Name: Sihang Wang

Group: ChatDB 77

Course: DSCI 551

Assignment : Group Project Proposal

Date: February 7, 2025

ChatDB - Natural Language Movie Database Management System

About Team

Sihang Wang is a first-year master's student majoring in Applied Data Science. He holds a bachelor's degree in Business Analytics and Information Management & Technology. As the sole project leader, coder, and author of his work, he has developed strong expertise in data-driven problem-solving. Sihang has hands-on experience with Python, R, and SQL, enabling him to tackle complex analytical challenges and build data-driven solutions.

Project Overview

Develop a natural language interface that enables users to interact with a movie database using natural language, achieving data query, exploration, and modification. Specifically, users can ask questions in natural language, such as "What movies has Tom Cruise starred in?", and the system will parse the user's question, translate it into a corresponding database query, and then return the query results in natural language.

Project Functions

4 Key Functions:

1. Support database architecture exploration: query tables and fields in the database
2. Support common SQL query operations: SELECT, FROM, WHERE, GROUP BY, JOIN. etc.
3. Support natural language requests to insert, delete, and update data.
4. Use natural language to generate SQL queries and execute and return results.

Database System and Architecture Design

Since this is a one-person group, which means that it will only contain one type of database. The selected database type is RDBMS, particularly MySQL. The them is a Movie Database. This database will contain the following tables:

1) movie_db (movie database): stores information related to movies.

- Tables: Movie, Actor, Genre, MovieActor (relationship table between movies and actors).

2) review_db (review database): stores user reviews and ratings.

- Table: review (contains movie reviews, ratings, and timestamps).

3) user_db (user database): stores user information and preferences.

- Table: user (contains user ID, name, and movie preferences).

The three datasets that will be used in this project must be movie-related, and it will be showing the join functions particularly to show the comprehensiveness of the database.

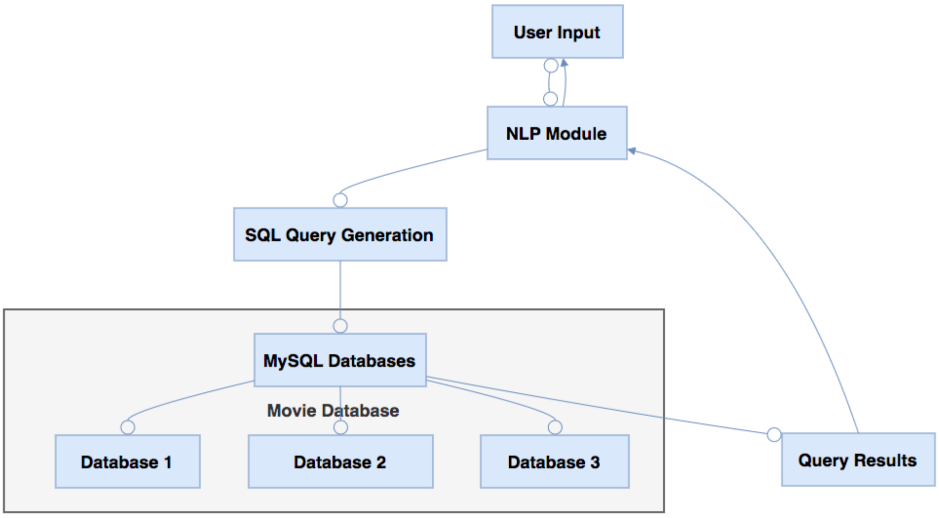
Technical Implementation and Development Plan

1. **Technical stack**
   a) MySQL as database
   b) Python as programming language for implementing the Natural Language Processing features.
   c) Libraries use: spaCy or NLTK for text processing and query interpretation; mysql-connector-python to connect and interact with MySQL.
   d) Use OpenAI API, might consider as GPT-3.5 Turbo or GPT-4.0o, which has powerful natural language understanding and generation capabilities (OpenAI, 2023), which can effectively convert users' natural language questions into database query statements.

   In order to improve query efficiency, the generated database query statements can be optimized by using indexes and optimized query algorithms. In addition, a cache mechanism can be used to cache commonly used query results to avoid repeated queries.

2. **System Architecture**
   To support the system functions defined in previous section, the system architecture is designed as below (Figure 1). The system architecture is mainly divided into three levels.
   (1) The first level is the customer input query requirements, such as which movies Tom Cruise has played;
   (2) The second level is Natural Language Processing, the system converts the customer's questions into SQL language so that I can interact with the database;
   (3) The third level is database processing, which queries the corresponding results according to SQL. The result will be returned to the user.

   Figure 1.



3. Project Plan

Noticed that all the project will be fully designed, coded, and performed by Sihang Wang, who is the only person in this group.

Week	Tasks & Development Phases	Milestone / Deliverable
Week 1-2 (Feb 1-14)	- Define the database schema and set up MySQL.	Project Proposal Submission (Feb 7)
	- Implement basic MySQL connection and test simple queries.	
	- Explore NLP techniques for natural language query processing.	
Week 3-4 (Feb 15-28)	- Develop NLP processing to interpret user queries.	N/A
	- Implement query translation from natural language to SQL.	
	- Test basic SQL queries (SELECT, WHERE, JOIN).	
Week 5-6 (Mar 1-15)	- Implement data modification (INSERT, UPDATE, DELETE).	Midterm Progress Report Submission (Mar 7)
	- Integrate NLP with MySQL for modifying database records.	
	- Implement error handling for invalid or ambiguous queries.	

Week 7-8 (Mar 16-30)	- Conduct testing & debugging to refine NLP-to-SQL accuracy.	N/A
	- Implement advanced queries (GROUP BY, ORDER BY, LIMIT).	
	- Improve error handling & query validation.	
	- Optimize system response time & efficiency.	
Final Week (Apr 1-7)	- Conduct user testing and finalize system features.	N/A
	- Finalize documentation and user instructions.	
	- Prepare for the live demo and ensure smooth performance.	
Live Demo (Apr 21 & 23)	- Present real-time demonstration of system capabilities.	Live Demo Presentation
	- Showcase schema exploration, query execution, and data modifications.	
	- Demonstrate handling of complex queries.	
Final Report (May 9)	- Submit a comprehensive final report detailing design, implementation, challenges, and solutions.	Final Report Submission
	- Include a Google Drive link to the project code and documentation.	
Project Implementation (Demo Date)	- Deliver a fully functional system with all features.	Final Project Submission
	- Upload project codebase & documentation before the demo.	
	- Ensure proper project submission & access permissions.	

4. Expected Outcomes

This project will create a user-friendly interface for interacting with MySQL databases using natural language. It will simplify database querying and management, making it accessible to a broader audience (Jurafsky & Martin, 2020). By the end of this project, users will be able to communicate with the database in a way that feels intuitive and natural, reducing the need for technical expertise in SQL.

References

Hernandez, M. and Koudsi, M., 2018. *Designing Relational Databases for Entertainment Industry Systems: A Case Study*. Journal of Database Systems, 29(4), pp. 54-68.

Jurafsky, D. and Martin, J.H., 2020. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd ed. Prentice Hall.

OpenAI, 2023. *GPT-3.5 Turbo: A Technical Overview*. OpenAI Documentation. Available at: https://openai.com/docs [Accessed 5 February 2025].