

Table of Contents

Preface	1
Chapter 1: What is Text Analysis?	9
What is text analysis?	9
Where's the data at?	14
Garbage in, garbage out	18
Why should you do text analysis?	20
Summary	22
References	22
Chapter 2: Python Tips for Text Analysis	25
Why Python?	25
Text manipulation in Python	28
Summary	32
References	33
Chapter 3: spaCy's Language Models	35
spaCy	35
Installation	38
Troubleshooting	38
Language models	39
Installing language models	40
Installation – how and why?	42
Basic preprocessing with language models	42
Tokenizing text	43
Part-of-speech (POS) – tagging	45
Named entity recognition	46
Rule-based matching	48
Preprocessing	48
Summary	50
References	50
Chapter 4: Gensim – Vectorizing Text and Transformations and n-grams	53
Introducing Gensim	53
Vectors and why we need them	55
Bag-of-words	55
TF-IDF	57
Other representations	58
Vector transformations in Gensim	58

n-grams and some more preprocessing	62
Summary	64
References	65
Chapter 5: POS-Tagging and Its Applications	67
What is POS-tagging?	67
POS-tagging in Python	73
POS-tagging with spaCy	74
Training our own POS-taggers	76
POS-tagging code examples	81
Summary	83
References	83
Chapter 6: NER-Tagging and Its Applications	85
What is NER-tagging?	85
NER-tagging in Python	90
NER-tagging with spaCy	93
Training our own NER-taggers	98
NER-tagging examples and visualization	104
Summary	106
References	106
Chapter 7: Dependency Parsing	109
Dependency parsing	109
Dependency parsing in Python	115
Dependency parsing with spaCy	117
Training our dependency parsers	122
Summary	129
References	129
Chapter 8: Topic Models	131
What are topic models?	131
Topic models in Gensim	133
Latent Dirichlet allocation	135
Latent semantic indexing	137
Hierarchical Dirichlet process	138
Dynamic topic models	141
Topic models in scikit-learn	141
Summary	145
References	145
Chapter 9: Advanced Topic Modeling	147
Advanced training tips	147
Exploring documents	151
Topic coherence and evaluating topic models	157

Visualizing topic models	160
Summary	165
References	166
Chapter 10: Clustering and Classifying Text	169
Clustering text	169
Starting clustering	171
K-means	174
Hierarchical clustering	176
Classifying text	178
Summary	182
References	182
Chapter 11: Similarity Queries and Summarization	185
Similarity metrics	185
Similarity queries	192
Summarizing text	194
Summary	201
References	201
Chapter 12: Word2Vec, Doc2Vec, and Gensim	203
Word2Vec	203
Using Word2Vec with Gensim	205
Doc2Vec	211
Other word embeddings	217
GloVe	218
FastText	219
WordRank	221
Varembd	222
Poincare	223
Summary	224
References	224
Chapter 13: Deep Learning for Text	229
Deep learning	229
Deep learning for text (and more)	231
Generating text	234
Summary	240
References	241
Chapter 14: Keras and spaCy for Deep Learning	243
Keras and spaCy	243
Classification with Keras	246
Classification with spaCy	254
Summary	264

References	264
Chapter 15: Sentiment Analysis and ChatBots	267
Sentiment analysis	267
Reddit for mining data	271
Twitter for mining data	273
ChatBots	275
Summary	285
References	285
Other Books You May Enjoy	289
Index	293
