gesis Leibniz Institute for the Social Sciences

Automatic Sampling and Analysis of YouTube Data

Recap, outlook, & practice

Johannes Breuer, Annika Deubel, & M. Rohangis Mohseni

February 15th, 2023

Leibniz Association

Course Recap (1)

Session	Example content
Introduction	Why is YouTube data interesting for research?
Tools for collecting YouTube data	Different options for collecting YouTube data
The YouTube API	API access, API requests, quota limits
Collecting YouTube data with R	Collecting channel/video stats & viewer comments
Processing and cleaning user comments	Data wrangling, string operations, emoji dictionaries

Course Recap (2)

Session	Example content
Basic text analysis of user comments	Counting and visualizing the frequencies of words and emojis in comments
Sentiment analysis of user comments	Assigning sentiment scores to words and emojis
Excursus: Retrieving video subtitles	Retrieving and parsing YouTube video subtitles

Where To Go From Here?

Some topics that we did not cover or only briefly touched upon that you might want to explore next/further:

- Analyses for more than one video: use for-loops, functions from the `apply` family or `map` functions from the `purrr` package
- Advanced text mining and NLP (going beyond bag-of-words approaches): check out the tutorials mentioned in the session on basic text analysis or this presentation by Cosima Meyer
- Alternatives to dictionary-based approaches for sentiment analysis: See, e.g., Boukes et al., 2019 and van Atteveldt et al., 2021
- Supervised machine learning for text analysis: The online book Supervised Machine Learning for Text Analysis in R by Emil Hvitfeldt and Julia Silge is an excellent resource here

Shameful Self-Promotion 

We have written a book chapter based on this course which should be published later this year:

Breuer, J., Kohne, J., & Mohseni, M. R. (2023). Using YouTube Data for Social Science Research. In J. Skopek (Ed.), *Research Handbook of Digital Sociology*. Edward Elgar Publishing.

If you are interested in working with *WhatsApp* data (and/or what else you can do with emojis and emoticons in text data), check out the `WhatsR` package (which is also still work in progress) by Julian Kohne.

Acknowledgements

All slides were created with `xaringan`. For the exercises, we used the `unilur` package. The structure and content of the workshop were built using the woRkshoptools package.

A substantial part of the code we used in this workshop - especially that for parsing/processing the *YouTube* data - was written by Julian Kohne.

We thank the *GESIS* Training team for taking good care of the organization of this workshop, and all of you for participating!

Any final questions or comments?

Practice time

You now have some time to start or continue working on your own *YouTube* data analysis project. We'll be around, so feel free to ask questions while you work on or get started with your projects.