DAB-402 Capstone Interim Project Report

AI Chatbot for St. Clair International Team

Group 11

Group Members
Parth Tripathi-0826520
Ruturaj Solanki-0827884
Chris Chhotai- 0826416
Isha Dhaduk-0827577
Manushree Venkatesh-0828892

Guided By,
Prof. Abiodun Sodiq Shofoluwe

Contents :

Abstract

In this ongoing project, we are dedicated to the development of an AI chatbot tailored for the St. Clair International Team, harnessing cutting-edge Natural Language Processing techniques. Our primary aim is to seamlessly facilitate interactions for individuals navigating the enrolment process at St. Clair College. Initially, we gather data from the St. Clair International Team and augment it through web scraping technology, comprehensively capturing information on the college's courses.

We are constructing a Large Language Model to drive the chatbot's functionality, prioritizing this approach amidst resource constraints. Subsequently, we explored the Transformer model available on Hugging Face for testing purposes. Challenges emerge when discrepancies between our text-based data and the model's requirements surface, prompting us to restructure our data into a columnar format.

Throughout the project's trajectory, we are maintaining a steadfast focus on fine-tuning hyperparameters to elevate the chatbot's performance continually. Employing an iterative methodology, we iteratively refine and test the chatbot to ensure it furnishes accurate and relevant responses to user inquiries. As we progress, our endeavors are centered on optimizing the chatbot's capabilities to deliver invaluable assistance to prospective students navigating the enrollment journey at St. Clair College.

Acknowledgments

We would like to thank the St. Clair International Team for providing us with the necessary data and assistance for this project. We also want to thank our instructor and mentors for their significant criticism and support during the project. In addition, we thank the open-source community for providing us with the tools and resources we needed to design and test our AI model. Finally, we want to thank everyone on the team for their hard work and collaboration in making this project a reality.

Introduction

In this ongoing endeavor, we are committed to the creation of an AI chatbot, specifically tailored for the St. Clair International Team, leveraging the power of advanced Natural Language Processing techniques. Our primary goal is to streamline the interaction process for individuals embarking on their enrollment journey at St. Clair College. Our initial steps involve gathering data from the St. Clair International Team and supplementing it with comprehensive information on the college's courses, obtained through web scraping technology.

Description of Dataset

The stakeholder has provided us with a comprehensive dataset containing information about international students from various countries. This dataset includes details such as the courses they are selecting and their home countries. This information was analyzed in the previous semester and provided valuable insights into the preferences and demographics of the international student body.

However, for the current semester, we realized that this dataset was not sufficient to train our Large Language Model (LLM). Therefore, we supplemented this dataset with additional information obtained from the college's website using web scraping techniques. This new data includes the Program Name, Overview, Admission Requirements, and other pertinent information related to each course offered at St. Clair College.

This enriched data set now provides us with all the necessary information to generate accurate and relevant outputs from our LLM model. It serves as a valuable resource for understanding the needs and preferences of prospective students, thereby enabling us to optimize the functionality of our AI chatbot. Overall, this dataset forms the backbone of our project, driving our efforts to enhance the enrolment experience at St. Clair College through AI technology.

DATA PRE-PROCESSING

Data Extraction

Web scraping, utilizing tools like Beautiful Soup in Python, allows us to systematically extract crucial information from college websites such as program details, admission requirements, faculty information, and course offerings. By navigating the HTML structure of web pages, we pinpoint specific elements, send HTTP requests, parse HTML content, and extract relevant data efficiently. This method is pivotal for our AI chatbot project for St. Clair College, enabling us to gather precise information for the bot's knowledge base. However, it's essential to approach web scraping ethically, ensuring compliance with websites' terms of use to uphold responsible and lawful data acquisition practices.

Data Cleaning

After employing web scraping techniques to extract data from college websites, we encountered a common challenge: the presence of excessive white spaces within the scraped data. To address this issue and ensure the cleanliness and consistency of our dataset, we systematically removed these unnecessary white spaces. This meticulous data preprocessing step significantly improved the quality of our extracted information, enhancing its readability and usability for subsequent analysis and integration into our AI chatbot project for St. Clair College. By eliminating these extraneous spaces, we optimized the efficiency of our data processing pipeline, ultimately facilitating more accurate and reliable responses from the chatbot.

TOKENIZATION

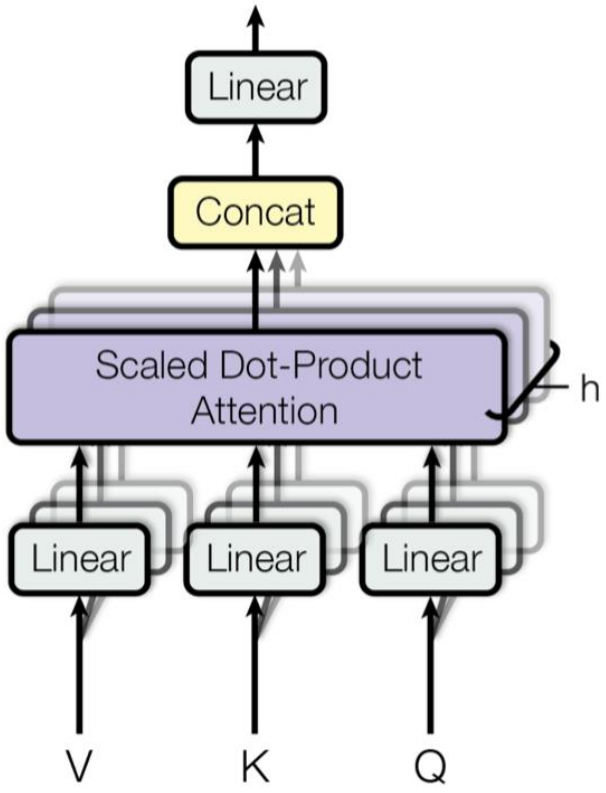
Tokenization serves as a fundamental preprocessing step in our project, where we harnessed both character and word tokenization techniques to analyse the data sourced from college websites via web scraping. Character tokenization involves breaking down the text into individual characters, offering insights into the granular details of each word. On the other hand, word tokenization dissects the text into meaningful words, presenting a more comprehensible representation of the content. A comprehensive study conducted as part of our project indicates that, in our specific case, word tokenization proves to be more beneficial. Word tokenization provides a more contextually relevant understanding of the data, aligning well with the nature of our project, which involves processing textual information related to college programs. This approach enhances the accuracy and interpretability of subsequent natural language processing tasks within our AI chatbot, facilitating a more effective and user-friendly interaction with the acquired data.

MODEL BUILDING (Large Language Models)

In our project, we aimed to develop a specialized language model tailored specifically for St. Clair College. To accomplish this, we conducted extensive training using a diverse dataset sourced from the college's websites. We employed both character and word tokenization techniques, ultimately finding that word tokenization yielded more nuanced insights into the college's programs, admission requirements, faculty information, and course offerings.

Throughout the training process, we curated a comprehensive dataset encompassing all relevant aspects of St. Clair College. Our objective was to imbue the model with a deep understanding of the college's context, enabling it to provide accurate and personalized responses to user queries. Our goal was to enhance the user experience by offering detailed and relevant information, thereby facilitating smoother interactions with St. Clair College.

We used multiheaded attention in building LLM from scratch. As it has multiple heads and each of them operating differently. Each attention head is responsible for capturing different patterns or relationships in the input data. The outputs from each head are typically concatenated or linearly combined to produce the final attention output. Ultimately, it will help the model to focus on different parts of the input sequence simultaneously, capturing a richer set of relationships.



HYPER PARAMETER TUNING – FALCON AND LLAMA 2

Recognizing the need for enhanced accuracy in our project, we made a strategic decision to transition from our custom-built language model to leveraging pre-trained models, namely Falcon and Llama 2. Despite our initial efforts, the intricacies of generating accurate outputs posed challenges, prompting us to tap into the capabilities of these pre-trained models. Falcon and Llama 2 are renowned for their proficiency in natural language understanding, having undergone extensive training on diverse datasets. By adopting these pre-trained models, we anticipate a significant improvement in the AI chatbot's ability to comprehend and respond to user queries related to St. Clair College. This shift allows us to capitalize on the advanced linguistic knowledge embedded in these models, providing a more robust and reliable foundation for our chatbot's performance in delivering precise and contextually appropriate information.

DEPLOYMENT

The deployment and integration phase of the project involved utilizing Flask, a Python web framework, and the Transformer library to create a user-friendly interface for the AI chatbot.Initially, the team encountered challenges in generating the necessary "safe.tensor" file required for deployment. However, through perseverance and adjustments to the model's parameters, the team successfully generated the file and deployed the chatbot on an HTML platform.

1. Front-end Development

The team dedicated efforts to developing a user-friendly front-end interface, ensuring a seamless and intuitive experience for users interacting with the AI chatbot. This involved creating a visually appealing and
responsive design, incorporating intuitive chat windows and input fields
for users to submit their queries.

2. Back-end Integration

On the back-end, the team integrated the trained language model with the Flask application, enabling real-time communication between the user
interface and the AI chatbot. This involved implementing APIs and establishing secure data transfer protocols to ensure efficient and secure processing of user queries.

3. Testing and Optimization

Prior to the final deployment, the team conducted rigorous testing and optimization processes. This involved simulating various user scenarios, evaluating the chatbot's responses, and identifying potential areas for

improvement. Performance metrics, such as response time and accuracy, were closely monitored and optimized for a seamless user experience.

Ethical Considerations and Responsible Development

As the team embarks on the development of an AI-powered chatbot, ethical considerations and responsible practices are of paramount importance. The team is committed to upholding the highest standards of data privacy, security, and transparency, ensuring that the chatbot's interactions with users are conducted in an ethical and trustworthy manner.

Data Privacy

The team prioritizes data privacy by implementing robust data protection measures, such as encryption, access controls, and anonymization techniques, to safeguard users' personal information and maintain their trust.

Fairness and Nondiscrimination

The chatbot's algorithms and models are designed to be unbiased and nondiscriminatory, ensuring fair and equitable treatment of all users regardless of their background or personal characteristics.

Transparency And Explainability

The team strives to maintain transparency by providing clear explanations about the chatbot's functionality, data sources, and decision-making processes, promoting trust and understanding among users.

Security and Robustness

Rigorous security measures are implemented to protect the chatbot from potential threats, such as cyber-attacks, data breaches, and malicious inputs, ensuring the system's robustness and integrity.

FUTURE SCOPE/CONCLUSION

In extending our project's scope, we aspire to transform our AI chatbot from a mere enrollment assistant into a comprehensive educational advisor for students at St. Clair College. The integration of personalized course recommendations represents a pivotal step towards this vision. By delving into machine learning algorithms, we aim to unlock the potential of vast datasets encompassing historical enrollment patterns, academic performance metrics, and user interaction logs. Through sophisticated analysis of this rich data landscape, our chatbot will evolve into a proactive guide, capable of discerning nuanced patterns and preferences to tailor recommendations that resonate with each student's unique academic journey.

In practical terms, this expansion will empower students to navigate the complex terrain of course selection with confidence and clarity. Leveraging insights

gleaned from past enrollment trends and user interactions, the chatbot will serve as a trusted advisor, steering students towards courses that not only align with their academic aspirations but also optimize their educational trajectory. Whether it's suggesting elective courses that complement their major, highlighting interdisciplinary opportunities, or identifying emerging fields of interest, the chatbot will offer personalized guidance that transcends traditional enrollment support. Ultimately, by harnessing the power of machine learning to deliver tailored recommendations, our project endeavors to enhance the student experience, foster academic success, and empower individuals to realize their full potential at St. Clair College and beyond.