Оглавление

Предисловие ... 17

Вступление ... 19

Благодарности ... 27
- Хобсон Лейн ... 29
- Ханнес Макс Хапке ... 29
- Коул Ховард ... 29

Об этой книге ... 30
- Дорожная карта ... 30
- Что вы найдете в книге ... 31
- О коде ... 32
- Дискуссионный форум liveBook ... 32
- Об авторах ... 33
- Об иллюстрации на обложке ... 34
- От издательства ... 34

Часть I. Машины для обработки слов

Глава 1. Знакомство с технологией NLP ... 37

1.1. Естественный язык в сравнении с языком программирования ... 38

1.2. Волшебство ... 39
- 1.2.1. Машины, которые общаются ... 40
- 1.2.2. Математика ... 40

1.3. Практическое применение ... 42

1.4. Язык глазами компьютера ... 44
- 1.4.1. Язык замков ... 45
- 1.4.2. Регулярные выражения ... 46

1.4.3. Простой чат-бот ... 48
1.4.4. Другой вариант ... 52
1.5. Краткая экскурсия по гиперпространству ... 56
1.6. Порядок слов и грамматика ... 58
1.7. Конвейер чат-бота на естественном языке ... 59
1.8. Углубленная обработка ... 62
1.9. IQ естественного языка ... 65
Резюме ... 67

Глава 2. Составление словаря: токенизация слов ... 68
2.1. Непростые задачи: обзор стемминга ... 70
2.2. Построение словаря с помощью токенизатора ... 71
2.2.1. Скалярное произведение ... 80
2.2.2. Измерение пересечений мультимножеств слов ... 81
2.2.3. Улучшение токенов ... 82
2.2.4. Расширяем словарь n-граммами ... 87
2.2.5. Нормализация словаря ... 94
2.3. Тональность ... 103
2.3.1. VADER — анализатор тональности на основе правил ... 104
2.3.2. Наивный байесовский классификатор ... 106
Резюме ... 110

Глава 3. Арифметика слов: векторы TF-IDF ... 111
3.1. Мультимножество слов ... 113
3.2. Векторизация ... 118
3.2.1. Векторные пространства ... 120
3.3. Закон Ципфа ... 126
3.4. Тематическое моделирование ... 129
3.4.1. Возвращаемся к закону Ципфа ... 132
3.4.2. Ранжирование по релевантности ... 133
3.4.3. Инструменты ... 136
3.4.4. Альтернативы ... 137
3.4.5. Okapi BM25 ... 138
3.4.6. Что дальше? ... 139
Резюме ... 139

Глава 4. Поиск смысла слов по их частотностям: семантический анализ ... 140
4.1. От частотностей слов до оценок тем ... 142
4.1.1. Векторы TF-IDF и лемматизация ... 142
4.1.2. Векторы тем ... 143

4.1.3. Мысленный эксперимент ... 144
4.1.4. Алгоритм оценки тем ... 149
4.1.5. LDA-классификатор ... 150
4.2. Латентно-семантический анализ ... 155
4.2.1. Воплощаем мысленный эксперимент на практике ... 158
4.3. Сингулярное разложение ... 160
4.3.1. U — левые сингулярные векторы ... 162
4.3.2. S — сингулярные значения ... 163
4.3.3. V^T — правые сингулярные векторы ... 164
4.3.4. Ориентация SVD-матрицы ... 165
4.3.5. Усечение тем ... 166
4.4. Метод главных компонент ... 168
4.4.1. PCA на трехмерных векторах ... 170
4.4.2. Хватит возиться с лошадьми, возвращаемся к NLP ... 171
4.4.3. Применение PCA для семантического анализа CMC ... 174
4.4.4. Применение усеченного SVD для семантического анализа CMC ... 176
4.4.5. Насколько хорошо LSA классифицирует спам ... 177
4.5. Латентное размещение Дирихле ... 180
4.5.1. Основная идея LDiA ... 181
4.5.2. Тематическая модель LDiA для CMC ... 183
4.5.3. LDiA + LDA = классификатор спама ... 186
4.5.4. Более честное сравнение: 32 темы LDiA ... 188
4.6. Расстояние и подобие ... 190
4.7. Стиринг и обратная связь ... 192
4.7.1. Линейный дискриминантный анализ ... 194
4.8. Мощь векторов тем ... 195
4.8.1. Семантический поиск ... 197
4.8.2. Дальнейшие усовершенствования ... 200
Резюме ... 200

Часть II. Более глубокое обучение: нейронные сети

Глава 5. Первые шаги в нейронных сетях: перцептроны и метод обратного распространения ошибки ... 203
5.1. Нейронные сети, список ингредиентов ... 204
5.1.1. Перцептрон ... 205
5.1.2. Числовой перцептрон ... 205
5.1.3. Коротко про смещение ... 206
5.1.4. Айда кататься на лыжах — поверхность ошибок ... 221

5.1.5. С подъемника — на склон ... 222
5.1.6. Проведем небольшую реорганизацию ... 223
5.1.7. Keras: нейронные сети на Python ... 224
5.1.8. Вперед и вглубь ... 228
5.1.9. Нормализация: «стильный» входной сигнал ... 228
Резюме ... 229

Глава 6. Умозаключения на основе векторов слов (Word2vec) ... 230
6.1. Семантические запросы и аналогии ... 231
6.1.1. Вопросы на аналогию ... 232
6.2. Векторы слов ... 233
6.2.1. Векторные умозаключения ... 237
6.2.2. Вычисление представлений Word2vec ... 240
6.2.3. Использование модуля gensim.word2vec ... 250
6.2.4. Как сгенерировать свои собственные представления векторов слов ... 252
6.2.5. Word2vec по сравнению с GloVe (моделью глобальных векторов) ... 255
6.2.6. FastText ... 256
6.2.7. Word2vec по сравнению с LSA ... 257
6.2.8. Визуализация связей между словами ... 258
6.2.9. Искусственные слова ... 264
6.2.10. Определение сходства документов с помощью Doc2vec ... 266
Резюме ... 268

Глава 7. Сверточные нейронные сети ... 269
7.1. Усвоение смысла ... 271
7.2. Инструментарий ... 272
7.3. Сверточные нейронные сети ... 273
7.3.1. Стандартные блоки ... 274
7.3.2. Размер шага (свертки) ... 275
7.3.3. Формирование фильтров ... 276
7.3.4. Дополнение ... 278
7.3.5. Обучение ... 279
7.4. Окна и правда узкие ... 280
7.4.1. Реализация на Keras: подготовка данных ... 282
7.4.2. Архитектура сверточной нейронной сети ... 288
7.4.3. Субдискретизация ... 288
7.4.4. Дропаут ... 291
7.4.5. Вишенка на торте ... 292

7.4.6. Приступаем к обучению ... 294
7.4.7. Применение модели в конвейере ... 296
7.4.8. Что дальше? ... 297
Резюме ... 299

Глава 8. Нейронные сети с обратной связью: рекуррентные нейронные сети ... 300
8.1. Запоминание в нейронных сетях ... 303
8.1.1. Обратное распространение ошибки во времени ... 308
8.1.2. Когда что обновлять ... 311
8.1.3. Краткое резюме ... 313
8.1.4. Всегда есть какой-нибудь подвох ... 313
8.1.5. Рекуррентные нейронные сети и Keras ... 314
8.2. Собираем все вместе ... 318
8.3. Приступим к изучению прошлого ... 321
8.4. Гиперпараметры ... 321
8.5. Предсказание ... 324
8.5.1. Сохранение состояния ... 325
8.5.2. И в другую сторону ... 326
8.5.3. Что это такое? ... 327
Резюме ... 328

Глава 9. Эффективное сохранение информации с помощью сетей с долгой краткосрочной памятью ... 329
9.1. Долгая краткосрочная память ... 330
9.1.1. Обратное распространение ошибки во времени ... 341
9.1.2. А как же проверка на практике? ... 343
9.1.3. «Грязные» данные ... 345
9.1.4. Возвращаемся к «грязным» данным ... 348
9.1.5. Работать со словами сложно. С отдельными буквами — проще ... 349
9.1.6. Моя очередь говорить ... 354
9.1.7. Моя очередь говорить понятнее ... 357
9.1.8. Мы научились, как говорить, но не что говорить ... 365
9.1.9. Другие виды памяти ... 365
9.1.10. Углубляемся ... 366
Резюме ... 368

Глава 10. Модели sequence-to-sequence и механизм внимания ... 369
10.1. Архитектура типа «кодировщик — декодировщик» ... 370
10.1.1. Декодирование вектора идеи ... 371
10.1.2. Знакомо, правда? ... 374

10.1.3. Диалог с помощью sequence-to-sequence ... 375
10.1.4. Обзор LSTM ... 376
10.2. Компонуем конвейер sequence-to-sequence ... 377
10.2.1. Подготавливаем набор данных для обучения модели sequence-to-sequence ... 378
10.2.2. Модель sequence-to-sequence в Keras ... 379
10.2.3. Кодировщик последовательностей ... 380
10.2.4. Декодировщик идеи ... 382
10.2.5. Формируем сеть sequence-to-sequence ... 383
10.3. Обучение сети sequence-to-sequence ... 384
10.3.1. Генерация выходных последовательностей ... 385
10.4. Создание чат-бота с помощью сетей sequence-to-sequence ... 386
10.4.1. Подготовка корпуса для обучения ... 386
10.4.2. Формирование словаря символов ... 388
10.4.3. Генерируем унитарные тренировочные наборы данных ... 388
10.4.4. Обучение нашего чат-бота sequence-to-sequence ... 389
10.4.5. Формируем модель для генерации последовательностей ... 390
10.4.6. Предсказание последовательности ... 391
10.4.7. Генерация ответа ... 391
10.4.8. Общаемся с нашим чат-ботом ... 392
10.5. Усовершенствования ... 393
10.5.1. Упрощаем обучение с помощью группирования данных ... 393
10.5.2. Механизм внимания ... 394
10.6. На практике ... 396
Резюме ... 398

Часть III. Поговорим серьезно. Реальные задачи NLP

Глава 11. Выделение информации: выделение поименованных сущностей и формирование ответов на вопросы ... 401
11.1. Поименованные сущности и отношения ... 401
11.1.1. База знаний ... 402
11.1.2. Выделение информации ... 405
11.2. Регулярные паттерны ... 405
11.2.1. Регулярные выражения ... 407
11.2.2. Выделение информации и признаков при машинном обучении ... 407
11.3. Заслуживающая выделения информация ... 409
11.3.1. Выделение GPS-координат ... 409
11.3.2. Выделение дат ... 410

11.4. Выделение взаимосвязей (отношений) ... 415
11.4.1. Частеречная (POS) разметка ... 416
11.4.2. Нормализация имен сущностей ... 420
11.4.3. Нормализация и выделение отношений ... 421
11.4.4. Паттерны слов ... 421
11.4.5. Сегментация ... 422
11.4.6. Почему не получится разбить по ('.!?') ... 424
11.4.7. Сегментация предложений с помощью регулярных выражений ... 425
11.5. На практике ... 427
Резюме ... 428

Глава 12. Начинаем общаться: диалоговые системы ... 429
12.1. Языковые навыки ... 430
12.1.1. Современные подходы ... 432
12.1.2. Гибридный подход ... 438
12.2. Подход сопоставления с паттернами ... 439
12.2.1. Сопоставляющий с паттернами чат-бот на основе AIML ... 440
12.2.2. Сетевое представление сопоставления с паттерном ... 447
12.3. Заземление ... 448
12.4. Информационный поиск ... 450
12.4.1. Проблема контекста ... 451
12.4.2. Пример чат-бота на основе информационного поиска ... 453
12.4.3. Чат-бот на основе поиска ... 456
12.5. Порождающие модели ... 459
12.5.1. Разговор в чате про nlpia ... 459
12.5.2. Достоинства и недостатки каждого из подходов ... 462
12.6. Подключаем привод на четыре колеса ... 462
12.6.1. Will — залог нашего успеха ... 463
12.7. Процесс проектирования ... 464
12.8. Маленькие хитрости ... 467
12.8.1. Задавайте вопросы с предсказуемыми ответами ... 467
12.8.2. Развлекайте пользователей ... 467
12.8.3. Если все остальное не дает результата — ищите! ... 468
12.8.4. Стремитесь к популярности ... 468
12.8.5. Объединяйте людей ... 468
12.8.6. Проявляйте эмоции ... 469
12.9. На практике ... 469
Резюме ... 470

Глава 13. Масштабирование: оптимизация, распараллеливание и обработка по батчам ... 471
13.1. Слишком много хорошего (данных) ... 472
13.2. Оптимизация алгоритмов NLP ... 472
13.2.1. Индексация ... 473
13.2.2. Продвинутая индексация ... 475
13.2.3. Продвинутая индексация с помощью пакета Annoy ... 477
13.2.4. Зачем вообще использовать приближенные индексы ... 481
13.2.5. Решение проблемы индексации: дискретизация ... 482
13.3. Алгоритмы с постоянным расходом RAM ... 483
13.3.1. gensim ... 484
13.3.2. Вычисления на графах ... 485
13.4. Распараллеливание вычислений NLP ... 486
13.4.1. Обучение моделей NLP на GPU ... 486
13.4.2. Арендовать или покупать? ... 487
13.4.3. Варианты аренды GPU ... 488
13.4.4. Тензорные процессоры ... 489
13.5. Сокращение объема потребляемой памяти при обучении модели ... 490
13.6. Как почерпнуть полезную информацию о модели с помощью TensorBoard ... 493
13.6.1. Визуализация вложений слов ... 493
Резюме ... 496

Приложения

Приложение А. Инструменты для работы с NLP ... 498
A.1. Anaconda3 ... 498
A.2. Установка пакета nlpia ... 499
A.3. IDE ... 500
A.4. Система управления пакетами Ubuntu ... 501
A.5. Mac ... 502
A.5.1. Система управления пакетами для Macintosh ... 502
A.5.2. Дополнительные пакеты ... 502
A.5.3. Настройки ... 502
A.6. Windows ... 504
A.6.1. Переходите в виртуальность ... 505
A.7. Автоматизация в пакете nlpia ... 505

Приложение Б. Эксперименты с Python и регулярные выражения 506
Б.1. Работа со строковыми значениями 507
Б.1.1. Типы строк: str и bytes 507
Б.1.2. Шаблоны в Python: .format() 508
Б.2. Ассоциативные массивы в Python: dict и OrderedDict 508
Б.3. Регулярные выражения 508
Б.3.1. | — OR 509
Б.3.2. () — группы 510
Б.3.3. [] — классы символов 511
Б.4. Стиль 511
Б.5. Овладейте в совершенстве 512

Приложение В. Векторы и матрицы: базовые элементы линейной алгебры 513
В.1. Векторы 513
В.2. Расстояния 515

Приложение Г. Инструменты и методы машинного обучения 520
Г.1. Выбор данных и устранение предвзятости 520
Г.2. Насколько хорошо подогнана модель 521
Г.3. Знание — половина победы 523
Г.4. Перекрестное обучение 524
Г.5. Притормаживаем модель 525
Г.5.1. Регуляризация 525
Г.5.2. Дропаут 526
Г.5.3. Нормализация по мини-батчам 527
Г.6. Несбалансированные тренировочные наборы данных 527
Г.6.1. Супердискретизация 528
Г.6.2. Субдискретизация 528
Г.6.3. Дополнение данных 529
Г.7. Метрики эффективности 530
Г.7.1. Оценка эффективности классификатора 530
Г.7.2. Оценка эффективности регрессора 532
Г.8. Советы от профессионалов 533

Приложение Д. Настройка GPU на AWS 535
Д.1. Шаги создания экземпляра с GPU на AWS 536
Д.1.1. Контроль затрат 547

Приложение Е. Хеширование с учетом локальности ... 549
Е.1. Векторы высокой размерности принципиально различны ... 550
Е.1.1. Индексы и хеши векторного пространства ... 550
Е.1.2. Мыслим многомерно ... 551
Е.2. Многомерная индексация ... 554
Е.2.1. Хеширование с учетом локальности ... 555
Е.2.2. Приближенный метод поиска ближайших соседей ... 555
Е.3. Предсказание лайков ... 556

Источники информации ... 558
Приложения и идеи проектов ... 558
Курсы и учебные руководства ... 560
Утилиты и пакеты ... 560
Научные статьи и обсуждения ... 561
Конкурсы и премии ... 564
Наборы данных ... 565
Поисковые системы ... 565

Глоссарий ... 569
Акронимы ... 570
Терминология ... 574