Midterm Progress Report: ChatDB on MySQL Database

ChatDB 87 Group

Qianshu Peng

1. Implementation Questions

1.1 Tech Stack:

Pymysql: This is used to connect to the MySQL server and run SQL queries.
Nltk: This is used for natural language processing (plan to use, not implemented yet).
Re: This is used for regex to match the patterns of languages to elements of queries.
Pandas: To take CSV datasets from users' uploads and analyze them before putting them into the MySQL database.
Os: to deal with the file path inputs.

1.2 Query Syntax

I only support SQL queries. In the queries model, I support generating sample queries and queries from the input. In this part, I use formatted strings for output query sentences (since SQL queries have relatively fixed structures) and let the functions take keyword arguments as input (the arguments should be analyzed in the NLP module). I write different formatted strings for different patterns of queries.

1.3 Database

The database I choose to use is MySQL. As this is a one-person team, this will be the only database I support.

2. Planned Implementation

I will complete this part of the report based on the proposal.

- User interface: This part is almost done. It is composed of a while loop and a lot of if-else structures. I use print to give the user simple instructions (and write specific ones in the documentation) and take the choices, natural language prompts, and user's dataset addresses using input(). I haven't tested it thoroughly since some of the called functions are not implemented yet.
- Database management: This part has a half-progress version, but I think there will be changes to it. I set the configuration of the MySQL database and connected to it using the pymysql package. I realized that the approach in the proposal is not practical since I

really have no idea about how to let the program identify relations and keys by itself. So I decided to manually construct the built-in database and ask the users to upload their database in the form of relational tables with the information like keys. Besides, the function of executing query in the database is in this part and not completed yet.

- Dataset analysis: This part is new and not in the proposal. I realized that this part needed to be independent in the progress of programming the database part, but I'm a bit confused with it. I wrote functions to identify variable types (categorical/quantitative), single-column unique keys in a table, and possible foreign key relationships across tables, although I'm doubtful about their time efficiency and tend to let users enter these data.
- NLP: This part is almost an empty .py file at this point. I plan to deal with the input sentence using nltk and re packages. The words will be tokenized, and regex search will extract the necessary variables mentioned in natural language prompts. Actually, I'm considering if I really need the nltk package, or if I can just give a relatively normative format of prompts (written in documentation). Also, I’m considering limiting the input of arguments to be the same with the column names, or it will be hard to “understand”, say, quantity is actually “transaction_qty”, or sales should be calculated as “transaction_qty * unit_price”. But I don’t think this is very correct.
- Queries: sample queries and query generators from prompt are similar in the way of creating queries (formatted strings with argument inputs). I’m still brainstorming about those complex queries involving multiple tables or aggregates: in these cases, it’s hard to guarantee that the generated queries are executable.

3. Status of the Project

Generally, the easy parts of each planned implementation part are completed and might need to have changes. The complex parts are still waiting for me to deal with, as said above. The progress is behind my expectations due to other coursework and job applications.

4. Challenges

To be honest, I think this project is very tricky and I encountered and solved countless problems in every part of it. One challenge I tackled was to allow every machine to connect to my MySQL database on EC2 from local command lines. I did some research and adjustments to my server to implement this.

Also, as I’m not very familiar with moduled programming, it was hard for me to decide on the whole structure of this project. I first created a file for each proposal part, then added or deleted files in the programming process according to my needs.

5. Timeline

I wish I could complete a task per week. The order of tasks might not go with the checkpoints listed below strictly.

10/28	Complete the user interface and database analysis parts.
11/4	Complete the natural language processing part.
11/11	Complete the NLP and query generation part.
11/18	Complete the database and execution part.
11/25	(if possible) Add more query templates; submit the program files in Google Drive; perform a final comprehensive test on the whole project.
11/26	The in-class demo day; everything should be completed and submitted before.