Sommario

Prefazione ..... **IX**

Capitolo 1 - Lo stack Python per la scienza dei dati **1**

Introduzione 1
Pacchetti e librerie di Python 2
IPython: una potente shell interattiva 2
Esercizio 1: Interagire con la shell di Python usando i comandi di IPython 3
Jupyter notebook 4
Esercizio 2: Primi passi con Jupyter Notebook 5
IPython o Jupyter? 7
Attività 1: IPython e Jupyter 8
NumPy 8
SciPy 9
Matplotlib 9
Pandas 9
Usare Pandas 10
Lettura dei dati 10
Esercizio 3: Leggere i dati con Pandas 11
Manipolazione dei dati 12
Esercizio 4: Selezionare i dati e usare il metodo .loc 14
Attività 2: Problemi con i dati 18
Conversione del tipo di dati 18
Esercizio 5: Esplorare i tipi di dati 19
Aggregazione e raggruppamento 21
Esercizio 6: Aggregare e raggruppare i dati 22
NumPy su Pandas 23

Esportare i dati da Pandas 23
Esercizio 7: Esportare i dati in diversi formati 24
Visualizzare i dati con Pandas 26
Attività 3: Tracciare dati con Pandas 26
Riepilogo 27

Capitolo 2 - Grafici statistici....29

Introduzione 29
Tipi di grafici e quando usarli 30
Esercizio 8: Tracciare il grafico di una funzione analitica 31
I componenti del grafico 32
Esercizio 9: Creare un grafico 33
Esercizio 10: Creare un grafico per una funzione matematica 35
Seaborn 36
Quale strumento utilizzare? 36
Tipi di grafici 36
Grafici a linee 37
Grafici di serie temporali 37
Esercizio 11: Creare grafici a linee utilizzando librerie diverse 38
DataFrame di Pandas e dati raggruppati 40
Attività 4: Grafici a linee con l'API orientata agli oggetti e i DataFrame di Pandas 41
Grafici a dispersione 42
Attività 5: Comprendere le relazioni delle variabili utilizzando i grafici a dispersione 44
Istogrammi 45
Esercizio 12: Creare un istogramma per la distribuzione della potenza 45
Grafici a scatola 47
Esercizio 13: Analizzare il comportamento di numero cilindri e potenza utilizzando un grafico a scatola 48
Cambiare la struttura del grafico: modificare i componenti 51
Titolo e configurazione delle etichette degli oggetti assi 51
Esercizio 14: Configurare il titolo e le etichette degli assi 51
Stili e colori delle linee 53
Dimensione della figura 54
Esercizio 15: Utilizzare i fogli di stile di Matplotlib 55
Esportare i grafici 56
Attività 6: Esportare un grafico in un file 59
Attività 7: Progettare la struttura completa di un grafico 59
Riepilogo 60

Capitolo 3 - Lavorare con i framework per i big data.....63
Introduzione 63
Hadoop 64
Manipolazione dei dati con HDFS 66
Esercizio 16: Manipolare i file in HDFS 66
Spark 67
SQL di Spark e DataFrame di Pandas 69
Esercizio 17: Eseguire operazioni DataFrame in Spark 69
Esercizio 18: Accedere ai dati con Spark 70
Esercizio 19: Leggere i dati dal file system locale e da HDFS 71
Esercizio 20: Scrivere dati in HDFS e PostgreSQL 72
Scrittura in file Parquet 73
Esercizio 21: Scrivere file Parquet 74
Aumentare le prestazioni delle analisi con Parquet e le partizioni 74
Esercizio 22: Creare un DataSet partizionato 75
Gestione dei dati non strutturati 76
Esercizio 23: Analizzare e ripulire il testo 76
Attività 8: Rimuovere le parole di stop dal testo 78
Riepilogo 79

Capitolo 4 - Un esame più approfondito di Spark.....81
Introduzione 81
Primi passi con i DataFrame di Spark 82
Esercizio 24: Specificare lo schema di un DataFrame 83
Esercizio 25: Creare un DataFrame partendo da un RDD esistente 84
Esercizio 26: Creare un DataFrame utilizzando un file CSV 85
Scrivere output dai DataFrame di Spark 86
Esercizio 27: Convertire un DataFrame di Spark in un DataFrame di Pandas 86
Esplorare i DataFrame di Spark 86
Esercizio 28: Visualizzare le statistiche di base del DataFrame 87
Attività 9: Primi passi con i DataFrame di Spark 89
Manipolazione dei dati con i DataFrame di Spark 89
Esercizio 29: Selezionare e modificare i nomi delle colonne del DataFrame 90
Esercizio 30: Aggiungere e rimuovere una colonna del DataFrame 90
Esercizio 31: Visualizzare e contare i valori distinti in un DataFrame 91
Esercizio 32: Rimuovere le righe duplicate e filtrare le righe di un DataFrame 92
Esercizio 33: Ordinare le righe di un DataFrame 93
Esercizio 34: Aggregare i valori di un DataFrame 94
Attività 10: Manipolare i dati con i DataFrame di Spark 95
Grafici in Spark 95
Esercizio 35: Creare un grafico a barre 96

Esercizio 36: Creare un grafico a modello lineare 97
Esercizio 37: Creare un grafico KDE e un grafico a scatola 98
Attività 11: Grafici in Spark 99
Riepilogo 102

Capitolo 5 - Gestire i valori mancanti e analizzare le correlazioni 103

Introduzione 103
Preparare Jupiter notebook 104
Valori mancanti 105
Esercizio 38: Contare i valori mancanti di un DataFrame 105
Esercizio 39: Contare i valori mancanti in tutte le colonne del DataFrame 106
Recuperare i record di valori mancanti dal DataFrame 107
Gestire i valori mancanti nei DataFrame di Spark 107
Esercizio 40: Rimuovere da un DataFrame i record che hanno valori mancanti 108
Esercizio 41: Rimpiazzare i valori mancanti con una costante in una colonna del DataFrame 108
Correlazione 109
Esercizio 42: Calcolare la correlazione 110
Attività 12: Gestire i valori mancanti e analizzare la correlazione con i DataFrame di PySpark 111
Riepilogo 114

Capitolo 6 - L'analisi preliminare 115

Introduzione 115
Definire un problema 116
Identificazione del problema 117
Raccolta dei requisiti 117
Elaborazione dei dati e flusso di lavoro 118
Identificare le metriche misurabili 118
Documentazione e presentazione 119
Traduzione del problema in metriche misurabili e analisi preliminare 119
Raccolta dei dati 119
Analisi della produzione di dati 120
Visualizzazione KPI 121
Rilevanza della caratteristica 121
Esercizio 43: Identificare la variabile target e i relativi KPI dai dati forniti per il problema 121
Esercizio 44: Generare la rilevanza della caratteristica per la variabile target ed effettuare l'analisi preliminare 128
Approccio strutturato al ciclo di vita del progetto di scienza dei dati 132
Le fasi del ciclo di vita di un progetto di scienza dei dati 133

Fase 1: Comprensione e definizione del problema 133
Fase 2: Accesso ai dati e osservazione 133
Fase 3: Struttura dei dati e pre-elaborazione 134
Attività 13: Eseguire la mappatura della distribuzione gaussiana delle caratteristiche numeriche a partire dai dati forniti 134
Fase 4: Sviluppo del modello 135
Riepilogo 136

Capitolo 7 - Riproducibilità nell'analisi dei big data 137

Introduzione 137
Riproducibilità con Jupyter notebook 138
Introduzione al problema 139
Documentare l'approccio e il flusso di lavoro 139
Spiegare il canale dati 140
Spiegare le dipendenze 140
Uso del controllo della versione del codice sorgente 140
Modularizzazione del processo 141
Raccogliere i dati in modo riproducibile 141
Funzionalità nelle celle markdown e nelle celle di codice 141
Descrivere il problema in markdown 143
Fornire un'introduzione dettagliata alla fonte dati 144
Descrivere gli attributi dei dati in markdown 144
Esercizio 45: Esecuzione della riproducibilità dei dati 146
Prassi e standard per il codice 149
Documentare l'ambiente 149
Scrivere codice leggibile con commenti 149
Effettiva segmentazione dei flussi di lavoro 150
Documentare il flusso di lavoro 151
Esercizio 46: Pre-elaborare i valori mancanti con elevata riproducibilità 151
Evitare le ripetizioni 154
Utilizzare funzioni e cicli per ottimizzare il codice 154
Sviluppare librerie/pacchetti per il riutilizzo di codici/algoritmi 155
Attività 14: Normalizzare i dati 155
Riepilogo 156

Capitolo 8 - Creazione di un rapporto d'analisi completo 157

Introduzione 157
Caricare in Spark i dati letti da diverse fonti 158
Esercizio 47: Leggere i dati da un file CSV utilizzando l'oggetto PySpark 158
Leggere i dati JSON usando l'oggetto PySpark 160
Operazioni SQL sui DataFrame di Spark 160
Esercizio 48: Leggere i dati in PySpark ed eseguire operazioni SQL 160

Esercizio 49: Creare e unire due dataFrame 164
Esercizio 50: Creare un sottoinsieme del DataFrame 166
Generare misure statistiche 167
Attività 15: Generare una visualizzazione mediante Plotly 169
Riepilogo 171

Appendice 173

Capitolo 1: Lo stack Python per la scienza dei dati 173
Attività 1: IPython e Jupyter 173
Attività 2: Problemi con i dati 174
Attività 3: Tracciare dati con Pandas 177
Capitolo 2: Grafici statistici usando Matplotlib e Seaborn 178
Attività 4: Grafici a linee con l'API orientata agli oggetti e i DataFrame di Pandas 178
Attività 5: Comprendere le relazioni delle variabili utilizzando i grafici a dispersione 180
Attività 6: Esportare un grafico in un file 181
Attività 7: Progettare la struttura completa di un grafico 183
Capitolo 3: Lavorare con i framework per i big data 185
Attività 8: Analisi del testo 185
Capitolo 4: Un esame più approfondito di Spark 187
Attività 9: Primi passi con i DataFrame di Spark 187
Attività 10: Manipolare i dati con i DataFrame di Spark 191
Attività 11: Grafici in Spark 195
Capitolo 5: Gestire i valori mancanti e analizzare le correlazioni 199
Attività 12: Gestire i valori mancanti e analizzare la correlazione con i DataFrame di PySpark 199
Capitolo 6: L'analisi preliminare 204
Attività 13: Eseguire la mappatura della distribuzione gaussiana delle caratteristiche numeriche a partire dai dati forniti 204
Capitolo 7: Riproducibilità nell'analisi dei big data 208
Attività 14: Normalizzare i dati 208
Capitolo 8: Creazione di un rapporto d'analisi completo 214
Attività 15: Generare una visualizzazione mediante Plotly 214

Indice analitico 217