Generative KI mit Python

RAG-Anwendungen und Agentensysteme mit Vektordatenbanken und LLMs

DAS VORWORT

Kapitel 1
Vorwort

Willkommen auf Ihrer Reise in die Welt der generativen KI.

In diesem Teil des Buches schaffen wir die Voraussetzungen dafür, was Sie mit diesem umfassenden Leitfaden zur generativen KI mit Python erarbeiten können.

Wir beginnen damit, in Abschnitt 1.1 das Ziel des Buches zu umreißen, das darin besteht, Ihnen ein tiefes Verständnis der anspruchsvollsten Aspekte der generativen KI zu vermitteln. Das reicht von den theoretischen Konzepten bis hin zur praktischen Anwendung.

Die Zielgruppe wird in Abschnitt 1.2 vorgestellt. Das Buch richtet sich an einen bestimmten Personenkreis, der einige Vorkenntnisse mitbringen muss.

Die spezifischen Voraussetzungen, die Sie erfüllen sollten, werden in Abschnitt 1.3 definiert. Der Abschnitt stellt sicher, dass Sie mit grundlegenden Kenntnissen in Python und idealerweise mit Know-how über das Maschinelle Lernen ausgestattet sind.

In Abschnitt 1.4 wird die Struktur des Buches von den Grundprinzipien bis hin zu fortgeschrittenen Anwendungen beschrieben, die auch immer wieder durch praktische Übungen und Beispiele unterbrochen wird.

In Abschnitt 1.5 biete ich Einblicke in die effektive Nutzung dieses Buches und empfehle einen aktiven Lernansatz mit eigener, praktischer Programmierung und einer kritischen Reflexion der Konzepte.

Ergänzend zu den Textinhalten werden herunterladbare Code-Beispiele und Anleitungen für die Systemeinrichtung bereitgestellt. Mehr dazu in Abschnitt 1.6.

In Abschnitt 1.7 erfahren Sie mehr über die Einrichtung Ihres Systems, so dass Sie meine Programmierbeispiele nachvollziehen und idealerweise eigene Ideen auf der Grundlage des neuen Wissens umsetzen können.

Die Danksagungen sind Ausdruck meiner Dankbarkeit gegenüber denjenigen, die die Entstehung dieses Buches unterstützt haben. Ich widme meinen Unterstützern und Unterstützerinnen den Abschnitt 1.8.

Der Abschnitt 1.9 macht Sie mit den typografischen Hinweisen vertraut, die eine reibungslose Navigation durch den Inhalt des Buches ermöglichen. Dieses Vorwort ist Ihr Fahrplan zur Maximierung des Nutzens, den Sie aus dieser Reise in die Welt der generativen KI mit Python ziehen können.

1.1 Zielsetzung des Buches

Willkommen zu einer umfassenden Erkundung der generativen KI mit einem Schwerpunkt auf einigen ihrer transformativsten und fortschrittlichsten Technologien. Das primäre Ziel dieses Buches ist es, Sie durch dieses dynamische Feld zu führen. Ein besonderes Augenmerk liegt dabei auf Large Language Models (LLMs), der nuancierten Kunst des Prompt Engineerings, dem Nutzen von Vektordatenbanken, den innovativen Prozessen der Retrieval-Augmented Generation (RAG) und dem immer größer werdenden Thema der Agentensysteme.

Große Sprachmodelle haben die Art und Weise, wie wir mit textbasierten KI-Systemen interagieren, revolutioniert. Sie verfügen über außergewöhnliche Fähigkeiten im Bereich des Sprachverständnisses und der Textgenerierung. Lassen Sie uns in die Architekturen dieser Modelle eintauchen und herausfinden, wie sie aus riesigen Datenmengen lernen, um Text zu erzeugen, der auch von einem Menschen geschrieben sein könnte.

Von dort aus werden wir uns mit dem Thema Prompt Engineering beschäftigen – einer wichtigen Fähigkeit im Zeitalter der LLMs. Sie lernen, wie man Eingabeaufforderungen (Prompts) erstellt, die es ermöglichen, effizient durch das Wissen und die Fähigkeiten des Modells zu navigieren.

Vektordatenbanken stellen den nächsten großen Schritt in der Organisation und dem Abruf von Daten dar. Wenn Sie diese Technologie verstehen, sind Sie bestens gerüstet, um mit hochdimensionalen Daten zu arbeiten und Systeme zu entwickeln, die schnellen und relevanten Zugriff auf Informationen ermöglichen. Lassen Sie uns gemeinsam die grundlegenden Konzepte, das Design und die Funktionsweise dieser Datenbanken erkunden und herausfinden, wie sie den Weg für anspruchsvolle KI-Anwendungen ebnen.

Das Konzept der Retrieval-Augmented Generation (RAG) verbindet die Suche nach relevanten Informationen mit der sofortigen Textgenerierung. RAG-Systeme stellen einen bedeutenden Meilenstein in der Entwicklung von KI dar, da sie Modellen helfen, genauere und Inhalte mit mehr Informationsgehalt zu erstellen. Das RAG-Kapitel wird einen tiefen Einblick in diese Mechanismen geben und zeigen, wie Sie diese Technik in generative Anwendungen integrieren können.

Anschließend tauchen wir in die Welt der agentischen Systeme ein. Das sind KI-Systeme, die in der Lage sind, autonom zu handeln, Entscheidungen zu treffen und Aufgaben zu übernehmen, die normalerweise menschliche Intelligenz erfordern. Wir werden die ethischen, technologischen und praktischen Aspekte solcher Systeme

erkunden. Dadurch werden Sie in der Lage sein, KI mit Autonomie und Eigenverantwortung auf eine verantwortungsvolle und innovative Weise zu gestalten.

Bis zu diesem Punkt werden Sie Anwendungen lokal entwickeln, aber irgendwann möchten Sie sich sicher auch mit der Welt teilen. Im Kapitel über die Bereitstellung von KI-Systemen (Deployment) werden wir uns noch intensiv mit diesem Thema beschäftigen.

Dieses Buch ist nicht nur darauf ausgelegt, Ihnen ein tiefes theoretisches Verständnis in diesen Bereichen zu vermitteln, sondern auch darauf, Ihnen praktische Fähigkeiten an die Hand zu geben. Durch eine Reihe sorgfältig ausgewählter Beispiele, Fallstudien und praktischer Projekte werde ich Sie dabei unterstützen, diese Konzepte in verschiedenen Szenarien anzuwenden – sei es für die berufliche Weiterentwicklung, die akademische Forschung oder einfach aus persönlicher Neugier.

Mein Ziel geht über bloßen Wissenstransfer hinaus. Ich möchte Lesende dazu befähigen, versierte Schöpfer und Innovatorinnen im Bereich der generativen KI zu werden. Ich will umfassend darauf vorbereiten, die Herausforderungen und Chancen, die diese Technologie in der modernen Welt mit sich bringt, anzugehen und zu meistern. Am Ende Ihrer Reise mit diesem Buch werden Sie nicht nur ein Verständnis für, sondern auch eine Meisterschaft über diese komplexen Werkzeuge der künstlichen Intelligenz erreicht haben.

1.2 Zielgruppe

Dieses Buch richtet sich an eine breite Leserschaft, von Softwareentwicklern und Datenwissenschaftlerinnen bis hin zu Studierenden und Forschenden, die sich für generative künstliche Intelligenz interessieren. Ein gewisses Maß an Programmierkenntnissen wird vorausgesetzt. Wenn Sie ein grundlegendes Verständnis von Python haben und ein großes Interesse an KI, insbesondere im Bereich der generativen KI-Modelle, mitbringen, wird Ihnen dieses Buch gute Dienste leisten.

Der Inhalt ist so gestaltet, dass er sowohl Anfänger anspricht, die ihre ersten Schritte in die Welt der generativen KI machen, als auch erfahrene Profis, die ihre Fähigkeiten und ihr Wissen verfeinern möchten. Die praktischen Beispiele und ausführlichen Erklärungen helfen dabei, die Konzepte und Techniken zu verstehen, die für die Entwicklung und Anwendung von generativen KI-Systemen entscheidend sind. Egal, ob Sie in Ihrem Bereich innovativ sein möchten, ein akademisches Projekt starten wollen oder KI einfach faszinierend finden – dieses Buch soll eine wertvolle Ressource auf Ihrem Weg sein.

1.3 Was Sie schon wissen sollten

Bevor wir in die Welt der generativen KI mit Python eintauchen, gibt es ein paar Voraussetzungen, die Sie beachten sollten, um eine reibungslose Reise zu gewährleisten. Ich habe sie in verpflichtende und optionale Voraussetzungen unterteilt.

1.3.1 Verpflichtende Voraussetzungen

Zuerst ist es wichtig, dass Sie ein gutes Verständnis von Python-Programmierung haben. Sie sollten sich wohlfühlen mit

- der Erstellung von Funktionen,
- der Arbeit mit verschiedenen Datenstrukturen wie Listen oder Dictionarys,
- der Manipulation dieser Strukturen,
- Ihrer Fähigkeit, Schleifen zu schreiben, hauptsächlich for-Schleifen,
- und der Nutzung von Bibliotheken wie `numpy` und `pandas`.

1.3.2 Optionale Voraussetzungen

Idealerweise haben Sie bereits ein grundlegendes Verständnis von Konzepten des maschinellen Lernens, zum Beispiel vom Trainieren von Modellen und dem Arbeiten mit Datensätzen.

Vertrautheit mit grundlegender Statistik und linearer Algebra wird ebenfalls von Vorteil sein, da sie viele KI-Algorithmen untermauern. Obwohl das Buch die notwendigen Theorien hinter generativer KI behandelt, werden Ihnen Erfahrungen mit neuronalen Netzen und Deep-Learning-Frameworks wie TensorFlow oder PyTorch helfen, die fortgeschritteneren Themen leichter zu meistern.

Wenn diese Voraussetzungen wie Sprachen klingen, die Sie sprechen, sind Sie bestens gerüstet, um diese aufregende Reise durch die generative KI anzutreten.

1.4 Struktur des Buches

Das Buch ist als praktischer Leitfaden für Python-Programmierer und -Programmiererinnen gedacht, die generative KI-Anwendungen entwickeln möchten. Die Struktur des Buches folgt einem schrittweisen Ansatz, beginnend mit einer Einführung in die grundlegenden Konzepte bis hin zu fortgeschritteneren Themen wie Vektordatenbanken und agentischen Systemen.

Ich ermutige Sie dazu, das Buch der Reihe nach durchzugehen, denn einige Kapitel bauen auf dem Wissen aus vorherigen Kapiteln auf – dazu folgt im nächsten Abschnitt noch eine Übersicht.

Praktische Python-Codebeispiele stehen zum Download bereit, um die Anwendung der besprochenen Konzepte zu veranschaulichen und zu festigen.

Einführung in die Generative KI

Dieses Kapitel führt Sie in die Grundlagen der generativen KI ein, einen Teilbereich der künstlichen Intelligenz, der sich mit der Erstellung neuer Inhalte beschäftigt.

Sie werden etwas über Modelle der natürlichen Sprachverarbeitung (NLP) lernen, insbesondere über große Sprachmodelle (LLMs). Auch aktuelle Entwicklungen wie große multimodale Modelle (LMMs) oder Denkmodelle werden kurz angesprochen.

Vortrainierte Modelle

In diesem Kapitel werden vortrainierte Modelle vorgestellt, insbesondere aus dem Bereich der natürlichen Sprachverarbeitung. Vortrainierte Modelle sind Sprachmodelle, die auf großen Datenmengen trainiert wurden und für verschiedene NLP- oder Computer-Vision-Aufgaben wiederverwendet werden können. Wir werden die beliebteste Plattform Huggingface besprechen, auf der Sie über eine Million Open-Source-Modelle für unterschiedlichste Aufgaben finden.

Die vorgestellten und diskutierten Modelle zielen hauptsächlich darauf ab, eine bestimmte Aufgabe zu lösen, wie zum Beispiel Textzusammenfassungen, Übersetzungen, Textklassifizierung oder Textgenerierung. Der Vorteil dieses Ansatzes ist, dass diese Modelle klein sind und lokal betrieben werden können. In diesem Abschnitt lernen Sie, wie Sie ein Modell auswählen und es auf dem eigenen System betreiben.

Large-Language Models

In diesem Abschnitt werden Sie lernen, wie Sie über Python-Code mit LLMs interagieren. Ich stelle Anbieter wie OpenAI oder Groq vor und erkläre, wie Sie deren Modelle für eigene Projekte nutzen können.

Es gibt verschiedene Möglichkeiten, diese Modelle zu verwenden. Hier lernen Sie, wie Sie mit LangChain arbeiten, einem Python-basierten Framework für die Interaktion mit LLMs. So implementieren Sie effiziente Ansätze – einschließlich Prompt-Vorlagen und vor allem Chains – um die LLMs zu den gewünschten Ausgaben zu leiten.

Prompt Engineering

Prompt Engineering ist die Gestaltung effizienter Eingaben für LLMs, um Ausgaben zu generieren. Sie werden lernen, wie man effektive Prompts erstellt, um die Leistung von Sprachmodellen zu optimieren und genauere Ergebnisse zu erzielen. Grundlegende Techniken wie Few-Shot-Prompting und Chain-of-Thought werden behandelt. Aber wir gehen auch darüber hinaus und schauen uns fortgeschrittenere Techniken wie Self-Feedback oder Reflexion an.

Vektordatenbanken

Dieser Abschnitt beschäftigt sich mit Vektordatenbanken, die entscheidend sind, um große Textsammlungen effizient zu speichern und abzufragen. Er beschreibt den gesamten Prozess von der Indexierungspipeline, die Daten vorverarbeitet, um sie schließlich in einer Vektordatenbank zu speichern, bis hin zur Datenspeicherung und Abfrage.

Wir werden uns hauptsächlich darauf konzentrieren, Daten in eine Vektordatenbank hinzuzufügen. Dieser Prozess besteht aus mehreren aufeinanderfolgenden Schritten, die wir nacheinander behandeln werden.

Es beginnt mit dem Laden von Dokumenten. Dafür stehen verschiedene LangChain-Dokumenten-Loader zur Verfügung, um aus nahezu jeder Art von Datenquelle strukturiert zu laden. Wir werden einige davon kennenlernen.

Bei der Datenaufteilung zeige ich, wie man Dokumente in sinnvolle, handliche Stücke strukturiert. Es gibt verschiedene Methoden, die je nach Art der Daten und dem Anwendungsfall verwendet werden können. Ich werde die gängigsten Ansätze zur Datenaufteilung vorstellen.

Die Textinformationen müssen in ein Format umgewandelt werden, das ein Computeralgorithmus »verstehen« kann. Da Computer mit Zahlen und nicht mit menschlicher Sprache arbeiten, werden die Informationen in eine numerische Darstellung umgewandelt, die als Embedding bezeichnet wird. Hier werden Sie lernen, was Einbettungen (Embeddings) sind, wie sie erstellt werden können und welche verschiedenen Typen es gibt.

Es existieren verschiedene Anbieter von Vektordatenbanken, die Sie nutzen können. Wir werden mit ChromaDB als Beispiel für eine Open-Source-Vektordatenbank arbeiten, die auf einem lokalen System laufen kann. Aber auch der webbasierte Anbieter Pinecone wird thematisiert.

Sobald die Daten in einer Vektordatenbank gespeichert sind, möchten wir die Datenbank nutzen, um Informationen abzurufen. Typischerweise haben Nutzer eine Abfrage und möchten die relevantesten Informationen, zum Beispiel ein Textdokument oder ein Bild, aus dem Speicher abrufen. Der Abschnitt zeigt, dass nicht nur Texte abgerufen werden können, sondern auch Bilder oder Audiodateien. Sie werden lernen, Daten basierend auf zusätzlichen Bedingungen, beispielsweise speziellen Eigenschaften, abzurufen. Diese sind typischerweise in Metadaten definiert.

Beim Finden relevanter Dokumente sind zwei Fragen zentral: Wie messe ich Relevanz und wie finde ich diese Dokumente? Diese Fragen behandeln wir im Abschnitt über Ähnlichkeitssuche. Einige der behandelten Konzepte sind die Kosinusähnlichkeit und die Maximum Margin Relevance.

In einem Abschlussprojekt bringen wir alle Teile in einem Projekt zusammen – beginnend mit einer spezifischen Datenquelle, der Vorverarbeitung der Daten und schließlich der Speicherung in einer Vektordatenbank.

Retrieval-Augmented Generation

Retrieval-Augmented Generation ist ein neues Paradigma, bei dem externe Wissensquellen, typischerweise aus einer Vektordatenbank, in den Generierungsprozess integriert werden. Dadurch wird die Qualität und Relevanz des erzeugten Textes verbessert. Das ist eines der relevantesten Features von generativer KI in einem Unternehmensumfeld. Sie werden in diesem Abschnitt herausfinden, wie mächtig diese Technologie ist. Der Abschnitt beleuchtet, wie man relevante Informationen aus einem großen Datenspeicher abruft, wie man diese Informationen in den Generierungsprozess integriert und wie man die Qualität und Vielfalt des erzeugten Textes bewertet.

Agentensysteme

Dieser Abschnitt führt in agentenbasierte beziehungweise agentische Systeme ein. Das sind KI-Systeme, die auf großen Sprachmodellen basieren und mit zusätzlichen Funktionen wie Werkzeugnutzung oder Gedächtnis ausgestattet sind, sodass diese Systeme in der Lage sind, Aufgaben autonom auszuführen und ihre Aktionen auf bestimmte Ziele auszurichten.

Das ist einer der spannendsten Bereiche in der KI-Entwicklung. Es gibt viele verschiedene Frameworks, und wir werden mit den relevantesten davon arbeiten: LangGraph, OpenAI Swarm, Microsoft AG2, Magentic One, TinyTroupe und Pydantic AI.

Deployment

Bis zu diesem Punkt haben wir Systeme lokal auf unseren Computern entwickelt, aber in diesem Abschnitt zeige ich, wie man generative KI-Modelle und -Systeme praktisch bereitstellt. Der Bereitstellungsprozess umfasst normalerweise die Verwendung von REST-APIs, um die Systeme für Benutzer und andere Systeme zugänglich zu machen, die diese Funktionen nutzen möchten.

Wir werden verschiedene Bereitstellungsoptionen und Designentscheidungen erkunden, wie zum Beispiel eine eigenständige Anwendung im Vergleich zu einer Frontend-Backend-Architektur.

Sie werden lernen, wie man einen Backend-REST-API-Dienst erstellt und auch, wie eine Anwendung bei verschiedenen Anbietern wie Heroku, streamlit.io oder render bereitgestellt wird.

Ausblick

Das Buch endet mit einem Blick auf zukünftige Entwicklungen in der generativen KI und den damit verbundenen Herausforderungen – wie zum Beispiel den Auswirkungen auf unser Berufsleben und andere gesellschaftliche Probleme – aber auch auf die Chancen, die sich bieten.