

KNIHOVNA UP Olomouc

Značka:

Contents

Acknowledgments	ix
Preface	xi
I Orientation	1
1 Text analysis	5
1.1 Enter science	5
1.2 Data analysis	7
1.2.1 Emergence of data science	7
1.2.2 Data science toolbelt	7
1.2.3 Quant everywhere	8
1.3 Language analysis	9
1.3.1 Qualities and quantities	9
1.3.2 The nature of data	10
1.4 Text analysis	12
1.4.1 Aims	12
1.4.2 Implementation	13
1.4.3 Applications	14
Activities	16
Summary	16
II Foundations	17
2 Data	21
2.1 Data	22
2.1.1 Populations and samples	22
2.1.2 Corpora	23
2.1.3 Other considerations	25
2.1.4 Access	25
2.2 Information	30
2.2.1 Organization	30
2.2.2 Transformation	34
2.3 Documentation	40

2.3.1	Data origin	41
2.3.2	Data dictionaries	41
Activities		42
Summary		43
3	Analysis	45
3.1	Describe	45
3.1.1	Central tendency	47
3.1.2	Dispersion	48
3.1.3	Distributions	50
3.1.4	Association	56
3.2	Analyze	60
3.2.1	Explore	60
3.2.2	Predict	62
3.2.3	Infer	64
3.3	Communicate	66
3.3.1	Report	66
3.3.2	Document	66
Activities		67
Summary		68
4	Research	69
4.1	Frame	70
4.2	Connect	70
4.2.1	Research area	70
4.2.2	Research problem	71
4.3	Define	73
4.3.1	Research aim	73
4.3.2	Research question	74
4.4	Blueprint	75
4.4.1	Plan	76
4.4.2	Scaffold	79
Activities		81
Summary		81
III	Preparation	83
5	Acquire	87
5.1	Downloads	88
5.1.1	Manual	88
5.1.2	Programmatic	92
5.2	APIs	99
Activities		109
Summary		109

6	Curate	111
6.1	Unstructured	112
6.1.1	Reading data	112
6.1.2	Orientation	112
6.1.3	Tidy the data	115
6.2	Structured	118
6.2.1	Reading datasets	118
6.2.2	Orientation	119
6.2.3	Tidy the dataset	121
6.3	Semi-structured	126
6.3.1	Reading data	126
6.3.2	Orientation	127
6.3.3	Tidy the data	129
6.4	Documentation	134
	Activities	135
	Summary	136
7	Transform	137
7.1	Preparation	137
7.1.1	Normalization	139
7.1.2	Tokenization	142
7.2	Enrichment	147
7.2.1	Generation	148
7.2.2	Recoding	152
7.2.3	Integration	154
	Activities	163
	Summary	164
IV	Analysis	165
8	Explore	169
8.1	Orientation	170
8.2	Analysis	171
8.2.1	Descriptive analysis	172
8.2.2	Unsupervised learning	192
	Activities	211
	Summary	212
9	Predict	213
9.1	Orientation	213
9.2	Analysis	216
9.2.1	Text classification	218
9.2.2	Text regression	246
	Activities	257
	Summary	258

10 Infer	259
10.1 Orientation	259
10.2 Analysis	263
10.2.1 Categorical	264
10.2.2 Numeric	285
Activities	303
Summary	304
 V Communication	 305
11 Contribute	309
11.1 Public-facing	309
11.1.1 Structure	310
11.1.2 Purpose	311
11.1.3 Strategies	312
11.2 Peer-facing	314
11.2.1 Structure	314
11.2.2 Purpose	316
11.2.3 Strategies	316
Activities	324
Summary	324
 References	 325
 Index	 335