Local LLMs

LLM Utility

I'm a big fan of Simon Willison's llm package. It works nicely with llama-cpp.

Installing `llm`

I didn't get on well with pipx in this use case so I used conda to create a virtual environments for LLM and then installed it in there.

Since I have an NVIDIA card I pass in CMAKE flags to have it build support for cuda:

```
conda create -y -n llm python=3.10
conda activate llm
pip install llm llm-llama-cpp
CMAKE_ARGS="-DLLAMA_CUBLAS=ON -DCMAKE_CUDA_COMPILER=/usr/local/cuda/bin/nvcc" FORCE_CMAKE=1
llm install llama-cpp-python

# alternatively if no NVIDIA support is available, this works well
# CMAKE_ARGS="-DLLAMA_OPENBLAS=on" FORCE_CMAKE=1 llm install llama-cpp-python
```

LangChain

LangChain is a FOSS library for chaining together prompt-able language models. I've been using it for building all sorts of cool stuff.

Revision #3
Created 1 October 2023 20:11:40 by James
Updated 21 January 2024 14:49:28 by James