Automatic Sampling and Analysis of YouTube Data

Introduction

Julian Kohne
Johannes Breuer
M. Rohangis Mohseni

2022-02-21

Goals of this course

After this course you should be able to…

- automatically collect *YouTube* data
- process/clean it
- do some basic (exploratory) analyses of user comments

About us

Julian Kohne

- M.Sc. in Social Psychology, University of Groningen (NL)
- Scientific advisor in GESIS presidential staff / Computational Social Science (CSS) department
  - Main area: New developments of GESIS in the area of digital behavioral data
- Ph.D. student at University of Ulm
  - Field: Social Psychology
  - Topic: Quantifying interpersonal relationships with chat log data (*WhatsApp*)

julian.kohne@gesis.org, @JuuuuKoooo, personal website

About us

Johannes Breuer

- Senior researcher in the team *Data Augmentation* in the department *Survey Data Curation* at *GESIS*
    - digital trace data for social science research
    - data linking (surveys + digital trace data)
- (Co-) Team leader of the team *Research Data & Methods* at the *Center for Advanced Internet Studies* (CAIS)
- Ph.D. in Psychology, University of Cologne
- Previously worked in several research projects investigating the use and effects of digital media (Cologne, Hohenheim, Münster, Tübingen)
- Other research interests
    - Computational methods
    - Data management
    - Open science

johannes.breuer@gesis.org, @MattEagle09, personal website

About us

M. Rohangis Mohseni

- Postdoctoral researcher (Media Psychology) at TU Ilmenau
- Ph.D. in Psychology, University Osnabrueck
- Ongoing habilitation "sexist online hate speech" 👿
- Other research interests
    - Electronic media effects
    - Moral behavior

rohangis.mohseni@tu-ilmenau.de, @romohseni

About you

- What's your name?
- Where do you work?
- What is your experience with `R`?
- Why/how do you want to use *YouTube* for your research?

Prerequisites for this course

- Working version of `R` >= 4.0.0 and a recent version of RStudio
- Some basic knowledge of `R`
- Interest in working with *YouTube* data

Workshop Structure & Materials

- The workshop consists of a combination of lectures and hands-on exercises
- Slides and other materials are available at

https://github.com/jobreu/youtube-workshop-gesis-2022

We also put the PDF versions of the slides and some other materials on the GESIS Ilias repository for this course.

Zoom Etiquette

- If possible, we invite you to turn on your camera (during the lecture and exercise parts); feel free to use a virtual background if you want to

Not so Exhausted Anymore, Ph.D.
@PhdExhausted

What presenting on Zoom feels like 😭

♡ 11K 2:43 PM - Nov 26, 2020

1,436 people are talking about this

Zoom Etiquette

- Please mute your microphones unless you are asking a question
- Asking questions:
    - If you have an immediate question, feel free to ask it via video/audio using the "raise hand" function in *Zoom* or via the text chat (private or public)
    - If you have a question that is not urgent and might be interesting for everybody, please wait until the end of the lecture part, then use the "raise hand" function and ask your question via audio/video
    - During the exercises you can also use "raise hand" + audio/video (if you have a question that might be interesting for others as well) or public or private text chat messages to ask questions
- We will also try to provide (one-on-one) "tech support" during the exercises if that is needed (please contact us via the text chat in case you have any technical issues/questions that we can solve)

Preliminaries: Base R vs. `tidyverse`

In this course, we will use a mixture of base `R` and `tidyverse` code as Julian prefers base `R`, Johannes prefers the `tidyverse`, and Ro is agnostic.

ICYC, here are some opinions for and against using/teaching the `tidyverse`.

Johannes' experience with learning and teaching the `tidyverse` is something like this...

The `tidyverse`

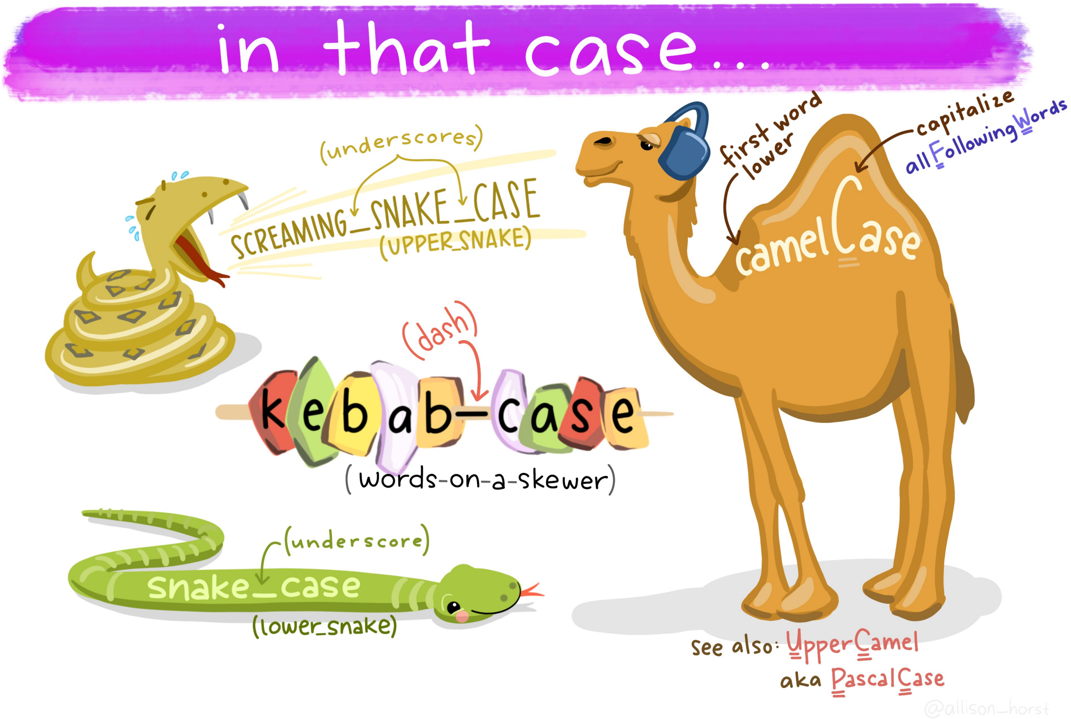
If you've never seen `tidyverse` code, the most important thing to know is the `%>%` (pipe) operator. Put briefly, the pipe operator takes an object (which can be the result of a previous function) and pipes it (by default) as the first argument into the next function. This means that `function(arg1 = x)` is equivalent to `x %>% function()`.

It may also be worthwhile to know/remember that `tidyverse` functions normally produce `tibbles` which are a special type of dataframe (and most `tidyverse` functions also expect dataframes/tibbles as input to their first argument).

If you want a short primer (or need a quick refresher) on the `tidyverse`, you can check out the blog post by Dominic Royé. For a more in-depth exploration of the `tidyverse`, you can, e.g., have a look at the workshop by Olivier Gimenez. And the book *R for Data Science* by Hadley Wickham and Garrett Grolemund (which is available for free online) provides a very comprehensive introduction to the `tidyverse`.

Preliminaries: What's in a name?

Another thing you might notice when looking at our code is that we love 🐍 as much as 🐪.



Artwork by Allison Horst

Course schedule

Monday, February 21st, 2022

When?	What?
10:00 - 11:00	Introduction
11:00 - 11:30	*Coffee break*
11:30 - 12:30	The YouTube API
12:30 - 13:30	*Lunch break*
13:30 - 15:00	Collecting data with the tuber package for R
15:00 - 15:30	*Coffee break*
15:30 - 17:00	Processing and cleaning user comments (in R)

Course schedule

Tuesday, February 22nd, 2022

When?	What?
09:00 - 10:30	Basic text analysis of user comments
10:30 - 11:00	*Coffee break*
11:00 - 12:00	Sentiment analysis of user comments
12:00 - 13:00	*Lunch break*
13:00 - 14:00	Excursus: Retrieving video subtitles
14:00 - 14:30	*Coffee break*
14:30 - 16:00	Practice session, questions, and outlook

Why is *YouTube* relevant?

- Important online video platform
  (Alexa Traffic Ranks, 2019; Konijn, Veldhuis, & Plaisier, 2013)

- Esp. popular among adolescents who use it to, e.g., watch movies & shows, listen to music, and retrieve information
  (Feierabend, Plankenhorn, & Rathgeb, 2016)

- For adolescents, *YouTube* partly replaces TV
  (Defy Media, 2017)

- YouTubers can be social media stars
  (Budzinski & Gaenssle, 2018)

Why is *YouTube* data interesting for research?

- Content producers and users generate huge amounts of data
- These data can be useful for research on media content, communicators, and user interaction
- The data are publicly available and relatively easy to retrieve via the *YouTube* API
- For some further reasons and examples, see Arthurs et al., 2019; Baertl, 2018

Research Examples

- Audience
  - Usage of YouTube
    (Defy Media, 2017)
  - Experiences with YouTube
    (Defy Media, 2017; Lange, 2007; Moor et al., 2010; Oksanen, et al. 2014; Szostak, 2013; Yang et al., 2010)
  - Video consumption
    (Montes-Vozmediano et al., 2018; Tucker-McLaughlin, 2013)
  - Radicalization
    (Ribeiro et al., 2020)
  - Community formation
    (Kaiser & Rauchfleisch, 2020)

Research Examples

- Content
  - Incivility / Hate Speech in comments
    (Döring & Mohseni, 2019a, 2019b, 2020; Obadimu et al, 2019; Spörlein & Schlueter, 2021; Wotanis & McMillan, 2014)
  - Commenter attributes
    (Literat & Kligler-Vilenchik, 2021; Röchert et al., 2020; Thelwall & Mas-Bleda, 2018)
  - Comment characteristics
    (Thelwall, 2018; Thelwall et al., 2012)
  - Video content
    (Kohler & Dietrich, 2021; Utz & Wolfers, 2020)

Research Examples

- Communicator
  - Video production
    (Utz & Wolfers, 2020)
  - Extremism / Ideology
    (Rauchfleisch & Kaiser, 2020, 2021; Dinkov et al., 2019; Ribeiro et al., 2020)
  - Gender / Diversity
    (Chen et al, 2021; Wegener et al., 2020; Thelwall & Mas-Bleda, 2018)
  - Economical aspects
    (Budzinski & Gaenssle, 2018)
  - Channel hierarchy / Ranking
    (Rieder et al., 2018; Rieder et al., 2020)

How to collect *YouTube* data

There are many different ways in which data from *YouTube* and other social media can be collected (see Breuer et al., 2020):

- Manually (e.g., via copy & paste and manual content analysis)
- Using existing data, such as *YouNiverse: Large-Scale Channel and Video Metadata from English YouTube* (also see the accompanying preprint by Ribeiro & West, 2021)
- Automatically via the *YouTube* API or web scraping

Overviews of tools for collecting *YouTube* data

- YouTube Tools collected by the Leibniz-HBI Social Media Observatory
- Social Media Research Tookit by the Social Media Lab at Ryerson University

Tools for the Automatic Sampling of *YouTube* Data without R

- YouTube Data Tools
- Facepager
- Webometric Analyst

Tools for the Automatic Sampling of *YouTube* Data with R

- `vosonSML` (formerly SocialMediaLab)
- `VOSONDash`
- `tuber`
- `youtubecaption`

In this course, we will work with the `tuber` package. The *voson* packages focus more on network data (and analysis) and the `youtubecaption` is for collecting captions (which we will also briefly discuss later on in the workshop).

Comparisons of Approaches for Collecting *YouTube* Data

Software	Type	Can collect	Comment Scope	Needs API Key
YouTube Data Tools 1.22	Website	Channel Info, Video Info, Comments	Only all	No
Webometric 4.1	Standalone app	Channel Info, Video Info, Comments, Video Search	100 most recent or all	Yes
Tuber 0.9.9	R package	Channel Info, Video Info, Comments, Subtitles, All searches	20-100 most recent or all	Yes
vosonSML 0.29.13	R package	Video IDs, Comments	1-x top-level	Yes
youtubecaption 1.0.0	R package	Subtitles	n/a	No

Exemplary Comparison of the Different Tools

Software	Ease of Use	Disadvantages	No. of Comments
YouTube Data Tools 1.22	High	Lacking flexibility, less information	52,243
Webometric 4.1	Low	Only first 5 follow-up comments, no error feedback, undetectable time-outs	49,150
Tuber 0.9.9	Low	Only first 5 follow-up comments	49,139
vosonSML 0.29.13	Low	Lacking flexibility, only comments	50,619

Example data source: Dayum Video

A note on using FOSS

The tools listed before are free and open source software (FOSS). Using FOSS has many advantages (availability, adaptability, etc.). However, one risk associated with using FOSS is that tools are not maintained anymore and, hence, cease to function. After all, people create and maintain these tools in their spare time or as side projects and this work is often not recognized enough (esp. within academia). For this reason it is especially important to acknowledge the work that goes into these tools, e.g., by properly citing them.

```r
citation("tuber")
```

```
##
## To cite package 'tuber' in publications use:
##
##   Gaurav Sood (). tuber: Access YouTube from R. R package version
##   0.9.9.
##
## Ein BibTeX-Eintrag für LaTeX-Benutzer ist
##
##   @Manual{,
##     title = {tuber: Access YouTube from R},
##     author = {Gaurav SOod},
##     note = {R package version 0.9.9},
##   }
```

Any questions so far?