

Введение

Данные все чаще используются для всевозможных целей, многие из которых ускользают от внимания, но каждый раз, когда вы заходите в Интернет, вы создаете их еще больше. Наука о данных превращает этот огромный объем данных в нечто полезное — то, что вы используете каждый день, чтобы решить набор различных задач или получить услуги от кого-то.

На самом деле вы, вероятно, уже используете науку о данных, причем такими способами, о которых даже не подозреваете. Например, когда вы использовали свою любимую поисковую систему этим утром для поиска чего-либо, она предложила альтернативные поисковые термины. Эти термины предоставлены наукой о данных. Когда на прошлой неделе вы обратились к врачу и оказалось, что обнаруженная опухоль не является раком, доктор, вероятно, сделал этот прогноз с помощью науки о данных. Таким образом, вы можете работать с наукой о данных каждый день и даже не знать об этом. Данная книга не только поможет вам начать использовать науку данных для решения множества практических задач, но также поможет понять, в скольких областях используется наука о данных. Зная, как решать задачи науки о данных и где ее применять, вы получаете значительное преимущество перед всеми остальными, увеличивая свои шансы на продвижение по службе или получение новой работы, которую вы действительно хотите.

О книге

Главная цель книги — устранить пугающий фактор и показать, что наука о данных не только действительно интересна, но и вполне выполнима с использованием языка Python. Возможно, вы думаете, что нужно быть гением информатики, чтобы решать сложные задачи, обычно связанные с наукой о данных, но это далеко от истины. Язык Python поставляется с множеством полезных библиотек, которые выполняют всю тяжелую работу за вас. Вы даже не понимаете, что происходит, и вам не нужно об этом заботиться. Все, что вам действительно нужно знать, — это то, что вы хотите решать конкретные задачи, и язык Python делает эти задачи достаточно доступными.

Частично акцент в книге сделан на использование правильных инструментов. Вы начнете с продукта Anaconda, который включает в себя IPython и Jupyter Notebook — два инструмента, которые исключают необходимость

работы с Python. Вы поэкспериментируете с IPython в полностью интерактивной среде. Код, который вы помещаете в Jupyter Notebook (называемый в этой книге также просто Notebook), обладает качеством презентации, и вы можете смешать несколько элементов презентации прямо в вашем документе. Это совсем не похоже на использование среды разработки. Чтобы упростить использование этой книги на альтернативных платформах, мы также рассмотрим такое приложение интерактивной среды разработки (Interactive Development Environment — IDE) как Google Colab, способное взаимодействовать с большинством, но не со всеми, примерами в книге, используя ваш любимый планшет или (при условии, что вы сможете достаточно сощуриться) смартфон.

В книге вы также обнаружите некоторые интересные приемы. Например, вы сможете создавать графики всех ваших экспериментов с данными, используя Matplotlib. Здесь также подробно описаны доступные ресурсы (такие, как пакеты), и как вы можете использовать пакет Scikit-learn для выполнения действительно интересных вычислений. Многие хотели бы знать, как осуществлять распознавание рукописного ввода, и если вы один из них, то в данной книге получите представление о процессе.

Конечно, вас все еще могут беспокоить проблемы среды программирования, и эта книга не оставит вас в неведении. В начале вы найдете полные инструкции по установке Anaconda, за которыми следует описание методов, необходимых для начала работы с наукой о данных с использованием Jupyter Notebook или Google Colab. Основной упор делается на то, чтобы вы как можно быстрее начали работать и чтобы примеры были простыми и понятными, чтобы код не стал камнем преткновения для обучения.

Во втором издании книги представлены обновленные примеры использования Python 3.x — самой современной версии языка Python. Кроме того, упрощены примеры и добавлен материал по глубокому обучению.

Соглашения, принятые в книге

Здесь используются соглашения, общепринятые в компьютерной литературе.

- » Новые термины в тексте выделяются *курсивом*. Чтобы обратить внимание читателя на отдельные фрагменты текста, также применяется *курсив*.
- » Текст программ, функций, переменных, URL веб-страниц и другой код представлен моноширинным шрифтом.
- » Все, что придется вводить с клавиатуры, выделено **полужирным моноширинным** шрифтом.

- » Знакоместо в описаниях синтаксиса выделено *курсивом*. Это указывает на необходимость заменить знакоместо фактическим именем переменной, параметром или другим элементом, который должен находиться на этом месте: `BINDSIZE= (максимальная ширина колонки) * (номер колонки)`.
- » Пункты меню и названия диалоговых окон представлены следующим образом: Menu Option (Пункт меню).

Текст некоторых абзацев выделен специальным стилем. Это примечания, советы и предостережения, которые обращают внимание на наиболее важные моменты в изложении материала и помогают избежать ошибок в работе.



СОВЕТ

Советы хороши тем, что позволяют сэкономить время или упростить решение некой задачи. Советы также содержат указатели на ресурсы, с которыми имеет смысл ознакомиться, чтобы получить максимум пользы от изучения языка Python применительно к науке о данных.



ВНИМАНИЕ!

Мы не хотим походить на строгих родителей или каких-то маньяков, но вам не следует делать то, что отмечено данной пиктограммой. В противном случае вы можете обнаружить, что ваше приложение не работает должным образом, вы получите неправильные ответы из, казалось бы, “пуленепробиваемых” уравнений или (в худшем случае) потеряете данные.



ТЕХНИЧЕСКИЕ
ПОДРОБНОСТИ

Увидев эту пиктограмму, знайте, что это дополнительный совет (или методика). Вы могли бы найти его очень полезным или слишком скучным, но он может содержать решение, необходимое для запуска программы. Пропускайте эти разделы, если хотите.



ЗАПОМНИ!

Если вы не вынесли ничего из какой-то главы или раздела, то запомните хотя бы материал, отмеченный этой пиктограммой. Здесь обычно содержится наиболее важная информация, которую следует знать для работы с языком Python или успешного решения задач, связанных с наукой о данных.

Глупые предположения

Возможно, вам будет трудно поверить, что мы о вас что-то думали — мы ведь даже не встречались! Хотя большинство предположений действительно глупы, мы сделали их, чтобы обеспечить некую отправную точку для книги.

Вы должны быть знакомы с платформой, которую хотите использовать, поскольку данная книга не предлагает никаких рекомендаций в этом отношении. (Однако в главе 3 приведены инструкции по установке Anaconda, а в главе 4 вы познакомитесь с Google Colab.) Чтобы предоставить вам максимум информации о языке Python и его применении в науке о данных, в этой книге не обсуждаются никакие проблемы, специфичные для платформы. Вам нужно знать, как устанавливать и использовать приложения и вообще работать с выбранной вами платформой, прежде чем начинать работу с этой книгой.

Вы должны знать, как работать с языком Python. В этой редакции книги больше не содержится учебника по языку Python, поскольку в Интернете вы можете найти множество таких учебных пособий (см. примеры на <https://www.w3schools.com/python/> и <https://www.tutorialspoint.com/python/>).

Настоящая книга не является учебником по математике. Да, в ней много примеров сложной математики, но основной упор делается на то, чтобы помочь вам использовать язык Python и науку о данных для решения задач анализа, а не преподавания математической теории. Главы 1 и 2 дадут вам лучшее представление о том, что нужно знать для успешного использования этой книги.

Мы также предполагаем, что вы можете получить доступ к Интернету. И в книге повсюду разбросаны многочисленные ссылки на сетевые материалы, которые расширят ваш опыт обучения. Но эти дополнительные источники полезны только в том случае, если вы их найдете и будете использовать.

Источники дополнительной информации

Эта книга — не конец вашего изучения языка Python или науки о данных, а только начало. Чтобы она стала для вас максимально полезной, мы предоставляем дополнительные источники информации. Получая от вас письма по электронной почте, мы сможем ответить на возникшие у вас вопросы, а также подсказать, как обновления Python или связанных с ним надстроек влияют на содержание книги. Вы также можете использовать следующие замечательные источники.

» **Шпаргалка.** Вы помните, как использовали шпаргалки в школе, чтобы получить лучшие оценки по контрольной? Да, это разновидность шпаргалки. В ней содержится ряд заметок о малоизвестных задачах, решаемых с помощью Python, IPython, IPython Notebook и науки о данных, известных не каждому. Шпаргалка находится в конце книги. В ней содержатся действительно полезная информация: наиболее

распространенные ошибки программирования, вызывающие у людей затруднения при использовании языка Python.

- » **Обновления.** Рано или поздно все изменяется. Например, мы могли не заметить грядущих изменений, когда смотрели в свои хрустальные шары во время написания этой книги. Когда-то это просто означало, что книга устарела и стала менее полезной, но теперь вы можете найти ее обновления по адресу www.dummies.com, если будете искать по английскому названию этой книги. Кроме этих обновлений, имеет смысл посетить блог автора по адресу <http://blog.johnmuellerbooks.com/>, содержащий ответы на вопросы читателей и связанные с книгой полезные материалы.
- » **Сопутствующие файлы.** Кто действительно хочет набрать весь код в книге и восстановить все эти графики вручную? Большинство читателей предпочитают тратить свое время на работу с Python, решение задач по науке о данных и просмотр чего-то интересного, а не на набор текста. К счастью для вас, примеры, используемые в книге, доступны для скачивания, поэтому все, что вам нужно сделать, — это прочитать книгу, чтобы изучить использование языка Python для методов науки о данных. Вы можете найти эти файлы на сайте www.dummies.com. Ищите по английскому названию этой книги и прокрутите вниз появившуюся страницу до изображения обложки книги и щелкните на нем. Затем щелкните на кнопке More about This Book (Подробнее об этой книге) и на открывшейся странице перейдите на вкладку Downloads (Загрузки). Сопутствующие файлы можно также загрузить со страницы русского издания книги по адресу: <http://www.dialektika.com/books/978-5-907203-47-1.html>

Что дальше

Если вы абсолютный новичок в Python и его использовании для задач по науке о данных, вам следует начать с главы 1 и продвигаться по книге со скоростью, позволяющей освоить как можно больше материала.

Если вы новичок, который спешит начать работу с языком Python для науки о данных как можно быстрее, вы можете перейти к главе 3, но имейте в виду, что некоторые темы впоследствии могут оказаться немного запутанными. Переход к главе 5 — это нормально, если у вас уже установлена Anaconda (программный продукт, описанный в книге), но обязательно просмотрите хотя бы главу 3, чтобы вы знали, какие предположения мы сделали при написании этой книги. Если вы планируете использовать для работы с этой книгой планшет, то

обязательно ознакомьтесь с главой 4, чтобы знать об ограничениях, накладываемых Google Colab при выполнении примера кода (не все примеры работают в этой IDE). Обязательно установите Anaconda с соответствующей версией Python 3.6.5, чтобы получить наилучшие результаты из исходного кода книги.

Читатели, которые знакомы с языком Python и имеют установленную программу Anaconda, могут сэкономить время, перейдя непосредственно к главе 5. В случае необходимости у вас всегда есть возможность вернуться к предыдущим главам. Однако вы должны понимать, как работает каждая техника, прежде чем переходить к следующей. Иначе вы можете пропустить жизненно важный момент, если начнете пропускать слишком много информации.

Ждем ваших отзывов!

Вы, читатель этой книги, и есть главный ее критик. Мы ценим ваше мнение и хотим знать, что было сделано нами правильно, что можно было сделать лучше и что еще вы хотели бы увидеть изданным нами. Нам интересны любые ваши замечания в наш адрес.

Мы ждем ваших комментариев и надеемся на них. Вы можете прислать нам электронное письмо либо просто посетить наш веб-сайт и оставить свои замечания там. Одним словом, любым удобным для вас способом дайте нам знать, нравится ли вам эта книга, а также выскажите свое мнение о том, как сделать наши книги более интересными для вас.

Отправляя письмо или сообщение, не забудьте указать название книги и ее авторов, а также свой обратный адрес. Мы внимательно ознакомимся с вашим мнением и обязательно учтем его при отборе и подготовке к изданию новых книг.

Актуальность ссылок не гарантируется.

Наши электронные адреса:

E-mail: info.dialektika@gmail.com

WWW: <http://www.dialektika.com>