NaturalViz

Applying natural language processing and dynamic code generation for Data Visualization and Exploration

Inspiration

Our inspiration for NaturalViz stems from the recognition of challenges in traditional data visualization tools. We observed barriers in accessibility and the reliance on specialized knowledge, hindering the broader audience from exploring and understanding datasets effectively. Inspired by the potential of language models, particularly Mixtral-8x7B-Instruct, we envisioned a solution that could make data exploration more inclusive, user-friendly, and dynamic. The goal was to create a tool that goes beyond conventional data visualization methods, leveraging natural language processing to enable users to interact with data in a conversational manner.

We believe there is a fundamental flaw in the current practice of having an analyst or middleman present data and findings to an audience. As professor Catherine D'Ignazio has mentioned in their paper "Data Visualization and the Politics of Illusion", data visualizations are not neutral. They are created by people who make choices about what data to include, how to represent it, and what stories to tell. Instead, the audience should be able to interact directly with the data. The current barrier to this interaction is the audience's potential lack of technical knowledge. However, if we can provide a natural language interface using Large Language Models (that can understand human input, generate dynamic code, and create visualizations on the fly, suddenly all of this becomes possible.

What it does

NaturalViz is an experimental solution that combines the power of Mixtral-8x7B-Instruct with data visualization capabilities to redefine how users explore and gain insights from datasets. The chatbot functionality allows users to pose natural language queries about the dataset, and, in response, the system generates Python code for analysis and visualization. This enables dynamic and infinite visualizations, empowering users to explore data from various perspectives effortlessly. NaturalViz promotes transparency by including the generated code in the analysis, making it accessible for users and judges to assess the technical proficiency and creativity involved.

How we built it

The development of NaturalViz involved a multi-faceted approach. We integrated Mixtral-8x7B-Instruct model with langchain , to process user queries and generate code snippets. The frontend utilizes Streamlit for creating an interactive and user-friendly interface. Matplotlib and Altair were employed for dynamic visualization, offering a diverse range of plots and charts. The project embraces an experimental mindset, refraining from extensive data cleaning to showcase the capabilities of language models in handling raw data.

The chatbot works like this:

It contains an agent wrapper with a list of tools. Main agent contains the given template,

```
You are a helpful assistant that can help users explore a dataset.
First 3 rows of the dataset:
{dataset_first_3_rows}
===="""
"""
TOOLS:
------
You has access to the following tools:

{tools}

To use a tool, please use the following format:

Thought: Do I need to use a tool? Yes
Action: the action to take, should be one of [{tool_names}]
Action Input: the input to the action
Observation: the result of the action

When you have a response to say to the Human, or if you do not need to use a tool, you MUST use
the format:

Thought: Do I need to use a tool? No
Final Answer: [your response here]

Begin!

New input: {input}
{agent_scratchpad}
```

Prompts

Direct NLP to Visualization

```
GENERATE_PLOT_TEMPLATE_PREFIX = """You are an expert in data visualization who can create suitable visualizations to find the required information. You have access to a dataset (dataset.csv) and you are given a question. Generate a python code with st.altair_chart to find the answer.
First 3 rows of the dataset:"""

DATASET = f"{dataset_first_3_rows}"

GENERATE_PLOT_TEMPLATE_SUFIX = """
Question:
{question}

Example for protein count of different products:
import altair as alt
import pandas as pd
import streamlit as st

# Read the dataset
df = pd.read_csv('dataset.csv')

# Calculate the protein count of different products
product_protein = df.groupby('name')['protein'].sum().reset_index()

# Create the chart
chart = alt.Chart(product_protein).mark_bar().encode(
            x=alt.X('name:N', title='Product Name'),
            y=alt.Y('protein:Q', title='Protein Count')
)

# Display the chart
st.altair_chart(chart, use_container_width=True)

Generated Python Code:
"""
```

```
RETRY_TEMPLATE_PREFIX = """Current code attempts to create a visualization of dataset.csv to meet the objective. but it has encountered the given error. provide a corrected code. if you are adding comments or explanations they should start with #.

Example:
import altair as alt
import pandas as pd
import streamlit as st

# Read the dataset
df = pd.read_csv('dataset.csv')

# Calculate the total social media followers for each region
region_followers = df.groupby('Region of Focus')[['X (Twitter) Follower #', 'Facebook Follower #', 'Instagram Follower #', 'Threads Follower #', 'YouTube Subscriber #', 'TikTok Subscriber #']].sum().reset_index()

# Melt the dataframe to convert it into long format
region_followers = region_followers.melt(id_vars='Region of Focus', var_name='Social Media', value_name='Total Followers')

# Create the chart
chart = alt.Chart(region_followers).mark_bar().encode(
            x=alt.X('Region of Focus:N', title='Region of Focus'),
            y=alt.Y('Total Followers:Q', title='Total Followers'),
            color=alt.Color('Social Media:N', title='Social Media')
)

# Display the chart
st.altair_chart(chart, use_container_width=True)

First 3 rows of the dataset:"""
            DATASET = f"{dataset_first_3_rows}"


            RETRY_TEMPLATE_SUFIX = """
Objective: {question}

Current Code:
{error_code}

Error Message:
{error_message}

Corrected Code:
"""
```

What we learned

The development of NaturalViz has been a learning journey. The project underscored the potential of language models in enhancing user interaction with data. The dynamic nature of the chatbot and its adaptability to various queries emphasized the importance of user-centric design. The project also highlighted the possibilities and challenges of minimal human interference in the data exploration process, relying on the capabilities of language models for insights generation.

What's next for NaturalViz

Looking ahead, the NaturalViz team envisions further refinement and expansion of the chatbot's capabilities. This includes exploring additional language models and incorporating user feedback to enhance the chatbot's understanding and response accuracy. The project aims to establish partnerships for real-world applications, potentially integrating NaturalViz into existing data analysis workflows. Continuous experimentation and improvement remain central to the project's future, with a commitment to shaping the next generation of data exploration tools.