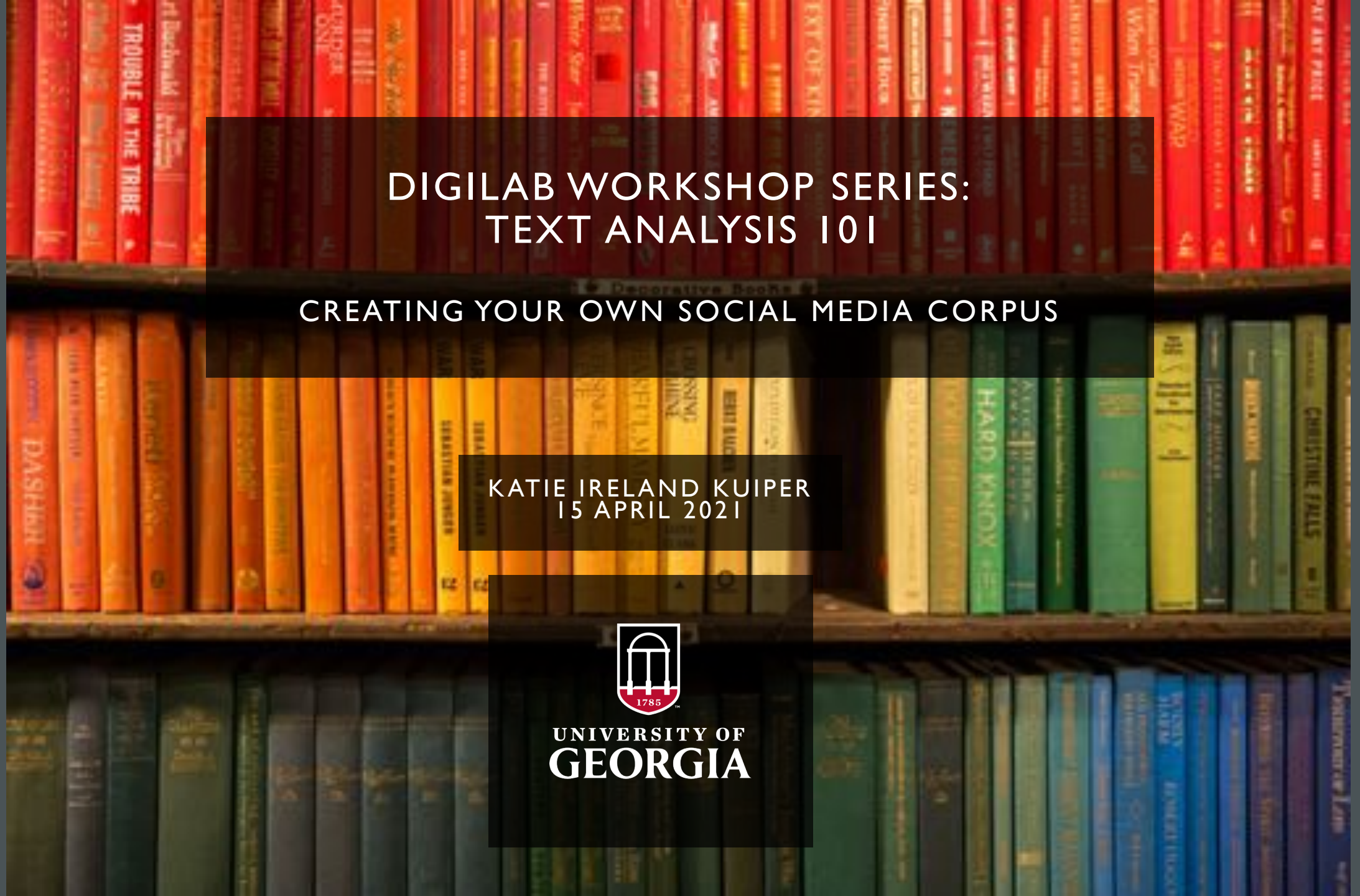


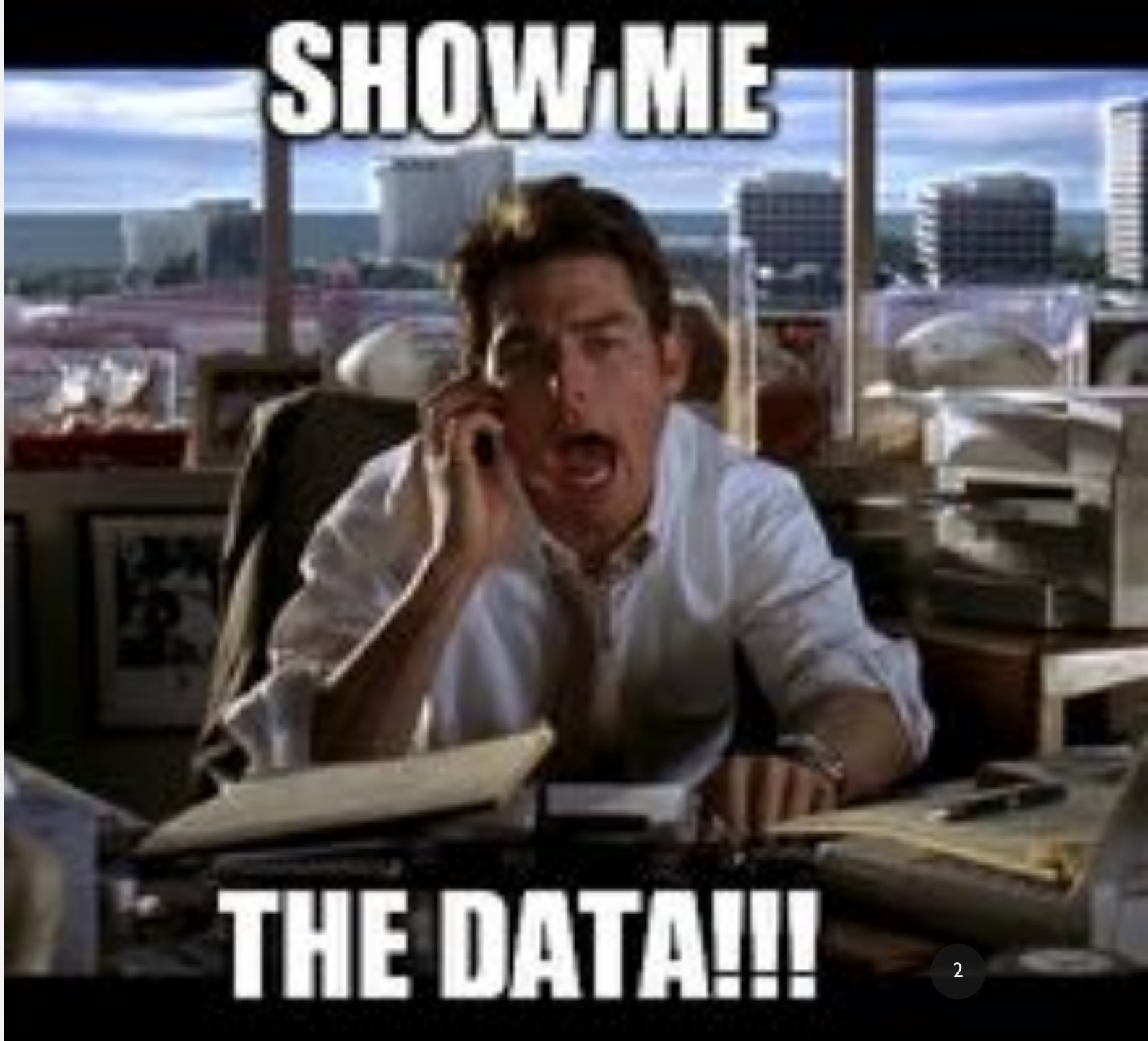
DIGILAB WORKSHOP SERIES: TEXT ANALYSIS 101

CREATING YOUR OWN SOCIAL MEDIA CORPUS

KATIE IRELAND KUIPER
15 APRIL 2021

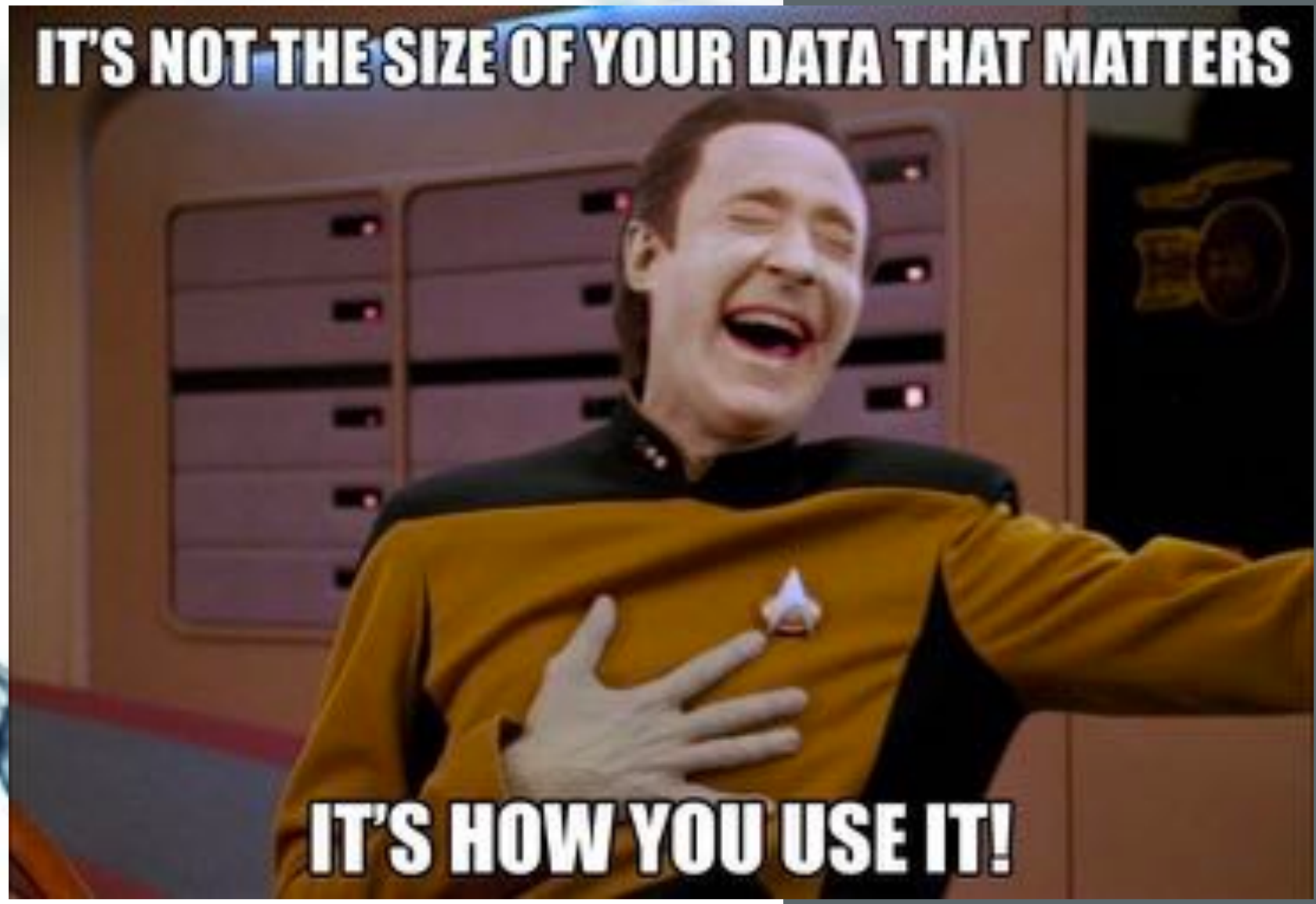
UNIVERSITY OF GEORGIA

IN PREPARATION



SOCIAL MEDIA & TEXT ANALYSIS

- There are now many different types of social media that are useful in a variety of disciplines and areas for research.
- Social media web sites contain various types of services and thus create different formats of data, including text, image, video (Hu & Lui 2012).
  - Social networks (facebook, linkedin), media sharing (youtube, Instagram),discussion forums (reddit, quora), microblogging (Twitter, Facebook), review networks
- Applications include: business, research, event detection, linguistic and language change, network analysis, opinion mining, bioscience, insights into community behavior, and more!



ON

DATA SOURCES

- Linguistic Data Consortium
- The World Wide Web
- Kaggle.com: many pages of social media datasets, including tweets, and others: example: disaster tweets dataset, Instagram data, emojis, reddit, and many many others.
- Stanford SNAP: large network dataset collection, including data from amazon, social media, Wikipedia and others
- Network Repository: including social networks, biological, graph data and tools for analyzing and comparing available datasets

METHODS: CORPUS CREATION

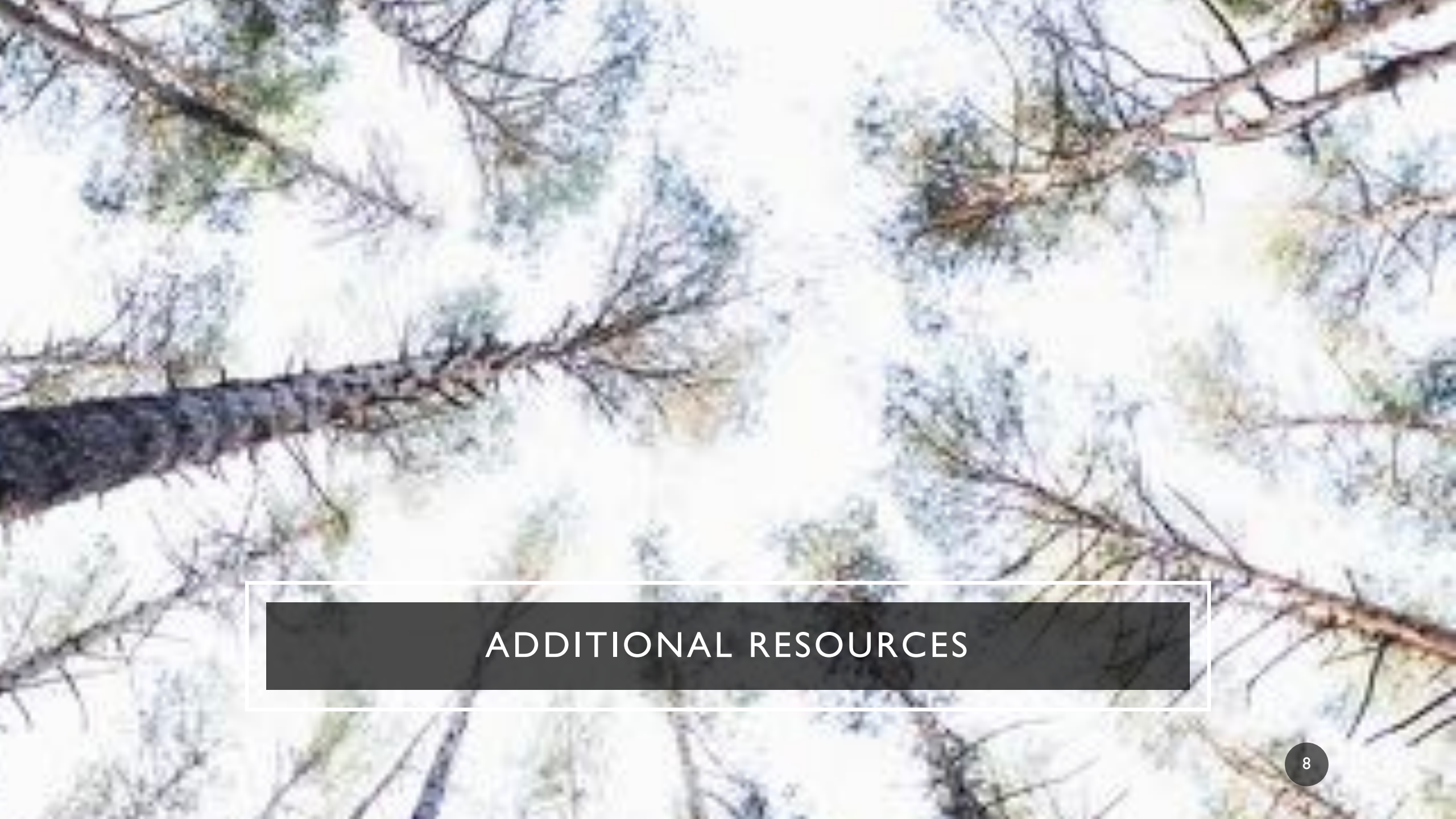
GATHER DATA

ORGANIZE METADATA & DATASET(S)

ANNOTATE AND FORMAT

ANALYZE

- **Tidytext**: helpful for data formatting and visualization; works well with other packages in the Tidyverse (Silge & Robinson 2016)
- **Quanteda**: incredibly useful package; includes preprocessing abilities, dtm function, as well as statistical analyses options like document classification and topic modeling
- **RedditExtractoR**: utilizes Reddit API to obtain posts, comments, and subreddit information
- **Rtweet**: useful package for getting Twitter data, with options for accessing followers, retweets, geolocation, and additional metadata.
- **Tokenizers**: useful options for text analytics, including tokenization and stemming.



ADDITIONAL RESOURCES

iScience Maps: web-based option for getting Twitter data, with options for sorting and analyzing the data

Naoyun: software for connecting Twitter data with Gephi, with options for visualizing "live Twitter activity"

Netlytic: uses APIs to collect public data from Twitter, YouTube, and RSS feeds. Includes free and paid user options, with network and text analytics

Socioviz: get and analyze Twitter data in this web-based environment

The Chorus Project: free web-based option for analyzing and obtaining Twitter data; based out of the UK

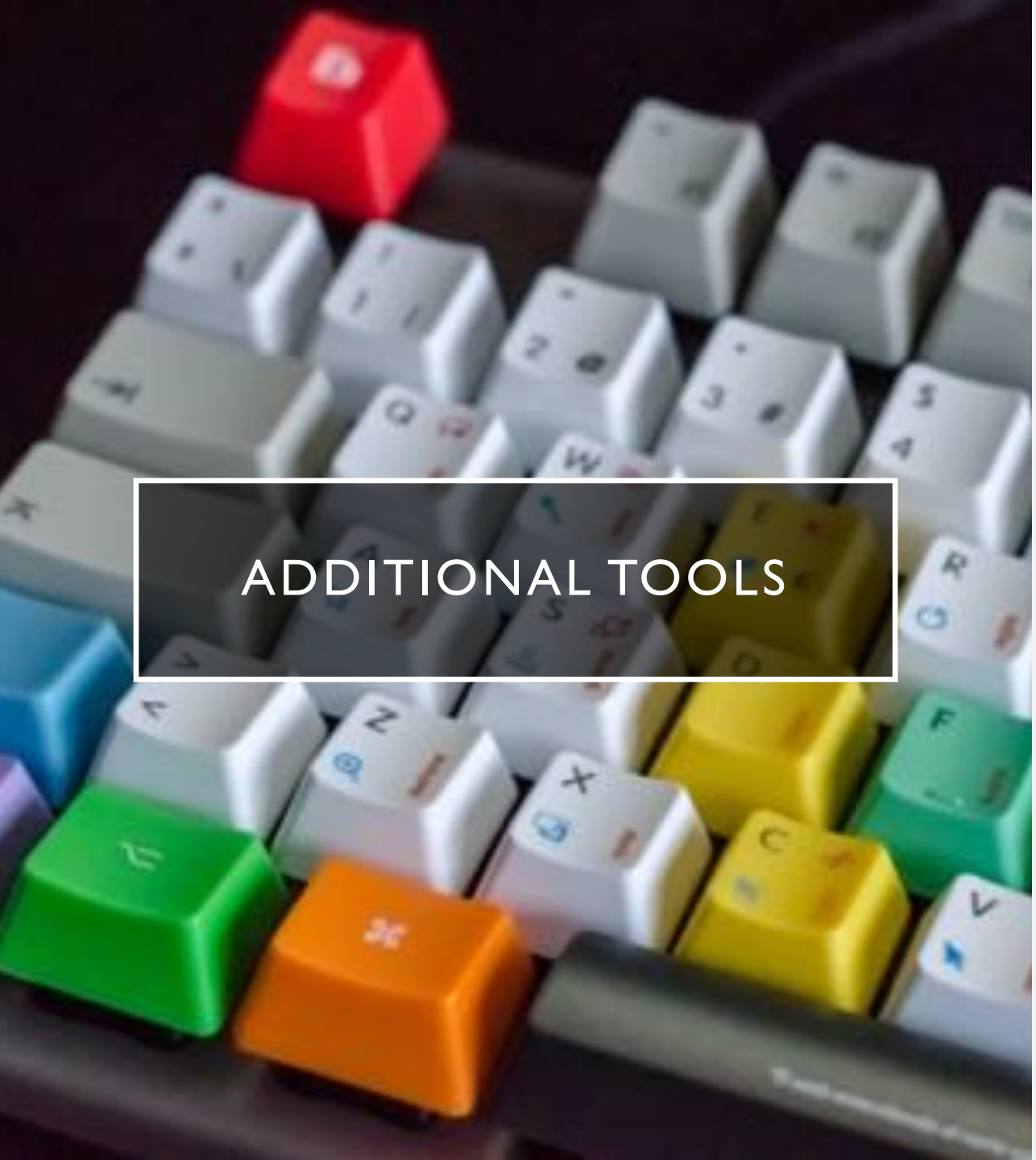
Webometric Analyst: free Windows-based program for gathering data, including Social media, from the Statistical Cybermetrics Research

Digital Footprints: obtain and analyze Facebook data; web-based service available for researchers, based out of Aarhus University

InfoExtractor: no longer maintained, but offers options for getting data from different URLs

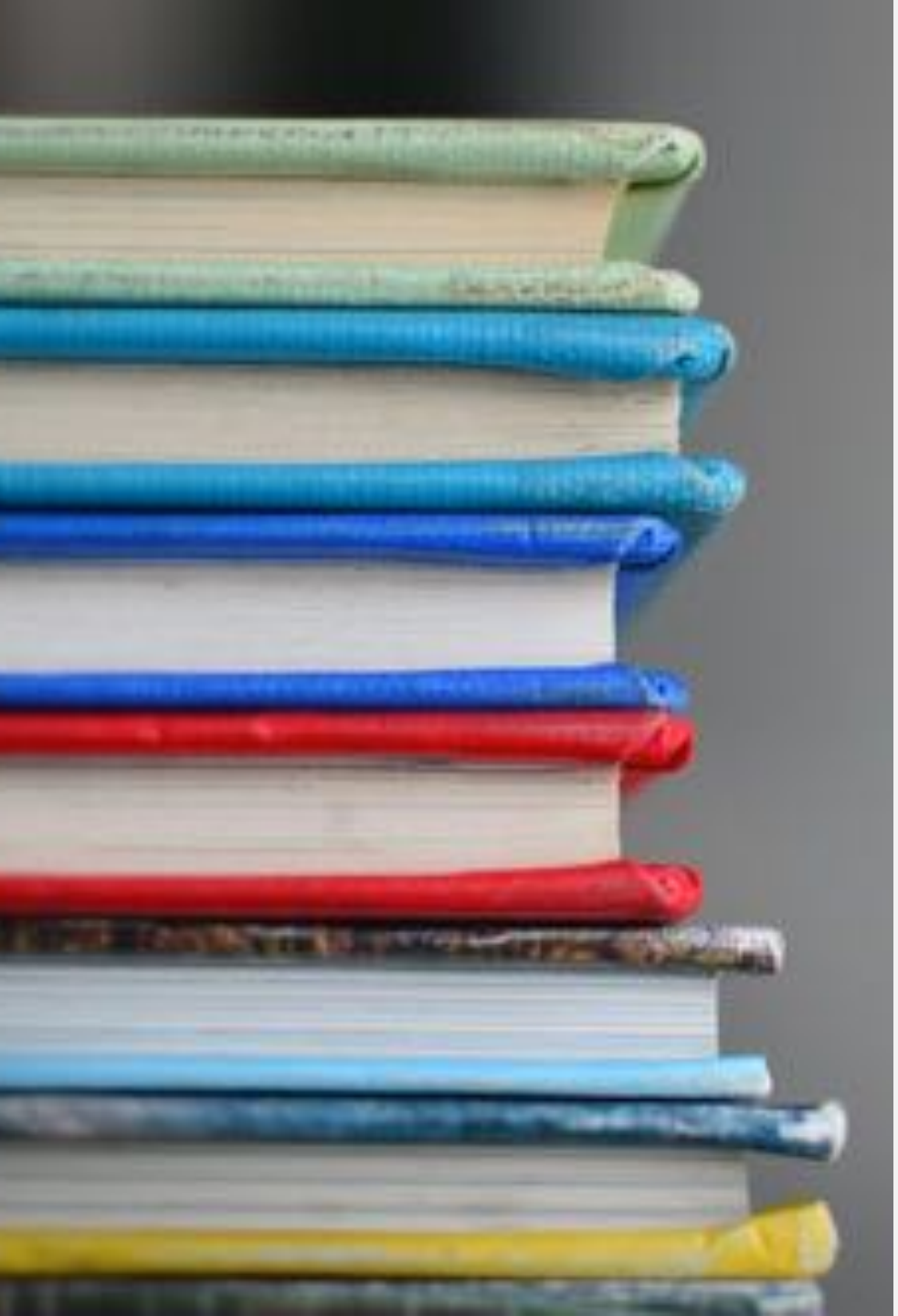
Snoopreport: free for researchers; focus on obtaining Instagram data

- streamR: Access to Twitter Streaming API via R
- twittR: also useful for getting twitter data in R
- Rfacebook: Rfacebook: Access to Facebook API via R
- **instaR:** access Instagram data via the Instagram API; an approved developer account is required



ADDITIONAL TOOLS

- Python libraries:
  - Facebook SDK: Facebook data scraper
  - Twitter scraper: for use with Python 3.6+; can get tweets based on user or other search terms
  - Reddit scraper: interacts with Reddit API and PRAW library to obtain Reddit data
  - Tweepy in Python will interact with Twitter API
- URS: Universal Reddit Scraper; command line tool to obtain Reddit data
- MOZDEH: Windows based programming for gathering social media data
- Webscraping: Chrome plugin
  - Beautiful Soup: useful python library for webscraping; better for smaller amounts of data
  - Scrapy: python library; best for larger datasets
  - Selenium: flexible, also beginner friendly library
  - R: xml2 and rvest work well in conjunction to harvest web data
  - Rcurl & RSelenium

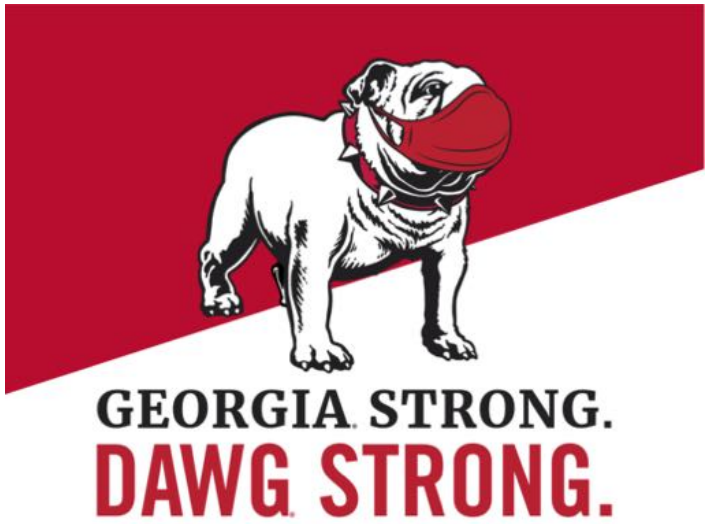


RECOMMENDED RESOURCES

- Silge and Robinson's Text Mining with R
- Beckman et al.'s Intro to Statistical Programming with R
- Social Media Research using R
- Python 3 tutorials
- Social Media Analytics: helpful overview of options and types of analyses
- Text Analysis Glossary
- Corpus Approaches to Social Media (Rüdiger & Dayter 2020)
- Linked-in Learning Tutorials

COURSES AT UGA

- This Fall 2021:
- Natural Language Processing: LING 4570/6570
- Style: ENGL/LING 4826/6826
- American English: ENGL/LING 4010/6010
- Note: These all count toward the Digital Humanities Undergraduate certificate!



DATA OFFICE HOURS

CONSULTATIONS FOR DATA CLEANING, STRUCTURING, AND VISUALIZING

Whether just starting your work, or trying to make sense of your research, schedule an appointment for our Data Office Hours and bring your data (text, archival information, numerical data, etc.) for advice and guidance on your project. Expertise in corpus linguistics, Excel, and R, among other tools for data structuring and visualization.

TUESDAYS • 4:00-5:00
WEDNESDAYS • 2:00-3:00

To schedule an appointment visit:
DIGI.UGA.EDU/RESOURCES

WILLSON CENTER DIGILAB

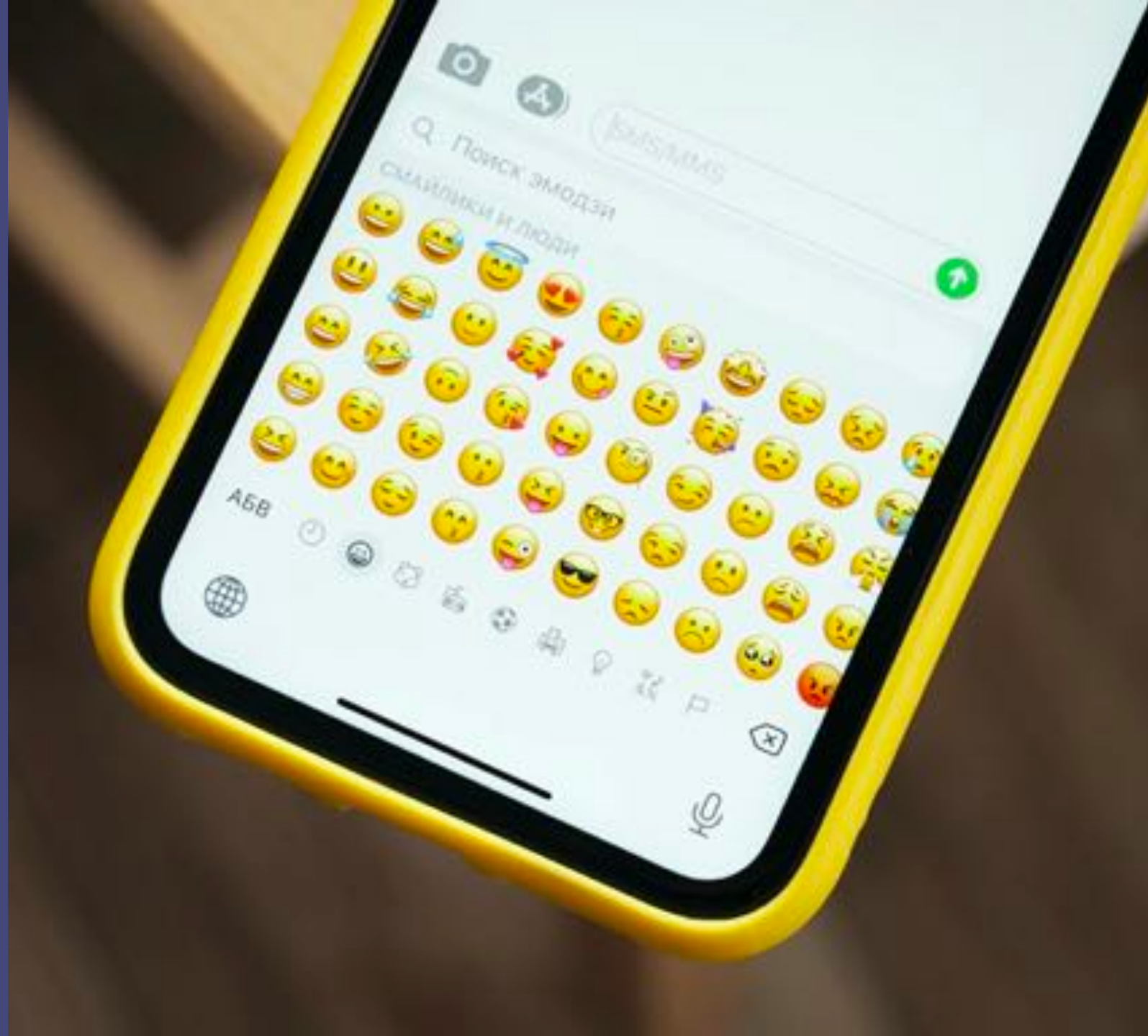
UP NEXT…

22 April: Text Analysis Applications: Social Media

THANKS FOR LISTENING!

KATHERINE.KUIPER25@UGA.EDU

PLEASE FILL OUT THIS SURVEY.



WORKS CITED

- Batrinca, Bogdan & Philip Treleaven. 2014. Social media analytics: a survey of techniques, tools, and platforms. *AI & Society.*
- Beckman, *Matthew, Stéphane Guerrier, Justin Lee, Roberto Molinari, Samuel Orso & Iegor Rudnytskyi. 2020. An Introduction to Statistical Programming with R.* https://smac-group.github.io/ds/index.html
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, William Lowe. (2018). "quanteda: An R package for the quantitative analysis of textual data." *Journal of Open Source Software*, 3(30), 774. doi: 10.21105/joss.00774, https://quanteda.io.
- Bird**,** Steven, Ewan Klein, and Edward Loper. 2019. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit
- Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics.*
- Brown, Simon. 2016. Tips for Computational Text Analysis. https://matrix.berkeley.edu/research/tips-computational-text-analysis
- Bussiere, Kirsten. 2018. Digital Humanities - A Primer.
- Evert, Stefan. 2007. Corpora and collocations. http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf
- Feinerer, Ingo. 2020. Introduction to the tm Package: Text Mining in R. https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
- Freelon, Deen. http://socialmediadata.wikidot.com/
- Han, Na-Rae. Python 3 tutorials. http://www.pitt.edu/~naraehan/python3/.
- Jockers, Matthew. 2020. Introduction to the Syuzhet Package. https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html
- Kearney, Matthew. 2018. R: Collecting and Analyzing Twitter Data: featuring {rtweet}. NiCAR 2018. https://mkearney.github.io/nicar_tworkshop/#1
- Kearney, Matthew, Andrew Heiss, and Francois Briatte. 2020. Package 'rtweet'. https://cran.r-project.org/web/packages/rtweet/rtweet.pdf
- Kuiper, Katie Ireland. 2021. Text Analysis Glossary. *DigiLab.*
- Laudun, John. Text Analytics 101. https://johnlaudun.org/20130221-text-analytics-101/
- Lincoln, Mullen,2018. https://cran.r-project.org/web/packages/tokenizers/vignettes/introduction-to-tokenizers.html

- Machlis, Sharon. 2020. How to search Twitter with rtweet and R. infoworld.com
- 2020.Modern Perl: Why Perl Rules for Text. https://somedudesays.com/2020/02/modern-perl-why-perl-rules-for-text/
- https://monkeylearn.com/text-analysis/
- Millot, Thomas. Photo. Unsplash
- Morikawa, Rei. 2019. 12 Best Social Media Datasets for Machine Learning. https://lionbridge.ai/datasets/12-best-social-media-datasets/
- Nordquist, R. 2019. "Definition and Examples of Text in Language Studies. https://www.thoughtco.com/text-language-studies-1692537
- Norman, Jeremy. Thomas Mendenhall Issues One of the Earliest Attempts at Stylomtery. Historyofinformation.com https://www.historyofinformation.com/detail.php?id=4120
- O'Connor, Brendan, David Bamman, and Noah Smith. 2011. Computational Text Analysis for Scoial Science: Model Assumptions and Complexity.
- Parlante, Nick. 2002. Essential Perl. http://cslibrary.stanford.edu/108/EssentialPerl.html.
- Rivera, Ian. 2019. package RedditExtractoR. https://cran.r-project.org/web/packages/RedditExtractoR/RedditExtractoR.pdf
- Rüdiger, Sophia, and Daria Dayter. 2020. Corpus Approaches to Social Media. In *Studies in Corpus Linguistics*.
- Silge, Julia, and David Robinson. 2017. Text Mining with R: A Tidy Approach. https://www.tidytextmining.com/
- Verma, Abhishek. 2021. Inspirational Quotes from GoodReads website. https://www.kaggle.com/abhishekvermasg1/goodreads-quotes/metadata
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Wiedemann, Gregor & Niekler, Andreas. 2017. Hands-on: A five day text mining course for humanists and social scientists in R. Proceedings of the 1st Workshop on Teaching NLP for Digital Humanities (Teach4DH@GSCL 2017), Berlin.
- Witten, Ian. 2004. Text mining. https://www.cms.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf