Write up Jai Evans Capstone

Introduction

My approach was to put together a combined application to leverage key concepts from the course. I wanted to load and analyze the table of data provided.

General architecture

The solution was built using the following technologies:

- Flask
- langchain
- transformers
- pandas
- sentence-transformers
- prompt engineering
- visualizations

Flask

Flask provides the web frame work along with an html page and active java script. This followed a model-view-controller pattern, where the model was very simply a json with answer and plot. The Javascript would manipuate the view depending on a response or error. The pyhthon in the back, driven by flask, would shape the data and perfomr the generative actions.

LLM

I utilized LangChain's HuggingFacePipeline. This wrapped the Hugging FAve languange model and interacted with it.

There was transformer pipelines to create the text generating pipeline.

I also played around with OpenAI integrations.

Data load

Pandas was used to load the csv file. Then there were various parsing and grouping functions to shape the data. At first I tried to send the whole data frame to the LLM, but this was too big. However pre-shaping the data was more useful to answer questions.

Sentence-Transformers

I initially tried to build the solution without sentence-transformers but ended up getting weird or inappropriate answers in many cases. Adding sentence transformers improved the quality and consistency of the query response. I will need to study these more.

Prompt Engineering

I also discovered that careful prompt engineering dramatically improved results. There is some syntax that I discovered that helped coach the engine to provide better responses.

Visualizations with Ploty

I used the ploty library for charts. This plotting was generally pure python data manipulation. I am interested in getting generative ai to do some of this data shaping for visualization, but at present I found that to be a limited approach.

Conclusion

It was relatively easy and fast to get some basic integration set up. However, I found I had to spend many hours tuning and manipulating the code to make it more useful. The RAG aspects are very interesting. It will take more work for me to better master this technology. It is clear that data preparation is key to truly leverage this technology.