Private GPT on Custom Data

This document provides a detailed overview of a project conducted for **AI47Labs**, which involved harvesting data from major hospital websites, training a specialized Private GPT model with this data, and assessing the model's effectiveness. The project utilized sophisticated natural language processing (NLP) techniques to analyse and derive insights from content on the websites of top healthcare providers.

Introduction

This document outlines how the project leveraged recent advancements in NLP and machine learning to extract and analyse substantial data from online sources. It specifically focused on gathering information from the websites of four renowned hospitals: Clínica Universidad de Navarra, Johns Hopkins Medicine, Singapore General Hospital, and Stanford Medicine Healthcare. The collected data included detailed profiles of doctors, their specialties, and departmental information, which were carefully scraped, cleaned, and formatted for training a Private GPT model. The document details the approaches used in data collection, model training, and evaluating the model's performance.

Data Collection

The initial phase of the project focused on the development and implementation of web scraping tools to facilitate the automated gathering of comprehensive information from selected hospital websites. This information spanned a broad spectrum, including healthcare providers' profiles, their areas of expertise, and specific departmental details. To organize the data effectively, the following column names were utilized: Page link, Link, Name, Category, Department, Department link, and Information. These labels were integral in categorizing the data for subsequent phases of the project.

A variety of Python libraries, including Beautiful Soup and Selenium, were utilized to effectively traverse and interpret the HTML frameworks of the websites. Subsequently, the extracted data was methodically compiled into a neatly arranged XLSX file (GPT2_finetune_data.xlsx), providing a well-structured dataset primed for additional analysis.

Model Training

Following the data collection phase, the dataset was pre-processed to convert it into a compatible format for the GPT model's training process. The proprietary GPT model, which builds upon the acclaimed GPT-2 framework, underwent refinement and adjustments using the carefully prepared dataset.

Training Parameters

- Base Model: GPT-2
- Number of Training Epochs: 3
- Batch Size: 4 per device
- Utilized Framework: PyTorch, supplemented with Hugging Face Transformers library

The training process tailored the GPT model to align with the healthcare context drawn from the hospital data. This customization allowed the model to produce content that was insightful and relevant to the specific context.

Model Training Progress and Performance Assessment

The following table details the progression of the model's training and validation losses over the course of three epochs. This data offers a glimpse into how the model's performance has enhanced as it learned from the data over time.

1st Training step

```
/usr/local/lib/python3.10/dist-packages/transformers/data/datasets/language_r
  warnings.warn(
[99/99 17:43, Epoch 3/3]
```

Epoch	Training Loss	Validation Loss
1	No log	1.331795
2	No log	1.071182
3	No log	1.017012

Training and Validation Loss:

- This image shows that no training loss logs were recorded for all three epochs, which might be an issue with logging the training loss or it could indicate a choice to focus on validation loss for performance assessment.
- **Epoch 1**: Validation loss is 1.331795, same as the first image.
- **Epoch 2 and 3**: No training loss is logged; however, the validation loss shows a decreasing trend, being 1.071182 in Epoch 2 and 1.017012 in Epoch 3.

2nd

```
/usr/local/lib/python3.10/dist-packages/transformers/data/datasets/language
  warnings.warn(
[99/99 17:20, Epoch 3/3]
```

Epoch	Training Loss	Validation Loss
1	No log	1.331795
2	2.107000	1.071182
3	2.107000	1.017012

Training and Validation Loss:

- **Epoch 1**: No training loss was recorded (indicated as "No log"), but a validation loss of 1.331795 is noted, suggesting that this is the starting baseline for model performance.
- **Epoch 2**: Shows a training loss of 2.107000 and a decreased validation loss of 1.071182, which indicates an improvement in the model’s ability to generalize to unseen data.
- **Epoch 3**: The training loss is the same as in Epoch 2, recorded at 2.107000, and the validation loss further decreases slightly to 1.017012, showing continued improvement.

Generated Output by the Model

```
# Pass a query to the chain
query = "Do you have a column called Name?"
response = chain({"question": query})
```

```
/usr/local/lib/python3.10/dist-packages/langchain_core/_api/deprecation.py:119: LangChainDeprecationWarning: The method `Chain.__call__` was deprecated in langchain 0.1.0 and will be removed in 0.2.0. Use invoke instead
  warn_deprecated(
```

```
print(response['result'])
```

```
Yes, all of the pieces of context provided include a name for the individual profiled on the Johns Hopkins Medicine website.
```

```
# Pass a query to the chain
query = "Will you provide me all the information about Dr.Fanny Meylin Caballeros Lam ?"
response = chain({"question": query})
```

```
print(response['result'])
```

```
Yes, Dr. Fanny Meylin Caballeros Lam is a specialist in the Radiology Service at the CUN (Clinica Universidad de Navarra) and her department link is https://www.cun.es/en/our-professionals/medical-services/radiology. Additic
```

1.Prompt: how many cancer specialist are there?

Generated: how many cancer specialist are there?

Cheryla Lai: I am an oncologist, and I treat patients at the National Cancer Institute of China Cancer Center, at Stanford Hospital.

Oncologist: Cun.org

2. Prompt: How to I get there website link?

Generated: How to I get there website link?

Websites for information about the Department's operations in Iraq and Syria: https://www.cun.es/en/our-professionals/doctors-medical-staff/prof-

3. Prompt: The specialist Dr. Ricardo Calderón Pelayo [SP] works on which category?

Generated: The specialist Dr. Ricardo Calderón Pelayo [SP] works on which category? -------------------------- Aphrodisiacal Medicine Department [SP] Status: Visiting Specialist | Department: Ophthalmology Department

4. Will you provide me all the information about Dr.Fanny Meylin Caballeros Lam ?

Generated: Yes, Dr. Fanny Meylin Caballeros Lam is a specialist in the Radiology Service at the CUN (Clinica Universidad de Navarra) and her department link is https://www.cun.es/en/our-professionals/medical-services/radiology. Additionally, you can find her professional profile and photo at the following page link: https://www.cun.es/nuestros-profesionales/profesionales/fanny-meylin-caballero-lam and her photo link: https://www.cun.es/.imaging/mte/newstyle/profesional_carrusel-80x80/dam/cun/imagen/profesionales/radiologia/dra-fanny-meylin-caballeros-lam/jcr:content/dra-fanny-caballeros.webp.

5. Provide me names for doctor who are specialist in Radiology Service?

Generated: Dr. Fanny Meylin Caballeros Lam, Dr. David Cano Rafart, Dr. Marta Calvo Imirizaldu.

These instances illustrate the model's refined comprehension and its capacity to formulate insightful replies, highlighting its viability as an instrument for the automation of content creation in the health sector.

Review and Insights

The model displayed a promising proficiency in fabricating text that adheres to the professional and informative style characteristic of medical provider descriptions. The obtained perplexity metrics were within satisfactory limits, signifying an elevated accuracy in predictions.

Summary and Actionable Insights

The initiative has successfully validated the potential of using GPT frameworks to analyse and synthesize significant content derived from health sector data. The custom-trained GPT model, informed by data from premier hospitals, showed a robust ability to generate coherent and context-sensitive narratives, emphasizing its possible value in assorted healthcare sector applications.

Prospective Avenues

- **Data Diversification**: Augmenting the dataset may refine the model's precision and adaptability.
- **Operational Integration**: Examining deployment approaches for the model's incorporation into health information ecosystems.
- **Ongoing Adaptation**: Instituting methods for the model's ongoing adaptation through continual learning from emergent data streams.

Acknowledgements

We express our sincere appreciation to AI47Labs for their confidence in commissioning this venture and acknowledge the development team for their commitment and technical acumen in overcoming obstacles to fulfil the project's goals.

Training Using LangChain

In training the new data, the LangChain framework was also utilized to enhance the GPT model's training and evaluation process. LangChain is designed to facilitate sophisticated interactions with language models, enabling a more nuanced and efficient engagement with the CSV dataset. This method led to a refined dataset exploration, which yielded insightful findings showcasing the model's sophisticated capabilities.

Approach

By employing LangChain, we constructed a set of specific queries and tasks aligned with the structured information in the CSV file. This approach allowed for a concentrated dataset exploration, extracting precise information and creating content related to hospital profiles, specialties, and other crucial details.

Engagement Synopses

LangChain was instrumental in directing the model to produce responses to a diverse array of prompts originating from the dataset's structure, enabling a vigorous assessment of the model's comprehension and its ability to apply the knowledge it has acquired to generate accurate and pertinent content.

- **Data-Driven Inquiries**: The formation of queries based on the dataset's data permitted the evaluation of the model's capacity for information retrieval and utilization, illustrating its practical application in real-world contexts.
- **Narrative Creation**: With the use of organized data, the model excelled in crafting detailed narratives, descriptions, and informational content that are of particular relevance to the healthcare industry.

Reflections

The outcomes from utilizing LangChain with the GPT model emphasize the model's adeptness in understanding and crafting responses from structured datasets. The model shows a clear understanding of the data's architecture and the ability to convey detailed, pertinent information when prompted. These capabilities affirm the model's suitability for tasks that demand sophisticated data interaction and bespoke content creation, signalling potential pathways for further innovation and refinement.

Summary of Outcomes

The LangChain-mediated engagements produced several noteworthy results:

- **Precision and Contextual Relevance**: The model consistently generated content that was not only precise but also profoundly aligned with the context of the inquiries, reflecting a solid comprehension of the underlying data.
- **Utility in Application**: The productive use of LangChain showcases the model's capacity for streamlining content generation, data mining, and insight derivation, presenting a solid case for its applicability across the healthcare landscape and other sectors.

Concluding Thoughts

Integrating LangChain with our dataset analysis has reinforced the adaptability and capability of the fine-tuned GPT model. This integration has carved out new pathways for employing advanced NLP methodologies with structured data, paving the way for upcoming endeavors to maximize the use of language models in data scrutiny and narrative creation.

Setting Up and Executing the Fine-Tuning of GPT Model on Hospital Data

To work with the files effectively, there are a couple of steps to follow. The 'main.ipynb' notebook contains the code for fine-tuning the model with the dataset, while 'web_scraping.ipynb' is responsible for extracting the data that will be used for fine-tuning. The data for this process is located in the 'GPT2_finetune_data.xlsx' file.

Before running 'main.ipynb', you must adjust the file paths to match your system's directory structure. This will ensure the code correctly converts the .xlsx file into a .txt format, which is then split into separate training and testing files for the machine learning process.

Once the data is prepared, you can proceed with the training, which may take some time. After training is complete, you'll calculate evaluation metrics to assess the model's performance. With the successful completion of these steps, you'll be set to deploy your custom-trained GPT model on the hospital data for insightful analysis and content generation.