README: Setting Up Marker and Ollama Environment on HPC

--------------------------------------------------------

This guide explains how to set up an environment on an HPC cluster to run Marker (OCR tool) and Ollama (for hosting large language models such as DeepSeek-R1). It includes steps for preparing the compute node, installing necessary Python tools, configuring the environment, and running Ollama.

Helpful Commands:

-----------------

- See all current jobs:

squeue -u $USER

- Cancel a specific job:

scancel <JOBID>

- Check current node:

hostname

1) Prepare Environment:

------------------------

a) Avoid Using the Login Node:

- Only perform lightweight tasks like editing files or submitting jobs on the login node.

- Do not run heavy programs (training/inference) on the login node. It may crash shared resources.

- To check if you are on the login node:

hostname

b) Start an Interactive Compute Node:

- Request GPUs and memory:

hpcshell --gpus=2 --mem=78G --account=AccountNameHere --time=2:00:00

Explanation of flags:

- --gpus=# : Number of GPUs requested.

- --mem=#G : Amount of memory in GB.

- --account=ACCNAME : Account under which the job runs.

- --time=#:##:## : Wall-clock time limit.

2) Change to Your Working Directory:

-------------------------------------

cd /your/working/directory

3) Set Up Ollama Models Directory:

-----------------------------------

- Create a models folder if needed:

mkdir models

- Export environment variable to tell Ollama where to store models:

export OLLAMA\_MODELS=$PWD/models

- Verify the environment variable:

echo $OLLAMA\_MODELS

Note: You must re-run the export command for every new session unless added to your .bashrc or .bash\_profile.

4) Load Required Modules:

--------------------------

- Load Spack environment:

. /opt/ohpc/pub/spack/v0.23.0/share/spack/setup-env.sh

- Load Python tools:

spack load py-torch py-virtualenv

5) Set Up Python Virtual Environment:

--------------------------------------

- Create a new virtual environment:

virtualenv marker

- Activate the virtual environment:

source marker/bin/activate

- Install Marker OCR:

pip install marker-pdf

- Verify Python installation:

which python

Marker GitHub Repository:

https://github.com/VikParuchuri/marker

6) Download and Set Up Ollama:

-------------------------------

a) Download Ollama Binary:

curl -L https://ollama.com/download/ollama-linux-amd64.tgz -o ollama-linux-amd64.tgz

b) Create Ollama Folder and Navigate Into It:

mkdir ollama

cd ollama

c) Extract Ollama Binary:

tar -xzf ../ollama-linux-amd64.tgz

7) Run Ollama and Download DeepSeek Models:

--------------------------------------------

a) Start Ollama Server:

- Inside your ollama/ folder:

bin/ollama serve

- This starts the Ollama server listening on 127.0.0.1:11434.

- Leave this terminal open.

b) Open a Second Terminal:

- You may land on a different (login) node.

- Check current jobs:

squeue -u $USER

- Identify the compute node and SSH into it:

ssh <nodeid>

- Navigate to your Ollama working directory:

cd /your/working/directory/ollama

- Re-export your model storage environment:

export OLLAMA\_MODELS=$PWD/models

c) Pull DeepSeek Models:

- Pull the DeepSeek-R1 base model:

bin/ollama run deepseek-r1:base

- Other available DeepSeek-R1 distilled models:

bin/ollama run deepseek-r1:1.5b

bin/ollama run deepseek-r1:7b

bin/ollama run deepseek-r1:8b

bin/ollama run deepseek-r1:14b

bin/ollama run deepseek-r1:32b

bin/ollama run deepseek-r1:70b

--------------------------------------------------------

End of README