TaoBao Products Analysis: Relationship Between Product Name and Price

Tianyi Wang: wtyjoyce@umich.edu Yixuan Jiang: yxjiang@umich.edu Dongming Yang: dongming@umich.edu
Department of Statistics, College of Literature, Science, and The Arts, University of Michigan

Introduction

Motivation

Our project investigates how product names and categories can lead to price disparities on online platforms like Amazon and Taobao. We specifically explore how seemingly disparate items, such as 'lipstick holders' and 'marker pen holders', often **share functional similarities but are priced differently** due to their respective categories. By uncovering these functional equivalents across diverse categories, **Our aim** is to guide consumers towards making more informed and budget-friendly choices by highlighting these functional equivalents in different categories.

Previous Works

- ***An Analysis of Factors Affecting on Online Shopping Behavior of Consumers[1]***:
  Studies exploring the psychological factors that impact consumer spending habits online.
- ***Dynamic Pricing under Competition on Online Marketplaces[2]:***
  Analyses revealing how e-commerce platforms use product naming and categorization as a tool for price differentiation.

Research Question

Investigating how product names influence pricing for functionally similar products.

Data

Raw Data

Web scraping products information including images from https://world.taobao.com/

Data Size

The Dataset contains **89374** rows with 10 columns. Columns including Product Name, Price, Categories, Image Link, Product Link and etc.

Data Pre-Processing

1. **Data Cleaning:** Convert to lower cases, Remove punctuations and numbers
2. **Tokenization:** Utilize nltk word tokenization and remove stop words.
3. **Lemmatization:** Reduce words to their base or root form

Example Data for Product Names and Images

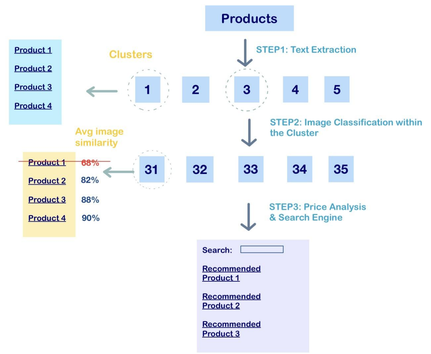
- **Ex1**. New science and technology velvet sofa living room modern simple wabisabi wind light luxury size apartment straight row puff fabric sofa
- **Ex2.** Long sleeve fitness tight men run sport set spring autumn high spring quick dry clothe basketball football run clothe

Ex3. Product Name:
Embroidery organizer box Storage box Cross stitch box Needle box winding board Plastic tools large box winding device

Ex.4 Product Name:
Multifunctional transparent plastic portable small storage box Treasure box cross-stitch sewing box Large storage box diy

Method

Diagram



Text Extraction

- **LDA & POS Tagging & BERTopic**: Leveraged Latent Dirichlet Allocation and Part-Of-Speech tagging to extract and categorize key functional terms from product description.
- **TF-IDF & K-means**: Utilized Term Frequency-Inverse Document Frequency combined with K-means clustering to identify groups of products with similar textual features, suggesting functional similarities.

Image Classification

- **CNN with VGG16**: Employed Convolutional Neural Network using the pre-trained VGG16 model to analysis product images. This method extracts complex visual patterns to understand and categorize products based on visual similarity.
- **Cosine Similarity**: Filtered image comparisons using cosine similarity, focusing on scores above 80% to ensure high visual similarness.

Search Engine Integration

- **Innovative Discovery & Filtering**: Combined textual and visual analytics into a sophisticated search platform with multidimensional filtering, from text to visuals to pricing, for comprehensive product searches.
- **User-Centric Design & Interaction**: Tailored the interface for personalized navigation, featuring reasonable search criteria and interactive, clickable results for a seamless transition from search to purchase.
- **Visual Insights & Decision Support**: Integrated dynamic visual tools for product comparisons, empowering users with immediate, visual insights for informed decision-making.

Result

Cluster Silhouette Score and Distribution:

- Demonstrated **high silhouette score** in product clusters, indicating effective grouping based on text and image features.
  - Silhouette score **for TF-IDF: 0.847392**
- Visualized cluster distribution in **a heatmap**, 8.

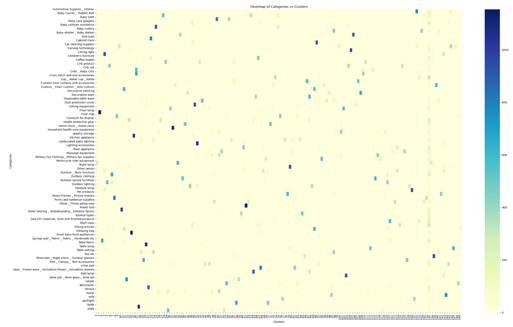
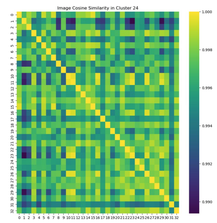
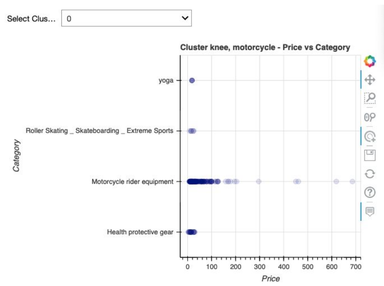


Image Classification Heatmap

- Constructed a **heatmap to visualize image classification**, highlighting product clusters with strong visual similarities.
- Only pairs with a cosine similarity above 80% are displayed, **enabling clear identification of closely related products**.



Price Analysis & Interactive Exploration

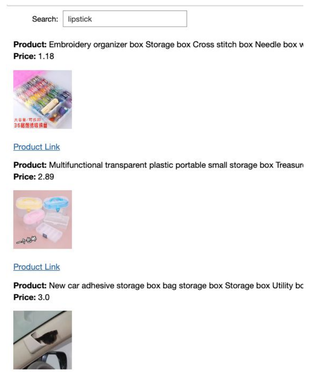


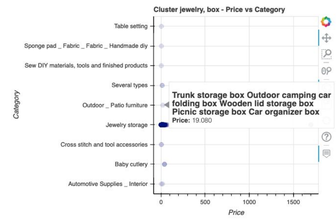
- Executed a price analysis over 150 product clusters, discovering **distinct pricing trends**.
- Crafting interactive visualizations for each cluster to enable dynamic investigation into the **correlations between price points and product categories**.

Result

Integrated Search Engine

- Implemented a **search engine** that combines text and image analyses, offering **functionally relevant product suggestions**.
- Augmented the search engine with **clickable interactive visualizations**, enabling users to be directed from comparative analysis to specific product pages for efficient navigation and informed purchasing





Conclusion and Future Work

Conclusion

- Analysis of image similarity indicates: the existence of cross-category products with similar functions but varying price, suggesting the **importance of a strategy for customers to identify cost-effective products**
- Within each newly generated cluster, **keywords of products** names were identified, **leading to options with lower prices**, as evidenced in the price analysis plots
- **A developed search engine** recommends more economical product options based on user-entered keywords, offering direct links to product pages

Future Work

- **Develop a Personalized Recommendation Algorithm**: Future work involves enhancing the search engine algorithm for more accurate personalization in product recommendations, considering user behavior and preference data
- **Price Analysis Expansion**: Future studies will broaden price analysis to include time series data, market trends, and consumer demands, offering a better recommendation to consumers

Reference

- *[1] Javadi M H M, Dolatabadi H R, Nourbakhsh M, et al. An analysis of factors affecting on online shopping behavior of consumers[J]. International journal of marketing studies, 2012, 4(5): 81.*
- *[2] Schlosser R, Boissier M. Dynamic Pricing under Competition on Online Marketplaces: A Data-Driven Approach[C]//KDD. 2018: 705-714.*