Sentiment Analysis

R case study

John Little

Cntr for Data & Viz

April 13, 2021

This PDF version of the slide deck is a PDF render for your convenience. It may not be up to date. This version was created on 2021-04-14

The actual presentation, created using the xaringan package, is available on GitHub. To access, download the GitHub repository (link in footer), unzip the repo, find the slides directory, open index.html

Otherwise, please continue to the next slide....

Packages

Sentiment Analysis: R case study

```
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 of the Shakori, Eno and Catawba people. This institution of higher education is built on land stolen from those peoples. These tribes were here before the colonizers arrived. Additionally this land has borne witness to over 400 years of the enslavement, torture, and systematic mistreatment of African people and their descendants. Recognizing this history is an honest attempt to breakout beyond persistent patterns of colonization and to rewrite the erasure of Indigenous and Black peoples. There is value in acknowledging the history of our occupied spaces and places. I hope we can glimpse an understanding of these histories by recognizing the origins of collective journeys.

Demonstration Goals

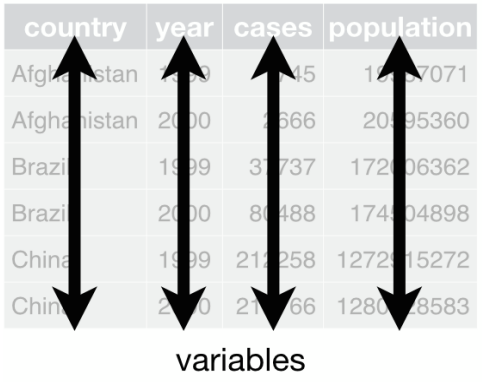
- Data *cleaning* & data *wrangling*
- Tokenize corpora (unit of analysis)
- Visualize word clouds (novelty)
- Sentiment analysis ()
- Analyzing word frequencies (tf-idf)

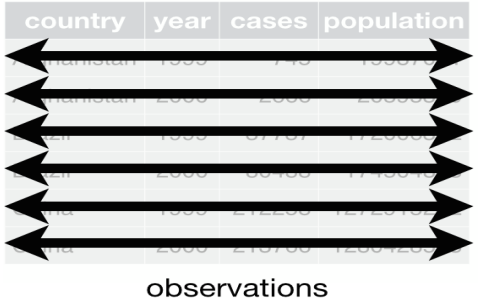
This is not a text analysis workshop. The foundations of text analysis require considerably more time that we have.
This is a demonstration on leveraging tidy packages (tidyverse and tidytext) and sharing resources.

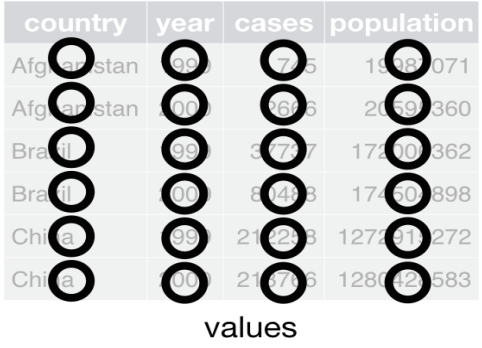
Text Mining with R

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. Chapter 12. *R for Data Science* by Wickham & Grolemund

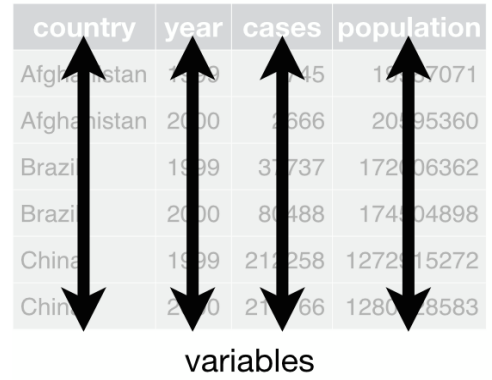
-- Wickham 2014

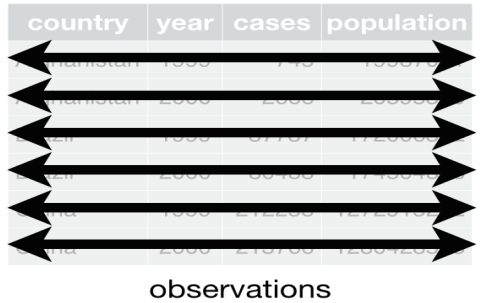
Tidy Text format

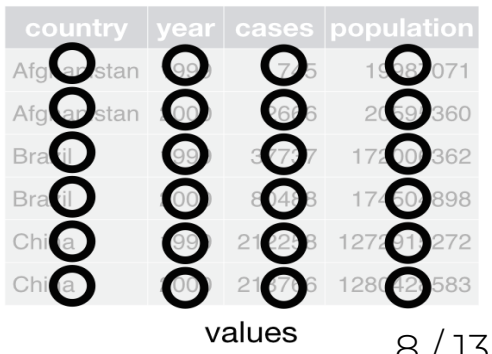
- A token is a meaningful unit of text
- Tokenization is the process of splitting text into tokens

  `tidytext::unnest_tokens()`

- A table with one-token-per-row







Other data structures

String

Text / character vectors

Corpus

Raw strings annotated with additional metadata

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

✤ Not covered in this case study

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

John R Little

Data Science Librarian
Center for Data & Visualization Sciences
Duke University Libraries

https://johnlittle.info
https://Rfun.library.duke.edu
https://library.duke.edu/data