ChatDB on MySQL Database

ChatDB 87 Group

Qianshu Peng

1. Team Member Background and Skills

My team is a one-person team, and I'm the only member. I'm Qianshu Peng, a first-year student in the Master of Applied Data Science program. I completed my undergraduate degree with Joint Major of Economics and Mathematics with a minor in Data Science at UCSD. I am proficient in Python programming skills with data analysis and statistical learning packages (Pandas, Numpy, Sklearn, etc.) and applied data analysis in the economics-related areas with Stata and R. I also have experience in Java programming and data management with SQLite.

2. Project Requirements

Based on the project guidelines, the project requires me to:

1) Create a MySQL database to store data and use Python to manipulate the data.
2) Create a user interact module and interface in Terminal which allows the program to take dataset files and natural language prompts. Also, it should be able to guide the user about how to use the DB.
3) Analyze the dataset and allow the program to have some understanding of its structure and characteristics. For example, the columns being categorical/quantitative, the relations between tables, the identifier, etc.
4) Create templates for SQL queries, including all kinds of keywords and cross-table queries. Check and make sure that all the queries are executable.
5) Understand the natural language prompts with certain patterns. Then, generate queries that fit the prompts' requirements.
6) If it's required, execute the queries on the dataset and output results.

For all the function requirements above, the DB should be tested on at least 3 datasets; I will choose at least 1 dataset that includes multiple tables. Also, the DB should have documentation for users and be uploaded to a Google Drive folder with the program files.

3. Planned Implementation

All implementation will be programmed in Python. MySQL will be used as the database. Since this is a one-person team, this will be the only database I support.

- User interaction: I almost know nothing about this part besides the sys module used in HW1 and the built-in input() function. I will learn to take files from their path and to build a readable and user-friendly interface in Terminal.
- Database and management: I will use MySQL modules in Python. As I never used MySQL, I will research how to manipulate the database in this way next week. I plan to check the datasets by keywords they have in the tables and create relation tables by SQL sentences. Also, the datasets will be checked to extract their characteristics and stored in the program.
- NLP: This part will basically become brainstorming on words. As I might only need to deal with some short sentences or word groups as prompts, I will do tokenization and stemming using relative packages. Then, certain keywords and patterns will be matched with query templates (e.g. total A by B → GROUP BY with SUM). This part will be done with regex and NLP packages (I only have a few experiences with NLP tasks so I'm not sure about what to use at this point; will figure it out this or next week).
- Queries: I'm not very clear about the structure of the program in this part yet. Maybe I will write several functions for certain keywords/templates (like example, where, having, group by, etc.), and they will take column name keywords to output the required queries. The functions should also be able to deal with joining tables by searching the column names in table information. I think regex will also be used in this part to form queries of certain formats. Also, the MySQL package will be used to execute the queries.

4. Team Responsibilities

Since this team is a one-person team, all tasks will be completed by myself.

5. Timeline

Rest days in September: outline the project, think about what files I will need, and frame the whole project, including the import relationships among Python files; learn some packages, such as NLP module, MySQLdb, SQLAlchemy or mysql.connector (still hesitating yet), input processing module, etc.

October before Midterm progress report (10/18): Program on database connection, dataset taking, prompt taking, and some basic SQL queries. Also, complete the midterm report.

Rest days in October and early November: Complete the more complicated queries and natural language processing part. Solve the issues in midterm report.

A week before the live demo (11/26): Find appropriate datasets to tune and test the code; prepare for the submission and in-class presentation.

Rest days before final submission (12/13): Write final report; make some trivial adjustments to the codes (if it is allowed to make changes after implementation is submitted)