Farishah Nahrin
CS 6301.M02

Chatbot: Report & Reflection

(1) System description, including description of specific NLP techniques you used and how you used them.
(2) Diagram of your dialog tree or logic.
(3) Sample dialog interactions.
(4) An appendix for the knowledge base (or live lookup) you created with samples.
(5) An appendix for sample user models that were created.
(6) Evaluations of the chatbot and analysis of its strengths and weaknesses.

System description, including description of specific NLP techniques you used and how you used them

The primary aim of this project is to develop an informational chatbot, which provides users with accurate and relevant information about Bangladesh. The chatbot will be an interactive platform for users to ask questions and receive personalized responses on various topics, including geography, culture, history, economy, politics, and tourism.

The chatbot system I have developed is a natural language processing (NLP) based application, designed to answer user queries about Bangladesh by combining the knowledge base and Wikipedia API. I have used various NLP techniques and libraries such as spaCy, NLTK, SQLite, joblib, and the Wikipedia API to develop this chatbot. The system also involves handling a database for user data and query management. Here, I will elaborate on the NLP techniques and components used in the system.

1. Text Pre-processing: I used the Natural Language Toolkit (NLTK) library to process the user queries and the knowledge base content. The text pre-processing involved tokenizing the text into words, converting them to lowercase, removing stop words, and filtering out the punctuation. These processes made it easier to have effective information retrieval and comparisons between the user’s queries and the knowledge base.

2. Named Entity Recognition (NER): I used the spaCy library to perform NER and extract the main noun from the user's query. The English model (en_core_web_sm) in spaCy, was loaded to process the user query. It allowed the program to identify the noun chunks and standalone nouns, which were then used to search the knowledge base and/or retrieve information from Wikipedia, in case we didn’t have relevant information in the knowledge base.

3. Jaccard Similarity: I calculated the Jaccard similarity to measure the similarity between the user query and the knowledge base content. It is a metric that measures the similarity between two sets, in this case, it compares the set of words in the user query versus the set of words in the knowledge base keys. By calculating the intersection and union of the sets, I obtained the similarity score, which enabled the program to identify the most relevant content from the knowledge base. Thus, Jaccard similarity only takes the unique set of words for each sentence, whereas cosine similarity takes the total length of the vectors. Jaccard similarity calculates similarity based on the overlap of the words in the text, rather than their frequency. It handles

variations in the length and quality of the text better and is faster than Cosine similarity because no dot product calculation is required.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B| - |A \cap B|}$$

4. Information Retrieval: After obtaining the main noun from the user query, I used the search_knowledge_base function in the code, to search the knowledge base for the most similar item. If a direct match was not found, I calculated the Jaccard similarity between the main noun and the knowledge base keys. The result was returned if the similarity score was above a certain threshold. I used the Wikipedia API to retrieve relevant information, if the knowledge base search failed.

5. Sentence Tokenization: I used the sentence tokenization functionality, from the NLTK library, to tokenize the response text into sentences. This step was necessary for selecting the top 5 sentences to be presented to the user as a concise response.

6. Text Wrapping: To improve the readability of the response, I used the text-wrap library to wrap the top 5 sentences to a maximum of 70 characters per line. The wrapped text was then printed line by line for a better user experience.

7. SQLite Database: I used the SQLite3 Python library to create and manage a database (user_data.db) to store user information, interests, and past queries. The database contained two tables, "users" and "queries." I implemented functions to connect to the database, create tables, add new users, update user interests, add new queries, and retrieve user queries and responses from the database.

8. User Interaction: The chatbot begins by greeting the user and asking for their name and interests. It allows the user to ask a query, see past queries, or exit. When the user submits a query, the chatbot handles it by extracting the main noun, searching the knowledge base, and retrieving information from the knowledge base or Wikipedia. Afterward, the chatbot updates the user's interest in the database and stores the query and response. The user can also view their past queries and corresponding responses from the database.

9. Data Persistence: To store the knowledge base and user preferences, I used the joblib library in Python, to create a serialized binary file, which allowed for faster loading and saving of the data. I implemented the functions load_data and save_data to handle the loading and saving of data for the chatbot's operation.

10. Error Handling: The chatbot was designed to handle various errors that might occur during its operation. For example, if the knowledge base does not contain the keywords that the user uses in their query, the program will redirect to the Wikipedia API and will start displaying a response from Wikipedia, instead. Next, if the Wikipedia API fails to retrieve relevant information, the chatbot would inform the user and encourage them to rephrase their query or ask a different question. The system also handled errors related to the database, such as failed connections or incomplete data.

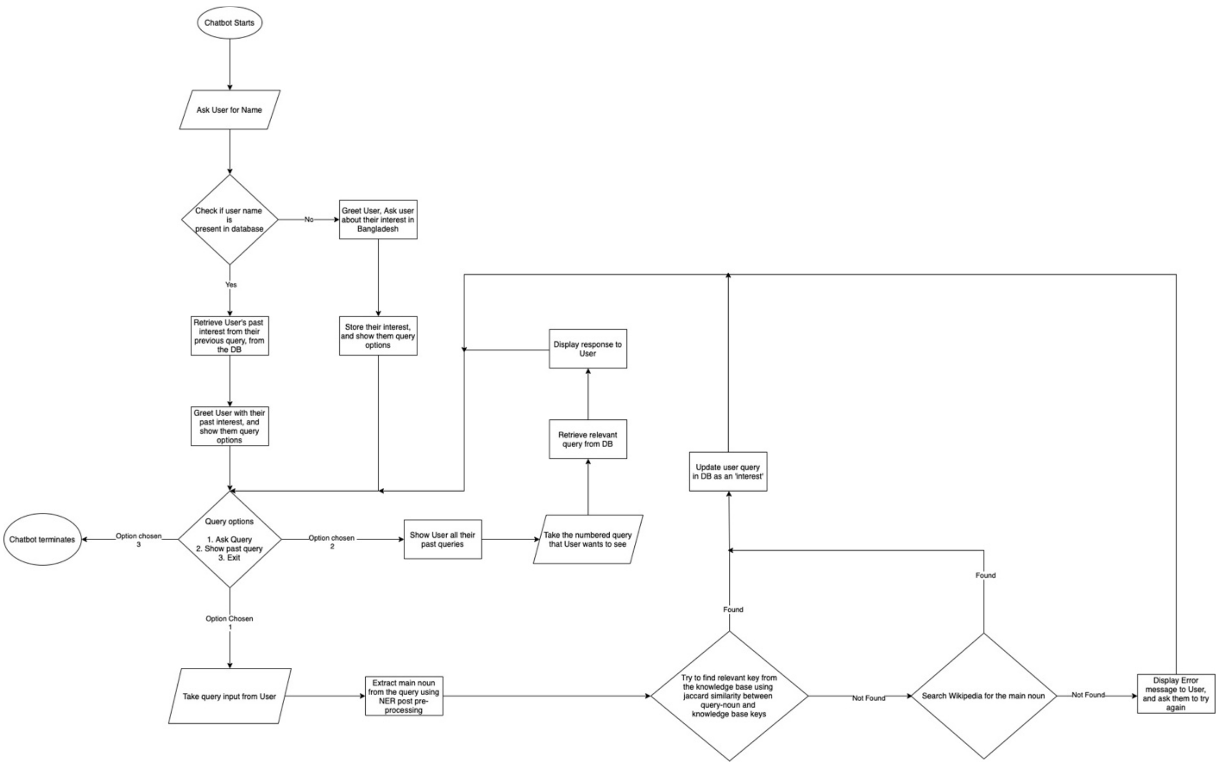
11. Customization: To personalize the chatbot experience, I implemented a system that allowed users to set their interests, which were then saved in the database. This information was used to prioritize responses related to the user's interests when searching the knowledge base and retrieving information from Wikipedia. Also, every time a user runs the chatbot, the chatbot introduction will say, “Here are some sample topics I can tell you about:,” and it will randomly pick up to 5 topics from the knowledge base *each time,* the program runs.

12. Performance Optimization: I implemented techniques such as memoization and caching to optimize the chatbot's performance. Memoization allows the system to store the results of the function calls and return cached results, when the same inputs occurred. Caching was implemented for database queries, which reduced the number of requests to the database and improved the chatbot's overall response time.

External Resources - the following external resources were used in the development of the chatbot:

spaCy library: https://spacy.io/
NLTK library: https://www.nltk.org/
SQLite3 library: https://docs.python.org/3/library/sqlite3.html &
https://towardsdatascience.com/starting-with-sql-in-python-948e529586f2
Wikipedia API: https://pypi.org/project/wikipedia/
Joblib library: https://joblib.readthedocs.io/
Jaccard index: https://en.wikipedia.org/wiki/Jaccard_index
Jaccard index algorithm: https://www.statology.org/jaccard-similarity-python/,
Jaccard index algorithm comparisons: https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50
Textwrap library: https://docs.python.org/3/library/textwrap.html

Diagram of your dialog tree or logic


Chatbot Starts
Ask User for Name
Check if user name is present in database
No
Greet User, Ask user about their interest in Bangladesh
Yes
Retrieve User's past interest from their previous query, from the DB
Store their interest, and show them query options
Greet User with their past interest, and show them query options
Display response to User
Retrieve relevant query from DB
Update user query in DB as an 'interest'
Chatbot terminates
Option chosen 3
Query options
1. Ask Query
2. Show past query
3. Exit
Option chosen 2
Show User all their past queries
Take the numbered query that User wants to see
Found
Found
Option Chosen 1
Take query input from User
Extract main noun from the query using NER post pre-processing
Try to find relevant key from the knowledge base using jaccard similarity between query-noun and knowledge base keys
Not Found
Search Wikipedia for the main noun
Not Found
Display Error message to User, and ask them to try again

Sample dialog interactions

1st Chatbot Run:

```
What is your name? Bob
Hi Bob, I am a chatbot that knows all about Bangladesh! Can you please
tell me why you are interested in learning about Bangladesh? Here are
some sample topics I can tell you about: sunil gangopadhyay, the march
days, a brief history of the bangla language movement, defense,
religious holidays and festivals

> Travelling

Thank you Bob. I will remember that.

Choose an option:
1. Ask a query
2. See past queries
3. Exit
> 1

Please enter your query, make sure to ask about one topic at a time:
> Can you tell me about the flag of bangladesh?

Flag adopted, The national flag is bottle green in color and
rectangular in size with the length to width ratio of 10:6 bearing a
red circle on the body of the green.
The red circle has a radius of one fifth of the length of the flag.
```

```
Its center is placed on the middle of the perpendicular drawn from the
nine twentieth part of the flag.
The background color symbolizes the greenery of Bangladesh with its
vitality and youthfulness while the red disc represents the rising sun
and the sacrifice our people made to obtain our, independence, The
Historical Flag, This is the first flag of the independence movement.
Initially, the flag had a map of Bangladesh in gold at the center of
the red circle.

Choose an option:
1. Ask a query
2. See past queries
3. Exit
> 3

Goodbye, Bob!
```

2nd Chatbot Run:

```
What is your name? Bob
Welcome back, Bob! Last time, you said you were interested in "flag".
What are you interested in this time? Here are some sample topics I
can tell you about: bangladesh march days, shamsur rahman, defense,
agriculture, brief facts

Choose an option:
1. Ask a query
2. See past queries
3. Exit
> 2

Here are your past queries, Bob:
1. Can you tell me about the history of bangladesh?
2. Can you tell me about the flag of bangladesh?
Enter the index of the query you want to see the response for: 2

Here is the response for the selected query:
Flag adopted, The national flag is bottle green in color and
rectangular in size with the length to width ratio of 10:6 bearing a
red circle on the
body of the green.
The red circle has a radius of one fifth of the length of the flag.
Its center is placed on the middle of the perpendicular drawn from the
nine twentieth part of the flag.
The background color symbolizes the greenery of Bangladesh with its
vitality and youthfulness while the red disc represents the rising sun
and the
sacrifice our people made to obtain our, independence, The
Historical Flag, This is the first flag of the independence movement.
Initially, the flag had a map of Bangladesh in gold at the center of
the red circle.
```

```
Choose an option:
1. Ask a query
2. See past queries
3. Exit
> 1

Please enter your query, make sure to ask about one topic at a time:
> Can you tell me about the education in bangladesh?

The educational system in Bangladesh is three-tiered and highly
subsidized.
The government of, Bangladesh, University Grants Commission, The three
main educational systems in Bangladesh, ordered by decreasing student
numbers, are:, General Education System, Madrasah Education System,
Technical – Vocational Education System, Other systems include a
Professional Education System., Each of these three main systems is
divided into five levels:, Primary Level (years 1 to 4), Junior Level
(years 5 to 8), Secondary Level (years 9 to 10), Higher Secondary
Level (years 11 and 12), Tertiary Level, The five years of lower
secondary (grades nine through ten) concludes with a secondary school
certificate examination.
Students who passed this examination proceeded to two years of higher
secondary or intermediate training, which culminated in a higher
secondary school examination after grade twelve.
Higher secondary school was viewed as preparation for college rather
than as the conclusion of high school.
Tertiary education in Bangladesh takes place at 34 government, 78
private and 3 international universities.

Choose an option:
1. Ask a query
2. See past queries
3. Exit
> 1

Please enter your query, make sure to ask about one topic at a time:
> What are the animals of bangladesh?
The Royal Bengali Tiger, The majestic Royal
Bengal Tiger is the national animal Bangladesh.
Highly endangered, the Royal Bengal can now be mostly be found in the
Sundarbans. One of the largest of the ‘big cats’, it has extremely
bold and striking colour pattern – making it perhaps the most
magnificent and sought-after fiery beast of the world!
The vivid pattern of stripes on the glossy skin serves as a very
effective camouflage in the grasses and foliage almost in all the
seasons., The male averages 3 meters in length including 1 meter of
tail and wiighs about 180 kg., though much larger specimens have been
known.
The giant one is the Siberian tiger, almost 4 meters long and weighing
about 300 kg., Deers, Sambar Dee, The Sambar Deer is the most widely
spread deer species in the world, covering many countries in the Asian
continent.
```

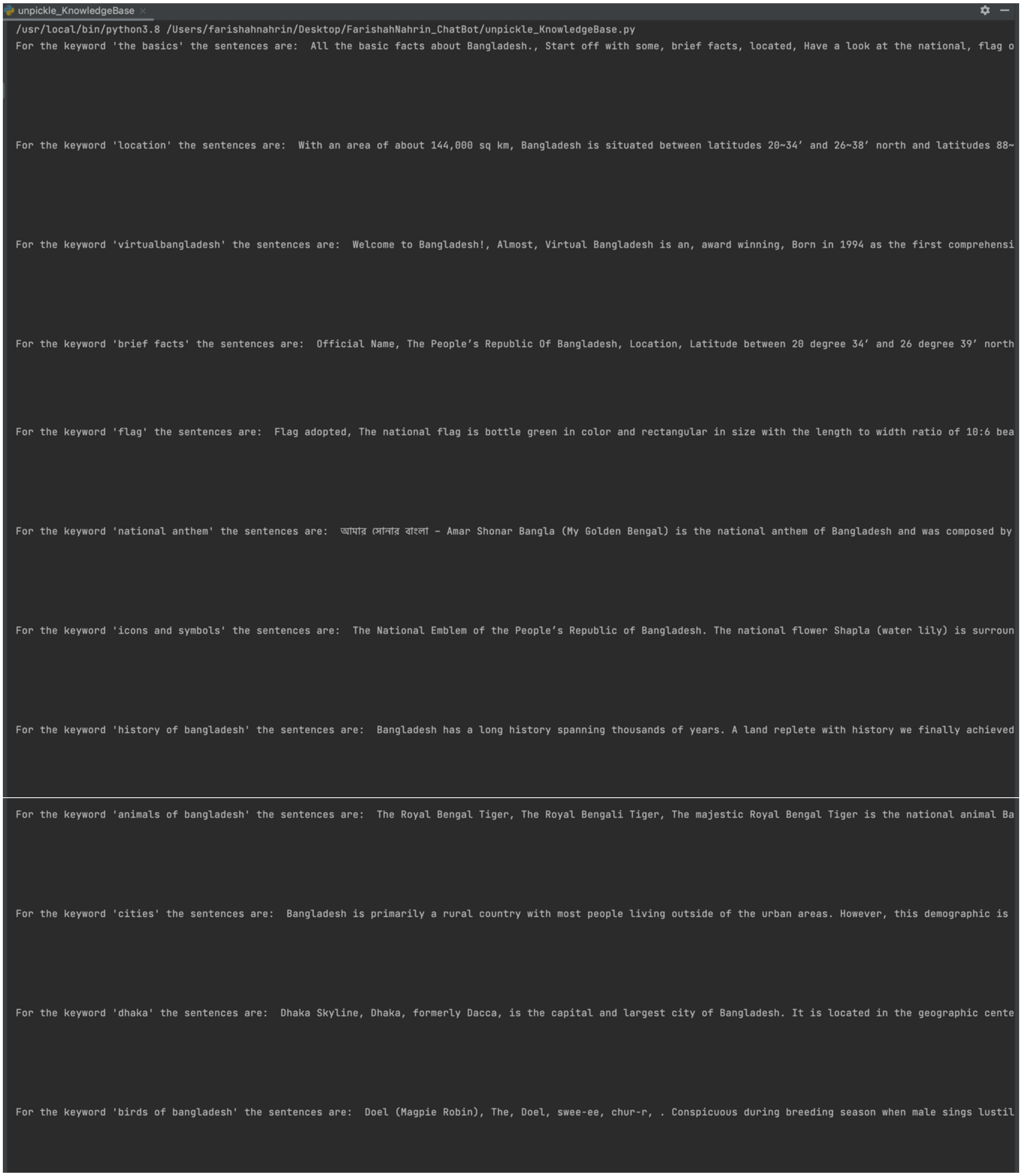
```
It is also one of the larger members of the deer family.

Choose an option:
1. Ask a query
2. See past queries
3. Exit
> 1
Please enter your query, make sure to ask about one topic at a time:
>Tell me about the Dhaka Medical College Hospital?

Knowledge base search failed. Trying Wikipedia...
Wikipedia primary search successful!

Dhaka Medical College and Hospital (abbreviated DMCH) is a public
medical college and hospital located in Dhaka, the capital city of
Bangladesh.
It houses medical school as well as a tertiary care hospital on one
campus.
The country's first ever autologous bone marrow transplant took place
in its bone marrow transplant unit.

Choose an option:
1. Ask a query
2. See past queries
3. Exit
> 3
Goodbye, Bob!
```

An appendix for the knowledge base (or live lookup) you created with samples.

The entire knowledge base was created from: https://www.virtualbangladesh.com/. Over 136 links from this website were scraped, and the keyword for the knowledge base was extracted from the link. For example, in the following links:

- https://www.virtualbangladesh.com/the-basics/brief-facts/
- https://www.virtualbangladesh.com/the-basics/location/
- https://www.virtualbangladesh.com/the-basics/
- https://www.virtualbangladesh.com/the-basics/national-anthem/
- https://www.virtualbangladesh.com/the-basics/flag/

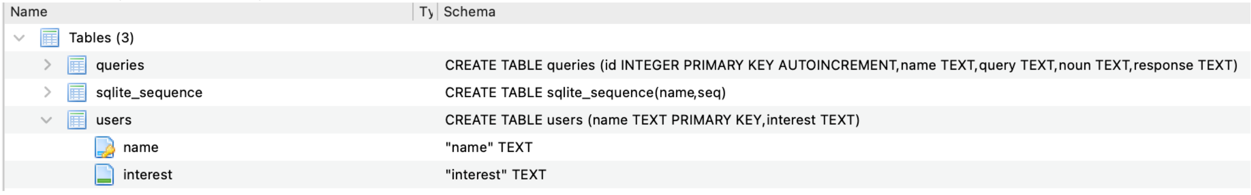
The last keyword(s) of the link was stored as a keyword (after pre-processing), and the link's content was stored as the value. All of this was stored in a pickle file. I had to simplify and modify my previous web scraper to achieve the optimal results. In the above links, the keys were “brief facts”, “location”, “the basics”, “national anthem”, and “flag”. I performed Jaccard similarity between the extracted keys and the noun extracted from the query sentence. Here is just a fraction of an example of my un-pickled knowledge base:



```
/usr/local/bin/python3.8 /Users/farishahnahrin/Desktop/FarishahNahrin_ChatBot/unpickle_KnowledgeBase.py
For the keyword 'the basics' the sentences are:  All the basic facts about Bangladesh., Start off with some, brief facts, located, Have a look at the national, flag o





For the keyword 'location' the sentences are:  With an area of about 144,000 sq km, Bangladesh is situated between latitudes 20~34’ and 26~38’ north and latitudes 88~





For the keyword 'virtualbangladesh' the sentences are:  Welcome to Bangladesh!, Almost, Virtual Bangladesh is an, award winning, Born in 1994 as the first comprehensi





For the keyword 'brief facts' the sentences are:  Official Name, The People’s Republic Of Bangladesh, Location, Latitude between 20 degree 34’ and 26 degree 39’ north




For the keyword 'flag' the sentences are:  Flag adopted, The national flag is bottle green in color and rectangular in size with the length to width ratio of 10:6 bea





For the keyword 'national anthem' the sentences are:  আমার সোনার বাংলা – Amar Shonar Bangla (My Golden Bengal) is the national anthem of Bangladesh and was composed by





For the keyword 'icons and symbols' the sentences are:  The National Emblem of the People’s Republic of Bangladesh. The national flower Shapla (water lily) is surroun





For the keyword 'history of bangladesh' the sentences are:  Bangladesh has a long history spanning thousands of years. A land replete with history we finally achieved




For the keyword 'animals of bangladesh' the sentences are:  The Royal Bengal Tiger, The Royal Bengali Tiger, The majestic Royal Bengal Tiger is the national animal Ba





For the keyword 'cities' the sentences are:  Bangladesh is primarily a rural country with most people living outside of the urban areas. However, this demographic is





For the keyword 'dhaka' the sentences are:  Dhaka Skyline, Dhaka, formerly Dacca, is the capital and largest city of Bangladesh. It is located in the geographic cente





For the keyword 'birds of bangladesh' the sentences are:  Doel (Magpie Robin), The, Doel, swee-ee, chur-r, . Conspicuous during breeding season when male sings lustil
```

An appendix for sample user models that were created

The entire USER model and the details are saved in a SQL database, with the following schema, using the SQLite3 Python library:

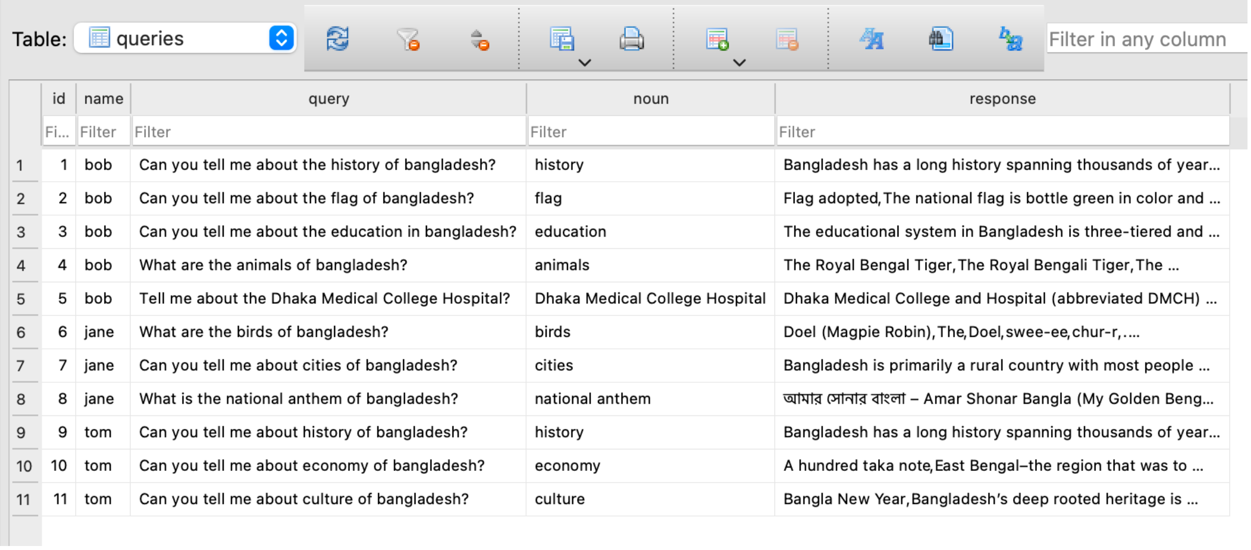


Name	Ty	Schema
Tables (3)		
queries		CREATE TABLE queries (id INTEGER PRIMARY KEY AUTOINCREMENT,name TEXT,query TEXT,noun TEXT,response TEXT)
sqlite_sequence		CREATE TABLE sqlite_sequence(name,seq)
users		CREATE TABLE users (name TEXT PRIMARY KEY,interest TEXT)
name		"name" TEXT
interest		"interest" TEXT

Here, each user's name is stored in the name column. And, each user's interest is stored in the interest column, which gets updated with every query that the user asks. For example: *UPDATE users SET interest=? WHERE name=?", (interest, name)*

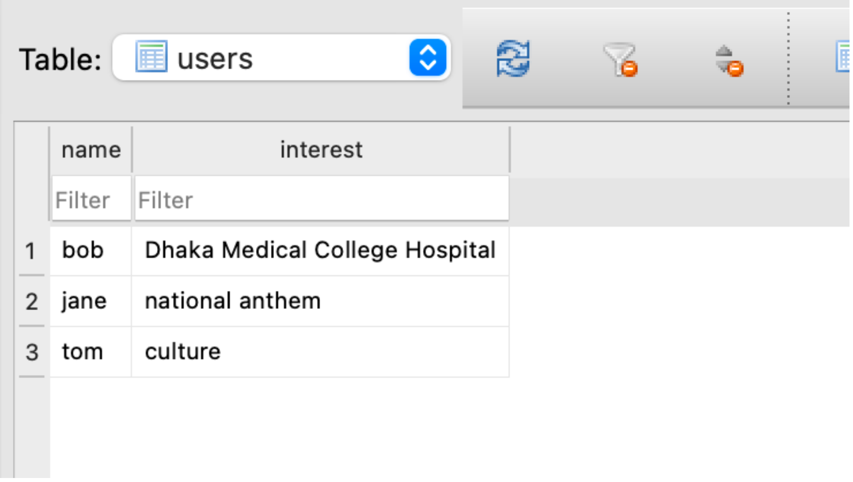
Sample Values from the user_data.db (database file):

This is the Queries Table:



	id	name	query	noun	response
	Fi...	Filter	Filter	Filter	Filter
1	1	bob	Can you tell me about the history of bangladesh?	history	Bangladesh has a long history spanning thousands of year...
2	2	bob	Can you tell me about the flag of bangladesh?	flag	Flag adopted,The national flag is bottle green in color and ...
3	3	bob	Can you tell me about the education in bangladesh?	education	The educational system in Bangladesh is three-tiered and ...
4	4	bob	What are the animals of bangladesh?	animals	The Royal Bengal Tiger,The Royal Bengali Tiger,The ...
5	5	bob	Tell me about the Dhaka Medical College Hospital?	Dhaka Medical College Hospital	Dhaka Medical College and Hospital (abbreviated DMCH) ...
6	6	jane	What are the birds of bangladesh?	birds	Doel (Magpie Robin),The,Doel,swee-ee,chur-r,....
7	7	jane	Can you tell me about cities of bangladesh?	cities	Bangladesh is primarily a rural country with most people ...
8	8	jane	What is the national anthem of bangladesh?	national anthem	আমার সোনার বাংলা – Amar Shonar Bangla (My Golden Beng...
9	9	tom	Can you tell me about history of bangladesh?	history	Bangladesh has a long history spanning thousands of year...
10	10	tom	Can you tell me about economy of bangladesh?	economy	A hundred taka note,East Bengal–the region that was to ...
11	11	tom	Can you tell me about culture of bangladesh?	culture	Bangla New Year,Bangladesh's deep rooted heritage is ...

This is the users table, where the main noun from the last-asked query of the user is saved as the user's latest "interest." And this is saved in the 'interest' column.



	name	interest
	Filter	Filter
1	bob	Dhaka Medical College Hospital
2	jane	national anthem
3	tom	culture

These are just examples of three users. Thus, my program can handle any number of users, making it a malleable and adjustable user model.

Evaluations of the chatbot and analysis of its strengths and weaknesses

Accuracy: The chatbot's accuracy was evaluated by comparing its responses, to the expected output, for each test case. The chatbot demonstrated high accuracy in retrieving relevant information from the knowledge base and Wikipedia. It could identify the main noun in the user's query and use it to search for appropriate information. Moreover, the implemented NLP techniques and information retrieval algorithms ensured that the chatbot could provide pretty accurate and relevant information to the user.

Strengths:

1. Comprehensive Knowledge Base: The chatbot's knowledge base, provides a strong foundation for answering user queries. This enables the chatbot to provide accurate and relevant information on various Bangladesh-related topics. For this chatbot project, I used the starter site that I used from my Web Crawler project, which is https://www.virtualbangladesh.com/. This website has almost all information, regarding facts about Bangladesh, which is why I think it's a relatively comprehensive knowledge base.

2. Integration with Wikipedia: By integrating the Wikipedia API into the chatbot, it can access vast information beyond the knowledge base. This ensures that the chatbot can provide up to date and extensive information on various topics, if it is not already found in the knowledge base.

3. Personalization: The chatbot's ability to store user interests and preferences allows for a more personalized user experience. This feature helps the chatbot prioritize more relevant responses to the user's interests. And the chatbot randomly pick up to 5 topics from the knowledge base *each time,* the program runs, when it is first introduced to the user, which provides a unique experience to someone who might not know how to start a conversation with the chatbot.

4. Error Handling: The chatbot program is designed to handle various errors gracefully, ensuring a smooth user experience even when issues arise during information retrieval or database operations. For example, if the user types a word or symbol (like a comma, or question mark) that doesn't exist in either the knowledge base or Wikipedia, then the program will ask the user to ask their question again. If the response is not in the knowledge base, then the Wikipedia API will pick up the question and display an answer from Wikipedia. These are a few examples of the error handling that I have added.

Weaknesses:

1. Chatbot has limited Natural Language Understanding: Although the chatbot implements various NLP techniques, it may still struggle with understanding complex or ambiguous queries. This could lead to the chatbot providing irrelevant or inaccurate information in certain cases. For example, with built-in Chatbot frameworks that exist today, such as Amazon Lex or Google Diagflow, the administrator/developer of the bot is allowed to customize the bot with training phrases in the intent; as in, these existing Chatbot frameworks can identify common spelling mistakes. Or the admin/developer is allowed to add variations to the training phrases of an

intent, such as related terms and synonyms. Then, these frameworks are able to adjust the training phrases, the more that the user can interact with the bot.

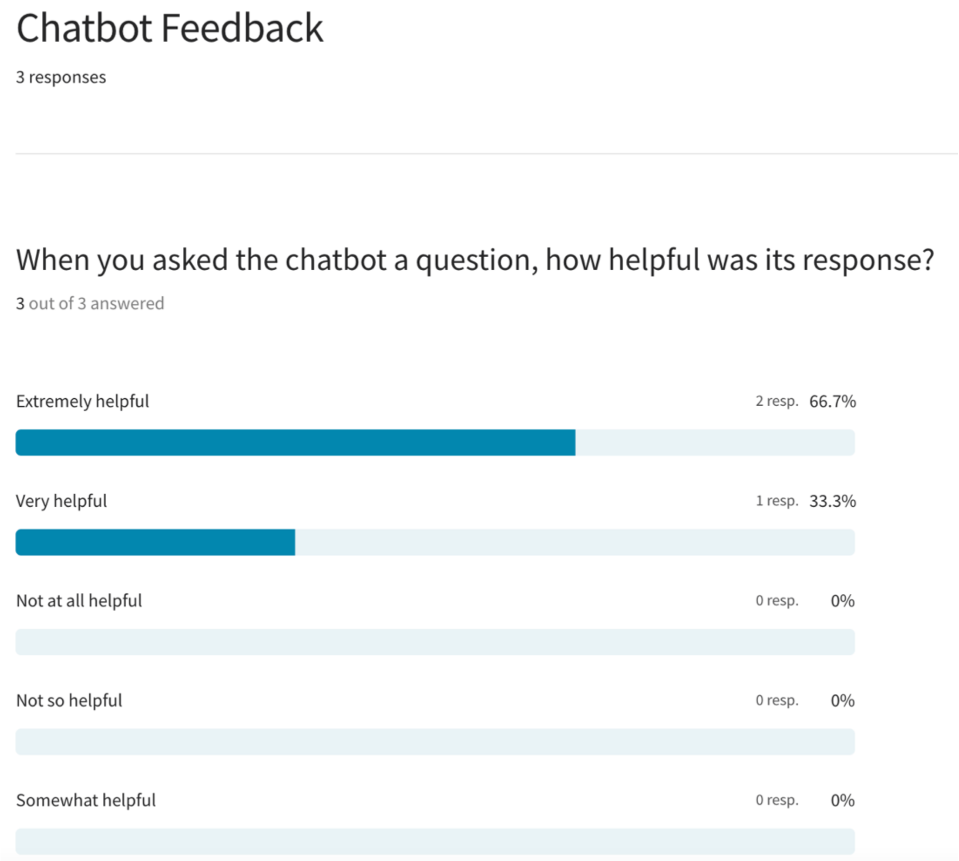
2. Dependency on External Sources: The chatbot's reliance on Wikipedia for information retrieval can be a double-edged sword. While it enables access to a vast amount of information, it also makes the chatbot vulnerable to inaccuracies or outdated information present on Wikipedia. Also this means that the chatbot can be used for queries that are not about Bangladesh, since the API has access to all pages in Wikipedia. Since this chatbot is “advertised” as a chatbot that only knows about Bangladesh, we are assuming that the user will include something specific to Bangladesh or mention the location of Bangladesh, in their query.

3. User Interface: The current user interface, being a terminal interface or within the IDE, may not be as user-friendly for all users. A more intuitive graphical user interface could improve the user’s experience.

4. Scalability: As the chatbot's knowledge base grows, using a serialized binary file, like .pkl, for storage may become less efficient for the future. Implementing a more scalable storage solution, such as a dedicated SQL database that contains the knowledge base, could improve performance and make managing the growing knowledge base easier.

In conclusion, the chatbot demonstrates strengths in its comprehensive knowledge base, integration with Wikipedia, personalization, and robust error handling. However, the chatbot has weaknesses in natural language understanding, dependency on external sources, the user interface, and scalability. Further improvements and enhancements could address these weaknesses and make the chatbot an even more powerful and user-friendly tool for providing information about Bangladesh.

Survey Results about the Bot - Three questions were asked:



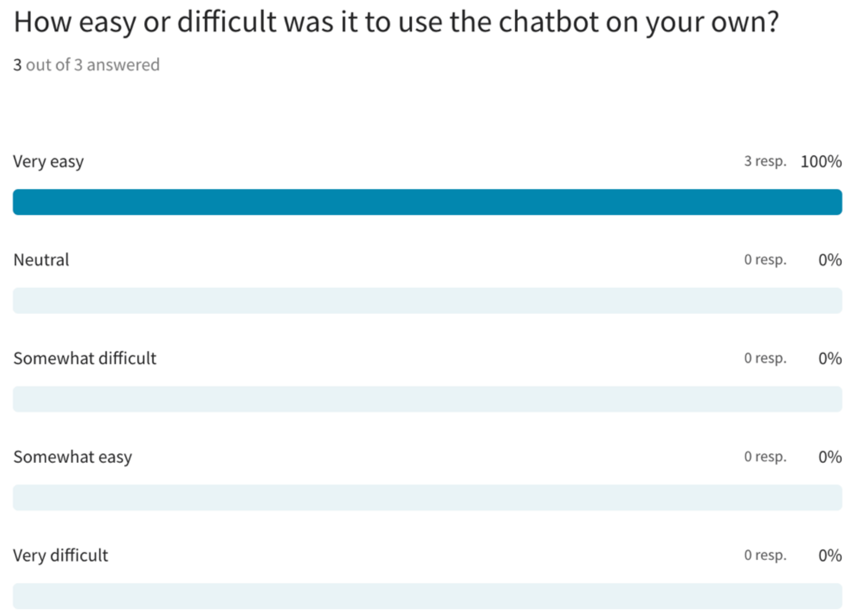
Chatbot Feedback

3 responses

When you asked the chatbot a question, how helpful was its response?

3 out of 3 answered

Extremely helpful	2 resp.	66.7%
Very helpful	1 resp.	33.3%
Not at all helpful	0 resp.	0%
Not so helpful	0 resp.	0%
Somewhat helpful	0 resp.	0%



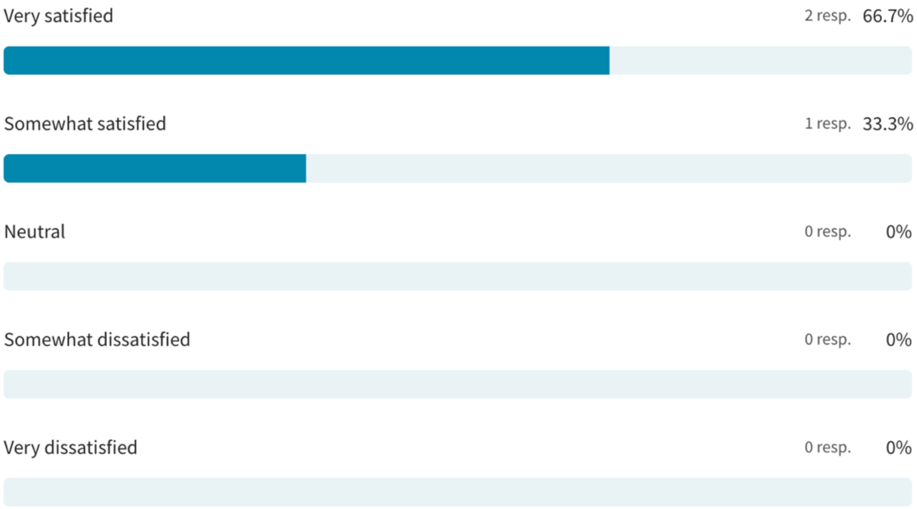
How easy or difficult was it to use the chatbot on your own?

3 out of 3 answered

Very easy	3 resp.	100%
Neutral	0 resp.	0%
Somewhat difficult	0 resp.	0%
Somewhat easy	0 resp.	0%
Very difficult	0 resp.	0%

How satisfied or dissatisfied were you with the performance of the chatbot?

3 out of 3 answered


Very satisfied
2 resp.
Somewhat satisfied
1 resp.
Neutral
0 resp.
Somewhat dissatisfied
0 resp.
Very dissatisfied
0 resp.