Retrieval-Augmented Generation

Duração do curso

Este curso tem uma duração de 5 dias, num total de 40 horas.

Descrição do Curso

Este Curso tem como objetivo dotar os participantes dos conhecimentos e técnicas necessários para a criação de aplicações especializadas de inteligência artificial generativa, quer seja para dotar os colaboradores com o know-how coletivo da empresa ou para automatizar o atendimento ao público.

A quem se destina

Programadores que queiram aprender a criar aplicações de inteligência artificial generativa.

Pré-requisitos

Conhecimentos das linguagens de programação python e SQL. Subscrição API da OpenAI.

Objectivos do curso

- Compreender e aplicar os conceitos fundamentais da engenharia de prompts.
- Criação de aplicações utilizando base de dados de vectores com recurso ao postgreSQL.
- Criação de aplicações RAG prontos para ambientes de produção.
- Criação de agentes RAG.

Tópicos do curso

Módulo 1: Introdução ao RAG

- O que são Large Language Models (LLMs)
- Large Language Models vs Small Language Models
- Prompts e Alucinações
- Visão geral do Retrieval Augmentation Generation: Definição, Importância e Casos de Uso
- RAG versus Finetuning versus Retraining

Módulo 2: Engenharia de prompt com recurso ao chatGPT

- Fundamentos da engenharia de prompt
- Padrões de solicitação
- Chain-of-thought prompting
- ReAct prompting

Módulo 3: Criação de aplicações inteligentes com recurso a base de dados de vectores

- O que são base de dados de vetores
- Utilização do postgreSQL como base de dados de vetores
- Representação de dados com vetores
- Algoritmos e medidas de similaridade entre vetores
- Pesquisas Sparse, Dense e Híbridas
- Criação de um sistema de pesquisa semântica de documentos utilizando postgreSQL
- Criação de um sistema de pesquisa semântica de imagens utilizando postgreSQL

Módulo 4: Criação de aplicações inteligentes com RAG

- Arquitetura e componentes de aplicações RAG
- Personalização de LLMs com postgreSQL
- Introdução e arquitetura do Llamaindex
- Criação de um chatbot “expert” em documentação proprietária

Módulo 5: Criação de aplicações RAG seguros e preparados para ambientes de produção

- Construção de pipelines RAG avançados com Llamaindex
- Avaliação e métricas RAG
- Sentence-window
- Auto-merging retrieval

Módulo 6: Criação de agentes RAG

- Fundamentos dos agentes RAG
- Encaminhamento de queries
- Selecção de ferramentas
- Construção de ciclos de raciocínio do agente
- Criação de um agente de pesquisa documental com Llamaindex e postgreSQL