



BUAN 6342.501

**APPLIED NATRUAL LANGUAGE
PROCESSING**

PROF. GASAN ELKHODARI

HARI SHANKER REDDY MADANA

Christopher Colley

Preethi Jerripothula

Sai Madhan Muthyam

Sofia Rajan

Hrushikesh Sriramaneni



Introduction

The purpose of this report is to outline the process of creating a chatbot and a knowledge base about a famous musicartist, Taylor Swift, using web crawling and natural language processing techniques. The knowledge base aims to capture relevant information and insights about Taylor Swift from publicly available sources, particularly Wikipedia.

Methodology for building Web Crawler:

Below are the step by step process of how the web crawler was built:

1. Web Crawling: Initiated the process by selecting Wikipedia as the primary source of information about Taylor Swift. Used Python libraries such as BeautifulSoup and requests to crawl Wikipedia pages related to Taylor Swift. The crawlingprocess involved extracting the URLs of relevant Wikipedia articles and storing them in a text file for further processing.
2. Text Extraction: After obtaining the URLs, text content was extracted and saved into individual text files, with eachfile representing a single Wikipedia page.
3. Text Pre-processing: Before building the knowledge base, text pre-processing to clean the text data was performed.This involved converting the text to lowercase, removing non-alphanumeric characters except for commas and full stops, removal of stop words and tokenizing the text into sentences.
4. Important Term Extraction: To identify key concepts and topics related to Taylor Swift, TF-IDF (Term Frequency-Inverse Document Frequency) analysis was used. This analysis was used to extract important terms from the pre- processed text corpus, considering their frequency and significance within the corpus.
5. Building the Knowledge Base: The list of important terms were used to search through the pre-processed text corpus to identify sentences containing these terms. These sentences were then stored in the knowledge base along with their corresponding terms, forming associations between terms and relevant textual contexts.
6. Storing the Knowledge Base: The knowledge base, comprising associations between terms and textual contexts, was stored in a structured format using the pickle library. This facilitated easy retrieval and analysis of information for future reference.

Screenshot of Knowledge base:

```
In [254]: # Load the knowledge base from the pickle file
with open('knowledge_base.pkl', 'rb') as f:
    knowledge_base = pickle.load(f)

# Print a limited number of sentences for each term
for term, sentences_info in knowledge_base.items():
    print("Term:", term)
    for i, sentence_info in enumerate(sentences_info[:3]): # Print only the first three sentences for each term
        print(f"Sentence {i+1}:", sentence_info)
    print() # Add an empty line between terms
```

```
Term: taylor
Sentence 1: The purchase of Big Machine Records by Scooter Braun 's company, Ithaca Holdings, in 2019 resulted in a highly publicized dispute and controversy with American singer-songwriter Taylor Swift regarding the ownership of the masters to her albums for the label.
Sentence 2: [3] Before he left DreamWorks, Borchetta approached Taylor Swift and her family after the musician p
erformed at the Bluebird Café in Nashville, Tennessee after first meeting her in 2004.
Sentence 3: [6]
Soon after, Big Machine released Swift's first ever recording, the single " Tim McGraw " and her debut album Tayl
or Swift .
```

```
Term: swift
Sentence 1: The purchase of Big Machine Records by Scooter Braun 's company, Ithaca Holdings, in 2019 resulted in a highly publicized dispute and controversy with American singer-songwriter Taylor Swift regarding the ownership of the masters to her albums for the label.
Sentence 2: [3] Before he left DreamWorks, Borchetta approached Taylor Swift and her family after the musician p
erformed at the Bluebird Café in Nashville, Tennessee after first meeting her in 2004.
Sentence 3: He made an offer to Swift and her parents, whereby he would recruit her to the new label's roster after it was established.
```

List of Important Terms:

```
print("Important terms:", important_terms_strings)

Important terms: ['time', '2016', 'may', 'american', 'january', 'folk', 'april', 'edit', 'new', 'video', 'nashville', 'march', 'july', 'swifts', '2015', 'album', '2017', 'february', 'pop', 'september', 'awards', 'december', 'records', 'august', '2022', 'rock', 'billboard', '2019', 'october', 'country', '2020', 'november', '2021', '2023', 'archived', 'original', 'music', 'retrieved', 'taylor', 'swift', 'Reputation', 'Lover', 'Red (album)', 'Grammy Awards', 'Country Music', 'Pop Music', 'Eras Tour', 'Billboard charts', '1989 (album)', 'Taylor Swift Foundation', 'Travis Kelce']
```

The important terms contains a mix of extracted terms using tf-idf and manually defined terms

Conclusion:

- The creation of the knowledge base for Taylor Swift provides a structured repository of information that can be used for various purposes such as text analysis, information retrieval, and knowledge discovery.
- By leveraging web crawling and natural language processing techniques, we were able to compile and organize valuable insights about Taylor Swift in a systematic manner.

Methodology for building Chatbot:

System Description:

The chatbot is designed to interact with users in a conversational manner, providing information and responding to user queries. It utilizes various Natural Language Processing (NLP) techniques to understand user input and generate appropriate responses.

Key NLP techniques used include:

- Tokenization: The process of splitting input text into individual words or tokens.
- Stop word Removal: Filtering out common words (stop words) that do not carry significant meaning in the context of the conversation.
- Cosine Similarity: Calculating the similarity between user input and entries in the knowledge base to retrieve relevant information.
- TF-IDF Vectorization: Transforming text data into numerical vectors to measure the importance of words in the context of the conversation.
- Knowledge Base Retrieval: Storing information in a knowledge base and retrieving relevant information based on user queries.
- Dialog Management: Handling greetings, farewells, and maintaining conversation flow.

Sample Dialog Interactions:

```
In [83]: chatbot()

- - - - -
AJ: Hello! I'm AJ chatbot.
- - - - -

AJ: Hi there! Please tell me your name.
User: Max
AJ: Nice to meet you, Max! What are your interests?
User: I like music
AJ: And what are your dislikes?
User: Bugs

AJ: Hi Max, feel free to ask me anything about Taylor Swift.

You: Has Taylor Swift been involved in any legal battles?
AJ: Swift donated to fellow singer-songwriter Kesha to help with her legal battles against Dr
AJ: Luke and to actress Mariska Hargitay 's Joyful Heart Foundation
AJ: Enter 'bye' to end the conversation.

You: Ok! how many private jets does she have
AJ: In February 2024, it was reported that Swift had sold one of her two private jets
AJ:
AJ: Enter 'bye' to end the conversation.

You: Her net worth?
AJ: ] She was the best-paid country music star in 2016 according to Forbes, with a net worth of $27
AJ: 5 million
AJ: Enter 'bye' to end the conversation.

You: Did she win a grammy in 2020
AJ: She became the first woman and fourth musician to win the Grammy Award for Album of the Year three times with h
er eighth studio album, Folklore , at the 63rd Annual Grammy Awards in 2021
AJ:
AJ: Enter 'bye' to end the conversation.
```

Our chatbot 'AJ' greets the user appropriately and takes input of their interests, likes and dislikes. It then requests user to input a query and displays answers appropriately. The bot also concludes the conversation with farewell message

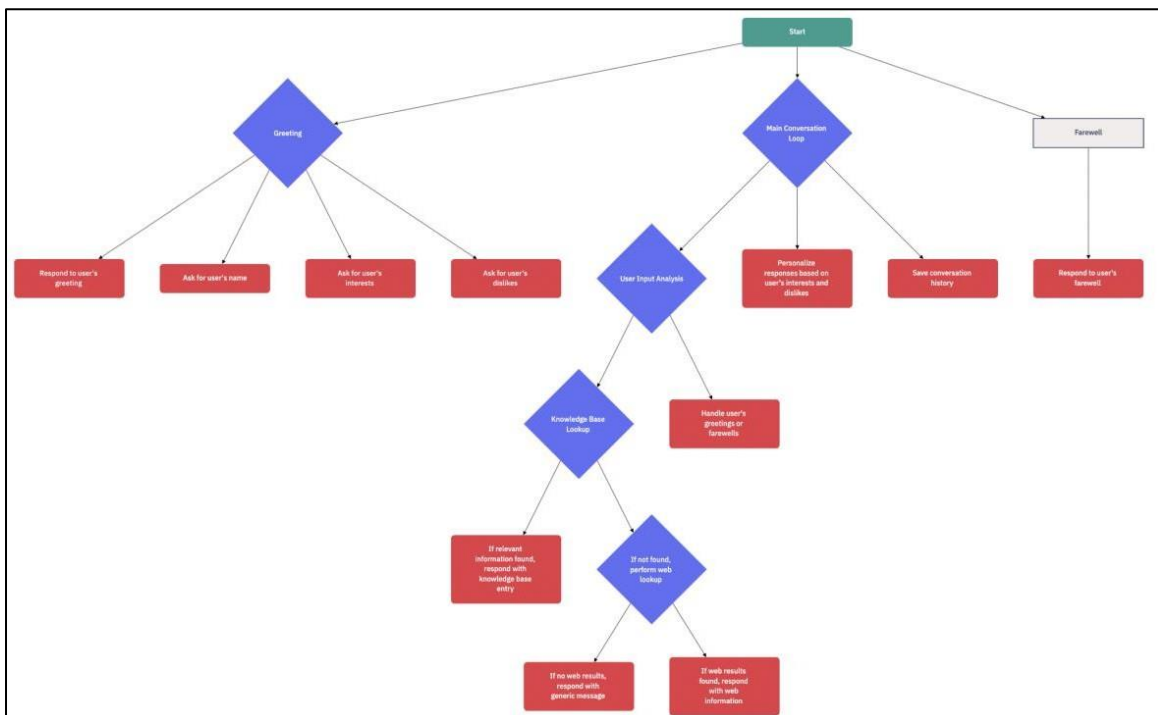
Dialog Logic Diagram:

Start

```

|
|---> Greeting
|
|   |---> Respond to user's greeting
|   |---> Ask for user's name
|   |---> Ask for user's interests
|   |---> Ask for user's dislikes
|
|---> Main Conversation Loop
|
|   |---> User Input Analysis
|   |   |---> Knowledge Base Lookup
|   |   |   |---> If relevant information found, respond with knowledge base entry
|   |   |   |---> If not found, perform web lookup
|   |   |       |---> If web results found, respond with web information
|   |   |       |---> If no web results, respond with generic message
|   |   |---> Handle user's greetings or farewells
|   |---> Personalize responses based on user's interests and dislikes
|   |---> Save conversation history
|
|---> Farewell
|
|   |---> Respond to user's farewell

```



Appendix - Knowledge Base:

```
Term: album
Sentence 1: The purchase of Big Machine Records by Scooter Braun 's company, Ithaca Holdings, in 2019 resulted in a highly publicized dispute and controversy with American singer-songwriter Taylor Swift regarding the ownership of the masters to her albums for the label.
Sentence 2: Big Machine's first album release was Ingram's Live: Wherever You Are , a re-release of a live album h e had previously issued independently in 2004.
Sentence 3: The album features two new studio tracks, "Wherever You Are" and "Love You", both of which were release d as singles.

Term: february
Sentence 1: [8]
In February 2010 , Swift won four Grammys –including Album of the Year ( Fearless ), Best Country Album ( Fe arless ), Best Country Song ( "White Horse" ), and Best Female Country Vocal Performance ("White Horse")—becomi ng the first Big Machine artist to win a Grammy Award.
Sentence 2: [11]
In a February 2015 interview, Borchetta refused to comment on the status of the label's distribution deal with Un iversal Music Group (UMG), which was up for renegotiation at the time.
Sentence 3: Borchetta then clarified in a February 2015 interview that Swift's catalog would be permitted on a stre aming service "that understands the different needs that we [Swift and Big Machine] have." wherebv "the choice to b
```

The sentences are tokenized from our scraped raw data. These sentences are ordered in our database if they contain the final list of important terms.

Above screenshot displays list of 3 sentences containing our extracted 'important term' as key.

Appendix - Sample User Models

```
602 User Jay's Interests: Tea
603 User Jay's Dislikes: Studying
604 User Jay: when did reputation release
605 Bot: "Taylor Swift Announces New Album, Reputation , for November Release"
606 Bot:
607 User Jay: Bye
608 Bot: Bye!
609 User Sid's Interests: Food
610 User Sid's Dislikes: Exams
611 User Sid: What was her first song
612 Bot: 1 on the Billboard Hot 100 " All Too Well " 2022 Highest Annual Earnings for a Female Musician Swift Most Nickelodeon Kids' Choice Awards nominations for an Individual Most Teen Choice Awards won by a Individual Most Teen Choice Awards won by a Musician Most Streamed Act on Spotify in 24 hours Most Day-one Streams of an Album on Spotify Midnights Most Streamed Album on Spotify in 24 hours Most Streamed Album on Spotify in One Week Most Day-one Streams of an Album on Spotify Most Streamed Album by a Female Artist in One Week Fastest-Selling Vinyl Album Most Streamed Album on Amazon Music in 24 Hours Most Albums with Million-Selling Weeks on the US Billboard 200 Swift Most Million-Selling Weeks on the US Albums chart † Most Cumulative Weeks at No
613 Bot: 1 on US Albums chart Most No
614 User Sid: bye
615 Bot: Bye!
616
```

All the chats with users are stored in conversation_history.txt file. The bot records User_name, user's likes and dislikes, followed by conversation containing user queries and bot's responses. For each user, user likes and dislikes are recorded once. For example, above image displays conversations with two users by the name Jay and Sid.

Evaluation:

Strengths:

- The chatbot engages users in conversational interactions and handles greetings and farewells appropriately.
- It is successfully able to perform a web lookup based on input query
- The use of TF-IDF vectorization and cosine similarity ensures that relevant information is retrieved even with variations in user input.

Weaknesses:

- Accuracy to be improved and ability of chatbot to rephrase answers according to user input
- Lack of support for handling ambiguous queries or questions requiring subjective opinions.

Survey Results:

Review 1:

Ease of Use: How easy was it to interact with the chatbot? (Likert scale: 1-5) : **5**

Accuracy of Responses: How accurate were the responses provided by the chatbot?

(Likert scale: 1-5) **3** Overall Satisfaction: How satisfied were you with the chatbot

experience? (Likert scale: 1-5) **5**

The survey results will help assess user satisfaction and identify areas for improvement in the chatbot's performance and usability.

Review 2:

Ease of Use: How easy was it to interact with the chatbot? (Likert scale: 1-5) 5

Accuracy of Responses: How accurate were the responses provided by the chatbot?

(Likert scale: 1-5) 4 Overall Satisfaction: How satisfied were you with the chatbot

experience? (Likert scale: 1-5) 5

The survey results will help assess user satisfaction and identify areas for improvement in the chatbot's performance and usability.