УДК 004.891
ББК 32.972.13
Т339

Теофили Т.
Т339 Глубокое обучение для поисковых систем / пер. с анг. Д. А. Беликова. – М.: ДМК Пресс, 2020. – 318 с.: ил.

ISBN 978-5-97060-776-3

В книге рассказывается о том, как использовать глубокие нейронные сети для создания эффективных поисковых систем. Рассматривается несколько компонентов поисковой системы, дается представление о том, как они работают, и приводятся рекомендации по использованию нейронных сетей в разных контекстах поиска. Особое внимание уделено практическому объяснению методов поиска и глубокого машинного обучения на базе примеров, большинство которых включает фрагменты кода.

Автор освещает основные проблемы, связанные с поисковыми системами, и указывает пути решения этих проблем. Он раскрывает принципы тестирования эффективности нейронных сетей, а также измерения их затрат и выгод.

Издание предназначено для читателей, владеющих программированием на среднем уровне и отлаживающих поисковые системы с целью повышения их эффективности, то есть выдачи наиболее релевантных результатов.

УДК 004.891
ББК 32.972.13

ISBN 978-1-617-29479-2 (анг.)
ISBN 978-5-97060-776-3 (рус.)

Содержание

Предисловие 10
От автора 11
Благодарности 13
Об этой книге 14
Об авторе 18
Об иллюстрации на обложке 19

Часть I. Поиск встречается с глубоким обучением 20
Глава 1. Поиск на основе нейронных сетей 21
1.1. Нейронные сети и глубокое обучение 23
1.2. Что такое машинное обучение? 25
1.3. Что глубокое обучение может сделать для поиска 27
1.4. Глубокое обучение: дорожная карта 30
1.5. Получение полезной информации 31
1.5.1. Текст, токены, термы и основы поиска 33
1.5.2. Релевантность прежде всего 41
1.5.3. Классические модели поиска 42
1.5.4. Точность и полнота 43
1.6. Нерешенные проблемы 43
1.7. Открываем черный ящик поисковой системы 45
1.8. Глубокое обучение спешит на помощь 46
1.9. Индекс, пожалуйста, познакомьтесь с нейроном 50
1.10. Обучение нейронной сети 51
1.11. Перспективы поиска на базе нейронных сетей 54
Резюме 54

Глава 2. Генерируем синонимы 56
2.1. Расширение синонимов. Введение 57
2.1.1. Почему синонимы? 58
2.1.2. Сопоставление синонимов на базе словаря 60
2.2. Важность контекста 69
2.3. Нейронные сети прямого распространения 71
2.4. Использование word2vec 75
2.4.1. Настройка word2vec в Deeplearning4j 83
2.4.2. Расширение синонимов на базе Word2vec 84
2.5. Оценки и сравнения 87
2.6. Соображения относительно продукционных систем 88
2.6.1. Синонимы против антонимов 90

Резюме 91

Часть II. Подключение нейронных сетей для использования их в поисковой системе 92

Глава 3. От простого поиска к генерации текста 93

3.1. Информационная потребность в сравнении с запросом: преодоление разрыва 94
3.1.1. Генерация альтернативных запросов 95
3.1.2. Подготовка данных 97
3.1.3. Подведем итог 104
3.2. Обучение на последовательностях 105
3.3. Рекуррентные нейронные сети 107
3.1.1. Внутреннее устройство и динамика РНС 110
3.3.2. Долгосрочные зависимости 113
3.3.3. LSTM-сети 114
3.4. LSTM-сети для генерации текста без контроля 115
3.4.1. Неуправляемое расширение запроса 122
3.5. От неконтролируемой до контролируемой генерации текста 126
3.5.1. Создание моделей sequence-to-sequence 126
3.6. Соображения относительно продукционных систем 129
Резюме 130

Глава 4. Более чувствительные поисковые подсказки 132

4.1. Генерация поисковых подсказок 133
4.1.1. Подсказки при составлении запросов 133
4.1.2. Подсказчики на базе словаря 134
4.2. Lookup API 135
4.3. Проанализированные подсказчики 138
4.4. Использование языковых моделей 145
4.5. Подсказчики на базе контента 149
4.6. Нейронные языковые модели 150
4.7. Нейронная языковая модель на базе символов для создания подсказок 152
4.8. Настройка языковой модели 155
4.9. Вносим разнообразие в подсказки, используя векторные представления слов 164
Резюме 166

Глава 5. Ранжирование результатов поиска с помощью векторных представлений слов 167

5.1. Важность ранжирования 168
5.2. Модели поиска 170
5.2.1. TF-IDF и модель векторного пространства 172
5.2.2. Ранжирование документов в Lucene 175
5.2.3. Вероятностные модели 178
5.3. Поиск информации на базе нейронных сетей 180
5.4. От векторов слов к векторам документов 180

5.5. Оценки и сравнения 186
5.5.1. Класс Similarity, основанный на усредненных векторных представлениях слов 188
Резюме 191

Глава 6. Векторные представления документов для ранжирования и рекомендаций 192

6.1. От векторных представлений слов к векторным представлениям документов 193
6.2. Использование векторов абзацев в ранжировании 196
6.2.1. ParagraphVectorsSimilarity 198
6.3. Векторные представления документов и сопутствующий контент 199
6.3.1. Поиск, рекомендации и сопутствующий контент 200
6.3.2. Использование частых термов для поиска похожего контента 201
6.3.3. Извлечение аналогичного контента с помощью векторов абзаца 210
6.3.4. Извлечение аналогичного контента с помощью векторов из моделей «кодер–декодер» 212
Резюме 214

Часть III. Шаг за пределы 215

Глава 7. Поиск по языкам 216

7.1. Обслуживание пользователей, говорящих на нескольких языках 216
7.1.1. Перевод документов в сравнении с переводом запросов 218
7.1.2. Поиск по нескольким языкам 220
7.1.3. Запросы на нескольких языках поверх Lucene 221
7.2. Статистический машинный перевод 223
7.2.1. Выравнивание 225
7.2.2. Перевод на основе фраз 226
7.3. Работа с параллельными корпусами 227
7.4. Нейронный машинный перевод 229
7.4.1. Модели кодер–декодер 230
7.4.2. Модель «кодер–декодер» для машинного перевода в DL4J 233
7.5. Векторные представления слов и документов для нескольких языков 240
7.5.1. Монолингвальные векторные представления с использованием линейной проекции 241
Резюме 246

Глава 8. Поиск изображений на основе контента 247

8.1. Содержимое изображения и поиск 248
8.2. Взгляд назад: поиск изображений на базе текста 251
8.3. Понять изображения 253
8.3.1. Представления изображений 255
8.3.2. Извлечение признаков 257
8.4. Глубокое обучение для представления изображений 266
8.4.1. Сверточные нейронные сети 267
8.4.2. Поиск изображений 275

8.4.3. Локально-чувствительное хеширование 280
8.5. Работа с непомеченными изображениями 283
Резюме 288

Глава 9. Взглянем на производительность 289
9.1. Производительность и перспективы глубокого обучения 290
9.1.1. От проектирования модели до производства 291
9.2. Индексы и нейроны работают вместе 306
9.3. Работа с потоками данных 309
Резюме 315
Глядя вперед 315

Предметный указатель 317