Workshop

RAG: Unveiling the Power of Retrieval-Augmented Generation

Célien Donzé[1], Jonathan Guerne[1], Henrique Marques Reis[1] and Pedro Costa[2]

HE-Arc Neuchâtel[1], Lumind[2]

Description

The Retrieval-Augmented Generation (RAG) framework represents a groundbreaking approach that seamlessly integrates two fundamental techniques, retrieval and generation, within a large language model (LLM). The result is the generation of more context-aware and informative responses, making RAG a valuable tool for companies with extensive documentation but lacking an efficient means to access specific information. This workshop endeavors to provide a comprehensive understanding of the RAG technology, emphasizing its applications and advantages. Through a technical introduction accompanied by concrete examples, participants will gain insights into how RAG can be effectively employed to address challenges related to information retrieval and contextual generation. The workshop will also facilitate discussions on the practical implementation of RAG in real-world scenarios, exploring its potential in enhancing knowledge management systems. Furthermore, the workshop will delve into the realm of self-hosted Large Language Models (LLMs), shedding light on the importance of data privacy and security in the deployment of generative AI technologies. Participants will be equipped with knowledge about the intricacies of hosting LLM models independently. By the conclusion of the workshop, participants will possess the skills to proficiently interact with a LLM, querying it about the contents of its associated documents. The overarching goal is to empower individuals with the expertise needed to harness the full potential of RAG and self-hosted LLMs, fostering a practical and informed approach towards the integration of these technologies in diverse real-world applications.

Schedule

Technical background, duration: 30 min

RAG overview

Hands on, duration: 1h

In Google collab (only a laptop is needed)

Applied examples, duration: 30min

Showcase real world implementation of RAG applications

Time for discussion, duration: 30min

References

Ollama Available: https://github.com/ollama/ollama
LangChain Available: https://github.com/langchain-ai/langchain