Web Scraping of Top 50 Hospitals and GPT Model Training

Contents

Code Files …… 1
Data Collection …….. 1
Model Training …… 1
Conclusion …… 2

Code Files

- **Main.py :** The main file contains all of the code from fetching the data to training the mode.
- **scrap_data.json :** The scrap data of top 50 hospital.
- **./gpt_finetuned :** The trained private GPT model is saved in the folder.

Data Collection

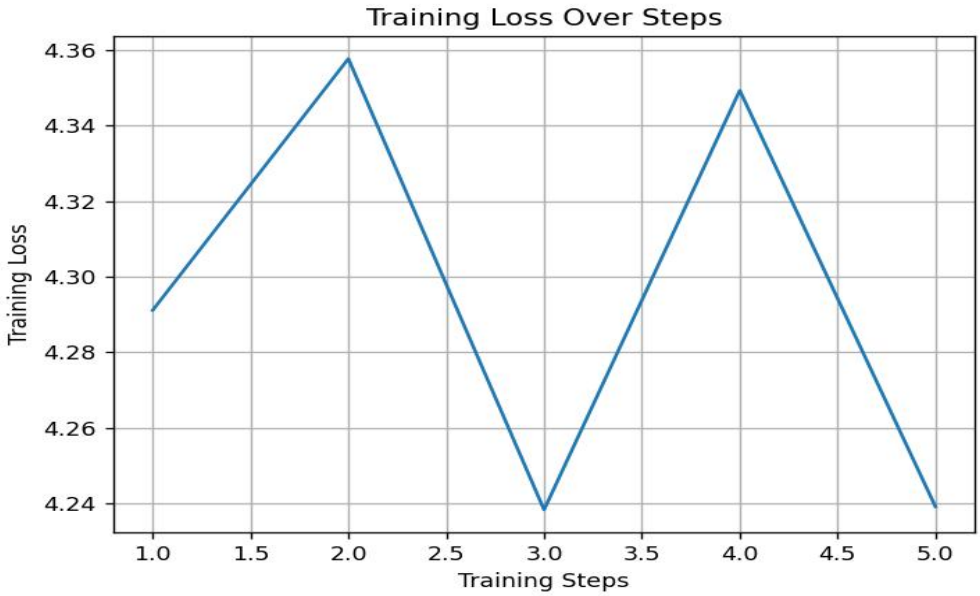
The code main.py initially retrieves all of the links from the website Newseek. Then we go through each link, scraping the data from each. We scrap only the <p> tags content. Finally, we clean the data by sending it via the clean function.

Model Training

After cleaning the data we train the model by passing the clean data.where we utilize he libraries like transformers. The GPT2 tokenizer from the transformers library was used to tokenize the text data.

Model Performance

After training, the trained model checkpoint was saved.



Conclusion

The project successfully accomplished its goal of web scraping the top hospital websites and training a proprietary GPT model on the scraped data.