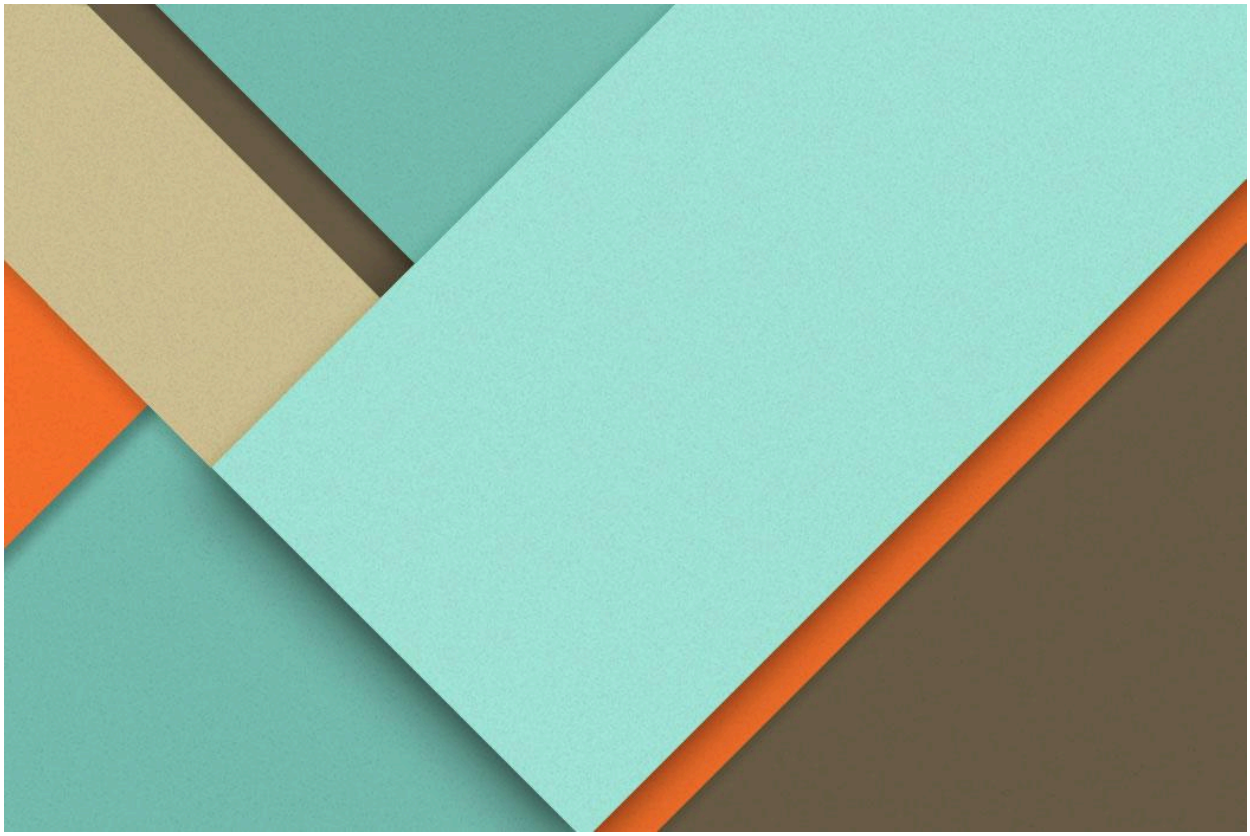


MedExtractAI

—

Web Scraping and AI Model Training

Project Overview

The goal of this project was to scrape data from the top 50 hospitals worldwide, clean and preprocess the data, and then train a **Private GPT** model on this dataset. The Private GPT model, sourced from GitHub, was fine-tuned to generate hospital-related content, including information about hospital services, departments, and general healthcare topics.

Steps

1. Data collection: Scraped relevant data from the websites of the top 50 hospitals.
2. Data cleaning: Cleaned and structured the scraped data for model training.
3. Model training: Fine-tuned a **Private GPT** model using the cleaned data.

1. Data Collection

- We began by identifying the top 50 hospitals in India using www.newsweek.com/rankings/worlds-best-hospitals-2024/india
- These rankings provided us with a list of hospitals across various specialties, which were then used to gather URLs for scraping.
- The script `extract_top_hospitals.py` was used to automatically fetch the URLs of the top 50 hospitals by scraping the list.
- Once we had the URLs of the hospitals, the next task was to extract relevant data from their websites.
- The script `hospital_data_scraper.py` was used to scrape.
- We used BeautifulSoup for scraping.
- The scraped data was saved in `scraped_data.json`.

2. Data Cleaning

- Load JSON Data: Read the data from `scraped_data.json`.
- Extract Text: Access the value of `'content']['p']` for each entry.
- Remove HTML Tags: Use `BeautifulSoup` to strip any embedded HTML tags.
- Normalize Whitespace: Replace all types of whitespace (`\s+`) with a single space.
- Remove Escape Sequences: Clean up characters like `\t`, `\n`, and zero-width spaces (`\u200b`).

- Remove Special Characters: Delete any non-alphanumeric characters except punctuation marks.
- Convert to Lowercase: Standardize text to lowercase.
- Trim Spaces: Remove leading and trailing spaces.
- Write Cleaned Data to File: Save the cleaned text to `cleaned_dataset.txt`.

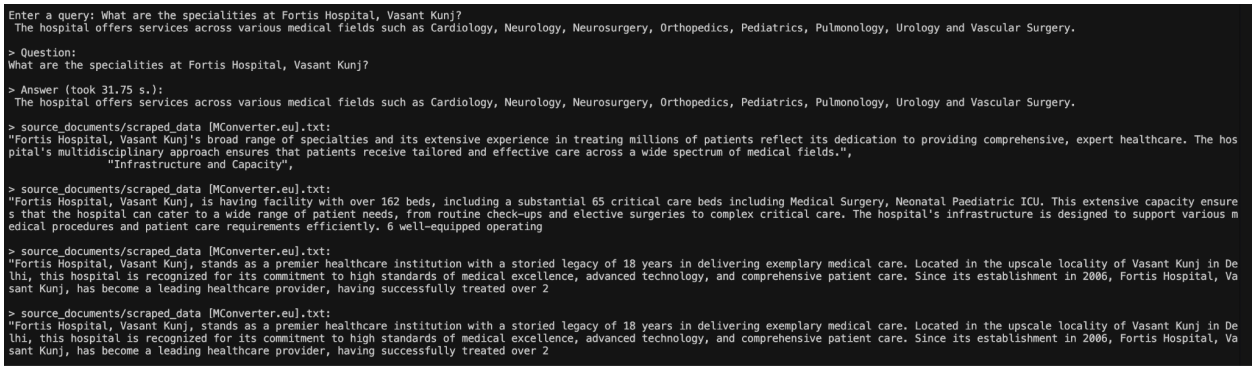
3. Model Training using Private GPT

- For this project, we used **PrivateGPT**, a private version of GPT that allows you to fine-tune a GPT model on your own dataset. PrivateGPT was sourced from GitHub and is designed to be flexible and customizable for various types of data.
- The fine-tuning was done on the cleaned hospital data to generate content relevant to healthcare and hospital-related inquiries.

Findings and Performance of the Model

- **Quality of Data**: The quality of the scraped data was directly linked to the performance of the trained model. Hospital websites with structured and detailed information resulted in more accurate and coherent outputs from the model.
- **Model Performance**: After fine-tuning, the Private GPT model was able to generate coherent responses to healthcare-related queries, including hospital information, services, and general healthcare knowledge.

Screenshots



```
Enter a query: Who is Dr. Amit Chandanisone?
 Dr. Amit Chandaniso ne of India's leading cardiothoracic surgeons. In his career, he has participated in over 10,000 heart surgeries. Apart from CIMS Hospital, Dr. Chanda n also visits K.K.Hospital
 at Bhuj.

> Question:
Who is Dr. Amit Chandanisone?

> Answer (took 49.65 s.):
 Dr. Amit Chandaniso ne of India's leading cardiothoracic surgeons. In his career, he has participated in over 10,000 heart surgeries. Apart from CIMS Hospital, Dr. Chanda n also visits K.K.Hospital
 at Bhuj.

> source_documents/scraped_data [MConverter.eu].txt:
"He is actively involved in many research projects and has many publications to his credit. He has also made several presentations at National conferences.",
                "Dr. Chandan has been teaching and training medical students, house officers, nurse surgical assistants as well as surgical trainees.",
                "Areas of Expertise :",

> source_documents/scraped_data [MConverter.eu].txt:
field of Cardiothoracic and vascular surgery.He is a member of theIndian Society for Heart and Lung Transplantation (INSHLT) andIndian Association of Cardiovascular-Thoracic Surgeons(IACTS).He is ac
tively involved in many research projects and has many publications to his credit. He has also made several presentations at National conferences.Dr. Chandan has been teaching and training medical s
tudents, house officers, nurse surgical assistants as well as surgical trainees.Areas of Expertise

> source_documents/scraped_data [MConverter.eu].txt:
"Dr. Chandan regularly participates in several workshops, training courses, as well as conferences worldwide in order to keep up on with the latest in the field of Cardiothoracic and vascular surger
y.",
                "He is a member of theIndian Society for Heart and Lung Transplantation (INSHLT) andIndian Association of Cardiovascular-Thoracic Surgeons(IACTS).",

> source_documents/scraped_data [MConverter.eu].txt:
"Cardiac Surgery",
                "MBBS, DNB (Cardiothoracic Surgery)",
                "Consultant Cardiothoracic and Vascular Surgeon",
                "Overview :",
                "Dr. Amit Chandanisone of India\u2019s leading cardiothoracic surgeons. In his career, he has participated in over 10,000 heart surgeries. Apart from CIMS Hospital, Dr. Chandan also
visits K. K. Hospital at Bhuj.",
```

```
Enter a query: Who is Dr. Naresh Trehan?
 Dr. Naresh Trehans a world-renowned cardiovascular and cardio thoracic surgeon who founded Medanta Gurugram with the mission to deliver advanced yet affordable medical services.

> Question:
Who is Dr. Naresh Trehan?

> Answer (took 37.46 s.):
 Dr. Naresh Trehans a world-renowned cardiovascular and cardio thoracic surgeon who founded Medanta Gurugram with the mission to deliver advanced yet affordable medical services.

> source_documents/cleaned_data.txt:
--------------------------------------------------
URL: https://www.medanta.org/gurugram/
Medanta Gurugram is founded byDr. Naresh Trehan, a world-renowned cardiovascular and cardiothoracic surgeon. It was conceived with the mission to deliver advanced yet affordable medical services.

> source_documents/scraped_data [MConverter.eu].txt:
"Important E-mail Addresses at AIIMS"
                ]
            }
        },
        {
            "url": "https://www.medanta.org/gurugram/",
            "content": {
                "p": [
                "Medanta Gurugram is founded byDr. Naresh Trehan, a world-renowned cardiovascular and cardiothoracic surgeon. It was conceived with the mission to deliver advanced yet affordable med
ical services.",

> source_documents/scraped_data [MConverter.eu].txt:
"For five years in a row, Medanta Gurugram has ranked as the Best Private Hospital in India (2020, 2021, 2022, 2023 and 2024). It also featured in the list of top 250 World's Best Hospitals Survey 2
024 by Newsweek.\u00a0...",
                "Medanta Gurugram is founded byDr. Naresh Trehan, a world-renowned cardiovascular and cardiothoracic surgeon. It was conceived with the mission to deliver advanced yet affordable med
ical services.",

> source_documents/scraped_data [MConverter.eu].txt:
"Md. Hussain Khan, a patient from Bangladesh speaks about his experience with Fortis Memorial Research Institute, Gurugram. He had been suffering from bone marrow disease. After undergoing microsurg
ery of the spine under the supervision of Dr. Sandeep Vaishya (HoD & Executive Director – Neurosurgery, Fortis Memorial Research Institute, Gurugr\u2026",

Enter a query:
```