tidytext: Text Mining and Analysis Using Tidy Data Principles in R

Julia Silge

Datassist

David Robinson

Stack Overflow

6 July 2016

Paper DOI: http://dx.doi.org/10.21105/joss.00037
Software Repository: https://github.com/juliasilge/tidytext
Software Archive: http://dx.doi.org/10.5281/zenodo.56714

Summary

The tidytext package (Silge, Robinson, and Hester 2016) is an R package (R Core Team 2016) for text mining using tidy data principles. As described by Hadley Wickham (Wickham 2014), tidy data has a specific structure:

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

Tidy data sets allow manipulation with a standard set of "tidy" tools, including popular packages such as dplyr (Wickham, Francois, and RStudio 2015), ggplot2 (Wickham, Chang, and RStudio 2016), and broom (Robinson et al. 2015). These tools do not yet, however, have the infrastructure to work fluently with text data and natural language processing tools. In developing this package, we provide functions and supporting data sets to allow conversion of text to and from tidy formats, and to switch seamlessly between tidy tools and existing text mining packages.

We define the tidy text format as being one-token-per-document-per-row, and provide functionality to tokenize by commonly used units of text including words, n-grams, and sentences. At the same time, the tidytext package doesn't expect a user to keep text data in a tidy form at all times during an analysis. The package includes functions to `tidy` objects (see the broom package (Robinson et al. 2015)) from popular text mining R packages such as tm (Ingo Feinerer and Meyer 2008) and quanteda (Benoit and Nulty 2016). This allows, for example, a workflow with easy reading, filtering, and processing to be done using dplyr and other tidy tools, after which the data can be converted into a document-term matrix for machine learning applications. The models can then be re-converted into a tidy form for interpretation and visualization with ggplot2.

The following is an example visualization made using tidytext's text mining and sentiment analysis tools.

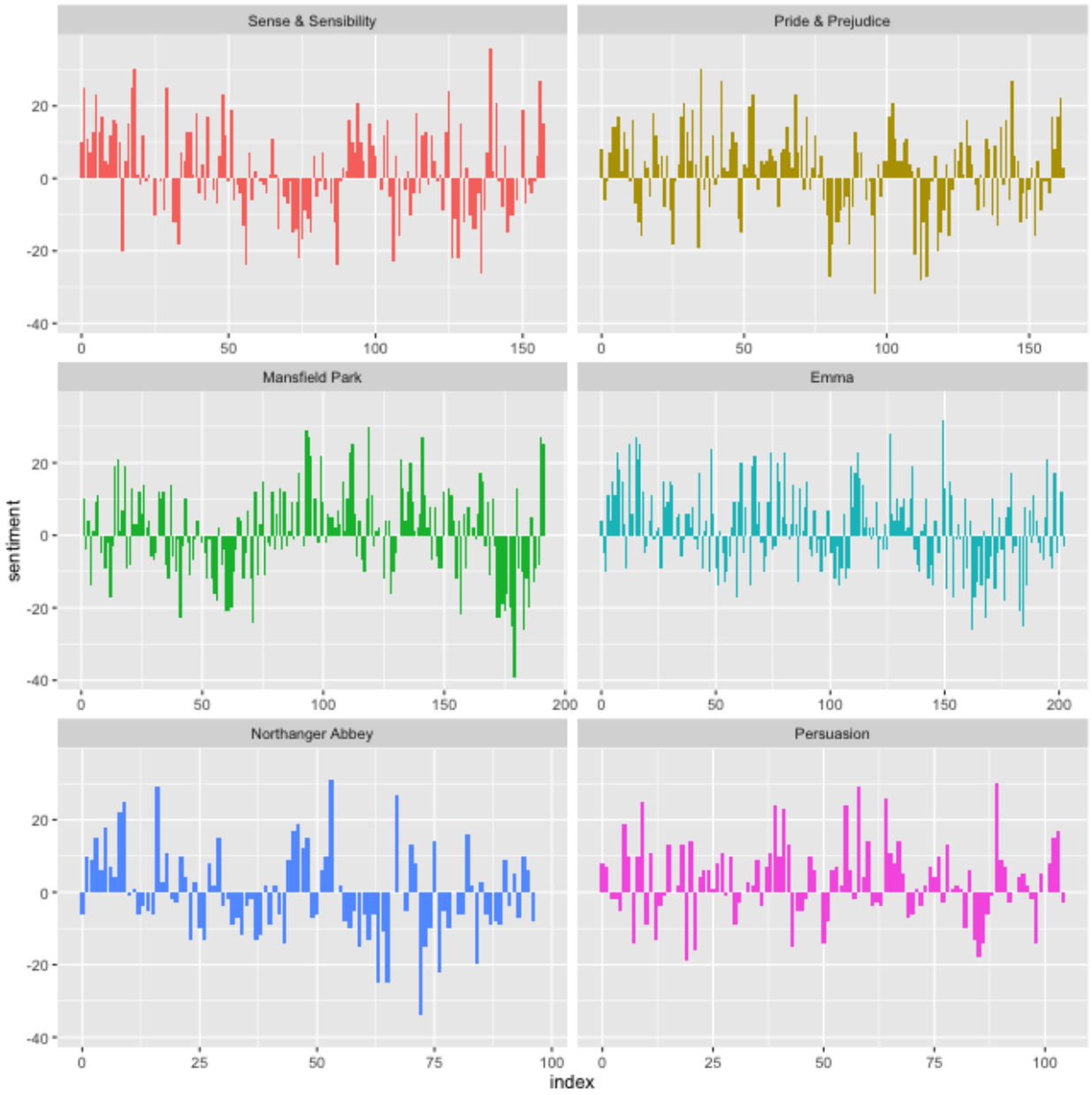

Figure 1: Sentiment in Jane Austen's Novels

References

Benoit, Kenneth, and Paul Nulty. 2016. *Quanteda: Quantitative Analysis of Textual Data*. https://CRAN.R-project.org/package=quanteda.

Ingo Feinerer, Kurt Hornik, and David Meyer. 2008. "Text Mining Infrastructure in R." *Journal of Statistical Software* 25 (5): 1–54. http://www.jstatsoft.org/v25/i05/.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Matthieu Gomez, Boris Demeshev, Dieter Menne, Benjamin Nutter, Luke Johnston, Ben Bolker, Francois Briatte, and Hadley Wickham. 2015. *Broom: Convert Statistical Analysis Objects into Tidy Data Frames*. https://CRAN.R-project.org/package=broom.

Silge, Julia, David Robinson, and Jim Hester. 2016. "Tidytext: Text Mining Using Dplyr, Ggplot2, and Other Tidy Tools." doi:10.5281/zenodo.56714.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (1): 1–23. doi:10.18637/jss.v059.i10.

Wickham, Hadley, Winston Chang, and RStudio. 2016. *Ggplot2: An Implementation of the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain Francois, and RStudio. 2015. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.