Documentation of Web Scraping, Data Cleaning, and training privateGPT model.

Overview

This documentation provides an overview of the web scraping and data cleaning project implemented using Python libraries Requests, BeautifulSoup, and Pandas. The project aims to scrape data from the MayoClinic website and clean the data for it to be used as private train data for privateGPT using gpt4all llm.

Table of Contents

6. Conclusion
    - Summary of the Project
7. Errors you might encounter
    - ERROR: Could not build wheels for hnswlib, which is required to install pyproject.toml-based projects
    - Other Errors while installing requirements or running ingest.py

1. Introduction

Purpose

This project aims to scrape data from the MayoClinic website and perform data cleaning operations to make the data usable for feeding into privateGPT as training data.

Scope

The scope of this project is limited to a MayoClinic website and involves scraping the name of doctors and their specialties from the website's pages. The project also includes data cleaning techniques using the Pandas library to preprocess the scraped data for feeding into training privateGPT.

Target Website

The target website for this project is the MayoClinic website (https://www.mayoclinic.org/appointments/find-a-doctor/search-results?searchterm=#edd114075cc94f35b9bccc081668c123). The website provides ordered list of doctors and their specialties.

Tools and Libraries

The following tools and libraries were used in this project: - Python (version 3.11) - Requests library (version 2.31.0) for making HTTP requests - BeautifulSoup4 library (version 4.12.2) for HTML parsing - data extraction - Pandas library (version 2.0.2) for data cleaning and manipulation - privateGPT (https://github.com/imartinez/privateGPT) for model creation.

2. Installation

Python

Ensure Python is installed on your system. You can download the latest version of Python from the official website (https://www.python.org).

Requests

To install the Requests library, use the following command:

```
pip install requests
```

Beautifulsoup4

To install the Beautifulsoup4 library, use the following command:

```
pip install beautifulsoup4
```

Pandas

To install the Pandas library, use the following command:

```
pip install pandas
```

PrivateGPT

To clone the privateGPT repository, use the following command:

```
git clone https://github.com/imartinez/privateGPT
```

To change the directory, use the following command:

```
cd privateGPT
```

To install requirements, use the following command:

```
pip3 install -r requirements.txt
```

3. Web Scraping

Fetching Web Page

The Requests library is used to send an HTTP request to the target website and retrieve the HTML content of the web page.

Parsing HTML

The BeautifulSoup library is used to parse the HTML content obtained from the web page. It provides convenient methods for navigating and searching the HTML structure.

Extracting Relevant Data

Using the parsed HTML, specific data fields are extracted using BeautifulSoup's find and find_all methods. The extracted data is stored in suitable data structures for further processing.

Handling Pagination

Since the target website has multiple pages, pagination logic is implemented to scrape data from all pages iteratively.

Storing Scraped Data

The scraped data is stored in CSV format.

4. Data Cleaning

Importing Data with Pandas

The Pandas library is used to import the scraped data into a DataFrame, a tabular data structure that allows for efficient data manipulation.

Handling Missing Values

Missing values in the scraped data are identified and handled appropriately. Techniques such as imputation or dropping rows/columns are applied based on the data characteristics.

Formatting Data

The data is formatted according to the desired data types, such as converting None to 'No Data' String value.

Exporting Cleaned Data

The cleaned data is exported to CSV file formats using Pandas.

5. Training privateGPT

Installing Requirements

To install requirements, use the following command:

```
pip3 install -r requirements.txt
```

Download gpt4all llm

Download gpt4all llm using ggml-gpt4all-j-v1.3-groovy.bin.

Changes in example.env file

If using any other LLM than mentioned above make necessary changes in 'example.env' file.

Rename 'example.env' to '.env'.

Placing Cleaned Data and LLM

The data is formatted according to the desired data types, such as converting None to 'No Data' String value.

Training Model

To train the model, use the following command:

```
python ingest.py
```

Run privateGPT.py

To run priavetGPT, use the following command:

```
python privateGPT.py
```

6. Conclusion

Summary of the Project

In this project, we successfully scraped data from the MayoClinic website using Python libraries Requests and BeautifulSoup4. The scraped data was then cleaned and processed using Pandas, resulting in a usable dataset for feeding into privateGPT as training data.

6. Error you might encounter

ERROR: Could not build wheels for hnswlib, which is required to install pyproject.toml-based projects

This error might occur while installing requirements for privateGPT. To resolve the error, use the following command:

```
export HNSWLIB_NO_NATIVE=1
```

Other Errors while installing requirements or running ingest.py

Make sure Python in use is version 3.10 or higher.