LLM Inference

Om Sai Krishna Madhav Lella

April 13, 2025

Overview

1. Install VSCode
2. Setup SSH Config
3. Connect to HPCC Development Node
4. Download and Copy Demo Code to HPCC
5. Create and Activate Python Virtual Environment
6. Create Hugging Face Access Token
7. Run Code
8. Schedule Job
9. Basic SLURM Commands

Install VSCode

Steps

→ Go to: `https://code.visualstudio.com/`

→ Click "Download for macOS"

→ Open the .zip file that downloads — it will extract to a Visual Studio Code.app.

→ Drag `Visual Studio Code.app` into the Applications folder.

→ Open it from Launchpad or Spotlight (Cmd + Space, then type "Visual Studio Code").

Setup SSH Config

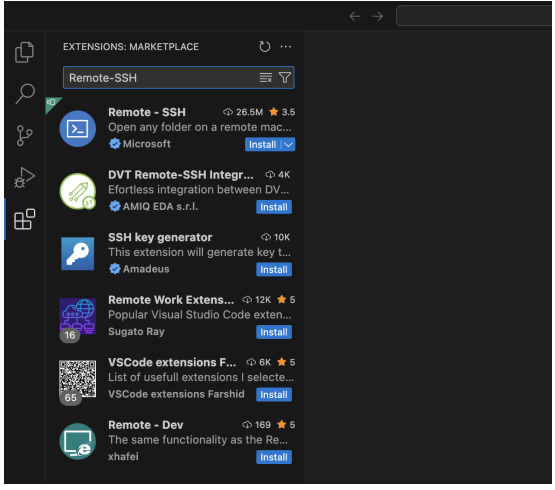
Install the Remote - SSH extension

→ Open VS Code.

→ Go to the Extensions view by clicking on the Extensions icon in the Activity Bar on the side of the window (or press Ctrl+Shift+X).

→ Search for "Remote - SSH" and click on the Install button.

Setup SSH Config



Setup SSH Config

Create/Edit Config file

→ Copy configuration from `https://github.com/madhav28/LLM-Demo/blob/main/SSH%20Config.md` and replace <netid>

→ Open Macbook Terminal.

→ Type **vim /Users/<username>/.ssh/config** and press i to insert text.

→ Paste the copied configuration, then press **esc**, type **:wq**, and press **Enter** to save and exit.

Connect to HPCC Development Node

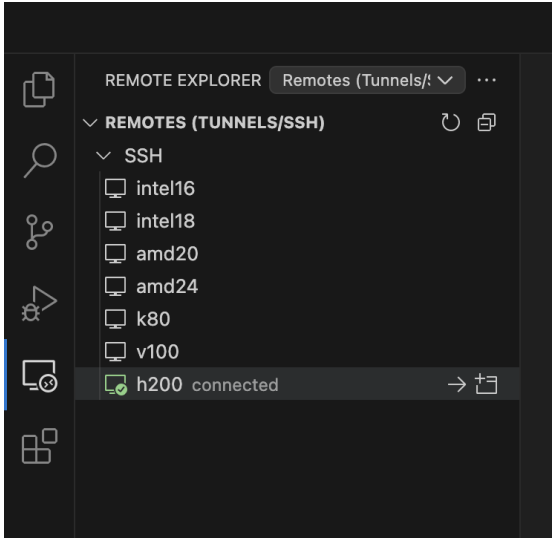
Steps

→ Open Visual Studio Code (VSCode).

→ In the sidebar, click on **Remote Explorer**.

→ Choose a development node and enter your password when prompted (you may need to enter it multiple times).

Connect to HPCC Development Node



Download and Copy Demo Code to HPCC

Download Demo Code

Link to the code and data: `https://github.com/madhav28/LLM-Demo/tree/main/LLM%20Inference%20Demo%20Code`

Copy and Paste Directory to HPCC

scp -r "/copy/path" "<netid>@hpcc.msu.edu:/mnt/home/<netid>/sample/path"

Create and Activate Python Virtual Environment

Create Python Virtual Environment

```
python -m venv /path/to/venv
```

Activate Python Virtual Environment

```
source /path/to/venv/bin/activate
```

Create Hugging Face Access Token

Steps

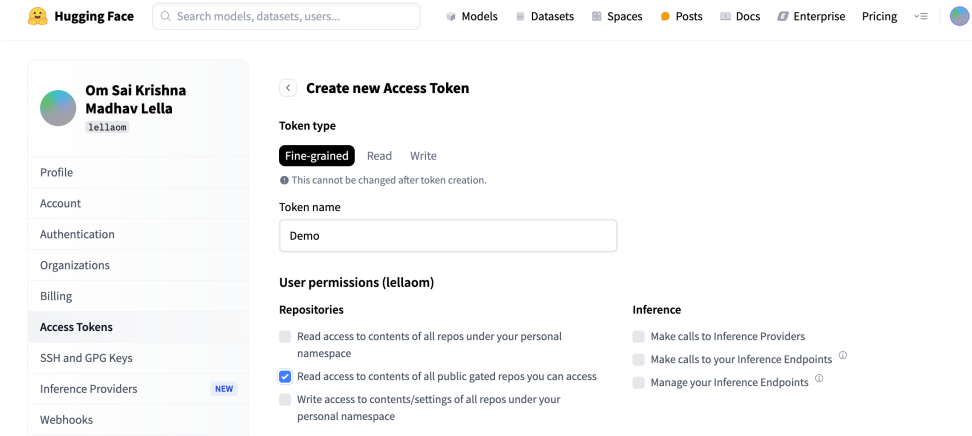
→ Go to `https://huggingface.co/settings/tokens`.

→ Click **Create new token**.

→ Enter token name click **Read access to contents of all public gated repos you can access**.

→ Click **Create token**, copy the token, and replace it in the code.

Create Hugging Face Access Token



Run Code

Command

```
python few_shot_classification.py
```

Schedule Job

Command

```
sbatch job_script.sh
```

Basic SLURM Commands

Job Status

squeue -u <netid>

Cancel a Job

scancel <job_id>

Cancel All Jobs

scancel -u <netid>

Thank You!