CS6200 Project Proposal

Yuting Sun

Background

It is always a challenging task to try out a new cosmetic product. Different people have different skin types and product preference. A product that is good for one person may not be suitable for another. Improper products would waste a lot of money and even cause severe skin problems. Since the COVID-19 pandemic, it's not convinient for people to go to the stores and try the products on themselves. Therefore, it would be helpful if we could build a searching tool for cosmetic products.

Ingredients are key to the property and effectiveness of a cosmetic product. For example, salicylic acid is widely used in treating acnes, titanium dioxide is used for UV absorbance in sunscreen products. Ingredient analysis can be very helpful for people to search for proper products. Some people likes the certain propety of a product, so s/he would propably like a product with similar ingredients. Some people are allergic to a certain ingredient, so it would be necessary to avoid products with the same ingredient. However, most cosmetic products contain more than 30 ingredients, which is difficult for people to read and analysis. It would be a good idea to combine information retrieval techniques with ingredient analysis. This would help customers find out new product that are most suitable for them, as well as getting more insights on cosmetic ingredients.

Solution

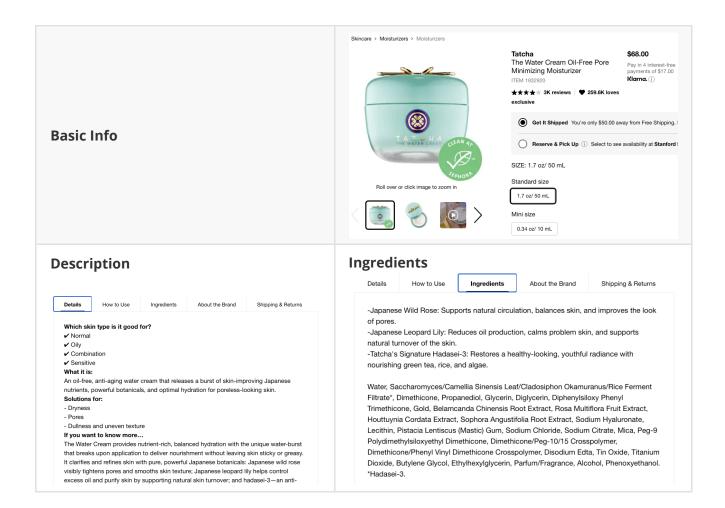
In this project, I would like to propose a searching tool for cosmetic products that contains two parts:

- 1) Search for relevant products based on product type, name, and description.
 - The category, brand, product name and description are indexed. Relevant results are can be sorted via multiple criteria such as best-selling, price, and ratings.
- 2) Recommend similar product based on ingredient analysis.
 - Ingredients of each product is vectorized. Similarity is calculated via cosine score.

Dataset

The Sephora dataset can be obtained from <u>Kaggle</u>. This dataset contains 9000+ products with product information, detailed description and ratings. More than 7500 products have ingredient information.

Product Information Example:



Future Improvement

The current dataset does not contain review text. The average number of reviews for each product is 282, which makes the total number of reviews to be \sim 2.7 million. We can parse the review text to conduct better search for relevant product.