

NLP ChatBot Report

Web Crawler

- How I created the Knowledge Base?

First and foremost was the topic selection:- I was trying to find a topic that might not need much data or isn't too sparse. I jumped multiple times, from stock market, Indian Cuisine, Spices, and then finally to Bollywood.

So, I mainly went by the book and created a web crawler, using 3 starting links for the crawler to start from. Each page was further scrapped and the data in 'p' tags were stored in files, and as for subsequent links, I created logic so that only relevant links are taken into consideration.

Logic -> firstly I allowed my scraper to collect a maximum of 10 links from each page, secondly I tested with many keywords in the link url, for eg- 'hindi', 'bollywood', 'khan', etc.. and also keeping a hard limiter on the maximum number of links to be crawled. In total I crawled to 17 relevant URLs (the number would have gone a lot higher if I didn't put a limiter on the type of links that the crawler can access).

The data stored in files are then pulled up and preprocessed - mainly removal of stopwords and punctuation, tokenization, Lemmatization and then I stored these cleaned up files.

Finally, using the Tf-Idf algorithm taught in class and modifying it a bit to work with the dimensions of my data, I computed the top important terms (terms with high Tf-Idf value). I then hand-picked 15 top terms to get rid of unnecessary words.

The Knowledge Base is pretty simple:- It is a dictionary of key: value pairs, where the keys are the important terms and the value is a list of processed sentences containing the term. Processed sentences are raw sentences along with tokenization, removal of '\n' and '[number]' characters to keep the knowledge base as clear as possible. Why did I do the processing of sentences before storing in the knowledge base? Because the output of the chatbot was very poor and more irrelevant if not done.

The knowledge base was pickled and stored to be used in chatBot notebook.

- Screenshot of the Knowledge Base

```
[20]: KB

[20]: defaultdict(list,
      {'horror': ['hindi-language horror films have been a subgenre of the hindi film industry in india since the birth of hindi film
ms.',
                 'these films tend to mimic the characteristics and themes of horror films around the world.',
                 'this is a list of indian horror films in hindi language.',
                 'the film succeeds in making us all witnesses to its horrors – and complicit in them when, like so many of the characters, w
e say nothing.',
                 'the big scenedevi and her gang attack a wedding party in search of recompense for the horrors she suffered in her youth.',
                 'hindi-language horror films have been a subgenre of the hindi film industry in india since the birth of hindi films.',
                 'these films tend to mimic the characteristics and themes of horror films around the world.',
                 'this is a list of indian horror films in hindi language.'],
      'film': ['hindi-language horror films have been a subgenre of the hindi film industry in india since the birth of hindi film
s.',
              'these films tend to mimic the characteristics and themes of horror films around the world.',
              'this is a list of indian horror films in hindi language.',
              'hindi cinema, popularly known as bollywood and formerly as bombay cinema, refers to the film industry based in mumbai, enga
ged in production of motion pictures in hindi language.',
              'the industry is a part of the larger indian cinema, which also includes south indian cinema and other smaller film industri
es.',
              'in 2017, indian cinema produced 1,986 feature films, of which the largest number, 364 have been in hindi.',
              'hindi cinema is one of the largest centres for film production in the world.',
              'hindi films sold an estimated 341 million tickets in india in 2019.',
              'earlier hindi films tended to use vernacular hindustani, mutually intelligible by speakers of either hindi or urdu, while
modern hindi productions increasingly incorporate elements of hinglish.',
              'the most popular commercial genre in hindi cinema since the 1970s has been the masala film, which freely mixes different ge
nres including action, comedy, romance, drama and melodrama along with musical numbers.',
              'masala films generally fall under the musical film genre, of which indian cinema has been the largest producer since the 1
960s when it exceeded the american film industry's total musical output after musical films declined in the west.',
              'dadasaheb phalke's silent film raja harishchandra (1913) is the first feature length film made in india.',
              'the first indian musical talkie was alam ara (1931), four years after the first hollywood sound film the jazz singer (192
7).',
              'alongside commercial masala films, a distinctive genre of art films known as parallel cinema has also existed, presenting r
ealistic content and avoidance of musical numbers.',
              'in more recent years, the distinction between commercial masala and parallel cinema has been gradually blurring, with an in
creasing number of mainstream films adopting the conventions which were once strictly associated with parallel cinema.'],
      'persian': []})
```

- My 15 important terms are:-

ImpTerms = ['horror', 'film', 'hindi', 'bollywood', 'kamal', 'drama', 'khan',
'hindustani', 'bhangra', 'director', 'shows', 'devanāgarī', 'indian', 'sanskrit',
persian]

Chat Bot

- System Description

For the Chat bot, first I unpickled the Knowledge Base file and the User Model file (if it exists, else created a user model file). Defined a few functions that I am going to use, for eg- Greetings, Text Pre-processing, Responses, Extracting user's name and so on.

For Greetings, I am looking for specific keywords such as 'hi', 'hey' and then output a random greetings response.

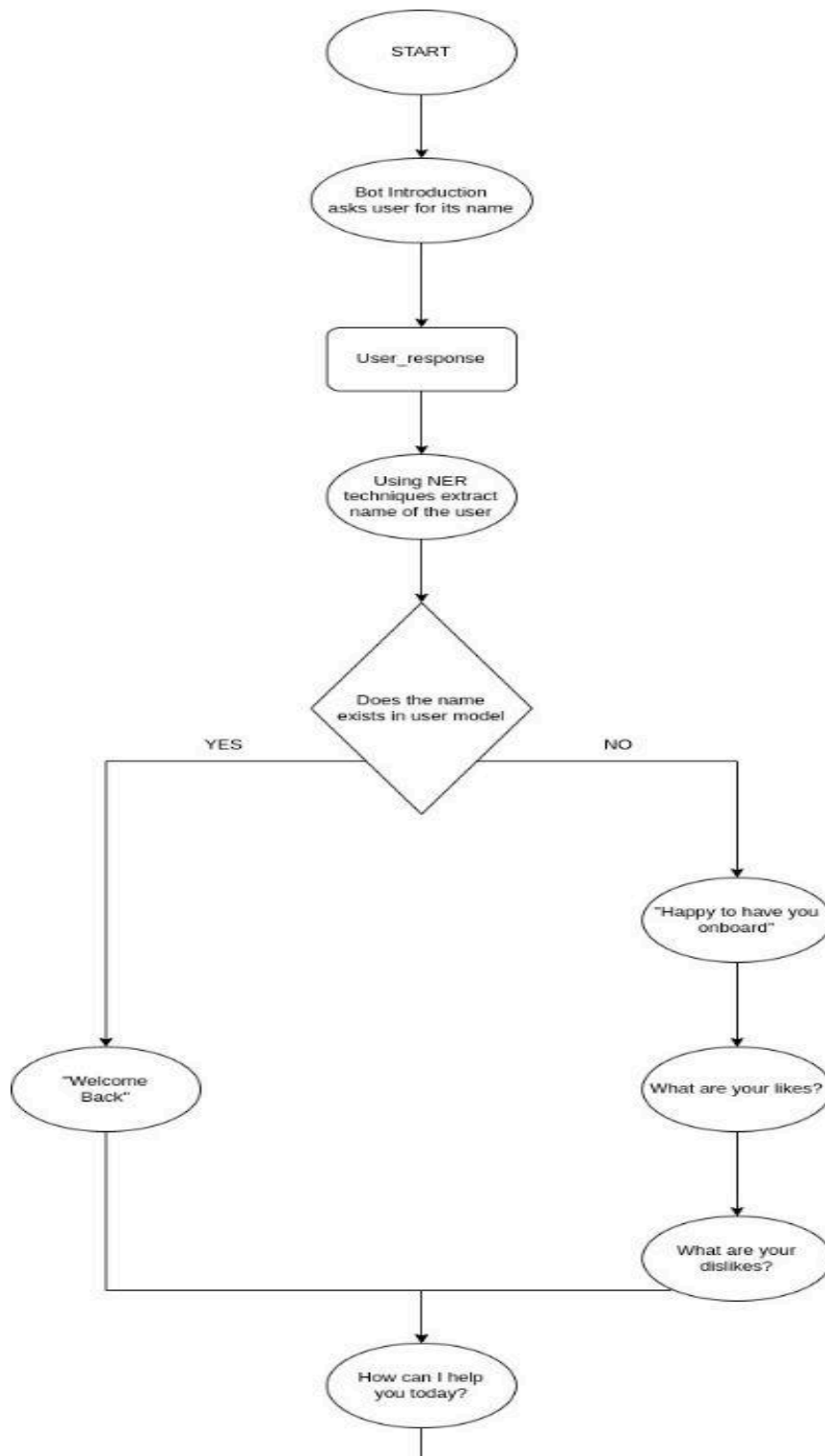
The main NLP techniques that I used are:- Named Entity Recognition (NER) to extract user's name from user input, Tf-Idf and Cosine Similarity to find sentences that are the most relevant to the user's input, Tokenization and Lemmatization to process the user input. Used WordNetLemmatizer to process text before creating the Tf-Idf matrix. I also used sklearn's Tf-Idf and Cosine similarity instead of hand coding these functions.

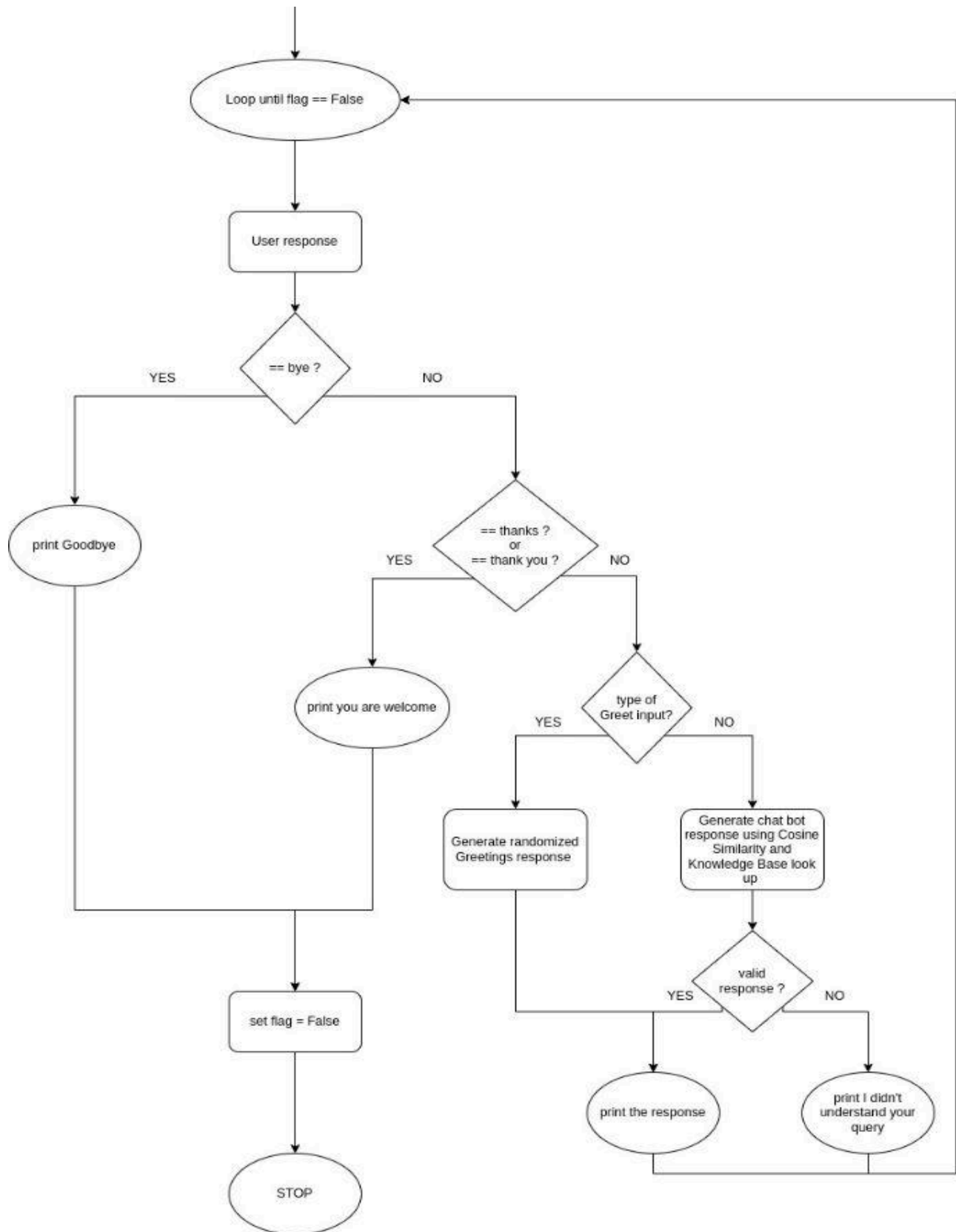
Stored the User Model as simple dictionaries:- Key is user name and Value is again a dictionary of "likes" : " " and "dislikes" : " " to store the likes and dislikes of users as we receive them.

This helps in remembering the user, and ask questions to know more about the user. I am also incorporating an external knowledge base that was created using the web crawling part.

If the user is not present in the userModel file then we will get the details and then save it in the file else we will directly start conversing with the FilmyBot.

- Diagram of dialog tree/logic





- Some Sample dialog interactions

```
[92]: # Call Filmy bot using the following command
FilmyBot()

Hello! I am FilmyBot, your guide on Bollywood. For ending the convo type bye!
For starters, what is your name?
samarth
Bot: Hi samarth. Happy to have you onboard this very Dhamakedar experience!
Bot: What do you like about Bollywood?
i like salman
Bot: Ok nice! Now what do you dislike about Bollywood?
I dislike ranvir
Bot: Fantastic! I am happy to learn more about you. Blast off any questions who have
do you have any recent moview?
Bot: I am sorry, unable to understand you.
moveis
Bot: I am sorry, unable to understand you.
movies
Bot: I am sorry, unable to understand you.
films?
Bot: to release their film theatrically or online in the country, every film maker first apply for film certification to the central board o
f film certification (cbfc) along film print, only after receiving cbfc certificate a film trailer or a film can be released in india, membe
rs of cbfc view the film, give rating-age restriction, suggest cuts on objectionable scenes or can ban the film from exhibition in anywhere
in the country.
thanks
Bot: You are Welcome..
```

Here you can see that FilmyBot is first asking the user for its name and finds out it is the user's first time and prompts the user to populate the user model with likes and dislikes. Next the user tries to ask questions to the bot, since the 'movie' is not in the knowledge base important term, the chatbot outputs "I am sorry, unable to understand you." On the other hand, since there is 'films' term in the knowledge base it outputs the sentence with the highest cosine similarity. Finally, it ends when the user says 'thanks'.

```
[28]: # Call Filmy bot using the following command
FilmyBot()

Bot: Hello! I am FilmyBot, your guide on Bollywood. For ending the convo type bye!
Bot: For starters, what is your name?
I am Harshil
Bot: Hi harshil. Welcome back! Happy to chat more about bollywood with you today!
can you tell me something about bollywood?
Bot: varun groverdirector: vidhu vinod chopracast: jackie shroff, anil Kapoor, nana patekargenre: crime, drama, romancethe quintessential bo
llywood crime film, 'parinda' tells of street-kid brothers kishen (shroff) and karan (kapoor).
how about khan
Bot: salman khan's Bajrangi Bhaijaan and Irrfan Khan's Hindi Medium were also Chinese hits in early 2018.
some Sanskrit shows
Bot: the international alphabet of Sanskrit transliteration (IAST) is the academic standard for the romanisation of Sanskrit.
bye
Bot: Goodbye!
```

In this interaction, the user is someone who has visited the filmy-bot before, and hence the chat bot welcomes the user and asks how it can help today. Then there are some questions asked by the user and responses generated by the chatbot using cosine similarity. Finally, it ends when the user says 'bye'.

- Evaluation of the Chatbot

Strengths:

- Personally, I see my chatbot's strength in how I have personalized the responses, for eg- Extracting the user name and greeting them by their name.
- It can handle various user queries and responses, thanks to the dynamic response generation mechanism. It looks up the knowledge base to output responses to the user queries.

Weakness:

- The main weakness that I can see is the vastness and depth of the knowledge base. Being a small rules based chatbot, it becomes pretty monotonous in generating responses.
- If it would have been a ML based chatbot or having created a larger knowledge base and to incorporate some language generation model such as ngram model, the response might become more dynamic but there would be a tradeoff with the sensibility of the responses.

As for the survey, I asked the below 3 questions to my friends and family:-

1. On a scale of 1 to 5, how satisfied are you with the overall performance of FilmyBot?
Average -> 3.5
2. Rate your satisfaction with the personalized interaction provided by FilmyBot. (Scale: 1 to 5)
Average -> 4
3. Rate the accuracy of the information provided by FilmyBot. (Scale: 1 to 5)
Average -> 3

Everyone loved the personalized interactions, but had varied responses for the performance and accuracy of the information provided. Some of the chatbot responses were accurate and catered to the user response while some were more ambiguous.