gesis

Leibniz Institute
for the Social Sciences

Automatic Sampling and Analysis of YouTube Data

Introduction

Johannes Breuer, Annika Deubel, & M. Rohangis Mohseni

February 14th, 2023

Leibniz Association

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

Prerequisites for this course

- Working version of `R` >= 4.0.0 and a recent version of RStudio
- Some basic knowledge of `R` (and, ideally, also the `tidyverse`)
- 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-2023

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

Online format

- If possible, we invite you to turn on your camera
- Feel free to ask questions anytime
    - If you have an immediate question during the lecture parts, please send it via text chat, publicly or privately (ideally to a person who is currently not presenting)
    - If you have a question that is not urgent and might be interesting for everybody, you can also use audio (& video) to ask it at the end of a lecture part or during the exercises (please use the use the "raise hand" function in *Zoom* for this)
- We would kindly ask you to mute your microphones when you are not asking (or answering) a question

Course schedule

Tuesday, February 14th, 2023

When?	What?
09:00 - 10:00	Introduction
10:00 - 11:00	The YouTube API
11:00 - 11:15	*Coffee Break*
11:15 - 12:15	Tools for collecting YouTube data
12:15 - 13:15	*Lunch Break*
13:15 - 14:45	Collecting YouTube data with R
14:45 - 15:00	*Coffee Break*
15:00 - 16:30	Processing and cleaning user comments

Course schedule

Wednesday, February 15th, 2023

When?	What?
09:00 - 10:30	Basic text analysis of user comments
10:30 - 10:45	*Coffee Break*
10:45 - 12:15	Sentiment analysis of user comments
12:15 - 13:15	*Lunch Break*
13:15 - 14:45	Excursus: Retrieving video subtitles
14:45 - 15:00	*Coffee Break*
15:00 - 16:30	Practice session, questions, and outlook

About us

Johannes Breuer

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

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

About us

Annika Deubel

- M.Sc. in Applied Cognitive and Media Sciences (University of Duisburg-Essen)
- Ph.D. candidate at the University of Duisburg-Essen
- Researcher in the team *Research Data and Methods* at the *Center for Advanced Internet Studies* (CAIS)
- Main area: health communication and information on social media
- Other research interests:
    - Data and Algorithm Literacy
    - Computational methods

annika.deubel@cais-research.de, @anndeub

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?

Preliminaries: Base `R` vs. `tidyverse`

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

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

Preliminaries: 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()`.

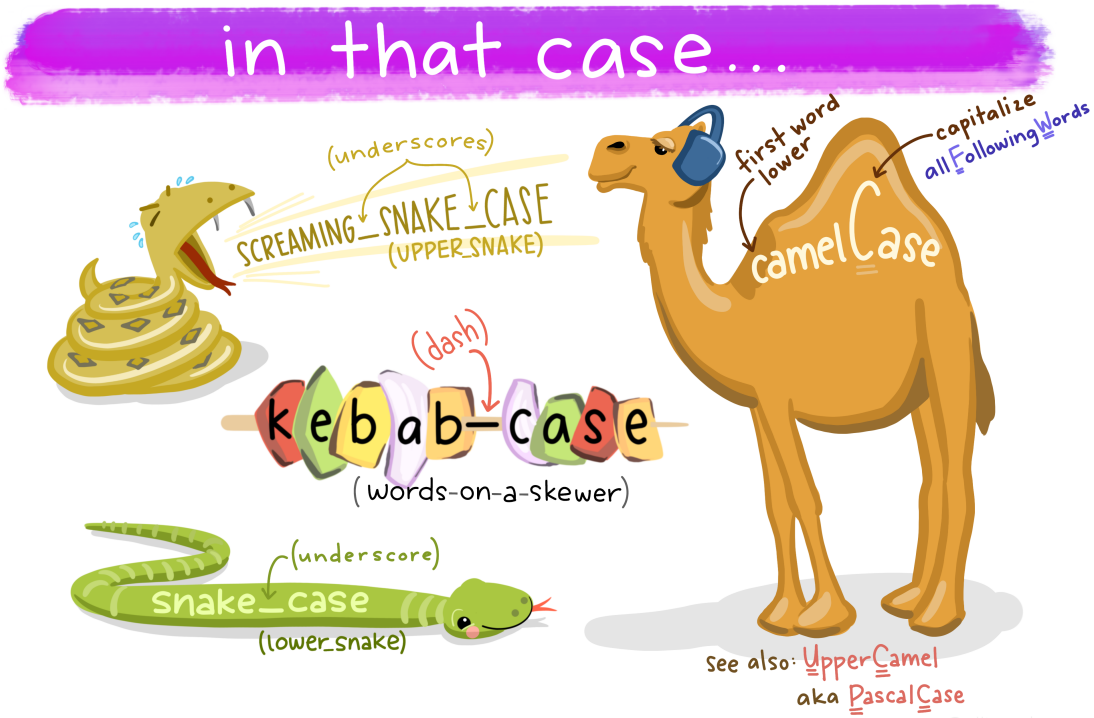
It may also be worthwhile to know/remember that `tidyverse` functions normally produce `tibbles` which are a special type of dataframe (and most `tidyverse` functions also expect dataframes/tibbles as input to their first argument).

Preliminaries: The `tidyverse`

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. 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

Disclaimer: API-based methods

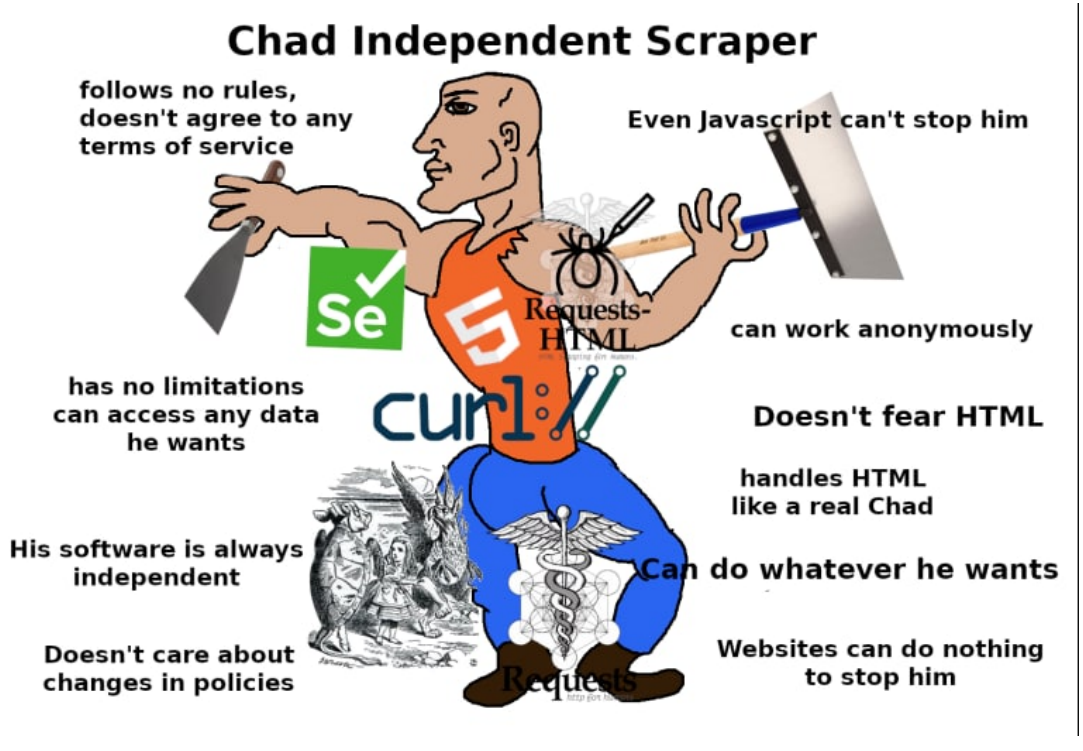
In this course, we will focus on getting data via the YouTube (Data) API. Another popular approach for getting data from social media (and other places online) is web scraping.

While YouTube has recently become more open with regard to data access for academic researchers by introducing the *YouTube Researcher Program*(more on that later), what happened with the *Facebook* APIs following the Cambridge Analytica case or the recent developments around the *Twitter* API show that relying (exclusively) on platform APIs for data access can be risky.

Accordingly, Freelon (2018) argues that researchers interested in social media and other internet data should know/learn how to scrape the web in what he calls the "post-API age" (and a paper by Mancosu & Vegetti, 2020 makes a similar point).

API vs. web scraping

> Get ready for the post api era pic.twitter.com/fbxjVDxWh2
>
> — David Schoch (@schochastics) February 3, 2023



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
    (Albadi et al., 2022; 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)

Any questions so far?