

## PROJECT REPORT

### PART 1: Web Crawler

For this project, I developed a web crawler to create a knowledge base about the most anticipated movies of 2024. The knowledge base was built by scraping relevant information from various websites, primarily "The Week" and IMDb.

#### Knowledge Base Creation:

1. **Web Scraping:** The web scraper was developed using Python's requests and BeautifulSoup libraries. The initial URL from "The Week" (<https://theweek.com/culture-life/film/2024-film-releases-most-anticipated-new-movies>) was used as a starting point to extract a list of movie titles and other relevant URLs.
2. **Data Extraction:** The scraper followed the extracted URLs, parsing the HTML content to collect movie-related information such as cast, directors, writers, ratings, runtime, release year, and IMDb ratings. This information was gathered from both "The Week" and IMDb.
3. **Text Cleaning:** The extracted text was cleaned by removing stopwords, punctuation, and performing lemmatization using the nltk library. This process aimed to retain only the most important words for further analysis.
4. **Term Importance:** The TfidfVectorizer from the scikit-learn library was used to identify the most important terms in the collected text. The top 50 important terms were generated out of which top 10 are presented:  
  
['2024', 'alien', 'also', 'cast', 'dune', 'film', 'first', 'gladiator', 'movie', 'new', 'one', 'part', 'role', 'star', 'two']
5. **Important Terms:** The final important terms that I manually picked for the knowledge base were to use the movie titles as the keys for the knowledge so that the database can be organized in a more proper manner and the information fetching process can be eased out. Below are the terms chosen as keys for the database dictionary:  
  
[Dune Part Two, The Taste of Things, The Iron Claw, The Zone of Interest, All of Us Strangers, The Holdovers, Society of the Snow, movies list, Other movies reviews]
6. **Knowledge Base Construction:** The extracted information was organized into a Python dictionary, where each movie title served as a key, and the associated details (cast, directors, writers, rating, runtime, year, and IMDb rating, reviews) were stored as values.

7. **Persistence:** The final knowledge base was serialized and saved using Python's pickle module for later use by the chatbot.
8. Screenshots of the Knowledge Base:

The screenshot displays a Windows File Explorer window with the address bar showing the path: `PS D:\UID\Spring 2024\MLP\Project 1>`. The left sidebar shows the 'PROJECT 1' directory structure, with the 'data' folder expanded. The 'data' folder contains 34 files named `filename1.txt` through `filename34.txt`. The right sidebar shows the 'OUTLINE' and 'TIMELINE' sections. The 'OUTLINE' section lists the following items:

- PROBLEMS
- OUTPUT
- TERMINAL
- PORTS
- SEARCH ERROR
- DEBUG CONSOLE

The 'TIMELINE' section shows the following items:

- chatotpy
- database.pkl
- DEscription.pdf
- user\_model.json
- web\_scrapper.py

The main pane displays the contents of the selected file, `filename1.txt`, which contains the following text:

```

TOPIC: BEST MOVIES 2024
Website: The Week & IMDb(for some facts)

LETS START SCRAPPING THE WEB

!!!Please be patient each process takes time!!!

Got 36 URLs, some of them are:
["https://collider.com/mufasa-the-lion-king-prequel-new-release-date-december-2024/", "https://eu.usatoday.com/story/entertainment/movies/2023/11/13/garfield-movie-2024-chris-pratt-summer-release/1570140007/", "https://go.redirectingat.com/?id=92X1679923&cust=theweek.us_122643964539030215&xs=1&url=https%3A%3Fwww.thesetimes.co.uk%2Farticle%2Fthe-holdovers-review-a-same-bug-from-the-team-behind-sideways-066b6b8d&ref=https%3A%3Ftheweek.com%2Fculture-life%2Ffilm%2F2024-film-releases-most-anticipated-new-movies/", "https://go.redirectingat.com/?id=92X1679923&cust=theweek.us_33827899286816612&xs=1&url=https%3A%3Fwww.thesetimes.co.uk%2Farticle%2Fthe-2-one-of-interest-revies-holocaust-film-masterfully-reinvents-martin-scorsese-novel-yggdrasil&ref=https%3A%3Ftheweek.com%2Fculture-life%2Ffilm%2F2024-film-releases-most-anticipated-new-movies/", "https://go.redirectingat.com/?id=92X1679923&cust=theweek.us_52561280362016551&xs=1&url=https%3A%3Fwww.thesetimes.co.uk%2Farticle%2Ffun-e-part-two-film-revies-lmthe-hes-chalmet-austin-butler-zendaya-gwr2d8&ref=https%3A%3Ftheweek.com%2Fculture-life%2Ffilm%2F2024-film-releases-most-anticipated-new-movies"]

Data Extracted from the urls and stored in the same folder
Cleaned the text for getting important words

Here's the list of 10 important words from the text
['2024', 'alien', 'also', 'cast', 'dune', 'film', 'first', 'gladiator', 'movie', 'new', 'one', 'part', 'role', 'star', 'two']

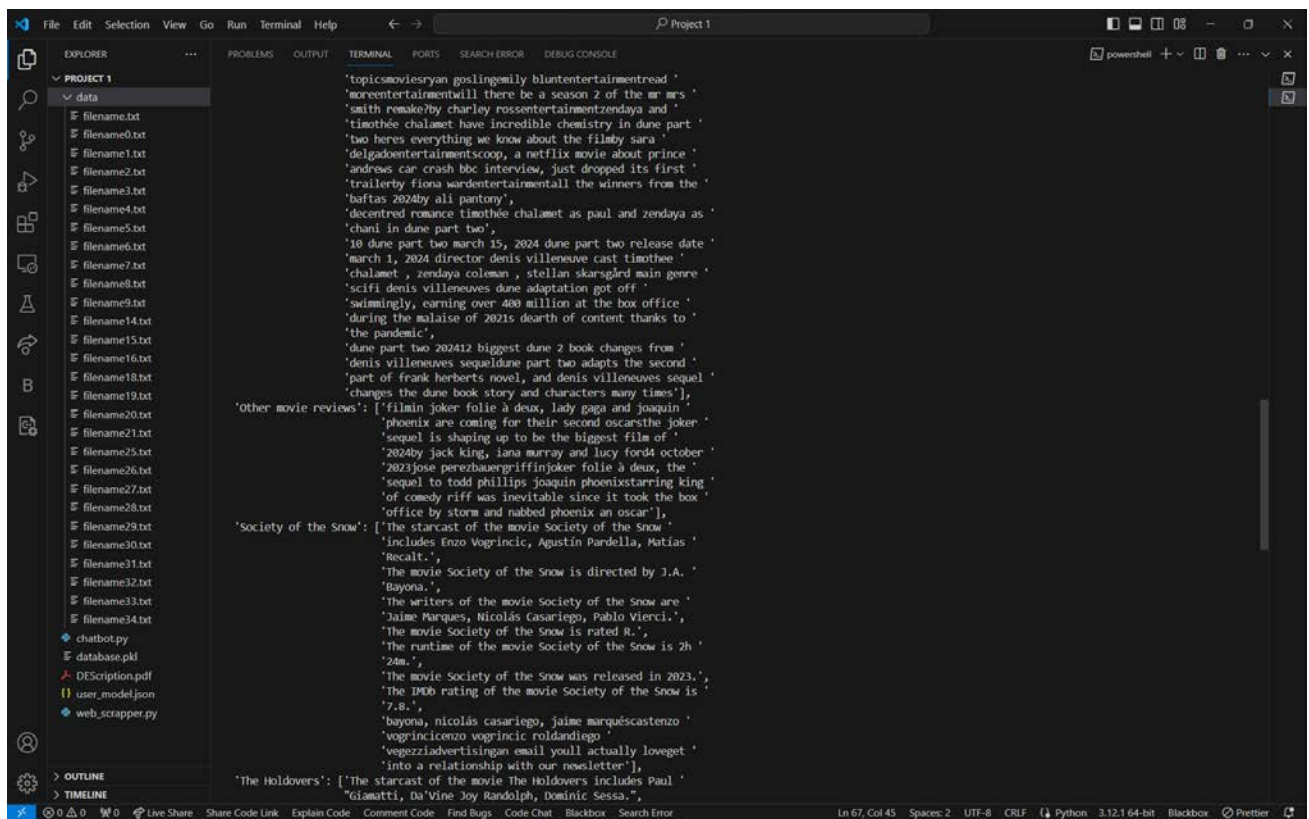
We are going with the movie titles as it organizes the knowledge base neatly

Now lets get some facts from the imdb site that will help in filtering info from websites that we scraped earlier

Lets add these facts to the database/ knowledge base

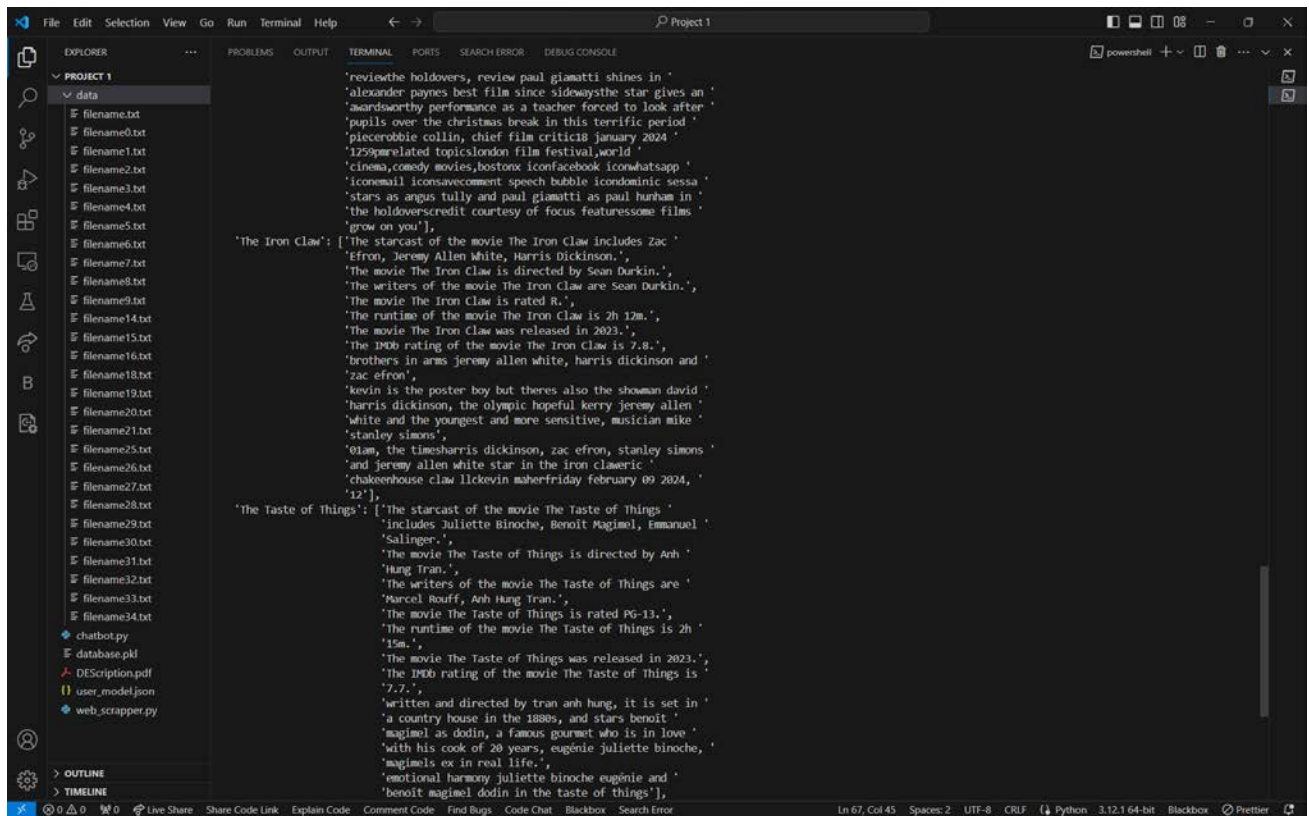
Database created successfully.
Now Lets Add content to the database

Finally, after a long period of processing...
Dumping the database into pickle file for chatbot usage
Here is your Movie Database:
  
```



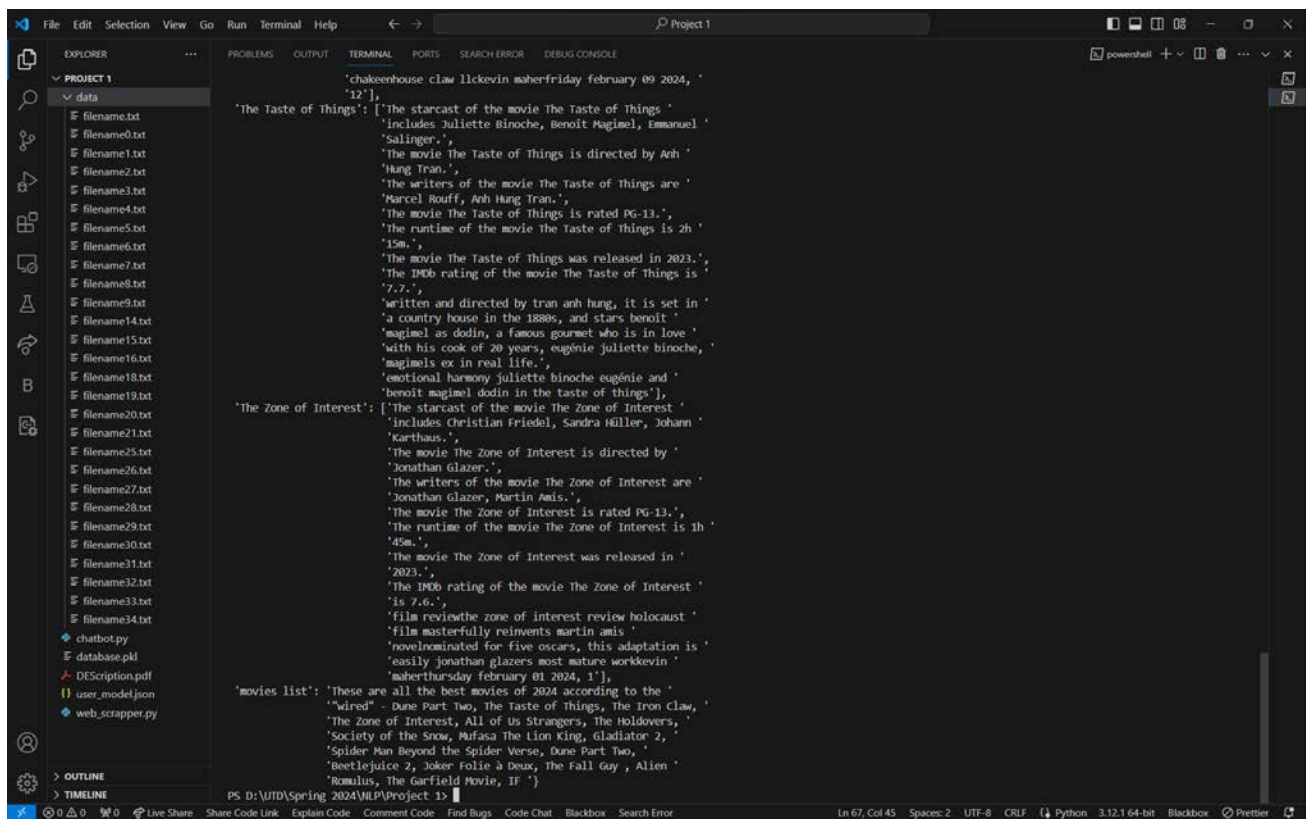
The screenshot shows a Visual Studio Code editor interface. The Explorer panel on the left displays a project named 'PROJECT 1' with a 'data' folder containing 34 files named 'filename0.txt' through 'filename34.txt'. Below the folder, there are four files: 'chatbot.py', 'database.pkl', 'DEscription.pdf', and 'user\_model.json'. The Terminal panel on the right shows a PowerShell session with a long string of text representing movie reviews. The text is wrapped across multiple lines. The status bar at the bottom indicates 'Ln 67, Col 45', 'Spaces: 2', 'UTF-8', 'CRLF', 'Python', '3.12.1 64-bit', 'Blackbox', and 'Prettier'.

```
'topicsmoviesryan goslingemily bluntenertainmentread '
'moreentertainmentwill there be a season 2 of the me mrs '
'smith remake?by charley rossentertainmentzendaya and '
'timothee chalamet have incredible chemistry in dune part '
'two heres everything we know about the filmy sara '
'delgadoentertainmentscoop, a netflix movie about price '
'andrews can crash bbc interview, just dropped its first '
'trailerby fiona wardentertainmentall the winners from the '
'baftas 2024by ali pantony',
'decentred romance timothee chalamet as paul and zendaya as '
'chani in dune part two',
'10 dune part two march 15, 2024 dune part two release date '
'march 1, 2024 director denis villeneuve cast timothee '
'chalamet , zendaya coleman , stellan skarsgard main genre '
'scifi denis villeneuves dune adaptation got off '
'swimmingly, earning over 400 million at the box office '
'during the malaise of 2021s dearth of content thanks to '
'the pandemic',
'dune part two 202412 biggest dune 2 book changes from '
'denis villeneuves sequeldune part two adapts the second '
'part of frank herberts novel, and denis villeneuves sequel '
'changes the dune book story and characters many times',
'Other movie reviews': ['filmin joker folie à deux, lady gaga and joaquin '
'phoenix are coming for their second oscarthe joker '
'sequel is shaping up to be the biggest film of '
'2024by jack king, lana murray and lucy ford4 october '
'2023jose perezbaugriffinjoker folie à deux, the '
'sequel to todd phillips joaquin phoenixstarring king '
'of comedy riff was inevitable since it took the box '
'office by storm and nabbed phoenix an oscar'],
'Society of the Snow': ['The starcast of the movie Society of the Snow '
'includes Enzo Vogrinicic, Agustin Pardella, Matias '
'Recalt.',
'The movie Society of the Snow is directed by J.A. '
'Bayona.',
'The writers of the movie Society of the Snow are '
'Jaime Marques, Nicolás Casariego, Pablo Vierci.',
'The movie Society of the Snow is rated R.',
'The runtime of the movie Society of the Snow is 2h '
'24m.',
'The movie Society of the Snow was released in 2023.',
'The IMDb rating of the movie Society of the Snow is '
'7.4.',
'bayona, nicolás casariego, jaime marquiscastenzo '
'vogrincicenzo vogrinicic rolandiego '
'vegezziadvertisingan email youll actually loveget '
'into a relationship with our newsletter'],
'The Holdovers': ['The starcast of the movie The Holdovers includes Paul '
'Giamatti, Da'Vine Joy Randolph, Dominic Sessa.',
```



The screenshot shows a Visual Studio Code editor interface. The Explorer panel on the left displays a project named 'PROJECT 1' with a 'data' folder containing 34 files named 'filename0.txt' through 'filename34.txt'. Below the folder, there are four files: 'chatbot.py', 'database.pkl', 'DEscription.pdf', and 'user\_model.json'. The Terminal panel on the right shows a PowerShell session with a long string of text representing movie reviews. The text is wrapped across multiple lines. The status bar at the bottom indicates 'Ln 67, Col 45', 'Spaces: 2', 'UTF-8', 'CRLF', 'Python', '3.12.1 64-bit', 'Blackbox', and 'Prettier'.

```
'reviewthe holdovers, review paul giamatti shines in '
'alexander paynes best film since sidewaysthe star gives an '
'awardsworthy performance as a teacher forced to look after '
'pupils over the christmas break in this terrific period '
'piecerobbie collin, chief film critic18 january 2024 '
'1259related topicslondon film festival,world '
'cinema,comedy movies,boothox iconfacebook icondatagp '
'iconemail iconsavecomment speech bubble icondominic sessa '
'stars as angus tully and paul giamatti as paul hunkin in '
'the holdoverscredit courtesy of focus featuressome films '
'grow on you'),
'The Iron Claw': ['The starcast of the movie The Iron Claw includes Zac '
'Efron, Jeremy Allen White, Harris Dickinson.',
'The movie The Iron Claw is directed by Sean Durkin.',
'The writers of the movie The Iron Claw are Sean Durkin.',
'The movie The Iron Claw is rated R.',
'The runtime of the movie The Iron Claw is 2h 12m.',
'The movie The Iron Claw was released in 2023.',
'The IMDb rating of the movie The Iron Claw is 7.8.',
'brothers in arms jeremy allen white, harris dickinson and '
'zac efron',
'kevin is the poster boy but theres also the shaman david '
'harris dickinson, the olympic hopeful kerry jeremy allen '
'white and the youngest and more sensitive, musician mike '
'stanley simons',
'01am, the timesharris dickinson, zac efron, stanley simons '
'and jeremy allen white star in the iron claweric '
'chakenhous claw ilckevin maherfriday february 09 2024, '
'12'],
'The Taste of Things': ['The starcast of the movie The Taste of Things '
'includes Juliette Binoche, Benoît Magimel, Emmanuel '
'Salinger.',
'The movie The Taste of Things is directed by Anh '
'Hung Tran.',
'The writers of the movie The Taste of Things are '
'Marcel Rouff, Anh Hung Tran.',
'The movie The Taste of Things is rated PG-13.',
'The runtime of the movie The Taste of Things is 2h '
'15m.',
'The movie The Taste of Things was released in 2023.',
'The IMDb rating of the movie The Taste of Things is '
'7.7.',
'written and directed by tran anh hung, it is set in '
'a country house in the 1880s, and stars benoit '
'magimel as dodin, a famous gourmet who is in love '
'with his cook of 20 years, eugénie juliette binoche, '
'magimels ex in real life.',
'emotional harmony juliette binoche eugénie and '
'benoit magimel dodin in the taste of things'],
```



```
'chakeenhouse claw llckevin maherfriday february 09 2024, '
'12'],
'The Taste of Things': ['The starcast of the movie The Taste of Things '
'includes Juliette Binoche, Benoit Magimel, Emmanuel '
'Salinger.',
'The movie The Taste of Things is directed by Anh '
'Hung Tran.',
'The writers of the movie The Taste of Things are '
'Marcel Rouff, Anh Hung Tran.',
'The movie The Taste of Things is rated PG-13.',
'The runtime of the movie The Taste of Things is 2h '
'15m.',
'The movie The Taste of Things was released in 2023.',
'The IMDb rating of the movie The Taste of Things is '
'7.7.',
'written and directed by tran anh hung, it is set in '
'a country house in the 1880s, and stars benoit '
'magimel as dodin, a famous gourmet who is in love '
'with his cook of 20 years, eugénie juliette binoche, '
'magimels ex in real life.',
'emotional harmony juliette binoche eugénie and '
'benoit magimel dodin in the taste of things'],
'The Zone of Interest': ['The starcast of the movie The Zone of Interest '
'includes Christian Friedel, Sandra Hüller, Johann '
'Karlhaus.',
'The movie The Zone of Interest is directed by '
'Jonathan glazer.',
'The writers of the movie The Zone of Interest are '
'Jonathan glazer, Martin Amis.',
'The movie The Zone of Interest is rated PG-13.',
'The runtime of the movie The Zone of Interest is 1h '
'45m.',
'The movie The Zone of Interest was released in '
'2023.',
'The IMDb rating of the movie The Zone of Interest '
'is 7.6.',
'film reviewthe zone of interest review holocaust '
'film masterfully reinvents martin amis '
'novelnominated for five oscars, this adaptation is '
'easily jonathan glazers most mature workkevin '
'maherthursday february 01 2024, 1'],
'movies list': 'These are all the best movies of 2024 according to the '
'wired' - Dune Part Two, The Taste of things, The Iron claw, '
'The Zone of Interest, All of us Strangers, The Holdovers, '
'Society of the Snow, Mufasa The Lion King, Gladiator 2, '
'Spider Man Beyond the Spider Verse, Dune Part Two, '
'Beetlejuice 2, Joker Folie à Deux, The Fall Guy, Alien '
'Romulus, The Garfield Movie, IF '}
```

## PART 2: Chatbot

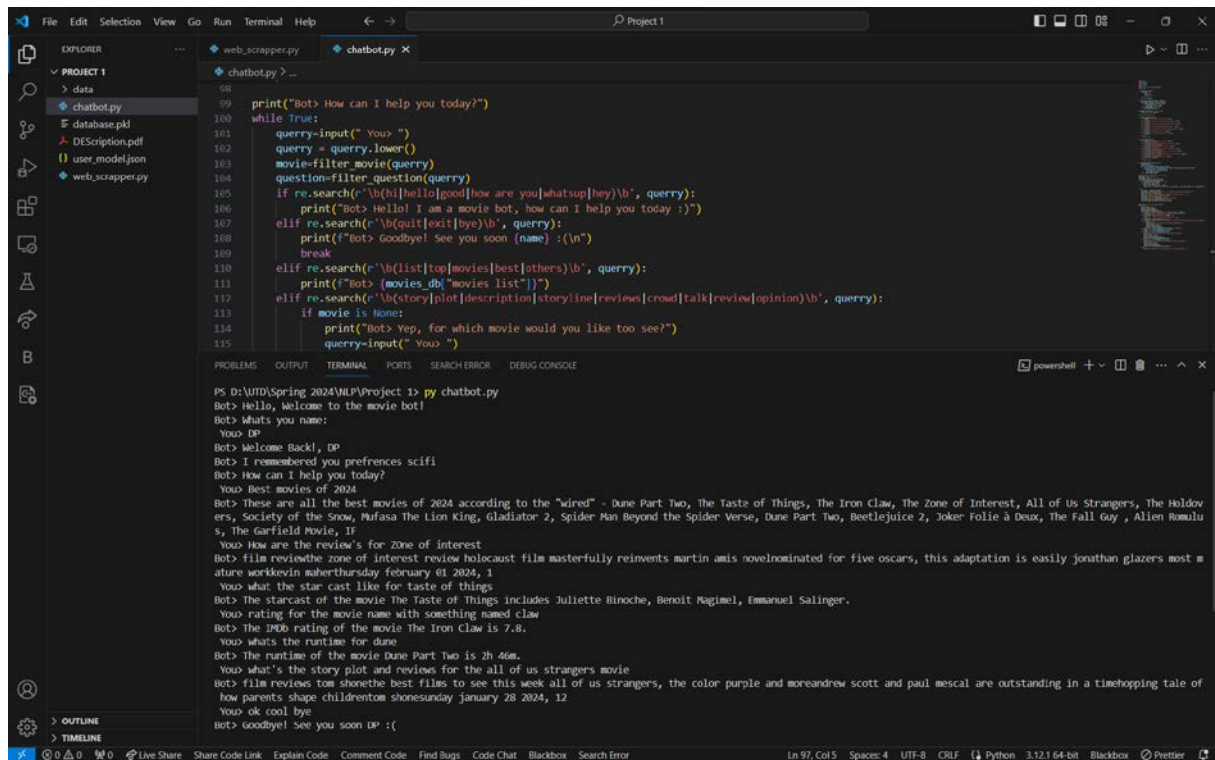
The chatbot system was developed in Python, utilizing natural language processing (NLP) techniques and the knowledge base created by the web crawler.

System Description and NLP Techniques:

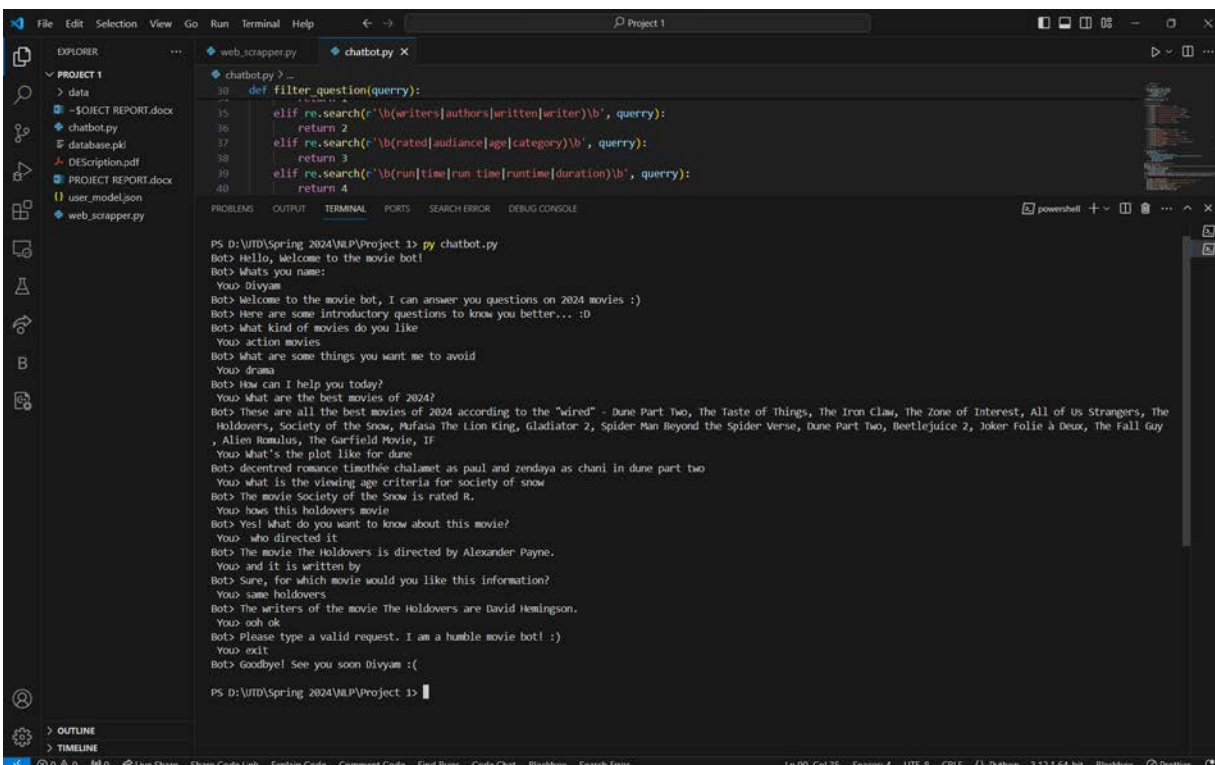
1. **Intent Recognition:** The chatbot used regular expressions to identify the user's intent from their input query. It could recognize intents such as inquiring about cast, directors, writers, ratings, runtime, release year, and IMDb rating.
2. **Entity Recognition:** Regular expressions were also employed to identify movie titles mentioned in the user's query. This allowed the chatbot to retrieve relevant information from the knowledge base.
3. **Response Generation:** Based on the recognized intent and movie entity, the chatbot would retrieve the corresponding information from the knowledge base and generate an appropriate response.



4. **Similarity Matching:** If the user's query did not directly match any information in the knowledge base, the chatbot employed a similarity matching technique using the SequenceMatcher from the difflib module. This allowed the chatbot to provide relevant responses even when the user's phrasing differed from the knowledge base entries.



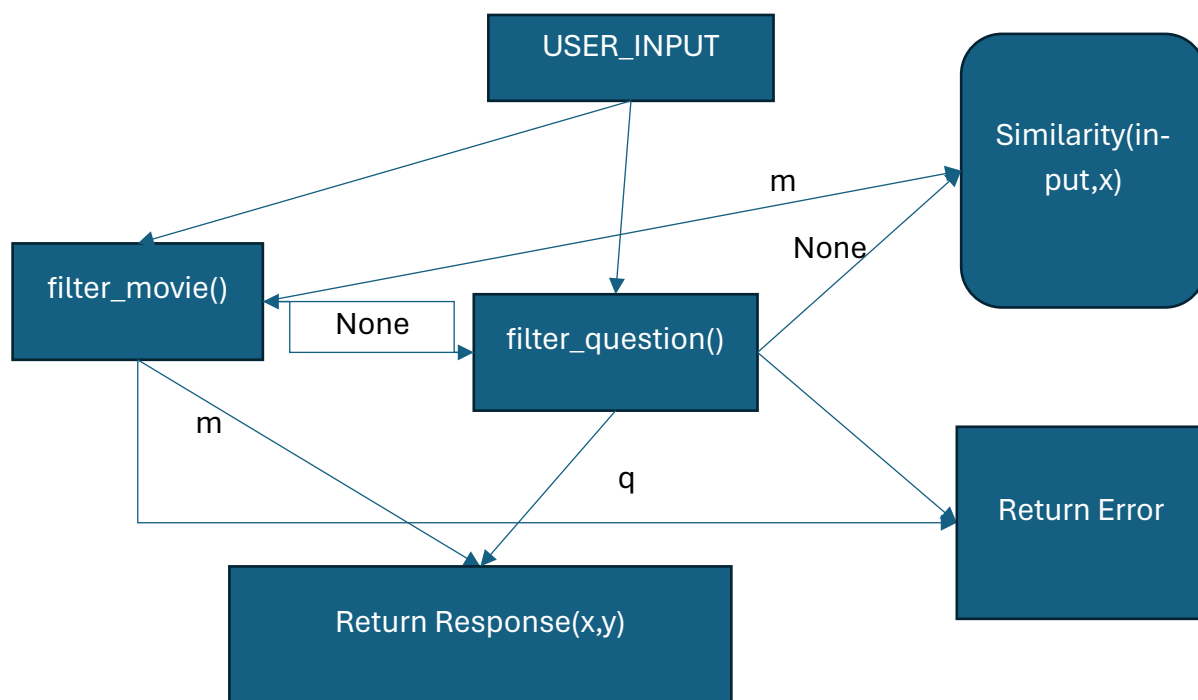
```
File Edit Selection View Go Run Terminal Help
Project 1
EXPLORER
PROJECT 1
  data
  chatbot.py
  database.pkl
  DESCRIPTION.pdf
  user_model.json
  web_scrapper.py
chatbot.py
98
99 print("Bot: How can I help you today?")
100 while True:
101     query = input("You: ")
102     query = query.lower()
103     movie = filter_movie(query)
104     question = filter_question(query)
105     if re.search(r'\b(hi|hello|good|how are you|whatsup|hey)\b', query):
106         print("Bot: Hello! I am a movie bot, how can I help you today :)")
107     elif re.search(r'\b(quit|exit|bye)\b', query):
108         print("Bot: Goodbye! See you soon (name) :(\n")
109         break
110     elif re.search(r'\b(list|top|movies|best|others)\b', query):
111         print("Bot: (movies_db["movies list"])")
112     elif re.search(r'\b(story|plot|description|storyline|reviews|crowd|talk|review|opinion)\b', query):
113         if movie is None:
114             print("Bot: Yep, for which movie would you like too see?")
115             query = input("You: ")
116
117 PS D:\UTD\Spring 2024\NLP\Project 1> py chatbot.py
Bot: Hello, Welcome to the movie bot!
Bot: What's your name:
You: DP
Bot: Welcome Back!, DP
Bot: I remembered your preferences scifi
Bot: How can I help you today?
You: Best movies of 2024
Bot: These are all the best movies of 2024 according to the "wired" - Dune Part Two, The Taste of Things, The Iron Claw, The Zone of Interest, All of Us Strangers, The Holdovers, Society of the Snow, Mufasa The Lion King, Gladiator 2, Spider-Man Beyond the Spider-Verse, Dune Part Two, Beetlejuice 2, Joker Folie à Deux, The Fall Guy, Alien Romulus, The Garfield Movie, IF
You: How are the reviews for Zone of Interest
Bot: film review the zone of interest review holocaust film masterfully reinvents martin amis novel nominated for five oscars, this adaptation is easily jonathan glazers most mature work kevin maher thursday february 01 2024, 1
You: What's the star cast like for taste of things
Bot: The starcast of the movie The Taste of Things includes Juliette Binoche, Benoit Magimel, Emmanuel Salinger.
You: Rating for the movie name with something named claw
Bot: The IMDb rating of the movie The Iron Claw is 7.8.
You: What's the runtime for dune
Bot: The runtime of the movie Dune Part Two is 2h 46m.
You: What's the story plot and reviews for all of us strangers movie
Bot: film review tom shone the best films to see this week all of us strangers, the color purple and more andrew scott and paul mescal are outstanding in a timehopping tale of how parents shape children tom shonesunday january 28 2024, 12
You: ok cool bye
Bot: Goodbye! See you soon DP :(:
```



```
File Edit Selection View Go Run Terminal Help
Project 1
EXPLORER
PROJECT 1
  -PROJECT REPORT.docx
  chatbot.py
  database.pkl
  DESCRIPTION.pdf
  PROJECT REPORT.docx
  user_model.json
  web_scrapper.py
chatbot.py
30 def filter_question(query):
31     if re.search(r'\b(writers|authors|written|writer)\b', query):
32         return 1
33     elif re.search(r'\b(rated|audience|age|category)\b', query):
34         return 2
35     elif re.search(r'\b(run|time|runtime|duration)\b', query):
36         return 4
37
38 PS D:\UTD\Spring 2024\NLP\Project 1> py chatbot.py
Bot: Hello, Welcome to the movie bot!
Bot: What's your name:
You: Divyam
Bot: Welcome to the movie bot, I can answer your questions on 2024 movies :)
Bot: Here are some introductory questions to know you better... :D
Bot: What kind of movies do you like
You: action movies
Bot: What are some things you want me to avoid
You: drama
Bot: How can I help you today?
You: What are the best movies of 2024?
Bot: These are all the best movies of 2024 according to the "wired" - Dune Part Two, The Taste of Things, The Iron Claw, The Zone of Interest, All of Us Strangers, The Holdovers, Society of the Snow, Mufasa The Lion King, Gladiator 2, Spider-Man Beyond the Spider-Verse, Dune Part Two, Beetlejuice 2, Joker Folie à Deux, The Fall Guy, Alien Romulus, The Garfield Movie, IF
You: What's the plot like for dune
Bot: decent romance timothee chalamet as paul and zendaya as chani in dune part two
You: What is the viewing age criteria for society of snow
Bot: The movie Society of the Snow is rated R.
You: How is this holdovers movie
Bot: Yes! What do you want to know about this movie?
You: who directed it
Bot: The movie The Holdovers is directed by Alexander Payne.
You: and it is written by
Bot: Sure, for which movie would you like this information?
You: same holdovers
Bot: The writers of the movie The Holdovers are David Hemingson.
You: ooh ok
Bot: Please type a valid request. I am a humble movie bot! :)
You: exit
Bot: Goodbye! See you soon Divyam :(:

PS D:\UTD\Spring 2024\NLP\Project 1> |
```

Dialog Tree/Logic Diagram:



#### Appendices:

- Appendix A: Knowledge Base (Samples)
  - As all the “movie\_titles” are keys so the content in them are related to that movie and it can be seen in the screenshots provided above.
- Appendix B: Sample User Models
  - The sample user models include the names, likes and dislikes of users who have interacted with the chatbot.

#### Evaluation and Analysis:

To evaluate the chatbot's performance, I conducted a survey with 4 participants (Uday, Arjun, Deepanshu, Kartik) who were not part of the development team. The survey included the following Likert-style questions:

1. How well did the chatbot understand your queries? (1 - Poor, 5 - Excellent)
2. How relevant and informative were the chatbot's responses? (1 - Poor, 5 - Excellent)
3. How natural and human-like did the chatbot's responses feel? (1 - Poor, 5 - Excellent)

The average scores from the survey were:

1. Understanding queries: 4.2
2. Response relevance and informativeness: 3.5
3. Natural and human-like responses: 3

Strengths:

- The chatbot effectively retrieved and provided relevant information from the knowledge base.
- The similarity matching technique allowed the chatbot to handle variations in user phrasing.
- The knowledge base covered a wide range of movie-related information, enabling the chatbot to answer diverse queries.

Weaknesses:

- The chatbot's responses were sometimes perceived as robotic or lacking natural language flow.
- The intent and entity recognition relied heavily on regular expressions, which can be brittle and inflexible.
- The knowledge base was limited to the information scraped from the selected websites, potentially missing important details or updates.
- The scrapped information was not proper even after cleaning there were many sentences with no spaces between them.

Overall, the rules-based chatbot successfully accomplished the task of providing information about 2024 movies. However, there is room for improvement, particularly in enhancing the natural language understanding and generation capabilities, as well as expanding and updating the knowledge base regularly.