Оглавление

От издательства **15**
О научном редакторе русского издания 15

Предисловие **16**
Структура книги 17
Для кого эта книга 20
Как работать с книгой 21
Требования к программному обеспечению, среде и ресурсам 21
Условные обозначения 21
Использование исходного кода примеров 22
Благодарности 22

Глава 1. Знакомство с векторными базами данных **24**
Для чего нужны векторные БД 24
Новый тип данных: вектор 25
Поиск по сходству 26
Особенности векторного типа данных 29
Где применяются векторные БД 30
Сравнение векторных и реляционных БД 32
Основы бизнес-математики: бухгалтерская арифметика 32
Векторное представление в реляционных БД 32
Потребность в специфичных возможностях векторов 33
Сравнение векторных и NoSQL баз данных 34
NoSQL БД как векторные хранилища 34
Ограничения векторных расширений для NoSQL БД 34
Когда имеет смысл использовать NoSQL БД 35
Гибридные подходы: объединение векторных и структурированных данных 35
Необходимость в векторах и метаданных 35
Ограничения чисто векторных хранилищ 36
Гибридная архитектура БД 36
Пример гибридного запроса 37
Преимущества гибридного подхода 37
Выводы 38

Глава 2. Эмбеддинги ... 39
Для чего нужны эмбеддинги ... 39
Word2Vec: прорыв, который все изменил ... 40
Doc2Vec: от слов к документам ... 42
От эмбеддингов к современным языковым моделям: подключение трансформеров ... 45
Трансформеры «только энкодер» (BERT и его варианты) ... 46
Трансформеры «только декодер » (семейство GPT) ... 46
Трансформеры «энкодер-декодер» (T5, BART) ... 47
Эмбеддинг-модели: специализированные генераторы эмбеддингов ... 48
Отличия от традиционных моделей ... 48
Роль в современных ИИ-приложениях ... 49
Практическое применение ... 50
Простой RAG-пайплайн ... 51
Библиотека SentenceTransformers: швейцарский нож для векторного представления текстов ... 54
Лучшие практики для работы с SentenceTransformers: подробное руководство ... 58
Эмбеддинг-слой: путь к zero-shot обучению ... 63
Анатомия эмбеддингов трансформера ... 63
Связь с zero-shot обучением ... 65
Условия, необходимые для zero-shot обучения ... 66
Ограничения и особенности ... 68
Последние разработки и тенденции ... 68
Векторная арифметика в Word2Vec: справочное руководство ... 69
Шаг 1: инсталляция и настройка ... 69
Шаг 2: загрузка предварительно обученной модели Word2Vec ... 69
Этап 3: реализация функций векторной арифметики ... 70
Этап 4: классический пример с королем и королевой ... 71
Этап 5: еще несколько интересных примеров ... 72
Этап 6: интерактивный инструмент для запуска примеров ... 72
И еще немного о векторной арифметике ... 73
Выводы ... 73

Глава 3. Поиск по сходству при помощи FAISS ... 74
Основы ... 74
Векторные представления ... 76
Меры близости ... 77
Правила выбора ... 79

Индексы FAISS ... 79
Плоские индексы (грубая сила) ... 79
IVF-индексы ... 80
Индексы на основе LSH ... 82
HNSW-индексы ... 82
Другие специализированные индексы ... 83
Составные и трансформирующие индексы ... 84
Выбор подходящего индекса ... 84
Квантование ... 87
Скалярное квантование ... 87
Продуктное квантование ... 89
Проблема ANN ... 93
Задача ... 95
Минимизация вычислений ... 95
Основные методы ANN в FAISS ... 95
Выбор индекса в FAISS ... 98
Пример кода ... 98
Общие сведения об HNSW-индексах ... 99
Что такое HNSW ... 99
Принцип работы HNSW ... 100
Основные параметры HNSW ... 101
Практический пример: создание системы поиска по сходству ... 103
Показатели производительности ... 104
Лучшие практики ... 105
Архитектура и компоненты FAISS ... 105
Основы ... 105
Основные принципы ... 107
Основные компоненты ... 108
Типовой рабочий процесс ... 110
Иллюстративный пример ... 110
Основные выводы ... 111
Дальнейшее изучение ... 111
Выводы ... 112

Глава 4. Семантический поиск при помощи SQLite3 ... 113

Знакомство с векторным расширением для SQLite ... 113
Обзор архитектуры ... 114
Ограничения ... 115

Настройка среды разработки ... 116
Установка зависимостей ... 116
Проверка установки ... 117
Рабочие прагмы ... 118
Проектирование схемы БД ... 118
Требования к схеме ... 118
Определение таблиц ... 119
Решения, принятые при проектировании схемы ... 120
Подключение к Reddit при помощи адаптера API Reddit для Python ... 121
Создание идентификационных данных для API Reddit ... 121
Реализация клиента PRAW ... 122
Пример применения ... 124
Извлечение и предварительная обработка контента ... 124
Пайплайн предварительной обработки текста ... 125
Фильтрация по качеству ... 128
Генерация и сохранение эмбеддингов ... 128
Генератор эмбеддингов ... 128
Сохранение в базе данных ... 130
Пайплайн пакетной обработки ... 134
Создание векторного индекса ... 136
Общие сведения об индексации в sqlite-vss ... 136
Управление индексом ... 137
Реализация семантического поиска ... 139
Контейнер для результатов поиска ... 139
Поисковый движок ... 140
Вся система целиком ... 146
Пример рабочего процесса ... 146
Пример результатов ... 148
Расширение: инкрементная индексация ... 149
Выводы ... 151

Глава 5. Поисковая система для ArXiv на основе PostgreSQL и pgvector ... 152

Трудности поиска по научной литературе ... 152
ArXiv — идеальный источник данных ... 152
Реальные сценарии использования ... 153
Обоснование выбора стека технологий ... 153
Обзор архитектуры ... 154
Компоненты системы ... 154
Поток данных ... 155

Философия дизайна ... 156
Настройка среды и зависимостей ... 157
Установка PostgreSQL и pgvector ... 157
Настройка среды Python ... 158
Структура папок и настройки ... 159
Проверка и тестирование ... 160
Проектирование базы данных для научных статей ... 160
Принципы проектирования схемы ... 160
Структура основных таблиц ... 162
Стратегия хранения векторов ... 165
Стратегия индексации ... 166
Интеграция с ArXiv и управление PDF-файлами ... 167
Реализация клиента API ArXiv ... 167
Пайплайн загрузки PDF ... 168
Система пакетной обработки ... 169
Извлечение текста из PDF и его обработка ... 171
Трудности извлечения текста из PDF ... 171
Интеллектуальный чанкинг текста ... 174
Генерация и сохранение эмбеддингов ... 175
Стратегия эмбеддинг-модели ... 175
Пайплайн пакетной обработки ... 176
Реализация системы поиска по сходству ... 177
Интерактивное приложение и пользовательский интерфейс ... 180
Упаковка в Docker для локального развертывания ... 182
Архитектура контейнера ... 182
Настройка Docker Compose ... 182
Скрипты для инициализации базы данных ... 184
Процесс развертывания ... 185
Вариант для развертывания в облаке ... 185
Базовая настройка производительности ... 186
Конфигурация индексов ... 186
Производительность запросов ... 186
Управление ресурсами ... 186
Дальнейшие действия ... 186
Имеющиеся ограничения ... 186
Предложения по улучшению ... 187
Результаты работы ... 187
Достоинства системы ... 188

Полученные технические навыки ... 188
Практический исследовательский инструмент ... 188
Основа для разработки более сложных систем ... 188
Потенциал системы ... 188
Выводы ... 189

Глава 6. RAG-система на основе SQLite VSS и Ollama ... 190

Обзор архитектуры ... 191
Фундамент: база данных с поддержкой векторного поиска ... 192
Настройка базы данных с поддержкой векторного поиска ... 192
Проектирование схемы БД для RAG ... 193
Создание поисковых индексов ... 194
Обработка текста и генерация эмбеддингов ... 195
Работа с эмбеддинг-моделью ... 195
Интеллектуальный чанкинг текста ... 196
Сохранение контента в эмбеддингах ... 197
Реализация гибридной поисковой системы ... 198
Алгоритм гибридного поиска ... 198
Компонент семантического поиска ... 199
Компонент поиска по ключевым словам ... 200
Слияние оценок и сортировка ... 201
Интеграция с LLM Ollama ... 202
Клиент API Ollama ... 202
Проверка состояния модели ... 203
RAG-пайплайн ... 203
Форматирование контекста ... 204
Вопросно-ответный пайплайн ... 204
Демонстрация и тестирование ... 206
Загрузка демонстрационных данных ... 206
Основная функция для демонстрации ... 207
Интерактивный интерфейс вопросно-ответной системы ... 208
Утилита для быстрого тестирования ... 209
Дальнейшие шаги: расширение системы ... 209
Улучшение работы с данными ... 209
Оптимизация производительности ... 211
Перенос системы в продакшен ... 212
Продвинутые RAG-паттерны ... 212
Выводы ... 213

Глава 7. Научная RAG-система на основе PostgreSQL и pgvector 214
Цели и возможности системы 215
Обзор архитектуры 215
База данных с расширением pgvector 217
Настройка базы данных 218
Проектирование схемы для научных статей 218
Высокоэффективные векторные индексы 219
Стратегия генерации эмбеддингов 220
Интеграция с ArXiv и обработка PDF-файлов 221
Поиск статей при помощи API ArXiv 221
Интеллектуальное извлечение текста из PDF 222
Интеллектуальный чанкинг текста 225
Пайплайн сохранения чанков и эмбеддингов 225
Многоуровневый семантический поиск 227
Поиск на уровне аннотаций 228
Поиск на уровне разделов 229
Подробнее о RAG-пайплайне 230
Интеграция с локальной LLM Ollama 230
Обнаружение модели и проверка работоспособности 231
Интеллектуальное получение контекста 232
Составление научного промпта 232
Полный RAG-пайплайн 233
Интерактивный интерфейс и демонстрация 235
Основная функция для демонстрации 235
Демонстрация поиска 236
Демонстрация RAG 237
Интерактивный интерфейс для поиска 238
Точка входа и выбор режима 239
Особенности HNSW 239
Оценка результатов 241
Дальнейшие шаги: расширение научной RAG-системы 241
Выводы 245

Глава 8. Полная RAG-система для поиска по сообщениям чатов 247
Цели и возможности системы 248
Обзор архитектуры 249
Цели работы 251
База данных для переписки 252
Проектирование схемы БД 252

Архитектура с тремя таблицами для оптимальной производительности ... 253
Высокоэффективный векторный индекс ... 254
Пайплайн импорта переписки и обработки данных ... 255
Надежный импорт JSON с обработкой ошибок ... 255
Обработка с атомарными транзакциями ... 256
Работа с метками времени и проверка данных ... 256
Восстановление после сбоев и логи ... 257
Эффективная генерация эмбеддингов и пакетная обработка ... 258
Паттерн «Одиночка» для работы с моделью ... 258
Стратегия инкрементной обработки ... 259
Пакетная обработка для оптимальной производительности ... 260
Добавление в базу данных с отслеживанием конфликтов ... 260
Контекстуальный поиск с пониманием диалогов ... 261
Семантический поиск по сходству ... 262
Несколько соединений для расширения контекста ... 262
Итоговое форматирование и структура ... 263
Определение контекста диалога ... 263
Вычисление контекстного окна ... 264
RAG-система для истории переписки ... 265
Управление структурированным контекстом ... 265
Интеграция с локальной LLM Ollama ... 265
Обнаружение модели и проверка работоспособности ... 266
Получение и формирование контекста ... 266
Составление промпта ... 267
Полный RAG-пайплайн с мониторингом производительности ... 268
Полный веб-API на основе FastAPI ... 269
Структура приложения на основе FastAPI ... 269
Модели запросов с проверкой ... 270
Реализация эндпоинта для поиска ... 270
Эндпоинт вопросно-ответной системы на основе RAG ... 271
Мониторинг и статистика работы системы ... 271
Настройка сервера ... 272
Примеры данных и демонстрация ... 273
Генерация реалистичных примеров данных ... 273
Примеры данных на различные темы ... 274
Пайплайн обработки примеров данных ... 275
Комплексная демонстрация системы ... 275
Поэтапная демонстрация возможностей системы ... 276

Демонстрация RAG-системы с условиями ... 276
Импорт в продакшен ... 277
Точки входа в приложение ... 277
Итог: полная персональная база знаний ... 278

Глава 9. Язык векторных запросов ... 279
Основные понятия ... 280
Модель данных ... 280
Структура базового синтаксиса ... 281
Векторные операции ... 282
Поиск по сходству ... 282
Гибридный поиск ... 283
Поиск по диапазону ... 283
Пакетные операции ... 283
Векторные функции и агрегации ... 284
Векторные функции ... 284
Векторные агрегации ... 285

Об авторе ... 287

Иллюстрация на обложке ... 288