Install Guide for GOR 2025 Workshop on Structured Information Extraction with LLMs

Paul Simmering

2025-03-25

Dear workshop participant,

Thank you for signing up to the GOR 2025 workshop on “Structured Information Extraction with LLMs”. I’m looking forward to meeting you in Berlin!

This guide will help you prepare for the workshop by setting up the required software. Please prepare your laptop before the workshop, so we can start right away. If you have any questions or run into issues, please contact me ahead of time at: paul.simmering@teamq.de.

The workshop will involve programming in Python. We will be using Python in a Jupyter Notebook environment. There are two ways to run it: 1) in the cloud using Google Colab, 2) locally on your laptop. If you have not worked with Python or Jupyter before, I recommend using Google Colab, as it is easier to set up.

Option 1: Google Colab

This option requires a Google account. Please create a free account if you do not have one yet.

Next, go to https://colab.research.google.com/drive/1lpQDRxTx3tnbChFqGUOW6vEF7CwZ3aw3 and log in by clicking on the “Sign in” button in the top right corner. Click on “File” in the top left and select “Open in playground mode”. Then connect the notebook to an NVIDIA T4 GPU by clicking on the “Connect T4” button in the top right corner. Now you’re ready to go. Hover over the code cells and click the “play” button to run them. The first cells will install the required Python packages and download a language model. These cells need to be run every time you open the notebook.

Option 2: Local installation

You can also run Python on your laptop. The requirements to run the workshop locally are:

- Python 3.10 or higher
- Jupyter
- A copy of the workshop notebook, available by downloading the Google Colab file. Click on the "File" menu, select "Download" and then choose "Download .ipynb".
- A Python environment with the following packages:
  - `instructor==1.7.2`
  - `polars==1.20.0`
  - `llama-cpp-python==0.3.4`, if using a local LLM (see below)

Open the workshop notebook in Jupyter or in a Jupyter-compatible IDE like VS Code or PyCharm.

There are two ways to run the an LLM: locally using llama-cpp-python or in the cloud.

Local LLM with llama-cpp-python

This option requires either a powerful CPU (for example a MacBook with M-series chip) or a GPU.

Follow the installation steps described in the llama-cpp-python readme. Use the "Metal" backend if you have a MacBook with M-series chip. The workshop notebook has runnable cells for an installation with CUDA 12.2 support.

Cloud LLM via an API

We are using the instructor for structured output. Set up instructor to connect to an LLM API. The example below shows how to connect to OpenAI:

```
import instructor
from openai import OpenAI
from functools import partial

client = instructor.from_openai(OpenAI(api_key="<your-api-key>"))
# Simplify the function to create a chat completion
create = partial(client.chat.completions.create, model="gpt-4o-mini")
```

You can place this chunk in your copy of the workshop notebook. The llama-cpp-python installation and related code is not needed in this case.

You can use your own API key from OpenAI or set up instructor to work with another provider that is compatible with instructor. This includes OpenAI, Anthropic, Google Cloud, and many others. See the instructor documentation for more details. If you don't have an API key, please contact me at: paul.simmering@teamq.de.