Ikhlaq Ahmad & Aloksai Choudari

Ixa190000 & axc190063

Dr. Karen Mazidi

CS 4395

April 15, 2023

Chatbot: Analysis Report

**Overview:**

The chatbot is a program written in python that utilizes some of the natural language processing techniques to determine a near accurate response to the user query. This program is primarily trained on a specific domain of knowledge with some basic and general conversation data. In this case, the data is primarily about the computer and security based; thus, it can only respond to those requests. Otherwise, it will simply respond with a predetermined prompt of not knowing the context.

**System Description:**

The system utilizes json data file to train the model. The file is read and stored for the data processing. The data processing removes the stop words, prefixes and/or suffixes, and any punctuation. Also, the data is lower cased and tokenized. After the data is processed, the Natural Language Processing techniques, TF-IDF (Term Frequency-Inverse Document Frequency) and Cosine similarities are used to precisely evaluate the user query.

TF-IDF (Term Frequency-Inverse Document Frequency) is a commonly used technique in Natural Language Processing (NLP) for determining the importance of a word in a document or a corpus. The technique considers two factors:

* Term Frequency (TF): which measures how often a term (i.e., a word or a group of words) appears in a document. The more often a term appears in a document, the higher its TF value will be.
* Inverse Document Frequency (IDF): which measures how important a term is across the entire corpus. The rarer the term is across the corpus, the higher its IDF value will be.

The TF-IDF score for a term in a document is calculated by multiplying the term frequency (TF) by the inverse document frequency (IDF) for that term across the entire corpus.

**TF-IDF = TF (term, document) x IDF (term)**, where TF (term, document) is the term frequency of the term in the document, and IDF (term) is the inverse document frequency of the term across the entire corpus. The formula for calculating the term frequency of a term in a document is:

**TF(term, document) = (Number of times the term appears in the document) / (Total number of terms in the document)**

The formula for calculating the inverse document frequency of a term across the entire corpus is:

**IDF (term) = log\_e(Total number of documents / Number of documents containing the term), where log\_e is the natural logarithm.**

The resulting score gives a measure of how relevant a term is to a particular document.

The other technique that is used in this program is Cosine Similarities. Cosine similarity is a measure of similarity between two non-zero vectors in a high-dimensional space. In NLP, cosine similarity is used to measure the similarity between two documents or between a document and a query. The documents and the query are represented as vectors in a high-dimensional space, where each dimension corresponds to a term in the vocabulary. The value of each dimension represents the importance of that term in the document or query, typically using TF-IDF weighting.

The cosine similarity between two vectors is calculated as the cosine of the angle between them. It is a value between -1 and 1, where 1 indicates that the two vectors are identical, 0 indicates that they are orthogonal (i.e., unrelated), and -1 indicates that they are diametrically opposed.

**System Structure:**

The program contains:

* Train.py
* Bot.py
* Users.json
* Training.json
* cosine\_similarites.npz
* user\_history (Directory)
  + <username>.json

**Train.py:**

This file contains all the logic used to train the chatbot model. The model uses the following libraries/dependencies:

* Json
* Spicy
* Sklearn
* NLTK
* String

The program starts by reading in the Training.json file. It contains several hundred json data in this format:

{

**"tag":** "tcp\_ip\_reference\_model",

**"patterns":** ["What is the TCP/IP reference model?", "Explain TCP/IP reference model in computer networks."],

**"responses":** ["The TCP/IP (Transmission Control Protocol/Internet Protocol) reference model … between devices connected to the Internet."]

},

Here, tag is the keyword; patterns are to match the similarities with the user request; and response is a predetermined answer for the matching query.

The TF-IDF and Cosine similarity is calculated, and a file is made and saved as an npz file

**Bot.py:**

Here, the trained cosine file is imported as a module from Train.py and stored. Then, user input is taken and processed in the similar way using TF-IDF to match the most accurate pattern in the json file. For multiple possible responses, randomization is used to give a random but a relevant response. If no similar response is found, a predetermined general response is sent. Furthermore, a user is authenticated using any name and password with no restriction. If the user and password combinations are not present, the user profile will be created in the user\_history directory as a json file. Users’ full chat history, including the name and other relevant information is stored as a json file in that directory.

**Users.json and Training.json:**

These files contain users’ chat history and train data of over 1000 lines respectively.

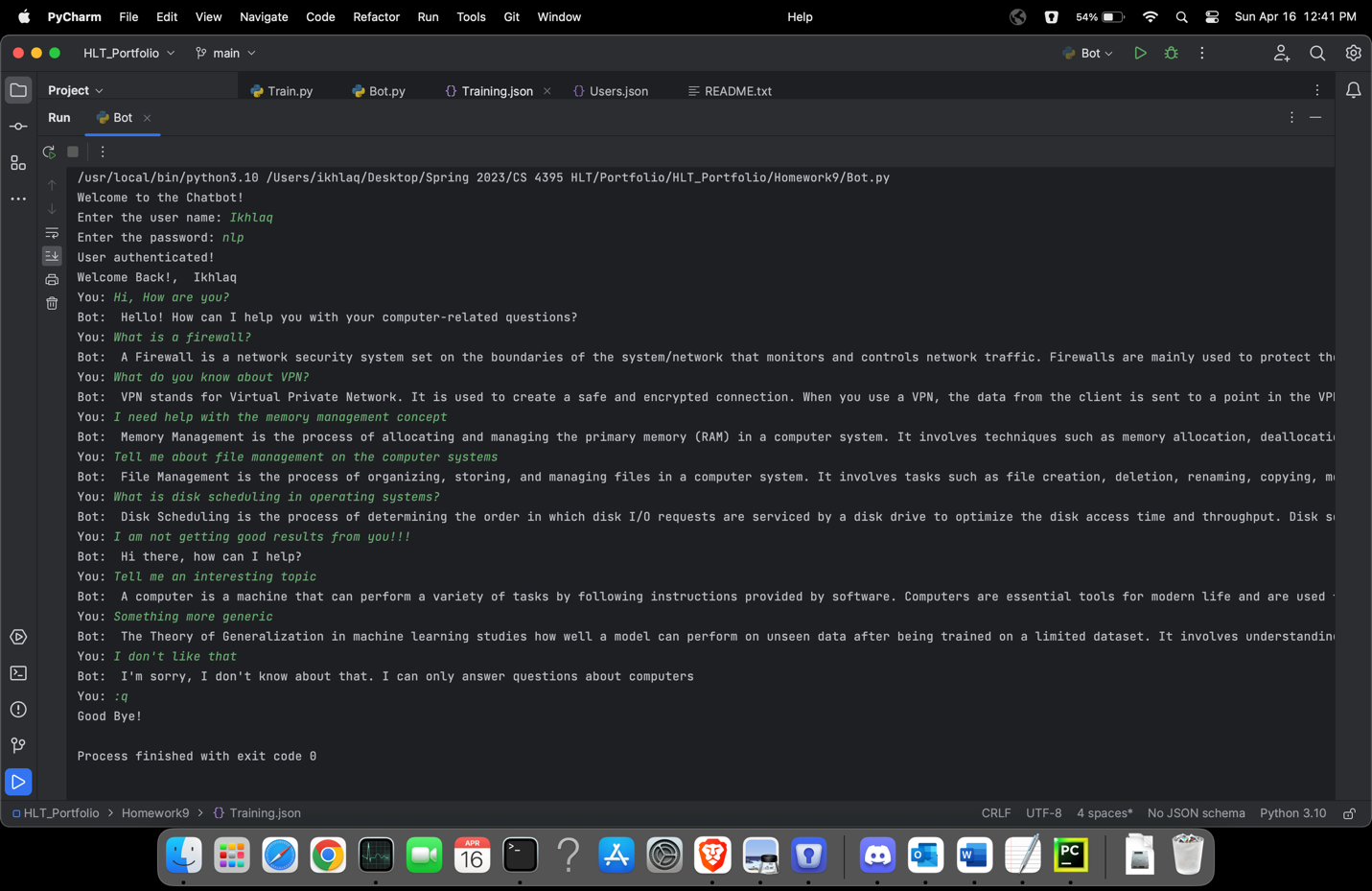
**cosine\_similarities.npz**

It is the trained file created by Train.py

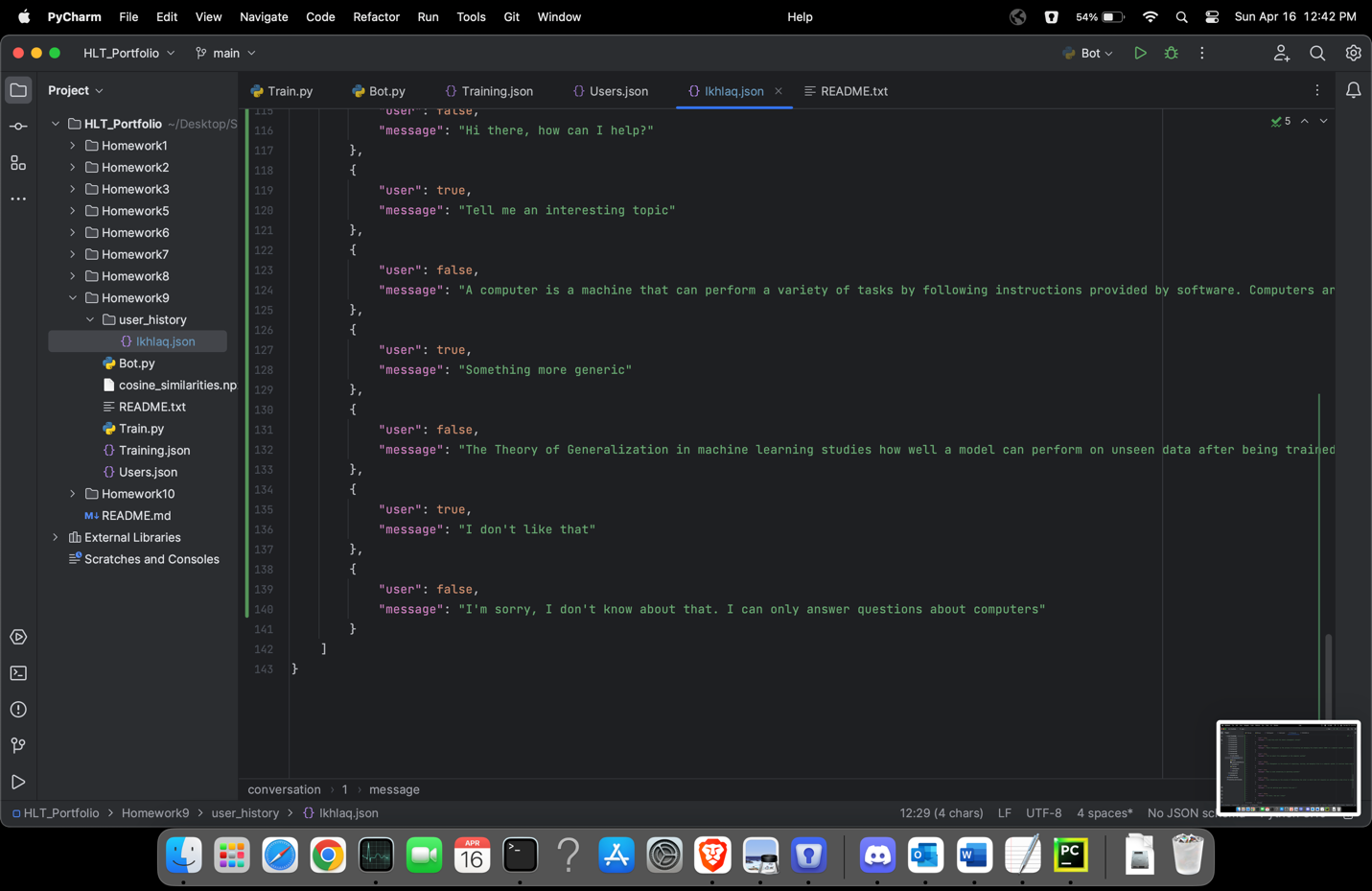
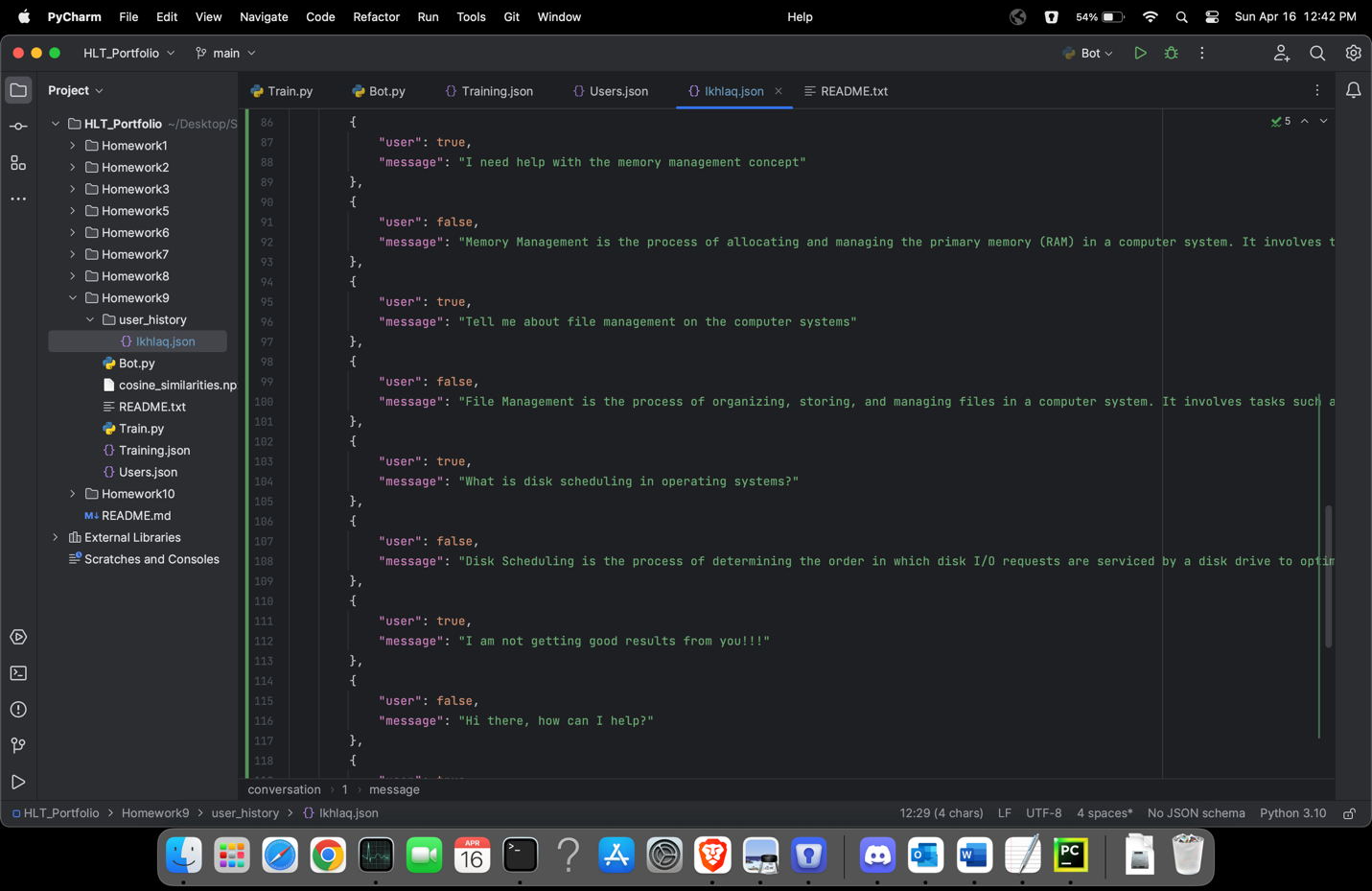
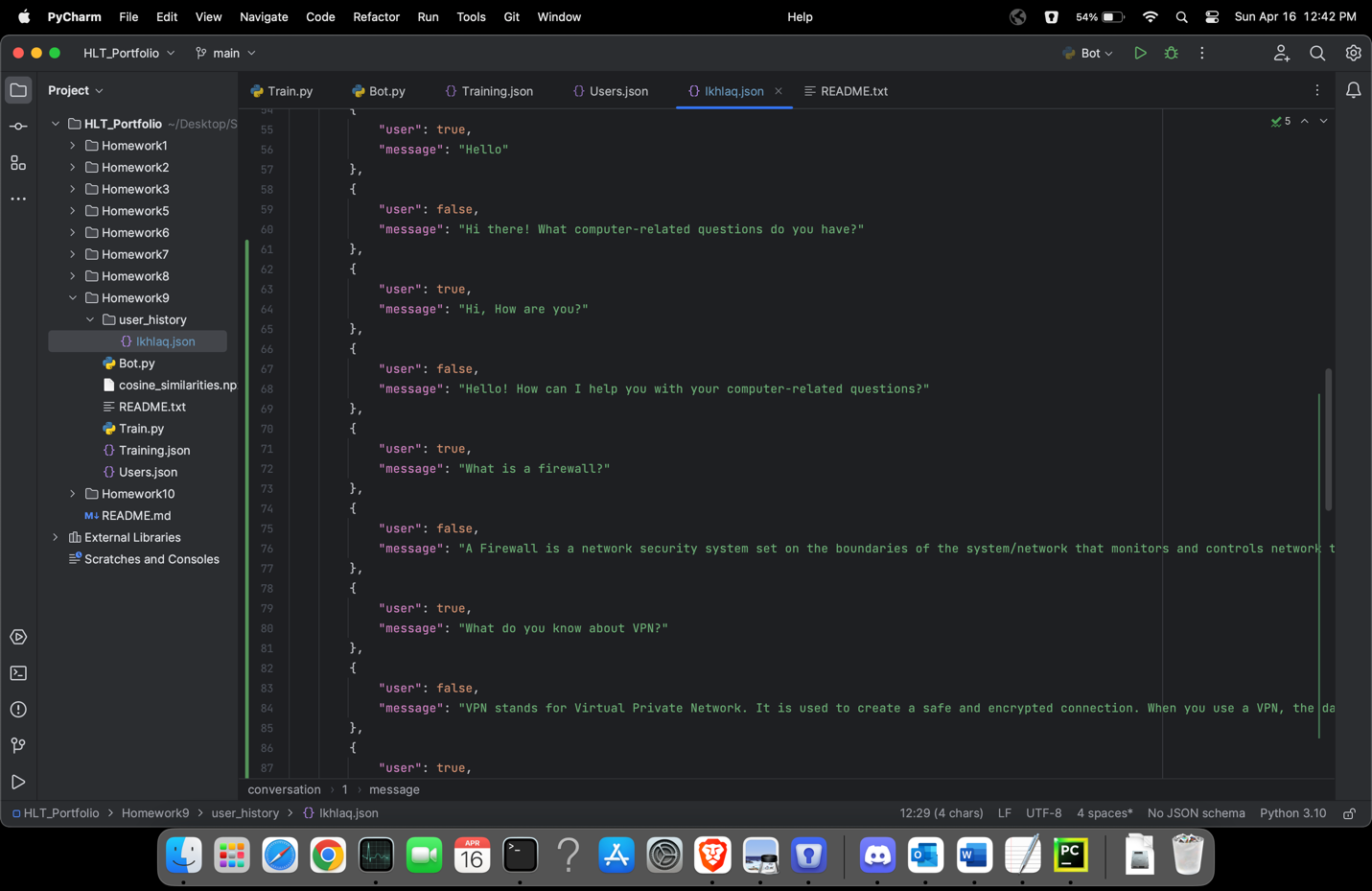
**Screenshots:**

Starting from the next page.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

These screenshots contain sample interaction with the bot and the saved model.

**Logic Diagram:**

Diagram

Description automatically generated

**Appendix of the knowledge base:**

Use these keywords in the query.

1. Error
2. Abstraction
3. Documentation
4. Testing
5. Datastore
6. Dynamic
7. Dll
8. Rnn (Recurrent Neural Networks)
9. Hyperparameters
10. API
11. Encryption
12. Virtual Reality
13. Database
14. Programming
15. Natural language processing
16. DevOps
17. Truth tables
18. Multi-dimensional Arrays
19. Recursion
20. Quicksort
21. Mealy and Moore machines

**Appendix of the sample user interaction model:**

"id": "Ikhlaq",

"name": "",

"age": 0,

"conversation": [

{

"user": true,

"message": "Hi, How are you?"

},

{

"user": false,

"message": "Hello! How can I help you with your computer-related questions?"

},

{

"user": true,

"message": "What is a firewall?"

},

{

"user": false,

"message": "A Firewall is a network security system set on the boundaries of the system/network that monitors and controls network traffic. Firewalls are mainly used to protect the system/network from viruses, worms, malware, etc. Firewalls can also be to prevent remote access and content filtering."

},

{

"user": true,

"message": "What do you know about VPN?"

},

{

"user": false,

"message": "VPN stands for Virtual Private Network. It is used to create a safe and encrypted connection. When you use a VPN, the data from the client is sent to a point in the VPN where it is encrypted and then sent through the internet to another point. At this point, the data is decrypted and sent to the server. When the server sends a response, the response is sent to a point in the VPN where it is encrypted and this encrypted data is sent to another point in the VPN where it is decrypted. And finally, the decrypted data is sent to the client. The whole point of using a VPN is to ensure encrypted data transfer."

},

{

"user": true,

"message": "I need help with the memory management concept"

},

{

"user": false,

"message": "Memory Management is the process of allocating and managing the primary memory (RAM) in a computer system. It involves techniques such as memory allocation, deallocation, and swapping to efficiently utilize the available memory for executing processes or threads, managing memory fragmentation, and providing memory protection to prevent unauthorized access."

},

{

"user": true,

"message": "Tell me about file management on the computer systems"

},

{

"user": false,

"message": "File Management is the process of organizing, storing, and managing files in a computer system. It involves tasks such as file creation, deletion, renaming, copying, moving, and organizing files into directories or folders. File management also includes file access permissions, file sharing, and file system integrity and consistency checking to ensure reliable and secure storage and retrieval of data."

},

{

"user": true,

"message": "What is disk scheduling in operating systems?"

},

{

"user": false,

"message": "Disk Scheduling is the process of determining the order in which disk I/O requests are serviced by a disk drive to optimize the disk access time and throughput. Disk scheduling algorithms determine the most efficient way to access data on a disk by reducing the seek time, rotational latency, and head movement. Common disk scheduling algorithms include First-Come, First-Served (FCFS), Shortest Seek Time First (SSTF), SCAN, C-SCAN, and LOOK, each with its own advantages, disadvantages, and suitability for different types of disk workloads."

},

{

"user": true,

"message": "I am not getting good results from you!!!"

},

{

"user": false,

"message": "Hi there, how can I help?"

},

]

**Analysis:**

This chatbot is the simple and basic approach to utilize some of the natural language processing techniques. It gives an accurate response is a query contains at least a keyword or something similar. This model, however, is trained on a very specific and small data. Therefore, it can only respond with something from the training model if it finds a similar textual query. On the flip side, real world chatbot are extremely complex and utilize a tremendous resources like computing power, teams of expert engineers, and huge number of libraries.

For this chatbot program, the downside is that it solely defendant on the raw data for its responses. If more time was available, the chatbot could be made with other techniques to train the model on dynamic data with an active internet connection. Also, several other techniques could be helpful in the response generation like neural networks, probabilistic models, and deep learning techniques etc. Overall, creating this chatbot from the scratch was an excellent opportunity to learn and hone my NLP skills.