

# Оглавление

<b>Вступление</b> .....	11
Сложности компьютерной обработки естественного языка .....	12
Лингвистические данные: лексемы и слова .....	12
Внедрение машинного обучения .....	14
Инструменты для анализа текста .....	15
О чем рассказывается в этой книге .....	16
Кому адресована эта книга .....	17
Примеры кода и репозиторий на GitHub .....	18
Типографские соглашения .....	19
Использование программного кода примеров .....	19
От издательства .....	20
Благодарности .....	20
<b>Глава 1. Естественные языки и вычисления</b> .....	22
Парадигма Data Science .....	23
Приложения данных, основанные на анализе естественного языка .....	25
Конвейер приложения данных .....	27
Тройка выбора модели .....	29
Язык как данные .....	31
Компьютерная модель языка .....	31
Лингвистические признаки .....	33
Контекстные признаки .....	36
Структурные признаки .....	38
В заключение .....	41
<b>Глава 2. Создание собственного корпуса</b> .....	42
Что такое корпус? .....	43
Предметные корпуса .....	43
Движок сбора данных Valeen .....	44
Управление корпусом данных .....	46
Структура корпуса на диске .....	48

---

Объекты чтения корпусов.....	51
Потоковый доступ к данным с помощью NLTK .....	53
Чтение корпуса HTML.....	56
Чтение корпуса из базы данных .....	60
В заключение .....	62
<b>Глава 3.</b> Предварительная обработка и преобразование корпуса .....	63
Разбивка документов.....	64
Выявление и извлечение основного контента .....	65
Разделение документов на абзацы .....	66
Сегментация: выделение предложений .....	68
Лексемизация: выделение лексем .....	70
Маркировка частями речи .....	71
Промежуточный анализ корпуса.....	73
Трансформация корпуса .....	74
Чтение предварительно обработанного корпуса.....	79
В заключение .....	81
<b>Глава 4.</b> Конвейеры векторизации и преобразования .....	82
Слова в пространстве .....	83
Частотные векторы .....	85
Прямое кодирование .....	87
Частота слова — обратная частота документа.....	90
Распределенное представление .....	93
Scikit-Learn API .....	97
Интерфейс BaseEstimator.....	97
Расширение TransformerMixin.....	99
Конвейеры.....	104
Основы конвейеров.....	105
Поиск по сетке для оптимизации гиперпараметров .....	106
Усовершенствование извлечения признаков с помощью объектов FeatureUnion.....	107
В заключение .....	110
<b>Глава 5.</b> Классификация в текстовом анализе .....	112
Классификация текста .....	113
Идентификация задач классификации.....	113

Модели классификации .....	115
Создание приложений классификации текста .....	117
Перекрестная проверка .....	118
Конструирование модели .....	122
Оценка модели .....	124
Эксплуатация модели .....	128
В заключение .....	129
<b>Глава 6.</b> Кластеризация для выявления сходств в тексте.....	130
Обучение на текстовых данных без учителя .....	131
Кластеризация документов по сходству .....	132
Метрики расстояния .....	133
Партитивная кластеризация .....	136
Иерархическая кластеризация .....	142
Моделирование тематики документов .....	146
Латентное размещение Дирихле .....	146
Латентно-семантический анализ .....	155
Неотрицательное матричное разложение .....	157
В заключение .....	159
<b>Глава 7.</b> Контекстно-зависимый анализ текста .....	161
Извлечение признаков на основе грамматики.....	162
Контекстно-свободные грамматики .....	163
Синтаксические парсеры .....	163
Извлечение ключевых фраз .....	165
Извлечение сущностей .....	168
Извлечение признаков на основе $n$ -грамм .....	169
Чтение корпуса с поддержкой $n$ -грамм .....	171
Выбор размера $n$ -грамм.....	173
Значимые словосочетания.....	174
Модели языка $n$ -грамм.....	177
Частота и условная частота.....	178
Оценка максимальной вероятности .....	181
Неизвестные слова: возвраты и сглаживание.....	184
Генерация языка .....	186
В заключение .....	188

---

<b>Глава 8. Визуализация текста.....</b>	<b>190</b>
Визуализация пространства признаков.....	191
Визуальный анализ признаков.....	191
Управление конструированием признаков.....	202
Диагностика моделей .....	210
Визуализация кластеров.....	211
Визуализация классов .....	213
Диагностика ошибок классификации .....	214
Визуальная настройка .....	218
Оценка силуэта и локтевые кривые.....	219
В заключение .....	222
<b>Глава 9. Графовые методы анализа текста.....</b>	<b>223</b>
Вычисление и анализ графов .....	225
Создание тезауруса на основе графа.....	225
Анализ структуры графа.....	227
Визуальный анализ графов .....	228
Извлечение графов из текста .....	229
Создание социального графа .....	230
Исследование социального графа .....	233
Разрешение сущностей.....	241
Разрешение сущностей в графе.....	242
Блокирование по структуре.....	244
Нечеткое блокирование .....	244
В заключение .....	247
<b>Глава 10. Чат-боты .....</b>	<b>249</b>
Основы диалогового взаимодействия .....	250
Диалог: непродолжительный обмен .....	253
Управление диалогом.....	256
Правила вежливой беседы .....	258
Приветствие и прощание.....	259
Обработка ошибок при общении .....	264
Занимательные вопросы.....	267
Анализ зависимостей.....	268
Анализ составляющих .....	269

Выявление вопроса .....	272
От столовых ложек к граммам .....	274
Обучение для рекомендаций .....	279
Соседство.....	281
Предложение рекомендаций .....	284
В заключение .....	286
<b>Глава 11. Масштабирование анализа текста .....</b>	<b>288</b>
Модуль multiprocessing .....	289
Запуск параллельных задач .....	292
Пулы процессов и очереди.....	297
Параллельная обработка корпуса.....	299
Кластерные вычисления с использованием Spark .....	301
Устройство заданий в Spark.....	302
Распределение корпуса .....	304
Операции RDD .....	306
Обработка естественного языка в Spark .....	308
В заключение .....	321
<b>Глава 12. Глубокое обучение и не только .....</b>	<b>323</b>
Прикладные нейронные сети.....	324
Нейронные модели языка .....	324
Искусственные нейронные сети.....	325
Архитектуры глубокого обучения .....	331
Анализ эмоциональной окраски .....	336
Глубокий анализ структуры .....	338
Будущее (почти) наступило .....	343
<b>Глоссарий.....</b>	<b>345</b>
<b>Об авторах.....</b>	<b>362</b>
<b>Выходные данные .....</b>	<b>364</b>