Fachkurs

Retrieval Augmented Generation

Nutzen Sie künstliche Intelligenz für die Wissens- und Datenintegration im Unternehmen mit Retrieval Augmented Generation (RAG).

RAG kombiniert bestehende Datenbank- und Informationssysteme mit Large Language Modellen (LLMs), um präzise und kontextbezogene Informationen zu liefern. Beispiele sind Kunden- und Mitarbeitenden-Dialoge mit Chatbots oder automatisierte Analysen und Berichte.

Berner
Fachhochschule

Portrait

Retrieval Augmented Generation (RAG) ist eine populäre und stark wachsende Technologie im Bereich der künstlichen Intelligenz. RAG kombiniert die Fähigkeiten grosser Sprachmodelle (LLMs) mit externen Datenquellen (bspw. Informationssysteme im Unternehmen), um präzisere und spezifischere Antworten zu generieren. RAG ermöglicht es, die Leistung von LLMs durch den Zugriff auf Kontextinformationen zu verbessern, ohne dass ein langes und teures Training von LLMs erforderlich ist.

In Dienstleistung, Forschung, Gesundheitswesen, Compliance und Finanzwesen kann RAG beispielsweise genutzt werden, um grosse Mengen von Daten zu durchsuchen und daraus Analysen, Berichte, Empfehlungen, usw. zu erstellen. Im Bereich Kundendienstleistungen können Chatbots via RAG auf Produktkataloge, Support-Dokumentationen und Kundenportfolio zugreifen, um Anfragen zu beantworten und interaktive Dialoge zu führen.

Dieser Kurs ermöglicht Ihnen, den Nutzen von RAG für Ihr Unternehmen einzuschätzen und anhand von ersten Beispielen eine konkrete technologische Umsetzung zu realisieren.

Zielpublikum

Dieser Kurs besteht aus zwei Teilen, die auch einzeln besucht werden können.

- Der erste Teil richtet sich an Interessierte und Fachspezialist*innen über alle Industriezweige, Projektverantwortliche, Business Developer und Entwickler*innen.

- Der zweite Teil richtet sich zusätzlich an Entwickler*innen, Data Scientists und Data Engineers.

Ausbildungsziele

In diesem Fachkurs lernen Sie in einem ersten Teil die Bedeutung und Einsatzmöglichkeiten von RAG im gesamten Umfeld von KI, Large Language Models und Chatbots kennen – mit verschiedenen Anwendungsbeispielen aus Industrie und Dienstleistung. Der zweite Teil vermittelt ihnen Umsetzungskompetenz mit aktuellen Tools der AI-Community.

Voraussetzungen

Für Teil 2 benötigen Sie Kenntnisse in maschinellem Lernen und Python. Bitte registrieren Sie ein Konto bei OpenAI (für ca. 25 USD) für die praktischen Teile des Kurses.

Steckbrief

Fachkurs	RAG-Anwendungen ⌙ Einführung und Entwicklung
Dauer	5 Tage
Unterrichtstage	Teil 1: Mo 31.3.25 / Di 1.4.25 Teil 2: Mo 7.4.25 / Di 8.4 .25 / Mi 9.4.25
Anzahl ECTS	4 ECTS-Credits
Unterrichtssprache	Deutsch, Unterlagen teilweise in Englisch

Durchführungsart	Hybrid (Vor Ort, Remote-Teilnahme möglich)
Kosten	Ganzer Kurs CHF 2'800, Teil 1: CHF 1'500, Teil 2: CHF 2'300
Anmeldeschluss	24. März 2025

Kursprogramm

Aufbau

Teil 1, 2 Tage: Technologie kennenlernen, Nutzen und Einsatzbereich von RAG einschätzen können RAG im Umfeld von KI, Large Language Modellen, Transformern und Chatbots. Aktueller Stand der KI, wichtige Begrifflichkeiten, Einsatzbereiche und Anwendungsbeispiele.

Teil 2, 3 Tage: Einstieg in die Anwendungsentwicklung, Kennenlernen von Tools und APIs Praktische Einführung in Retrieval Augmented Generation und Entwicklung von Anwendungen mit Transformern, Architektur von RAG und Transformer-Systemen, Anbieter von API und Tools.

Inhalt

Teil 1
KI, Machine Learning, Generative KI, RAG, Large Language Modelle, Transformer und Chatbots. Aktueller Stand der KI, wichtige Begrifflichkeiten, Einsatzbereiche und Anwendungsbeispiele.

Tag 1: Einführung

1. Entwicklung der KI
2. Abgrenzung: KI vs. maschinelles Lernen, mathematische Optimierungstechniken und Statistik: Unterschiede und Überschneidungen
3. Vollständige Vernetzung als das Hauptunterscheidungsmerkmal von KI
4. Der Aufstieg der Generativen KI
5. Deskriptive vs. Generative KI
6. Autonome KI-Agents
7. Arten der Generativen KI: von Chatbots bis zu Musikgenerierung
8. Technologiewahl
   - Ist KI immer die beste Wahl?
   - KI-Ressourcen und -Portale
9. Programmiersprachen und Frameworks
   - Organisation und Durchführung der KI-Projekte
   - Von der Idee zur Umsetzung: Design-Thinking-Workshops
   - Wo finde ich die KI-Spezialist*innen?
   - Mit welchen Software-Kosten muss ich rechnen?
   - Welche Infrastruktur wird benötigt und was wird sie kosten?
   - Projekt-Team: welche Rollen und Skills brauche ich in KI-Projekten?
   - Qualitätssicherung

Tag 2: Anwendungs-Beispiele aus unterschiedlichen Industrien

1. Lagerhaltung- und Nachfrage Planung im Gross- und Einzelhandel – KI für Demand-Forecasting
2. KI in Medizin und Labordiagnostik: Blutkrebs-Erkennung
3. KI für Projekt-Planung: Forecasting von Projekt-Verspätungen und Budget-Overruns
4. Volkswirtschaftliche Planung: Bestimmung der essenziellen Rohstoffe und ihrer Zukunfts-Preise
5. Autonomes Fahren, Robotik
6. Abschluss KI-Projekt zu einem Wahl-Thema
   - Rollen Verteilung
   - Workshop zur Lösungsfindung und Umsetzungs-Planung
   - Festlegung der Technologie und Ressourcen
   - Mitarbeiter-Suche und Interviews
   - Abschliessende Präsentation der Idee und deren Umsetzung

Teil 2
Praktische Einführung in RAG (Retrieval-Augmented Generation) und die Entwicklung von Anwendungen mit Transformern. Für diesen Kursteil benötigen Sie Grundkenntnisse in maschinellem Lernen und Python. Bitte registrieren Sie ein Konto bei OpenAI mit minimalem Betrag (für ca. 25 USD).

Tag 1: Transformer

1. Was sind Transformer?
2. Was ist RAG und warum wird es benötigt?
   - Warum ist ein Abruf von relevanten Informationen aus einer Datenbank eine bessere Lösung als grosser mitgegebener Prompt-Kontext?
   - Warum verringert RAG die sogenannten «Halluzinationen»?
   - Bereitstellung relevanter und aktueller Informationen in der RAG-Pipeline
   - Nutzung von unternehmenseigenen Inhalten/Wissensbasis
3. Transformer Anwendungen
   - Text Vervollständigung, Code Vervollständigung
   - Image Generation
   - Einbettungen («Embeddings»)
4. Einführung in die Transformer-Architektur
   - Das Encoder-Decoder Framework
   - Attention Mechanismus
   - Hugging Face Transformer
5. Tour durch die Transformer-Anwendungen
   - Text Classification, Named Entity Recognition
   - Question Answering
   - Summarization
   - Translation
   - Text Generation
6. Hugging Face: Hub, Models, Tokenizers, Datasets
7. Praxis: Aufbau einer vollständigen Transformer-Anwendung in Python (Teamarbeit)

Tag 2: RAG

1. Beispiel: «Verwandeln von PDF-Dokumenten in eine RAG-Applikation»
2. Diskussion der Problem-Stellung: Umwandlung von unstrukturierten Daten mit wertvollen Informationen zum Bestandteil der RAG-Pipeline
3. Zieldefinition: RAG-Pipeline mit Benutzeranfragen auf der Grundlage der in PDF-Dokumenten enthaltenen Informationen
4. Prozess
   - Text-Extrahierung
   - Konvertierung der Dokumente in Bilder
   - Analyse der Seiten mit GPT
   - Generierung von Text-Embedding-Tabellen
   - Erstellen von RAG, Prompt-Engineering für RAG

Tag 3: Begleitete Erstellung einer RAG-Applikation

1. Umsetzung der Prozess-Schritte aus Tag 2
2. Test der Anwendung an einem Live OpenAI Terminal
3. Präsentation der Anwendung
4. Übersicht über weitere RAG-Architekturen: RAG und Graph-Datenbank Neo4J
5. Q&A und Abschluss

Organisation

Kursleitung:
Arno Schmidhauser
E-Mail: arno.schmidhauser@bfh.ch

Kursadministration:
Andrea Moser
E-Mail: andrea.moser@bfh.ch

Berner Fachhochschule
Weiterbildung
Aarbergstrasse 46 (Switzerland Innovation Park Biel/Bienne)
2503 Biel

Telefon +41 31 848 31 11
E-Mail: weiterbildung.ti@bfh.ch
Web: bfh.ch/ti/weiterbildung