SENTIMENT ANALYSIS

R case study

John R Little
Duke University

PACKAGES

```
install.packages(c("tidyverse", "tidytext",
"janeaustenr", "wordcloud2"))
```

DUKE UNIVERSITY: LAND ACKNOWLEDGEMENT

I would like to take a moment to honor the land in Durham, NC. Duke University sits on the ancestral lands

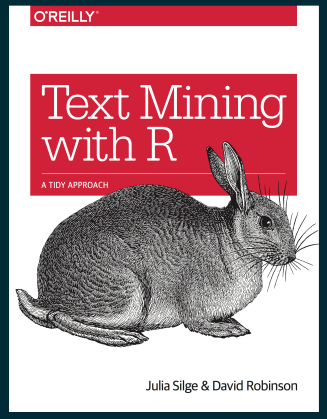
DEMONSTRATION GOALS

- Data *cleaning* & data *wrangling*
- Tokenize corpora (unit of analysis)
- Visualize word clouds (novelty)
- Sentiment analysis
- Analyzing word frequencies (tf-idf)

This is not a text analysis workshop. The foundations of text analysis require considerably more time that we have. This is a demonstration on leveraging tidy packages (tidyverse and tidytext) and sharing resources.

BOOK & RESOURCES

Text Mining with R



BY SILGE & ROBINSON

- www.tidytextmining.com
- juliasilge.github.io/tidytext
- github.com/juliasilge/janeaustenr

TIDY DATA

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

Tidy Data

TIDY TEXT FORMAT

- A token is a meaningful unit of text
- Tokenization is the process of splitting text into tokens `tidytext::unnest_tokens()`
- A table with **one-token-per-row**

OTHER DATA STRUCTURES

STRING

- Text / character vectors

CORPUS

- Raw strings annotated with additional metadata
- A collection of documents

DOCUMENT-TERM MATRIX

- A sparse matrix describing a collection of documents (i.e. *corpus*) with one row for each document and one column for each term. (tf-idf)

OTHER PACKAGES

- tm – *Text Mining Infrastructure in R*
- quanteda – *Package for managing and analyzing textual data*
- gutenbergr – public domain text from **Project Gutenberg**

FURTHER STUDY

Read more of *Text Mining with R: A Tidy Approach*

1. The tidy text format
2. **Sentiment analysis with tidy data**
3. **Analyzing word and document frequency: tf-idf**
4. Relationships between words: n-grams and correlations
5. Converting to and from non-tidy formats
6. **Topic modeling** (unsupervised classification)
7. Case study: comparing Twitter archives
   plus more case studies

FURTHER STUDY

Summer Institute for Computational Social Science
co-founded by Chris Bail & Matthew Salganik

SICSS Text Analysis curriculum