DIGILAB WORKSHOP SERIES

INTRO TO TEXT ANALYSIS WITH R

KATIE IRELAND KUIPER
18 FEBRUARY 2021

UNIVERSITY OF GEORGIA
INSTALL R AND R STUDIO
• 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.
<table>
<thead>
<tr>
<th>Packages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tidyverse: tidytext</td>
</tr>
<tr>
<td>tokenizers</td>
</tr>
<tr>
<td>readtext</td>
</tr>
<tr>
<td>udpipe</td>
</tr>
</tbody>
</table>
• **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:
DIGL.UGA.EDU/Resources

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


• Evert, Stefan. 2003. The CQP Query Language Tutorial.


• HathiTrust. https://www.hathitrust.org/about.

• 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.


• Wijffels, Jan. 2020. UD Pipe Natural Language Processing Annotation. https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html#udpipe_the_R_package