A guide on extracting and tidying Tweets with R

Julia Bahia Adams

IEL/UNICAMP

CAPES (88887.479810/2020-00)

September 24, 2021

Ethics in social media research

- The untapped possibilities for research (D'ARCY; YOUNG, 2012);
- The challenge of establishing guidelines;
- The terms of use and privacy policies (ZIMMER, 2010).

Ethics in social media research

- The untapped possibilities for research (D'ARCY; YOUNG, 2012);
- The challenge of establishing guidelines;
- The terms of use and privacy policies (ZIMMER, 2010).

Ethics in social media research

- The untapped possibilities for research (D'ARCY; YOUNG, 2012);
- The challenge of establishing guidelines;
- The terms of use and privacy policies (ZIMMER, 2010).

Ethics in social media research

- The role of institutional review boards (see D'ARCY; YOUNG, 2012; ZIMMER, 2010);
- How to establish informed working relationships with users of social media that could potentially be participants in research? (ECKERT, 2014; SOBO; DE MUNCK, 1998);
- The overlap between ethical and methodological issues.

Ethics in social media research

- The role of institutional review boards (see D'ARCY; YOUNG, 2012; ZIMMER, 2010);
- How to establish informed working relationships with users of social media that could potentially be participants in research? (ECKERT, 2014; SOBO; DE MUNCK, 1998);
- The overlap between ethical and methodological issues.

Ethics in social media research

- The role of institutional review boards (see D'ARCY; YOUNG, 2012; ZIMMER, 2010);
- How to establish informed working relationships with users of social media that could potentially be participants in research? (ECKERT, 2014; SOBO; DE MUNCK, 1998);
- The overlap between ethical and methodological issues.

Motivation for Adams (2020)

- It is a variationist sociolinguistic study about stranded prepositions and syntactic variation in Brazilian Portuguese.

1. a. Mas acho que é algo que vale a pena conversar **sobre**.
   b. Fazer sentido da vida é algo que todo mundo volta e meia sente necessidade **de**.
   c. O José não é uma pessoa razoável para se fazer uma tatuagem **com**.
   d. A língua russa parece muito interessante, mas não sei nada **de**.

- How to address the issue of data scarcity?

Motivation for Adams (2020)

- It is a variationist sociolinguistic study about stranded prepositions and syntactic variation in Brazilian Portuguese.

1. **a.** Mas acho que é algo que vale a pena conversar **sobre**.
   b. Fazer sentido da vida é algo que todo mundo volta e meia sente necessidade **de**.
   c. O José não é uma pessoa razoável para se fazer uma tatuagem **com**.
   d. A língua russa parece muito interessante, mas não sei nada **de**.

- How to address the issue of data scarcity?

Motivation for Adams (2020)

- It is a variationist sociolinguistic study about stranded prepositions and syntactic variation in Brazilian Portuguese.

1. **a.** Mas acho que é algo que vale a pena conversar **sobre**.
   b. Fazer sentido da vida é algo que todo mundo volta e meia sente necessidade **de**.
   c. O José não é uma pessoa razoável para se fazer uma tatuagem **com**.
   d. A língua russa parece muito interessante, mas não sei nada **de**.

- How to address the issue of data scarcity?

Results from Adams (2020)

- The script in `R`, based on the `rtweet` package (KEARNEY, 2018; R CORE TEAM, 2020);
- A corpus of ten million words, consisting of roughly 450,000 Tweets;
- Data anonymization strategies.

Disclaimer

This is not an introduction to...

- `R` for linguists (see OUSHIRO, 2014);
- Data science (see WICKHAM; GROLEMUND, 2017);
- `tidyverse` packages (see WICKHAM et al., 2019).

Data harvesting

The script provides four functions to extract and tidy Twitter data.

- `puxaTWEET` extracts Tweets through the Twitter APIs;
- `limpaTWEET` cleans the data to remove the variables that are lists;
- `criaTABELA_principal` creates a CSV file where the new data will be attached to;
- `atualizaTWEETS_principal` adds more data to the main file.

Data harvesting

Access to the Twitter API

create_token(
app = “NameOfApplication”,
consumer_key = ”YourConsumerKey”,
consumer_secret =”YourConsumerSecret”,
access_token = ”YourAccessToken”,
access_secret = ”YourAccessSecret”)

`https://developer.twitter.com/portal`

Data harvesting

- Google Maps API key via Google Cloud Platform Console (see KAHLE; WICKHAM, 2013).

```
# Looking up the coordinates to specify that only Tweets
## published in Brazil should be extracted
br = lookup_coords("brazil")
```

Data harvesting

```
# Searching Tweets
tweets = search_tweets(
q = "'YourChoiceOfQuery"',
n = 10000,
type = "mixed",
lang = "pt",
include_rts = FALSE,
geocode = bz,
max_id = NULL,
parse = TRUE,
token = NULL,
retryonratelimit = TRUE,
verbose = FALSE,
show_col_types = FALSE
)
```

Data harvesting

TEXT
SCREEN_NAME
CREATED_AT
SOURCE
DISPLAY_TEXT_WIDTH
REPLY_TO_SCREEN_NAME
IS_QUOTE
PLACE_NAME
PLACE_TYPE
LOCATION
FOLLOWERS_COUNT
FRIENDS_COUNT
FAVOURITES_COUNT

Data harvesting

- source("researchTwitter.R")
- `atualizaTWEETS_principal`
- Rate limit

References

ADAMS, J. B. **Um estudo sobre preposition stranding e orphaning em falantes de português brasileiro**. Relatório final do processo no. 18/24511-8, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), 2020.

D'ARCY, Alexandra; YOUNG, Taylor Marie. Ethics and social media: Implications for sociolinguistics in the networked public. **Journal of Sociolinguistics**, v. 16, n. 4, p. 532-546, set. 2012. DOI 10.1111/j.1467-9841.2012.00543.x.

ECKERT, Penelope. "Ethics in linguistic research". In: PODESVA, Robert J.; SHARMA, Deyvani. (Eds.). **Research Methods in Linguistics**. New York: Cambridge University Press, 2014. p. 11-26.

KAHLE, David; WICKHAM, Hadley. ggmap: Spatial Visualization with ggplot2. **The R Journal**, v. 5, n. 1, p. 144-161, 2013.

KEARNEY, Michael W. **rtweet: Collecting Twitter data**. Comprehensive R Archive Network, 2018. DOI 10.5281/zenodo.2528481.

OUSHIRO, Livia. "TRATAMENTO DE DADOS COM O R PARA ANÁLISES SOCIOLINGUÍSTICAS", p. 134-177. In: Raquel Meister Ko. Freitag (Organizadora). **Metodologia de Coleta e Manipulação de Dados em Sociolinguística**, São Paulo: Editora Edgard Blücher, 2014.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2020.

SOBO, Elisa J.; DE MUNCK, Victor C. "The Forest of Methods". In: DE MUNCK, Victor C.; SOBO, Elisa J. (Eds.) **Using Methods in the Field: a practical introduction and casebook**. Walnut Creek: Altamira Press, 1998. p. 13-37.

WICKHAM, Hadley; GROLEMUND, Garrett. **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data**. 1. ed. Sebastopol: O'Reilly Media, Inc., 2017.

WICKHAM, Hadley et al. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019. DOI 10.21105/joss.01686.

ZIMMER, Michael. "But the data is already public": on the ethics of research in Facebook. **Ethics and Information Technology**, v. 12, n. 4, pp. 313-325, 2010. DOI 10.1007/s10676-010-9227-5.

Thank you!

j176760@dac.unicamp.br
jubaadams