Ling 460
Making sense of Big Data:
Textual Analysis with R

Spring 2022

M, W, F 2:30pm-3:20pm
Dey Hall, 304

Satisfies “Data & Statistical Thinking”
core requirement for Data Science Minor; elective credit for Ling major

Instructor: K. Pertsova
pertsova@unc.edu

There are no prerequisites for this class. It is
targeted towards students with little or no prior
programming experience. Most of the datasets
we will work on will be of linguistic interest, but for
your final project you’re welcome to work on any
problem involving textual data analysis even if it
is not linguistic in nature.

Linguists and other social scientists rely
heavily on textual data, such as large collections
of texts (corpora), databases, crowdsourced
surveys, experiments etc. In this course you will
learn the basic principles and techniques in Data
Analysis using the software R. Learn how to:

- Organize, structure and clean your data
- Visualize your data using graphs: histograms, boxplots, scatter plots and more
- Perform basic statistic analysis of your data:
  measures of central tendencies and spread,
  testing differences between two samples,
  computing correlation between two variables
  and more.
- Find patterns and model them using:
  dictionary-based approaches, decision trees
  and Bayesian classifiers.