# Cosmetics Recommendation System

Shrisha Hedge, Soham Chandrakant Kasar, Srikanth Nimmagadda, Sai Sri Harshini Kosuri

San Jose State University

***Abstract*— Consumer interest in cosmetics has risen globally in recent years, with a focus on skincare. Consumers have previously relied on best-seller products or in-store counter recommendations. However, because everyone's skin is different, these are ineffective ways of determining whether a product and a person are compatible. The goal of this proposal is to create a skincare product recommendation system based on the user's skin type and the product's ingredient composition. Product chemical components are identified and products with similar constituent compositions are found using content-based filtering.**

## I. Introduction

With an annual increase in sales, skincare has surpassed makeup as the "fastest-growing category internationally". According to Trefis, a financial research and analysis business, skincare product sales in the United States increased by 13% in 2018, whereas makeup sales only increased by 1%. Trefis also expects the worldwide skincare market to reach $180 billion in the next five years, up more than 30% from where it is now. Customers' desire for natural beauty, as well as men's rising interest in skincare products, are driving this expansion. Anti-aging products are also being added by companies in order to keep women as their key customers as they age.

More customers began visiting the cosmetics counter for product recommendations, necessitating the need for improved technologies. This method, however, is frequently ineffectual and time-consuming. The overwhelming amount of information available on the internet has also made it difficult for consumers to make informed decisions. The quantity of product information and feedback is seen as beneficial. However, it prohibits consumers from selecting appropriate information and making decisions based on their requirements. As a result of these difficulties, there is a pressing demand for customized solutions that can make data access easier.

Researchers have proposed different recommender systems to resolve the information overloading problem and facilitate the selection process. The two most adopted methods are collaborative filtering and content-based filtering. Recently, a hybrid approach that combines the two techniques was introduced to maximize the benefits of both methods while covering their weaknesses.

There could be some people who share a very similar taste for cosmetics. And with user-user collaborative filtering, we can recommend new products based on the ranking values of this neighboring group. However, the skin type and feature of a person is a more sensitive and tricky problem than just recommending your tonight's movie show. To get the reliability and stability in the recommendation, we need to focus on the real content of each product, or the ingredients of products, and get the similarities based on them. Although a hybrid approach seems to have potential in the skincare domain, it requires a data set that involves both the behavioral information of the user as well as the product information. Such data set, however, is scarce in skincare.

We are going to implementing the content based recommendation system and TF_IDF recommendations system, compare the results and concluded.

### *A. Dataset*

The data was scraped from sephora.com. This website offers beauty products from multiple brands. Among the six were extracted to focus on skincare products. These categories include moisturizing creams, facial treatments, cleansers, facial masks, eye treatments, and sun protection. Initially, we scrapped the product links from the search page for each category, and then information about each product is extracted from the links gathered. The data set includes information about the brand, name, price, rank, skin types, and chemical components of each product. The extraction is done using the library 'Selenium' that allows data mining from different websites. The data was extracted using the XPath of the element on the HTML page. This data set will be used specifically to evaluate the efficiency of this method after the implementation of the content-based recommender system.

| **Feature name** | **Type** | **Description** |
|---|---|---|
| Label | Categorical | Type of skincare product |
| Name | Categorical | Name of the product |
| URL | Categorical | The Sephora link for the product |
| Brand | Categorical | The brand name of the product |
| Rank | Numerical | Ratings of the product |
| Skin Type | Categorical | Favorable skin type for the product |

|  |  |  |
|---|---|---|
| Ingredients | Categorical | Ingredients in the product. |

### B. *Exploratory Data Analysis*

#### 1. Data Preprocessing

Data preprocessing is a data mining technique that is used to transform the raw data into a useful and efficient format.

**Steps Involved in Data Preprocessing:**

#### (A) Data Cleaning

The raw data may contain a variety of meaningless and missing items. To deal with this, data cleaning is used. It requires coping with data that is absent or noisy.

1. Remove duplicate and unwanted data: Removed unwanted columns like URLs from the dataset.
2. Structural errors must be corrected: Structural error columns Label, ingredients, skin_type, price, and rating were fixed.
3. Handling missing values: We decided to handle the missing data on the bases of columns in which the data is missing. In columns like ingredients, we have dropped the row. And in columns like Skin_type, we replaced it with mode.

#### 2. Feature Scaling

This step is conducted to convert the data into a format that can be used in the mining process. In order to use the skin type data, we apply one-hot encoding on that column.

#### 3. Data Visualizations

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

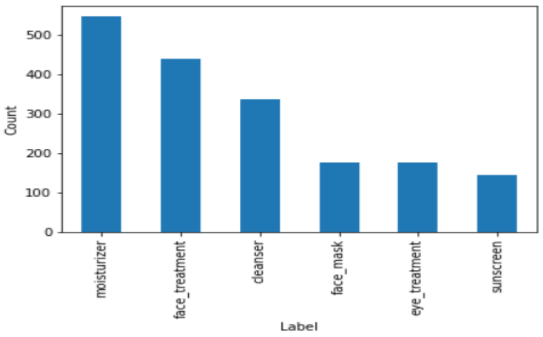


Fig 1: Distribution of products

***Observation:*** *Fig 1 shows the distribution of products in different categories and it seems that moisturizer has the highest number of products whereas sunscreen holds the least number. This may influence our final prediction.*

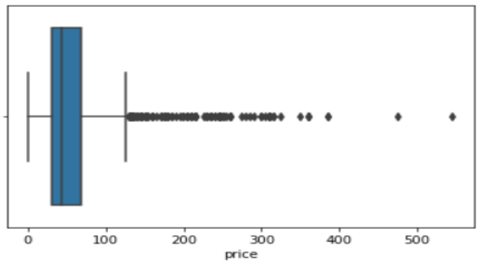


Fig 2: Price distribution

***Observation:*** *In Fig 2 Most of the products lie in the range of 0 to $120 but for some premium products we can see that the price exceeds $500. Usually, the outliers are dropped or imputed but in this case, as the model is based on ingredients we can't impute the price column.*

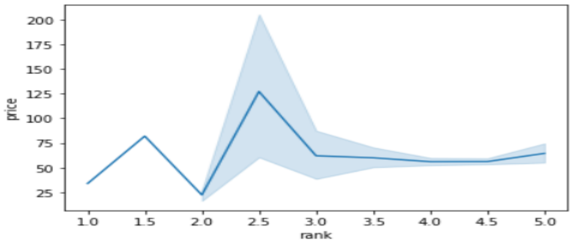


Fig 3: Rank Vs Price

***Observation:*** *In Fig 3 the graph depicts that the higher price of the product does not imply a better rank or rating. Variability in the rating of products is high from rank 2 to 3.*

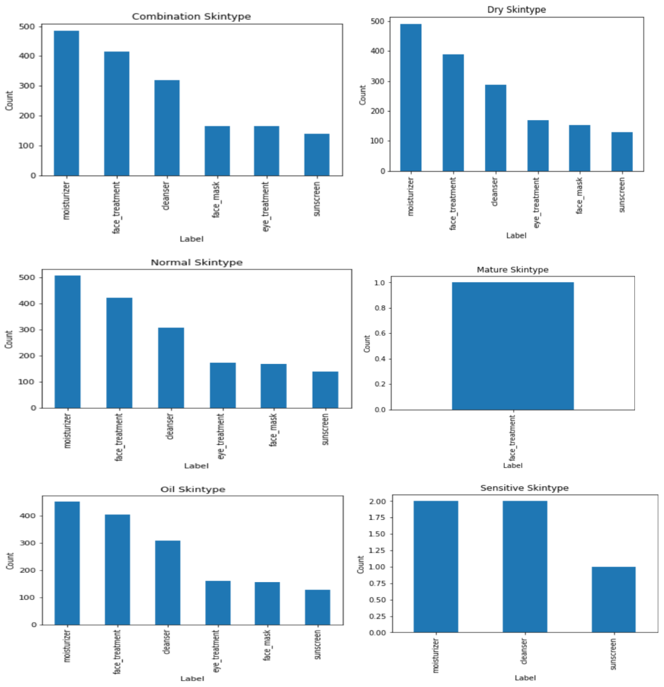
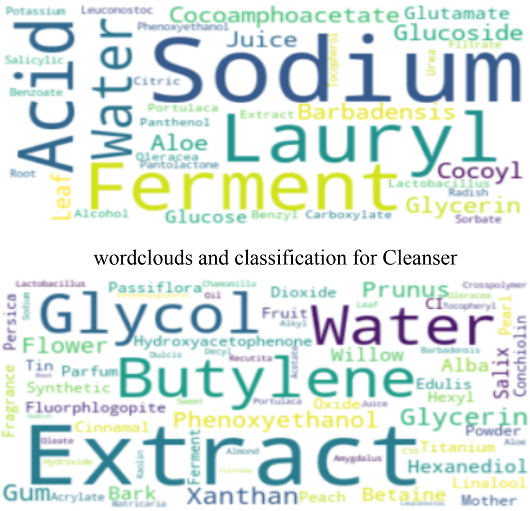


Fig 4: Product distribution based on Skintype

***Observation:*** *In fig 4 shows the distribution of products for each skin type. we can observe that there are negligible amounts of products for Mature and Sensitive skin types.*



wordclouds and classification for Cleanser

wordclouds and classification for Mask

Fig 5: wordclouds and classification

***Observation:*** *Fig 5 depicts that cleanser mainly contains Sodium, Lauryl, Acid, Water, and Ferment as their main ingredients.And Mask mainly contains Extract, Butylene, Glycol, and Water.*

### 4. Ingredient Extraction

The list of all ingredients is extracted from the data set's ingredients column and divided into tokens. After the duplication has been checked, each chemical element is assigned a unique index that will be stored in a dictionary.

The document term matrix (DTM) is then created between the products and their corresponding ingredients. An empty matrix is created and then filled with zeros. The number of rows here represents the number of skincare products, while the number of columns represents the total number of ingredients. Then, depending on the presence of ingredients in each product, one-hot encoding is used to fill in the cosmetic-ingredient matrix with either 1 (present) or 0 (not present).

### 5. Dimensionality Reduction

Dealing with very high-dimensional data is a common issue for data scientists nowadays (lots of features). Most algorithms for classification and prediction that work well with low-dimensional datasets don't work as well with many features, a problem known as the curse of dimensionality. To address this, we can reduce the number of dimensions through feature selection or feature extraction, typically by handpicking the most relevant features or using Principal Component Analysis (PCA) prior to feeding the data into our preferred model. Even though PCA is excellent in most cases, it is still a linear model that may be insufficient for some datasets. So, we have used t-SNE to reduce the dimensionality of the dataset.

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a very popular and state-of-the-art dimensionality reduction technique that is usually used to map high dimensional data to 2 or 3 dimensions in order to visualize it. It does so by computing affinities between points and trying to maintain these affinities in the new, low-dimensional space.

Using t-SNE on the cosmetic-ingredient matrix we have reduced the dimensionality of the cosmetic-ingredient matrix from 452 to 2.

## II. METHODS

### 1. Content-based Filtering:

The chemicals, together with the user's skin type, are submitted into the recommender system after they have been extracted and processed. This method compares the similarity of item constituent composition using cosine similarity. It's used to rank cosmetics with similar qualities to the original product by generating recommendations for different product categories.

$$Cos\theta = \frac{\vec{a}\cdot\vec{b}}{||\vec{a}||\;||\vec{b}||} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2}\;\sqrt{\sum_1^n b_i^2}}$$

Using the resulting matrix, all cosmetic elements are vectorized into two-dimensional coordinates using t-SNE. The cosine similarity equation is used to calculate the distances between various locations using these coordinates. Finally, these values are ranked from most similar to least similar in ascending order. The method is repeated to filter the data by adding more product categories. The system can recommend products from a variety of categories by categorizing the data.

### 2. TF-IDF Filtering:

There is also a different technique that can be used to recommend skincare products based on the product entered. The top recommendations can be identified by calculating the TF-IDF values. The values are derived using the below equations.

$$TF\,(i,\,X) = \sum_{p=1}^{m} \frac{n_p - \alpha_{p,i}}{n_p}$$

*np*: the number of unique ingredients included in product *p* in
beauty effect group *X*
*m*: the number of products in beauty effect group *X*
*αp*,*i*: the rank of ingredient *i* listed in product *p*

$$IDF\,(i) = log\frac{N}{pf\,(i)}$$

*N*: the number of products in the data set
*pf* (*i*): the number of products including ingredient *i*

$$TF\_IDF(i, X) = TF(i, X) \times IDF(i)$$

The similarity of the products represented as TF-IDF vectors is computed using the linear kernel. The products that include such ingredients are then filtered. Finally, the top recommendations are returned to the user.

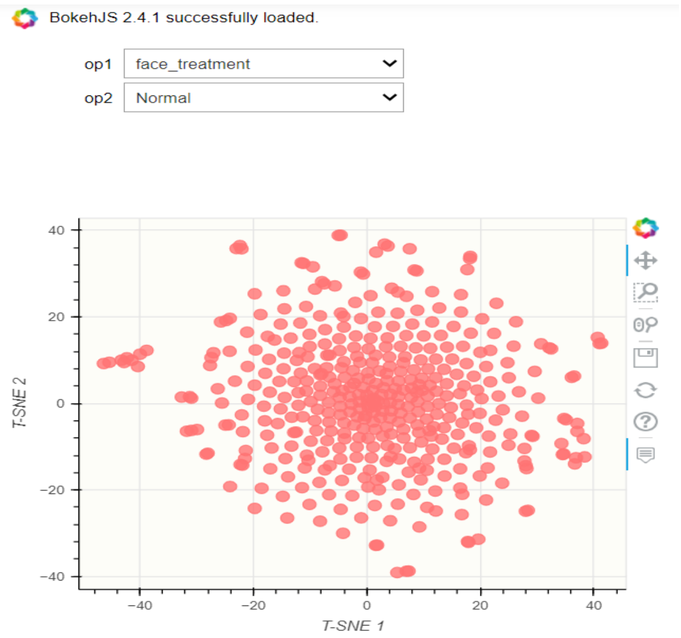
## III. COMPARISONS

The data that was gathered using web scraping consisted of information on ingredients and product type, no information was available about users or customers of the product. So quantifying recommender systems is not possible this time. However, we have manually compared the ingredients of the input product with the recommendations generated and found quite a lot of similarities and found that the TF IDF method of recommendation was slightly better.

There are surveys usually conducted to test this kind of data [1]. In this paper, the user liking is compared with the recommendations and it was found that the TF IDF method was slightly better than the Content-Based recommendation system.

## IV. Conclusion and Results

This report talks about the implementation of Content-Based recommendations and TF IDF-based recommendations. The bokeh plot in figure 6 shows the plot face_treatment with normal skin type. Each data point in the plot indicates the item and items with similar ingredients appear closer and if we hover over an item tooltip shows the name, brand, cost, and rating. By this, we can see what items are similar.

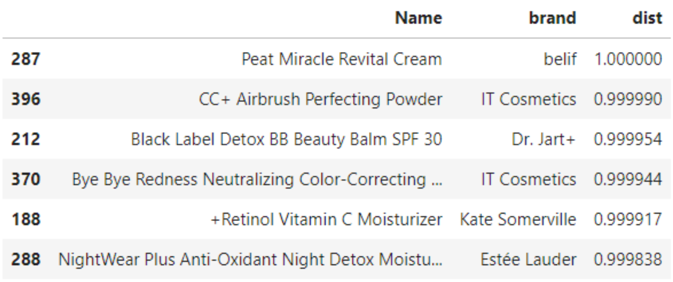


BokehJS 2.4.1 successfully loaded.

op1 face_treatment

op2 Normal

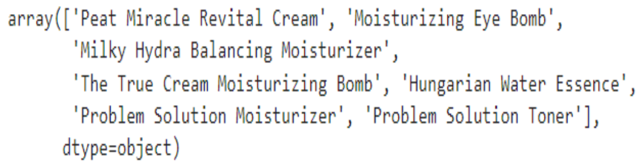
Fig 6: Virtualization using Bokeh



| | Name | brand | dist |
|---|---|---|---|
| 287 | Peat Miracle Revital Cream | belif | 1.000000 |
| 396 | CC+ Airbrush Perfecting Powder | IT Cosmetics | 0.999990 |
| 212 | Black Label Detox BB Beauty Balm SPF 30 | Dr. Jart+ | 0.999954 |
| 370 | Bye Bye Redness Neutralizing Color-Correcting ... | IT Cosmetics | 0.999944 |
| 188 | +Retinol Vitamin C Moisturizer | Kate Somerville | 0.999917 |
| 288 | NightWear Plus Anti-Oxidant Night Detox Moistu... | Estée Lauder | 0.999838 |

Fig 7: Result of Content-Based Recommendation

In figure 7 we see the results of the content-based recommendation system. The input is “Peat Miracle Revital Cream” top 5 items were chosen. The first one is the exact same because the cosine similarity to itself is 1.



```
array(['Peat Miracle Revital Cream', 'Moisturizing Eye Bomb',
       'Milky Hydra Balancing Moisturizer',
       'The True Cream Moisturizing Bomb', 'Hungarian Water Essence',
       'Problem Solution Moisturizer', 'Problem Solution Toner'],
      dtype=object)
```

Fig 8: Result of TF_IDF recommendation

In figure 8 we see the results for the TF IDF recommendation system. The ingredients of recommended items were manually compared with the item selected and it was found that TF IDF was slightly better.

## IV. Limitations and Future Research

Lack of access to customer information or any survey-based user liking was the main challenge. If this data were available then we can quantify the recommendation systems that are implemented. Once the user data is available we can also implement collaborative filtering and compare the recommendation among all the three recommendation systems.

## V. REFERENCES

**[1] Gyeongeun Lee. 2019. A Content-based Skincare Product Recommendation System https://portfolios.cs.earlham.edu/wp-content/uploads/2020/05/Gyeongeun_Lee_Paper.pdf**

**[2] Shlomo Berkovsky and Jill Freyne. 2015. Web Personalization and Recommender Systems. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Aug. 2015). https://doi.org/10.1145/2783258. 2789995**

**[3] Ahiza Garcia. 2019. The skincare industry is booming, fueled by informed consumers and social media. https://www.cnn.com/2019/05/10/business/skincareindustry-trends-beauty-social-media/index.html**

**[4 ]Linda Hansson. 2015. Product Recommendations in E-commerce Systems using Content-based Clustering and Collaborative Filtering. Master’s thesis. Lund University, Lund, Sweden**