

Humanities Data Analysis



Case Studies with Python

Folgert Karsdorp, Mike Kestemont & Allen Riddell

PRINCETON UNIVERSITY PRESS

PRINCETON AND OXFORD

Contents

Preface	ix
I Data Analysis Essentials	1
Chapter 1 Introduction	3
1.1 Quantitative Data Analysis and the Humanities	3
1.2 Overview of the Book	5
1.3 Related Books	6
1.4 How to Use This Book	7
1.4.1 What you should know	8
1.4.2 Packages and data	12
1.4.3 Exercises	13
1.5 An Exploratory Data Analysis of the United States’ Culinary History	13
1.6 Cooking with Tabular Data	14
1.7 Taste Trends in Culinary US History	18
1.8 America’s Culinary Melting Pot	26
1.9 Further Reading	30
Chapter 2 Parsing and Manipulating Structured Data	32
2.1 Introduction	32
2.2 Plain Text	33
2.3 CSV	36
2.4 PDF	40
2.5 JSON	43
2.6 XML	46
2.6.1 Parsing XML	48
2.6.2 Creating XML	51
2.6.3 TEI	56
2.7 HTML	57
2.7.1 Retrieving HTML from the web	64

2.8	Extracting Character Interaction Networks	65
2.9	Conclusion and Further Reading	74
Chapter 3 Exploring Texts Using the Vector Space Model		78
3.1	Introduction	78
3.2	From Texts to Vectors	79
3.2.1	Text preprocessing	81
3.3	Mapping Genres	90
3.3.1	Computing distances between documents	97
3.3.2	Nearest neighbors	107
3.4	Further Reading	111
3.5	Appendix: Vectorizing Texts with NumPy	113
3.5.1	Constructing arrays	113
3.5.2	Indexing and slicing arrays	117
3.5.3	Aggregating functions	120
3.5.4	Array broadcasting	122
Chapter 4 Processing Tabular Data		126
4.1	Loading, Inspecting, and Summarizing Tabular Data	127
4.1.1	Reading tabular data with Pandas	130
4.2	Mapping Cultural Change	136
4.2.1	Turnover in naming practices	136
4.2.2	Visualizing turnovers	146
4.3	Changing Naming Practices	149
4.3.1	Increasing name diversity	150
4.3.2	A bias for names ending in <i>n</i> ?	153
4.3.3	Unisex names in the United States	158
4.4	Conclusions and Further Reading	162
II Advanced Data Analysis		165
Chapter 5 Statistics Essentials: Who Reads Novels?		169
5.1	Introduction	169
5.2	Statistics	170
5.3	Summarizing Location and Dispersion	171
5.3.1	Data: Novel reading in the United States	171
5.4	Location	175
5.5	Dispersion	179
5.5.1	Variation in categorical values	184
5.6	Measuring Association	188
5.6.1	Measuring association between numbers	188
5.6.2	Measuring association between categories	192
5.6.3	Mutual information	195
5.7	Conclusion	197
5.8	Further Reading	198

Chapter 6 Introduction to Probability	201
6.1 Uncertainty and Thomas Pynchon	202
6.2 Probability	203
6.2.1 Probability and degree of belief	205
6.3 Example: Bayes's Rule and Authorship Attribution	208
6.3.1 Random variables and probability distributions	213
6.4 Further Reading	225
6.5 Appendix	227
6.5.1 Bayes's rule	227
6.5.2 Fitting a negative binomial distribution	228
Chapter 7 Narrating with Maps	229
7.1 Introduction	229
7.2 Data Preparations	230
7.3 Projections and Basemaps	233
7.4 Plotting Battles	236
7.5 Mapping the Development of the War	238
7.6 Further Reading	244
Chapter 8 Stylometry and the Voice of Hildegard	248
8.1 Introduction	248
8.2 Authorship Attribution	250
8.2.1 Burrows's Delta	252
8.2.2 Function words	254
8.2.3 Computing document distances with Delta	257
8.2.4 Authorship attribution evaluation	260
8.3 Hierarchical Agglomerative Clustering	262
8.4 Principal Component Analysis	266
8.4.1 Applying PCA	268
8.4.2 The intuition behind PCA	271
8.4.3 Loadings	274
8.5 Conclusions	280
8.6 Further Reading	280
Chapter 9 A Topic Model of United States Supreme Court Opinions, 1900–2000	285
9.1 Introduction	285
9.2 Mixture Models: Artwork Dimensions in the Tate Galleries	287
9.3 Mixed-Membership Model of Texts	294
9.3.1 Parameter estimation	300
9.3.2 Checking an unsupervised model	304
9.3.3 Modeling different word senses	309
9.3.4 Exploring trends over time in the Supreme Court	313
9.4 Conclusion	317

9.5 Further Reading	318
9.6 Appendix: Mapping Between Our Topic Model and Lauderdale and Clark (2014)	320
Epilogue: Good Enough Practices	323
Bibliography	325
Index	333