Welcome

Prof. Eric A. Suess

March 15, 2021

Welcome

Welcome to Stat 653 Statistical Natural Language Processing

Terms

- Dictionary, words, tokens
- Bag-of-Words
- Corpus, Documents
- Document Term Matrix
- Tidy Format, one-token-per-row
- TF-IDF
- Sentiment Analysis
- Topic Modeling

R packages Text Mining

- tm
- quanteda
- tidytext

R package for accessing Text data

- janeautenr
- gutenbergr
- harrypotter

R function for processing text data

- unnest_tokens()
- anti_join() remove stop words
- count()

Jane Austen

Read Chapter 1 of our book and compare Jane Austen to the Bronte sisters and H.G. Wells.

Harry Potter

Install the harrypotter R package from the author's github.

Take a look at the books, tidying the words from each book, counting the words and comparing the uses of words in each book the overall rates.

Sentiment Analysis

There are several sentiment lexicons.

- sentiments
- AFINN
- bing
- nrc
- try get_sentiments() for each