Final Project Proposal

**Data Source**

In this final project, I plan to analyze data to help users decide which clothing item is most suitable for them. I intend to crawl data from Amazon on clothing products and collect product names, prices, ratings, rating counts, and product URLs as my raw data for analysis.

**Tree or Network**

When searching and filtering products, I use two different data structures to help organize and process data. Firstly, I store each product as a dictionary, where each key represents a different product attribute, such as "product name", "price", "rating", "rating count", and so on. These dictionaries are organized into a list, where each dictionary represents a product. Secondly, I use a tree-based data structure to store and organize all products by category, gender, and subcategory. I use the treelib library to construct and store this tree structure, which allows us to quickly find products that match user inputs.

When determining user preferences, I use another data structure: tuples. I store the price, rating, and rating count information for each product as a tuple and organize these tuples into a list so we can sort and compare products to determine the best match for the user's preferences. The use of these data structures helps us better organize and process data, allowing our program to quickly find products that meet user requirements.

**Data Processing**

After crawling data from Amazon and writing it to a CSV file, I will group the data into categories based on clothing types such as T-shirts, shirts, and down jackets. Each product will have its own attributes such as price, rating, rating count, product URL, and available colors.

**Data Presentation**

I will create a command-line interface to interact with users. Specifically, users can specify the type and color of clothing they are looking for, as well as their preferred price range, rating requirement, and rating count. Based on their preferences, I will recommend the most suitable clothing items to users and provide product names and URLs.

|  |  |  |
| --- | --- | --- |
| Data Source |  | Challenge Score |
| Web API you haven’t uesd before that requires API key or HTTP Basic authorization | Amazon | 6 |
| CSV or JSON file you haven’t used before with > 1000 records |  | 2 |
| Scraping a new single page |  | 4 |