CRITICALRIVER

Querying Tabular Data Via LLM

Problem Statement & Solution

Problem Statement:

In many cases, when working with large tabular datasets (CSV, SQL, etc) making sense of the data without knowledge and/or use of advanced statistics software is usually not possible. For example, if we want to find random occurrences of certain keywords that fulfil a condition, we would need to use Excel or Pandas for example. This is a task that someone who is not very proficient with computers may struggle to do.

Solution:

Utilize LLM's ability to parse through text queries or understand natural language to chat or query your data as if it is a person. This would allow anyone to be able to make complex queries on data.

Steps Involved:

1. Vectorized and Split Data
2. Save data to Vector DB
3. Setup LLM and GUI
4. Use LLM to understand user query via prompt engineering
5. Extract keywords and information from user query using engineered LLM prompt
6. Query DB and feed context into LLM when appropriate
7. Provide User with final answer

Challenges

- Context window Size
  - Initially used Flan-T5-XXL as LLM for project
  - Context window was too small, experimented around with lots of querying methods and data splitting methods
  - Eventually came to conclusion that the window was too small, switched to GPT 3.5-16K as it has a much larger window
- We had to breakdown user queries into set of steps and attributes
  - Used LLM itself to extract key information from user query
  - Used information extracted programmatically
- Tried to use table based model
  - TAPAS couldn't work with the python library we were using (Langchain), it gave us an even larger reason to use GPT as it was supported in the library way more.

Use case – Process Flow

- User would upload their CSV file to GUI
- Software will Vectorize it and prepare it for querying
- User will make queries on file within app in Chat like interface

This whole process will happen on a web browser.

Results

Shown in Demo Video

Benefits

- User doesn't require any knowledge of data science or need to use any advanced software
- User can directly interact with data without any person needing to look over it as well
- Simple solution that is up and ready in minutes
- Minimal Hardware requirements

Future Work

This could be very useful to businesses who want to have a clearer idea of their data but do not want to dive deep into data analytics. Also great for companies with very large spreadsheets who want to save time and effort skimming through their data.

THANKS!

Pradyun Magal
Pradyum.magal@criticalriver.com