A Natural Language Query System for Data Visualization

CS Capstone Project

Brandon Wilson

Tufts University
Fall 2024

High-level Requirements

1. NLP Understanding

- Implement keyword and pattern matching system.
- Have predefined set of query templates for common data analysis questions.
- Use regex or tokenization and part-of-speech tagging to interpret user input.

2. Guided Query Building

- Implement step-by-step query builder that feels conversational.
- Start with broad categories, (“Would you like to do a comparison, trend or distribution?”) etc.
- Based on user’s choice, provide more specific options (“Which fields do you want to compare?”)

3. Query Disambiguation

- If input is ambiguous, provide multiple choice options
- For example if the user just says sales, based on the data, ask to clarify (“total sales”, “unit sales”, etc.)

4. Predefined Visualization Rules

- Create a set of rules that map query types to appropriate visualizations.
- For example, comparisons default to bar charts, trends or time series to line charts etc.

5. Template Based Explanatory Text

- Ex. “This bar chart shows how {variable} changes over {time period}.”

4. User Feedback Loop

- Implement simple feedback mechanism where users can indicate if the visualization is helpful.
- Use feedback to refine mapping over time.

Target Audience and Users

1. Primary Users

- Business analysts without deep technical expertise
- Sales organizations
- Executives and decision-makers

2. User Characteristics

- Limited or no knowledge of SQL or complex data querying
- Need quick data insights
- Varied levels of data literacy

3. Use Cases

- Ad hoc data exploration
- Regular business reporting and KPI tracking
- Presentation preparation
- Quick fact-checking and hypothesis testing

4. Industries

- This could be applied to a wide range of industries. - Retail, finance and banking, sales orgs, healthcare, education, tech and software companies, etc.

Future Possibilities

I'd love to turn this into a useful product to sell down the road. With that said, another useful implementation that I want to incorporate (but may be out of scope for this particular project) is data reporting. Letting the user set up specific queries about their data that they would like to monitor. The software then generates reports on the user-indicated cadence and emails the report to them.

Engineering Diagram