RAG + Embedding with AnythingLLM and Ollama

AnythingLLM - is an all-in-one AI application that simplifies the interaction with Large Language Models (LLMs) for business intelligence purposes. It allows users to chat with any document, such as PDFs or Word files, using various LLMs, including enterprise models like GPT-4 or open-source models like Llama and Mistral.

Ollama - we introduce this in the last blog. We will use its ability to service multiple Open Source LLMs via its API interface.

Here is a quick outline:

- Details about docker containers for Ollama (Platform/Server) and AnythingLLM (front end/chat/uploading documents).
- Explore:
  - Embedding a news article about recent US political events.
  - Vector database (Lance DB) - you can spin up chroma if you like, but Lance DB comes bundled with AnythingLLM.
  - Query the LLM about the news article and assess how well it did.

A lot of the above is built into AnythingLLM.

Components used

- Ollama Server - a platform that make easier to run LLM locally on your compute.
- Open WebUI - a self-hosted front end that interacts with APIs that presented by Ollama or OpenAI compatible platforms. I am using to download new LLMs much easier to manage than connecting to the ollama docker container and issuing 'ollama pull'.
- AnythingLLM - an all-in-one AI application that simplifies the interaction with Large Language Models (LLMs).
- Linux Server or equivalent device - spin up three docker containers with the Docker-compose YAML file specified below.

Code break down

Analysis of Docker-Compose.yml file

```
services:
  ollama-server:
    image: ollama/ollama:latest
    container_name: ollama-server
```

```
    ports:
      - "11434:11434"
    volumes:
      - ./ollama_data:/root/.ollama
    restart: unless-stopped

  ollama-webui:
    image: ghcr.io/ollama-webui/ollama-webui:main
    container_name: ollama-webui
    restart: unless-stopped
    environment:
      - 'OLLAMA_BASE_URL=http://ollama-server:11434'
    volumes:
      - ./webui:/app/backend/data
    ports:
      - "3010:8080"
    extra_hosts:
      - host.docker.internal:host-gateway

  anything-LLM:
    image: mintplexlabs/anythingllm:latest
    container_name: anything-llm
    cap_add:
      - SYS_ADMIN
    restart: unless-stopped
    environment:
      - SERVER_PORT=3001
      - UID='1000'
      - GID='1000'
      - STORAGE_DIR=/app/server/storage
      - LLM_PROVIDER=ollama
      - OLLAMA_BASE_PATH=http://ollama-server:11434
      - OLLAMA_MODEL_PREF='phi3'
      - OLLAMA_MODEL_TOKEN_LIMIT=4096
      - EMBEDDING_ENGINE=ollama
      - EMBEDDING_BASE_PATH=http://ollama-server:11434
      - EMBEDDING_MODEL_PREF=nomic-embed-text:latest
      - EMBEDDING_MODEL_MAX_CHUNK_LENGTH=8192
      - VECTOR_DB=lancedb
      - WHISPER_PROVIDER=local
```

```
      - TTS_PROVIDER=native
      - PASSWORDMINCHAR=8
    volumes:
      - ./anythingllm_data/storage:/app/server/storage
      - ./anythingllm_data/collector/hotdir/:/app/collector/hotdir
      - ./anythingllm_data/collector/outputs/:/app/collector/outputs
    ports:
      - "3001:3001"
    extra_hosts:
      - host.docker.internal:host-gateway
```

Line 6 - Ollama Server exposes port 11434 for its API.

Line 8 - maps a folder on the host ollama_data to the directory inside the container /root/.ollama - this is where all LLM are downloaded to.

Line 16 - environment variable that tells Web UI which port to connect to on the Ollama Server. Since both docker containers are sitting on the same host we can refer to the ollama container name 'ollama-server' in the URL.

Line 18 - maps a folder on the host webui to the directory inside the container /app/backend/data - storing configs.

Line 20 - Connect to the Web UI on port 3010.

Line 21-22 - Avoids the need for this container to use 'host' network mode.

Line 30 - Environmental variable that are used by AnythingLLM - more can be found at ENV variables Note the Base_Path to ollama refers to the ollama container listed above in the docker compose file.

Line 47 - AnythingLLM uses a lot of volume mapping. They may make changes to this later, the last two collector was a recent addition, so it will depend on the version of the docker image that gets pulled. Since it is set to 'latest'

My directory structure in the folder where docker compose exist. Create these folders before starting the 'docker compose' commands.

Issue 'docker compose up -d' from the folder where your docker compose YAML file sits, to install and start the containers. Once the containers are up, you can browse to the AnythingLLM on port 3001 - example http://x.x.x.x:3001

Ins0mniA

Revision #5
Created 2025-08-02 02:22:37 EEST by Green
Updated 2025-09-10 19:44:49 EEST by Green