Chapter 1

Querying Numerical Databases in Natural Language: GPT3.5 implementation

Femi Adeniran

The idea behind this project is to use a natural language interpreter to extract information from a database and convert it into code, which can then be used to retrieve the desired data. My inspiration for this project came from OPENAI's natural language to SQL feature.

1 Data Collection

FBref is a website that provides a comprehensive database of football statistics, covering a wide range of leagues and competitions from around the world. I used a specialized scraper to get data of the 2022/2023 English Premiership Season. Once the data was extracted, I combined tables as some are stored in different pods.I formatted this data from csv imported into a Pandas dataframe to make it easy to query in Python.

2 Method

At its core, our objective is to create a natural language interface that allows users to easily interact with databases without requiring any prior knowledge of programming languages or complex query languages. To achieve this, we are utilizing OpenAI's advanced natural language processing capabilities to extract meaning and context from user prompts relating to football and translate them

into executable Python code. This involves providing a clear description of the database being queried, along with examples of the types of queries that will be made, and the desired output. Once the query is submitted, we expect to receive a Python code reply from the OpenAI servers, which we then intercept and execute on the database as described, before returning the results to the user via a user-friendly interface. Overall, this approach allows users to easily access and analyze data without needing to possess any specialized database skills or expertise.

3 How to

To replicate this project from scratch, the first thing you need is an OPENAI API key to use the query via API. Go to the OpenAI quickstart page and choose your preferred programming language. Clone the repository, it comes with a virtual environment. In my case I used the python quickstart. Before doing anything else, we want to install the requirements using pip install -r requirements.txt Next we create a .env file from the .env.example file to contain your environmental variables, including your API key. Go into the virtual environment, and run the 'flask run' command to launch the app in its default mode. The preview app is a pet name generator, which should run on your local machine now. You can modify the prompt making it easy to get hands-on experience with prompt engineering and GPT3 app building. To test changes incrementally, copy the contents of app.py which is where the code and prompt really is into file into a Jupyter notebook, and make changes to it incrementally. Now that we understand how it works, in our code we read our already cleaned database. I Started by formatting the data into a pandas dataframe from a couple of CSV files that measure several metrics. The most important ingredient, apart from the architecture, is the prompt.

4 THE PROMPT

The prompt aims to be general and clear but specific enough in some instances.

The database in described in general terms, providing some associative information. For example, I mentioned this is a football database from FBREF, to ensure that the prompt can guess what the columns mean from the abbreviations, even if it has not been trained on the data. Apart from explaining the shape and type of data, I created the prompt template of the Question and Answer. Simple question in English and the answer replied in Python code that would deliver the answer. There are many things to consider when making prompts brevity

should be one of them as it saves cost. The approach would be to aim for clarity first, and then remove words one by one to test the prompt's breaking point.

5 Findings

The practical implementation of this natural language interface has exceeded my initial expectations in terms of its accuracy and ability to understand not only the literal meaning of queries, but also the more nuanced and colloquial aspects of language. In fact, the system has shown impressive capabilities in being able to correctly identify slang terms and nicknames used in the context of sports, such as "Red Devils" being the informal name of the Manchester United football club, or "Naija" being the informal name for Nigerians . It has an high order of complexity management also, for exampe it returns the right answer to " who are the top 10 highest tackling forwards who never played in midfield or defense this season". It carefully filters out what is needed. This level of sophistication and flexibility in understanding human language makes the interface more accessible and intuitive for users, who can interact with databases using the same language they would use in conversation, without having to learn complex query languages or programming syntax. One of the most remarkable aspects of the project is its versatility and adaptability. Initially developed with a focus on football data, it became apparent that the underlying natural language processing capabilities could be applied to virtually any numerical database. Although the project was named "Football Genie," the concept of a genie as a wish-granter can be applied to any domain, and the system is capable of interpreting and executing queries in a wide range of contexts. This flexibility and scalability make the project not only innovative but also highly practical and potentially transformative for a wide range of industries and use cases.

6 Furthermore...

To facilitate the integration of new databases into the system without requiring extensive customization of the prompt for each new user and database, we plan to implement a script that can automatically detect the data schema and update a glossary of columns. This will enable the system to quickly adapt to new datasets and ensure that the natural language processing capabilities remain accurate and effective.

OpenAI 2022 OPEN AI 2023 Together 2023 Willison 2023 Ransom 2023

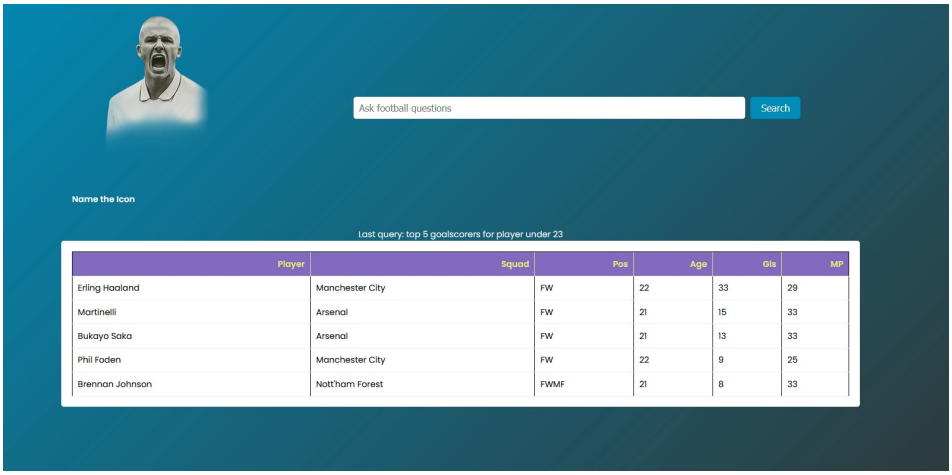

Figure 1: Caption for the image

References

OPEN AI. 2023. *Gpt-3.5 with browsing (alpha) now available for gpt plus users*. https://openai.com/blog/gpt-3-5-alpha-now-available-for-gpt-plus-users/.

OpenAI. 2022. *Chatgpt: optimizing language models for dialogue*. https://openai.com/blog/chatgpt-optimizing-language-models-for-dialogue/.

Ransom, Tyler. 2023. *DScourseS23: repository for data science course*. https://github.com/tyleransom/DScourseS23. Accessed: May 10, 2023.

Together. 2023. *Redpajama-data: an open source recipe to reproduce llama training dataset*. https://github.com/together/RedPajama-Data.

Willison, Simon. 2023. *Large language models are having their stable diffusion moment*. https://simonwillison.net/2023/Mar/11/large-language-models/.