Leo Nguyen - ldn190002

Amol Perubhatla - AVP180003

Chatbot-Report and Evaluation

1. Introduction

- Our chatbot is a ruled-base system bot which will provide basic information on cyberpunk genre based on the user's querry. This is a basic chatbot, so all of our data was collected from https://www.neondystopia.com/.
- How to run:
    - Open the command-line
    - Type in py chatbot.py

2. System Description

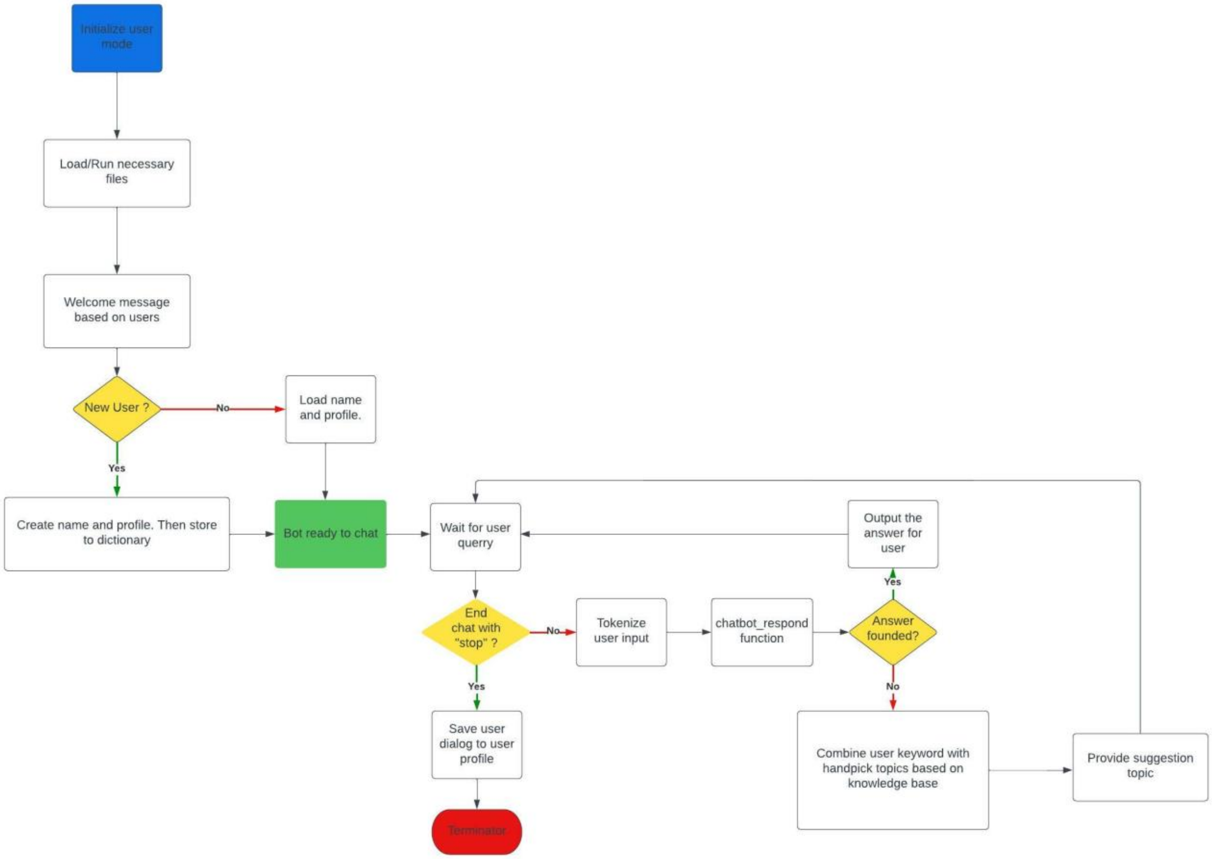
- The chatbot system will require 3 files: chatbot.py, knowledge_base.py, and stopwords.txt
- The stopword.txt will contain extensive English stopword list from kaggle.com
- The chatbot.py create and maintain the user's interface. It will start by asking the username. If it is a new user the program will create a new user's profile. If it is an existed user the program will open user profile.
- For the system to be able to provide the basic answers regarding cyberpunk related topic. It will need a knowledge base. This is when the knowledge base become handy. If the knowledge base file "k_base.p" already exists on the the system, it will autoload by the "chatbot.py" if not, the chatbot file will call the "knowledge_base.py" to extract the data from https://www.neondystopia.com/ and create the "k_base.p" file.

- The chatbot utilize serveal NLP techniques:
    - To collect and create the knowledge base raw data, the sytem use **webcrawling** and **webscraping** techniques. First, it will perform the webcrawling from the stater url https://www.neondystopia.com/ to extract 10 most relavent urls from the starter url. Then for each relavent url, a raw data will be collected by webcrapping technique.
    - To be able to use the knowledge base, this draw data need to be preprocess to remove unnecessary elements such as extra space and special characters using **regex** technique Then use the **sent_tokenize** technique to break into separate sentences. Finally, save and output this sentences collection into a **pickle** file to be used for the next step. This help to reduce the time for chatbot systems to initialize.
    - After the knowledge base pickle is ready. The system will be intialized with different user mode. Each user will be maintained by name and user profile using **dictionary** techique with username is the key, and 2 values: name and user's keyword.
    - The system is ready when the user data was created/loaded and the knowledge base file loaded. Each user input will be tokenized by **word_tokenizer** technique then further processing to remove the **stopword** and none alphabet characters. Then the final keyword will be feed into the chatbot respond function to find the most relavant to the querry.
    - To be able to find the relavant answer for the querry, the system utilize 2 main techniques: **TfidfVectorizer** and **cosine_similarity**. The first technique- TfidfVectorizer will convert the corpus of knowledge base into vector of

sentences based on their frequency and weighting that frequency in term of the level of impact. Then use the cosine_similariry technique to compare the user input with with sentences vectors to find the most relavant to the keyword.

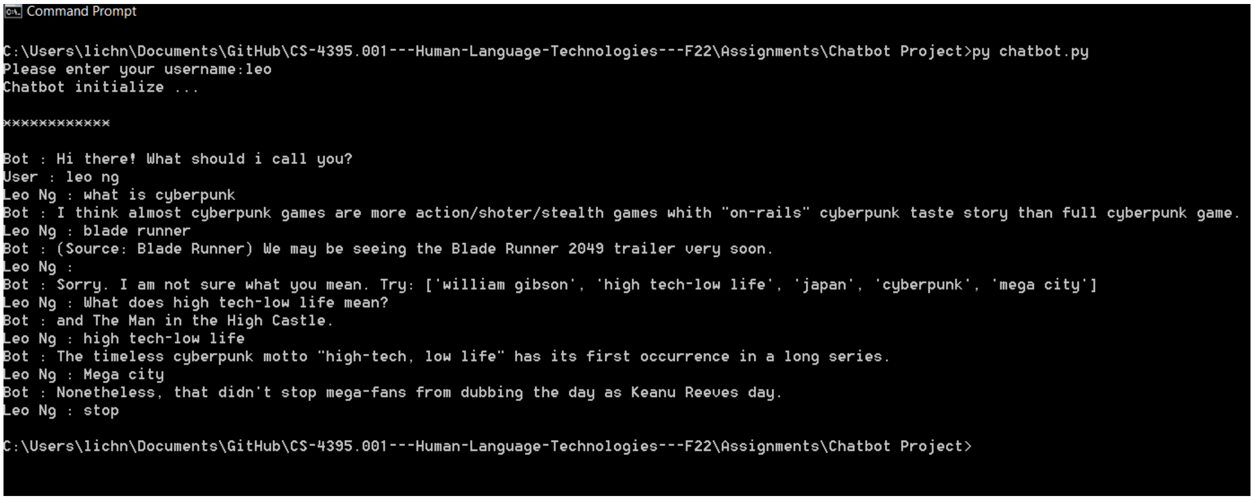
- Finally, print the the most relavant answer for the querry. The system will then save this querry into user profile in user dictionary. So it can provide better suggestions when user come back next time.

3. Diagram of chatbot dialog tree and logic.



4. Sample dialog interaction

- New user mode:



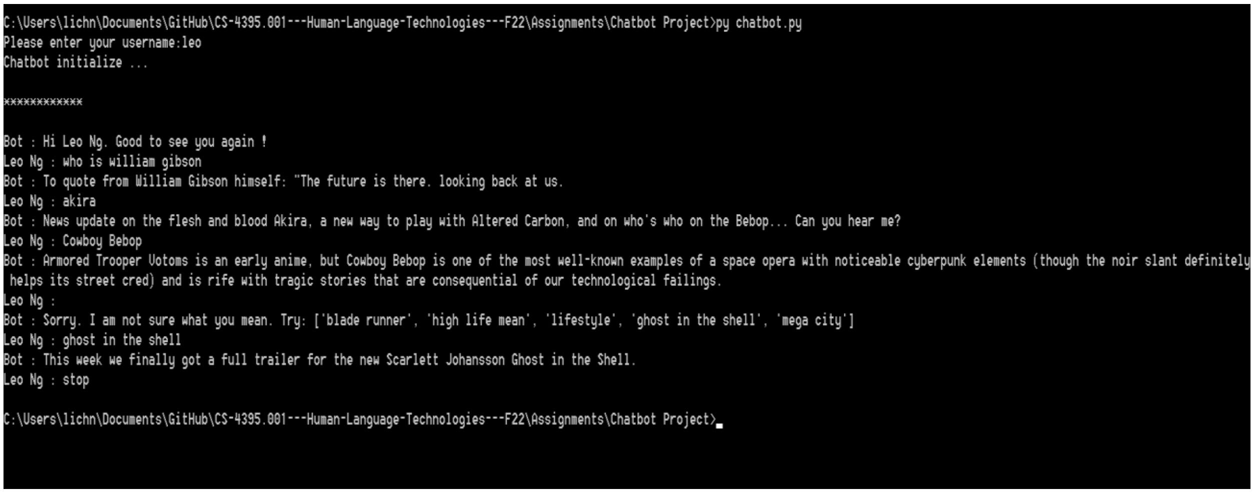
```
C:\Users\lichn\Documents\GitHub\CS-4395.001---Human-Language-Technologies---F22\Assignments\Chatbot Project>py chatbot.py
Please enter your username:leo
Chatbot initialize ...

************

Bot : Hi there! What should i call you?
User : leo ng
Leo Ng : what is cyberpunk
Bot : I think almost cyberpunk games are more action/shoter/stealth games whith "on-rails" cyberpunk taste story than full cyberpunk game.
Leo Ng : blade runner
Bot : (Source: Blade Runner) We may be seeing the Blade Runner 2049 trailer very soon.
Leo Ng :
Bot : Sorry. I am not sure what you mean. Try: ['william gibson', 'high tech-low life', 'japan', 'cyberpunk', 'mega city']
Leo Ng : What does high tech-low life mean?
Bot : and The Man in the High Castle.
Leo Ng : high tech-low life
Bot : The timeless cyberpunk motto "high-tech, low life" has its first occurrence in a long series.
Leo Ng : Mega city
Bot : Nonetheless, that didn't stop mega-fans from dubbing the day as Keanu Reeves day.
Leo Ng : stop

C:\Users\lichn\Documents\GitHub\CS-4395.001---Human-Language-Technologies---F22\Assignments\Chatbot Project>
```

- Existed user mode:



```
C:\Users\lichn\Documents\GitHub\CS-4395.001---Human-Language-Technologies---F22\Assignments\Chatbot Project>py chatbot.py
Please enter your username:leo
Chatbot initialize ...

************

Bot : Hi Leo Ng. Good to see you again !
Leo Ng : who is william gibson
Bot : To quote from William Gibson himself: "The future is there. looking back at us.
Leo Ng : akira
Bot : News update on the flesh and blood Akira, a new way to play with Altered Carbon, and on who's who on the Bebop... Can you hear me?
Leo Ng : Cowboy Bebop
Bot : Armored Trooper Votoms is an early anime, but Cowboy Bebop is one of the most well-known examples of a space opera with noticeable cyberpunk elements (though the noir slant definitely
 helps its street cred) and is rife with tragic stories that are consequential of our technological failings.
Leo Ng :
Bot : Sorry. I am not sure what you mean. Try: ['blade runner', 'high life mean', 'lifestyle', 'ghost in the shell', 'mega city']
Leo Ng : ghost in the shell
Bot : This week we finally got a full trailer for the new Scarlett Johansson Ghost in the Shell.
Leo Ng : stop

C:\Users\lichn\Documents\GitHub\CS-4395.001---Human-Language-Technologies---F22\Assignments\Chatbot Project>
```

5. Appendix for Knowledge base

- Data was mainly collected from https://www.neondystopia.com/
- Ten main urls extracted from starter url:
  1. https://www.neondystopia.com/cyberpunk-movies-anime/
  2. https://www.neondystopia.com/cyberpunk-politics-philosophy/
  3. https://www.neondystopia.com/cyberpunk-books-fiction/
  4. https://www.neondystopia.com/cyberpunk-games/
  5. https://www.neondystopia.com/cyberpunk-music/
  6. https://www.neondystopia.com/cyberpunk-technology/
  7. https://www.neondystopia.com/cyberpunk-art-photography/
  8. https://www.neondystopia.com/cyberpunk-fashion-lifestyle/
  9. https://neondystopia.com/cyberpunk-games-database/
  10. https://www.neondystopia.com/what-is-cyberpunk/
- Then from each url above, the system will extract 10 more urls and scrape the data from that 10 urls to build the knowledge base

6. Appendix for user mode

- Each user will be stored to *dict_username.p* with username as key. Each key contains 2 values: name and user's profile. The user profile will content a string seperated by underscore _ for each keyword user had been used before.
- User mode with 2 user currently:leo and amol with their name and profile perspectedly {'leo': ['leo ng', '_cyberpunk_blade runner_high life mean_mega city_william gibson_akira_cowboy bebop_ghost shell'], 'amol': ['amol per', '_matrix_cyberspace_corporation_cyberpunk lifestyle']}

7. Chatbot Evaluation

- **Strength:** It is a very simple rule-based chatbot system. If the knowledge base file already exist, it can run by itself without internet or help from other system, API. The system have create and maintain their own files for both knowledge base and user mode file. It can run offline most of the time. The chatbot only require internet to crawling and scraping the starter url to create the knowledge base.
- **Weakness:** Because the system run mainly from the pre collected knowledge base, there is a limited number of answer it provide based on the size of the knowledge base file.