

HLT – Chatbot Project

Report

System description

This chatbot can answer questions about well-known bands in music history, such as who their members are, their origin, and the years they have been active. It also saves some of the user's data, including their name and whether they like a band or not. The chatbot reads band names out of full sentences based on keyboards and previously trained English pipelines and scrapes data from Wikipedia to answer the user's questions.

NLP techniques used:

I have used multiple NLP techniques that we have learned over the course of the semester in this program. These include:

- Pickles

Used to save user's data to store information about previous conversations, including names and likes or dislikes. This is useful so that a conversation can be picked up where it was left off last time, as seen in current chatbots, such as ChatGPT, which saves old user data for future conversations.

- Regular expressions

These are used mostly to format data so that specific functions can successfully read input. An example is extracting certain words from sentences.

- Sentiment analysis

This was used in the making of the chatbot to check if a user likes or dislikes a certain band. This is helpful to keep information about users to future uses of the chatbot.

- BeautifulSoup

This library was used to scrape data from the web. This is probably the most important part of the program, as well as the most useful one. Any data about any band can be pulled from Wikipedia using this, giving the user the option to ask about millions of bands.

- Requests

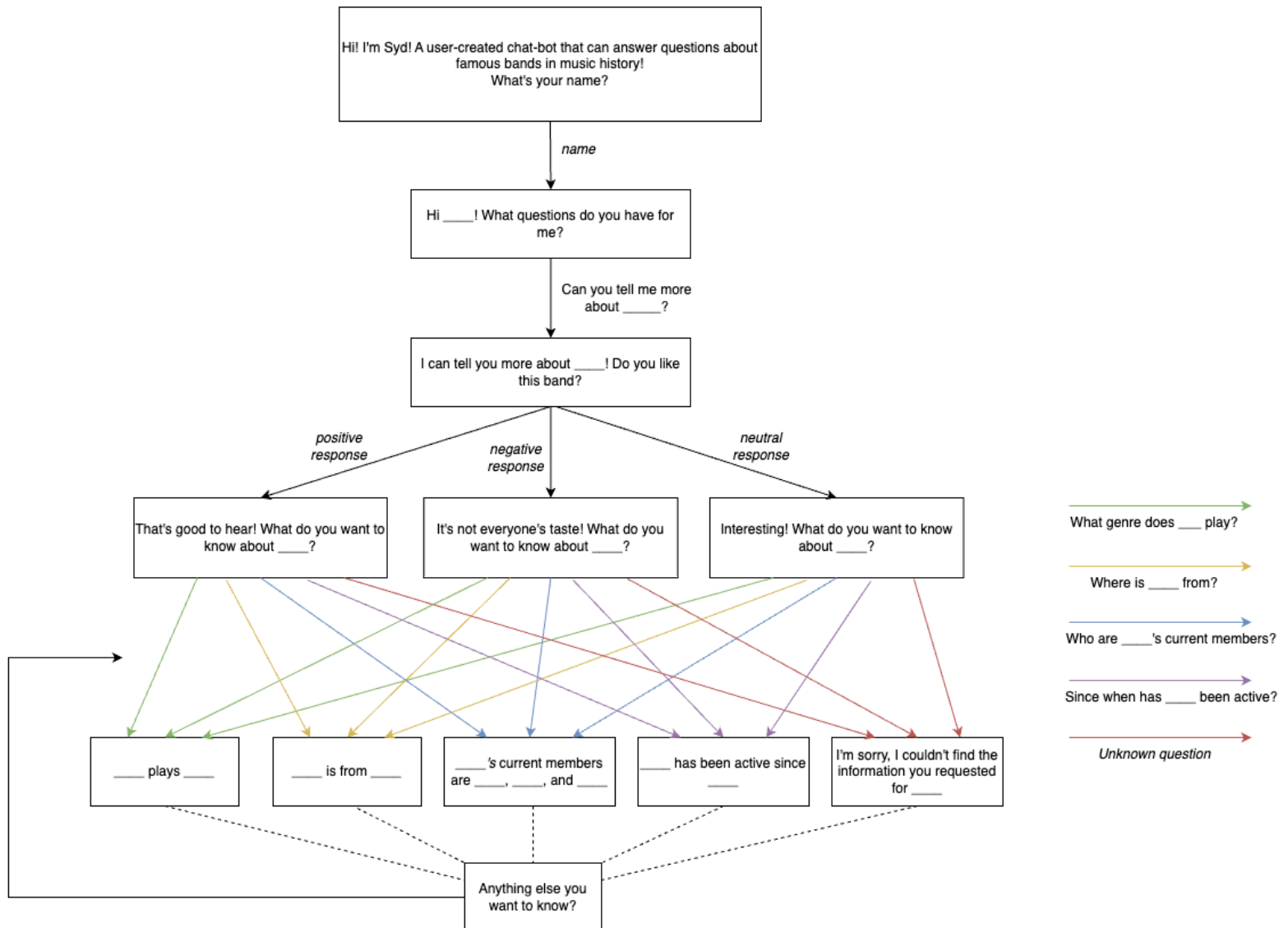
This is also used to pull data from the web, as previously used in the WebScraper project. Specifically, to get a URL used to scrape data.

- Text processing

Basic built-in python string conversion functions, such as “replace”, “join”, or “lower” are used to process text, as previously used in most python projects.

Without these simple tools, it would be hard to process the user’s input and pull any information out.

Dialog Tree



Sample dialog interaction

Sample 1:

Syd: Hi! I'm Syd! A user-created chat-bot that can answer questions about famous bands in music history!

Syd: What's your name?

You: Hi Syd! I'm John!

Syd: Hi John! What questions do you have for me?

You: Can you tell me more about the band Led Zeppelin?

Syd: I can tell you more about Led Zeppelin! Do you like this band?

You: Yes! I love them.

Syd: That's good to hear! What do you want to know about Led Zeppelin?

You: Where are they from?

Syd: Led Zeppelin is from London, England.

Syd: Anything else you want to know?

You: Since when have they been active?

Syd: Led Zeppelin was active from 1968 to 1980.

Syd: Anything else you want to know?

Sample 2:

Syd: Hi! I'm Syd! A user-created chat-bot that can answer questions about famous bands in music history!

Syd: What's your name?

You: Hi! My name is Alex

Syd: Hi Alex! What questions do you have for me?

You: Tell me more about the band The Beatles

Syd: I can tell you more about The Beatles! Do you like this band?

You: No, not really

Syd: It's not everyone's taste! What do you want to know about Led Zeppelin?

You: Who are their members?

Syd: The Beatles's current/latest members are John Lennon, Paul McCartney, George Harrison, Ringo Starr, and (see Personnel section for others).

Syd: Anything else you want to know?

You: Where are they from?

Syd: The Beatles are from Liverpool, England.

Syd: Anything else you want to know?

Appendix for knowledge base/live lookup

The program reads the band name by checking every token in the SpaCy “en_core_web_sm” trained pipeline if it is a name from a band. It then saves it in the dictionary of bands. With the band name, it looks up the Wikipedia page using:

```
url = f"https://en.wikipedia.org/wiki/{context['bands'][band.lower()]}"
```

It then scrapes data from this website to get information, such as the band’s origin, members, or active years. It uses the BeautifulSoup-library’s html-parser and looks for the infobox containing all the information about the band.

This gives the user a higher chance to find useful knowledge, compared to a static knowledge base containing information about just a single topic. I could have potentially broadened this program to be able to answer questions about not just bands, but also movies or TV shows, since they have similar Wikipedia pages.

Appendix for sample user models created:

The program saves the user’s name, as well as for each band if they liked it or not. This information is saved in a python dictionary and then exported as a pickle. For future usage, this can be very useful, when more information of users is necessary to make more precise decisions. In the current state of this chatbot, it’s not as important yet, since we don’t need as much background information to hold basic conversations. It would be useful to improve the program by evaluating previous answers and connect machine learning algorithms to actually build a decent chatbot.

Evaluation

This is obviously not in any way comparable to current chatbots, like ChatGPT. For a normal class project, this gave decent results, given the expectations. The strengths include the live web scraping, making the range of information in the given domain of topics almost impossible. And as previously mentioned, it is easily improvable to be able to talk about other subjects, such as movies or TV shows. Given the somewhat low expectations, it is most definitely more effective than building a simple static knowledge base, as we have done in a previous project, since this not only gives a limited domain of things to talk about, but also limited information in this domain with predefined answers. This project was scrapped and restarted multiple times. If I could have decided on a topic right away, I would have had more time to implement a wider domain of questions to answer and even possibly open it up to different topics, not just music. The program does require keywords to be mentioned to even come to a solution, but it is good at finding not only single words, but also combinations of multiple words in a sentence to extract information. Overall, as mentioned, it is absolutely nothing compared to some of the new chatbots, but given the low expectations, it is not as bad as expected when this project was started. It is possible to hold infinitely long conversations with a variety of things to talk about.