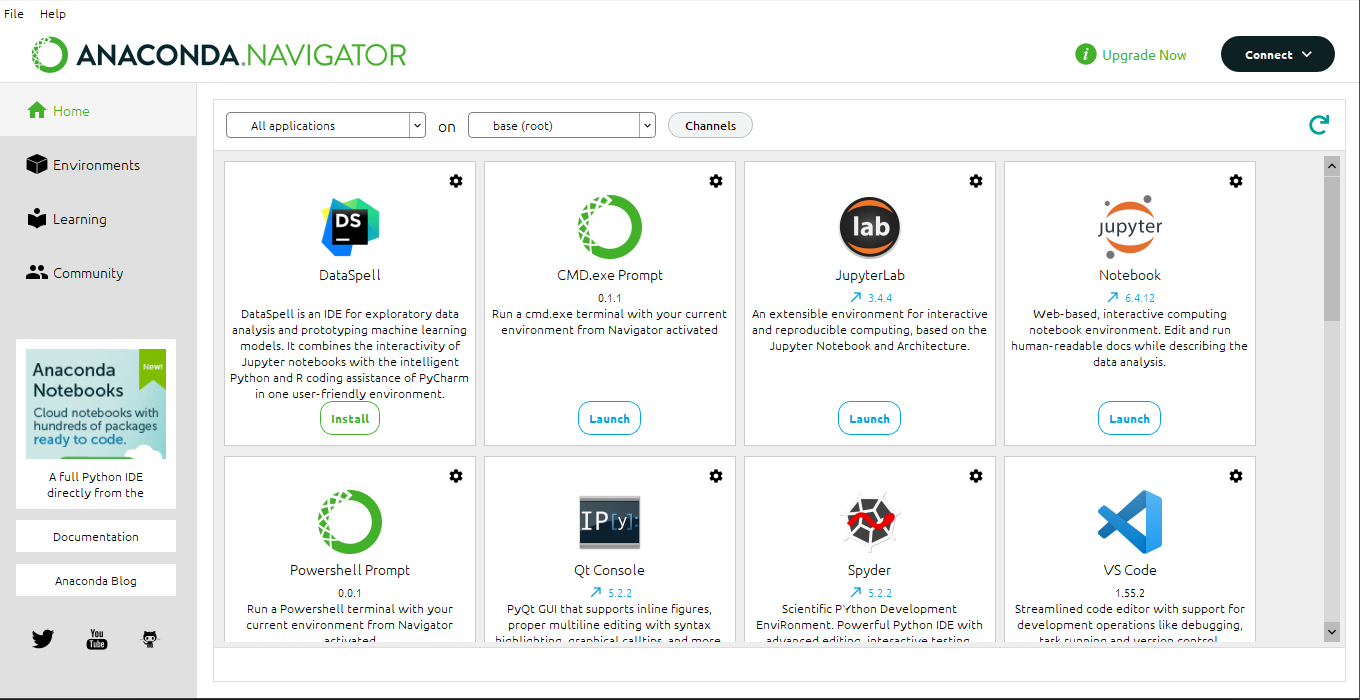
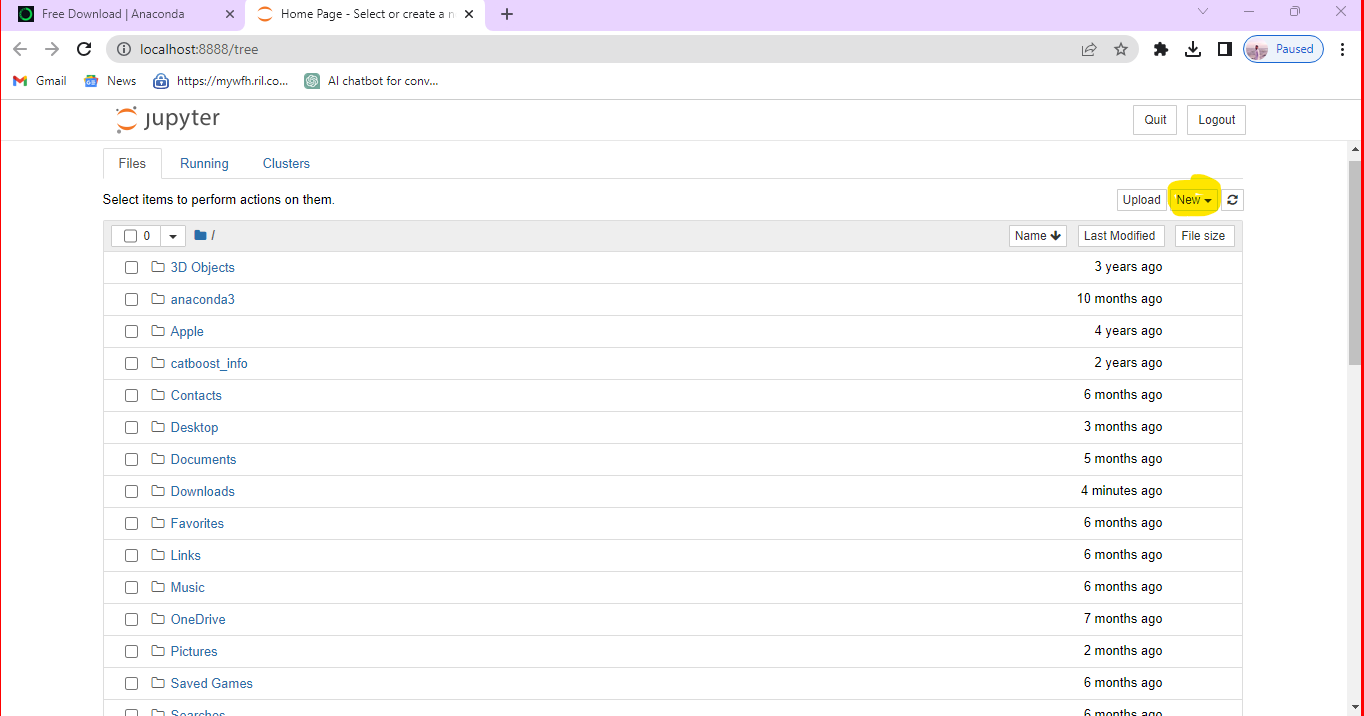
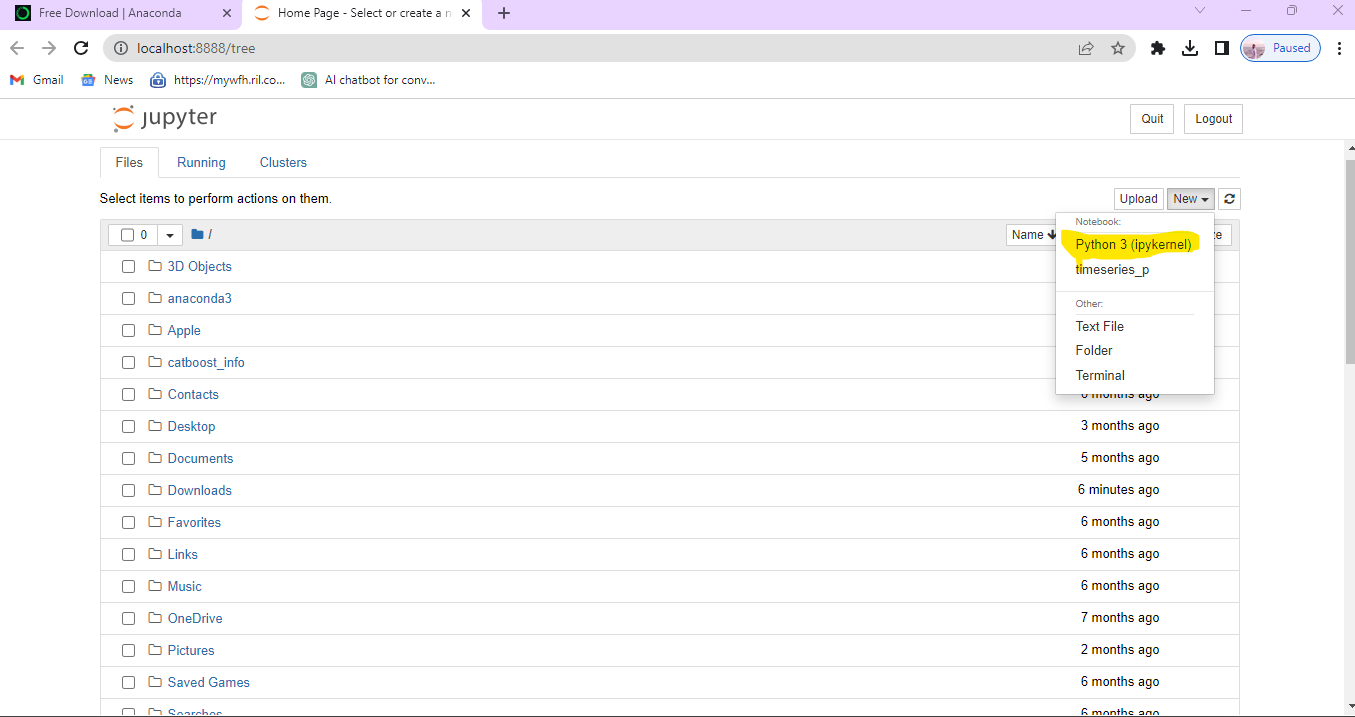
**Steps to setup Llama 2 7B model in a Windows machine**

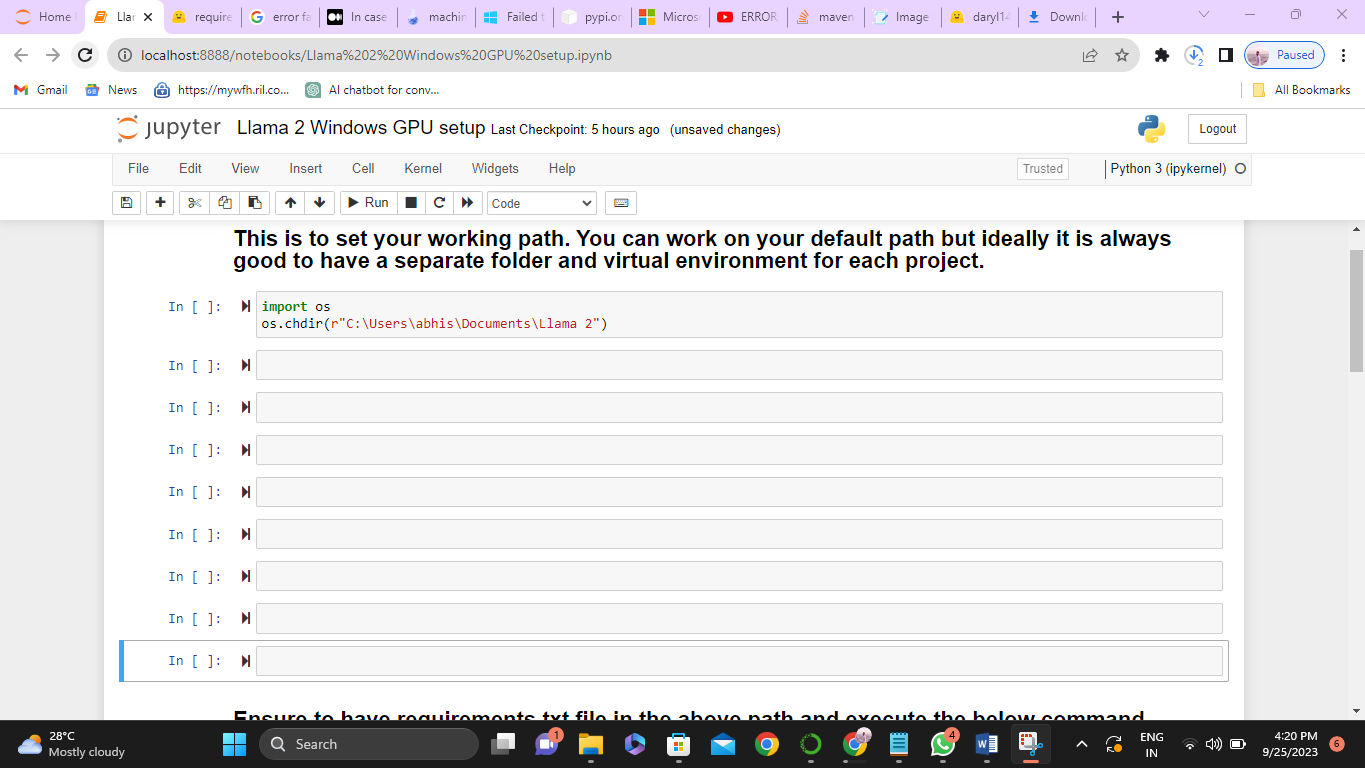
1. Install Anaconda tool kit using this link - <https://www.anaconda.com/download>
2. After downloading anaconda toolkit, search for Anaconda Navigator in search tab and launch the Jupyter notebook (First row fourth column) from the below screenshot.
3. This will open a tab in your default web browser and choose a new file by clicking on the highlighted text. Click on New -> Python3 (ipykernel)

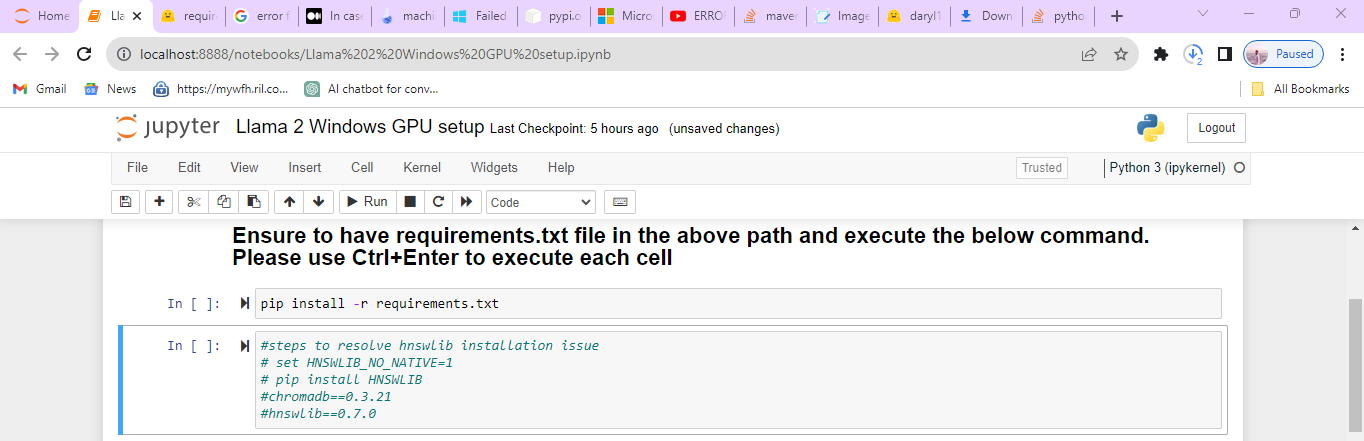


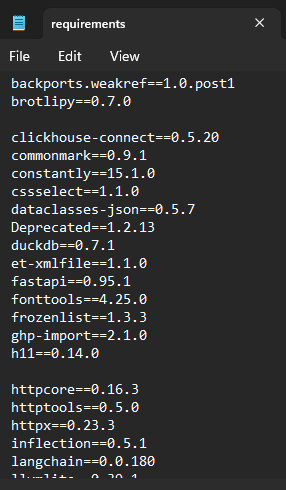
1. In he below screenshot click on the highlighted text which will enable us to open a new jupyter notebook in a default browser.



1. Set your working path as per the below. You can work on your default path but ideally it is always a good practice to have a separate folder and virtual environment for each project.

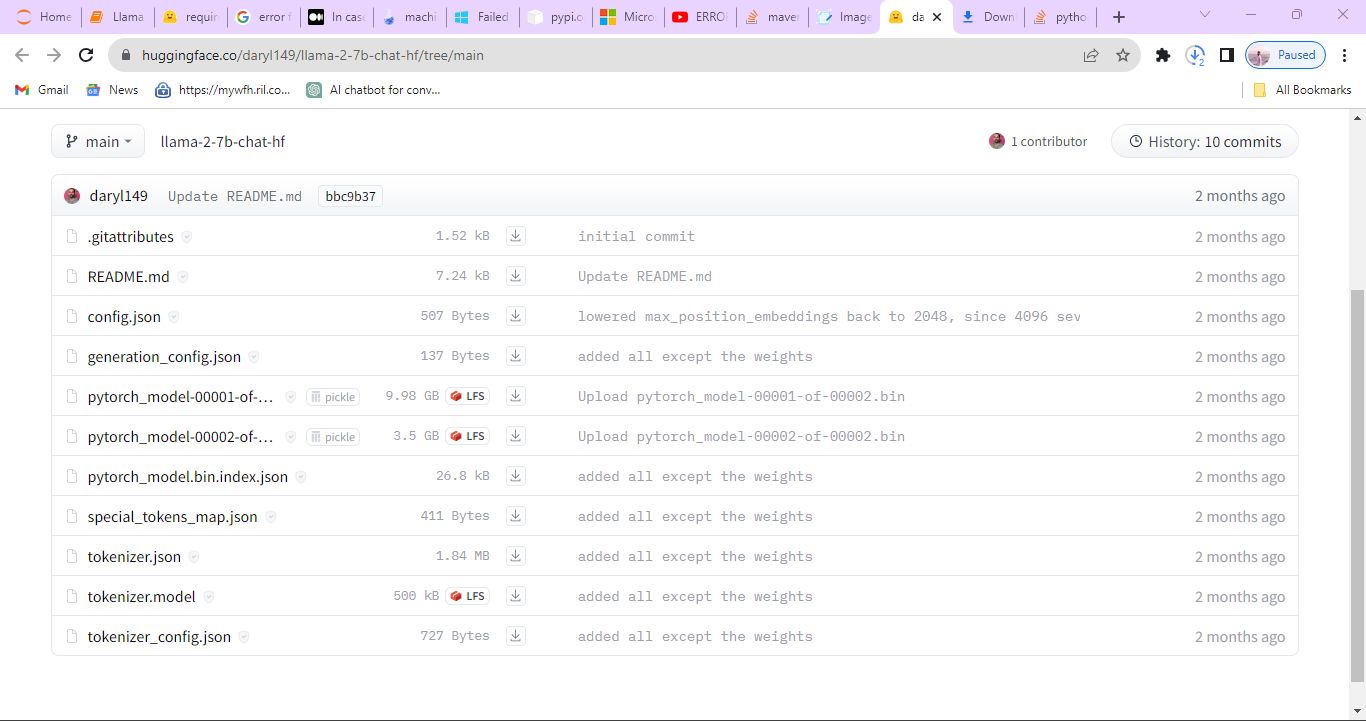


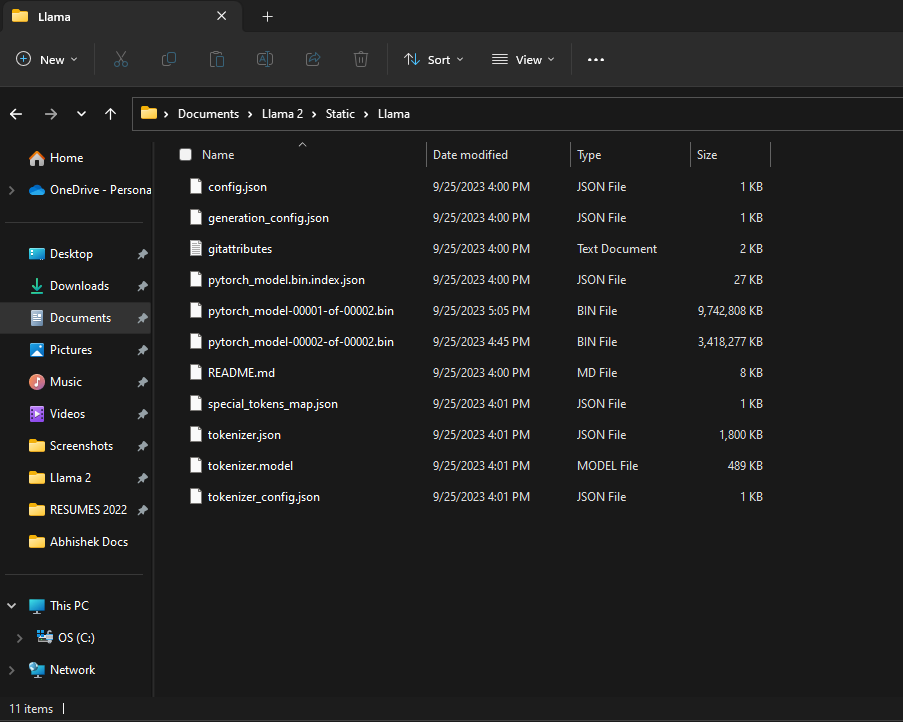
1. After setting the current working directory, we have to install the python libraries that are needed to implement our project. Usually we use pip (Preferred Installer Program) to install libraries. For e.g., *!pip install pandas* is used to install pandas library which is used to manipulate the tabular data. But if we have multiple libraries, we save all those libraries along with their versions in a .txt file and by using the below command. 



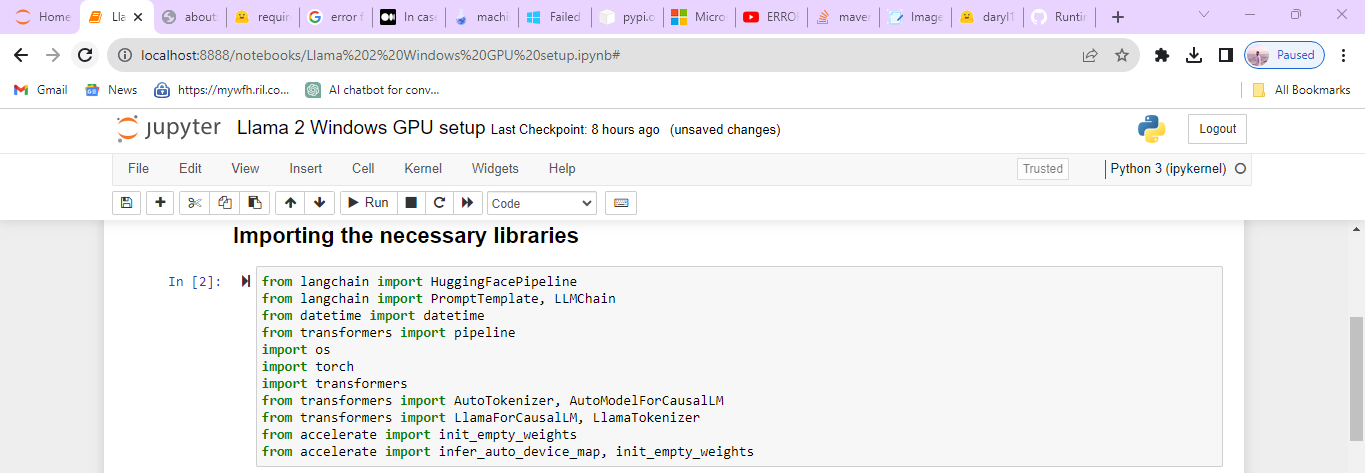
1. Download the model weights and store it in a folder. Ensure that this folder is in our current working path. These weights can be downloaded from HuggingFace hub. For e.g., model 7B weights can be downloaded from the below link and ensure to download all the files that are shown in the below screenshot.

<https://huggingface.co/daryl149/llama-2-7b-chat-hf/tree/main>

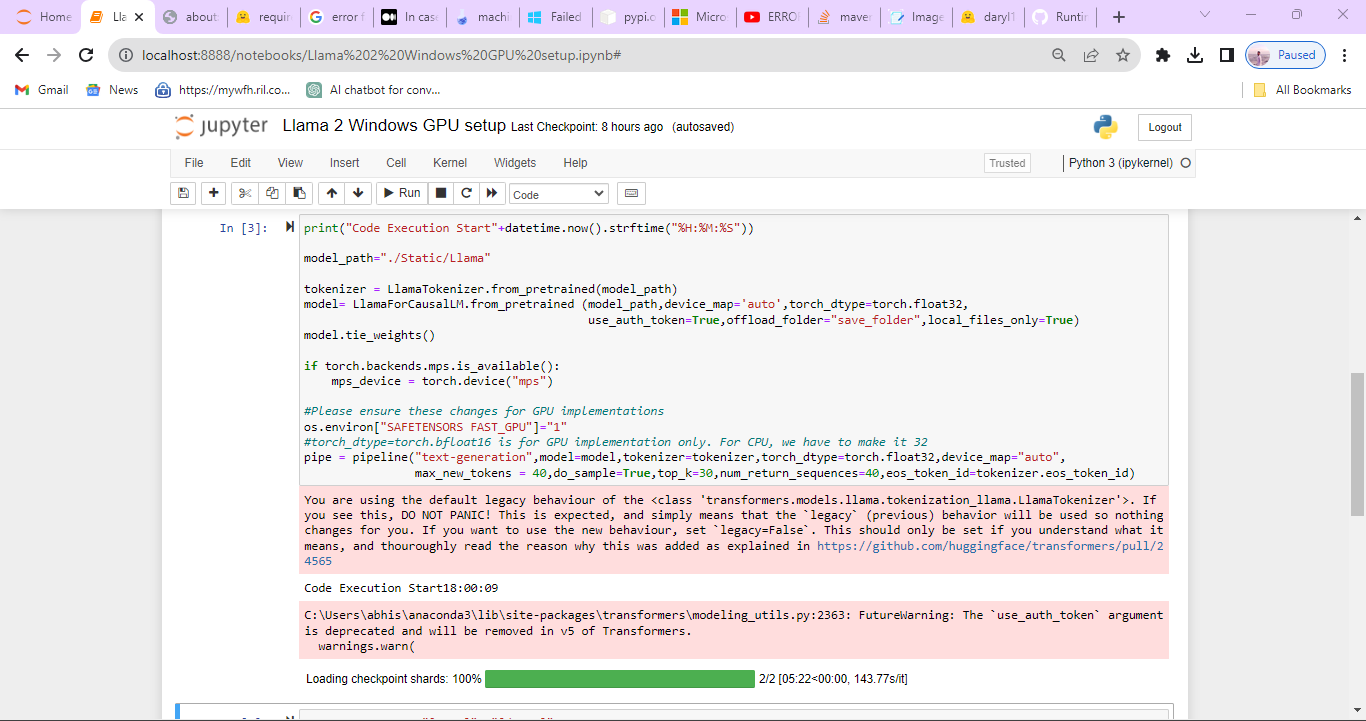




1. Import all the necessary libraries that are to be used for implementing the project. Also, define all the custom functions that can be reused as and when required.



1. Define the model path – path where we have downloaded and stored the model weights. Also, we have to instantiate the tokenizer, model, and other mandatory terms. For GPU based machines we have to assign to one environment variable as mentioned below. Create a pipeline to codify and automate the workflow as per the steps instantiated and needed for the task.



1. Set the default system prompt and the system prompt and create a prompt template for the custom inputs – like name, location and etc. Create the chain with the prompt and an LLM. Now when we pass the text to the chain, it will hit to the LLM and result us the response as per the instruction.

