Pratyush Verma
20CS8092

A
Project Report on

“DataSpeak Engine: Conversational Database Interface”

Submitted by
Pratyush Verma
(20CS8092)

Under the Guidance of
Prof. Suvrojit Das

Department of Computer Science and Engineering

National Institute of Technology,

Durgapur

(2023 – 2024)

Pratyush Verma
20CS8092

Introduction

The DataSpeak Engine: Conversational Database Interface is a pioneering project that blends natural language interaction with database querying. Utilizing cutting-edge technologies such as Google Palm and Langchain , this initiative seeks to redefine user experiences by seamlessly integrating human language with MySQL database interactions.

Tailored for the context of a T-shirt store, the project addresses the specific needs of users, particularly store managers, enabling them to effortlessly communicate queries in natural language. Through intelligent processing, the system translates these language inputs into precise SQL statements, ensuring efficient execution on the MySQL database. This capability provides real-time insights into crucial aspects such as inventory levels, sales data, and applied discounts.

Pratyush Verma
20CS8092

Functionalities

Natural Language Interaction:

- Users can inquire about various aspects of the T-shirt store using natural language.
- Example Queries:
  "How many white color Adidas t-shirts do we have left in the stock?"
  "What is the projected sales if we sell all extra-small size t-shirts after applying discounts?"

Query Generation:

- The system is intelligent enough to understand and interpret natural language queries.
- Converts user queries into accurate SQL statements for database interaction.

Database Interaction:

- Executes SQL queries on the MySQL database to retrieve relevant information.
- Supports retrieval of data related to inventory levels, sales projections, and discounted prices.

Use Case Example:

- Store managers can effortlessly seek critical information by asking questions in plain language.
- Example Scenario:
  Manager: "How much sales our store will generate if we can sell all extra-small size t-shirts after applying discounts?"
  System: Translates the query into SQL, executes it on the database, and provides the calculated sales projection.

Pratyush Verma
20CS8092

Approach and Process Overview

1. Embedding Generation:
   The training dataset undergoes transformation into embedding vectors utilizing the Hugging Face library. The resulting embedding vectors encapsulate the semantic information of the dataset.

2. Chroma DB Storage:
   The generated embedding vectors are stored in Chroma DB, a database designed for efficient vector storage.

3. Few-Shot Embedding Creation:
   Embeddings are created for a few-shot scenario, capturing the semantic essence of a limited set of examples.

4. Retrieval Process:
   During retrieval, the system queries Chroma DB for the topmost semantically close example to a given vector.The retrieved example is indicative of the semantic proximity based on the stored embeddings.
   .

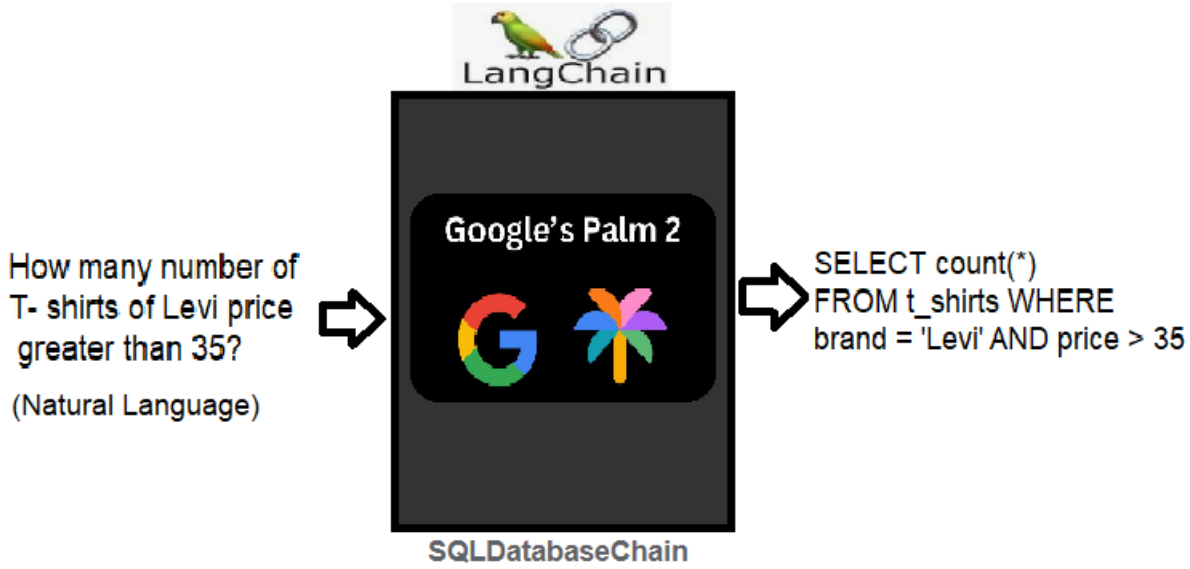
Pratyush Verma
20CS8092

Technologies and Architecture

Technologies Used:

1. Google Palm: Leveraged for advanced language understanding and processing.
2. Langchain: Utilized for seamless interaction between natural language queries and SQL generation.
3. Huggingface Library: To convert training dataset into embedding vector.
4. Chroma :It is a vector database used to store embedding vector.
5. MySQL: For retrieving data

Architecture:



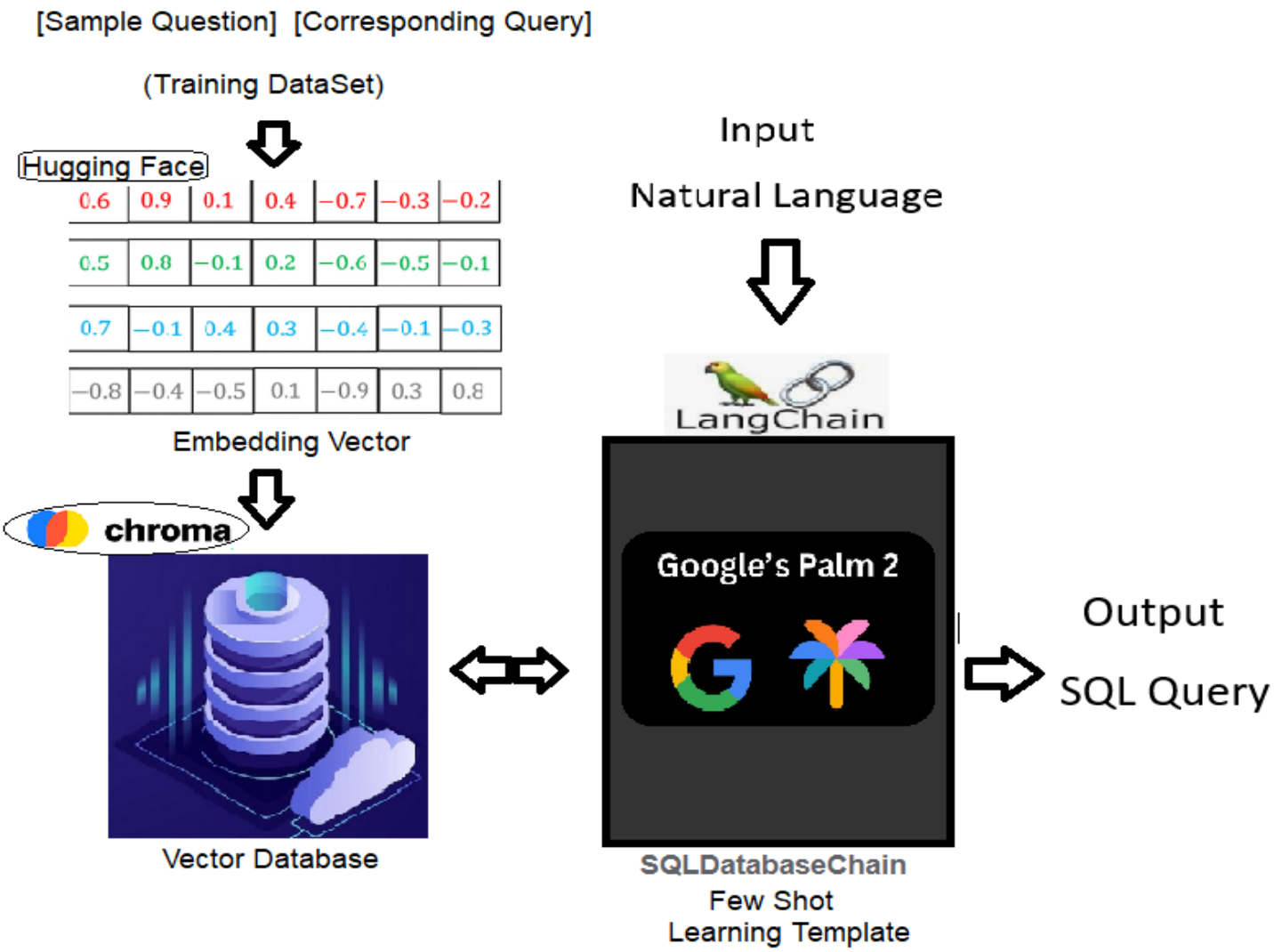
Pipeline (for Simple Query)

Note:

It works for Simple Query but fails for complex queries. For complex queries we use the concept of few shot learning. It refers to a machine learning paradigm where a model is trained to make

accurate predictions with only a small number of examples per class.

For Complex Queries:



Pipeline

Pratyush Verma
20CS8092

Results:

```
> Entering new SQLDatabaseChain chain...
How much is the price of all white color levi t shirts?
SQLQuery:SELECT sum(price*stock_quantity) FROM t_shirts WHERE brand = 'Levi' AND color = 'White'
SQLResult: [(Decimal('9768'),)]
Answer:9768
> Finished chain.
```

Results given by LLM

```
7    SELECT * FROM t_shirts WHERE brand = 'Levi' AND color = 'White'
```

Result Grid | Filter Rows: | Edit: | Export/Import:

t_shirt_id	brand	color	size	price	stock_quantity
30	Levi	White	S	15	69
49	Levi	White	M	44	91
9	Levi	White	L	33	93
8	Levi	White	XL	20	83
NULL	NULL	NULL	NULL	NULL	NULL

Actual Results

Pratyush Verma
20CS8092

Conclusion

The DataSpeak Engine stands as an intelligent and user-friendly system that bridges the gap between natural language communication and database interaction. By providing a conversational interface, it streamlines the process of retrieving vital information related to the T-shirt store's operations. The project successfully combines language understanding technologies and database interaction, offering a promising solution for enhancing efficiency in managing inventory, predicting sales, and implementing discounts. Future enhancements could include expanding support for additional databases and further refining natural language processing algorithms for even more complex queries.