

INTRODUCTION TO TEXT ANALYSIS IN R

April 18, 2019

GOALS

- Introduce the Tidy Text format and package for analyzing texts in R
- Explore three (3) methods for importing text
 - Manual entry
 - Using the gutenbergr package
 - Importing .csv files
- Demonstrate and practice basic methods of text analysis in R
 - Basic text preparation
 - Word frequencies
 - Basic sentiment analysis

TIDY TEXT

What is tidy data?

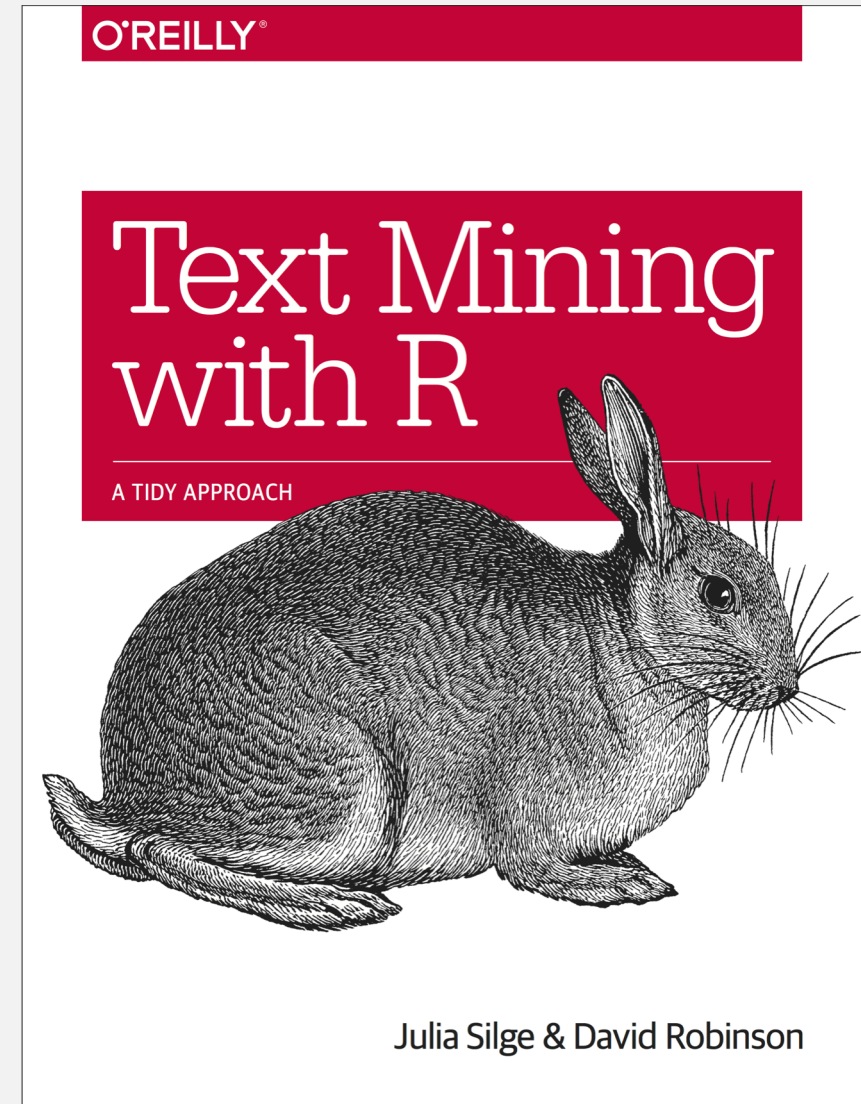
- Each variable is a column
- Each observation is a row
- Each type of observational unit is a table

(Hadley Wickham, 2014)

What is tidy text format?

- A table with one-token-per-row

(Julia Silge & David Robinson, 2019)



<https://www.tidytextmining.com/>

TIDY TEXT

What is tidy data?

- Each variable is a column
- Each observation is a row
- Each type of observational unit is a table

(Hadley Wickham, 2014)

What is tidy text format?

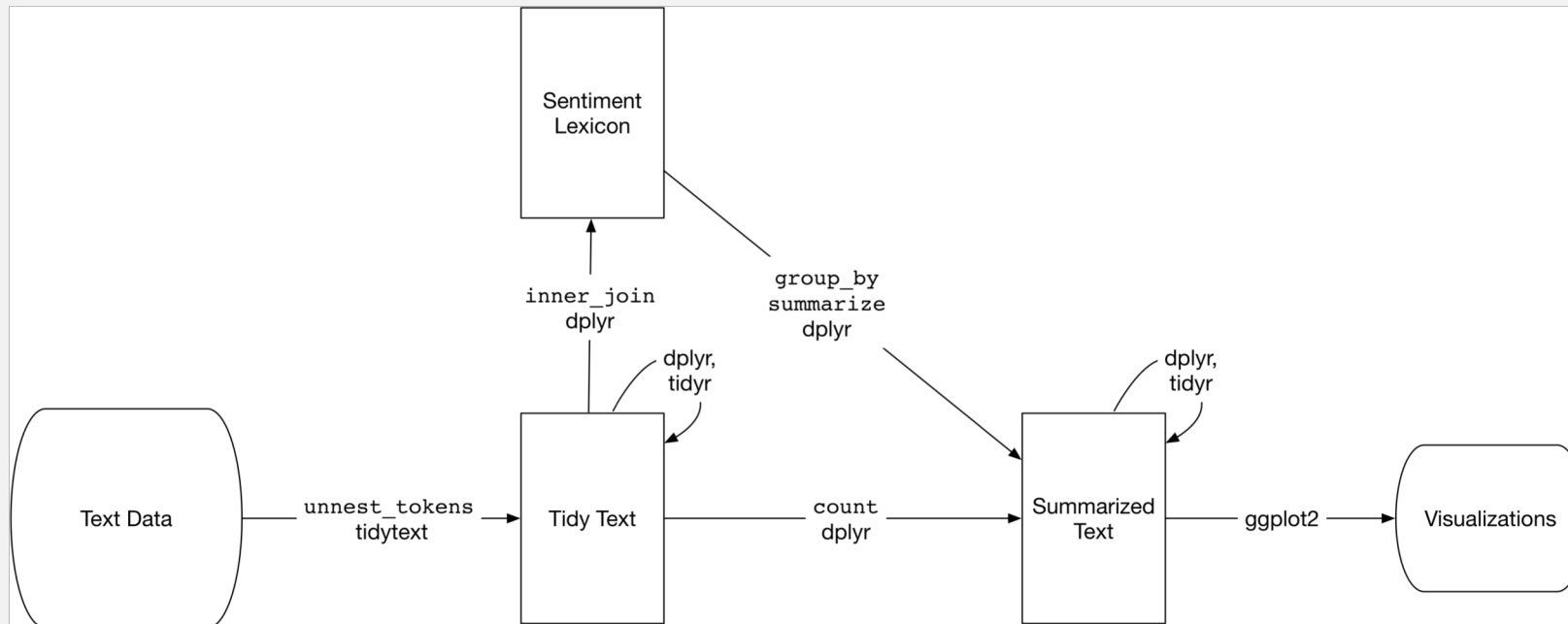
- A table with one-token-per-row

(Julia Silge & David Robinson, 2019)

“A **token** is a meaningful unit of text, such as a word, that we are interested in using for analysis, and tokenization is the process of splitting text into tokens. This one-token-per-row structure is in contrast to the ways text is often stored in current analyses, perhaps as strings or in a document-term matrix. For tidy text mining, the **token** that is stored in each row is most often a single word, but can also be an n-gram, sentence, or paragraph. In the tidytext package, we provide functionality to tokenize by commonly used units of text like these and convert to a one-term-per-row format.”

<https://www.tidytextmining.com/>

SENTIMENT ANALYSIS



(Julia Silge & David Robinson, 2019)

ASSUMPTIONS FOR TODAY

- Focusing on the Tidy Text approach to text analysis in R
- Working with single words as tokens
- While we will be removing stop words (i.e. and, the, a, an, etc.), we will not be concerned word stems (roots), or adjusting for pluralization or verb tense

PACKAGES TO BE USED

- **tidytext** for text tokenization, accessing stopwords and sentiment lexicons
- **dplyr** for data manipulation
- **gutenbergr** for accessing texts via Project Gutenberg
- **ggplot2 / ggthemes** for data visualization
- **stringr** for recognizing patterns using regular expressions
- **tidyr** for creating tidy data

SYNTAX REMINDERS

- R uses `<-` to assign values to variables
- R uses `==` to represent equality of values
- R uses `%>%` to represent a pipe that moves the output of one operation to another operation.

ADDITIONAL RESOURCES

- Basic Text Processing in R (Taylor Arnold & Lauren Tilton):
<https://programminghistorian.org/en/lessons/basic-text-processing-in-r>
- Stop word lists in languages other than English:
 - Example: <https://rdr.io/cran/lisa/man/stopwords.html> (German, Dutch, French, Polish & Arabic)