Automatic Sampling and Analysis of YouTube Data

Introduction

Julian Kohne
Johannes Breuer
M. Rohangis Mohseni

2021-02-24

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 presidentidal staff / CSS department
  - Main area: developments of GESIS in the area of digital behavioral data
- PhD Student at University of Ulm / Stanford Social Media Lab
  - Field: Social Psychology
  - Topic: Quantifying interpersonal relationships with chat log data (WhatsApp)

julian.kohne@gesis.org

About us

Johannes Breuer

gesis Leibniz-Institut
für Sozialwissenschaften

- Senior researcher in the team Data Linking & Data Security at the GESIS Data Archive
  - digital trace data for social science research
  - data linking (surveys + digital trace data)
- 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

TECHNISCHE UNIVERSITÄT
ILMENAU

- 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` and 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-2021

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

Zoom Etiquette

- Please mute your microphones unless you are asking a question
- Asking questions:
    - If you have an immediate question during the lecture parts, please send it via text chat to one of the two people who are currently not presenting (the person you contact will either be able to answer your question directly or later forward it to the presenter)
    - 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 in *Zoom* 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 try to provide (one-on-one) "tech support" during the exercises (please contact us via the text chat if 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`

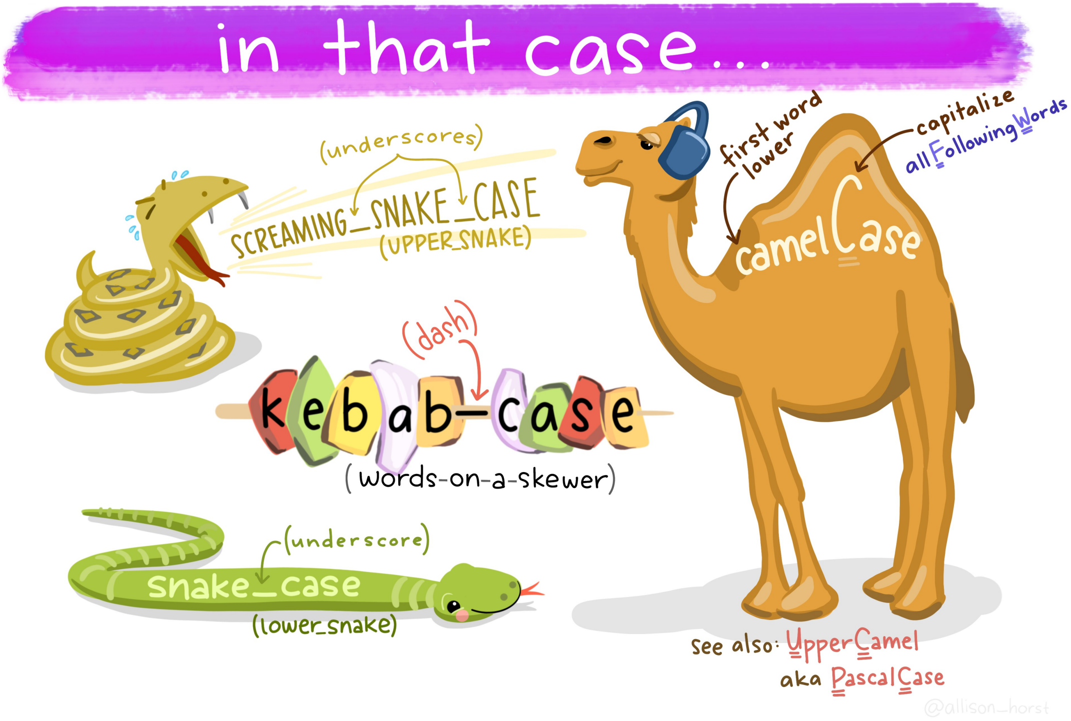
If you've never seen `tidyverse` code, the most important thing to know is the `%>%` (pipe) operator. Put briefly, the pipe operator takes an object (which can be the result of a previous function) and pipes it as the first argument into the next function. This means that `function(data = x, value =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

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 posts by Martin Frigaard or Dominic Royé. For a more in-depth exploration you can, e.g., also have a look at the materials for the workshop *Data Wrangling & Exploration with the Tidyverse in R* that Johannes taught together with Stefan Jünger at GESIS in May 2019.

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

Wednesday, February 24th, 2021

When?	What?
10:00 - 11:00	Introduction: Why is YouTube data interesting for research?
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

Thursday, February 25th, 2021

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?

- Largest / most important online video platform
  (Alexa Traffic Ranks, 2019; Konijn, Veldhuis, & Plaisier, 2013)

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

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

Why is *YouTube* data interesting for research?

- Content producers and users generate huge amounts of data
- Useful for research on media content, communicators, and user interaction
- Data publicly available
- Relatively easy to retrieve via *YouTube* API

Research Examples

- What do people write (content)?
    - Sexist Online Hate Speech
      (Doering & Mohseni, 2019a, 2019b,2020; Thelwall & Mas-Bleda, 2018; Wotanis & McMillan, 2014)
    - Comment characteristics
      (Thelwall, Sud, & Vis, 2012)
    - Subtopics, sentiments, & gender differences
      (Thelwall, 2017; Röchert, Neubaum, Ross, Brachten, & Stieglitz, 2020)

Research Examples

- Who writes it (communicator)?
    - User experiences
      (Defy Media, 2017; Lange, 2007; Moor, Heuvelman, & Verleur, 2010; Oksanen, Hawdon, Holkeri, Naesi, & Raesaenen, 2014; Szostak, 2013; Yang, Hsu, & Tan, 2010)
    - Radicalization
      (Ribeiro et al., 2020)
    - Filter bubble/Structural hierarchy
      (Kaiser & Rauchfleisch, 2020; Rieder et al., 2020)

Tools for the Automatic Sampling of YouTube Data without R

- YouTube Data Tools
- Facepager
- Webometric Analyst

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 with R

- vosonSML (formerly SocialMediaLab)
- tuber

In this course, we will work with the `tuber` package.

Comparisons of Approaches for Collecting *YouTube* Data

Method	Manual Coding	Webometric Analyst	YouTube Data Tools	tuber
Type	n/a	Program	Web service	Package for R
Platforms	All	Win	All	Win, Mac, Linux, Unix
Collected Features	Depends on coding scheme	Channel Info, Video Info, Comments, Video Search	Channel Info, Video Info, Comments, Video List	Channel Info, Video Info, Comments, Subtitles, All searches
Scoping	Depends on coding scheme	100 most recent or all comments	All comments	20-100 most recent or all comments

Pros and Cons of Different Approaches

Method	Manual Coding	Webometric Analyst	YouTube Data Tools	tuber
Need API Key?	No	Yes	No	Yes
Disadvantages	Time-consuming	Only first 5 follow-up comments, no error feedback, undetectable time-outs	Lacking flexibility, fewer infos	Only first 5 follow-up comments due to bug
Ease of Use	High	Low	High	Low
License	n/a	Free for n/c	Open Source	Open Source
Example: Dayum Video (22-02-2019, 2pm)	47,163	44,828	47,153	44,810

Dayum Video / tuber issue

Any questions so far?