

O'REILLY®



Прикладной  
анализ текстовых  
данных на Python

МАШИННОЕ ОБУЧЕНИЕ И СОЗДАНИЕ ПРИЛОЖЕНИЙ  
ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

 ПИТЕР®

Бенджамин Бенгфорт  
Ребекка Билбро и Тони Охеда

---

# Applied Text Analysis with Python

*Enabling Language-Aware Data Products with  
Machine Learning*

*Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda*

Beijing • Boston • Farnham • Sebastopol • Tokyo

**O'REILLY**®

Бенджамин Бенгфорт  
Ребекка Билбро и Тони Охеда

# Прикладной анализ текстовых данных на Python

---

МАШИННОЕ ОБУЧЕНИЕ И СОЗДАНИЕ ПРИЛОЖЕНИЙ  
ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА



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

2019

20

ББК 32.973.233-018

УДК 004.62

Б46

### **Бенгфорт Бенджамин, Билбро Ребекка, Охеда Тони**

**Б46** Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. — СПб.: Питер, 2019. — 368 с.: ил. — (Серия «Бестселлеры O'Reilly»).

ISBN 978-5-4461-1153-4

Технологии анализа текстовой информации стремительно меняются под влиянием машинного обучения. Нейронные сети из теоретических научных исследований перешли в реальную жизнь, и анализ текста активно интегрируется в программные решения. Нейронные сети способны решать самые сложные задачи обработки естественного языка, никого не удивляет машинный перевод, «беседа» с роботом в интернет-магазине, перефразирование, ответы на вопросы и поддержание диалога. Почему же Сири, Алекса и Алиса не хотят нас понимать, Google находит не то, что мы ищем, а машинные переводчики веселят нас примерами «трудностей перевода» с китайского на албанский? Ответ кроется в мелочах — в алгоритмах, которые правильно работают в теории, но сложно реализуются на практике. Научитесь применять методы машинного обучения для анализа текста в реальных задачах, используя возможности и библиотеки Python. От поиска модели и предварительной обработки данных вы перейдете к приемам классификации и кластеризации текстов, затем приступите к визуальной интерпретации, анализу графов, а после знакомства с приемами масштабирования научитесь использовать глубокое обучение для анализа текста.

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

ББК 32.973.233-018

УДК 004.62

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. 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 978-1491963043 англ.

Authorized Russian translation of the English edition of Mobile Applied Text Analysis with Python, ISBN 9781491963043 © 2018 Benjamin Bengfort, Rebecca Bilbro  
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 978-5-4461-1153-4

© Перевод на русский язык ООО Издательство «Питер», 2019  
© Издание на русском языке, оформление ООО Издательство «Питер», 2019  
© Серия «Бестселлеры O'Reilly», 2019

# Краткое содержание

<b>Вступление</b> .....	11
<b>Глава 1.</b> Естественные языки и вычисления .....	22
<b>Глава 2.</b> Создание собственного корпуса.....	42
<b>Глава 3.</b> Предварительная обработка и преобразование корпуса .....	63
<b>Глава 4.</b> Конвейеры векторизации и преобразования.....	82
<b>Глава 5.</b> Классификация в текстовом анализе .....	112
<b>Глава 6.</b> Кластеризация для выявления сходств в тексте.....	130
<b>Глава 7.</b> Контекстно-зависимый анализ текста .....	161
<b>Глава 8.</b> Визуализация текста.....	190
<b>Глава 9.</b> Графовые методы анализа текста.....	223
<b>Глава 10.</b> Чат-боты .....	249
<b>Глава 11.</b> Масштабирование анализа текста .....	288
<b>Глава 12.</b> Глубокое обучение и не только .....	323
<b>Глоссарий</b> .....	345