

Обработка естественного языка в действии

ВТОРОЕ ИЗДАНИЕ

ХОБСОН ЛЕЙН, МАРИЯ ДЫШЕЛЬ



Санкт-Петербург • Москва • Минск

2026

Краткое содержание

Часть I. Обработка слов: векторные модели естественного языка

Глава 1. Модели, которые читают и пишут: знакомство с технологией NLP	32
Глава 2. Знаки мысли: слова естественного языка	80
Глава 3. Арифметика слов: векторы TF-IDF	132
Глава 4. Поиск смысла слов по их частоте: семантический анализ	180

Часть II. Более глубокое обучение: нейронные сети

Глава 5. Словесный мозг: нейронные сети.....	230
Глава 6. Умозаключения на основе векторных представлений слов	267
Глава 7. Поиск крупниц знаний в тексте с помощью CNN	322
Глава 8. Сокращение, повторное использование и переработка слов: RNN и LSTM.....	373

Часть III. Поговорим серьезно: реальные задачи NLP

Глава 9. Нарастиваемое глубокое обучение: трансформеры	422
Глава 10. Большие языковые модели в реальном мире.....	481
Глава 11. Извлечение информации и графы знаний	550
Глава 12. Начинаем общаться: диалоговые системы.....	596

Приложения

Приложение А. Инструменты NLP.....	652
Приложение Б. Эксперименты с Python и регулярные выражения	664
Приложение В. Векторы и линейная алгебра.....	673
Приложение Г. Инструменты и методы машинного обучения	690
Приложение Д. Развертывание контейнерных микросервисов NLU.....	710
Приложение Е. Глоссарий.....	729

Оглавление

Предисловие	14
Благодарности	17
Введение	21
Об авторах	28
Иллюстрация на обложке	30

Часть I. Обработка слов: векторные модели естественного языка

Глава 1. Модели, которые читают и пишут тексты: знакомство с технологией NLP	32
1.1. Языки программирования и NLP	33
1.1.1. Понимание естественного языка	36
1.1.2. Генерация естественного языка	37
1.1.3. Объединение всех компонентов ИИ для достижения положительного эффекта	39
1.2. Волшебство естественного языка	40
1.2.1. Язык и мышление	42
1.2.2. Машины, которые общаются	44
1.2.3. Математика	45
1.3. Практическое применение	47
1.3.1. Обработка языков программирования с помощью NLP	51
1.4. Язык глазами компьютера	52
1.4.1. Язык замков	53
1.4.2. Регулярные выражения	53
1.5. Создание простого чат-бота	56
1.5.1. Распознавание приветствий с помощью ключевых слов	57
1.5.2. Распознавание намерений с использованием регулярных выражений	58
1.5.3. Другой способ распознавания приветствий	62
1.6. Краткая экскурсия по гиперпространству	66
1.7. Порядок слов и грамматика	69
1.8. Конвейер чат-бота на естественном языке	70
1.9. Углубленная обработка	73
1.10. IQ естественного языка	76
1.11. Проверьте себя	78
Резюме	79

Глава 2. Знаки мысли: слова естественного языка	80
2.1. Токены и токенизация	81
2.1.1. Набор инструментов для создания токенизатора	82
2.1.2. Простейший токенизатор	82
2.1.3. Токенизация на основе правил	83
2.1.4. SpaCy	85
2.1.5. Поиск самого быстрого токенизатора слов	87
2.2. За пределами токенов	89
2.2.1. Токенизаторы WordPiece	91
2.3. Улучшение словаря	97
2.3.1. Расширяем словарь n-граммами	97
2.3.2. Нормализация словаря	102
2.4. Сложные токены: обработка логографических языков	112
2.4.1. Сложности лемматизации и стемминга в китайском языке	113
2.5. Векторы токенов	115
2.5.1. Унитарные векторы	115
2.5.2. Векторы в форме «мешка слов»	119
2.5.3. Почему не используются векторы в форме «мешка символов»	121
2.6. Тональность	123
2.6.1. VADER — анализатор тональности на основе правил	125
2.6.2. Наивный байесовский классификатор	126
2.7. Проверьте себя	131
Резюме	131
Глава 3. Арифметика слов: векторы TF–IDF	132
3.1. Мультимножество слов	134
3.2. Векторизация текста с помощью конструктора DataFrame	140
3.2.1. Подсчет токенов быстрее, удобнее, проще	143
3.2.2. Векторизация кода	145
3.2.3. Векторное пространство TF–IDF	148
3.3. Расстояние между векторами и сходство векторов	152
3.3.1. Скалярное произведение	154
3.4. Подсчет частот TF–IDF	157
3.4.1. Анализ пакета this	159
3.5. Закон Ципфа	164
3.6. Обратная частота документа (IDF)	167
3.6.1. Возвращаемся к закону Ципфа	170
3.6.2. Ранжирование по релевантности	172
3.6.3. Математика сглаживания	175
3.7. Использование оценок TF–IDF в чат-боте	175
3.8. Что дальше	178
3.9. Проверьте себя	179
Резюме	179
Глава 4. Поиск смысла слов по их частоте: семантический анализ	180
4.1. От частот слов до оценок тем	181
4.1.1. Ограничения векторов TF–IDF и лемматизации	182
4.1.2. Векторы тем	183
4.1.3. Мысленный эксперимент	184
4.1.4. Алгоритм оценки тем	188

4.2. Задача обнаружения недоброжелательности.....	189
4.2.1. Классификатор на основе линейного дискриминантного анализа.....	191
4.2.2. За пределами линейности.....	197
4.3. Понижение размерности.....	197
4.3.1. Введение в метод главных компонент.....	199
4.3.2. Сингулярное разложение.....	201
4.4. Латентный семантический анализ.....	203
4.4.1. Погружение в семантический анализ.....	205
4.4.2. TruncatedSVD или PCA?.....	208
4.4.3. Насколько эффективен метод LSA для выявления недоброжелательных комментариев?.....	209
4.4.4. Другие способы понижения размерности.....	210
4.5. Латентное размещение Дирихле.....	211
4.5.1. Основная идея LDiA.....	212
4.5.2. Тематическая модель LDiA для комментариев.....	214
4.5.3. Определение недоброжелательности с помощью LDiA.....	217
4.6. Расстояние и сходство.....	219
4.7. Стиринг и обратная связь.....	221
4.8. Мощь векторов тем.....	222
4.8.1. Семантический поиск.....	223
4.9. Добавление в бот поддержки семантического поиска.....	226
4.10. Проверьте себя.....	227
Резюме.....	228

Часть II. Более глубокое обучение: нейронные сети

Глава 5. Словесный мозг: нейронные сети.....	230
5.1. Зачем нужны нейронные сети.....	231
5.1.1. Нейронные сети для обработки слов.....	232
5.1.2. Нейроны как механизмы конструирования признаков.....	234
5.1.3. Биологические нейроны.....	237
5.1.4. Перцептрон.....	239
5.1.5. Перцептрон на Python.....	241
5.2. Пример логистического нейрона.....	246
5.2.1. Логика обработки кликбейтных заголовков.....	247
5.2.2. Половое воспитание.....	248
5.2.3. Логика определения пола.....	249
5.2.4. Новый изящный нейрон PyTorch.....	257
5.3. Спуск по склону ошибок.....	261
5.3.1. Спрыгиваем с подъемника на склон: градиентный спуск и локальные минимумы.....	262
5.3.2. Встряхивание: стохастический градиентный спуск.....	263
5.4. Проверьте себя.....	265
Резюме.....	266
Глава 6. Умозаключения на основе векторных представлений слов.....	267
6.1. Это твой мозг под действием слов.....	268
6.2. Практическое применение.....	270
6.2.1. Поиск смысла.....	271
6.2.2. Объединение векторных представлений слов.....	274
6.2.3. Вопросы на нахождение аналогии.....	275
6.2.4. Инновации Word2Vec.....	276
6.2.5. Искусственный интеллект опирается на векторные представления.....	277

6.3. Word2Vec	278
6.3.1. Рассуждения по аналогии.....	281
6.3.2. Обучение векторным представлениям слов	285
6.3.3. Обучение значениям слов без словаря	290
6.3.4. Использование модуля gensim.word2vec	296
6.3.5. Как сгенерировать векторные представления слов.....	299
6.4. Альтернативы моделям Word2Vec	301
6.4.1. GloVe.....	302
6.4.2. FastText.....	303
6.4.3. Word2Vec по сравнению с LSA	305
6.4.4. Статические и контекстуальные векторные представления.....	306
6.4.5. Визуализация связей между словами	307
6.4.6. Установление связей.....	314
6.4.7. Искусственные слова	319
6.5. Проверьте себя	320
Резюме	321
Глава 7. Поиск крупниц знаний в тексте с помощью CNN	322
7.1. Закономерности в последовательностях слов.....	324
7.2. Свертка.....	326
7.2.1. Трафареты для текста на естественном языке	327
7.2.2. Еще немного о трафаретах	329
7.2.3. Корреляция и свертка.....	332
7.2.4. Свертка как функция преобразования	332
7.2.5. Пример свертки на Python	333
7.2.6. Одномерная CNN для анализа четырехмерных векторных представлений.....	338
7.2.7. Примеры из природы	342
7.3. Азбука Морзе	343
7.3.1. Декодирование морзянки с помощью свертки.....	346
7.4. Создание сверточной нейронной сети с помощью PyTorch	350
7.4.1. Обрезка и дополнение	353
7.4.2. Применение векторных представлений слов	355
7.4.3. Перенос обучения.....	359
7.4.4. Повышение устойчивости CNN с помощью прореживания	362
7.5. PyTorch CNN для обработки постов о чрезвычайных ситуациях.....	363
7.5.1. Архитектура сети.....	363
7.5.2. Субдискретизация.....	367
7.5.3. Линейные слои	368
7.5.4. Подготовка к обучению.....	369
7.5.5. Настройка гиперпараметров.....	369
7.6. Проверьте себя	371
Резюме	372
Глава 8. Сокращение, повторное использование и переработка слов: RNN и LSTM.....	373
8.1. Где можно использовать RNN.....	375
8.1.1. Обработка последовательностей в RNN.....	378
8.1.2. Рекуррентные сети запоминают все, что получают на входе.....	378
8.1.3. RNN скрывают свое понимание	381
8.1.4. Рекуррентные сети запоминают все, что вы им говорите	384
8.2. Определение национальности по фамилии	385
8.2.1. Создание RNN с нуля.....	393
8.2.2. Обучение RNN по одному токену за раз	394

8.2.3. Анализ результатов.....	397
8.2.4. Многоклассовая классификация и маркировка несколькими метками.....	401
8.3. Обратное распространение ошибки во времени	402
8.3.1. Инициализация скрытого слоя в RNN	405
8.4. Запоминание в рекуррентных сетях.....	407
8.4.1. Модели языка на уровне слов	408
8.4.2. Управляемый рекуррентный блок	410
8.4.3. Долгая краткосрочная память.....	413
8.5. Прогнозирование.....	417
8.6. Проверьте себя	419
Резюме	420

Часть III. Поговорим серьезно: реальные задачи NLP

Глава 9. Нарастиваемое глубокое обучение: трансформеры	422
9.1. Рекурсия и рекуррентность	424
9.1.1. Внимание — это не все, что вам нужно	426
9.1.2. Набор LEGO для изучения языка	428
9.2. Заполнение пробелов во внимании	436
9.2.1. Позиционное кодирование.....	436
9.2.2. Соединение всех частей.....	437
9.2.3. Трансформер-переводчик.....	440
9.3. Двухнаправленное обратное распространение ошибки и BERT.....	465
9.3.1. Токенизация и предварительное обучение	466
9.3.2. Тонкая настройка	467
9.3.3. Реализация.....	467
9.3.4. Тонкая настройка предварительно обученной модели BERT для классификации текста	470
9.4. Проверьте себя	480
Резюме	480
Глава 10. Большие языковые модели в реальном мире	481
10.1. Большие языковые модели.....	483
10.1.1. Масштабирование.....	486
10.1.2. Более умные и меньшие по размеру LLM.....	494
10.1.3. Семантическая маршрутизация и защитные механизмы	495
10.1.4. Тестирование моделей на безопасность и устойчивость.....	504
10.2. Генерирование слов с помощью своей LLM.....	506
10.2.1. Создание своей генеративной модели LLM	507
10.2.2. Тонкая настройка генеративной модели.....	513
10.2.3. Бессмыслица — галлюцинация	517
10.3. Повышение IQ модели с помощью поиска	519
10.3.1. Поиск слов — полнотекстовый поиск	520
10.3.2. Поиск смысла — семантический поиск	522
10.3.3. Масштабирование семантического поиска.....	524
10.3.4. Поиск приблизительного ближайшего соседа.....	525
10.3.5. Выбор индекса.....	526
10.3.6. Квантование	528
10.3.7. Собираем все воедино с помощью Haystack.....	532
10.3.8. Воплощаем замысел в реальность.....	533
10.3.9. Haystack — хранилище знаний.....	533
10.3.10. Ответы на вопросы.....	537
10.3.11. Сочетание семантического поиска с генерированием текста.....	537
10.3.12. Развертывание приложения в облаке.....	539

10.3.13. Лучше обслуживайте своих пользователей.....	542
10.3.14. Этика и безопасность ИИ.....	547
10.4. Проверьте себя.....	548
Резюме.....	549
Глава 11. Извлечение информации и графы знаний.....	550
11.1. Заземление.....	552
11.1.1. Старомодный подход — извлечение информации с помощью паттернов.....	553
11.2. Первый шаг: разделение текста на предложения.....	554
11.2.1. Почему split('!?') не работает?.....	555
11.2.2. Сегментация предложений с помощью регулярных выражений.....	556
11.2.3. Семантика предложений.....	560
11.3. Конвейер извлечения знаний.....	564
11.4. Распознавание сущностей.....	566
11.4.1. Распознавание объектов с помощью регулярных выражений: извлечение координат GPS.....	567
11.4.2. Распознавание поименованных сущностей с помощью spaCy.....	568
11.5. Разрешение кореференций.....	571
11.5.1. Разрешение кореференций с помощью spaCy.....	572
11.5.2. Нормализация имен сущностей.....	576
11.6. Анализ зависимостей.....	576
11.6.1. Анализ грамматики составляющих с помощью beeperag.....	580
11.7. От анализа зависимостей к извлечению отношений.....	581
11.7.1. Извлечение отношений с помощью паттернов.....	582
11.7.2. Извлечение отношений с помощью нейронных сетей.....	585
11.8. Создание базы знаний.....	587
11.8.1. Большой граф знаний.....	589
11.9. Поиск ответов в графе знаний.....	591
11.9.1. От вопросов к запросам.....	594
11.10. Проверьте себя.....	595
Резюме.....	595
Глава 12. Начинаем общаться: диалоговые системы.....	596
12.1. Чат-боты повсюду.....	597
12.1.1. Чат-боты разные, а инструменты одни и те же.....	599
12.1.2. Проектирование диалогов.....	601
12.1.3. Первое дерево диалогов.....	602
12.1.4. Критерии хорошего диалога.....	605
12.1.5. Как сделать чат-бот хорошим слушателем: неявные и явные подтверждения.....	606
12.1.6. Использование элементов графического интерфейса.....	608
12.2. Поиск смысла во вводе пользователя: понимание естественного языка.....	608
12.2.1. Распознавание намерений.....	608
12.2.2. Многозначная классификация.....	612
12.3. Генерирование ответа.....	613
12.3.1. Подход на основе шаблонов.....	613
12.3.2. Графы диалогов.....	614
12.3.3. Хранение графа в реляционной базе данных.....	617
12.3.4. Масштабирование контента: подход на основе поиска.....	619
12.3.5. Проектирование сложной логики: программный подход.....	619
12.4. Генеративный подход.....	621
12.5. Фреймворки чат-ботов.....	624
12.5.1. Создание чат-бота на основе намерений с помощью Rasa.....	625
12.5.2. Добавление LLM в чат-бот с помощью LangChain.....	629

12.6. Сопровождение чат-бота	637
12.7. Оценка чат-бота	639
12.7.1. Выбор оценок эффективности чат-бота	639
12.7.2. Оценка качества NLU	640
12.7.3. Оценка пользовательского опыта	641
12.7.4. Что дальше?	644
12.8. Проверьте себя	649
Резюме	650

Приложения

Приложение А. Инструменты NLP	652
A.1. Установка nlria2 в виртуальной среде	653
A.2. Батарейки в комплекте	654
A.3. Anaconda3	654
A.4. Установка пакета nlria2	655
A.5. Интегрированная среда разработки	656
A.6. Система управления пакетами Debian	657
A.7. Mac	658
A.8. Настройки	658
A.9. Windows	660

Приложение Б. Эксперименты с Python и регулярные выражения	664
Б.1. Среда для экспериментов	665
Б.2. Работа со строковыми значениями	667
Б.3. Ассоциативные массивы в Python: dict и OrderedDict	668
Б.4. Регулярные выражения	669
Б.5. Стиль	671
Б.6. Овладейте в совершенстве	672

Приложение В. Векторы и линейная алгебра	673
В.1. Векторы	673
В.2. Матрицы	681
В.3. Тензоры	688
В.4. Углубляемся в линейную алгебру	689

Приложение Г. Инструменты и методы машинного обучения	690
Г.1. Смена парадигмы	690
Г.2. Насколько хорошо должна быть подогнана модель	696
Г.3. Оптимизация модели	703
Г.4. Оценка модели и меры качества	704
Г.5. Советы профессионалов	709

Приложение Д. Развертывание контейнерных микросервисов NLU	710
Д.1. Многозначный классификатор намерений	710
Д.2. Постановка задачи и обучающие данные	712
Д.3. Архитектура микросервисов	715
Д.4. Запуск и тестирование микросервиса прогнозирования	723

Приложение Е. Глоссарий	729
Е.1. Акронимы	729
Е.2. Терминология	734