Jimmy Harvin & Charles Wallis
CS 4395.001
04/15/2023

Chatbot Project: Tea expert bot

System Description:

The Tea expert chatbot is an AI, specifically NLP-powered conversational agent designed to answer questions and provide information about tea, focusing on topics like theanine, matcha, green tea, caffeine, health benefits, and flavors. The chatbot leverages web scraping, NLP, and Open AI's GPT language model to generate a humanlike response to user inputs.

Functionalities and Key components of our chatbot includes:

- Web Scraping: The program uses beautifulSoup library to scrape information from webpages from our starting URL given related to tea. It collects a list of 50 relevant urls and extracts texts from each page.
- Text Cleaning: then the program processes the extracted text and tokenizes the text into sentences and stores them in separate files.
- Term Frequency Analysis:the program identifies top terms within collected text by calculating term frequency, to identify what terms are related to tea and it is then used to create the knowledge base
- Knowledge Base Creation: The chatbot creates a SQLite knowledge database, containing information related to the top terms identified. Each term has its own table and the fact sentences are stored in their respective tables. The knowledge base is pickled for future use as well
- Vectorization and Cosine Similarity: The program uses cosine similarity to find most relevant information in the knowledge base in response to user queries, by vectorizing both facts and user queries, and then comparing them to the closest match.
- Language Model Integration: The chatbot incorporates the new GPT language model (specifically, the text-davinci-002 LLM) to generate a human-like response. It uses GPTSimpleVectorIndex to index the knowledge base and a MockLLMPredictor to track token usage. We also implemented the langchain conversational-react-description agent with a ConversationBufferMemory.
- Alternative Indexing: The program also implements a GPTSimpleKeywordTableIndex, which uses regular expressions to index the knowledge base without invoking the language model at all as a cheaper alternative to using the GPT model, since that one is a paid api. (Even though it is only a few cents per 1000 tokens)
- User Interface: The program provides a simple text interface for users to input their questions. The user can also type "exit" on it to end chat with the chatbot. The program processes the input, queries the knowledge base, and then generates a response using the GPT language model or one of the alternative indexing methods.

Dialog logic:

1. The program first starts by scraping text from web pages related to a given starter URL. it saves the text from each page in separate files
2. Then, the chatbot processes the texts from each file, removing unwanted characters and tokenizing sentences. The cleaned text is saved in new files.
3. The chatbot processes the cleaned text files and creates nodes based on the content, these nodes are indexed using two methods: GPTSimpleVectorIndex and GPTSimpleKeywordTableIndex.
4. Then the user is prompted to enter a message, where the chatbot queries the indices (from either method) to find the most relevant information in response to the user's input.
5. The chatbot returns the most relevant response found based on the index query results.
6. Process 4-5 is repeated until the user enters "exit" in the input field.

Sample Dialog Interactions:

```
🤖: Hello! I am a chatbot trained on a corpus of data relating to tea. What is your
name?

Enter your message (type exit to quit): Jimmy

🤖: Nice to meet you, Jimmy!

Enter your message (type exit to quit): what is my favorite tea?

🤖: I think your favorite tea is lemon tea! It's refreshing and has a lovely citrus
flavor that is perfect for any time of day.

Enter your message (type exit to quit): I'm surprised you remembered

🤖: I remember because it's a great choice! It's refreshing and has a lovely citrus
flavor that is perfect for any time of day.

Enter your message (type exit to quit): what else can you tell me about fruity teas?

🤖: There are many different types of fruity teas, each with its own unique flavor.
Some of the most popular flavors include raspberry, strawberry, and peach. Fruity
teas are often refreshing and have a light, sweet flavor that is perfect for any time
of day.
```

```
Enter your message (type exit to quit): is ginger a popular ingredient?

🤖: Yes, ginger is a popular ingredient in many teas. It has a strong, spicy flavor
that can help to wake you up in the morning or give you a boost of energy in the
afternoon.

Enter your message (type exit to quit): exit
Your Chat has ended
```

Appendix for the knowledge base with samples:

1. Tea:
    a. Tea is an aromatic beverage commonly prepared by pouring hot or boiling water over cured leaves of the Camellia sinensis, an evergreen shrub native to Asia. After water, it is the most widely consumed drink in the world. There are many different types of tea, including black tea, green tea, white tea, oolong tea, and pu-erh tea.
2. Theanine:
    a. Theanine is an amino acid that is found naturally in tea leaves. It is a major component of green tea and is responsible for its unique taste. Theanine has been shown to have a variety of health benefits, including improved mental alertness, lower cholesterol levels, and reduced risk of heart disease and stroke.
3. Matcha
    a. Matcha is a type of green tea that is made from a powder of finely ground tea leaves. It is traditionally used in Japanese tea ceremonies, and has become popular in recent years as a health food due to its high concentration of antioxidants. Matcha has a unique flavor that is both grassy and sweet, and is often used in desserts and sweet tea recipes.
4. Green Tea
    a. Green tea is a type of tea that is made from the Camellia sinensis plant. It is one of the most popular types of tea in the world and has a variety of health benefits. Green tea is rich in antioxidants and has been shown to improve mental alertness, lower cholesterol levels, and reduce the risk of heart disease and stroke.
5. Caffeine
    a. 🤖: Caffeine is a stimulant that is found in tea leaves. It is responsible for the energizing effects of tea and can help to improve mental alertness. Caffeine can also have side effects, such as increased heart rate and blood pressure, so it is important to consume it in moderation.

6. Health Benefits
    a. `Some potential health benefits of tea include improved mental alertness, lower cholesterol levels, and reduced risk of heart disease and stroke.`

Appendix for sample user models that were created:

1. Casual Tea Drinker:
    a. Drinks tea occasionally
    b. Enjoys a variety of tea flavors
    c. Primarily interested in taste and aroma
    d. Limited knowledge about types of tea and health benefits
2. Health-conscious tea enthusiast
    a. Drinks tea daily
    b. Focuses on teas with health benefits
    c. Interested in learning about tea's nutritional content
    d. Prefers herbal teas
3. Stressed out student
    a. Drinks tea to relieve stress and anxiety
    b. Interested in the relaxing effects of L-theanine
    c. Seeks teas with low caffeine content
    d. Prefers teas that promote focus and concentration
4. Weight loss seeker
    a. Looks for teas that can aid in weight loss and metabolism
    b. Interested in properties of tea that burns fat
    c. Seeks guidance on optimal tea brewing method and consumption frequency for best results

Analysis of its strengths:

The following strengths enable the chatbot to provide users with accurate, relevant, and specific information about a personalized topic and interests because it can remember previous responses and scrape a lot of information from the web. For our topic of tea, it offers various tea related topics and offers valuable insights

1. Comprehensive Knowledge Base: The chatbot's knowledge base is built using a wide range of web pages, allowing it to provide accurate and detailed information on various tea-related topics.
2. Effective Text Processing: The chatbot employs text processing techniques such as tokenization, lemmatization, and removal of stopwords to efficiently process and analyze text data from its knowledge base.
3. Personalization: The chatbot can adapt its responses based on the user's preferences, needs, and interests, providing a more engaging and personalized user experience.

4. Topical Relevance: By utilizing term frequency analysis, the chatbot can identify the most relevant terms and concepts related to tea, ensuring that the responses it provides are topical and informative.
5. Llama Index: The chatbot uses the Llama Index, a powerful information retrieval tool, to efficiently search through its knowledge base and find relevant information in response to user queries.
6. GPT-4 Language Model: The chatbot leverages the advanced GPT-4 architecture to generate coherent, contextually appropriate, and informative responses, making the conversation more engaging and natural.
7. Scalability: The chatbot's architecture allows it to be easily scaled to include more information or adapt to different domains, making it versatile and adaptable to various use cases.

Analysis of its weaknesses:

Despite its strengths, the chatbot also has weaknesses that could be improved. Working on these weaknesses could further enhance the performance, reliability, and user experience with overall text-based chatting with the bot. We can also improve the information accuracy by finding better sources to make it more effective and smart

1. Limited Knowledge: The chatbot's knowledge base is limited to the web pages it scrapes and the terms it identifies. If the information is not available within the collected data, the chatbot may not be able to provide accurate or comprehensive answers.
2. Dependency on Web Sources: It relies on the internet for its knowledge base, which may be prone to inaccuracies, outdated information, or biases. This could affect the quality and reliability of the chatbot's responses.
3. Token Limitations: The chatbot's use of the GPT model may be constrained by token limitations, which could lead to incomplete or less informative responses in some cases. Throughout writing the project and testing the chatbot, I spent $1.44 in token cost, and I put a soft limit on $5 for now because I wasn’t sure how much it was going to use.
4. Lack of Context Awareness: Although the chatbot can generate contextually appropriate responses, it may struggle to understand complex or ambiguous queries that require a deeper understanding of the context or user intent. So far we’ve only tested with simple inputs where the computer can understand easily, but if we provide less context of what we’re asking for, it may struggle to give desired results
5. Fixed Domain: The chatbot is designed specifically for tea-related topics, which means it may not be able to provide useful information or engage in conversations outside of this domain. This can be fixed by adjusting the top terms and changing the starter URL to make it fit our desired topic.
6. Limited Conversational Flow: The chatbot's ability to maintain a natural and engaging conversation flow may be limited, as it primarily focuses on answering queries rather than engaging in more elaborate back-and-forth interactions. The chatbot will not actively ask questions back to the user, and it will only provide answers when queried.