

Оглавление

Предисловие	16
Примечание Мэтью Рассела	16
README.1st	17
Предвосхищая ожидания	17
Технологии на основе Python	20
Новое в третьем издании	22
Этические аспекты добычи данных	24
Типографские соглашения	26
Использование примеров программного кода	27
Благодарности к третьему изданию	28
Благодарности ко второму изданию	28
Благодарности к первому изданию	29
От издательства	30
ЧАСТЬ I. ЭКСКУРСИЯ ПО СОЦИАЛЬНЫМ СЕТЯМ	31
Вступление	32
Глава 1. Twitter: исследование актуальных тем, о чем говорят люди и многое другое	34
1.1. Обзор	34
1.2. Причины популярности Twitter	35
1.3. Twitter API	38
1.3.1. Базовая терминология Twitter	38
1.3.2. Подключение к Twitter API	41
1.3.3. Исследование актуальных тем	46
1.3.4. Поиск твитов	51

1.4. Анализ 140 (или более) символов	58
1.4.1. Извлечение сущностей из твита.....	60
1.4.2. Исследование твитов и сущностей в них с применением частотного анализа.....	62
1.4.3. Определение лексического разнообразия твитов.....	65
1.4.4. Исследование шаблонов в ретвитах.....	68
1.4.5. Визуализация частот с помощью гистограмм	71
1.5. Заключительные замечания.....	75
1.6. Упражнения.....	76
1.7. Онлайн-ресурсы.....	78
Глава 2. Facebook: анализ фан-страниц, исследование дружественных связей и многое другое.....	79
2.1. Обзор	80
2.2. Facebook Graph API	81
2.2.1. Знакомство с Graph API.....	83
2.2.2. Знакомство с Open Graph Protocol	88
2.3. Анализ связей в социальном графе.....	96
2.3.1. Анализ страниц в Facebook	100
2.3.2. Манипулирование данными с помощью pandas.....	113
2.4. Заключительные замечания.....	122
2.5. Упражнения.....	123
2.6. Онлайн-ресурсы.....	124
Глава 3. Instagram: компьютерное зрение, нейронные сети, распознавание объектов и лиц	126
3.1. Обзор	127
3.2. Instagram API.....	128
3.2.1. Выполнение запросов к Instagram API	129
3.2.2. Извлечение своей ленты постов из Instagram	132
3.2.3. Извлечение медиафайлов по хештегу	134
3.3. Анатомия поста в Instagram	135
3.4. Краткое введение в искусственные нейронные сети	138
3.4.1. Обучение нейросети «рассматриванию» изображений.....	140
3.4.2. Распознавание рукописных цифр.....	142
3.4.3. Распознавание объектов на фотографиях с помощью предварительно обученных нейросетей	148

3.5. Применение нейронных сетей для анализа постов в Instagram	152
3.5.1. Классификация содержимого изображения.....	154
3.5.2. Определение лиц на изображениях	154
3.6. Заключительные замечания.....	156
3.7. Упражнения	157
3.8. Онлайн-ресурсы.....	158
Глава 4. LinkedIn: классификация по профессиям, группировка коллег и многое другое.....	161
4.1. Обзор	162
4.2. LinkedIn API	163
4.2.1. Выполнение запросов к LinkedIn API.....	163
4.2.2. Загрузка файла с информацией о контактах в LinkedIn.....	168
4.3. Краткое введение в приемы кластеризации данных.....	169
4.3.1. Нормализация данных для анализа.....	171
4.3.2. Измерение степени сходства.....	185
4.3.3. Алгоритмы кластеризации	188
4.4. Заключительные замечания.....	204
4.5. Упражнения	205
4.6. Онлайн-ресурсы.....	206
Глава 5. Анализ текстовых файлов: определение сходства документов, извлечение словосочетаний и многое другое.....	208
5.1. Обзор	209
5.2. Текстовые файлы.....	209
5.3. Краткое введение в TF-IDF	211
5.3.1. Частота слова	212
5.3.2. Обратная частота документа	214
5.3.3. TF-IDF.....	216
5.4. Оценка запросов данных на естественном языке с использованием TF-IDF... ..	220
5.4.1. Введение в Natural Language Toolkit	221
5.4.2. Вычисление оценки TF-IDF для текста на естественном языке.....	225
5.4.3. Поиск похожих документов	227
5.4.4. Анализ биграмм на естественном языке.....	234
5.4.5. Размышления об анализе данных на естественном языке	246
5.5. Заключительные замечания.....	248

5.6. Упражнения	249
5.7. Онлайн-ресурсы	250
Глава 6. Анализ веб-страниц: использование методов обработки естественного языка, обобщение статей из блогов и многое другое	251
6.1. Обзор	252
6.2. Скрапинг, парсинг и обход сайтов в интернете	253
6.2.1. Обход страниц методом поиска в ширину	257
6.3. Определение семантики декодированием синтаксиса	261
6.3.1. Пошаговая иллюстрация обработки естественного языка	264
6.3.2. Выделение предложений из данных на человеческом языке.....	268
6.3.3. Обобщение документов	273
6.4. Анализ сущностей: смена парадигмы	286
6.4.1. Определение общего смысла данных на человеческом языке	291
6.5. Оценка качества при анализе данных на человеческом языке	297
6.6. Заключительные замечания.....	300
6.7. Упражнения	301
6.8. Онлайн-ресурсы.....	303
Глава 7. Анализ электронной почты: кто кому пишет, о чем, как часто и многое другое.....	305
7.1. Обзор.....	307
7.2. Получение и обработка корпуса с почтовыми сообщениями.....	307
7.2.1. Пример почтового ящика UNIX.....	307
7.2.2. Получение корпуса Enron.....	313
7.2.3. Преобразование почтового корпуса в формат mbox.....	316
7.2.4. Преобразование почтовых ящиков UNIX в объекты DataFrames	317
7.3. Анализ корпуса Enron.....	320
7.3.1. Запрос по диапазону времени	322
7.3.2. Анализ закономерностей во взаимодействиях отправителей и получателей.....	325
7.3.3. Поиск писем по ключевым словам.....	330
7.4. Анализ собственных почтовых данных	332
7.4.1. Доступ к почтовому ящику Gmail с использованием OAuth	334
7.4.2. Извлечение и парсинг электронных писем	337
7.4.3. Визуализация закономерностей в электронных письмах с помощью Immersion	339

7.5. Заключительные замечания	340
7.6. Упражнения	341
7.7. Онлайн-ресурсы	342
Глава 8. Анализ GitHub: особенности сотрудничества при разработке ПО, графы интересов и многое другое	344
8.1. Обзор	345
8.2. GitHub API.....	346
8.2.1. Подключение к GitHub API	348
8.2.2. Выполнение запросов к GitHub API	352
8.3. Моделирование данных с помощью графов свойств.....	355
8.4. Анализ графов интересов в GitHub.....	359
8.4.1. Начало создания графа интересов.....	359
8.4.2. Вычисление мер центральности графа.....	364
8.4.3. Расширение графа интересов ребрами «следования» между пользователями	367
8.4.4. Использование узлов в качестве точек опоры для увеличения эффективности запросов.....	380
8.4.5. Визуализация графа интересов.....	387
8.5. Заключительные замечания.....	388
8.6. Упражнения	390
8.7. Онлайн-ресурсы	391
ЧАСТЬ II. СБОРНИК РЕЦЕПТОВ ДЛЯ TWITTER	393
Глава 9. Сборник рецептов для Twitter	394
9.1. Доступ к Twitter API для целей разработки	395
9.1.1. Задача	395
9.1.2. Решение	395
9.1.3. Пояснение	395
9.2. Использование OAuth для доступа к Twitter API в промышленных целях	397
9.2.1. Задача	397
9.2.2. Решение	397
9.2.3. Пояснение	397
9.3. Поиск актуальных тем.....	402
9.3.1. Задача	402
9.3.2. Решение	402
9.3.3. Пояснение	402

9.4. Поиск твитов	403
9.4.1. Задача	403
9.4.2. Решение	403
9.4.3. Пояснение	403
9.5. Конструирование удобных вызовов функций	405
9.5.1. Задача	405
9.5.2. Решение	405
9.5.3. Пояснение	406
9.6. Запись и чтение текстовых файлов с данными JSON	407
9.6.1. Задача	407
9.6.2. Решение	407
9.6.3. Пояснение	407
9.7. Сохранение данных JSON в MongoDB и доступ к ним	408
9.7.1. Задача	408
9.7.2. Решение.....	408
9.7.3. Пояснение.....	408
9.8. Получение выборки из потока твитов с использованием Streaming API.....	411
9.8.1. Задача	411
9.8.2. Решение	412
9.8.3. Пояснение	412
9.9. Сбор временных последовательностей данных	413
9.9.1. Задача	413
9.9.2. Решение	414
9.9.3. Пояснение	414
9.10. Извлечение сущностей из твитов	415
9.10.1. Задача	415
9.10.2. Решение	416
9.10.3. Пояснение	416
9.11. Поиск самых популярных твитов в коллекции	417
9.11.1. Задача	417
9.11.2. Решение	417
9.11.3. Пояснение	418
9.12. Поиск самых популярных сущностей в коллекции твитов.....	419
9.12.1. Задача	419
9.12.2. Решение	419
9.12.3. Пояснение	419

9.13. Вывод результатов частотного анализа в табличной форме.....	420
9.13.1. Задача	420
9.13.2. Решение	421
9.13.3. Пояснение	421
9.14. Поиск пользователей, ретвитнувших статус	422
9.14.1. Задача	422
9.14.2. Решение.....	422
9.14.3. Пояснение	422
9.15. Определение автора твита.....	425
9.15.1. Задача	425
9.15.2. Решение	425
9.15.3. Пояснение	425
9.16. Выполнение надежных запросов к Twitter	426
9.16.1. Задача	426
9.16.2. Решение	426
9.16.3. Пояснение	427
9.17. Получение информации из профиля пользователя.....	429
9.17.1. Задача	429
9.17.2. Решение.....	429
9.17.3. Пояснение.....	429
9.18. Извлечение сущностей твитов из произвольного текста.....	431
9.18.1. Задача	431
9.18.2. Решение	431
9.18.3. Пояснение	431
9.19. Получение всех друзей и последователей пользователя	432
9.19.1. Задача	432
9.19.2. Решение	432
9.19.3. Пояснение	432
9.20. Анализ друзей и последователей пользователя	435
9.20.1. Задача	435
9.20.2. Решение	435
9.20.3. Пояснение	435
9.21. Извлечение твитов пользователя.....	436
9.21.1. Задача	436
9.21.2. Решение	437
9.21.3. Пояснение	437

9.22. Обход графа дружбы	439
9.22.1. Задача	439
9.22.2. Решение	439
9.22.3. Пояснение	439
9.23. Анализ содержимого твитов.....	441
9.23.1. Задача	441
9.23.2. Решение	441
9.23.3. Обсуждение	441
9.24. Обобщение целевых ссылок.....	443
9.24.1. Задача	443
9.24.2. Решение.....	443
9.24.3. Пояснение	443
9.25. Анализ избранных твитов пользователя.....	446
9.25.1. Задача	446
9.25.2. Решение	446
9.25.3. Пояснение	447
9.26. Заключительные замечания	448
9.27. Упражнения	448
9.28. Онлайн-ресурсы.....	450
ЧАСТЬ III. ПРИЛОЖЕНИЯ.....	451
Приложение А. Информация о виртуальной машине с примерами для этой книги.....	452
Приложение Б. Основы OAuth.....	454
Обзор	454
OAuth 1.0a.....	455
OAuth 2.0.....	457
Приложение В. Советы и рекомендации для Python и Jupyter Notebook	460
Об авторах	461
Об обложке	462