PyLLMCore

PyLLMCore is a python library for working with a variety of LLM models and it supports both OpenAI and Local models.

Setup on Linux

Install the `llama-cpp-python` library first so that you can ensure that the nvidia dependencies are all pre-configured.

```
CMAKE_ARGS="-DLLAMA_CUBLAS=ON -DCMAKE_CUDA_COMPILER=/usr/local/cuda/bin/nvcc" pip install
llama-cpp-python
pip install py-llm-core
```

Put models in the correct location

The library seems quite fussy about model location. They must be in the `~/.cache/py-llm-core/models/` folder inside your user profile. Since I am already using SimonW's LLM (as described here) I symlink the zephyr model from there:

```
ln -s ~/.config/io.datasette.llm/llama-cpp/models/zephyr-7b-alpha.Q5_K_M.gguf\
 ~/.cache/py-llm-core/models/zephyr-7b-alpha.Q5_K_M.gguf
```

I realised I had done this wrong because I passed in a full filename to a model elsewhere and got an error like this:

```
Traceback (most recent call last):
  File "/home/james/workspace/raf/llmcore.py", line 24, in <module>
    book = parser.parse(text)
  File "/home/james/miniconda3/envs/raf/lib/python3.10/site-packages/llm_core/parsers.py",
line 20, in parse
    completion = self.model_wrapper.ask(
  File "/home/james/miniconda3/envs/raf/lib/python3.10/site-
packages/llm_core/llm/llama_cpp_compatible.py", line 65, in ask
    self.sanitize_prompt(prompt=prompt, history=history, schema=schema)
  File "/home/james/miniconda3/envs/raf/lib/python3.10/site-packages/llm_core/llm/base.py",
line 29, in sanitize_prompt
```

```
    required_ctx_size = len(codecs.encode(complete_prompt, self.name))
LookupError: unknown encoding: /home/james/.config/io.datasette.llm/llama-cpp/models/zephyr-
7b-alpha.Q5_K_M.gguf
```

Revision #4
Created 29 October 2023 14:31:59 by James
Updated 21 January 2024 14:49:29 by James