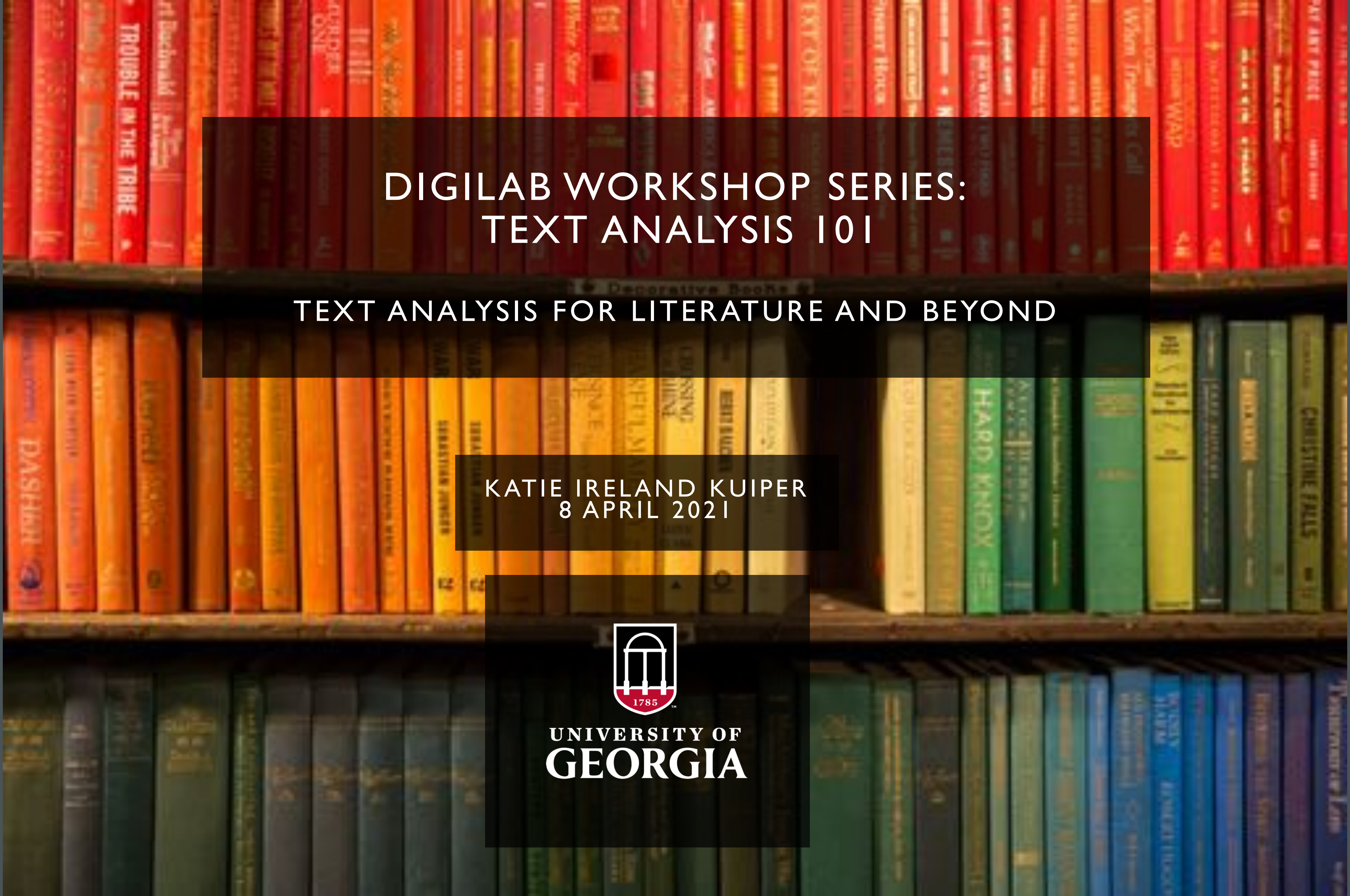


DIGILAB WORKSHOP SERIES: TEXT ANALYSIS 101

TEXT ANALYSIS FOR LITERATURE AND BEYOND

KATIE IRELAND KUIPER
8 APRIL 2021

1785
UNIVERSITY OF
GEORGIA

IN PREPARATION

Download and install:

R and R Studio

THE MEANING OF A WORD IS ITS USE IN THE LANGUAGE

Ludwig Wittgenstein

LITERATURE & TEXT ANALYSIS

- Literature and text analysis make excellent companions!
- Busa’s work in 1946: encoding 11 million tokens of Thomas Aquinas’ writing on IBM punch cards (Sula & Hill 2019)
- on literary dialect (Ellis 1994) including Southern American English, Scots (Burns poetry studied in 1930s by Snyder)
- Revisiting claims of literary scholars:
  - themes in Jane Austen (Fischer-Starke 2010)
- Culpepper on Shakespeare (2002), Klein’s work on slavery texts (2013), semantics, pragmatics (Biber 2011) and much more!!

REPRESENTATION & DRAWING CONCLUSIONS

- Like any data sample, a corpus can be evaluated for the extent to which it represents a ‘population’ (Biber 2011; Brezina 2018; Baker 2006, 2011)

DATA SOURCES

- University of Georgia Corpus Server
- Linguistic Data Consortium
- The World Wide Web
- UGA Library Databases
- The Hathi Trust Digital Library
- Project Gutenberg
- CLARIN historical corpora
- ELTeC: European Literary Text Collection

METHODS

- Frequency analysis
- Analysis of Multiword units/ngrams
- Keyword-in-context; dispersion
- Sentiment analysis

- **Tidytext**: helpful for data formatting and visualization; works well with other packages in the Tidyverse (Silge & Robinson 2016)
- **Syuzhet:** package created specifically for sentiment analysis by Jockers
- **Quanteda**: incredibly useful package; includes preprocessing abilities, dtm function, as well as statistical analyses options like document classification and topic modeling
- **Ggplot2**: great way to visualize your data

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!

TOOLS

- AntConc: A free corpus analysis toolkit for concordancing and corpus-based methods
- Voyant Tools: web-based text reading and analysis environment
- Google Books Ngram Viewer: online search engine that charts the frequencies of any set of comma-delimited search strings
- Wordseer: text analysis environment that combines visualization, information retrieval, and nlp methods
- Tapor: web-based set of text analysis tools
- SketchEngine: text mining app based out of the EU; includes options for your own corpora and includes 500+ other corpora
- JSTOR Lab's Text Analyzer Tool: Includes options for analyzing your texts, identifying topics, and built-in recommendations.
- DataBasic: suite of tools for visualizing and analyzing text (and other types of data)

ADDITIONAL TOOLS

- MALLET: Maps patterns across texts with various tools.
- Perl: was originally created to be a general purpose programming language to help with reports; includes many excellent text-specific functions; supports powerful regular expressions, string processing, and parsing
- Python: StanfordNLP, CoreNLP, gensim, and spacy are all useful libraries.

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:

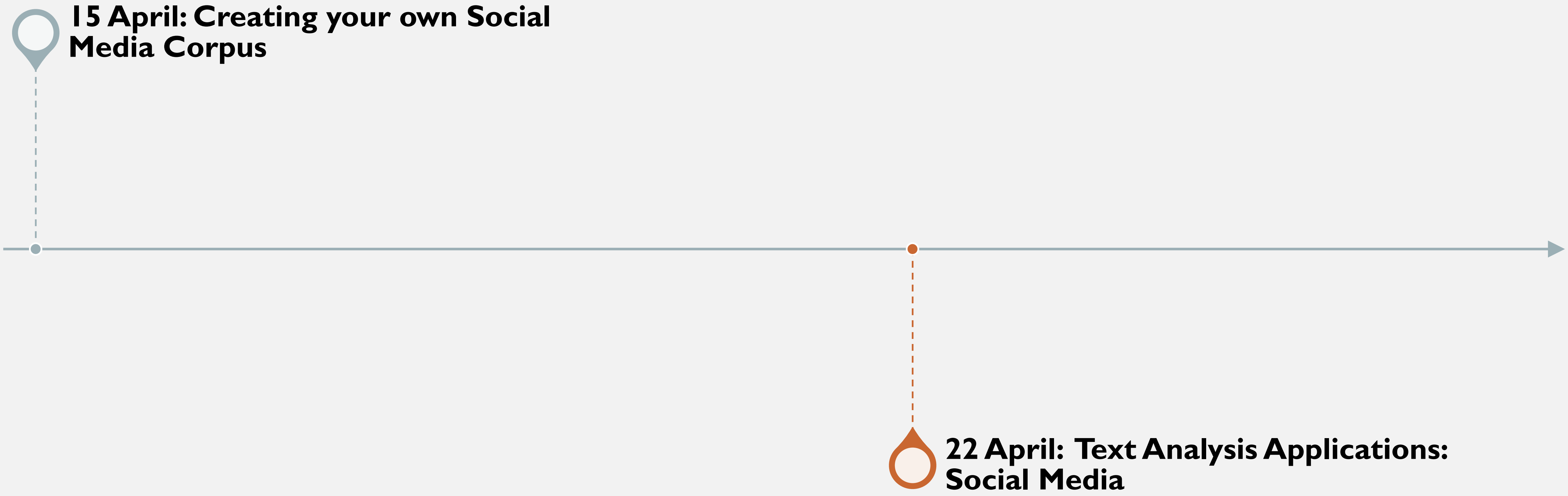
DIGI.UGA.EDU/RESOURCES

WILLSON CENTER DIGILAB

RECOMMENDED RESOURCES

- Language and Literature Journal
- Stylistics: a practical coursebook
- Dialect and Dichotomy
- Evert's work on collocations and corpus methods
- Silge and Robinson's Text Mining with R

COMING UP NEXT…



15 April: Creating your own Social Media Corpus

22 April: Text Analysis Applications: Social Media

THANKS FOR LISTENING!

KATHERINE.KUIPER25@UGA.EDU

PLEASE FILL OUT THIS SURVEY.

WORKS CITED

- Biber, Douglas. 2011. Corpus linguistics and the study of literature: Back to the future? *Scientific Study of Literature.*
- Bird, Steven, Ewan Klein, and Edward Loper. 2019. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit
- Blaette, Andreas. 2020. Introducing the 'polmineR'-package. https://cran.r-project.org/web/packages/polmineR/vignettes/vignette.html.
- 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.
- Cohen Minnick, Lisa. 2004. Dialect and Dichotomy: Literary Representations of African American Speech.
- Evert, Stefan and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference,* University of Birmingham, UK.
- Evert, Stefan. 2003. The CQP Query Language Tutorial.
- Evert, Stefan. 2007. Corpora and collocations. http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf
- Feinerer et al. 2008.
- Feinerer, Ingo. 2020. Introduction to the tm Package: Text Mining in R. https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
- Fischer-Starke, Bettina. 2010. Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries.
- Firth, JR. 1957. Papers in Linguistics. London: OUP.
- Han, Na-Rae. Python 3 tutorials. http://www.pitt.edu/~naraehan/python3/.
- HathiTrust. https://www.hathitrust.org/about.
- Jockers, Matthew. 2020. Introduction to the Syuzhet Package. https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html
- Kuiper, Katie Ireland. 2021. Text Analysis Glossary. *DigiLab.*
- Kretzschmar, William, C. Darwin, C. Brown, D. Rubin, D. Biber. Looking for the Smoking Gun: Principled Sampling in Creating the Tobacco Industry Documents Corpus. *Journal of English Linguistics.* 32:1.
- Laudun, John. Text Analytics 101. https://johnlaudun.org/20130221-text-analytics-101/
- Loria, Steven. 2020. TextBlob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/

- 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
- 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.
- Project Gutenberg. https://www.gutenberg.org
- Sankoff, D. & Sankoff, G. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In Darnell R. (ed.) *Canadian Languages in their Social Context* Edmonton: Linguistic Research Incorporated. 1973. 7–64.
- 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