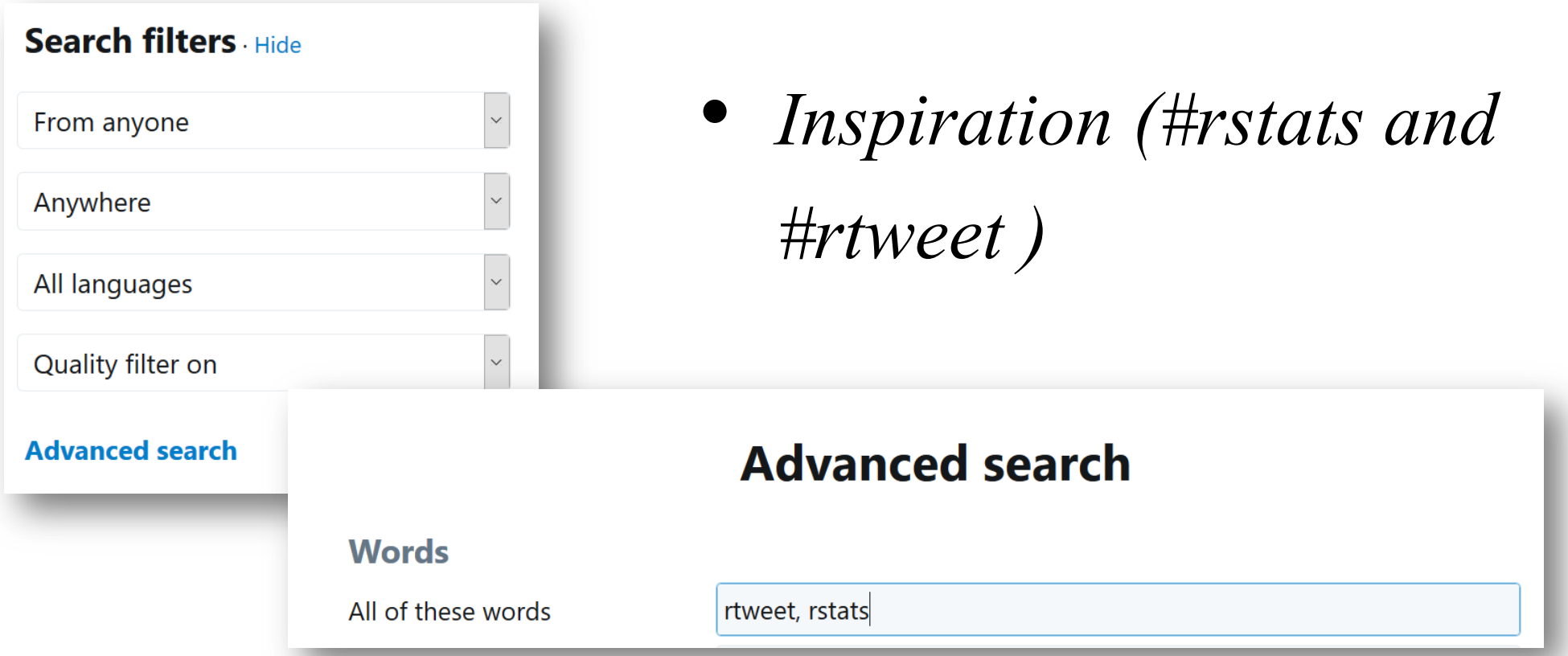

```
ggpubr::ggpie(pie_words,
                    "n",
                    label = "sentiment",
                    fill = "sentiment",
                    color = "white",
                    palette = "Spectral")
```

Pie chart

```
ggpubr::ggpie(pie_words,
              "n"
              labe
              fill
              colo
              pale
```

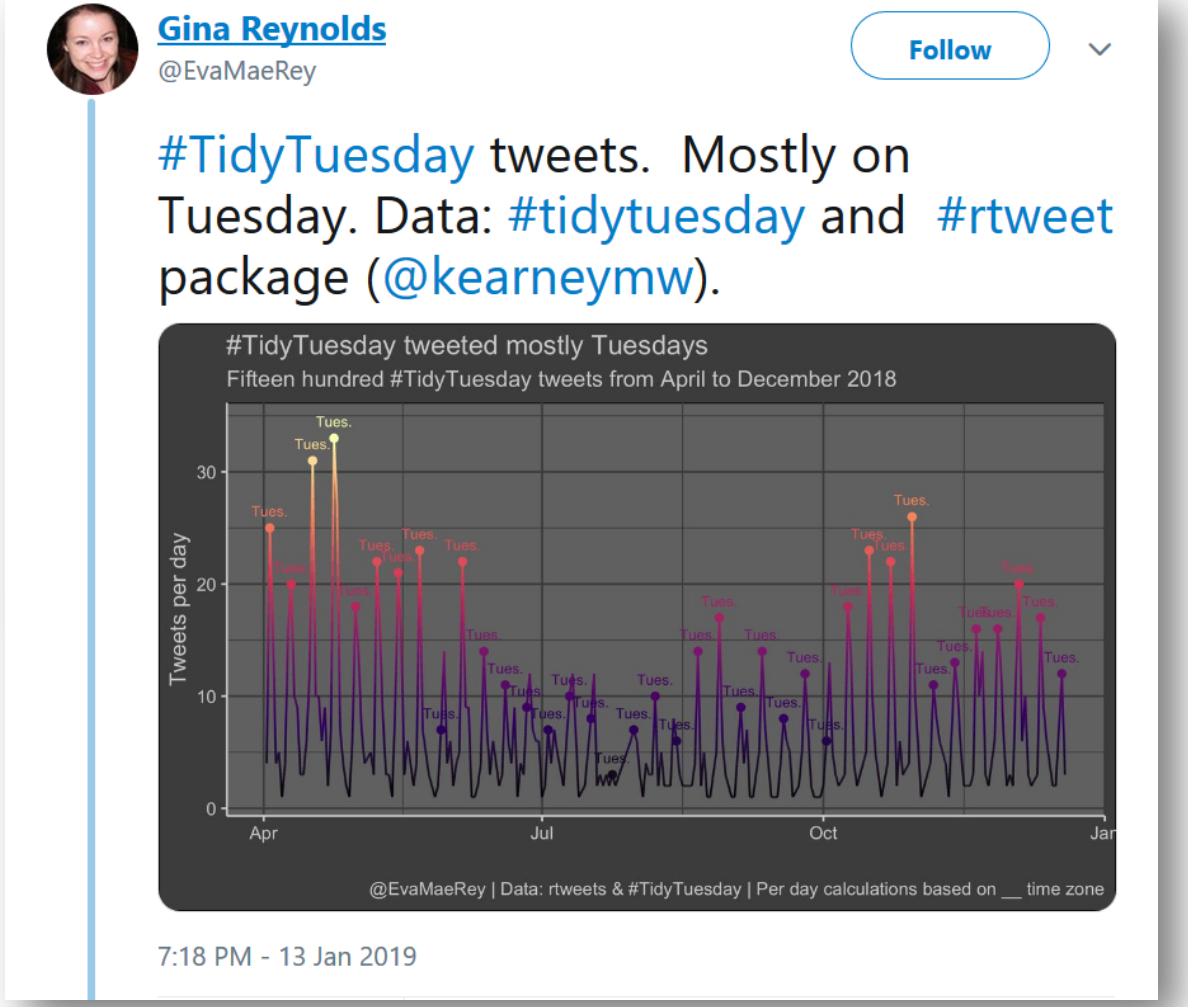


Twitter as a learning resource



- *Inspiration (#rstats and #rtweet)*

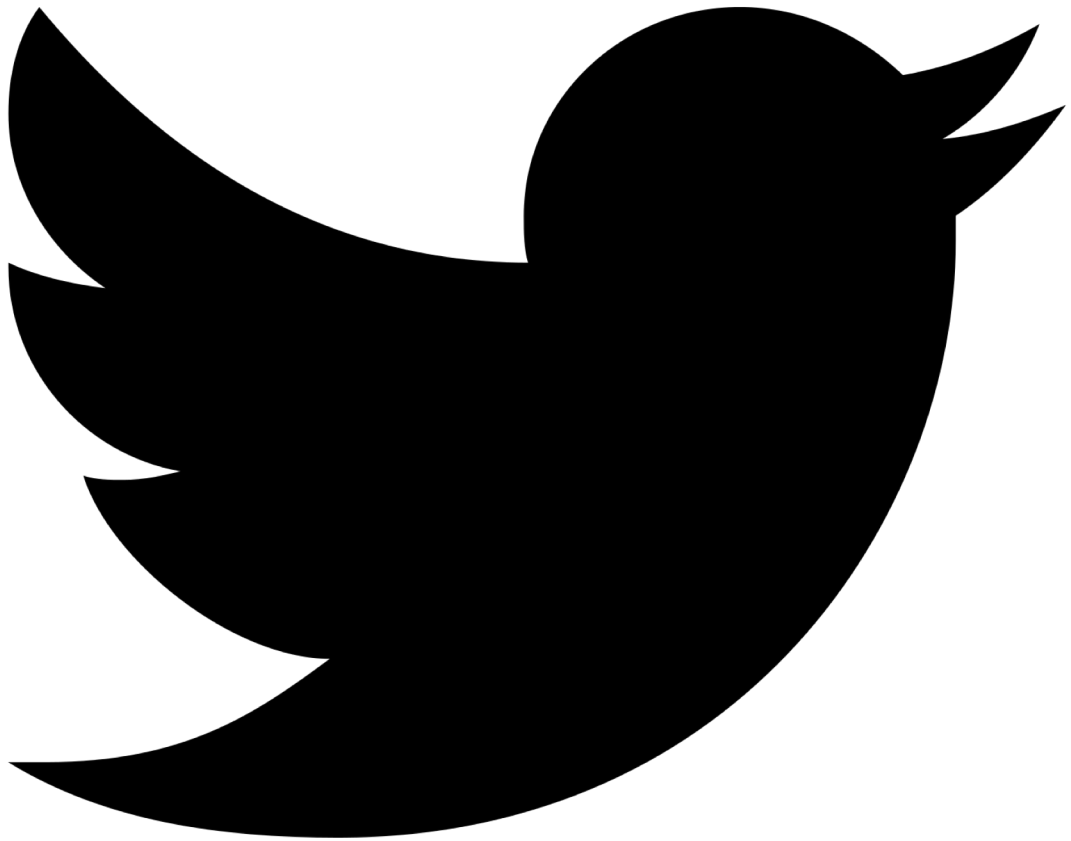
Twitter as a learning resource



Twitter as a learning resource

- *Inspiration (#rstats and #rtweet)*
- *#tidytuesday*
- *Get help & join the community*

You can do many more cool things:



You can do many more cool things:

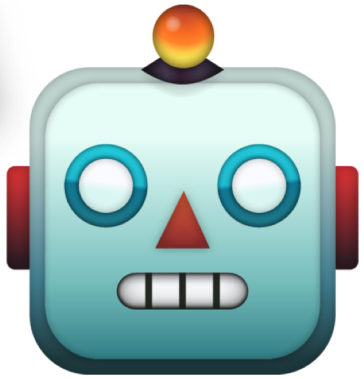
@GlasgowGIST

ggwordcloud

You can do many more cool things:

My most frequently used emoji is …

	emo	n
1	🤖	46
2	😍	35
3	😊	23
4	🎉	22
5	❤️	21
6	💙	18
7	🍀	15
8	💡	13
9	🧠	12
10	🙈	12



emo

You can do many more cool things:

Shiny

[*Trump Tweet Time*]

Let me know about your next rtweet project!

@AnnaHenschel

References

- *Carrillo, M., Han, Y., Migliorati, F., Liu, M., Gazzola, V., & Keysers, C. (2019). Emotional Mirror Neurons in the Rat's Anterior Cingulate Cortex. Current Biology.*

- *Taylor, J., & Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges?. Research Ethics, 14(2), 1-39.*

- *Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. Sociology, 51(6), 1149-1168.*

Links

- *Datenkraken, https://en.wiktionary.org/wiki/datenkraken*
- *Rtweet introduction by Michael W. Kearney, https://mkearney.github.io/nicar_tworkshop/#1*
- *Introduction to tidytext by Julia Silge and David Robinson, https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html*
- *LSE Impact Blog: "Academic journals with a presence on Twitter are more widely disseminated and receive a higher number of citations."*
- *Lego Grad Student*

Thanks to the SGSSS for supporting this workshop.

Slides available via the Open Science Framework

@AnnaHenschel

Anna.Henschel@glasgow.ac.uk