GUVI

Data Harvest Maestro

PROJECT TITLE : Myntra T-Shirt Image Extraction Application

A.MOWLIDHARAN
BATCH - M19

PROBLEM STATEMENT

Image Extraction:

- Implement web scraping using tools like Selenium, BeautifulSoup, or any other suitable tools to retrieve T-shirt images from Myntra.

Data Organization:

- Structure the extracted images and associated metadata into a format suitable for analysis and further use.

User Interface Development:

- Optionally, create an intuitive and user-friendly interface using Streamlit/Flask/Django for users to interact with and explore the extracted images.

TOOL USED

- **Web Scraping:**
  - Selenium
- **Data Visualization:**
  - Plotly
- **Data Organization:**
  - Pandas (Organization)
  - Lance DB (Storage and Vector Search for Structured and Unstructured Data)

APPROACHES

Web Scraping:

- Images and their attributes are collected from the Myntra website using Selenium, which automates web browsers.

Data Visualization:

- Bar chart is created with Plotly, which is a library for interactive graphs, to show the brand and discount percentage distribution on the Myntra website.

Data Organization:

- Pandas, which is a tool for data analysis and manipulation, is used to manage and process the extracted data.
- Lance DB, which is a database that can store and retrieve both structured and unstructured data, is used to provide images based on queries using vector search.

IDEAS

- Images and their details from 8 product categories on Myntra are scraped using Selenium, and features such as category, brand, price, discount, material, and description are obtained for 2000+ records.
- A bar chart of brands and discounts on Myntra is created using Plotly, which is color-coded and filterable by brand category.
- Lance DB stores the data as vectors in various formats using Pandas dataframes. The data is embedded with the open-clip model, a method that aligns image and text representations. It shows an image that matches the product description query, based on the cosine similarity score, a measure of the angle between two vectors.