Scraping and visualizing Twitter data

@AnnaHenschel
Anna.Henschel@glasgow.ac.uk
A short introduction to Twitter (and rtweet).
Mike Kearney
@kearneymw

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

Tweet your reply

Hao Ye @Hao_and_Y · 12 Jan 2017

Replying to @kearneymw

Awesome! Will check it out!
Social media + data science == super trendy, right? Well then install my R package, rtweet, and start figuring it out. #rstats
Many scientists are on Twitter!
Why though?
New papers / preprints

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-biology

Hey #scienctwitter, 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

The effect of interpersonal synchrony with a robot on likeability and social motivation osf.io/q9ku8/
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.
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
And ... data!

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!
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
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!
# 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)
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.
Composing an annual report for a grant, the grad student describes an alternate universe where everything is going according to plan.
Jamming gym clothes into his luggage for a conference, the grad student gets the only workout he will have during the entire conference.
Enjoying his work, the grad student solemnly ponders whether he has fallen victim to Stockholm syndrome.
Getting (almost all) tweets of a user

```r
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)
**Tidy tweets** = one word per row format

```r
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

```r
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
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)

<table>
<thead>
<tr>
<th>word</th>
<th>lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>https</td>
<td>twitter</td>
</tr>
<tr>
<td>t.co</td>
<td>twitter</td>
</tr>
<tr>
<td>rt</td>
<td>twitter</td>
</tr>
<tr>
<td>amp</td>
<td>twitter</td>
</tr>
<tr>
<td>rstats</td>
<td>twitter</td>
</tr>
<tr>
<td>gt</td>
<td>twitter</td>
</tr>
</tbody>
</table>
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)), <environment>): NAs
## introduced by coercion
Removing stop words with \texttt{anti_join()}

\begin{verbatim}
# Get rid off all stop words
no_stop_words <- no_numbers %>%
    anti_join(all_stop_words, by = "word")
\end{verbatim}
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

Lego Grad Student @legogradstudent · 15 Oct 2018

*Hoping* to get a phone call for a job interview, the grad student sinks deeper and deeper into *despondency* with every passing hour.
Sentiment analysis
Sentiment analysis
Sentiment analysis

Text Mining with R by Julia Silge & David Robinson
# Add sentiments by using a lexicon
nrc_words <- no_stop_words %>%
  inner_join(get_sentiments("nrc"),
             by="word")

view(nrc_words)
Sentiment analysis?

```r
pie_words <- nrc_words %>%
group_by(sentiment)
```
pie_words<- nrc_words %>%
group_by(sentiment) %>%
tally
Sentiment analysis?

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

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

ggpubr::ggpie(pie_words, "n", label = "sentiment", fill = "sentiment", color = "white", palette = "Spectral"
Twitter as a learning resource

• Inspiration (#rstats and #rtweet)
Twitter as a learning resource

• Inspiration (#rstats and #rtweet)

• #tidytuesday
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:

<table>
<thead>
<tr>
<th>emo</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
</tr>
</tbody>
</table>

My most frequently used emoji is ...
You can do many more cool things:

Trump Tweet Time

When did Trump tweet...

“I will never testify against Trump.” This statement was recently made by Roger Stone, essentially stating that he will not be forced by a rogue and out of control prosecutor to make up lies and stories about “President Trump.” Nice to know that some people still have “guts!”

During...

Meeting  Travel  Executive Time

Based on White House schedules released by Axios.
App by @grrrck using rtweet, Shiny, and ncessy.
Let me know about your next rtweet project!

@AnnaHenschel
References


• Rtweed 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