1. **Thông tin sinh viên/nhóm sinh viên thực hiện**

* Trần Thị Thanh Xuân – lớp 19IT3 – MSV 19IT216

1. **Tên đề tài**

* XÂY DỰNG CHATBOT TỰ ĐỘNG VỚI PYTHON

1. **Liệt kê các thư viện sử dụng (mô tả 1 dòng về chức năng của mỗi thư viện) và Mã nguồn tham khảo chính (Cung cấp Link Github hoặc link đến mã nguồn mà nhóm tham khảo)**

* Thư Viện Numpy: hỗ trợ cho việc tính toán các mảng nhiều chiều, có kích thước lớn với các hàm đã được tối ưu áp dụng lên các mảng nhiều chiều đó.
* JSON là tên viết tắt của cụm từ JavaScript Object Notation, nó là một kiểu định dạng dữ liệu tuân theo một quy tắc nhất định mà hầu hết mọi ngôn ngữ lập trình hiện nay đều có thể đọc được. JSON là một trong những tiêu chuẩn mở giúp trao đổi dữ liệu có trên website. Ở bài này em dùng json để đọc dữ liệu file Json và lưu vào trong các mảng

Text

Description automatically generated

* Thư viện Python NLTK là từ viết tắt của Natural Language Toolkit. Đây là một bộ thư viện cho phép chúng tôi thực hiện Xử lý ngôn ngữ tự nhiên (NLP).
  + steming: Loại bỏ dấu câu, chữ thường và loại bỏ trùng nhau
  + tokenization: đọc câu và tách câu, (Vd: "Cung cấp những gói du lịch nào?" -> "Cung","cấp","những","gói","du","lịch","nào", "?")
  + bag of word: tạo thành túi từ để dự đoán câu hỏi người dùng nhập vào

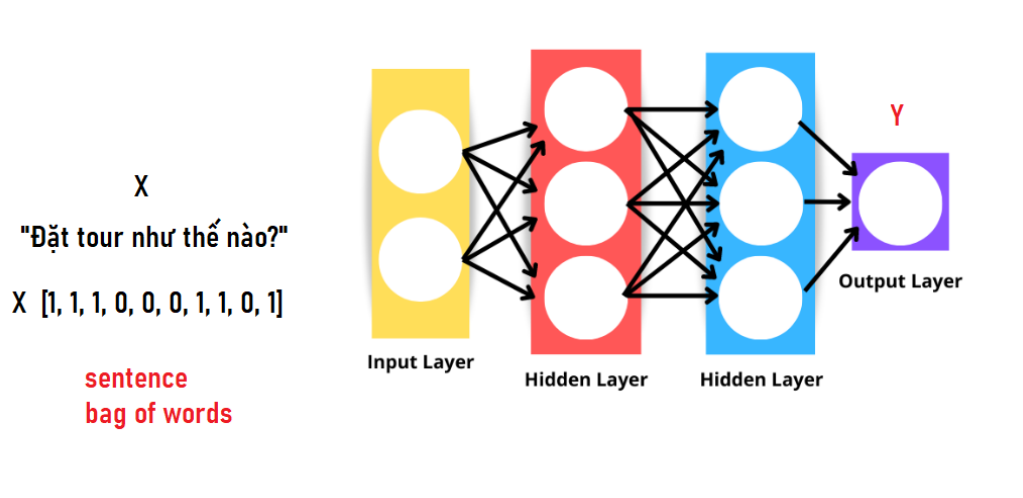
Text

Description automatically generatedText

Description automatically generated

* Pytorch chính là một framework hỗ trợ Deep Learning được phát triển bởi Facebook. Đây là package sử dụng các thư viện của CUDA và C/C++ hỗ trợ các tính toán trên GPU nhằm gia tăng tốc độ xử lý của mô hình. 2 mục tiêu chủ đạo của package này hướng tới là:
  + Thay thế kiến trúc của numpy để tính toán được trên GPU.
  + Deep learning platform cung cấp các xử lý tốc độ và linh hoạt.Text

    Description automatically generated
* Dataset là một tập hợp dữ liệu. Nói cách khác, dataset tương ứng với nội dung của một bảng cơ sở dữ liệu hoặc một ma trận dữ liệu thống kê, trong đó mỗi cột của bảng đại diện cho một biến cụ thể và mỗi hàng tương ứng với một thành viên nhất định của tập dữ liệu được đề cập. Trong các dự án Máy học, chúng ta cần một tập dữ liệu đào tạo. Đây là tập dữ liệu thực tế được sử dụng để huấn luyện mô hình thực hiện các hành động khác nhau.
* Neural Network với Pytorch: Pytorch hỗ trợ thư viện torch.nn để xây dựng neural network. Nó bao gồm các khối cần thiết để xây dựng nên 1 mạng neural network hoàn chỉnh. Mỗi layer trong mạng gọi là một module và được kế thừa từ nn.Module. Mỗi module sẽ có thuộc tính Parameter
  + Input layer: Dữ liệu đầu vào
  + Hide layer: Các lớp nằm giữa lớp đầu vào và lớp đầu ra được gọi là lớp ẩn
  + Output layer: Lớp đầu ra được sử dụng để đưa ra dự đoán



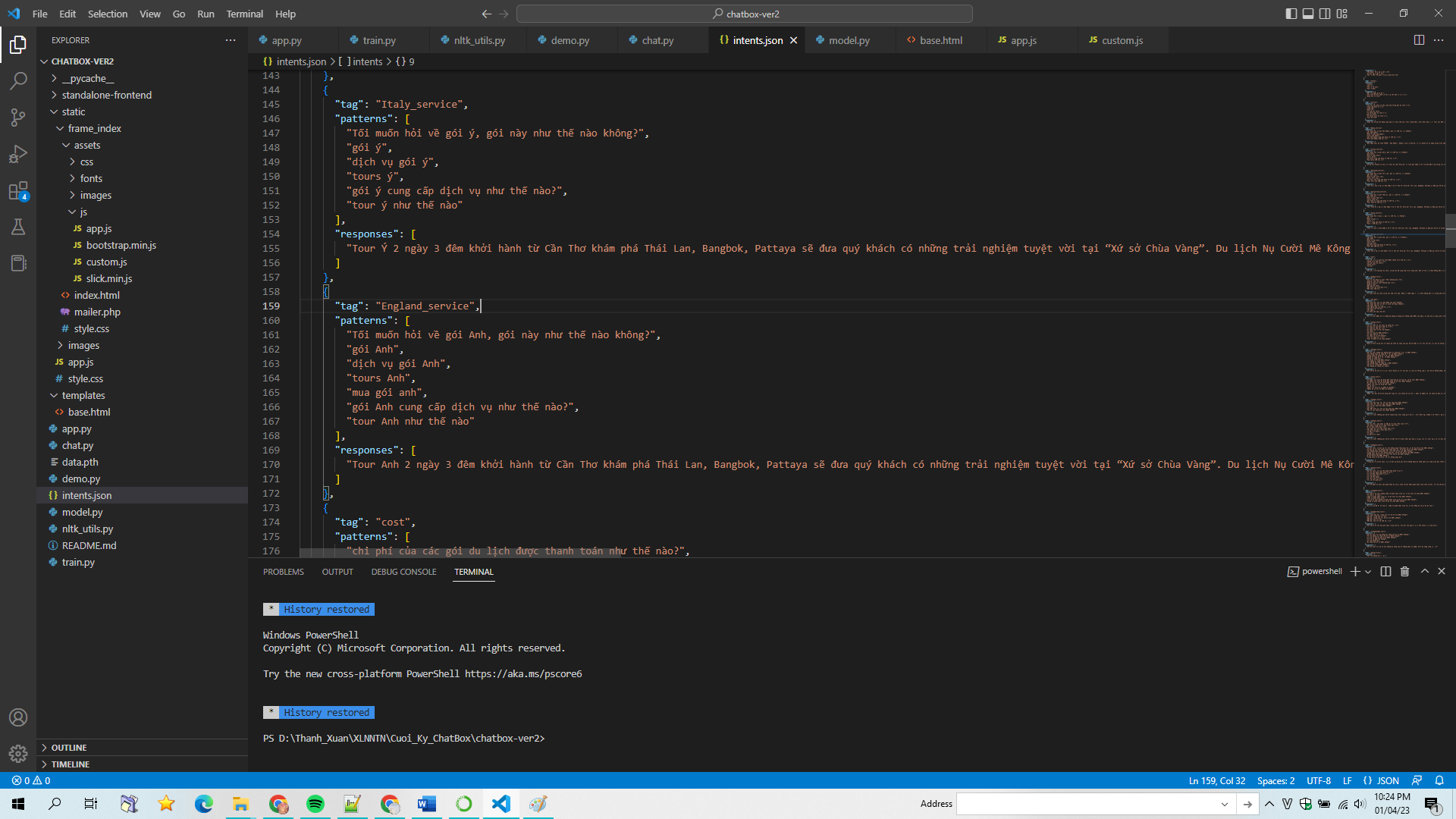
A screenshot of a computer

Description automatically generated with medium confidence

* Link Github mã nguồn nhóm tham khảo: <https://github.com/patrickloeber/pytorch-chatbot>

1. **Mô tả ngắn gọn dữ liệu: Số lượng, File dữ liệu nằm ở đâu, cấu trúc dữ liệu. Nếu dữ liệu lấy từ các nguồn có sẵn thì cung cấp Link.**

* Số lượng câu hỏi: với đề đề tài xây dựng hệ thống chatbot tự động, nhóm phải tạo thu thập nhiều câu hỏi và câu trả lời về một chủ đề tạo dữ liệu một cách thủ công. Số lượng câu hỏi mà nhóm thu thập là 29 tag (nhóm câu hỏi), và trong mỗi tag có tầm 7 đến 8 câu
* File dữ liệu: đây là vị trí lưu dữ liệu



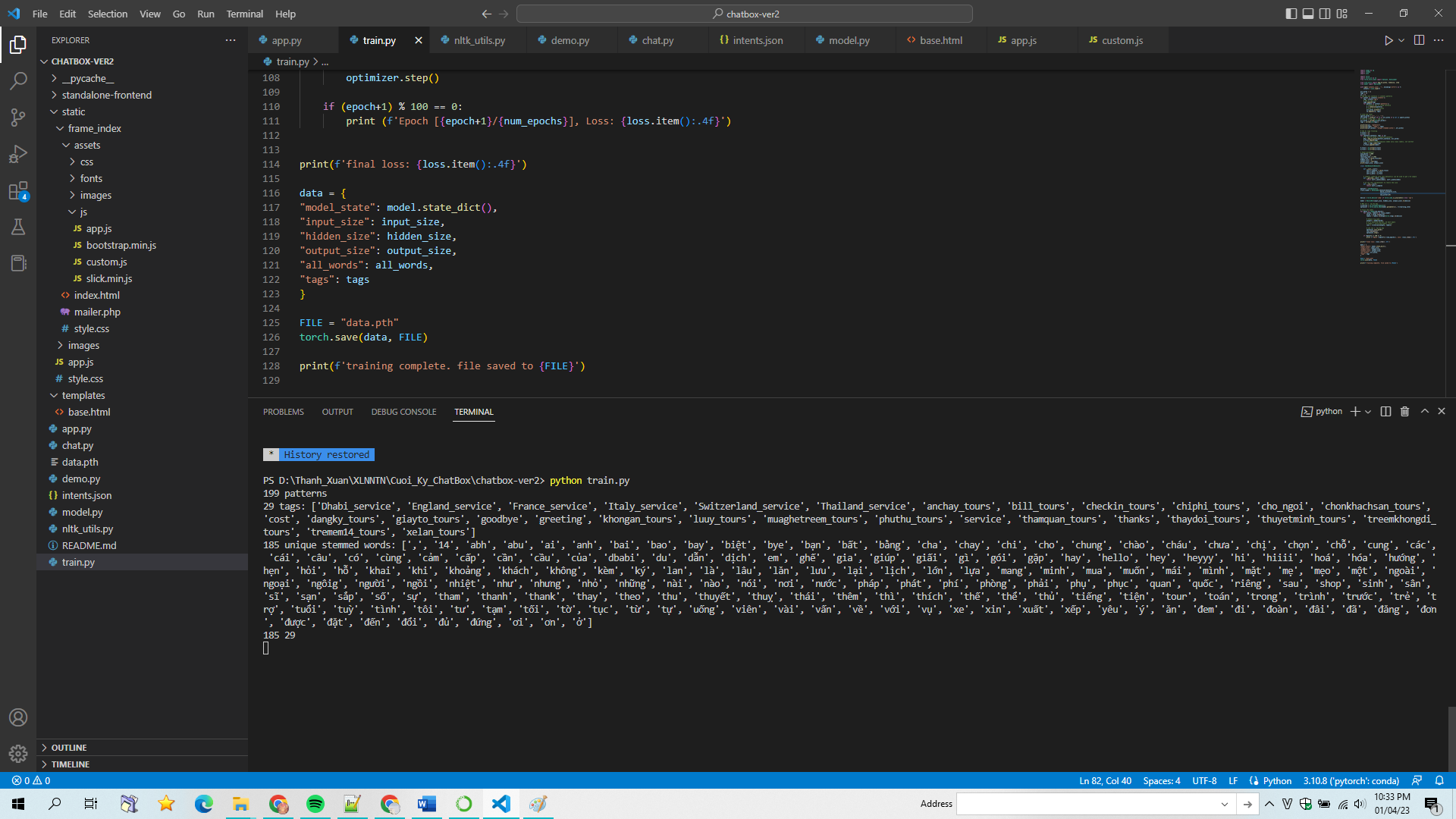
* Cấu trúc như sau:
  + tag: tên của nhóm câu hỏi
  + patterns: mẫu câu mà người dùng nhập
  + responses: phần trả lời

Text

Description automatically generated

1. **Các bước cài đặt và chạy chương trình (cả phần huấn luyện, dự đoán)**

* Cài đặt môi trường:
  + pip install nltk
  + $ python
  + >>> import nltk
  + >>> nltk.download('punkt')
* Chạy chương trình: chạy file train.py để training dữ liệu. Và đây là kết quả gồm có: 29 tag, và túi từ đê dự đoán kết quả gồm 185 từ gốc độc đáo được phân tích từ gốc.



A screenshot of a computer

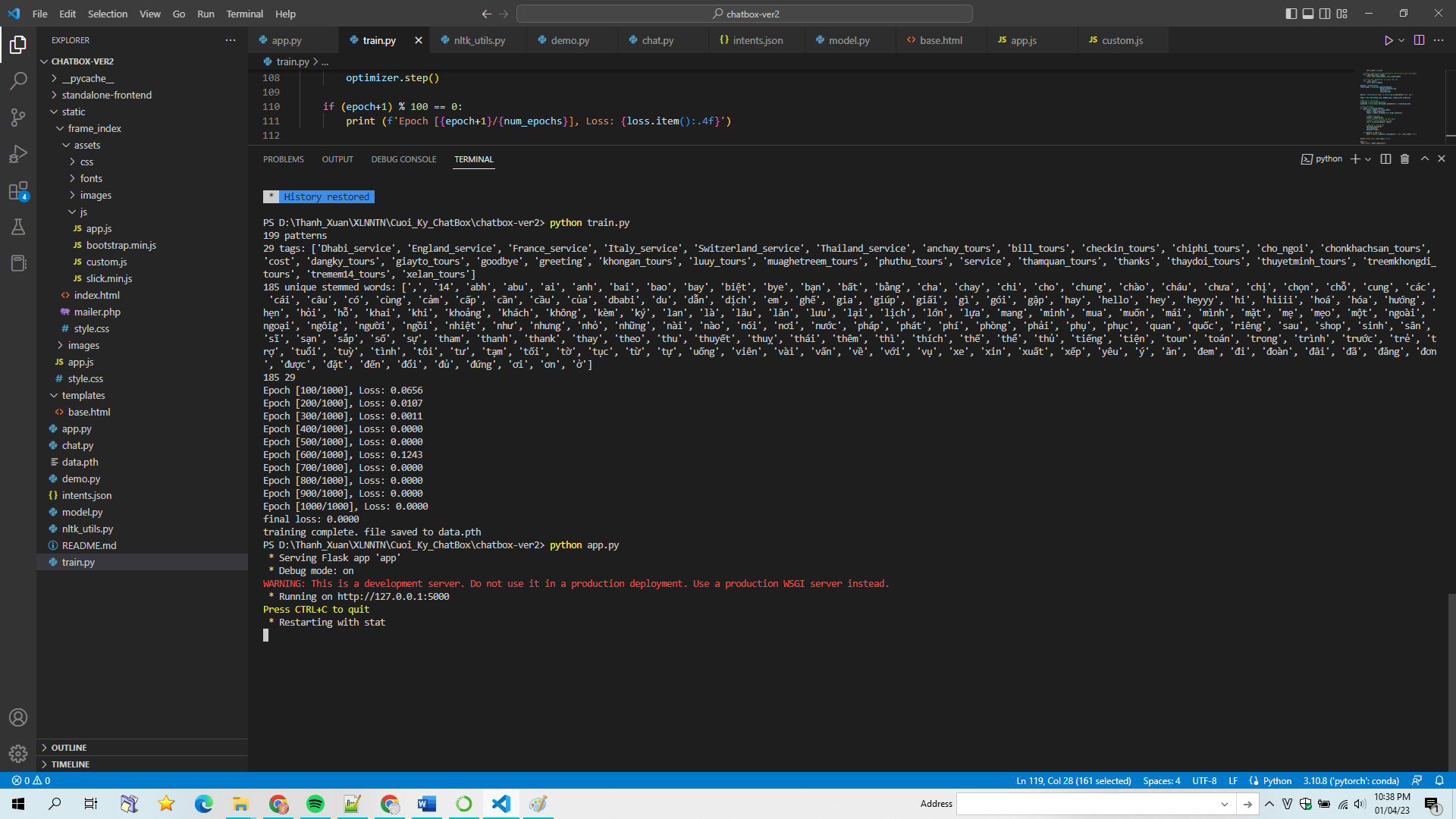
Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Một số hình ảnh tiêu biểu khi chạy chương trình (chụp lại màn hình, chỉ cần 2 hình hoặc có thể nhiều hơn)**

* Sau khi chạy file train.py
* Giờ thì chạy file app.py



* Giao diện chính

Graphical user interface, website

Description automatically generated

* Phần giải đáp thắc mắc, một số câu hỏi phổ biến

Graphical user interface, application

Description automatically generated

* Khi nhấp vào bóng chat

Graphical user interface, application

Description automatically generated

* Khi trò chuyện

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated