

Obsah

Předmluva	13
O autorovi	15
Poděkování	16
O odborných korektorech	17
Úvod	19
Co kniha popisuje	19
Co budete potřebovat	20
Komu je kniha určena	20
Styly	21
Zpětná vazba od čtenářů	22
Errata	22
KAPITOLA 1	
Začínáme	
Informatika	23
Umělá inteligence	24
Strojové učení	24
Statistika	24
Matematika	25
Znalostní domény	25
Data, informace a znalosti	25
Povaha dat	26

Proces datové analýzy	27
Problém	27
Příprava dat	28
Průzkum dat	28
Prediktivní modelování	28
Vizualizace výsledků	29
Kvantitativní a kvalitativní datová analýza	29
Význam vizualizace dat	30
A co big data?	31
Senzory a fotoaparáty	32
Analýza sociálních sítí	33
Nástroje a hračky v této knize	34
Proč Python?	34
Proč mlpy?	34
Proč D3.js?	35
Proč MongoDB?	35
Shrnutí	36

KAPITOLA 2

Práce s daty	37
Datové zdroje	38
Otevřená data	39
Textové soubory	39
Soubory aplikace Excel	39
SQL databáze	40
NoSQL databáze	41
Multimédia	42
Data z webu	42
Čištění dat	45
Statistické metody	45
Rozložení textu	46
Převádění dat	47
Datové formáty	47
CSV	48
JSON	49
XML	50
YAML	51

Začínáme s OpenRefine	52
Textová fazeta	52
Clustering	53
Textové filtry	54
Číselné fazety	54
Převod dat	55
Export dat	56
Historie akcí	57

Shrnutí	58
----------------	-----------

KAPITOLA 3

Vizualizace dat	59
------------------------	-----------

Dokumenty založené na datech (D3)	60
--	-----------

HMTL	60
DOM	61
CSS	61
JavaScript	61
SVG	61

Začínáme s D3.js	62
-------------------------	-----------

Sloupcový graf	62
Koláčový graf	67
Bodové vykreslení	70
Spojnicový graf	72
Spojnicový graf s více křivkami	76

Interakce a animace	79
----------------------------	-----------

Shrnutí	81
----------------	-----------

KAPITOLA 4

Třídění textu	83
----------------------	-----------

Učení a třídění	83
------------------------	-----------

Bayesiánské třídění	84
----------------------------	-----------

Naivní Bayesův algoritmus	85
---------------------------	----

Ověřování předmětu e-mailu	85
-----------------------------------	-----------

Data	86
------	----

Algoritmus	88
Přesnost třídění	91
Shrnutí	94
KAPITOLA 5	
Rozpoznávání podobných obrázků	95
Vyhledávání obrázků na základě podobnosti	95
Dynamické borcení času (DTW)	96
Zpracování obrázků v datové sadě	98
Implementace dynamického borcení času	99
Analýza výsledků	101
Shrnutí	103
KAPITOLA 6	
Simulace cen akcií	105
Finanční časové řady	105
Simulace nahodilé chůze	106
Postupy Monte Carlo	107
Generování nahodilých čísel	108
Implementace v D3.js	109
Shrnutí	116
KAPITOLA 7	
Předvídání cen zlata	117
Práce s daty v časových řadách	117
Komponenty časových řad	119
Čištění časových řad	120
Data – historické ceny zlata	123
Nelineární regrese	123
Hřebenová regrese jádra	123
Čištění časových řad cen zlata	125

Předvídání z vyčištěných časových řad	126
Srovnávání předpovězených hodnot	128

Shrnutí	129
----------------	------------

KAPITOLA 8

Práce s podpůrnými vektorovými stroji	131
--	------------

Úvod do datových sad s více proměnnými	132
---	------------

Redukce dimenzionality	135
-------------------------------	------------

Lineární diskriminantní analýza	136
---------------------------------	-----

Analýza hlavních komponent	136
----------------------------	-----

Začínáme pracovat s podpůrnými vektorovými stroji	139
--	------------

Funkce jádra	140
--------------	-----

Problém dvojité spirály	140
-------------------------	-----

Podpůrné vektorové stroje a knihovna mlpy	141
---	-----

Shrnutí	144
----------------	------------

KAPITOLA 9

Modelování infekčních nemocí prostřednictvím buněčných automatů	147
--	------------

Úvod do epidemiologie	148
------------------------------	------------

Epidemiologický trojúhelník	149
-----------------------------	-----

Epidemiologické modely	150
-------------------------------	------------

Model SIR	150
-----------	-----

Řešení běžných diferenciálních rovnic modlu SIR pomocí SciPy	151
--	-----

Model SIRS	152
------------	-----

Modelování pomocí buněčných automatů	153
---	------------

Buňka, stav, mřížka a sousedství	153
----------------------------------	-----

Globální stochastický kontaktní model	154
---------------------------------------	-----

Simulace modelu SIRS v CA prostřednictvím rozhraní D3.js	155
---	------------

Shrnutí	163
----------------	------------

KAPITOLA 10

Práce se sociálními grafy	165
----------------------------------	------------

Struktura grafu	165
------------------------	------------

Neorientovaný graf	166
Orientovaný graf	166
Analýza sociálních sítí	167
Tvorba grafu z Facebooku	167
Netvizz	167
Reprezentace grafů v Gephi	170
Statistická analýza	172
Poměr mužů a žen	172
Rozložení směrů	174
Histogram grafu	175
Středovost	176
Převod GDF na JSON	177
Vizualizace grafu v D3.js	179
Shrnutí	183

KAPITOLA 11

Analýza nálad na Twitteru	185
Anatomie dat z Twitteru	186
Tweet	186
Sledující	186
Témata trendů	187
Přístup k API Twitteru pomocí OAuth	187
Úvod do modulu Twython	189
Jednoduché vyhledávání	190
Práce s časovými osami	193
Práce se sledujícími	195
Práce s místy a trendy	197
Třídění nálad	198
Pravidla pro anglická slova	199
Textový korpus	199
Úvod do NLTK	200
Dávka slov	201
Naivní Bayesův klasifikátor	201
Analýza nálad tweetů	203
Shrnutí	204

Zpracování a agregace dat v MongoDB 205**Úvod do databáze MongoDB 206**

Databáze 206

Kolekce 207

Dokument 208

Příkazový řádek Mongo 208

Vkládání, aktualizace a odstraňování 209

Dotazy 209

Příprava dat 211

Převod dat pomocí OpenRefine 211

Vkládání dokumentů pomocí PyMongo 214

Skupina 217**Agregační rozhraní 219**

Kanály 220

Výrazy 221

Shrnutí 222**Práce s modelem MapReduce 223****Úvod do modelu MapReduce 224****Programovací model 225****Model MapReduce a databáze MongoDB 226**

Funkce map 226

Funkce reduce 227

Příkazový řádek Mongo 227

Práce s nástrojem UMongo 229

Práce s nástrojem PyMongo 232

Filtrování vstupní kolekce 233**Seskupování a agregace 234****Vizualizace nejběžnějších slov ve tweetech v mraku slov 236****Shrnutí 241**

Online analýza dat s nástroji Ipython a Wakari 243**Úvod do nástroje Wakari 244**

Založení účtu ve Wakari 244

Úvod do nástroje IPython Notebook 246

Vizualizace dat 248

Úvod do zpracování obrázků knihovnou PIL 250

Otevření obrázku 250

Histogram obrázku 251

Filtrování 252

Operace 254

Převody 256

Úvod do knihovny Pandas 257

Práce s časovými řadami 257

Práce s datovými sadami s více proměnnými a s objektem DataFrame 261

Seskupování, agregace a korelace 264

Multiprocessing s nástrojem IPython 266

Pool 267

Sdílení poznámkového bloku 267

Data 267

Shrnutí 270

PŘÍLOHA

Zavádění infrastruktury 271**Instalace a spuštění prostředí Python 3 272**

Instalace a spuštění prostředí Python v Ubuntu 272

Instalace a spuštění prostředí IDLE v Ubuntu 272

Instalace a spuštění prostředí Python 3.2 ve Windows 273

Instalace a spuštění prostředí IDLE ve Windows 274

Instalace a spuštění knihovny NumPy 275

Instalace a spuštění knihovny NumPy v Ubuntu 275

Instalace a spuštění knihovny NumPy ve Windows 276

Instalace a spuštění knihovny SciPy 277

Instalace a spuštění knihovny SciPy v Ubuntu 277

Instalace a spuštění knihovny SciPy ve Windows 278

Instalace a spuštění knihovny mlpy	279
Instalace a spuštění knihovny mlpy v Ubuntu	279
Instalace a spuštění knihovny mlpy ve Windows	279
Instalace a spuštění nástroje OpenRefine	280
Instalace a spuštění nástroje OpenRefine v Linuxu	280
Instalace a spuštění nástroje OpenRefine ve Windows	281
Instalace a spuštění databáze MongoDB	281
Instalace a spuštění databáze MongoDB v Ubuntu	282
Instalace a spuštění databáze MongoDB ve Windows	283
Připojení Pythonu k databázi MongoDB	285
Instalace a spuštění nástroje UMongo	286
Instalace a spuštění nástroje Umongo v Ubuntu	287
Instalace a spuštění Umongo ve Windows	288
Instalace a spuštění Gephi	289
Instalace a spuštění Gephi v Linuxu	290
Instalace a spuštění Gephi ve Windows	290
Rejstřík	291