BASF Hiring Challenge: AI DevOps Engineer

Task 1 - Containers, AI and communicating services

Your task is to deploy a tiny LLM as an API in a containerized solution such as Docker-Compose or Kubernetes.

Hints:

- You should provide us with access to a private GitHub repository that has a proper structure and is well-documented.
- You should write a Dockerfile, a requirements.txt file, a Python script to load the model into a web server capable of running it, and, with your selected deployment solution, either a Kubernetes deployment, service, and ingress configuration or a docker-compose.yml file.
- You can find tiny models on Huggingface.com such as Phi-3, TinyLlama… these are only 2 examples:
    - https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v0.4
    - https://huggingface.co/microsoft/Phi-3-mini-4k-instruct
- Bonus: Choose a production-ready web server for deploying your model inside a separate container, and provide a user-friendly way to interact with the selected LLM model.
- Bonus: Provide a solution working in Kubernetes and Docker-Compose.
- Bonus: Provide a diagram to explain your solution.

Task 2 - Migration Mentorship for AI Developers

Design a mentorship plan for a team of developers who have completed their applications in a Proof of Concept (PoC) phase on their local machines. These applications are written in Python and integrate a language model (LLM) without being containerized. The team needs your expertise to transition these applications to a production-ready environment using best practices in DevOps and cloud architecture.

Version: 24th June 2024