Local RAG using FAISS (GPU)

Requirements:

- Python=3.11.5 (will be installed through Anaconda)
- CUDA 11 or 12 (if using faiss-gpu)
- Anaconda (or miniconda3)
- OpenAI API-key
- LLamaCloud API-key

Acquire and setup API-keys:

- OpenAI: Follow this guide
- LLama Cloud: Get your key here and then setup for Python.

Build Instructions:

- Run the following command (in terminal): `conda create -n kulocalbot python=3.11.5`
    - `kulocalbot` is just a name for the environment, it can be changed.
- Activate the virtual environment using `conda activate kulocalbot`
- Install uv pip: `pip install uv`
- Install required libraries for the chatbot: `uv pip install -r requirements.txt`
- Lastly, install faiss-gpu (Linux Only)

  ```
  conda install -c pytorch -c nvidia faiss-gpu=1.8.0
  ```

How To Use:

- Place your PDF file(s) in a directory called `data` inside the project folder.
- Open a terminal in the project folder, i.e. `local-gpu-rag`.
- Run `python main.py`. This parses the documents and creates a FAISS-Index.
- If you're running the script again using the same document(s), you can instead run

  ```
  python main.py --load_existing
  ```

  This loads the already existing index stored in `faiss_index`, instead of parsing the same documents again.

Notes:

- More detailed instructions on how to install `faiss-gpu` on your device are available at their github page.
- This project was built and tested using Python 3.11.5 (Other versions may require different dependency versions).