

Spis treści |

O autorze	13
O korektorach merytorycznych	14
Wstęp	15

CZĘŚĆ 1. Wprowadzenie do generatywnej sztucznej inteligencji i frameworka LlamaIndex

ROZDZIAŁ 1.

Duże modele językowe	23
Wprowadzenie do generatywnej sztucznej inteligencji i dużych modeli językowych	24
Czym jest generatywna sztuczna inteligencja?	24
Czym jest duży model językowy?	24
Rola modeli LLM we współczesnej technologii	26
Wyzwania związane z modelami LLM	28
Wzbogacanie modeli LLM za pomocą techniki RAG	32
Podsumowanie	33

ROZDZIAŁ 2.

LlamaIndex — ukryty skarb.

Wprowadzenie do ekosystemu LlamaIndex	35
Wymagania techniczne	35
Optymalizacja modeli językowych — dostrajanie, RAG i LlamaIndex	36
Czy RAG jest jedynym rozwiązaniem?	36
Co robi LlamaIndex?	38
Zalety stopniowego ujawniania złożoności	40
Ważny aspekt do uwzględnienia	41
System PITS — praktyczny projekt z użyciem LlamaIndexu	41
Sposób działania PITS	41

Przygotowanie środowiska programistycznego	43
Instalacja Pythona	44
Instalacja Gita	44
Instalacja LlamaIndexu	45
Rejestracja klucza API OpenAI	45
Odkrywanie Streamlita — idealnego narzędzia	
do szybkiego tworzenia i wdrażania	48
Instalacja Streamlita	48
Ostatnie przygotowania	49
Ostatnia kontrola	49
Struktura bazy kodu w LlamaIndexie	50
Podsumowanie	51

CZĘŚĆ 2. Rozpoczęcie pracy nad pierwszym projektem z użyciem frameworka LlamaIndex

ROZDZIAŁ 3.

Rozpoczęcie pracy z LlamaIndexem	55
Wymagania techniczne	55
Podstawowe elementy LlamaIndexu: dokumenty, węzły i indeksy	56
Dokumenty	56
Węzły	59
Ręczne tworzenie obiektu węzła	60
Automatyczne wyodrębnianie węzłów z dokumentów	
za pomocą separatorów	61
Węzły nie lubią być same — pragną relacji	62
Dlaczego relacje są ważne?	64
Indeksy	64
Czy to już wszystko?	67
Jak to właściwie działa?	67
Krótki przegląd kluczowych koncepcji	68
Budowanie pierwszej interaktywnej aplikacji z użyciem	
dużego modelu językowego	69
Wykorzystanie funkcji rejestru w LlamaIndexie	
do zrozumienia logiki i debugowania aplikacji	71
Dostosowywanie modelu LLM używanego przez LlamaIndex	72
Łatwe jak 1, 2, 3	72

Parametr temperatury	73
Jak używać Settings do dostosowywania modeli?	75
Rozpoczęcie pracy nad projektem PITS — ćwiczenie praktyczne	76
Kod źródłowy	77
Podsumowanie	80

ROZDZIAŁ 4.

Wprowadzanie danych do przepływu pracy RAG	81
Wymagania techniczne	81
Wprowadzanie danych za pomocą LlamaHuba	82
Wprowadzenie do LlamaHuba	83
Stosowanie łańcuchów danych z LlamaHuba do wprowadzania treści	84
Wprowadzanie danych ze stron internetowych	84
Wprowadzanie danych z bazy danych	86
Masowe wprowadzanie danych ze źródeł z wieloma formatami plików	87
Podział dokumentów na węzły	91
Proste narzędzia do dzielenia tekstu	91
Stosowanie bardziej zaawansowanych parserów węzłów	93
Stosowanie parserów relacyjnych	96
Parsery węzłów i dzielniki tekstu to to samo?	97
Parametry chunk_size i chunk_overlap	97
Uwzględnianie relacji za pomocą parametru include_prev_next_rel	99
Praktyczne sposoby wykorzystania modeli tworzenia węzłów	100
Praca z metadanymi w celu poprawy kontekstu	101
SummaryExtractor	103
QuestionsAnsweredExtractor	103
TitleExtractor	104
EntityExtractor	105
KeywordExtractor	106
PydanticProgramExtractor	106
MarvinMetadataExtractor	107
Definiowanie własnego ekstraktora	107
Czy posiadanie wszystkich metadanych jest zawsze potrzebne?	108
Szacowanie kosztów użycia ekstraktorów metadanych	109
Najlepsze praktyki minimalizowania kosztów	109
Oszacuj maksymalne koszty przed uruchomieniem rzeczywistych ekstraktorów	110

Ochrona prywatności z ekstraktorami metadanych i nie tylko	112
Usuwanie danych osobowych i innych wrażliwych informacji	113
Stosowanie przepływu wprowadzania danych do poprawy wydajności	115
Obsługa dokumentów zawierających mieszankę tekstu i danych tabelarycznych	118
Praktyka: wprowadzanie materiałów szkoleniowych do aplikacji PITS	119
Podsumowanie	121

ROZDZIAŁ 5.

Indeksowanie z LlamaIndexem	122
Wymagania techniczne	122
Indeksowanie danych — spojrzenie z lotu ptaka	123
Wspólne cechy wszystkich typów indeksów	123
VectorStoreIndex	125
Prosty przykład użycia indeksu VectorStoreIndex	125
Osadzenia	127
Wyszukiwanie podobieństwa	129
Jak LlamaIndex generuje osadzenia?	133
Jak wybrać model osadzający?	134
Przechowywanie i ponowne używanie indeksów	136
StorageContext	137
Różnica między magazynami wektorów a wektorowymi bazami danych	139
Inne typy indeksów w LlamaIndexie	140
SummaryIndex	140
DocumentSummaryIndex	142
KeywordTableIndex	144
TreeIndex	147
KnowledgeGraphIndex	151
Budowanie indeksów na bazie innych indeksów za pomocą grafu ComposableGraph	154
Jak używać grafu ComposableGraph?	155
Więcej szczegółów na temat grafu ComposableGraph	156
Szacowanie potencjalnych kosztów budowy i przeszukiwania indeksów	157
Indeksowanie materiałów do nauki PITS — praktyka	161
Podsumowanie	162

CZĘŚĆ 3. Przeszukiwanie i praca ze zindeksowanymi danymi

ROZDZIAŁ 6.

Zapytania do własnych danych, część 1.

— wyszukiwanie kontekstu	165
Wymagania techniczne	165
Mechanika zapytań — przegląd	166
Podstawowe mechanizmy wyszukiwania	166
Mechanizmy wyszukiwania dla indeksu VectorStoreIndex	167
Mechanizmy wyszukiwania dla indeksu SummaryIndex	170
Mechanizmy wyszukiwania dla indeksu DocumentSummaryIndex	172
Mechanizmy wyszukiwania dla indeksu TreeIndex	174
Mechanizmy wyszukiwania dla indeksu KnowledgeGraphIndex	180
Wspólne cechy mechanizmów wyszukiwania	184
Wydajne wykorzystanie mechanizmów wyszukiwania — operacja asynchroniczna	184
Budowanie bardziej zaawansowanych mechanizmów wyszukiwania	185
Prosta (naiwna) metoda wyszukiwania	185
Implementacja filtrów metadanych	186
Użycie selektorów do bardziej zaawansowanej logiki decyzyjnej	189
Narzędzia	191
Przekształcanie i przeformułowywanie zapytań	192
Tworzenie trafniejszych podzapytań	194
Gęste i rzadkie wyszukiwanie	196
Wyszukiwanie gęste	197
Wyszukiwanie rzadkie	198
Implementacja wyszukiwania rzadkiego w LlamaIndex	200
Inne zaawansowane metody wyszukiwania	204
Podsumowanie	204

ROZDZIAŁ 7.

Zapytania do własnych danych, część 2. —

postprocessing i synteza odpowiedzi	206
Wymagania techniczne	206
Ponowne sortowanie, przekształcanie i filtrowanie węzłów za pomocą postprocesorów	207

Sposoby filtrowania, przekształcania i ponownego sortowania węzłów przez postprocesory	208
SimilarityPostprocessor	210
KeywordNodePostprocessor	211
PrevNextNodePostprocessor	214
LongContextReorder	215
PIINodePostprocessor i NERPIINodePostprocessor	216
MetadataReplacementPostProcessor	216
SentenceEmbeddingOptimizer	218
Postprocesory oparte na czasie	219
Postprocesory ponownie sortujące	221
Uwagi końcowe dotyczące postprocesorów węzłów	226
Synteзаторы odpowiedzi	226
Implementacja technik parsowania wyników	230
Wydobywanie ustrukturyzowanych wyników za pomocą parserów	231
Wydobywanie ustrukturyzowanych wyników za pomocą programów Pydantic	234
Budowanie i stosowanie silników zapytań	235
Metody budowania silników zapytań	235
Zaawansowane zastosowania interfejsu QueryEngine	236
Praktyka — budowanie quizów w aplikacji PITS	244
Podsumowanie	246

ROZDZIAŁ 8.

Budowanie czatbotów i agentów za pomocą LlamaIndexu	247
Wymagania techniczne	247
Czatboty i agenty	248
Silnik czatu	250
Tryby czatu	251
Implementacja strategii agentowych w aplikacjach	261
Tworzenie narzędzi i klas ToolSpec dla agentów	261
Pętle rozumowania	264
OpenAI Agent	265
ReActAgent	270
Jak wchodzimy w interakcję z agentami?	272
Udoskonalanie agentów za pomocą narzędzi użytkowych	272

Wykorzystanie agenta LLMCompiler do bardziej zaawansowanych scenariuszy	276
Wykorzystanie niskopoziomowego API Agent Protocol	279
Praktyka — implementacja śledzenia przebiegu rozmów w aplikacji PITS	281
Podsumowanie	286

CZĘŚĆ 4. Dostosowywanie, inżynieria promptów i końcowe uwagi

ROZDZIAŁ 9.

Dostosowywanie i wdrażanie projektu stworzonego

za pomocą LlamaIndexu	289
Wymagania techniczne	289
Dostosowywanie komponentów RAG	290
Jak LLaMA i LLaMA 2 zmieniły krajobraz modeli otwartoźródłowych?	290
Uruchamianie lokalnego modelu LLM za pomocą LM Studio	292
Routing między modelami LLM za pomocą takich usług jak Neutrino lub OpenRouter	298
A co z dostosowywaniem modeli osadzania?	300
Wygodne i gotowe do użycia Llama Packs	301
Interfejs wiersza poleceń LlamaIndexu	303
Użycie zaawansowanych technik śledzenia i oceny	305
Śledzenie przepływów RAG za pomocą Phoenixa	306
Ocena systemu RAG	309
Wprowadzenie do wdrażania z użyciem frameworka Streamlit	315
Praktyka — przewodnik krok po kroku dotyczący wdrażania	316
Wdrażanie projektu PITS na Streamlit Community Cloud	318
Podsumowanie	322

ROZDZIAŁ 10.

Wytyczne i najlepsze praktyki inżynierii promptów	323
Wymagania techniczne	323
Dlaczego prompty są Twoją tajną bronią?	324
Wykorzystanie promptów przez LlamaIndex	327

Dostosowywanie domyślnych promptów	330
Wykorzystanie zaawansowanych technik tworzenia promptów w LlamaIndexie	333
Złote zasady inżynierii promptów	334
Dokładność i jasność wyrażenia	334
Ukierunkowanie	335
Jakość kontekstu	335
Ilość kontekstu	335
Wymagany format wyjściowy	336
Koszt wnioskowania	337
Ogólne opóźnienie systemu	337
Wybór odpowiedniego modelu LLM do zadania	337
Powszechne metody tworzenia skutecznych promptów	341
Podsumowanie	344

ROZDZIAŁ 11.

Zakończenie i dodatkowe źródła wiedzy	345
Inne projekty i dalsza nauka	345
Zbiór przykładów na stronie LlamaIndexu	345
Przyszłość — nagrody Replita	349
W grupie siła — społeczność LlamaIndexu	350
Kluczowe wnioski i słowo końcowe	350
O przyszłości RAG w szerszym kontekście generatywnej sztucznej inteligencji	352
Krótka filozoficzna myśl	355
Podsumowanie	356