IdealRatings

Senior AI assessment task

Extract Specific Categorical and Numerical Information from Scraped Book Data

Objective

Develop a solution to extract specific **categorical** and **numerical** information from scraped book data obtained from https://books.toscrape.com. The extracted data should be organized in a structured format, and the solution should be capable of handling the diversity of books, categories, and content on the site.

Specific Questions to Address

Categorical Questions (Yes/No)

1. Are there any books in the "Travel" category that are marked as "Out of stock"?
2. Does the "Mystery" category contain books with a 5-star rating?
3. Are there books in the "Classics" category priced below £10?
4. Are more than 50% of books in the "Mystery" category priced above £20?

Numerical Data to Extract

1. What is the average price of books across each category?
2. What is the price range (minimum and maximum) for books in the "Historical Fiction" category?
3. How many books are available in stock across the four categories?
4. What is the total value (sum of prices) of all books in the "Travel" category?

Hybrid Questions (Categorical + Numerical)

1. Which category has the highest average price of books?
2. Which categories have more than 50% of their books priced above £30?
3. Compare the average description length (in words) across the four categories.
4. Which category has the highest percentage of books marked as "Out of stock"?

Task Requirements

1. **Input**:
   A dataset containing scraped information about books in the **Travel**, **Mystery**, **Historical Fiction**, and **Classics** categories from https://books.toscrape.com. The dataset should include details like book titles, categories, prices, availability status, ratings, and descriptions.
2. **Output**:
   - **Categorical answers**: Yes/No answers for each of the specified questions, including justification.
   - **Numerical answers**: Extracted numerical values for average prices, price ranges, stock counts, etc., for each of the specified questions, including justification.

What to Deliver

1. **Code**:
   A script or pipeline to scrape the website, preprocess the data, and answer the specified questions. The code should be modular and efficient, capable of handling future data expansions.
2. **Documentation**:
   A detailed explanation of your approach, including:
   - How you scraped and processed the data.
   - Methods used to answer the specific Yes/No and numerical questions.
   - Key challenges faced and how you overcame them.
3. Interface:

- A simple **Streamlit** or **Gradio** interface that allows users to input questions (from the provided list or custom ones) and get answers directly.
- The interface should be deployed using **Hugging Face Spaces,** allowing easy access and interaction.

Helpful notes:

- The data may include books with varied prices, availability statuses, descriptions, and ratings. Ensure your solution is robust and capable of handling this diversity.
- Make sure the numerical answers include proper context, such as the sample size or relevant data points (e.g., book count, price range).
- Consider both rule-based and AI-driven methods to ensure the accuracy of answers and efficient retrieval of data.

Evaluation Criteria

1. **Accuracy**: How well the solution extracts the correct answers, both categorical and numerical, and how the references support those answers.
2. **Scalability**: The ability of the solution to handle large datasets or new books being added to the site.
3. **Clarity**: The readability and modularity of your code, as well as the quality of your documentation.
4. **Efficiency**: The speed and computational efficiency of your solution.
5. **Innovation**: Creative and effective approaches to solving the task, especially for handling edge cases or ambiguous data.