Introduzione

Devo parlarvi di scienza, fra poco, giusto per distrarvi dai vostri pensieri.
Marie Corelli, romanziera britannica

Questo libro è stato ispirato da un corso introduttivo alla scienza dei dati in Python che ho tenuto nell'estate del 2015 a un piccolo e selezionato gruppo di studenti della Suffolk University di Boston. Il corso era stato concepito per essere il primo di una sequenza di due corsi, con un'enfasi sull'ottenimento, la "ripulitura", l'organizzazione e la visualizzazione dei dati, con una spruzzata di elementi di statistica, machine learning e analisi dei dati di rete.

Ben presto mi sono reso conto che l'abbondanza di sistemi e moduli Python coinvolti da queste operazioni (database, framework per l'elaborazione del linguaggio naturale, parser JSON e HTML e strutture dati ad alte prestazioni, solo per citarne alcune) sarebbe stata davvero eccessiva non solo per uno studente universitario, ma anche per un professionista esperto. In realtà, devo confessare che, mentre mi occupavo dei miei progetti di ricerca nei campi della scienza dei dati e dell'analisi dei dati di rete, dedicavo molto tempo, fin troppo, a richiamare la funzione `help()` e a sfogliare le discussioni online su Python. Inoltre, devo ammettere di aver avuto anche alcuni momenti imbarazzanti durante la lezione, quando sembravo aver completamente dimenticato il nome di alcune funzioni o di alcuni parametri opzionali.

Nell'ambito del materiale del corso, ho anche compilato una serie di appunti e schemi sui vari argomenti, da usare come riferimento. Tali appunti e schemi si sono evoluti fino a diventare questo libro. Auspicabilmente, il fatto di avere questo libro sulla vostra scrivania vi aiutera a riflettere più sulla scienza dei dati e l'analisi dei dati che sui nomi di funzioni e parametri opzionali.

Informazioni su questo libro

Questo libro si occupa di acquisizione, pulizia, memorizzazione, ricerca, trasformazione, visualizzazione dei dati, con elementi di analisi dei dati avanzata (l'analisi dei dati di rete), statistica e machine learning. Non è un'introduzione alla scienza dei dati o una guida di riferimento generale alla scienza dei dati, anche se troverete una panoramica introduttiva nel Capitolo 1, *Che cosa si intende per scienza dei dati?*. Presuppongo quindi che abbiate

già appreso altrove i metodi tipici della scienza dei dati e della statistica. L'indice analitico che si trova alla fine del libro fa riferimento alle implementazioni Python dei concetti chiave, ma in molti casi scoprirete che tali concetti vi sono già familiari.
Nel Capitolo 2, *Elementi di base di Python per la scienza dei dati*, troverete un riepilogo delle strutture di dati di Python; funzioni per stringhe, file e il Web; espressioni regolari; manipolazione delle liste. Tale capitolo ha lo scopo di rinfrescarvi la memoria su questi argomenti, non di insegnarli. Sono disponibili molti eccellenti testi su Python e una buona conoscenza di tale linguaggio è assolutamente importante per diventare veri esperti di scienza dei dati.
La prima parte del libro mostra la manipolazione di vari tipi di dati testuali, inclusa l'elaborazione di testi strutturati e non strutturati, l'elaborazione di dati numerici con i moduli `NumPy` e `Pandas` e l'analisi dei dati di rete. Altri tre capitoli si occupano di vari aspetti dell'analisi: l'uso di database relazionali e non-relazionali, la visualizzazione dei dati e l'analisi predittiva.
Questo libro ha un suo filo logico ma è anche una guida di riferimento. A seconda delle vostre esigenze, potete leggerlo sequenzialmente o saltare direttamente all'indice, trovare la funzione o il concetto che vi interessa e ricercare le descrizioni e gli esempi pertinenti. Nel primo caso, se siete già esperti programmatori Python, potete tranquillamente saltare il Capitolo 2, *Elementi di base di Python per la scienza dei dati*. Se non pensate di dover interagire con dei database esterni (come MySQL), potete ignorare il Capitolo 4, *Utilizzare i database*. Infine, il Capitolo 9, *Probabilità e statistica*, presuppone che non abbiate alcuna idea della statistica. Se già conoscete l'argomento, saltatene pure le prime due Unità e partite dall'Unità 47, *Statistica in Python*.

A chi si rivolge questo libro

A questo punto, potreste chiedervi se non sia il caso di far diventare questo libro una presenza stabile nella vostra libreria.
Il libro è rivolto a studenti e laureati, ai docenti di scienza dei dati, ai professionisti alle prime armi nel campo della scienza dei dati (in particolare a coloro che stanno passando da R a Python) e agli sviluppatori alla ricerca di una guida di riferimento che li aiuti a ricordare tutte le funzioni e le opzioni di Python.
Se rientrate in questo breve ritratto, questo libro fa proprio per voi.

Questioni di software

Nonostante alcune controversie sulla transizione da Python 2.7 a Python 3.3 e alle versioni successive, sono decisamente convinto sostenitore della più recente versione di Python. La maggior parte del nuovo software per Python è stata sviluppata per la versione 3.3 e anche la maggior parte del software precedente è stata portata alla versione 3.3. Tenendo conto di questa tendenza, sarebbe poco saggio aggrapparsi a una versione obsoleta, indipendentemente dalla sua popolarità.
Tutti gli esempi in Python di questo libro possono funzionare con i moduli menzionati

nella Tabella I.1. Tutti questi moduli, con l'eccezione del modulo `community` che deve essere installato separatamente (`https://pypi.python.org/pypi/python-louvain/0.3`) e dello stesso interprete Python, sono inclusi nella distribuzione Anaconda, distribuita da Continuum Analytics e disponibile gratuitamente (`https://www.continuum.io`).

Tabella I.1 Componenti software usati nel libro.

Pacchetto	Versione usata	Pacchetto	Versione usata
`BeautifulSoup`	4.3.2	`community`	0.3
`json`	2.0.9	`html5lib`	0.999
`matplotlib`	1.4.3	`networkx`	1.10.0
`nltk`	3.1.0	`numpy`	1.10.1
`pandas`	0.17.0	`pymongo`	3.0.2
`pymysql`	0.6.2	`python`	3.4.3
`scikit-learn`	0.16.1	`scipy`	0.16.0

Se pensate di provare a usare (o di usare effettivamente, per lavoro) i database, dovrete anche scaricare e installare MySQL (`https://www.mysql.com`) e MongoDB (`https://www.mongodb.com`). Entrambi i database sono gratuiti e funzionano perfettamente in Linux, Mac OS e Windows.

Note sugli apici

Python consente di racchiudere le stringhe di caratteri in apici `'singoli'`, `"doppi"`, `'''tripli'''` e perfino `"""tre doppi apici"""` (le ultime due possibilità sono utili per le stringhe multiriga). Tuttavia, quando stampa le stringhe, impiega sempre la notazione a singolo apice, indipendentemente dal numero di apici usati nel programma.
Molti altri linguaggi (C, C++, Java) usano gli apici singoli e doppi in modo molto differente: singoli per i caratteri singoli, doppi per le stringhe di caratteri. Come tributo a questa differenziazione, anche in questo libro useremo i singoli apici per i caratteri e i doppi apici per le stringhe di caratteri.

Il forum del libro

Il forum della community di questo libro si trova sul sito dell'editore inglese The Pragmatic Programmers (`https://pragprog.com/book/dzpyds/data-science-essentials-in-python`). Qui potete porre domande, inviare commenti e segnalare errori.
Un'altra ottima risorsa per le domande e le risposte (ma non relative in modo specifico a questo libro) è il recente forum Data Science Stack Exchange (`https://datascience.stackexchange.com`).

Scarica il codice degli esempi

Il codice sorgente di molti esempi presentati nel testo è disponibile sul sito dell'editore originale inglese The Pragmatic Programmers. L'indirizzo per scaricare gli archivi ZIP o TGZ è: `https://pragprog.com/titles/dzpyds/source_code` (o per comodità `http://bit.ly/pp-code-dsp`).

Esercitazioni

Alla fine di ogni capitolo si trova una particolare Unità chiamata "Esercitazioni". Questa Unità presenta vari progetti che potete svolgere per vostro conto (o anche insieme ad altri) per rafforzare e "concretizzare" le vostre conoscenze sul materiale teorico proposto. I progetti contrassegnati con una stella (☆) sono i più semplici. Per risolverli vi basterà una buona conoscenza delle funzioni menzionate nei capitoli precedenti. Potete pensare di completare tali progetti in una trentina di minuti. Le soluzioni di tali esercitazioni si trovano nell'Appendice B, *Soluzioni dei progetti di tipo* (☆).

I progetti contrassegnati con due stelle (☆☆) sono più complessi. La loro risoluzione può portarvi via un'ora o più, a seconda delle vostre abilità e abitudini di programmazione. I progetti a due stelle richiedeono l'uso di strutture di dati intermedie e di algoritmi ben ragionati.

Infine, i progetti a tre stelle (☆☆☆) sono i più difficili. Alcuni progetti a tre stelle potrebbero perfino non prevedere una soluzione davvero "perfetta", pertanto non preoccupatevi se non riuscite a trovarla! Ma il fatto di lavorare su questi progetti, migliorerà certamente le vostre abilità di programmatori e vi renderà anche migliori esperti di scienza dei dati. I docenti possono anche considerare i progetti a tre stelle come suggerimenti per assegnazioni di attività pratiche ai loro studenti.

E ora… cominciamo!

Dmitry Zinoviev
`dzinoviev@gmail.com`
Agosto 2016