STAT 437 Final Project Proposal

Yuzhou Fu

Business Senario (What)

Act as a data scientist in the beauty product company, apply NLP and unsupervised learning technique to mine online customer's product reviews and customer data, deliver business insights to the relevant stakeholders.

- Two main business objectives:
  1. **Product-centric**: Based on customer product reviews, cluster the reviews to discover the strengths and weaknesses of products that are in the same cluster, and analyze clusters by using available features to deliver insights to the product and marketing teams.
  2. **Customer-centric**: Create a customer dataframe (customer as the primary key) containing aggregated features such as avg_rating, total_reviews and preferred_category. Perform clustering on the customers to identify customer segments.

  Analyze both clustering results, expecting the existence of cluster intersections, for example, one cluster of customers corresponds to one cluster of product reviews, which can be leveraged to design targeted customer actions.

Project Purpose (Why)

- Connecting the machine learning model results with business needs is crucial for data science practitioners in the industry. This project, which simulates a business scenario using machine learning techniques on a realistic Sephora dataset, provides a great chance for me to practice this ability.
- This proposed project is closely related to data science work in the fast-moving consumer goods industry, which is my intended career path. It provides an opportunity to develop a comprehensive project that can be included in my data science portfolio.

Project Phase (How)

- **11/11 ~ 16/11**

  EDA

  Data-preprocessing, text cleaning, text-preprocessing
- **17/11 ~ 23/11**

  Data-preprocessing, text cleaning, text-preprocessing

  Feature engineering:

  - word embedding: SBERT + NMF (tentative models, recommended by AI)
- **24/11 ~ 30/11**

  Feature engineering:

  - word embedding: SBERT + NMF
  - Customer dataframe creation

  Apply unsupervised learning

  - clustering

- **01/12 ~ 07/12**

  Apply unsupervised learning

  - clustering

  Analyze result

  Project report

- **08/12 ~ 14/12**

  Analyze result

  Project report

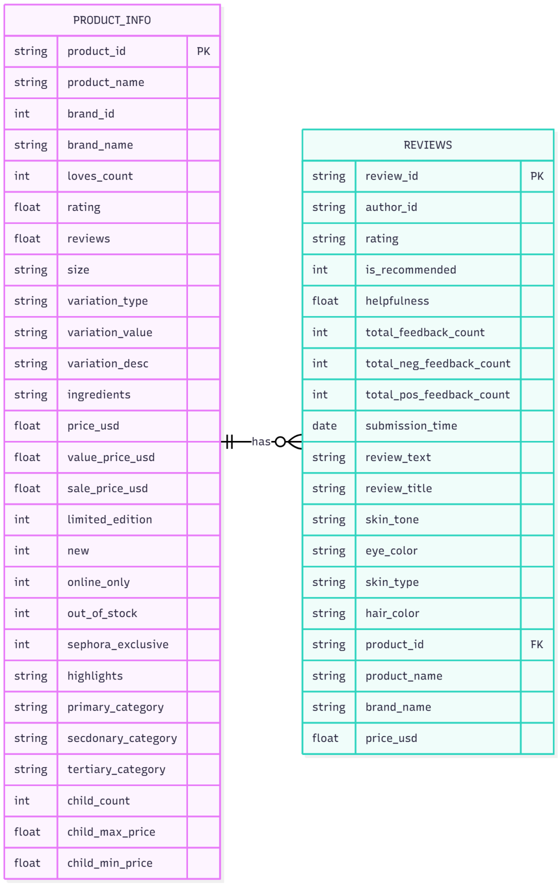
Dataset

- **Sephora Products and Skincare Reviews**

  Primary key: product_id, review_id

  Foreign key: product_id

  - Two tables:
    - product information (8494 products in total): product name, price . . .
    - customer information (1,094,411 reviews in total, dates ranging from 2018 to 2023, may not use all the data): reviews, customer profile . . .



PRODUCT_INFO		
string	product_id	PK
string	product_name	
int	brand_id	
string	brand_name	
int	loves_count	
float	rating	
float	reviews	
string	size	
string	variation_type	
string	variation_value	
string	variation_desc	
string	ingredients	
float	price_usd	
float	value_price_usd	
float	sale_price_usd	
int	limited_edition	
int	new	
int	online_only	
int	out_of_stock	
int	sephora_exclusive	
string	highlights	
string	primary_category	
string	secdonary_category	
string	tertiary_category	
int	child_count	
float	child_max_price	
float	child_min_price	

REVIEWS		
string	review_id	PK
string	author_id	
string	rating	
int	is_recommended	
float	helpfulness	
int	total_feedback_count	
int	total_neg_feedback_count	
int	total_pos_feedback_count	
date	submission_time	
string	review_text	
string	review_title	
string	skin_tone	
string	eye_color	
string	skin_type	
string	hair_color	
string	product_id	FK
string	product_name	
string	brand_name	
float	price_usd	

Reference

- text cleaning and text preprocessing reference
- NLP with Disaster Tweets - EDA, Cleaning and BERT
- N-grams
- A Comprehensive Guide to Word Embeddings in NLP
- Twitter sentiment Extaction-Analysis,EDA and Model
- Word Clouds
- Skincare Products EDA & Sentiment Analysis (Exactly the same dataset using)
- From Raw Text to Insightful Analysis: NLP Text Preprocessing Explained
- BERT Word Embeddings Tutorial