Data Collection and Prompt Engineering

Goal: In this study, EVs are considered as vehicles that use electric motors for propulsion and include all types of EVs. In order to save time and automate the collection of information due to the adequate amount of consultancy farms and EV producer companies, several web crawlers have to be designed to collect vehicle information. Collection of monthly or quarterly sales data, production data or inventory data from EV producer companies to align with our sentiments of end users or any other promises that have been achieved. Used prompt engineering with ChatGPT to predict the sentiments from a statement of end-users and capture the past,present future goals/achievements/promises.

Approach for data extraction:

When extracting data from consultancy firms and EV producer companies, the first step is to identify the specific websites or web pages that contain the desired data. Once the target websites are identified, Python can be used to initiate a connection to the websites and retrieve the HTML content.

After obtaining the HTML content, employed data scraping libraries like BeautifulSoup, Selenium to parse the HTML and extract the required data elements. These libraries provide functions and methods to navigate through the HTML structure, locate specific elements based on their tags, classes, or other attributes, and extract the associated data by their tags. Extracting the tags will help us in summarization or keyword extraction for E.g. Some paragraphs under a <p> tag will have a <h2> or <h3> tag associated with it as a heading. Removed unnecessary content from each article/page checking similar content across all the pages. Scraped the data from each page and kept the date,source(if applicable) separate for each article for faster access. Collected share price of tatamotors from https://www.nseindia.com/.

Work with prompt engineering:

1. To extract the sentiments of the users data scraped from social media websites.
2. Extract the summary if there are any promises in the future or any goals achieved in present or past containing the numbers associated with those promises or achievements. (Couldn't experiment more due to billing issues of the api usage).

Achieved:

1. Scraped whole data regarding EV from 2 consultancy firms with various keywords like 'EV','electric vehicle','green vehicle','no emission','no emission automobile', 'net zero' etc.
2. Scraped EV related data from the tata motors website.
3. Scraped some twitter for capturing end-users data. (But after the new billing charges with twitter api, we collected some manual twitter data.)

These datasets are being used with state-of-the-art language models with encoder-decoder based transformer architecture for sentiment extraction, keyword extraction, alignment checking, text summarization, keyword ranking, alignment checking for the prediction with original text (evaluating the language models performance).