
Table of Contents

Preface.....	vii
1. The Tidy Text Format.....	1
Contrasting Tidy Text with Other Data Structures	2
The unnest_tokens Function	2
Tidying the Works of Jane Austen	4
The gutenbergr Package	7
Word Frequencies	8
Summary	12
2. Sentiment Analysis with Tidy Data.....	13
The sentiments Dataset	14
Sentiment Analysis with Inner Join	16
Comparing the Three Sentiment Dictionaries	19
Most Common Positive and Negative Words	22
Wordclouds	25
Looking at Units Beyond Just Words	27
Summary	29
3. Analyzing Word and Document Frequency: tf-idf.....	31
Term Frequency in Jane Austen's Novels	32
Zipf's Law	34
The bind_tf_idf Function	37
A Corpus of Physics Texts	40
Summary	44
4. Relationships Between Words: N-grams and Correlations.....	45
Tokenizing by N-gram	45

Counting and Filtering N-grams	46
Analyzing Bigrams	48
Using Bigrams to Provide Context in Sentiment Analysis	51
Visualizing a Network of Bigrams with ggraph	54
Visualizing Bigrams in Other Texts	59
Counting and Correlating Pairs of Words with the widyr Package	61
Counting and Correlating Among Sections	62
Examining Pairwise Correlation	63
Summary	67
5. Converting to and from Nontidy Formats.....	69
Tidying a Document-Term Matrix	70
Tidying DocumentTermMatrix Objects	71
Tidying dfm Objects	74
Casting Tidy Text Data into a Matrix	77
Tidying Corpus Objects with Metadata	79
Example: Mining Financial Articles	81
Summary	87
6. Topic Modeling.....	89
Latent Dirichlet Allocation	90
Word-Topic Probabilities	91
Document-Topic Probabilities	95
Example: The Great Library Heist	96
LDA on Chapters	97
Per-Document Classification	100
By-Word Assignments: augment	103
Alternative LDA Implementations	107
Summary	108
7. Case Study: Comparing Twitter Archives.....	109
Getting the Data and Distribution of Tweets	109
Word Frequencies	110
Comparing Word Usage	114
Changes in Word Use	116
Favorites and Retweets	120
Summary	124
8. Case Study: Mining NASA Metadata.....	125
How Data Is Organized at NASA	126
Wrangling and Tidying the Data	126
Some Initial Simple Exploration	129

Word Co-occurrences and Correlations	130
Networks of Description and Title Words	131
Networks of Keywords	134
Calculating tf-idf for the Description Fields	137
What Is tf-idf for the Description Field Words?	137
Connecting Description Fields to Keywords	138
Topic Modeling	140
Casting to a Document-Term Matrix	140
Ready for Topic Modeling	141
Interpreting the Topic Model	142
Connecting Topic Modeling with Keywords	149
Summary	152
9. Case Study: Analyzing Usenet Text.....	153
Preprocessing	153
Preprocessing Text	155
Words in Newsgroups	156
Finding tf-idf Within Newsgroups	157
Topic Modeling	160
Sentiment Analysis	163
Sentiment Analysis by Word	164
Sentiment Analysis by Message	167
N-gram Analysis	169
Summary	171
Bibliography.....	173
Index.....	175