Marek Gągolewski
Maciej Bartoszuk
Anna Cena

Przetwarzanie i analiza danych w języku Python

PWN

Przetwarzanie i analiza danych w języku Python

Marek Gągolewski
Maciej Bartoszuk
Anna Cena

Przetwarzanie i analiza danych w języku Python

PWN

Projekt okładki **Hubert Zacharski**

Ilustracja na okładce **shutterstock/phipatbig**

Wydawca **Łukasz Łopuszański**

Redaktor prowadzący **Iwona Lewandowska**

Redaktor **Ewa Ławrynowicz**

Koordynator produkcji **Anna Bączkowska**

Skład i łamanie **FixPoint**

ISBN: 978-83-01-18940-2

Wydanie I
Warszawa 2016

Wydawnictwo Naukowe PWN SA
02-460 Warszawa, ul. Gottlieba Daimlera 2
tel. 22 69 54 321, faks 22 69 54 288
infolinia 801 33 33 88
e-mail: pwn@pwn.com.pl; reklama@pwn.pl
www.pwn.pl

Druk i oprawa: OSDW Azymut Sp. z o.o.

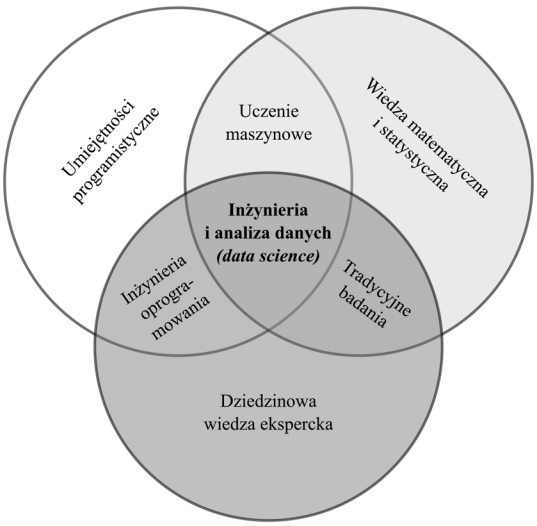
SPIS TREŚCI

II Przetwarzanie danych

PRZEDMOWA

Żyjemy w fascynującej dobie przeciążenia informacyjnego: dla korporacji, instytucji i zainteresowanych jednostek zarówno samo zdobycie różnorakich danych, jak i późniejsze ich przechowywanie nie stanowi już właściwie żadnego problemu – jeśli tylko dysponują odpowiednimi środkami. Wyzwaniem pozostaje jednak przetworzenie tego oceanu nic niemówiących ciągów bitów na użyteczną wiedzę, np. plan marketingowy firmy, zoptymalizowany pod różnymi względami sposób rekomendowania treści użytkownikom portalu internetowego, taktykę inwestowania na giełdzie, dalszy kierunek badań R&D. Z tego rodzaju wyzwaniami mierzy się m.in. względnie nowa, ale prężnie rozwijająca się dziedzina *inżynierii i analizy danych* (ang. *data science*), która wymaga od zajmującej się nią profesjonalistów nie tylko dziedzinowej wiedzy eksperckiej oraz szeroko pojętej kreatywności i ciekawości świata, ale także kompetencji matematycznych (w szczególności statystycznych) oraz umiejętności programistycznych; por. rys. 0.1.

Na rynku dostępnych jest wiele narzędzi do obliczeń analizodanowych, jednak w ostatnim czasie obserwujemy, że wielu specjalistów coraz chętniej sięga po otwarte



Rysunek 0.1. Trzy główne obszary kompetencji specjalistów inżynierii i analizy danych

i wolnodostępne rozwiązania. Aby oprogramowanie spełniało swoją funkcję w wyróżnionym przez nas obszarze zastosowań, musi zawierać pokaźny zestaw najbardziej pożytecznych, sprawdzonych, gotowych do użycia metod, służących m.in. do sporządzania wykresów, wyznaczania modeli regresji i klasyfikacji czy weryfikowania hipotez statystycznych. Co ważne, musi być ono oparte także na „pełnowymiarowym" *języku programowania*, tak by korzystający z niego nie czuł się w żaden sposób ograniczony – prawdziwe wyzwania *data science* bardzo rzadko wpisują się w proste szablony. Dzięki temu możliwe jest także automatyzowanie procesów przetwarzania danych i sprawianie, by przeprowadzane analizy były odtwarzalne.

O jednym z takich niezmiernie popularnych narzędzi – środowisku R (zob. np. [16]) – mówi się, że zostało stworzone przez statystyków dla statystyków. Mimo że R oferuje coraz lepszą łączność ze wszystkimi popularnymi systemami baz danych, wsparcie dla systemów *big data* (HDFS, Spark itd.), możliwość tworzenia rozszerzeń w językach Java i C++, bardzo rzadko jest on doceniany przez osoby niemające wykształcenia okołoanalizodanowego. W szczególności — z racji tego, że semantyka języka R została zainspirowana różnymi pochodnymi języka Lisp, który współcześnie nie jest już zbyt często używany — obserwujemy, że trudno jest do niego przekonać informatyków: ich opinia jest o tyle ważna, że np. implementowane przez specjalistów *data science* modele predykcyjne stają się potem częścią większych rozwiązań projektowych.

Środowisko Python jest z kolei oparte na języku o znacznie szerszych zastosowaniach. Stworzone przez programistów dla programistów pozwala na tworzenie wszelkiej maści projektów informatycznych – od szybkich prototypów rozwiązań (zgodnie z ideą *rapid development*) po duże, wielomodułowe aplikacje składające się z kilkuset klas. Jego wysoka przydatność jest powszechnie znana od wielu lat w takich obszarach, jak programowanie aplikacji sieciowych, internetowych i mobilnych, przetwarzanie obrazów i sygnałów audio, tworzenie gier komputerowych itp. Co ważne, od jakiegoś czasu w szerokiej gamie powiązanych ze sobą pakietów i bibliotek możemy odnaleźć wiele nowoczesnych i wydajnie zaimplementowanych algorytmów uczenia maszynowego (ang. *machine learning*), statystycznych systemów uczących się, wizualizacji danych itd. Wszystko to sprawia, że to właśnie Python staje coraz częściej narzędziem z wyboru dla specjalistów *data science* – ma ono bowiem ogromne możliwości.

Wśród dostępnych na polskim rynku wydawniczym pozycji poświęconych wprowadzeniu do programowania w „bazowym" języku Python możemy wymienić m.in. [1, 8, 18, 27, 34, 35, 46]. Zainteresowani szeroko pojętą analizą danych do tej pory zmuszeni byli już jednak odwoływać się do lepszych lub gorszych angielskojęzycznych tytułów [4, 9, 20, 26, 40, 41, 43, 48]. Niniejsza książka jest pierwszą polską pozycją poświęconą w całości przetwarzaniu i analizie danych z wykorzystaniem środowiska Python.

Cel książki i jej adresaci. Książka stawia sobie za cel przygotować Czytelnika do samodzielnego przeprowadzenia procesu analizy danych, od pobrania i załadowania zbioru danych, poprzez jego wstępne przetworzenie i wyczyszczenie, aż po samą analizę.

Czytelnik będzie potrafił dokonać wizualizacji danych oraz przedstawić wyniki analizy w formie raportów. Ponieważ wierzymy, że lepiej dać przysłowiową wędkę niż złowioną rybę, kładziemy nacisk na dokładne omówienie samego języka Python 3 i najważniejszych pakietów towarzyszących, tak by Czytelnik był przygotowany na twórcze mierzenie się z nowymi problemami oraz dalsze samodzielne zgłębianie literatury przedmiotu. Wiemy, że pewne rozwiązania, które stworzy Czytelnik, będą przeznaczone do wielokrotnego użytku i tym samym będą zasługiwać na wdrożenie w ramach większych projektów informatycznych. Z tego powodu omawiamy także zestaw dobrych praktyk z zakresu inżynierii oprogramowania.

Książka jest przeznaczona dla szerokiego grona odbiorców, m.in. (obecnych i przyszłych) analityków danych, *data scientists*, specjalistów *business intelligence*, naukowców, badaczy i programistów. Szczególnie polecamy ją:

— Osobom, które już mają wiedzę teoretyczną i umiejętności w zakresie wizualizacji danych, statystyki lub uczenia maszynowego, które wykorzystywały oprogramowanie typu R, SAS czy SPSS w analizie danych i które pragną wreszcie poznać, jak wykonać podobne czynności w środowisku Python.

— Programistom języka Python zainteresowanym inżynierią i analizą danych w całej swej okazałości. Proponowany kurs zapoznaje Czytelnika z najistotniejszymi technikami, dając przy tym inspirację i podstawy do dalszego – samodzielnego już – zgłębiania wiedzy teoretycznej i rozwoju umiejętności praktycznych. Pozwala też zrozumieć wyzwania stające przed analitykami danych i znaleźć z nimi „wspólny język".

— Osobom, które znają podstawy programowania w jakimś innym języku, a które chcą poznać język Python w kontekście, którego nie oferują inne podręczniki. Książka bowiem zawiera całościowy kurs języka – problemy analizy danych dla niektórych mogą być po prostu niewinnym pretekstem do wykonania całej serii przyjemnych ćwiczeń. Ciekawe zbiory danych prowokują do stawiania pytań i poszukiwania niebanalnych odpowiedzi, do których droga prowadzi przez programowanie. Po zapoznaniu się z przedstawionym tutaj materiałem Czytelnika nie zdziwi żadna zawiłość ani cecha języka, będzie on w stanie swobodnie posługiwać się dokumentacją różnych pakietów, a także samodzielnie utworzyć większy fragment oprogramowania.

Pisząc ten podręcznik, mieliśmy również na uwadze potrzeby studentów i wykładowców kierunków takich jak matematyka, statystyka, informatyka, fizyka czy ekonometria – jako podręcznik wspomagający zajęcia ze statystyki lub uczenia maszynowego bądź będący podstawą kursu wprowadzającego do szeroko rozumianej inżynierii i analizy danych. Czerpaliśmy z doświadczenia zdobytego w trakcie prowadzenia zajęć na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej oraz szkoleń z cyklu *Data Science Retreat* w Berlinie.

Struktura książki. Materiał podzieliliśmy na cztery następujące części. Ufamy, że ten przejrzysty układ pozwoli nie tylko lepiej uporządkować przyswajaną wiedzę, ale i później – każdorazowo w razie potrzeby – łatwo wyszukiwać potrzebne informacje.

Podstawy języka Python 3. Zaczynamy od omówienia instalacji dystrybucji Anaconda i sposobów pracy z notatnikami Jupyter w rozdz. 1. Następnie omawiamy w wyczerpujący sposób podstawy „bazowego" języka Python – w każdym razie w zakresie potrzebnym do rozpoczęcia prawdziwej przygody z *data science*. W szczególności interesują nas najważniejsze typy danych: wartości skalarne (rozdz. 2), listy, krotki i inne typy sekwencyjne oraz iterowalne (rozdz. 3), słowniki i zbiory (rozdz. 4), a także instrukcje sterujące (rozdz. 5) i sposoby definiowania własnych funkcji (rozdz. 6).

Przetwarzanie danych. W drugiej części zajmujemy się zagadnieniami związanymi ze wstępnym przetwarzaniem danych i przygotowaniem ich do analizy. Omawiamy szczegółowo pakiet numpy, który udostępnia wektory, macierze i inne *n*-wymiarowe tablice a także szeroką gamę metod i funkcji operujących na nich (rozdz. 7). Dalej skupiamy się na opartym na numpy pakiecie pandas, przy którego użyciu możemy reprezentować i przekształcać rekordy zapisane w postaci tabelarycznej (rozdz. 8). Nie zapominamy przy tym o innych ważnych zagadnieniach: przetwarzaniu napisów i wydobywaniu wiedzy z informacji tekstowych (rozdz. 9), obsłudze plików i automatycznym zbieraniu informacji z internetu (rozdz. 10), a także łączeniu się z bazami danych SQL (rozdz. 11). Co więcej, na zakończenie tej partii materiału przedstawiamy zestaw kilkudziesięciu ćwiczeń, które są poświęcone najczęściej wykonywanym w praktyce operacjom na ramkach danych, m.in. wyszukiwaniu informacji, przekształceniu zmiennych, filtrowaniu wierszy i kolumn, agregacji zmiennych w podgrupach utworzonych przez kombinacje wielu czynników oraz złączaniu tabel.

Analiza danych. W trzeciej części nasza uwaga jest skupiona na szeroko pojętej analizie danych, czyli na różnorakich metodach, które pozwalają przekuwać surowe informacje na użyteczną wiedzę. Najpierw poznajemy pakiety matplotlib i seaborn, na których podstawie będziemy dokonywać wizualizacji różnych aspektów udostępnionych nam danych oraz wyników przeprowadzanych analiz (rozdz. 12). Następnie przechodzimy do opisu dostępnych w środowisku Python metod statystycznych (rozdz. 13) – w szczególności problemów estymacji nieznanych parametrów i charakterystyk rozkładów oraz weryfikacji hipotez. Dzięki nim będziemy potrafili odpowiadać na pewne istotne pytania w sytuacji, gdy mamy do czynienia z *nie-wielkimi* próbkami, np. czy wpływ określonego czynnika na zachowanie się pewnej zmiennej jest rzeczywiście istotny. Z kolei w rozdz. 14 omawiamy trzy najważniejsze grupy algorytmów maszynowego uczenia się: regresji, klasyfikacji i analizy skupień. Przy ich użyciu możemy modelować różne rodzaje zależności między zmiennymi, przewidywać wartości kluczowych charakterystyk dla jeszcze niezaobserwowanych próbek oraz dokonywać automatycznej segmentacji (podziału) zbioru danych na ciekawe podgrupy.

Tworzenie własnego oprogramowania. Ostatnią część książki poświęcamy zagadnieniom z dziedziny inżynierii oprogramowania – dobrze działający proces przetwarzania czy modelowania danych nierzadko należy wdrożyć jako część większego projektu informatycznego. I tak w rozdz. 15 poznajemy sposoby tworzenia własnych modułów, pakietów i skryptów, a w 16 – własnych klas, czyli nowych typów danych.

Dzięki nim możemy efektywniej panować nad złożonością pisanych przez nas programów i wygodniej dzielić się efektami naszej pracy z innymi.

WAŻNE

Materiały uzupełniające do książki, m.in. zbiory danych, przykładowe skrypty i erraty, udostępniliśmy na stronie github.com/gagolews/Analiza_danych_w_jezyku_Python/.

Podziękowania. Chcielibyśmy serdecznie podziękować naszym współpracownikom i przyjaciołom: Barbarze Żogale-Siudem (Instytut Badań Systemowych PAN) i Grzegorzowi Siudemowi (Wydział Fizyki Politechniki Warszawskiej) za liczne uwagi, komentarze i erraty do wstępnej wersji niniejszej książki. Dziękujemy także naszym wyjątkowym studentom na Wydziale Matematyki i Nauk Informacyjnych PW – w szczególności Natalii Potockiej, Martynie Śpiewak oraz Małgorzacie Dobkowskiej – za uczestnictwo w jej „beta testach" oraz p. Izabeli Mika (Instytut Podstawowych Problemów Techniki PAN) za wiele cennych porad redakcyjnych i językowych.

Marek Gągolewski
Maciej Bartoszuk
Anna Cena

Warszawa, sierpień 2016 r.