

**Cosmetics Recommendation System- A personalized recommender  
system for the cosmetic business**

**A PROJECT REPORT**

*Submitted by*

**SRI GOPIKA. S**

**CB.SC.I5DAS18036**

*In partial fulfillment of the requirements for the award of the degree of*

**Integrated. MASTER OF SCIENCE**

**IN**

**DATA SCIENCE**



**AMRITA SCHOOL OF ENGINEERING**

**AMRITA VISHWA VIDYAPEETHAM**

**COIMBATORE – 641 112 (INDIA)**

**JUNE – 2022**

**Cosmetics Recommendation System- A personalized recommender  
system for the cosmetic business**

**A PROJECT REPORT**

*Submitted by*

**SRI GOPIKA. S**

**CB.SC.I5DAS18036**

*In partial fulfillment of the requirements for the award of the degree of*

**Integrated. MASTER OF SCIENCE**

**IN**

**DATA SCIENCE**



**AMRITA SCHOOL OF ENGINEERING**

**AMRITA VISHWA VIDYAPEETHAM**

**COIMBATORE – 641 112 (INDIA)**

**JUNE – 2022**

**AMRITA VISHWA VIDYAPEETHAM**  
**AMRITA SCHOOL OF ENGINEERING, COIMBATORE, 641112**



**BONAFIDE CERTIFICATE**

This is to certify that the project report entitled “**Cosmetics Recommendation System - A personalized recommender system for the cosmetic business**” submitted by **Ms. SRI GOPIKA. S (Reg. No: CB.SC.I5DAS18036)** in partial fulfilment of the requirements for the award of the **Degree of Integrated Master of Science in Data Science** is a Bonafide record of the work carried out.

**Signature of the Project Coordinator**

**Signature of the Chairperson**

**The project was evaluated by us on:**

**Internal Examiner**

**AMRITA SCHOOL OF ENGINEERING**  
**AMRITA VISHWA VIDYAPEETHAM**  
**COIMBATORE - 641 112**  
**DEPARTMENT OF MATHEMATICS**

**DECLARATION**

I, Ms. Sri Gopika. S (Reg.No: CB.SC.I5DAS18036), hereby declare that this dissertation entitled Data Science job market Segmentation, is the record of the original work done by me. To the best of knowledge this work has not formed the basis for the award of any degree/diploma/associateship/fellowship/or a similar award to any candidate in any University.

**Place: Coimbatore**

**Signature of the Student**

**Date: 06/06/2022**

**COUNTERSIGNED**

**Dr. Prakash P**

**Class Advisor**

**Department of Mathematics**

## **ACKNOWLEDGEMENTS**

I would like to thank the faculty members of Department of Mathematics for allowing me to work for this Project.

Dr.PRAKASH. P, Department of Mathematics, Amrita Vishwa Vidyapeetham, Coimbatore, for constant encouragement and prudent suggestions during the course of my study and preparation of the final manuscript of this Project.

My heartfelt thanks to all my friends for their invaluable co-operation and constant inspiration during my Project work.

I owe a special debt gratitude to my revered parents for their blessings and inspirations.

Coimbatore,  
June 2022

SRI GOPIKA. S

## CONTENTS

<b>1. INTRODUCTION.....</b>	<b>8</b>
<b>2. RELATED WORK.....</b>	<b>9</b>
2.1 Collaborative Filtering.....	9
2.2 Content-based Filtering.....	9
2.3 Hybrid Approach.....	10
<b>3. WORK AND IMPLEMENTATION.....</b>	<b>10</b>
3.1 Data Collection.....	10
3.2 Tokenizing the Ingredients.....	11
3.3 DTM (document-term matrix) .....	11
3.4 Counter Function.....	11
3.5 Cosmetic- Ingredient Matrix.....	11
<b>4. DIMENSION REDUCTION WITH t-SNE.....</b>	<b>12</b>
4.1 Results.....	12
<b>5. CONCLUSION.....</b>	<b>14</b>

## BIBLIOGRAPHY

## **ABSTRACT**

In recent years, consumer interest in cosmetics has been increasing globally with a focus on skincare. In the past, consumers have depended on best-seller products or in-store recommendations from the counter. However, everyone has different skin conditions, so these are not effective methods to judge compatibility between a product and a user. This proposal focuses on designing a content-based recommendation system where the ‘content’ is the ingredients of cosmetics. Specially, I processed ingredients lists for cosmetics on Sephora via word embedding, then visualized ingredients similarity using machine learning method (called t-SNE). This method also allows users to input their desired beauty effect instead of a product name if they lack knowledge or have not found a product they like.

## 1. INTRODUCTION

Hailed as the “fastest-growing category globally”, skincare has taken over makeup with its increasingly sale each year. According to Trefis, a financial research and analysis firm, sales of skincare products in the U.S. surged by 13% in 2018, while makeup sales only grew by 1%. Trefis also estimates the global skincare market to reach \$180 billion, an increase of over 30% from the current stage, in the next five years. This growth is mainly due to the customers pursuit of natural beauty as well as men’s increased interest in skincare products. Companies are also adding anti-aging products for women to keep them as their primary consumers as they get older.

The need for advanced technology accompanied such growth as more customers started visiting the cosmetics counter to get product recommendations. However, this process is often ineffective and time-consuming. The overwhelming quality of accessible online information has also made it difficult for the users to make correct choices. The abundance of product information and reviews are perceived to be valuable. But at the same time, it prevents users from picking out desired information and making decision based on their needs. Such difficulty has evoked the pressing need for personalized systems that could ease the access of data.

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

However, it is still unclear which technique best measures the suitability of products for each customer. In fact, lots of online cosmetics stores still recommend bestsellers to customers regardless of their individual skin condition. Hence, there is a need for further investigation and improvement in the recommender systems for personal care products.

This proposal presents a content-based recommendation method that evaluates the similarity of ingredient composition within products. Instead of recommending within the same category, the new system recommends products across different categories to allow more effective recommendations. It also gives an option for the user to provide minimum input to get suggestions for skincare products.

In this project, I created a content-based recommendation system where the 'content' is the ingredients of cosmetics. Specifically, I processed ingredient lists for 1472 cosmetics on Sephora



via word embedding, then visualized ingredient similarity using a machine learning method called t-SNE.

t-SNE or t-distributed stochastic neighbor embedding is an unsupervised machine learning method to find a low-dimensional representation of the data. This representation, or embedding is then used to find similar products

## **2. RELATED WORK**

This section discusses existing recommendation methods, including collaborative filtering, content-based filtering, and hybrid approach that mixes both. It focuses on the details of each technique, motivation, contribution, and the experimental or theoretical nature of the research. Moreover, the analysis and comparison of different algorithms used within each recommender system are presented with their results.

### **2.1 Collaborative Filtering**

Collaborative filters use the information provided by users, such as clicks, likes, purchases, etc. Although they face a cold-start problem, they work well with enough amount of behavioral data. Researchers who use collaborative filtering believe that identifying similar users can help match suitable products with the customer. Matsunami et al. and Okuda et al. both adopted user similarity calculating method and analyzed reviews of cosmetic items. They used automatic scoring and k-means clustering to extract not only ratings but also textual reviews that contain individual preference and opinion. Ye also used collaborative filtering but focused on improving the weakness of the traditional method. She attempted to alleviate the data sparsity problem and personalize the technique by making it more item-based. Matsunami et al. and Okuda et al. depended heavily on users' reviews to perform textual analyses while Ye focused more on the aspects of the items instead of users to recommend products. The new content-based recommender proposed in the later section is similar to the research of Matsunami et al. and Okuda et al. in that it incorporates ratings of users. Although it does not filter products based on ratings, it uses them to validate the results. Like in Ye's research, the system not only considers users' opinions but also items' properties by taking their ingredients and comparing them with others.

### **2.2 Content-based Filtering**

Another standard recommendation method is content-based filtering, which takes into account the descriptions of the items as well as user preferences. Content-based filters tend to have an overspecialization problem, which is if someone is buying a mouse, the system will likely miss out on recommending a mouse pad or a keyboard. Putriany et al. felt the need to personalize the skincare product recommendation as much as possible and adopted content-based filtering for their research. The system was based on the items that were rated, liked, or chosen in the past by a particular user. Patty et al., like Putriany et al., focused on targeting the user profile but included factors such as cosmetic type, skin type, usage, price, description, and pictures for

enhanced recommendation. This approach can help personalize the method by going beyond their purchase habits. Unlike Putriany et al., Sato et al. tried to grasp other users' influence while using content-based filtering. However, content-based filtering works only for active users, and it is hard to give a good recommendation if the information is not easy to categorize.

### **2.3 Hybrid Approach**

Noticing the problems in the traditional methods, companies like Netflix and Google started adopting hybrid recommender systems. Experiments on the live traffic of the website done by Google suggest that the hybrid method improves the quality of recommendation. With similar assumptions, some researchers have used hybrid filtering to maximize the benefits of both collaborative and content-based filtering. Hansson proposed a hybrid recommender for online products using k-means++. Using the data set from an online book retailer and fashion retailer, she obtained the values for precision and recall for each method tested on both data sets. She concluded that her algorithms do not have the same functionality across different data sets, and combinations of strong algorithms do not produce better results. James and Rajkumar also proposed a hybrid method and added the time sequence method for collaborative filtering. Unlike Hansson's, their research is theory-based and was not tested on an actual data set. They suggested three different directions for their algorithm: item similarity, bipartite projection, and spanning tree. Since the time sequence model learns the change in data over time, the authors concluded that it will generate higher accuracy by using static data. 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, so the proposed method only includes content-based filtering.

## **3. WORK AND IMPLEMENTATION**

*Content-based Filtering.* A user provides one of five skin types (combination, dry, normal, oily, and sensitive), select a product category from one of six types (moisturizer, cleanser, treatment, face mask, eye cream, sun protect), select brand, and select the product (cream, serum, etc...). this skin type, brand, product, and product category directly maps to the recommender system while ingredients are extracted from the product, then the skin type of the user, brand they need, type of product needed and other products provided by the selected brand will be sent to the content-based recommender system along with Sephora data, which contains the information about other products. In this method, recommendation is done by providing the top k number of option which got the similar ingredients by evaluating in all brands.

### **3.1 Data Collection**

An existing data set on cosmetics was used in this project. The data was scraped from sephora.com, a website that offers beauty products from multiple brands. Among many categories of personal care items, only six were extracted to focus on skincare products. These six categories include moisturizing cream, facial treatments, cleanser, facial mask, eye treatment, and sun protection. The data set consists of 1472 items which includes information about the

brand, name, price, rank, skin types, and chemical components of each product. Additionally, star ratings for all 1472 items will be extracted from sephora.com along with the reviewers' skin types. The extraction will be done using a tool called Scrape Storm that allows data mining from different websites. This data set will be used specifically to evaluate the efficiency of this method after the implementation of the content-based recommender system.

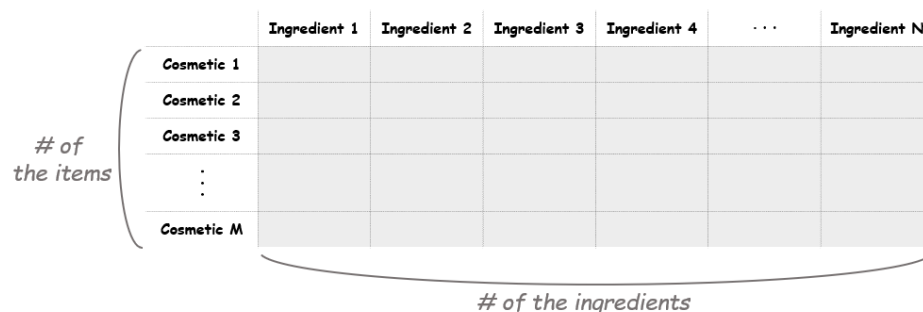
### 3.2 Tokenizing the Ingredients

To get to our end goal of comparing ingredients in each product, we first need to do some preprocessing tasks and bookkeeping of the actual words in each product's ingredients list. The first step will be tokenizing the list of ingredients in Ingredients column. After splitting them into tokens, we'll make a binary bag of words. Then we will create a dictionary with the tokens, ingredient index, which will have the following format:

```
{"ingredient": index value, ...}
```

### 3.3 DTM (document-term matrix)

Next, initialization of document-term matrix (DTM). Here each cosmetic product will correspond to a document, and each chemical composition will correspond to a term. This means we can think of the matrix as a “cosmetic-ingredient” matrix. To create this matrix, we'll first make an empty matrix filled with zeros. The length of the matrix is the total number of cosmetic products in the data. The width of the matrix is the total number of ingredients. After initializing this empty matrix, we'll fill it in the following tasks.



The diagram illustrates a Document-Term Matrix (DTM) as a grid. The rows represent individual cosmetic products, labeled 'Cosmetic 1', 'Cosmetic 2', 'Cosmetic 3', and 'Cosmetic M'. The columns represent different ingredients, labeled 'Ingredient 1', 'Ingredient 2', 'Ingredient 3', 'Ingredient 4', and 'Ingredient N'. A bracket on the left side of the rows is labeled '# of the items', indicating the total number of cosmetic products. A bracket at the bottom of the columns is labeled '# of the ingredients', indicating the total number of unique ingredients. The matrix cells are currently empty, representing a zero-filled matrix ready for data entry.

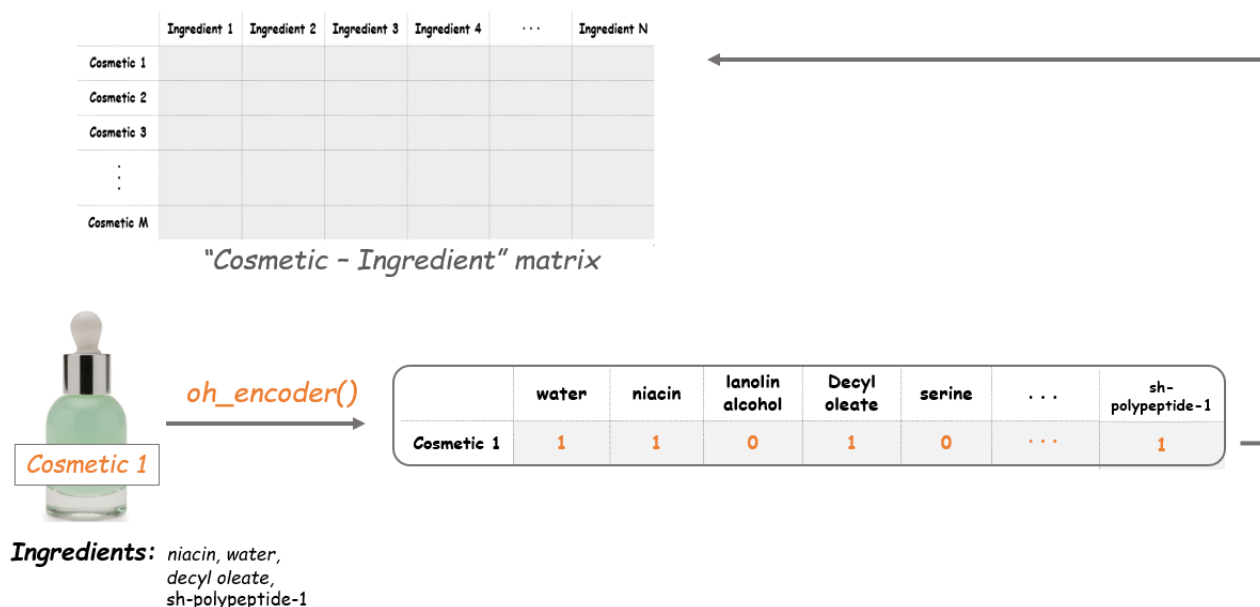
	Ingredient 1	Ingredient 2	Ingredient 3	Ingredient 4	...	Ingredient N
Cosmetic 1						
Cosmetic 2						
Cosmetic 3						
⋮						
Cosmetic M						

### 3.4 Counter Function

Creating a counter function. Before we can fill the matrix, let's create a function to count the tokens (i.e., an ingredients list) for each row. Our end goal is to fill the matrix with 1 or 0: if an ingredient is in a cosmetic, the value is 1. If not, it remains 0. The name of this function, one-hot encoder, will become clear next.

### 3.5 Cosmetic-Ingredient matrix

Cosmetic-Ingredient matrix. Now we'll apply the one-hot encoder function to the tokens in corpus and set the values at each row of this matrix. So, the result will tell us what ingredients each item is composed of. For example, if a cosmetic item contains water, niacin, decyl aleate and sh-polypeptide-1, the outcome of this item will be as follows:



This is what we called one-hot encoding. By encoding each ingredient in the items, the Cosmetic-Ingredient matrix will be filled with binary values.

## 4. DIMENSION REDUCTION WITH t-SNE

The dimensions of the existing matrix are (190, 2233), which means there are 2233 features in our data. For visualization, we should downsize this into two dimensions. We'll use t-SNE for reducing the dimension of the data here.

**T-distributed Stochastic Neighbor Embedding (t-SNE)** is a nonlinear dimensionality reduction technique that is well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, this technique can reduce the dimension of data while keeping the similarities between the instances. This enables us to make a plot on the coordinate plane, which can be said as vectorizing. All of these cosmetic items in our data will be vectorized into two-dimensional coordinates, and the distances between the points will indicate the similarities between the items.

After having these two coordinates we know the distance point. Using this distance point we will find the Euclidean distance between them and arranging them in descending order to get the top k matches. The visualization of the coordinates is done using Bokeh plotting.

In short t-SNE or t-distributed stochastic neighbor embedding is an unsupervised machine learning method to find a low-dimensional representation of the data. This representation, or embedding is then used to find similar products.

### 4.1 Results

	index	Label	Brand	Name	Price	Ingredients	distance
1	56	Moisturizer	FRESH	Crème Ancienne®	290	Limnanthes Alba (Meadowfoam) Seed Oil, Water, ...	[[0.0010394990003981376]]
2	34	Moisturizer	LA MER	The Renewal Oil	245	Limnanthes Alba (Meadowfoam) Seed Oil, Dimethi...	[[0.0389609832995188]]
3	87	Moisturizer	GUERLAIN	Midnight Secret Late Night Recovery Treatment	29	Visit the Guerlain boutique	[[0.10171046238336909]]
4	32	Moisturizer	CAUDALIE	Grape Water	10	Vitis Vinifera (Grape) Fruit Water*, Vitis Vin...	[[0.30561077506528417]]
5	71	Moisturizer	FARMACY	Sleep Tight Firming Night Balm with Echinacea ...	48	Prunus Amygdalus Dulcis (Sweet Almond) Oil, Co...	[[0.3555075019294346]]

Top k matches recommended as per the distance measure would be this.

I even created an app with streamlit to make it interactive for the user. So, basically Streamlit is an open-source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc.

The app would look like this:

## Find the Right Skin Care for you!!!

Hi there! 🤖 If you have a skincare product you currently like I can help you find a similar one based on the ingredients.

Please select a product below so I can recommend similar ones

My dataset contains 1400+ products but unfortunately it is possible that I do not have the product you are looking for 😊

Select a product category

Moisturizer

Select a brand

ALGENIST

Select the product

Firming & Lifting Neck Cream

Select your skin type

Combination

Find similar products!

Here the user can select their product category, brand they need, selecting the product the brand provides, and the skin type to get the products with similar ingredients across all brands. This will help the user to choose their products according to their personal needs without any chemical background.

## 5. CONCLUSION

This project used ingredient as a base recommending the product based on the personal preference of the user. Recommendation based on product properties can be applied to any kinds of domains. It could be also applied to a book or a wine recommendation. We can make various maps with the product features. Visualizing products on the coordinate plane helps us to understand the relations between the items in an intuitive way. With this analysis, we can go further toward more refined services. It could show the images and the reviews of the item or provide a direct link to the page for ordering. We can also make a different plot whose axes represent chemical attributes. For example, if we add some expert chemical knowledge, we can set the axes for hydration or toxicity. In that case, it will be able to see how much toxicity each item is. These kinds of application will offer a higher level of product analysis.

## BIBLIOGRAPHY

- [1] 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>
- [2] 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>
- [3] Hongwu Ye. 2011. A Personalized Collaborative Filtering Recommendation Using Association Rules Mining and Self-Organizing Map. Journal of Software 6, 4 (April 2011). <https://doi.org/10.4304/jsw.6.4.732-739>

