Project 1 Report

HeartWise – a chatbot on heart disease information by Adithya Viswanathan

Part 1: Web Crawler

1. Web Scraping and clean-up

   First, I scraped a couple starter URLs to start my URL collection. Let's call this Round 1. Then, I iterated through Round 1 URLs and looked for more URLs in each of them. These are Round 2 URLs. Scraping was done with BeautifulSoup library in Python.

   Round 1 websites were chosen in a way that once scraped, their "child" URLs would contain a variety of links leading to different topics, domains, but still centered around heart disease. This was done through a blacklist and whitelist filtering. URLs with keywords such as Espanol, Facebook, Instagram, or pdf were discarded, while URLs with keywords such as heart, cardiology, cardio and disease were kept.

   Initially, I had 2 starter websites. I experimented with adding one more starter URL so my knowledge base can grow and have more information. But, this resulted in more unnecessary, irrelevant, and off-topic sentences in my knowledge base. So, I went back to having only 2 starter URLS. This resulted in about 45 websites scraped. The exact number depends on which websites it successfully scrapes; http errors might prevent all the collected URLs from being scraped. The site numbers, lengths and URLs can be found in log.txt.

   The next step was to clean up the scraped files. Let's call the input files "raw scrape" and the output "cleaned up" files. These files can be found in subdirectories with similar names. First, I did some **regex** to remove newline characters and clean up some Unicode characters. Then, I **tokenized** into sentences and returned this in list form. This was sufficient to transform the raw scraped files into plain sentence files. These can be found in the "cleaned_up" subfolder.

2. Keywords

   Now that all website information is cleaned up and in sentence form, it is time to evaluate the vocabulary and find keywords to populate our knowledge base with. For this, I employed **TF-IDF** measurement. With the help of Professor's detailed TF-IDF code on GitHub, I was able to find the most important words across my corpus. The preliminary clean-up techniques I used was **word-tokenizing**, **lower-casing**, and **removing stopwords and punctuation**. The keywords outputted by TF-IDF were, in general, better than I expected. It caught many important words pertaining to my domain, such as "cholesterol", "valve", "pulmonary", "inflammation", etc. But some words were not of the best use, such as “Update”, “depend”, “Many”, “Learn”, etc. In the end, I aggregated about 19 keywords, some of which were manually added based on my domain knowledge and anticipation of what vocabulary a potential user would use in their chat with my chatbot.

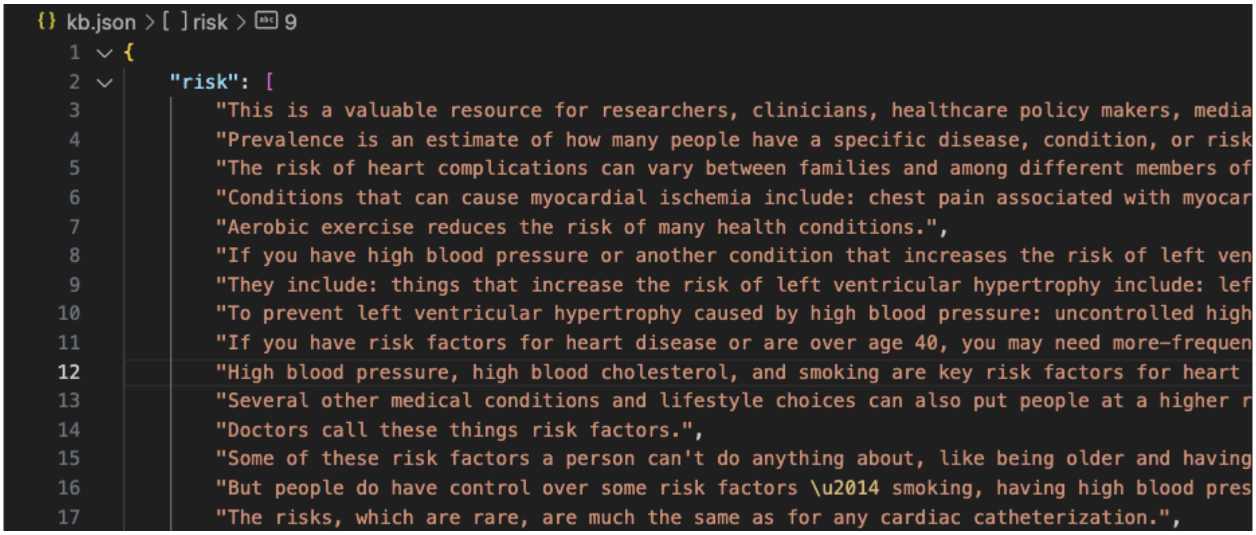
Keywords: ['risk', 'prevent', 'cure', 'medicine', 'symptom', 'attack', 'treatment', 'cause', 'type', 'foods', 'hypertension', 'ischemia', 'help', 'failure', 'valve', 'pain', 'congenital', 'cholesterol', 'definition']

3. Knowledge base

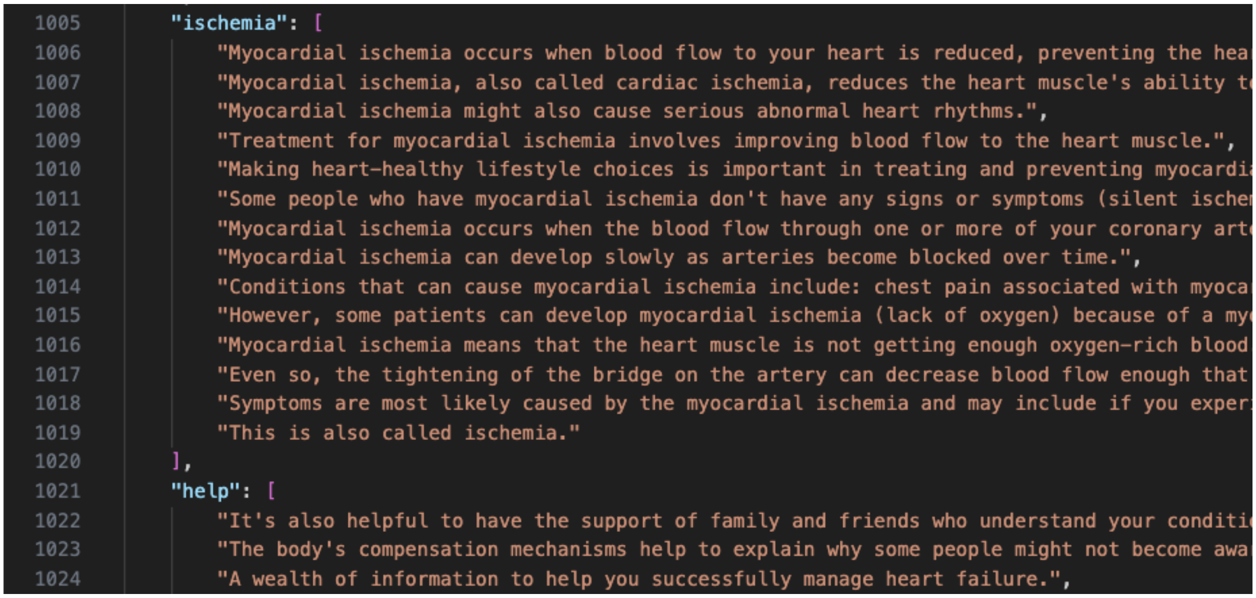
Once I had this list of keywords, it was time to start populating my knowledge base. My approach was parsing through the cleaned-up files, tokenizing each document into sentences, and iterating through each sentence in each document. This was when I found which of the keywords were in this sentence and added it to that keyword’s sentence bank. In the process, I filtered out sentences which I commonly found unnecessary, such as sentences beginning with “Read More”, “Learn more”, etc. One keyword I would like to elaborate on is “definition”. This had some extra manipulation done before being added as a keyword to the knowledge base. This keyword actually started out as “ term “ (yes, with the spaces). The reason for this keyword was because I noticed that most sentences with definitions of important terms contain the word “term” somewhere in the sentence. So, I found sentences containing “ term “ and added them to a keyword bank. Then, I renamed the key of this dictionary element as “definition”. This way, when a user asks for a definition, I can fetch an answer from the “definition” word bank.

Finally, I wrote the knowledge to a json file and pickle file. The former was for visual inspection purposes, and the latter was for efficient access by chatbot module.

4. Screenshots of Knowledge Base (KB)



```
{
    "risk": [
        "This is a valuable resource for researchers, clinicians, healthcare policy makers, media
        "Prevalence is an estimate of how many people have a specific disease, condition, or risk
        "The risk of heart complications can vary between families and among different members of
        "Conditions that can cause myocardial ischemia include: chest pain associated with myocar
        "Aerobic exercise reduces the risk of many health conditions.",
        "If you have high blood pressure or another condition that increases the risk of left ven
        "They include: things that increase the risk of left ventricular hypertrophy include: lef
        "To prevent left ventricular hypertrophy caused by high blood pressure: uncontrolled high
        "If you have risk factors for heart disease or are over age 40, you may need more-frequen
        "High blood pressure, high blood cholesterol, and smoking are key risk factors for heart
        "Several other medical conditions and lifestyle choices can also put people at a higher r
        "Doctors call these things risk factors.",
        "Some of these risk factors a person can't do anything about, like being older and having
        "But people do have control over some risk factors \u2014 smoking, having high blood pres
        "The risks, which are rare, are much the same as for any cardiac catheterization.",
```



```
    "ischemia": [
        "Myocardial ischemia occurs when blood flow to your heart is reduced, preventing the hea
        "Myocardial ischemia, also called cardiac ischemia, reduces the heart muscle's ability t
        "Myocardial ischemia might also cause serious abnormal heart rhythms.",
        "Treatment for myocardial ischemia involves improving blood flow to the heart muscle.",
        "Making heart-healthy lifestyle choices is important in treating and preventing myocardi
        "Some people who have myocardial ischemia don't have any signs or symptoms (silent ische
        "Myocardial ischemia occurs when the blood flow through one or more of your coronary art
        "Myocardial ischemia can develop slowly as arteries become blocked over time.",
        "Conditions that can cause myocardial ischemia include: chest pain associated with myoca
        "However, some patients can develop myocardial ischemia (lack of oxygen) because of a my
        "Myocardial ischemia means that the heart muscle is not getting enough oxygen-rich blood
        "Even so, the tightening of the bridge on the artery can decrease blood flow enough that
        "Symptoms are most likely caused by the myocardial ischemia and may include if you exper
        "This is also called ischemia."
    ],
    "help": [
        "It's also helpful to have the support of family and friends who understand your conditi
        "The body's compensation mechanisms help to explain why some people might not become awa
        "A wealth of information to help you successfully manage heart failure.",
```

Part 2: Chatbot

System Description (including NLP techniques – **in bold**)

My chatbot is called HeartWise and holds information on heart disease.

User profile:
Creates/retrieves user json file (using user's name) to store personal information such as name, questions asked, likes, dislikes, and feedback for chatbot. Ex: adithya.json. From now on, chatbot refers to user by their name.

Chatbot asks user for query (about 2/3 of the time) or personal question (1/3 of the time)
When chatbot asks for query:

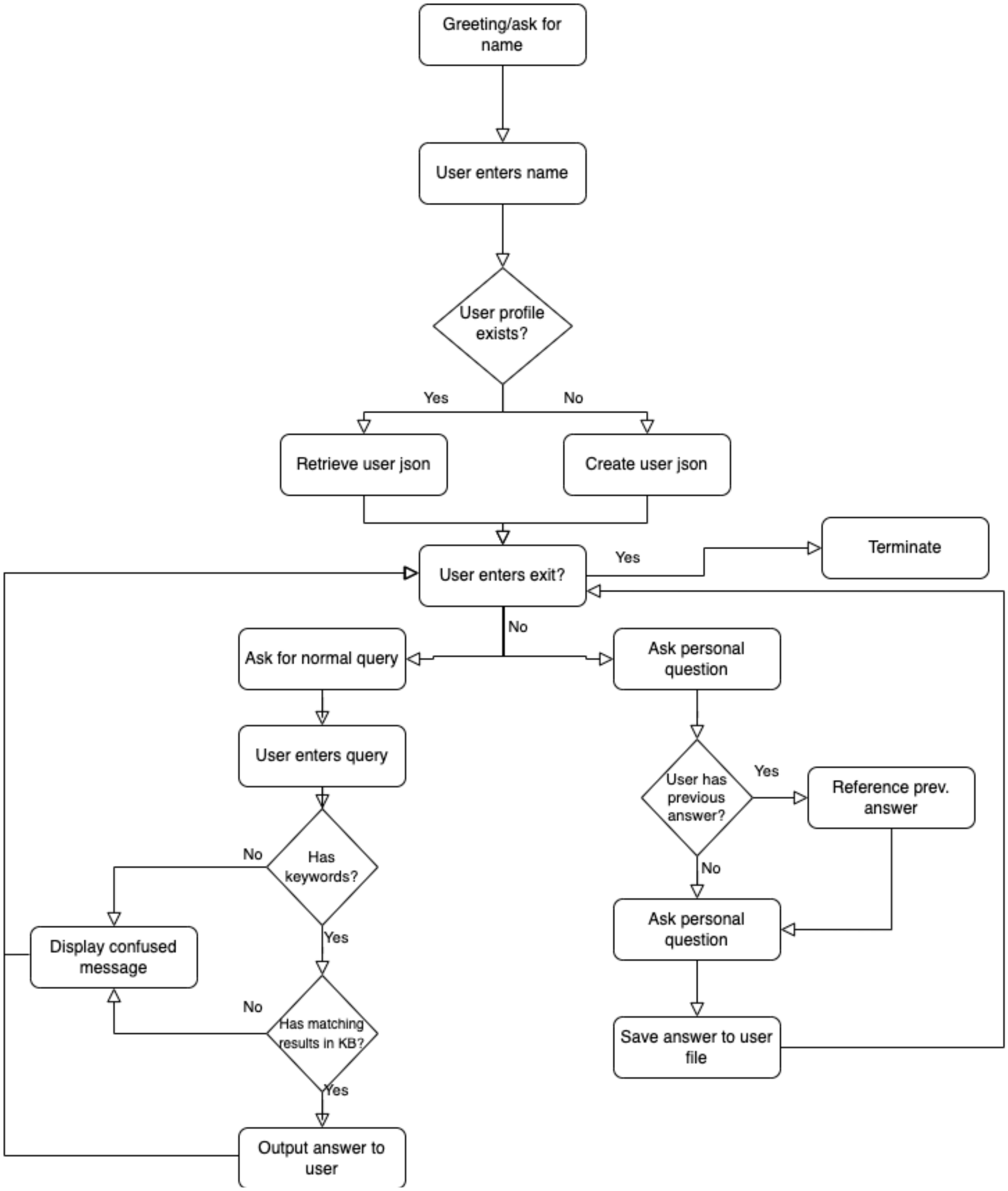
- Writes query to questions list in user file
- Gets keywords from user query
    - NLP Techniques: **TF-IDF keyword extraction**, **word tokenize**, **lemmatizing**, **lowercasing**, **filtering out non alpha-numeric words and stopwords**
    - Other: **Difflib** (sequence matcher library)
    - Keywords which match a key in knowledge base are labeled as "confirmed keywords". Others are labeled as "unconfirmed keywords".
- Selecting best answer from knowledge base from keywords
    - Since knowledge base (KB) is json, confirmed keyword acts as a key and value is list of sentences. This is done to get to the general topic of the answer.
    - Unconfirmed keywords given in user query are used to narrow down the list of sentences. This makes sure a relevant sentence is returned.
    - Final answer is chosen randomly if filtering above results in more than 1 sentence.

When chatbot asks personal question

- Choose from a pool of questions
    - 3 categories of question: general likes, general dislikes, and feedback about chatbot
    - Customized for each user: if user has previous answers, it references user's previous answer and asks if they would like to add on.
- Save answer to appropriate field in user json file

EXIT command will terminate the program.

Dialog tree or logic



Sample dialog interactions

SAMPLE 1

```
Welcome! My name is HeartWise, your heart companion. What is your name?
> Adithya

Hello Adithya! Let's chat about heart health and how you can maintain a healthy heart.
Enter your question or 'exit' to exit the program
> What is heart disease?
The term heart disease refers to several types of heart conditions, including coronary artery disease and heart attack.
> What are the symptoms of heart disease?
The symptoms of a heart valve problem are similar to symptoms of other heart diseases, and include:  in the case of a valve that does not close all the way, a doctor may recommend surgery to repair the leaflets of the valve.
> What are the causes of heart disease?
Heart disease is the leading cause of death in the united states, but there are ways to prevent and manage many types of heart disease.
> What are the causes of heart attack?
Call 911 or your local emergency number if you have signs and symptoms of a potential heart attack: an enlarged heart (cardiomegaly) can be caused by damage to the heart muscle or any condition that makes the heart pump harder than usual, including pregnancy.
> What foods should I eat?
The american heart association recommends limiting red meat in favor of lean meats such as fish or chicken, and watching portions on carbohydrate-heavy foods, such as pasta and rice.

 Sorry to interrupt adithya. Can I get some feedback about my chatbot skills?
> Answers are well-written and relevant.
Thank you for your response!

 Sorry to interrupt adithya. Can I get some feedback about my chatbot skills? You previously said your feedback was "Answers are well-written and relevant.". What else would you like to add on?
> Answers are correct length, not too long, and not too short
Thank you for your response!
> How big is the heart?

Apologies, I'm not sure I follow. Can you clarify or ask a different question?
> Is there cure for heart disease?
Even with successful treatment, there is no outright cure for carotid artery disease.

 Before I forget, I'm curious to know about any activities, foods, or preferences you don't like?
> I don't like donuts
```

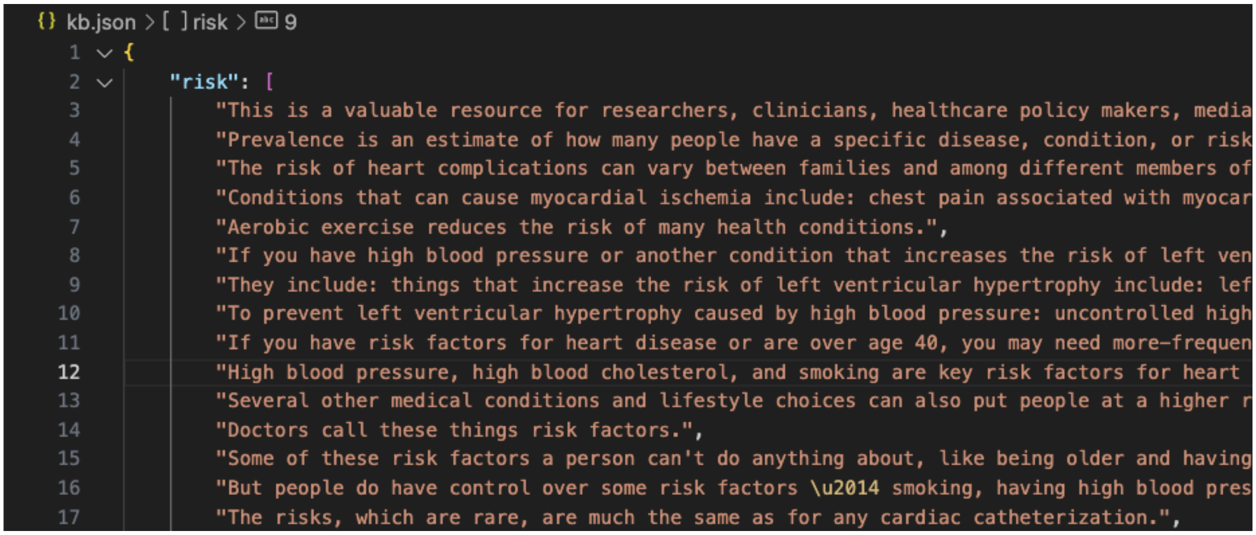
```
Thank you for your response!
> exit
```

SAMPLE 2

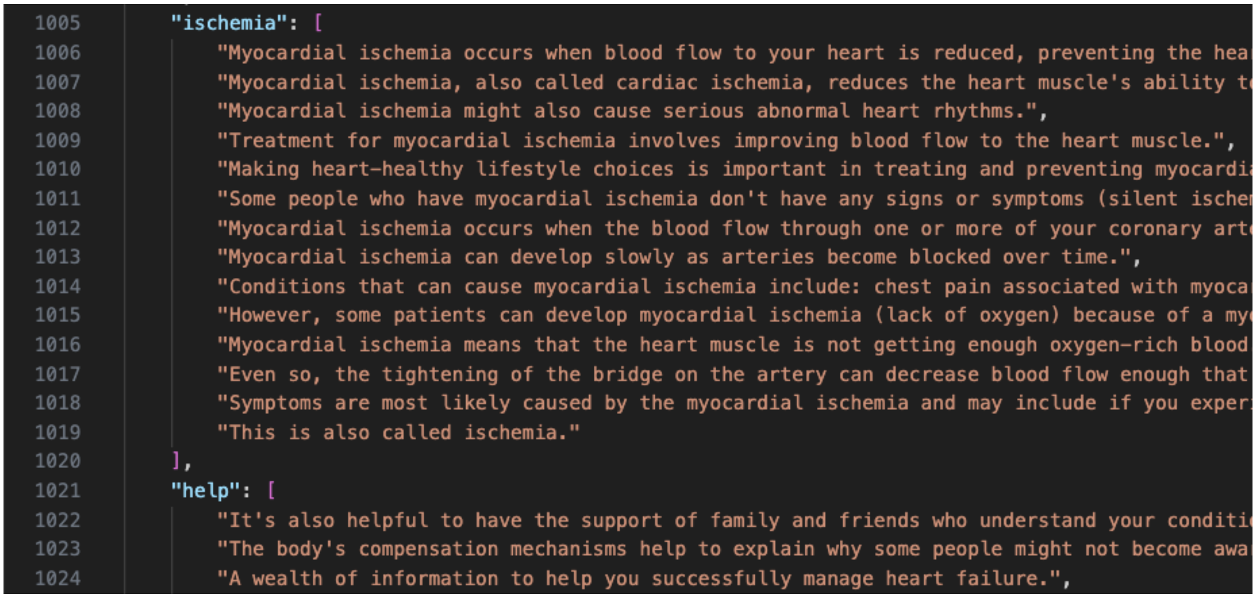
```
Welcome! My name is HeartWise, your heart companion. What is your name?
> Vlad

Hello Vlad! I'm here to provide guidance on understanding and managing heart disease. Let's get started.
Enter your question or 'exit' to exit the program
> What can prevent heart disease?
Non-study author dr. sonia rivera-martinez, do, an american osteopathic association board certified physician specializing in family medicine, offered the following speculation:   the identification of biomarkers indicating arterial inflammation can lead to development of biological inhibitors of the specific biomarker causing the arterial inflammation and thereby preventing the development of cardiovascular disease.
> What medicine should I take?
If tachycardia occurs, your pediatric cardiologist and obstetrician will recommend medicine for you to take to help regulate your fetus's heart rhythm.

 Just wondering vlad, what are your interests?
> I like walking my dog
Thank you for your response!

 Sorry to interrupt vlad. Can I get some feedback about my chatbot skills?
> I like the level of personal connection
Thank you for your response!
> Is there a cure for heart disease?
Even with successful treatment, there is no outright cure for carotid artery disease.
> Is there a risk of death?
They are also a risk for heart failure, stroke, and sudden death.
> What type of surgery do I need?
The treatment options will vary depending on the type of heart disease a person has, but some common strategies include making lifestyle changes, taking medications, and undergoing surgery.
> exit
```

An appendix for the knowledge base you created with samples

Note: Knowledge base is not modified in any way from the output of web crawler code. The following screenshots are same as previously mentioned.

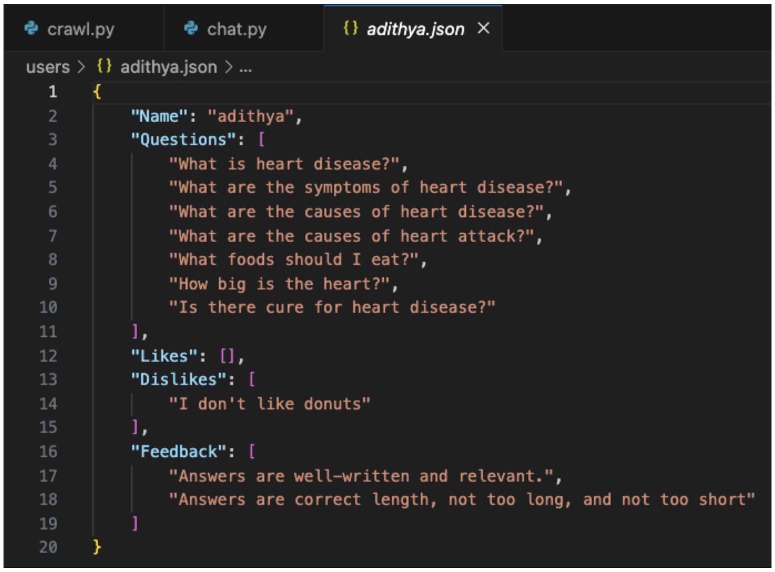


```
{
    "risk": [
        "This is a valuable resource for researchers, clinicians, healthcare policy makers, media
        "Prevalence is an estimate of how many people have a specific disease, condition, or risk
        "The risk of heart complications can vary between families and among different members of
        "Conditions that can cause myocardial ischemia include: chest pain associated with myocar
        "Aerobic exercise reduces the risk of many health conditions.",
        "If you have high blood pressure or another condition that increases the risk of left ven
        "They include: things that increase the risk of left ventricular hypertrophy include: lef
        "To prevent left ventricular hypertrophy caused by high blood pressure: uncontrolled high
        "If you have risk factors for heart disease or are over age 40, you may need more-frequen
        "High blood pressure, high blood cholesterol, and smoking are key risk factors for heart
        "Several other medical conditions and lifestyle choices can also put people at a higher r
        "Doctors call these things risk factors.",
        "Some of these risk factors a person can't do anything about, like being older and having
        "But people do have control over some risk factors \u2014 smoking, having high blood pres
        "The risks, which are rare, are much the same as for any cardiac catheterization.",
```

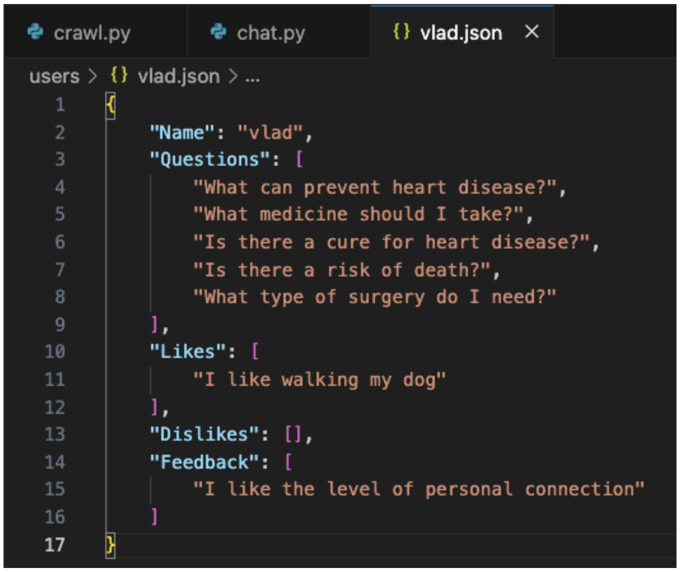


```
    "ischemia": [
        "Myocardial ischemia occurs when blood flow to your heart is reduced, preventing the hea
        "Myocardial ischemia, also called cardiac ischemia, reduces the heart muscle's ability t
        "Myocardial ischemia might also cause serious abnormal heart rhythms.",
        "Treatment for myocardial ischemia involves improving blood flow to the heart muscle.",
        "Making heart-healthy lifestyle choices is important in treating and preventing myocardi
        "Some people who have myocardial ischemia don't have any signs or symptoms (silent ische
        "Myocardial ischemia occurs when the blood flow through one or more of your coronary art
        "Myocardial ischemia can develop slowly as arteries become blocked over time.",
        "Conditions that can cause myocardial ischemia include: chest pain associated with myoca
        "However, some patients can develop myocardial ischemia (lack of oxygen) because of a my
        "Myocardial ischemia means that the heart muscle is not getting enough oxygen-rich blood
        "Even so, the tightening of the bridge on the artery can decrease blood flow enough that
        "Symptoms are most likely caused by the myocardial ischemia and may include if you exper
        "This is also called ischemia."
    ],
    "help": [
        "It's also helpful to have the support of family and friends who understand your conditi
        "The body's compensation mechanisms help to explain why some people might not become awa
        "A wealth of information to help you successfully manage heart failure.",
```

An appendix for sample user models that were created.



```json
{
    "Name": "adithya",
    "Questions": [
        "What is heart disease?",
        "What are the symptoms of heart disease?",
        "What are the causes of heart disease?",
        "What are the causes of heart attack?",
        "What foods should I eat?",
        "How big is the heart?",
        "Is there cure for heart disease?"
    ],
    "Likes": [],
    "Dislikes": [
        "I don't like donuts"
    ],
    "Feedback": [
        "Answers are well-written and relevant.",
        "Answers are correct length, not too long, and not too short"
    ]
}
```



```json
{
    "Name": "vlad",
    "Questions": [
        "What can prevent heart disease?",
        "What medicine should I take?",
        "Is there a cure for heart disease?",
        "Is there a risk of death?",
        "What type of surgery do I need?"
    ],
    "Likes": [
        "I like walking my dog"
    ],
    "Dislikes": [],
    "Feedback": [
        "I like the level of personal connection"
    ]
}
```

Sample user files: adithya.json and vlad.json
"Adithya" and "Vlad" refers to the name that the user entered when program prompted them in the beginning. This file holds the following fields and information:

- Name: self-explanatory
- Questions: questions that user has asked
- Likes: information that user has provided about their interests
- Dislikes: information that user has provided about their disinterests

Fields start as empty lists, and can grow according to what questions are asked by user.

Evaluations of the chatbot and analysis of its strengths and weaknesses

Strengths:

- When a user's query has keyword matching a key in the knowledge base dictionary, the general answer is correct.
- Saving user's information is flawless.
- Getting user's previous answers and asking new question is working well.

Weaknesses:

- Answers (even through a valid keyword) are not relevant (due to web crawler and sentences that don't have to do with the keyword mentioned in them)
- Can be more personalized. Currently relying on few versions of greetings, confusion, and personal questions.

Likert-Style Feedback

1. On a scale of 1 to 5, how satisfied are you with the responsiveness of the chatbot?
   1 - Very Dissatisfied | 2 – Dissatisfied | 3 - Neutral | 4 – Satisfied | 5 - Very Satisfied

2. How helpful did you find the chatbot's responses?
   1 - Not Helpful at All | 2 - Slightly Helpful | 3 - Moderately Helpful | 4 – Helpful | 5 - Extremely Helpful

3. To what extent did the chatbot understand your queries?
   1 - Didn't Understand at All | 2 - Partially Understood | 3 - Understood Somewhat | 4 - Mostly Understood | 5 - Completely Understood

4. How satisfied are you with the accuracy of the information provided by the chatbot?
   1 - Very Inaccurate | 2 - Somewhat Inaccurate | 3 – Neutral | 4 - Somewhat Accurate | 5 - Very Accurate

Responses:

Surya, my family friend
5
4
2
4

Aparna, my sister
4
3
3
4

Shankari, my mom
4
4
3
4