CSE 535 Information Retrieval Final Project Report: IR Chat-bot

Name	UBIT NAME
Krishna Vineeth Puchalapalli	puchalap
Indra teja pidathala	indratej

Contents

1 Introduction **2**

2 Methodology **2**

2.1 Architecture 2

2.2 Components 3

2.2.1 Wiki Scraper 3

2.2.2 chit-chat bot 4

2.2.3 Training 4

2.3 Visualizations Chat UI 4

3 Conclusion **4**

4 contributions **5**

5 References **5**

1 Introduction

This project provides an intelligent and interactive chatbot capable of providing users with informative responses and engaging in chit-chat conversations. The project combines critical components and enhances the user experience by gathering the knowledge base from wikipedia, summarizing, contextual understanding and multi-lingual support by provisioning a user-centric approach. The critical components of the solution include Wiki Scraper, Analysis, Bot, Exception Handling and Visualization. Implementation of a web scraper makes it capable of extracting data from Wikipedia and it targets a collection of a minimum of 50,000 documents for 10 main topics which includes Health, Environment, Sports, Economy, Entertainment, Politics, Education, Travel, Food and Technology. Subtopics on each of the main topics can also be explored like Global warming related to Environment. The solution is designed to provide relevant answers to the topic by classifying the queries. Wiki Q/A Bot retrieves relevant documents and generates coherent summaries based on a given user query and the classified topic by scraping the data. Generative models enhance the summary quality by potentially earning bonus points. The concept of Exception Handling prevents chatbot crashes and provides information to handle the errors when the chatbot fails to understand user queries. By this means a complete life cycle of data retrieval and processing is done. The solution provides a user-friendly web interface (UI) for interacting with the chatbot and is then hosted as a web application which is accessible from outside of the local environment.

The Wikipedia Question Answering System project represents a comprehensive approach to build a highly functional and user-friendly chatbot. By systematically implementing web scraping, topic analysis, question answering, error handling, and visualization, we aim to provide users with a valuable and engaging experience. The project's success will be measured by its ability to cater to diverse user interests and deliver informative responses seamlessly. This report outlines the project's objectives, methodologies, and components, setting the stage for the development of an intelligent chatbot that harnesses the vast knowledge available on Wikipedia to serve users effectively. Challenging issues of the system which involve non-classification of the model, compliance and ethical consideration can be improvised in the future to provide a better solution of the same.

2 Methodology

2.1 Architecture

pipeline of our Chatbot is shown in fig:1.

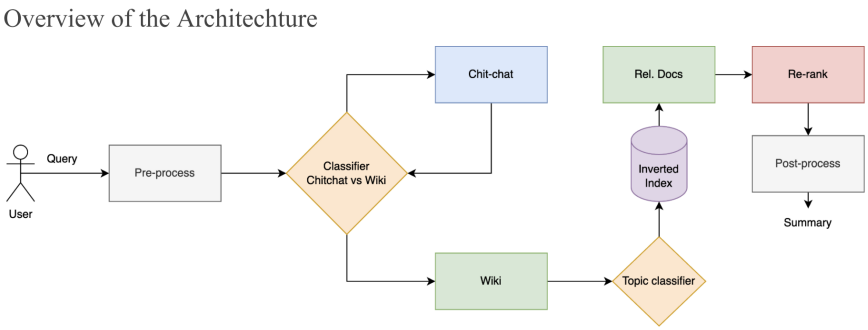

Figure 1: Architecture

2.2 Components

the IR chatbot is comprised of 2 main components/functionalities. mainly Wiki-Scraper and chit-chat component. both pf them are integrated during training

2.2.1 Wiki Scraper

we aim to scrape data with Wiki scraper API. Before mining and analyzing such content, collecting, cleaning, and storing such data is a crucial step. collect online data and efficiently store them, making data retrieval and analytics easier for downstream applications.

Scrape 50,000 document data from Wiki related to the for each Topic from below list.

1. Health: Common diseases, global health statistics, mental health trends...
2. Environment: Global warming, endangered species, deforestation rates...
3. Technology: Emerging technologies,AI advancements...
4. Economy: Stock market performance, job markets, cryptocurrency trends...
5. Entertainment: Music industry, popular cultural events, streaming platforms...
6. Sports: Major sporting events, sports analytics...
7. Politics:Elections, public policy analysis, international relations...
8. Education: Literacy rates, online education trends, student loan data...
9. Travel: Top tourist destinations, airline industry data, travel trends...
10. Food: Crop yield statistics, global hunger and food security...

Save all the scraped documents as a csv or just a simple text file and use them to train the bot for specific informative knowledge conversations.

2.2.2 chit-chat bot

we used **dialog-GPT** to make this chit-chat component.DialogGPT is a variant of the GPT (Generative Pre-trained Transformer) model developed by OpenAI. It is specifically fine-tuned for generating human-like responses in a conversational context. The model is trained on a massive amount of internet text and is capable of understanding and generating coherent and contextually relevant responses.

Unlike traditional chatbots that rely on rule-based systems, DialogGPT uses a deep learning approach, allowing it to handle a wide range of topics and engage in more natural and dynamic conversations. It is based on the transformer architecture, which excels at capturing long-range dependencies in sequential data

we use this chat-bot model for general chit-chat conversations.

2.2.3 Training

we have trained the chit-chat bot on the wiki scraper conversations to make the chatbot more broad, robust and effective in it's responses to wiki-related chat.

2.3 Visualizations Chat UI

DialogGPT based general chit-chat is shown in fig:2. fig:3 shows the chat for health topic based chat.

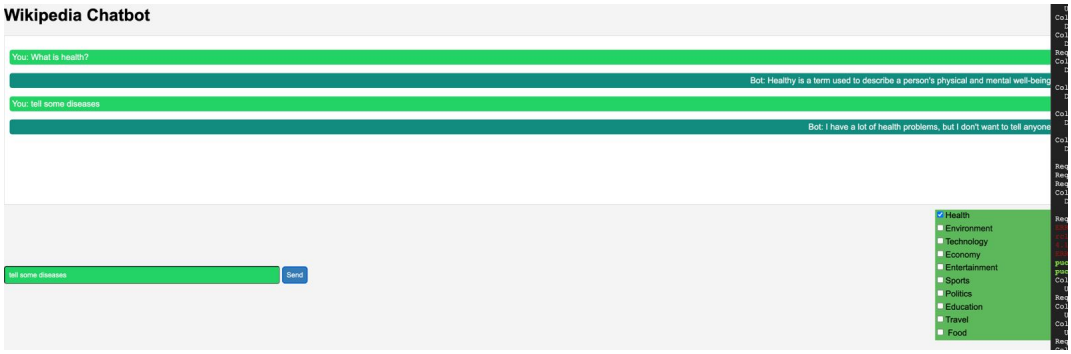

Figure 2: Chi-chat

3 Conclusion

In Conclusion, we are able to create an IR chat-bot which can do generic conversational chats and also can chat with high-level knowledge based on scraped data from wiki.

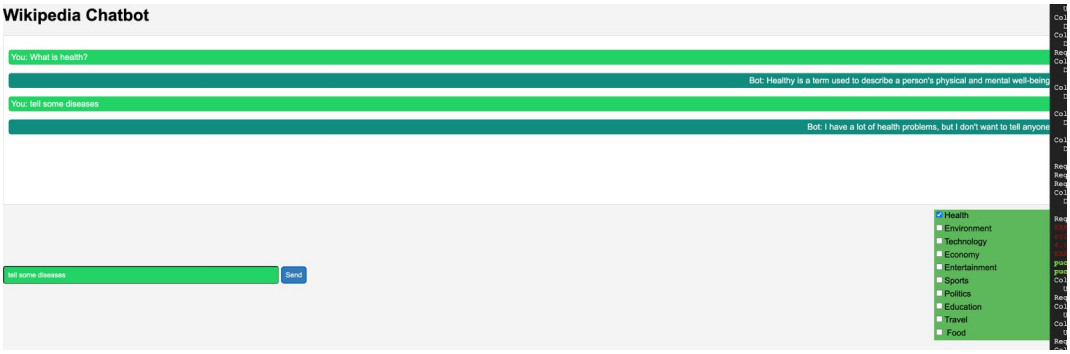

Figure 3: Wiki-health-chat

4 contributions

we are 2 in our group. contributions are as follows.

Name	UBIT NAME	percentage
Krishna Vineeth Puchalapalli	puchalap	60
Indra teja pidathala	indratej	40

5 References

- Project Manual
- Article
- Intel Article
- Overleaf(LateX)
- Wikipedia