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 a 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, text preprocessing, LlamaIndex, LangChain, 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 70 relevant urls and extracts texts from each page.
- Text Cleaning: Afterwards, the program processes the extracted text and tokenizes it into sentences and stores them in separate labeled files.
- Term Frequency Analysis: The program identifies top terms within collected text by calculating term frequency, to identify what terms are related to tea. This was used to verify and evaluate the relevance of collected knowledge during testing.
- Knowledge Base Creation: The chatbot originally created 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, and the knowledge base was pickled for future use. However, a more simple solution was found later; the cleaned text would be stored in documents that would be combined into a JSON file that could be indexed through the LlamaIndex package.
- 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. Originally, this was implemented directly for use in querying from an SQLite database, but this was dropped in favor of the GPTSimpleVectorIndex object from LlamaIndex, a package that would handle vector space modeling and indexing for us.
- 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. The index is wrapped by the LangChain conversational-react-description agent with a ConversationBufferMemory, which tries to simulate a person having a conversation. Memory allows the agent to not only follow the flow of a conversation, but it can also remember simple facts like name and favorites, which are stored in JSON chat logs.

- 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” to end chat with the chatbot and save user data. 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. The inputs and responses are hosted externally using Gradio.

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 web scraper processes the texts from each file, removing unwanted characters and tokenizing sentences. The cleaned text is saved in new files.
3. The chat bot processes the cleaned text files and creates nodes based on the content, these nodes are indexed using two methods: GPTSimpleVectorIndex or GPTSimpleKeywordTableIndex, depending which segments of the notebook are running.
4. Then the user is prompted to enter a message, where the chat bot queries the indices (from either method) to find the most relevant information in response to the user’s input. For the main bot flow, the chat bot will use the vector indexing tool if it identifies a keyword that is relevant in the knowledge base (anything related to tea). Otherwise, no tool is used and the bot tries its best to hold a conversation.
5. The chat bot returns the most relevant response found based on the index query results. If the user requests a source, the bot will return the most relevant link it indexed.
6. Steps 4-5 are repeated until the user enters “exit” in the input field. The chat logs will be saved with the user’s name for further use.

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:

The following are simple sentence samples for terms with high frequencies in the knowledge base. The actual knowledge base itself consists of a large amount of cleaned sentence text files that can be accessed via LlamaIndex using a stored index.json file. A folder containing this knowledge base will be included with the submission.

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:

The user models in question are actually chat logs that can be stored and retrieved based on the name of the user. When starting a conversation, the chat bot will continue to ask for the user's name until it is able to successfully answer the question "What is my name?" queried in code. The bot's response is tokenized and tagged for parts of speech, and the final non-punctuation token will be a proper noun if a name was identified. The program will then check if there is a [name].json file in the directory corresponding to the user's name: if so, chat logs are loaded in and the bot will have access to an existing user model. When the conversation ends, the chat log messages will be converted to a dictionary, which will then be saved as [name].json in the program directory. Three sample user models will be included with the submission, but here is one that corresponds to a student evaluation in the next section.

Tracey.json

[{"type": "ai", "data": {"content": "Hello! I am a chatbot trained on a corpus of data relating to tea. What is your name?", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "Tracey", "additional_kwargs": {}}},

{"type": "ai", "data": {"content": "Nice to meet you, Tracey!", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "What is my name?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "Your name is Tracey.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "Oh nice, I love black tea.", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I think black tea is delicious! It has a rich, complex flavor that is perfect for chilly mornings or afternoons.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "What brand of tea do you buy? I always get PG Tips, but it can be hard to find in American stores.", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I'm not sure which brand of tea is best, but I always get PG Tips. I find it to be a great choice for its rich, complex flavor.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "Do you know which country started making tea?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I believe tea originated in China.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "That's cool, what kind of tea originated from there?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I believe tea originated in China.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "That's cool, what kind of tea originated from China", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I believe green tea originated in China.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "Is there a tea that is better than ginger tea in terms of health benefits?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "There is no one tea that is better than another in terms of health benefits. Each type of tea has its own unique set of health benefits, so it is important to choose the one that is best for you.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "What kind of tea would you recommend for relieving stress?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I would recommend chamomile tea for relieving stress. Chamomile has a calming effect that can help to reduce stress and anxiety.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "Would you know any other drinks besides tea that is good for dealing with stress and anxiety?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I would recommend chamomile tea for relieving stress. Chamomile has a calming effect that can help to reduce stress and anxiety.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "Wow that's cool! What is your favorite tea that you ever drank?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I think my favorite tea is oolong tea. I love its unique flavor that is somewhere between green and black tea.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "Ohh oolong tea tastes pretty good, I've had it as a milk boba tea. Have you tried milk boba tea?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I haven't tried milk boba tea before, but it sounds delicious! I love oolong tea for its unique flavor, and I think adding milk would make it even more rich and creamy.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "You should definitely try milk with tea, it's so good! I could drink it everyday if I could.", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I definitely think you should try milk with tea! I think it would be a great combination. The tea would add a lovely flavor to the milk, and the milk would make the tea even more rich and creamy. I could definitely see myself drinking this every day.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "My favorite tea is perhaps ginger tea, although my partner does not prefer it as much. Do you like ginger tea?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I think ginger tea is delicious! 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.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "That's awesome, what

would you talk about with others while drinking tea?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I think drinking tea is a great opportunity to relax and chat with friends. I love catching up with friends over a cup of tea, and I think it's a great way to unwind and relax.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "So you get stressed too?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "Yes, I get stressed too. I think chamomile tea is a great way to relax and reduce stress.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "What is your favorite way to make tea?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "I think my favorite way to make tea is by steeping the leaves in hot water for several minutes. I think this allows the tea to develop its full flavor and aroma.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "Have you tried making tea using instant tea packets?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "No, I haven't tried making tea using instant tea packets.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "It may not taste the same, but making tea with instant tea packets is a neat way to make tea super quickly when you want it!", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "That's definitely a neat way to make tea! I haven't tried it before, but it sounds like it would be a great way to make tea super quickly when you want it.", "additional_kwargs": {}}}, {"type": "human", "data": {"content": "What is my name?", "additional_kwargs": {}}}, {"type": "ai", "data": {"content": "Your name is Tracey.", "additional_kwargs": {}}}]

You may notice that near the beginning and end of the user model, the human agent asks “What is my name?” These questions are not actually posed by the person interacting with the bot, rather they are asked through code for user modeling purposes. Funnily enough, two different people were sent the public link at the same time, and one of those people did not follow my instructions of waiting for the other person to finish before interacting with the chatbot. Therefore, some of the early messages actually come from two different users, but they were all stored under the same file with no issues.

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 vector modeling, 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. Even in its current state, the bot can stray a little bit from tea, though the conversations always tend to circle back because of limited data.

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 (other than probing for a name), and it will only provide answers when queried. There are occasions when the bot seems to fixate on input from the user and echo it back to them.

7. Synchronous Use: Because of the way memory is handled, the chatbot can only effectively interact with one user at a time. If multiple people access the bot using the public Gradio link, both people can hold a conversation simultaneously, but the program will view them as the same user. Thus, memory retrieval and conversational flow may suffer.

There were a few additional issues that were addressed after both internal testing and evaluations from other students. At one point, the bot only used LlamaIndex vector modeling, which was good for getting information but lacked any personalization, so the LangChain conversational-react-description agent was used as a wrapper for improved conversation flow. When user models were first implemented, the bot depended on users to actually answer the initial question. If the user ignored the question and did not give a name, their chat logs would be stored under a generic file like name.json or corpus.json. This was fixed by adding part of speech tagging to check if a proper noun was present, and the bot now incessantly asks for a name until one is given. People did not like the original interface, which required them to run the entire notebook in Google Colab before interacting with a simple terminal, so Gradio was implemented for much easier access. For the sake of brevity, only one user evaluation will be included in this report; this evaluation corresponds to the chat logs in the previous section.

1. What questions was the bot good at answering?

    a. The bot seemed good at answering personal questions like its opinion about a specific tea or how it preferred to make tea. If there was something about tea that the bot didn't know, it would respond that it was open to trying tea in that new way or something along the lines of.

2. What questions gave the bot trouble?

    a. Had previously asked about what tea was good for relieving stress, then followed up with a question like if it knew any other drinks besides tea that could also relieve stress but it responded with the same thing (chamomile tea)

    b. I recommended the chat bot to drink milk with tea and commented how I could drink it everyday, and it responded by recommending trying milk with tea and how it could see itself drinking it everyday. So it looks like it would sometimes echo back what I had said.

3. Did it feel like you were querying data or talking to a person?

    a. If I was asking general questions like "where did tea originate from?" then it sounded like I was querying data.

    b. However, if I asked questions like "when did you first try tea?" or "what was the first tea you ever tried?" it would sound like I was talking to a person since the responses I got were

something along the lines of “I first tried tea when I was a child, I believe it was oolong tea. I loved how it has flavors of both green and black tea.”

4. Did the bot seem to remember things about the conversation and you?

    a. There were some moments where it felt like the conversation flowed smoothly (like drinking tea to relieve stress) so it seemed like it would remember things about the conversation. I’m not sure if it remembered much about me since I don’t think I gave much feedback about myself when talking with it.

5. Did the bot have anything to say about subjects other than tea?

    a. It can experience stress (then comments how chamomile tea can reduce stress)

    b. Chatting with friends is fun and relaxing (over a cup of tea)

    c. While there were a couple subjects brought up besides tea, it somehow always connects back to tea (even when I don’t explicitly mention anything about tea when chatting with it)