

Мэтью Рассел  
Михаил Классен

# Data Mining

---

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ  
ИЗ FACEBOOK, TWITTER,  
LINKEDIN, INSTAGRAM, GITHUB



Санкт-Петербург • Москва • Екатеринбург • Воронеж  
Нижний Новгород • Ростов-на-Дону  
Самара • Минск

2020

ББК 32.973.233-018+32.988.02  
УДК 004.62+004.738.5  
P24

**Рассел Мэтью, Классен Михаил**

P24 Data Mining. Извлечение информации из Facebook, Twitter, LinkedIn, Instagram, GitHub. — СПб.: Питер, 2020. — 464 с.: ил. — (Серия «IT для бизнеса»).

ISBN 978-5-4461-1246-3

В недрах популярных социальных сетей — Twitter, Facebook, LinkedIn и Instagram — скрыты богатейшие залежи информации. Из этой книги исследователи, аналитики и разработчики узнают, как извлекать эти уникальные данные, используя код на Python, Jupyter Notebook или контейнеры Docker.

Сначала вы познакомитесь с функционалом самых популярных социальных сетей (Twitter, Facebook, LinkedIn, Instagram), веб-страниц, блогов и лент, электронной почты и GitHub. Затем приступите к анализу данных на примере Twitter.

**16+** (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.973.233-018+32.988.02  
УДК 004.62+004.738.5

Права на издание получены по соглашению с O'Reilly. Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги. Издательство не несет ответственности за доступность материалов, ссылки на которые вы можете найти в этой книге. На момент подготовки книги к изданию все ссылки на интернет-ресурсы были действующими.

ISBN 978-1491985045 англ.

Authorized Russian translation of the English edition of Mining the Social Web, 3rd Edition. ISBN 9781491985045 © 2019 Matthew A. Russell, Mikhail Klassen  
This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

ISBN 978-5-4461-1246-3

© Перевод на русский язык ООО Издательство «Питер», 2020  
© Издание на русском языке, оформление ООО Издательство «Питер», 2020  
© Серия «IT для бизнеса», 2020



# Краткое содержание

<b>Предисловие .....</b>	<b>16</b>
--------------------------	-----------

## **ЧАСТЬ I. ЭКСКУРСИЯ ПО СОЦИАЛЬНЫМ СЕТЯМ**

<b>Вступление .....</b>	<b>32</b>
-------------------------	-----------

<b>Глава 1.</b> Twitter: исследование актуальных тем, о чем говорят люди и многое другое.....	34
---	----

<b>Глава 2.</b> Facebook: анализ фан-страниц, исследование дружественных связей и многое другое.....	79
--	----

<b>Глава 3.</b> Instagram: компьютерное зрение, нейронные сети, распознавание объектов и лиц .....	126
--	-----

<b>Глава 4.</b> LinkedIn: классификация по профессиям, группировка коллег и многое другое .....	161
---	-----

<b>Глава 5.</b> Анализ текстовых файлов: определение сходства документов, извлечение словосочетаний и многое другое.....	208
--	-----

<b>Глава 6.</b> Анализ веб-страниц: использование методов обработки естественного языка, обобщение статей из блогов и многое другое .....	251
---	-----

<b>Глава 7.</b> Анализ электронной почты: кто кому пишет, о чем, как часто и многое другое .....	305
--	-----

<b>Глава 8.</b> Анализ GitHub: особенности сотрудничества при разработке ПО, графы интересов и многое другое .....	344
--	-----

## ЧАСТЬ II. СБОРНИК РЕЦЕПТОВ ДЛЯ TWITTER

**Глава 9.** Сборник рецептов для Twitter ..... 394

## ЧАСТЬ III. ПРИЛОЖЕНИЯ

**Приложение А.** Информация о виртуальной машине с примерами для этой книги..... 452

**Приложение Б.** Основы OAuth..... 454

**Приложение В.** Советы и рекомендации для Python и Jupyter Notebook ..... 460

**Об авторах**..... 461

**Об обложке** ..... 462