TalkNet: Interacting with Websites via Langchain Communication

A Machine Learning Lab Report

Submitted in partial fulfilment of the
requirements for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

Computer Science and Engineering (AIML)

by

Sanya Dureja

(Registration No.:219310040)

Prisha Goyal

(Registration No.:219310071)

under the supervision of

Dr. Amit Kumar Bairwa

(Assistant Professor (Selection Grade), Department of AIML)

MANIPAL UNIVERSITY
JAIPUR

School of Computer Science and Engineering

Department of Artificial Intelligence and Machine Learning

MANIPAL UNIVERSITY JAIPUR

JAIPUR-303007

RAJASTHAN, INDIA

Jan-May 2024

Certificate

This is to certify that the project entitled ”TalkNet: Interacting with Websites via Langchain Communication” is a bonafide work carried out as part of the course AI3230, Machine Learning Lab Project, under my guidance, by Sanya Dureja and Prisha Goyal, students of Computer Science (AIML) & VIth Semester at the Department of Artificial Intelligence and Machine Learning, Manipal University Jaipur, during the academic semester VI, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering (AIML), at MUJ, Jaipur.

Sanya Dureja, 219310040 & Prisha Goyal, 219310071

Dr. Amit Kumar Bairwa

(Project Guide)

Date: Jan-May 2024

Acknowledgment

Firstly, I would like to express my deep sense of gratitude and indebtedness to my supervisor ***Prof. Amit Kumar Bairwa*** for his constant involvement and guidance throughout the research work. Indeed, his technical insight and intuitions have helped me in exploring several interesting problems in Artificial Intelligence. From him, I have also come to know about different research directions in the Integrated Machine Learning Approach. This project would not have been possible without close interaction, inspiration, and advice from numerous individuals.

My sincere thanks and gratitude to ***Dr. Santosh Kumar Vishwakarma (HOD-AIML)***, for enabling me with the capability to complete the research work and showing the right path at the right time.

I must also extend my gratitude to ***Dr. Sandeep Chaurasia, Director (SCSE)***, for his constant inspiration and for bringing into my notice many latest scientific advancements in the world not only in the field of my research interest but beyond this topic also. I extend my thanks and gratitude to all my team members for their constant involvement, guidance, and motivation in this journey.

Jan-May 2024

Manipal University Jaipur

Abstract

In the current era, efficient and user-friendly web interaction has become imperative. The primary objective is to address the limitations of conventional web browsing by enabling users to engage in conversational interactions with websites.

The project presents TalkNet, a Python-based solution for seamlessly communicating with web content, facilitating dynamic querying and information retrieval. By utilizing Langchain Communication, a bidirectional framework facilitating real-time interaction between users and websites, TalkNet eliminates the need for manual navigation. As a Machine Learning lab project, TalkNet presents exciting challenges and opportunities for advancing the state of the art in NLP, communication frameworks, and user interface design, with the potential to reshape the future of online interaction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 What is TalkNet

In the modern era, efficient and user-friendly web interaction has become paramount. Traditional web browsing methods often entail manual navigation and limited engagement, posing challenges for users in accessing and interacting with online content seamlessly. Recognizing this need for improvement, this project introduces TalkNet—a Python-based solution designed to revolutionize web interaction.

The project is a comprehensive guide to building an application capable of interacting with websites, extracting information, and communicating in a user-friendly manner. It leverages the power of LangChain and integrates it with a Streamlit GUI for an enhanced user experience.

Chapter 2

Literature review

2.1 Brief Literature Review

The literature review explores the transformation of web interaction paradigms from conventional methods to innovative solutions like TalkNet, focusing on the challenges faced, advancements made, and the potential impact on user experience.

The evolution of natural language processing (NLP) and machine learning (ML) techniques has spurred innovative applications across various domains. Notably, recent research has focused on leveraging LangChain, an emerging NLP framework, to enhance web interaction and information retrieval.

[1] Introduces a novel approach for question generation from PDFs using LangChain, demonstrating its efficacy in automatically deriving meaningful questions from document content. This work underscores LangChain's potential in educational assessments, fostering deeper student engagement and comprehension.

In a related context, [2] proposes an effective query system utilizing LangChain and large language models (LLMs) to streamline PDF document querying. Their solution highlights LangChain's capability to extract essential information and improve search efficiency within unstructured PDFs.

[3] Explores revolutionizing retrieval-augmented generation by enhancing PDF structure recognition. This study emphasizes the importance of accurate document parsing in professional knowledge-based question answering, showcasing LangChain's integration with advanced techniques to enhance retrieval accuracy.

Additionally, [4] presents a pioneering application of LangChain in automating customer service, demonstrating its potential to reshape traditional support methods with personalized, context-aware interactions. This work signifies LangChain's adaptability across industries, particularly in customer-

centric ecosystems.

Collectively, these studies illustrate LangChain's versatility and efficacy in advancing NLP capabilities, ranging from educational assessments and document querying to customer service automation. They collectively contribute to the ongoing evolution of ML-driven applications, offering insights into the transformative potential of LangChain within diverse domains.

2.1.1 A study of previous works

2.1.2 Literature Review Summary

Table 2.1: Literature Review: Summary Table

S. No.	Author Name	Year of Publication	Contribution	Scope	Limitations
1.	Keivalya Pandya and Mehfuza Holia	2023	Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations[4].	Customer service automation	Focuses on chatbot application; applicability in broader automation contexts may require additional adaptation and testing.

2.	Madhav, Dinesh and Nijai, Sanskruti and Patel, Urvashi and Champanerkar, Komal	2024	Question Generation from PDF using LangChain[1].	Educational assessments	Limited to PDF-based question generation; may require further adaptation for diverse document formats and applications.
3.	Adith Sreeram A S and Pappuri Jithendra Sai	2023	An Effective Query System Using LLMs and LangChain[2].	Document querying	Focuses primarily on PDFs; potential challenges with scalability and diverse document structures.
4.	Demiao Lin	2024	Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition[3].	Professional knowledge-based question answering.	Relies on PDF structure; may encounter issues with varied PDF layouts and formats.

Chapter 3

Problem Definition

3.1 Problem Statement

In the current digital era, conventional web browsing methods often hinder efficient and user-friendly interaction with online content. Manual navigation and limited engagement impede users from seamlessly accessing and retrieving information from websites. There is an increasing demand for an intuitive and dynamic approach that enables users to engage in conversational interactions with web content, thereby overcoming the limitations of traditional browsing methods. This project aims to address this challenge by developing TalkNet, a Python-based solution that leverages Langchain Communication, a bidirectional framework enabling real-time interaction between users and websites. TalkNet aims to revolutionize web interaction by facilitating dynamic querying and information retrieval, eliminating the need for manual navigation. As a Machine Learning lab project, TalkNet offers exciting opportunities to advance the state of the art in Natural Language Processing (NLP), communication frameworks, and user interface design. The successful implementation of TalkNet has the potential to reshape the future of online interaction, providing users with a seamless and personalized browsing experience.

Chapter 4

Methodology and Framework

The methodology and framework of ”TalkNet” can be summarized as follows:

1. **Web Scraping with BeautifulSoup**

   The process begins by scraping a website to gather the HTML content. BeautifulSoup is a Python library used to pull data out of HTML and XML files. It transforms a complex HTML document into a complex tree of Python objects which you can search and modify.

2. **Text Splitting**

   The scraped HTML content is then split into smaller chunks of text. These chunks can be considered separate documents or sections of the original HTML content.

3. **Vectorization**

   Each chunk of text is converted into numerical vectors called embeddings. This is done to transform the textual information into a format that can be processed by machine learning algorithms. Embeddings capture the semantic meaning of the text.

4. **Semantic Search**

   When a query is made, the query itself is also converted into an embedding. This process is known as question embedding. The semantic search involves comparing the query embedding with the embeddings from the vector database to find the most relevant pieces of text.

5. **Vector Database**

   The embeddings created from the chunks of text are stored in a vector database, Chroma. This database is structured to allow efficient retrieval of entries that are semantically similar to a given query embedding.

6. **Ranked Results**

   The semantic search yields results ranked by relevance to the query embedding. This ranking is typically based on the similarity of the embeddings, with the most similar embeddings appearing at the top of the ranked results.

7. **Language Model (LLM)**

   A language model, like GPT-4, takes the ranked results as input. The LLM generates a human-like text response by considering the information from the most relevant chunks of text.

8. **Answer Generation**

   Finally, the language model generates an answer to the query. This answer is based on the information retrieved and ranked as relevant in the previous steps.

4.1 Framework

The diagram illustrates the process for RAG (Retrieval-Augmented Generation) using a website's content.

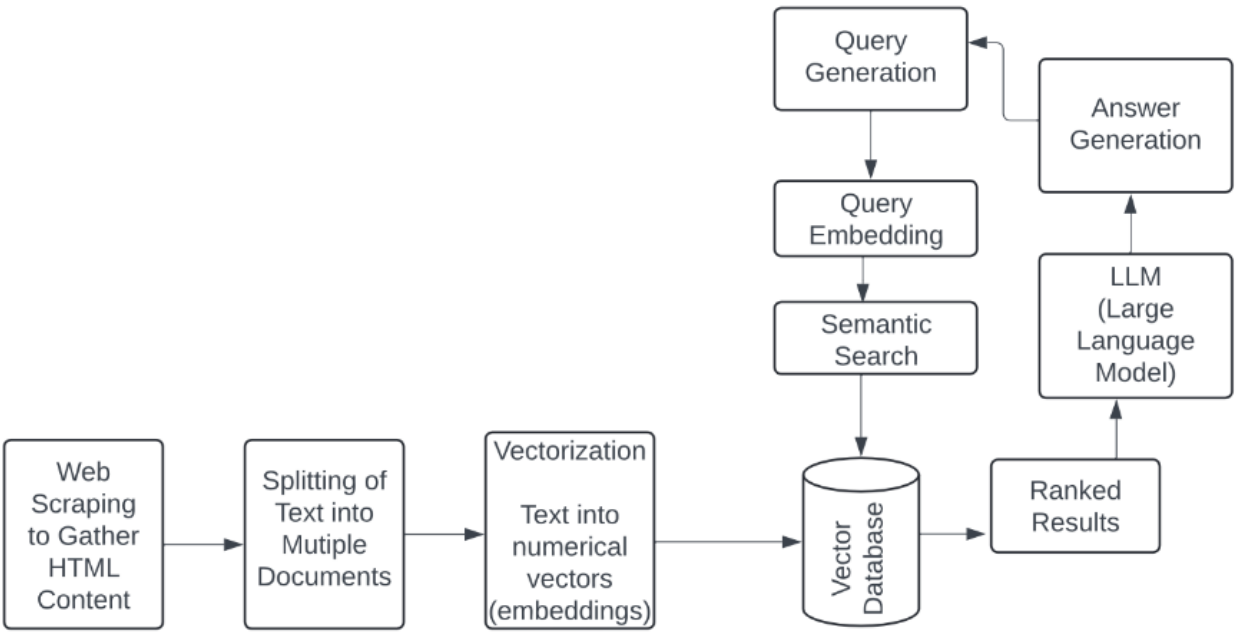

Figure 4.1: Proposed Methodology Architecture

Chapter 5

Implementation

The implementation of TalkNet involves several stages (given in the methodology): Web Scraping with BeautifulSoup, Text-Splitting, Vectorization, Semantic Search, Vector Database, Ranked Results, Language Model (LLM), and Answer Generation.

We have completed the implementation till the Text-Splitting part where the scraped HTML content is split into smaller chunks of text which are considered separate documents or sections of the original HTML content.

In the Vectorization stage, each chunk of text is converted into numerical vectors called embeddings using the OpenAI embeddings model. These embeddings capture the semantic meaning of the text.

However, during the Vectorization phase, we encountered a significant challenge with the rate limit errors from the OpenAI API, which temporarily halted our progress.

Moving forward we will overcome the rate limit challenges so that an answer can be generated for a particular query entered by the user.

5.1 Graphical User Interface

Figure 5.1 depicts the basic GUI of TalkNet.

5.2 Text-Splitting

Figure 5.2 illustrates the splitting of text into multiple documents for a random URL.

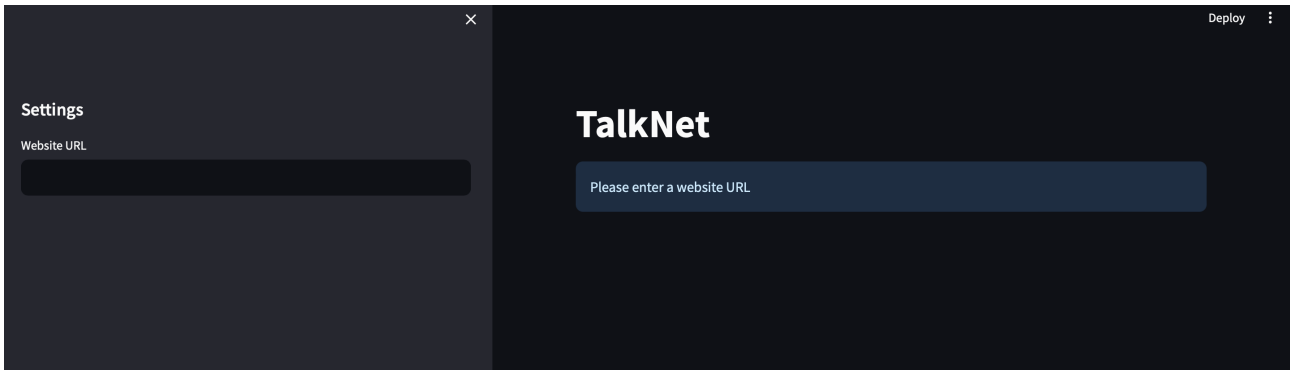

Figure 5.1: GUI Using Streamlit

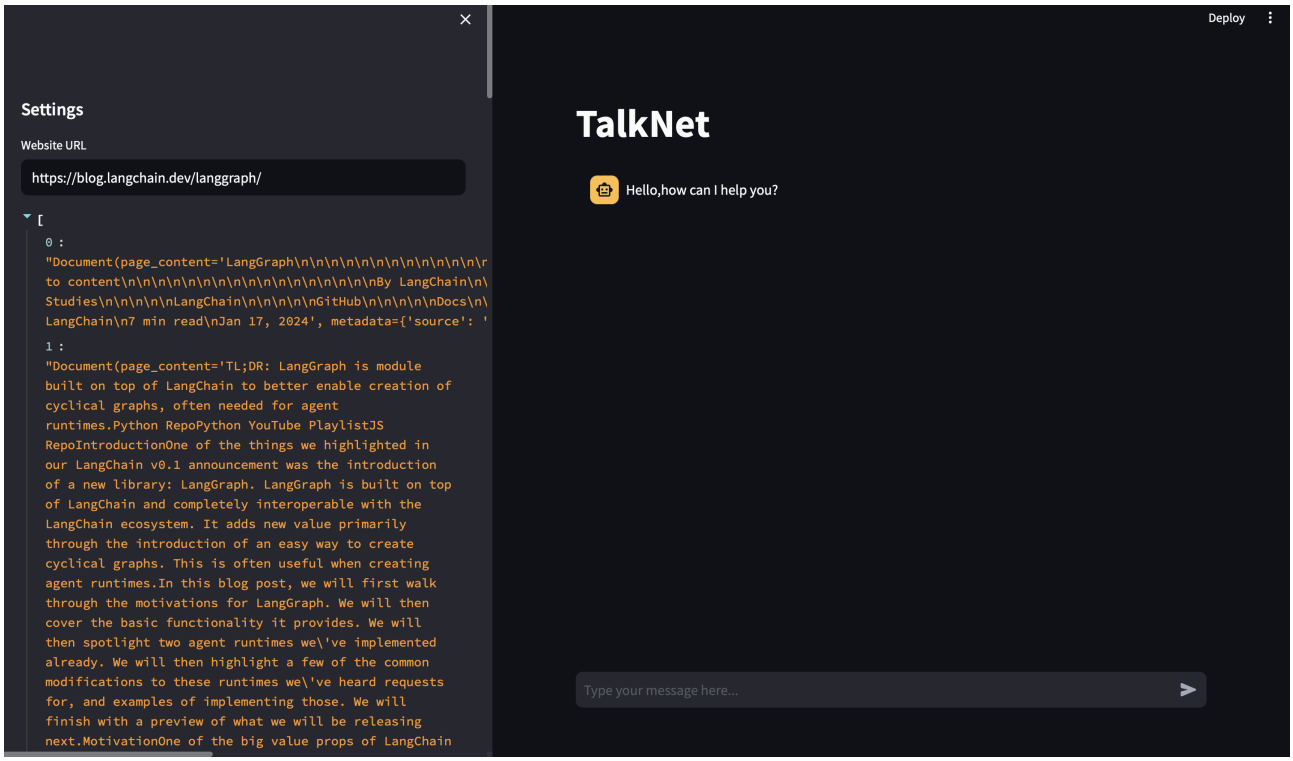

Figure 5.2: Splitting of Text into Multiple Documents

5.3 Vectorization: OpenAI Embeddings Rate Limit Error

Figure 5.3 illustrates the significant challenge with the rate limit error from the OpenAI API, which temporarily halted our progress.

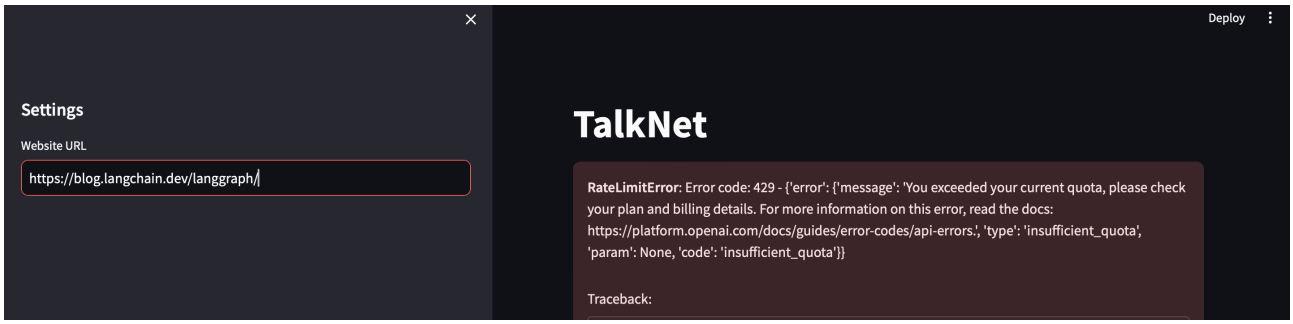

Figure 5.3: OpenAI Rate Limit Error

Chapter 6

Conclusion

TalkNet represents a significant step forward in the field of web interaction, leveraging advanced NLP techniques and Machine Learning to provide a conversational browsing experience. The project has laid a strong foundation for transforming how users interact with digital content. The future enhancements outlined promise to not only refine the system but also broaden its applicability and effectiveness. The potential of TalkNet to revolutionize web interaction is immense, offering a glimpse into the future of digital communication where information retrieval is seamless, intuitive, and conversational. This project underscores the importance of continual innovation in the NLP and AI fields, highlighting the transformative impact of these technologies on everyday digital interactions.

Chapter 7

Future Work

In the future, the following enhancements are planned in order to create a fully functional application:

1. **Optimization of API Usage**
   Overcome the rate limit challenges so that an answer can be generated for a particular query entered by the user.

2. **Improvement of Semantic Search Algorithms**
   Refine the algorithm used for semantic search to improve the accuracy and relevance of the search results.

3. **Scalability and Parallel Processing**
   Implement a more robust system capable of handling multiple URLs concurrently, significantly increasing the throughput of data processing and allowing for more comprehensive and faster response generation.

These enhancements will not only address current limitations but also expand the capabilities of TalkNet to serve a wider array of user needs.

References

[1] D. Madhav, S. Nijai, U. Patel, and K. Champanerkar, "Question generation from pdf using langchain," in *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2024, pp. 218–222.

[2] A. S. A. S and P. J. Sai, "An effective query system using llms and langchain," *International Journal of Engineering Research & Technology (IJERT)*, vol. 12, no. 06, June 2023.

[3] D. Lin, "Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition," 2024.

[4] K. Pandya and M. Holia, "Automating customer service using langchain: Building custom open-source gpt chatbot for organizations," 2023.