ragtime

Release 1

jan jansen

Oct 18, 2024

CONTENTS:

CHAPTER

ONE

INTRO

These are some building blocks in what should become a local LLM for financial info. I do like investing, but I rather not read all the publications.

This software should be able to :

- leech reports
- embed them into a vectordb
- use a local LLM to retrieve and bundle information
- generate a report

1.1 Homelab: (did not want to create separate repo)

I have an old HP dl380p gen8, which I modified: - removed a SAS controller card, and have disks on internal controller - add a 1TB NVME M2 disk, on PCI-e 3 adapter - installed Proxmox

todo: - boot on sata disk on cdrom port - remove RAID1 and use disks as such: space over security - insert NVIDIA P4 single slot

CHAPTER

TWO

INSTALL SOMETHING FROM DOCKER

https://docs.docker.com/engine/install/ubuntu/ https://download.docker.com/linux/ubuntu/dists/jammy/pool/stable/amd64/

2.1 using portainer:

sudo docker volume create portainer_data sudo docker run -d -p 8000:8000 -p 9443:9443 –name portainer –restart=always -v /var/run/docker.sock:/var/run/docker.sock -v portainer_data:/data portainer/portainer-ce:2.21.3

2.2 leecher:

custommade : pdf copy to directory leech

2.3 building pdf to text:

docker build -t pdf-text-converter . docker run -v $(pwd)/leech:/app/leech -v $(pwd)/pages:/app/pages pdf-text-converter

2.4 pdfconverter2

docker build -t pdf-watcher . docker run -v ($pwd)/leech:/leech -v ($pwd)/pages:/pages pdf-watcher

CHAPTER

THREE

STORING VECTOR + METADATA

PostgreSQL + pgvector Example If you prefer using PostgreSQL, you can use the pgvector extension, which allows you to store vector embeddings in a PostgreSQL table alongside metadata.

Steps:

- Install pgvector: First, install the pgvector extension.
- Create a table: Create a table that stores both vectors and metadata.
- Insert vectors and metadata: Insert each chunk's vector along with its metadata.

sql

CHAPTER

FOUR

A POOR MAN'S RIG

I theory it should be possible to use cheap hardware.

I have an old server (proliant gen8), but it has 256GB of RAM. I theory I can store a big LLM on RAMDISK and use this trick:

https://huggingface.co/blog/lyogavin/run-llama-405b-on-4gb-vram