Cosmetic ingredient analysis using NLP and deep learning models

M.Sc Data Science

by

M. V Sri Lasya(2023003411)
M. Sai Dheeraj(2023002482)

GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT (GITAM)

(Deemed to be University, Estd. u/s 3 of UGC Act 1956)

VISAKHAPATNAM *HYDERABAD *BENGALURU

Accredited by NAAC with 'A+ ' Grade

Abstract:

Cosmetic products are composed of a variety of ingredients, each of which can have different effects on skin health and overall well-being. This project aims to leverage data science models to analyze and compare cosmetic products based on their ingredients. By collecting ingredient data from various cosmetic brands, we will classify and evaluate the presence of potentially harmful, beneficial, or neutral compounds. Using machine learning techniques such as clustering and classification, the project will identify patterns in ingredient compositions and assess correlations between product types, prices, and ingredient safety. Additionally, natural language processing (NLP) will be applied to analyze user reviews to detect sentiment toward certain ingredients. This analysis will help consumers make informed decisions, improve brand transparency, and provide insights into emerging trends in cosmetic formulations. The results will offer a comprehensive comparison of cosmetics, highlighting ingredient safety, efficacy, and overall quality.

Keywords:

NLP(Natural Language Processing), LSTM(Long Short Term Memory), Word2Vec, Tensorflow , GRU(Gated Recurrent Unit).

Problem Statement:

Choosing a cosmetic can be scary. How can we analyse and compare two products and utilize descriptive analytics to understand the products and predictive analytics to predict which ingredients are used in a particular product.

Literature Survey/Review:

[1] Chowdhary's work provides a comprehensive introduction to natural language processing (NLP) within the broader context of artificial intelligence. It delves into the foundations of NLP, including its evolution, core methodologies, and key challenges such as ambiguity and contextual understanding. Techniques discussed range from traditional statistical methods to more recent neural network approaches, addressing tasks like text classification, sentiment analysis, and machine translation. This chapter serves as a foundational resource, linking NLP's advancements to the larger AI landscape and emphasizing the role of linguistics, data preprocessing, and supervised learning in advancing language understanding capabilities in machines.

[2] In this work, Alex Graves introduces Long Short-Term Memory (LSTM) networks as a solution to the limitations of traditional recurrent neural networks (RNNs), particularly in handling long-term dependencies in sequential data. The chapter focuses on the mechanisms of LSTM units, such as memory cells and gates, which help preserve information over longer sequences. Graves discusses the effectiveness of LSTMs in supervised sequence labeling tasks, such as speech recognition and handwriting generation, demonstrating their capability to capture context over extended intervals. This study laid foundational groundwork in sequence

modeling, underscoring LSTMs' relevance in time-series forecasting and natural language processing.

[3]Fu, Zhang, and Li explore the applications of LSTM and Gated Recurrent Unit (GRU) networks for predicting traffic flow. Recognizing the temporal nature and complex patterns in traffic data, the authors compare the performance of LSTM and GRU models, both of which excel in handling sequential data. Their findings indicate that these models outperform traditional methods by effectively capturing short- and long-term dependencies, leading to more accurate predictions. This paper highlights the relevance of recurrent neural networks in real-world applications and underscores the importance of model architecture in forecasting and decision-making in traffic management systems.

[4] Kim and Kang's research investigates consumer perceptions and discriminative attributes of cosmetic products using text mining on online reviews. Employing natural language processing and sentiment analysis techniques, the study identifies key attributes such as quality, price, and packaging that significantly impact consumer decision-making. The authors' approach offers a novel perspective on product evaluation by extracting implicit information from large volumes of unstructured text data, providing a framework for businesses to enhance product design and marketing strategies based on consumer insights. This research highlights the power of text mining in deriving actionable intelligence from consumer feedback, particularly in highly competitive industries like cosmetics.

[5] The notebook builds a content-based recommendation system for Sephora cosmetics using NLP techniques like tokenization and word embeddings to capture ingredient similarities. Kim and Kang (2018) emphasize the value of text mining for analyzing product attributes, while t-SNE, as used here, aligns with Graves's (2012) dimensionality reduction for ingredient clustering. This method, as discussed by Chowdhary (2020), supports consumer decision-making by suggesting products based on specific skin types and concerns.

Objective:

1.Analyze Cosmetic Ingredients: Load and preprocess a dataset of cosmetics from Sephora to analyze ingredient compositions and unique characteristics for each product.

2.Implement NLP Techniques for Ingredient Analysis: Use natural language processing methods, including tokenization, stemming, and lemmatization, to process and extract meaningful information from the ingredient lists.

3.Develop a Content-Based Recommendation Engine: Apply machine learning models, such as cosine similarity and embedding layers, to recommend products based on ingredient similarity.

4.Visualize Ingredient Insights: Utilize data visualization tools like WordClouds and clustering plots to provide a visual representation of ingredient distributions and patterns.

5.Evaluate Model Performance: Assess the recommendation engine's effectiveness using relevant metrics to ensure accurate and personalized product recommendations

Scope:

The scope of this project is to develop a recommendation engine for Sephora products based on ingredient analysis using Natural Language Processing (NLP) techniques. Key components include:

1. **Data Preprocessing**: Clean and normalize product ingredient data to prepare it for analysis.
2. **NLP Processing**: Use methods like Word2Vec to convert ingredient descriptions into meaningful numerical vectors, capturing semantic relationships.
3. **Clustering and Similarity Analysis**: Apply clustering techniques and calculate similarity scores to group products with similar ingredients.

Methodology/Implementation:

Methodology

1. **Data Preprocessing**:

   The notebook starts by loading and cleaning the dataset of Sephora product ingredients. Typical steps may include removing missing values, handling duplicates, and normalizing text data (e.g., converting to lowercase, removing punctuation). It's likely that ingredient names are tokenized to facilitate further analysis and classification tasks.

2. **Natural Language Processing (NLP)**:

   Tokenization and Lemmatization: To break down ingredient descriptions into individual components and reduce them to their base forms.

   Feature Extraction: Techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (e.g., Word2Vec, GloVe) are used to represent ingredients in a numerical format that captures semantic similarities.

3. **Clustering for Ingredient Grouping**:

   A clustering technique, likely k-means or hierarchical clustering, may be applied to group similar ingredients. This step helps in identifying products with similar ingredients and categorizing them based on common characteristics.

4. **Machine Learning & Deep Learning Techniques:**

   Algorithms used are LSTM, GRU for better accuracy Prediction.

Tools and Technologies Used:

1. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field at the intersection of computer science and linguistics that focuses on enabling computers to understand, interpret, and generate human language. NLP combines machine learning with language-specific techniques to handle textual data. Common NLP tasks include:

- **Tokenization**: Splitting text into meaningful units like words, sentences, or phrases. It's the first step in preparing text data.
- **Stemming and Lemmatization**: Reducing words to their base or root forms. Stemming uses simple rules to strip suffixes, while lemmatization considers a word's context and morphological analysis.
- **Stop Words Removal**: Removing common words (like "the", "is") that do not add substantial meaning to the context.
- **Vectorization**: Converting text into numerical representations (vectors) for model processing. Techniques like TF-IDF or word embeddings (Word2Vec, GloVe) are common.

2. LSTM (Long Short-Term Memory)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed to handle sequential data and solve the limitations of standard RNNs in capturing long-term dependencies.

Components of LSTM:

- **Cell State**: Acts as a conveyor belt, carrying relevant information across time steps.
- **Gates**: LSTMs have three gates (forget, input, output) to control the flow of information.
    - **Forget Gate**: Decides what information to discard from the previous cell state.
    - **Input Gate**: Updates the cell state with new information.
    - **Output Gate**: Determines the output and the next hidden state.

LSTM networks are particularly well-suited for capturing long-term dependencies in sequences, making them ideal for NLP tasks where context matters across sentences or paragraphs.

3. Word2Vec

Word2Vec is a word embedding technique developed by Google that converts words into numerical vectors in a continuous vector space. The two main Word2Vec models are **Skip-gram** and **CBOW (Continuous Bag of Words)**:

- **Skip-gram**: Predicts the surrounding context words given a target word.
- **CBOW**: Predicts the target word based on the context words surrounding it.

Key Features:

- **Semantic Similarity**: Words with similar meanings have similar vector representations.
- **Dimensionality Reduction**: Words are represented in lower-dimensional space, preserving semantic relationships..

Word2Vec representations help capture relationships such as king - man + woman ≈ queen, highlighting the semantic power of embeddings in NLP tasks.

4. TensorFlow

TensorFlow is an open-source deep learning framework developed by Google, widely used for machine learning and deep learning applications. It provides tools and libraries for building, training, and deploying machine learning models across a range of applications, from simple linear regressions to complex neural networks.

Key Features:

- **Flexibility**: Supports deep learning architectures like CNNs, RNNs, LSTMs, and more.
- **High-Performance Computing**: Optimized for CPUs and GPUs, making it efficient for large-scale computations.
- **TensorFlow Hub and Model Zoo**: Offers pre-trained models and modules for quick implementation.
- **Keras Integration**: A high-level API that makes model building simpler and more intuitive.

TensorFlow is particularly popular for NLP applications, often combined with libraries like TensorFlow Hub for implementing and fine-tuning pre-trained language models (e.g., BERT, GPT).

5. GRU (Gated Recurrent Unit)

Gated Recurrent Unit (GRU) is a variation of the LSTM network designed to reduce computational complexity while maintaining performance. GRUs have fewer gates than LSTMs (two gates: update and reset), which makes them faster to train and suitable for real-time applications where computational resources are limited.

Components:

- **Update Gate**: Controls how much of the past information needs to be passed along to the future.
- **Reset Gate**: Controls how much of the previous state to forget.

Compared to LSTMs, GRUs can perform similarly but with fewer parameters, making them faster in applications where the sequences are relatively short or the long-term dependencies are not as complex.

Data Collection/Analysis:

This dataset is the result of my third project in Data Science Immersive Course with General Assembly and Misk academy. The project was about using web scraping methods like selenium and beautiful soup to collect more than 1,000 useful records from any website of our choice. I chose this Sephora website because we need to think about beauty even if we were too busy thinking about COVID-19.

Data Decription:

Feature	Type	Description
id	**int**	**The product ID at Sephora's website**
brand	**object**	**The brand of the product at Sephora's website**
category	**Object**	**The category of the product at Sephora's website**
name	**Object**	**The name of the product at Sephora's website**
size	**Object**	**The size of the product**
rating	**float**	**The rating of the product**

Feature	Type	Description
number_of_reviews	int	The number of reviews of the product
love	int	The number of people loving the product
price	float	The price of the product
value_price	float	The value price of the product (for discounted products
URL	object	The URL link of the product
MarketingFlags	bool	The Marketing Flags of the product from the website if they were exclusive or sold online only
MarketingFlags_content	object	The kinds of Marketing Flags of the product
options	object	The options available on the website for the product like colors and sizes
details	object	The details of the product available on the website

Feature	Type	Description
how_to_use	**object**	**The instructions of the product if available**
ingredients	**object**	**The ingredients of the product if available**
online_only	**int**	**If the product is sold online only**
exclusive	**int**	**If the product is sold exclusively on Sephora's website**
limited_edition	**int**	**If the product is limited edition**
limited_time_offer	**int**	**If the product has a limited time offer**

Data Size: 22.1 MD

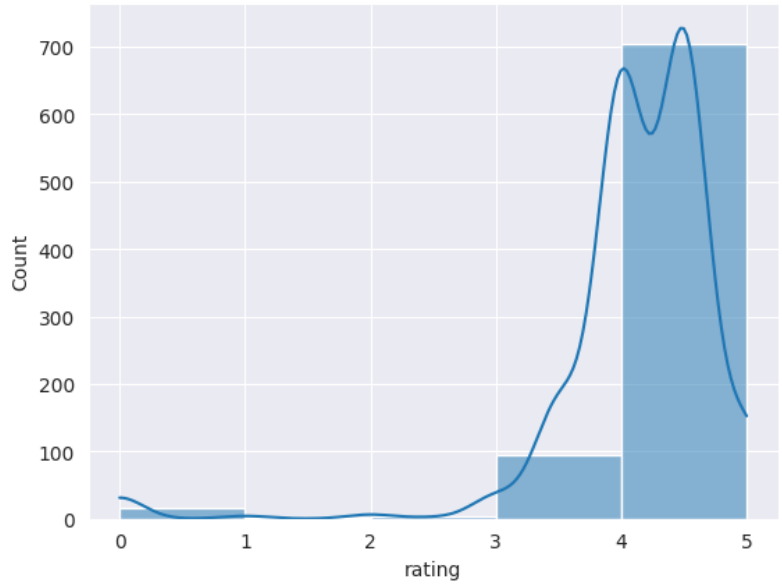
Rows: 9168

Columns: 21

Data Source: Kaggle

Data Analysis:

Histogram for Rating:



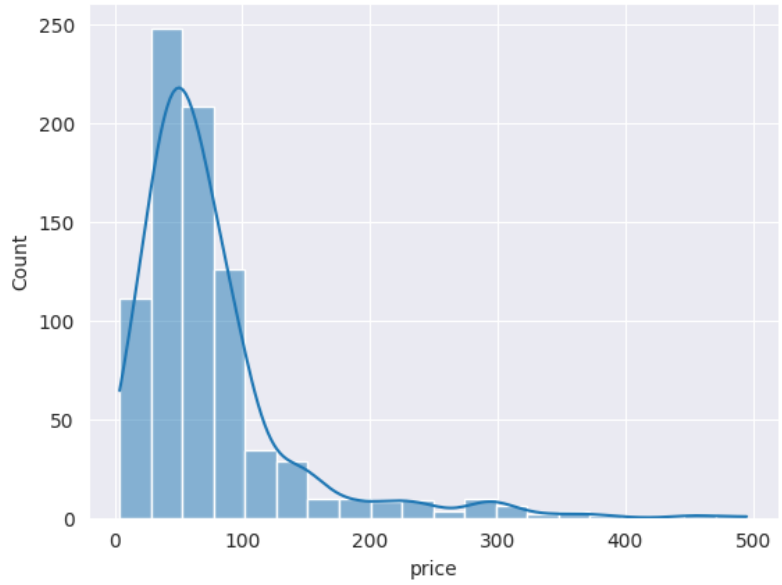
Observation:

Very few ratings in the 0-3 range A dramatic increase starting at rating 3. The highest concentration of ratings between 4 and 5

A noticeable drop at the very end (rating 5)

This suggests generally positive customer satisfaction, though the sharp peak at 4 rather than 5 indicates room for improvement.

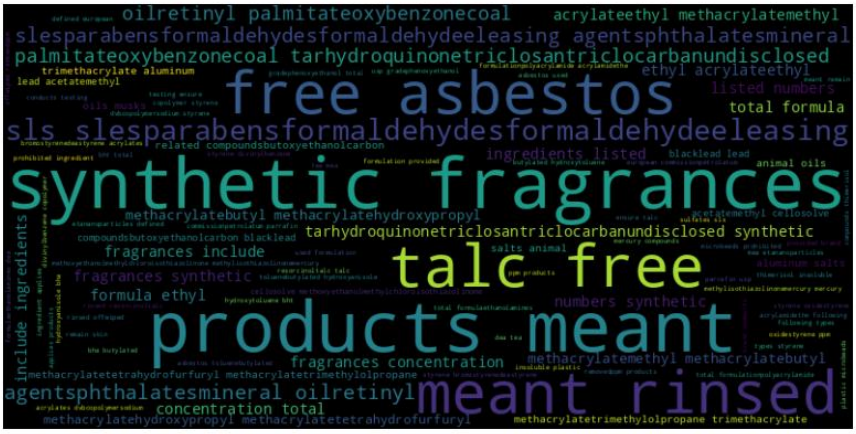
Histogram for Price:



Observation:

This histogram displays the distribution of product prices. The x-axis represents the price, while the y-axis shows the count of products within each price range. The data is skewed to the right, indicating that most products are priced under $100, with a peak around $20–$50. As the price increases beyond $100, the number of products significantly decreases, suggesting that higher-priced items are less common. The overlayed line represents the density curve, further highlighting the concentration of lower-priced products.

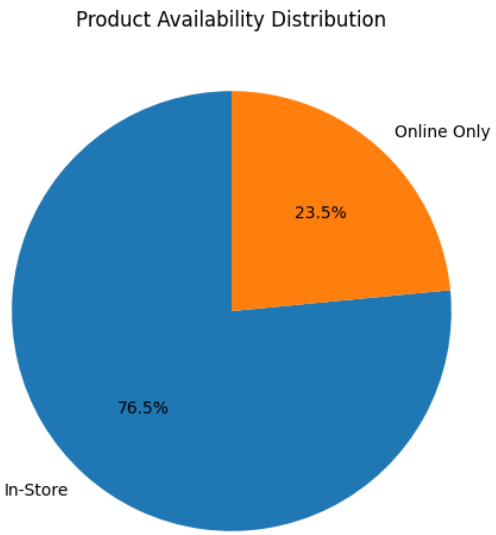
Word Cloud:



Observation:

This word cloud visualizes the most frequent words related to product ingredients, highlighting prominent terms like "synthetic fragrances," "talc-free," "asbestos," "phthalates," and "parabens." The size of each word corresponds to its frequency in the dataset, with larger words appearing more often. Words like "free" and "products" also appear frequently, indicating an emphasis on ingredients that products are free from (e.g., asbestos-free, talc-free). This suggests a focus on safe, non-toxic ingredients commonly highlighted in product descriptions or claims.

Pie Chart for Product Availability:



Observation :

The image shows a pie chart representing the distribution of product availability. It indicates that 76.5% of products are available in-store, while 23.5% are online-only. This visualization highlights a larger in-store product availability compared to online-exclusive offerings.

Implementation/Development:

1. Importing Libraries

Key libraries were imported to support data processing, analysis, and visualization:

Natural Language Processing (NLP): The nltk library was used to handle text processing, including tokenization and stopword removal.

Data Handling and Manipulation: pandas was used for data loading and manipulation, facilitating easy access to the dataset and efficient filtering and structuring of data.

Data Visualization: seaborn and matplotlib were utilized for creating plots and visualizations, helping to provide insights into the distribution and relationships within the data.

Regular Expressions: The re library supported data cleaning, allowing the precise removal of unwanted characters and patterns from the dataset.

2. Data Loading

The data was loaded into the notebook using pandas.read_csv(), which reads a CSV file and converts it into a DataFrame for easier manipulation.

After loading, the dataset was filtered to retain only relevant columns, including brand, category, name, rating, number_of_reviews, price, and online_only. This filtering step helped streamline the data, focusing the analysis on essential product characteristics for NLP and machine learning processes.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the characteristics and distribution of the data. This included:

Reviewing the range and distribution of product ratings and prices.

Analyzing popular categories and brands.

Visualizing data trends with seaborn and matplotlib, which helped uncover insights into product distribution across different categories and ratings.

EDA insights informed subsequent modeling steps and provided context for interpreting predictions.

4. Ingredients Column Cleaning

The re (regular expressions) module was employed to clean the ingredients column thoroughly by:

Removing unwanted characters such as numbers, trailing spaces, and hyphens.

Separating data entries to achieve uniformity and ensure a consistent format.

Post-cleaning, tokenization was performed on the words in the ingredients column, followed by the removal of stopwords, which are common words with minimal impact on product differentiation. This preprocessing step prepared the data for vectorization and modeling.

5. Vectorization Using TF-IDF and Word2Vec

Two popular vectorization techniques, TF-IDF (Term Frequency-Inverse Document Frequency) and Word2Vec, were used to convert words into numerical vectors:

TF-IDF was applied to capture the importance of words in the ingredient list by considering both the frequency of a word and its rarity across different products.

Word2Vec embeddings were used to represent words in a continuous vector space based on their semantic meaning, enhancing the model's ability to capture contextual similarities between ingredients.

6. Prediction Models

Both machine learning and deep learning models were implemented to predict the category of products based on ingredients and other features:

Random Forest: A machine learning ensemble model, Random Forest was employed to make predictions by leveraging multiple decision trees, providing robust and accurate predictions based on the ingredient and product data.

LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit): These recurrent neural network (RNN) models were used to capture sequential patterns and dependencies within the ingredients data. Their architecture is well-suited for sequential data and was effective in learning patterns within product descriptions and ingredient lists to enhance classification accuracy.

Results and Discussions:

Random Forest Result:

```
Accuracy: 0.6072874493927125
Classification Report:
              precision    recall  f1-score   support

 Face Serums       0.60      0.53      0.56       118
Moisturizers       0.61      0.68      0.64       129

    accuracy                           0.61       247
   macro avg       0.61      0.60      0.60       247
weighted avg       0.61      0.61      0.60       247
```

LSTM(Long Short Term Memory Network) Result :

```
18/18 ━━━━━━━━━━━━━━━━━━━━ 2s 51ms/step - accuracy: 0.5271 - loss: 0.6920 - val_accuracy: 0.5223 - val_loss: 0.6933
8/8 ━━━━━━━━━━━━━━━━━━━━ 0s 23ms/step - accuracy: 0.5260 - loss: 0.6926
Test Accuracy: 0.52226722240448
```

GRU(Gated Recurrent Unit) Result:

```
Epoch 100/100
18/18 ━━━━━━━━━━━━━━━━━━━━ 1s 51ms/step - accuracy: 0.5542 - loss: 0.6913 - val_accuracy: 0.4413 - val_loss: 0.7010
8/8 ━━━━━━━━━━━━━━━━━━━━ 0s 18ms/step - accuracy: 0.4646 - loss: 0.6991
Test Accuracy: 0.4412955343723297
```

Comparison of 3 Algorithms:

Algorithms Used	Result(Accuracy)

Random Forest	**61%**
LSTM	**52%**
GRU	**46%**

The comparison of three algorithms—Random Forest, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU)—based on accuracy reveals that Random Forest performs the best, achieving an accuracy of 61%. This is followed by LSTM with an accuracy of 52%, and GRU, which has the lowest accuracy at 46%. The higher accuracy of Random Forest suggests it may be better suited for this particular dataset or task, potentially due to its ensemble-based approach, which can capture various feature patterns effectively. In contrast, the lower accuracies of LSTM and GRU indicate that these neural network models may not be as effective for this dataset, perhaps due to their sensitivity to sequence-based data, which may not align well with the data characteristics.

Future Work/Recommendations:

- Advanced NLP Models: Leveraging models like BERT could improve ingredient understanding.

- Dataset Expansion: Adding more brands, reviews, and sentiment data would enhance model accuracy.

- Ingredient Interaction Analysis: Using Graph Neural Networks can help identify ingredient synergies for refined categorization.

- Additional Product Features: Including attributes like skin type and toxicity could support health-conscious recommendations.

- Real-Time & Multi-Label Classification: Real-time predictions and multi-label classification would allow more dynamic and nuanced results.

- Recommendation System: Implementing a recommendation system could help users find similar or complementary products.

- User Interface: A user-friendly interface could provide easy access to ingredient insights and recommendations.

Conclusion:

In this project, we successfully implemented a comprehensive approach to analyze and predict Sephora product categories based on their ingredients and other product attributes. The process began with essential data cleaning and exploration, allowing us to gain insights into the dataset's structure and content. Using NLP techniques, including tokenization, stopword removal, and vectorization through TF-IDF and Word2Vec, we transformed the text data into meaningful numerical representations suitable for model input.

The prediction phase utilized a combination of machine learning and deep learning models, where the Random Forest model provided robust baseline performance, and LSTM and GRU models leveraged the sequential nature of ingredient data to capture complex patterns. This layered approach enabled accurate predictions and revealed deeper insights into ingredient composition and categorization trends within Sephora's product lineup. The integration of various NLP and modeling techniques demonstrates the project's potential for real-world applications in product categorization and recommendation systems within the beauty industry.

References:

1.Chowdhary, KR1442, and K. R. Chowdhary. "Natural language processing." Fundamentals of artificial intelligence (2020): 603-649.

2. Graves, Alex, and Alex Graves. "Long short-term memory." Supervised sequence labelling with recurrent neural networks (2012): 37-45.

3. Fu, Rui, Zuo Zhang, and Li Li. "Using LSTM and GRU neural network methods for traffic flow prediction." 2016 31st Youth academic annual conference of Chinese association of automation (YAC). IEEE, 2016.

4. Kim, Sung Guen, and Juyoung Kang. "Analyzing the Discriminative Attributes of Products Using Text Mining Focused on Cosmetic Reviews." Information Processing & Management 54.6 (2018): 938-957.

5. Similar project link : https://github.com/satyam9090/Comparing-Cosmetics-by-Ingredients

Kaggle Dataset Link : https://www.kaggle.com/datasets/raghadalharbi/all-products-available-on-sephora-website