[DM2025] Lab 2 – Ollama Setup

Hi everyone,

This document provides detailed instructions for setting up the environment required for Lab 2's optional notebook using Ollama for LLM Open-Source Models' usage.

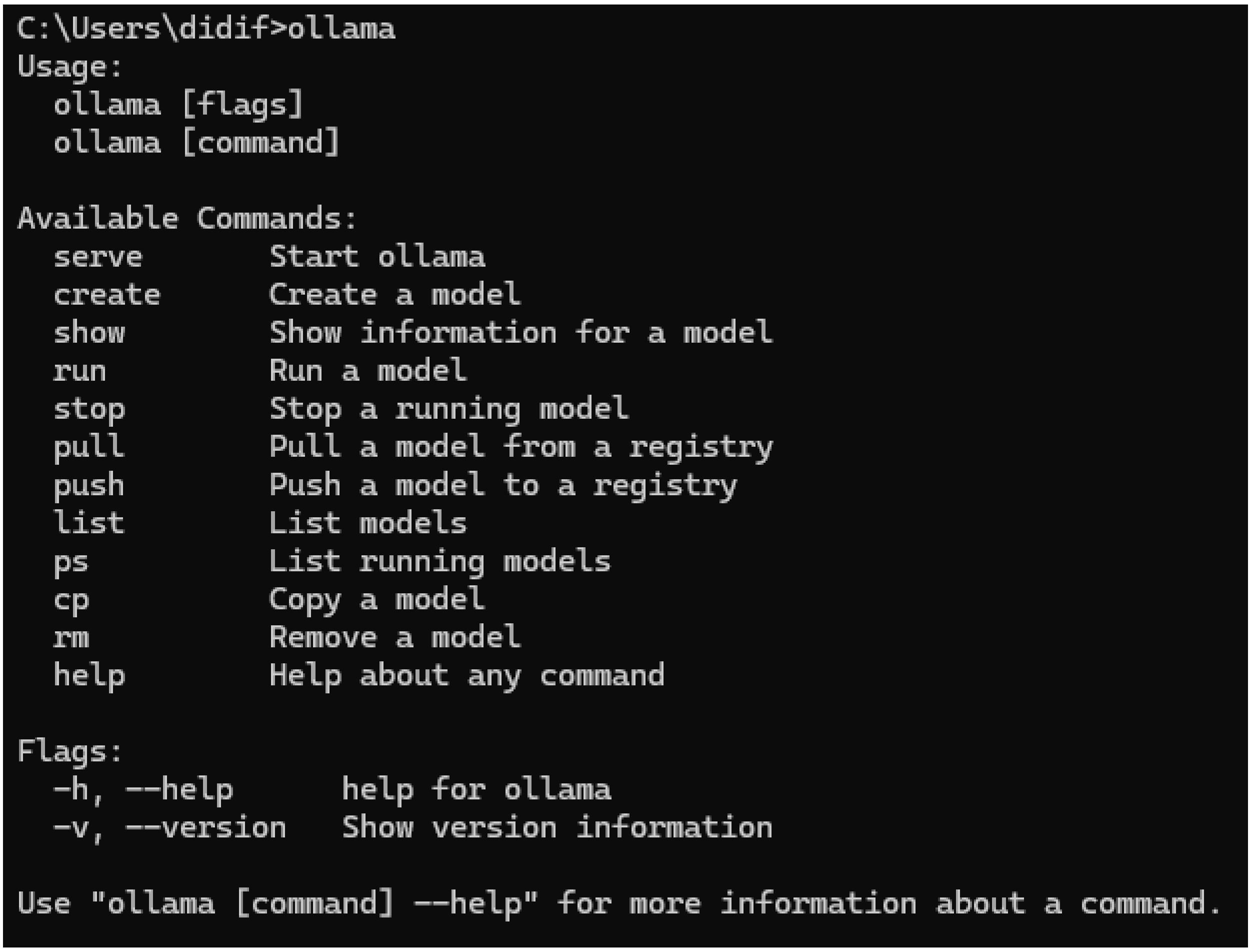
Using Ollama to use Open Source LLMs (Optional):

This part is `not worth any points`, it is just optional material if you want to learn how to use this technology.

Using Ollama locally:

We will be using some small open-source LLMs that will be running in your device with `Ollama`, please enter the website, download and install it in your device: Ollama website

After the installation is done, go to your terminal and type: **ollama** You should be getting the following information if the installation was correct:



```
C:\Users\didif>ollama
Usage:
  ollama [flags]
  ollama [command]

Available Commands:
  serve       Start ollama
  create      Create a model
  show        Show information for a model
  run         Run a model
  stop        Stop a running model
  pull        Pull a model from a registry
  push        Push a model to a registry
  list        List models
  ps          List running models
  cp          Copy a model
  rm          Remove a model
  help        Help about any command

Flags:
  -h, --help      help for ollama
  -v, --version   Show version information

Use "ollama [command] --help" for more information about a command.
```

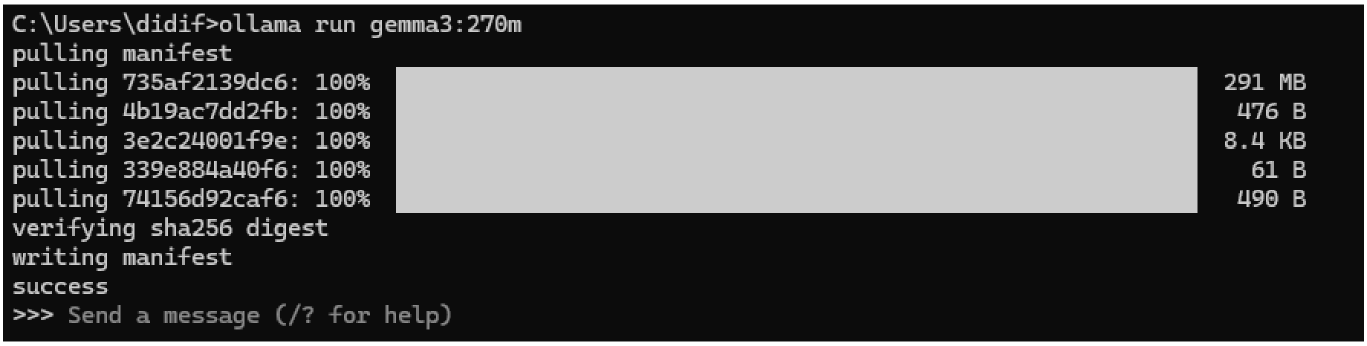
We will be using 4 Open-source LLMs here, for that it is recommended to have at least **4 GB of VRAM, 16 GB of RAM and multi-core processor** to run them in the **most optimal way**, although they can be run with less computing resources, but they will be slower in response. To download and install them you will need to type the following commands in the terminal:

- ollama run gemma3:4b
  - It is a version with **4 billion parameters**. We will be using this one for multi-modal prompting with images, and for some advanced text-based tasks.
- ollama run gemma3:270m
  - We will use this for simple text prompting, it is a small LLM model of **270 million parameters**.
- ollama run llama3.2:1b
  - We will use this for tool calling, it is a version of **1 billion parameters**.
- ollama run embeddinggemma
  - Model with **300 million parameters.** We will use this model to obtain text embeddings from our data.

Just for reference, GPT-4 from OpenAI has **1.8 trillion parameters**, so these are just very small models in comparison.

Feel free to explore the ollama library for other models, you can change their name in the code and re-run our provided material to compare the outputs.

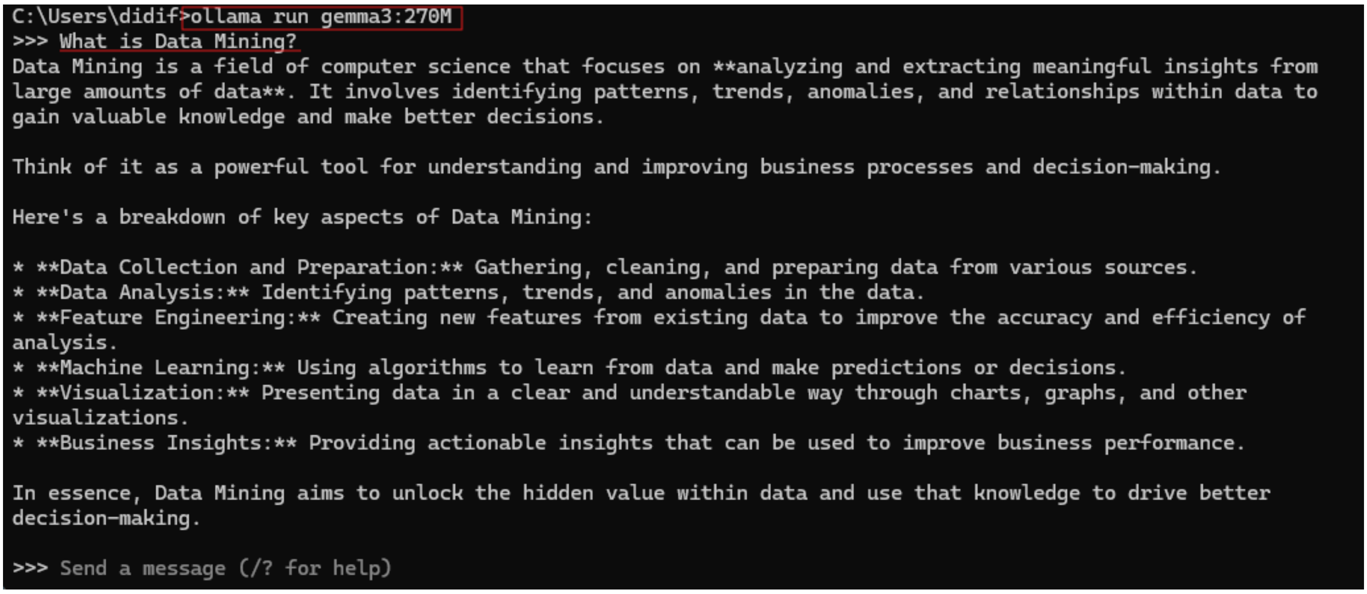
After you run one of the commands the model will start to download in this way:



```
C:\Users\didif>ollama run gemma3:270m
pulling manifest
pulling 735af2139dc6: 100%                                         291 MB
pulling 4b19ac7dd2fb: 100%                                         476 B
pulling 3e2c24001f9e: 100%                                         8.4 KB
pulling 339e884a40f6: 100%                                          61 B
pulling 74156d92caf6: 100%                                         490 B
verifying sha256 digest
writing manifest
success
>>> Send a message (/? for help)
```

So download and install all of the models **one by one.**

After finishing you can verify each model by asking something in a prompt in the terminal:



```
C:\Users\didif>ollama run gemma3:270M
>>> What is Data Mining?
Data Mining is a field of computer science that focuses on **analyzing and extracting meaningful insights from
large amounts of data**. It involves identifying patterns, trends, anomalies, and relationships within data to
gain valuable knowledge and make better decisions.

Think of it as a powerful tool for understanding and improving business processes and decision-making.

Here's a breakdown of key aspects of Data Mining:

* **Data Collection and Preparation:** Gathering, cleaning, and preparing data from various sources.
* **Data Analysis:** Identifying patterns, trends, and anomalies in the data.
* **Feature Engineering:** Creating new features from existing data to improve the accuracy and efficiency of
analysis.
* **Machine Learning:** Using algorithms to learn from data and make predictions or decisions.
* **Visualization:** Presenting data in a clear and understandable way through charts, graphs, and other
visualizations.
* **Business Insights:** Providing actionable insights that can be used to improve business performance.

In essence, Data Mining aims to unlock the hidden value within data and use that knowledge to drive better
decision-making.

>>> Send a message (/? for help)
```

Using Ollama on the cloud (Kaggle | Google Colab):

We also provide a tested set of notebooks to run this additional material in Kaggle and Colab:

Kaggle | LLM Master Examples with Ollama

Colab | LLM Master Examples with Ollama

`Enter the links, copy and run them accordingly.`

Note that for Kaggle because the notebook size is greater than 1 Mb we cannot save a version there, only a draft, so if needed just download after you finish trying the notebook out.

Run the notebooks with `GPU`, both Kaggle and Google Colab offer `T4 GPUs` to run notebooks. Kaggle offers 30 hrs of GPU usage every week, while Google Colab is dynamic, depends on the current available resources. Kaggle can offer more Quota if you link your account to Colab Pro.

After this you need to run the following on the top of your notebook:

```
# This command installs pciutils, necessary to locate the NVIDIA GPUs in the cloud container
!apt-get install pciutils -y -qq > /dev/null 2>&1
```

```
#Download ollama
!curl -fsSL https://ollama.com/install.sh | sh
import subprocess
process = subprocess.Popen("ollama serve", shell=True) #runs on a different thread
```

```
#Download model gemma3:4b
!ollama pull gemma3:4b
```

```
#Download model llama3.2:1b
!ollama pull llama3.2:1b
```

```
#Download model gemma3:270m
!ollama pull gemma3:270m
```

```
#Download model embeddinggemma
!ollama pull embeddinggemma
```

Now, you can try to run the code inside `DM2025-Lab2-Optional-Ollama.ipynb`

Note that it might take some time to download and load everything in the cloud, so expect to wait 5 to 10 mins.