

УДК 316.77:004.738.5:004.438Python
ББК 60.52с5
Б81

Бонцанини М.

Б81 Анализ социальных медиа на Python / пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2018. – 288 с.: ил.

ISBN 978-5-97060-574-5

Язык программирования Python является оптимальным выбором для исследователей-аналитиков, поскольку позволяет создавать прототипы, визуализировать и анализировать наборы данных малого и среднего размера. Бесчисленное количество предприятий обращается к Python для решения задач, связанных с выявлением особенностей поведения потребителей и превращением исходных данных в действенную информацию о клиентах. Настоящая книга рассказывает, как с помощью научного инструментария Python получать и анализировать данные из наиболее популярных сетей, таких как Facebook, Twitter, Stack Exchange и др. В русскоязычное издание добавлено приложение об анализе данных из сети «ВКонтакте».

Издание предназначено для специалистов по анализу данных, а также будет полезно всем разработчикам на Python, желающим извлекать коммерческую пользу из социальных сетей.

УДК 316.77:004.738.5:004.438Python
ББК 60.52с5

Copyright © Packt Publishing 2016. First published in the English language under the title «Mastering Social Media Mining with Python – (9781783552016)».

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN 978-1-78646-689-1 (англ.)
ISBN 978-5-97060-574-5 (рус.)

Copyright © 2016 Packt Publishing
© Оформление, издание, перевод, ДМК Пресс, 2018

Содержание

Об авторе	8
О рецензенте	9
Предисловие	10
Глава 1. Социальные медиа, социальные данные и Python	21
Начало	21
Социальные медиа – проблемы и возможности	22
Возможности	23
Проблемы	25
Технология анализа социальных данных	28
Инструменты Python для науки о данных	31
Настройка среды разработки Python	32
Эффективный анализ данных	35
Машинное обучение	39
Обработка естественного языка	43
Анализ социальных сетей	48
Визуализация данных	49
Обработка данных в Python	51
Создание составных конвейеров данных	53
Резюме	54
Глава 2. Твиттер – хештеги, темы и временные ряды	55
Начало работы	55
Twitter API	56
Ограничение частоты запросов	56
Поиск и потоковая обработка	57
Выборка данных из Twitter	58
Получение твитов из ленты	60
Структура твита	62
Применение потокового интерфейса Streaming API	66
Анализ твитов – сущности	69
Анализ твитов – текст	73
Анализ твитов – временные ряды	79
Резюме	82
Глава 3. Пользователи, читатели и сообщества в Twitter	83
Пользователи, друзья и читатели	83
Возвращаясь к интерфейсу Twitter API	83
Структура профиля пользователя	85
Загрузка профилей друзей и читателей	87
Анализ связей	89
Измерение степени влияния и вовлеченности	94
Анализ читателей	98
Анализ диалога	104

Привязка твитов к географической карте	107
От твитов к GeoJSON	108
Простота создания карт с Folium.....	110
Резюме.....	116

Глава 4. Сообщения, страницы и взаимодействие пользователей

в Facebook	117
Интерфейс Facebook Graph API	117
Регистрация приложения	118
Аутентификация и безопасность	119
Доступ к Facebook Graph API из Python	121
Анализ сообщений	124
Структура сообщения.....	127
Частотно-временной анализ	127
Анализ страниц Facebook.....	129
Получение сообщений со страницы	131
Измерение степени вовлеченности	135
Визуализация сообщений в виде облака слов.....	141
Резюме.....	142

Глава 5. Тематический анализ в Google+

Начало работы с Google+ API	144
Поиск в Google+	147
Вывод результатов поиска в веб-интерфейсе	149
Декораторы в Python.....	150
Маршруты и шаблоны Flask.....	151
Заметки и действия со страницы Google+	154
Анализ текстов и статистическая мера TF-IDF для заметок	157
Получение словосочетаний при помощи n-грамм	163
Резюме.....	163

Глава 6. Вопросы и ответы в сети Stack Exchange

Вопросы и ответы	165
Начало работы с Stack Exchange API.....	168
Поиск вопросов с тегами	170
Поиск пользователя	172
Обработка дампов данных из Stack Exchange	175
Классификация текстов по тегам в вопросах	180
Обучение с учителем и классификация текстов	180
Алгоритмы классификации	184
Оценка.....	187
Классификация текстов на данных из сети Stack Exchange	189
Встраивание классификатора в приложение реального времени.....	193
Резюме.....	198

Глава 7. Блоги, RSS, Википедия и обработка естественного языка

Блоги и обработка естественного языка	199
Получение данных из блогов и веб-сайтов	200
WordPress.com API	200

Blogger API	203
Каналы RSS и Atom	206
Получение данных из Википедии	207
Несколько слов о выборке данных из веба	210
Основы обработки естественного языка	210
Предварительная обработка текста	211
Извлечение информации	220
Резюме	225
Глава 8. Анализ других данных	226
Большое количество социальных API	226
Анализ видео на YouTube	226
Анализ открытого программного обеспечения на GitHub	231
Анализ сведений о местных предприятиях в Yelp	238
Создание собственного клиента на Python	243
Простой интерфейс для вызовов по протоколу HTTP	243
Резюме	245
Глава 9. Связанные данные и Семантическая паутина	247
Паутина данных	247
Словарь Семантической паутины	249
Микроформаты	252
Связанные данные и открытые данные	254
Среда описания ресурса RDF	255
Формат данных JSON-LD	256
Инициатива Schema.org	257
Анализ связей из DBpedia	258
Анализ географических координат	260
Извлечение геоданных из Википедии	260
Нанесение геоданных на карты Google Maps	263
Резюме	267
Приложение А. Анализ данных из социальной сети «ВКонтакте»	269
Анализ сообщества и определение его типичного участника	270
Определение центральных узлов социального графа	276
Отображение центральностей на графике	277
Прочие операции	279
Предметный указатель	281