**Project Report: Cosmetic Ingredient Analysis and Recommendation System**

**Introduction**

Choosing the right cosmetic products can be daunting due to complex ingredient lists and the potential for adverse skin reactions. To address this challenge, we developed a **content-based recommendation system** that analyzes the chemical components of cosmetics. This project focuses on **moisturizers for dry skin** to predict and recommend suitable products using **data science techniques** such as word embedding, t-SNE for dimensionality reduction, and interactive visualization with Bokeh.

**Objective**

The primary goal is to:

1. Create a content-based recommendation system for cosmetic products.
2. Analyze and visualize ingredient similarity between products.
3. Enable users to explore cosmetic options interactively.

**Dataset Description**

The dataset contains information about 1472 cosmetic products listed on Sephora, with the following key attributes:

* **Label:** Product category (e.g., moisturizers, cleansers, face masks).
* **Skin Types:** Suitability for skin types (e.g., dry, oily, combination).
* **Ingredients:** List of chemical ingredients for each product.
* **Brand, Price, Rank:** Additional metadata.

For this project, the data was filtered to include **moisturizers for dry skin** only.

**Methodology**

**1. Data Preparation**

* **Filtering Data:** Focused on moisturizers suitable for dry skin.
* **Tokenizing Ingredients:** Split ingredient lists into tokens and created a dictionary mapping each ingredient to a unique index.
* **Bag of Words Encoding:** Created a **Cosmetic-Ingredient matrix** where each row represents a product, and each column represents an ingredient. A binary value (1 or 0) indicates the presence or absence of an ingredient.

**2. Dimensionality Reduction**

* The resulting matrix had high dimensionality (190 products × 2233 ingredients).
* Used **t-SNE (t-distributed Stochastic Neighbor Embedding)** to reduce the dimensions to 2, preserving ingredient similarity for visualization.

**3. Visualization**

* Plotted the t-SNE features using **Bokeh** to create an interactive scatter plot.
* Added a hover tool to display details about each product, such as:
  + Product Name
  + Brand
  + Price
  + Rank (Rating)

**Results**

**1. Ingredient Analysis**

The **Cosmetic-Ingredient matrix** allowed us to identify:

* Frequently occurring ingredients across moisturizers for dry skin.
* Unique ingredients contributing to the similarity or distinction between products.

**2. Visualization**

The **interactive scatter plot** provides:

* A visual representation of ingredient similarity among moisturizers.
* The ability to explore individual products by hovering over points to view details (name, brand, price, and rank).

**3. Insights**

* Products with similar ingredient compositions cluster together on the t-SNE plot.
* Distant points indicate products with distinct ingredient lists, potentially catering to niche requirements.

**Tools and Technologies**

* **Python Libraries:**
  + **Pandas & NumPy:** For data manipulation and matrix operations.
  + **Scikit-learn:** For t-SNE dimensionality reduction.
  + **Bokeh:** For interactive data visualization.
* **Techniques:**
  + Tokenization and One-Hot Encoding.
  + Dimensionality reduction using t-SNE.

**Limitations**

1. **Interpretability of Axes:** The axes in the t-SNE plot lack direct interpretability as t-SNE is primarily a visualization tool.
2. **Ingredient Information:** Some ingredient names may not provide enough context for laypersons to judge their effects.

**Future Work**

1. **User Personalization:** Expand the recommendation system to include multiple product categories and skin types.
2. **Ingredient Effects:** Incorporate external data about ingredient properties to evaluate product effectiveness.
3. **Sentiment Analysis:** Use customer reviews to enhance the recommendation accuracy.

**Conclusion**

This project demonstrates the power of data science in simplifying the cosmetic selection process. By visualizing ingredient similarities, users can make informed decisions and identify products suited to their skin type. The interactive visualization provides an engaging and user-friendly interface to explore cosmetic options.

