DIGILAB WORKSHOP SERIES

INTRO TO TEXT ANALYSIS WITH R

KATIE IRELAND KUIPER
18 FEBRUARY 2021

UNIVERSITY OF GEORGIA

INSTALL R AND R STUDIO

R

- Extremely useful programming language; includes a wide variety of packages for working with text(s) and corpora.
- Many packages support working with additional languages besides English, as well as regular expressions and data cleaning.

PACKAGES

- Tidyverse: tidytext
- tokenizers
- readtext
- udpipe

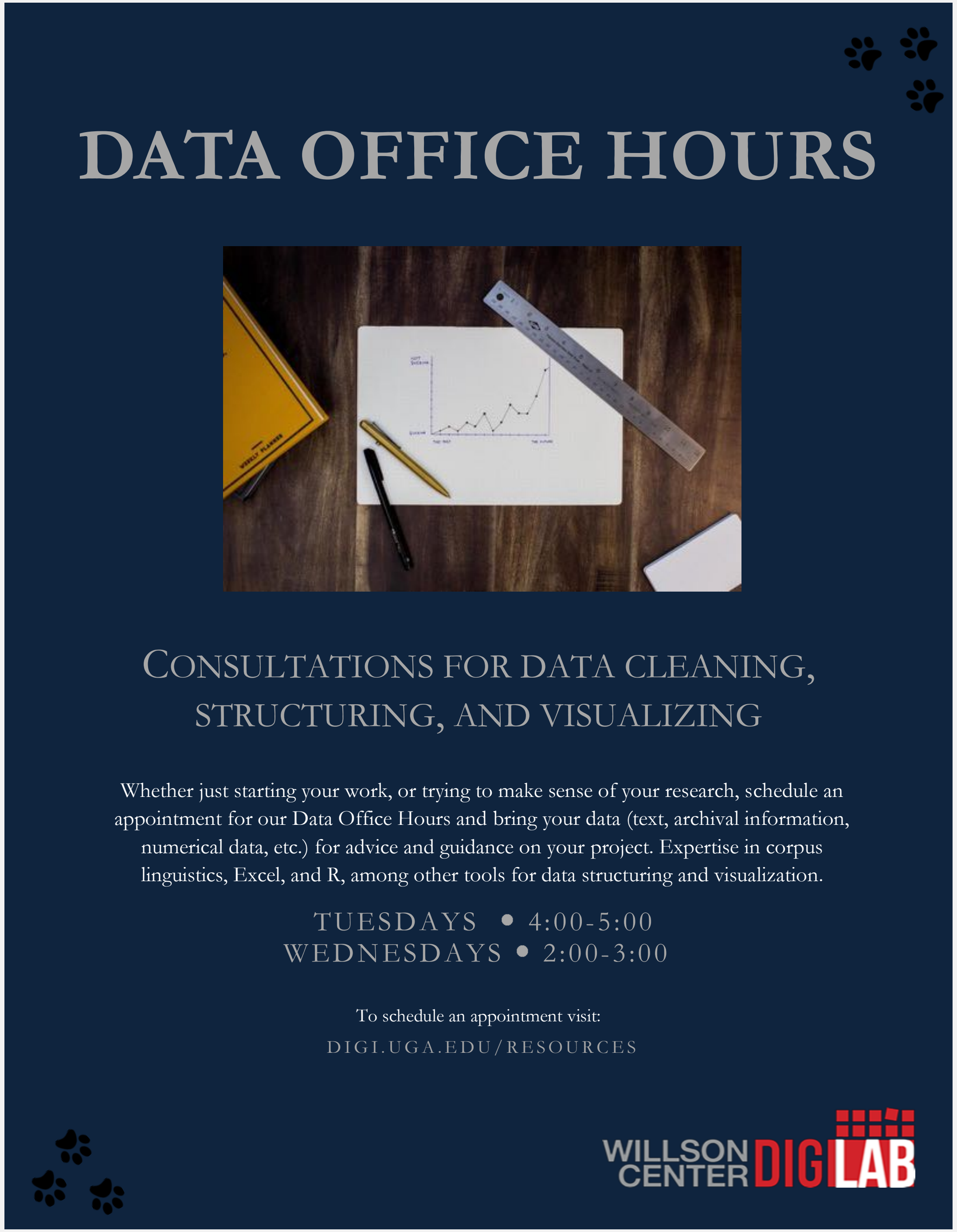
- **Tidytext**: helpful for data formatting and visualization; works well with other packages in the Tidyverse (Silge & Robinson 2016)
- **Textmining/tm:** includes options for data processing, metadata management, and creation of term-document matrices (Feinerer 2020; Feinerer et al. 2008)
- **Syuzhet:** package created specifically for sentiment analysis by Jockers
- **Text2vec**: dtm, vectorizing data, supports topic modeling and collocational analysis, too
- **StringR**: supports regex, pattern matching, useful for string manipulation
- **spacyR**: NLP package originally created for Python; useful for tokenization and works well with quanteda and tidytext
- **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

RESOURCES AT UGA

- Corpus Server
- Upcoming Courses
- Digilab Resources
- Data Office Hours

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

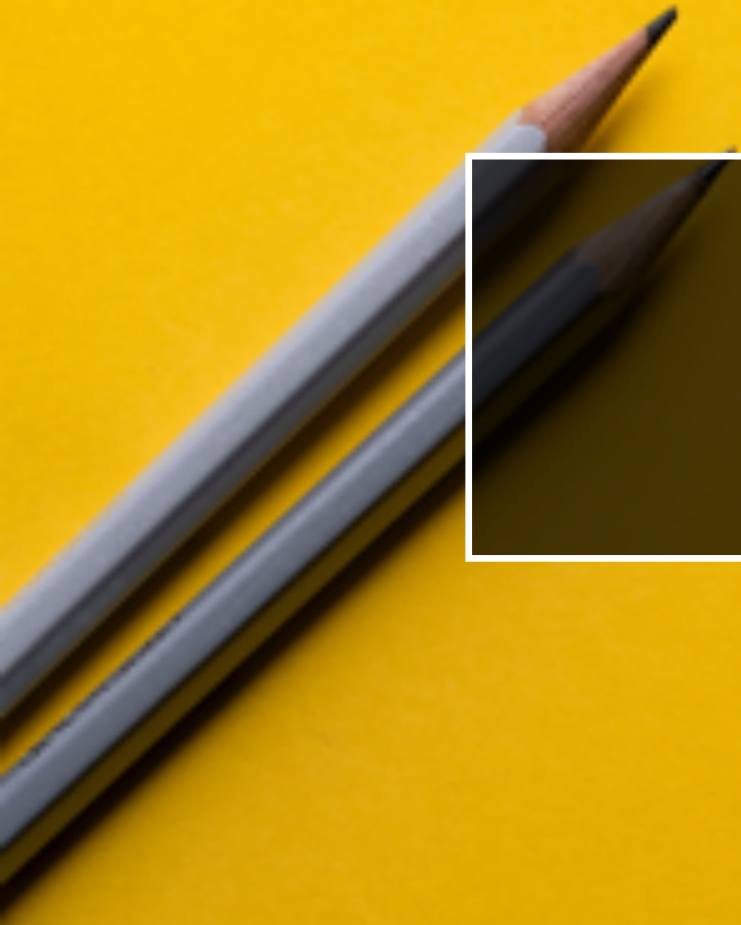
WILLSON CENTER DIGILAB

RECOMMENDED RESOURCES

- Data office hours!
- For more on pos-tagging, check out this tutorial : UDPipe Natural Language Processing Annotation.
- Tidyverse tutorial
- Tokenizers package tutorial

COMING UP NEXT…

- 25 Feb. Advanced R for text analysis



THANKS FOR LISTENING!

KATHERINE.KUIPER25@UGA.EDU

WORKS CITED

- Arnold,Taylor, and Tilton,Lauren. 2017. Basic Text Processing in R. The Programming Historian https://programminghistorian.org/en/lessons/basic-text-processing-in-r.
- Benoit, K, Obeng, A, Watanabe, Kohei, Matsuo, A, Nulty, Paul, and Muller, Stefan. 2020. readtext R package. https://cran.r-project.org/web/packages/readtext/readtext.pdf
- Benoit, K, and Matsuo, Akitaka. 2020. A Guide to Using spacyr. https://cran.r-project.org/web/packages/spacyr/vignettes/using_spacyr.html
- Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A. 2018. quanteda: An R package for the quantitative analysis of textual data. https://quanteda.io. Journal of Open Source Software, 3(30), 774. doi: 10.21105/joss.00774
- Brown, Simon. 2016. Tips for Computational Text Analysis. https://matrix.berkeley.edu/research/tips-computational-text-analysis
- 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.
- Feinerer, Ingo. 2020. Introduction to the tm Package: Text Mining in R. https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
- HathiTrust. https://www.hathitrust.org/about.
- Laudun, John. Text Analytics 101. https://johnlaudun.org/20130221-text-analytics-101/

- Millot, Thomas. Photo. Unsplash
- Mullen, Lincoln. 2018. Introduction to the tokenizers package. https://cran.r-project.org/web/packages/tokenizers/vignettes/introduction-to-tokenizers.html
- Mullen, Lincoln, Keyes, Os, Selivanoc, Dmitriy, Arnold, Jeffrey, Kenneth, Benoit. 2018. tokenizers R package.https://cran.r-project.org/web/packages/tokenizers/index.html
- Project Gutenberg. https://www.gutenberg.org
- Silge, Julia, and David Robinson. 2016. tidytext R package.
- Silge, Julia, and David Robinson. 2020. Text Mining with R: A Tidy Approach. https://www.tidytextmining.com/preface.html
- Wickam, Hadley et al. 2019. Welcome to the tidyverse. https://tidyverse.tidyverse.org/authors.html
- Wijffels, Jan. 2020. UDPipe Natural Language Processing Annotation. https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html#udpipe_the_R_package
- Witten, Ian. 2004. Text mining. https://www.cms.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf