GOOGLE SUMMER OF CODE

ABOUT-ME:

NAME: Sanjay Srikanth

EMAIL: sanjay.sriksun@gmail.com

College: New Horizon College of Engineering

YEAR: 2

LOCATION: India (Time zone GMT+5:30)

SKILLS: Python, C, basics of OOPs and java

PROBLEM: **Add an AI chatbot to the Chat Activity**

PROFILE: https://www.linkedin.com/in/sanjay-sriksun/

WHY THIS PROJECT:

As a beginner to the AI and ML course stream, the '**Add an AI chatbot to the Chat Activity**' problem statement has piqued my interest. Sugar presents an excellent platform for facilitating children's learning across different skills through various activities. This has caught my interest, giving me an opportunity to contribute to this project.

The chat activity within Sugar holds significant potential as a fundamental communication tool for children. It enhances a child's communication skills and their vocabulary.

Reasons why I would like to contribute to this project:

1. I have always found interest in AI and I believe this is a great opportunity for me to do something in this field.

2. The chatbot to the chat activity is a wonderful idea for children to increase their communication skills and vocabulary.

3. This will be a learning experience for me in the field of AI and I hope to understand better on how language models truly work.

Besides this, I would like to build this chatbot as a complete useful asset for the children apart from my learning experience.

HOW WILL SUGARLABS BE AFFECTED?

1. It will enable general engagement between children and sugarlabs.
2. It will also enhance children's communication capabilities.
3. Sugarlabs can now be a far better tool for a child and boost their skills.
4. Young users can learn basic words, how to speak in formal and informal ways.

APPROACH TO PROBLEM:

The chat activity is going to be based on fine tuning an already existing pre-trained language model. There exists a couple of free and open LLMs that can be fine-tuned. **Pytorch** offers a pre-trained language model to work with. I intend to use Pytorch to fine tune the model with different datasets and build a new model architecture for the model that aims for the younger audience.

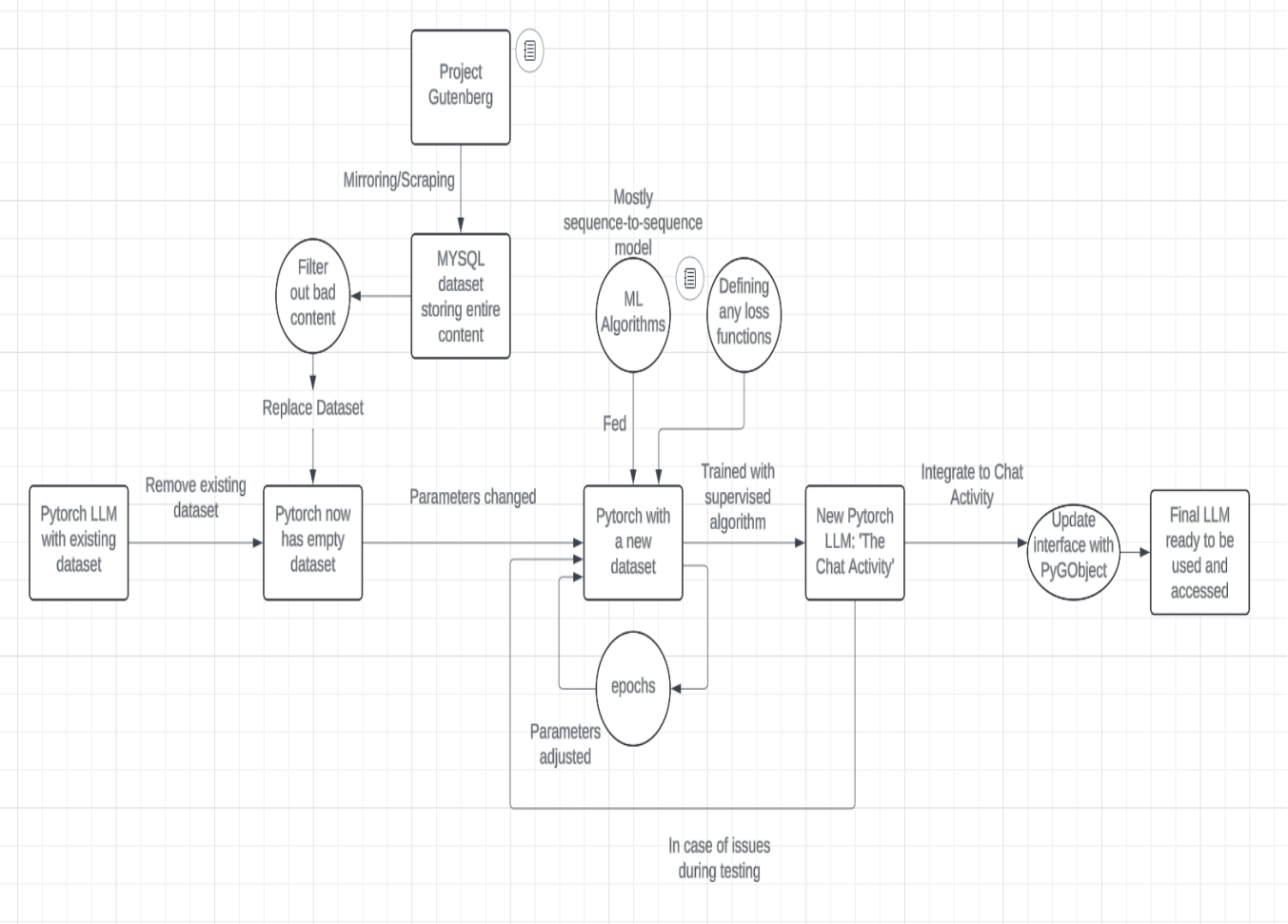
Project Gutenberg can be used as a dataset since it has a variety of e-books, it's files are free to use and is a large text corpus. Moreover, the content is in plain text format which makes it a better source for a machine to learn. I intend to establish a good engagement of conversation and since e-books have various content, I believe it will be easier for the model to train on this dataset. Hence, I will use Project Gutenberg for machine learning and text generation.

Any inappropriate content will be filtered out and the model will be trained. The model will be trained on supervised algorithms and reinforced learning. This is to ensure better response outputs and outputs based on feedback. Once trained, the model gets integrated to the chat activity.

I intend to use **SQL** to store all the data for the datasets and use querying to filter out the content not required. I have knowledge on MYSQL and querying and have some experience on using SQL. I also have knowledge on **python** which I would most likely use for the project.

I hope my skills are enough to contribute and I'd love working on this project.

Over here is a block-chart diagram of how my entire LLM during the creation and training process would look like:



LINK: https://lucid.app/lucidchart/36fb63c9-8f04-49c2-a554-60b4f98d169e/edit?viewport_loc=-56%2C-97%2C2219%2C1076%2C0_0&invitationId=inv_caabbd63-d599-4ea0-bfb8-1ac8b6e7d81a

(Lucid website for creating block diagrams and flow charts)

TIMELINE:

I plan to adhere closely to this timeline for the project, ensuring thorough follow-through to complete the project and fulfill my contribution.

The 175-hour project will be divided in weeks as follows:

WEEK 1:

Checking out the chat activity codebase and testing it out.

Using Pytorch language model and test out its resources.

WEEK 2:

Use Pytorch to sample create a language model with a small dataset and test it out.

(This includes creating the model with a sample database, replace the existing Pytorch dataset with a sample dataset and designing how the model should look like along with a model architecture.)

WEEK 3 and 4:

Creating the actual dataset with **MYSQL** by creating a private mirror that copies all file data (**rsync** and **GNU wget**) and possibly use scraping tools (like python's **beautifulsoup**) to collect data from Project Gutenberg

This data will be used as the learning source for the AI.

In this phase, the dataset will be queried to filter out any inappropriate/unnecessary data out and finalize the dataset to be used.

WEEK 5:

Replace the sample dataset created earlier in week 2 with real dataset used in week 3 and 4. The hyperparameters are also tuned

Here I will also define all the loss functions using Pytorch itself.

WEEK 6 and 7:

The model will undergo training which may take a couple of days.

The model will also undergo numerous tests to check its working status and debugged if there is an issue.

WEEK 8:

After a desirable outcome of the model, it will finally be integrated to the chat activity source code.

The interface will also be updated. The interface in the chat activity code seems to be a version of GTK+ 3. I intend to use **pyGObject** to work on the interface.

During the course of the project, at the end of every week my progress will be shown to the mentor and the timeline will be followed strictly.