**UNIVERSITY OF NEW HAVEN**

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

A logo of a university

AI-generated content may be incorrect.

# Capstone Project TITTLE:

# **University of New Haven Chatbot**

**INSTRUCTOR:**

**Assistant professor Dr. Muhammad Aminul Islam and**

**Professor Dr.Ardiana Sula**

**TEAM**

1. Nanduvardhan Reddy Amireddy
2. Sirisha Kasetty
3. Madhavi Kancham
4. Anusha Katapally
5. Chetana Balagam
6. Sana Tabassum Shaik
7. Kaleswara Manikanta Daddanala
8. Lavanya Gurram
9. Akhila Immanni
10. Surya Manjri

**Report 2: University of New Haven Chatbot (week2)**

**Objectives of week 2:**

* **Data collection –** collected data from university course catalogue data and from surveys.
* **Web scrapping -**used **beautiful** soup , scrapy to extract data from webpage.
* **Data cleaning –** we removed irrelevant data, duplicates, handled missing data using python .
* **Data storage –** we created vector database chromaDB for storing the data.
* **Preparing questions and answers –** we converted the data chunked data into questions and answers and stored in Json format.

1. **Data collection:**

* we collected data from various resources mainly from university course catalogue page that contains the required course information.
* we also collected data from student survey from our university.

**2.Web scrapping:**

* Collected course-related data such as course names, codes, descriptions, prerequisites, credits, instructors, and schedules.
* Used web scraping tools like Beautiful Soup, Scrapy to extract data from the identified HTML elements.

**3. Data cleaning:**

* weremoved duplicates, irrelevant data, or HTML tags.
* Prepared the data in Standardize formats (e.g., dates, course codes, credits).
* Handled the incomplete data and filled the gaps between the data .
* Later converted the cleaned text into json format.

**4. Data storage:**

* We created the vector database chromaDB to Store the collected course data in a structured and efficient manner for easy retrieval and processing.
* chromaDB can handle large datasets efficiently. We used it because it is scalable
* It is Ease of Use and Simple API for storing and querying data.
* We broke the datasets into smaller, manageable chunks for better processing and storage.
* Use libraries like nltk or spaCy to split text data into sentences or paragraphs.

**5. preparing Q/A:**

* Transformed the chunked data into a structured JSON format suitable for training the chatbot.
* Define a schema for the JSON structure, including fields like question, answers like course code, topic , Prerequisites .
* Map each chunk of data to a question-answer pair.

**6. challenges and solutions :**

**6.1 – chunking the data**

* **Question –** Splitting course descriptions into chunks without losing context and also determining the optimal size of the chunk.
* **Solution** - Experimented with different chunk sizes and evaluate their impact on chatbot performance.

**6.2 -Json Q/A conversion issue**

**Question** - Mapping Chunks to Q&A Pairs - Creating meaningful questions and answers from chunked data. Resolving ambiguous or incomplete data that may not fit neatly into Q&A pairs.

* **Solution -** Use predefined templates to generate questions (e.g., "What are the prerequisites for [DS -6001]?").

**7. Keyoutcomes :**

* Successfully scraped course data (e.g., course names, codes, descriptions, prerequisites, credits, instructors, and schedules) from the university website.
* Cleaned and standardized the scraped data, removed irrelevant content, resulting in a high-quality dataset ready for further processing.
* Stored the cleaned data efficiently using ChromaDB also Organized data into collections for easy retrieval and querying.
* Transformed chunked data into a structured **JSON Q&A format**, making it suitable for training the chatbot.

**8 . Next steps :**

The following steps are – we have to train the LLM model with RAG based model and also fine tuning.