TUNKU ABDUL RAHMAN UNIVERSITY COLLEGE

BEYOND EDUCATION

BACS2003 ARTIFICIAL INTELLIGENCE

202301 Session, Year 2022/23

Assignment Documentation

Full Name: Oh Boon Suen
Student ID: 21WMR02965
Programme: RSW2S3
Tutorial Class: Group 2
Project Title: Product Recommender System
Module In-Charged: Content-Based Filtering Recommender System

Other team members' data

No	Student Name	Module In Charge
1	Chai Wei Qi	Simple Recommender System
2	Tan Cherng Ming	Collaborative Filtering Recommender System

Lecturer: Dr. Goh Ching Pang	Tutor : Dr. Ho Chuk Fong

1. Introduction

1.1. Problem Background

MyElectronic is an e-commerce business that specializes in selling a wide range of electronic products through its website. However, the users of the website have provided feedback that they have experienced difficulty discovering products that meet their specific needs and preferences, due to the overwhelming number of products available to them. Hence, the business is facing challenges in facilitating seamless product discovery and purchase experiences for users. The business is committed to addressing this issue and improving the user-friendliness and efficiency of the website to ensure that their customers can easily browse, select, and purchase desired beauty products with ease and convenience. Besides, customers have provided feedback that scrolling through the product list in the product catalog page is often time-consuming and less interesting in finding the desired products. It results in customer frustration and loss of retention, leaving a suboptimal shopping experience for the users.

To optimize user experience and align with the purpose of online shopping, it is crucial to address the issues of product discovery and streamline the purchasing process. It can be solved by implementing a recommender system (RS) onto the website. With RS, users can receive recommendations of the products based on various mediums and mechanisms. For example, providing a “Similar Products You May Like” section that will list out products relevant to the users’ interests.

Recommender systems (RS) have become ubiquitous in many major companies such as Netflix, Spotify, Shopee, and Lazada, as they effectively guide users to relevant products, music, movies, etc. Product recommendation systems can provide personalized recommendations, boost sales, refine browsing experiences, retain customers, enhance overall purchasing experiences, and reduce information overloading (Vaidya & Khachane, 2017). Therefore, it is imperative to implement a recommender system on the website to facilitate product discovery browsing experience for customers.

In this project, we will implement 3 different types of recommender systems for different scenarios, which are simple, content-based filtering, and collaborative filtering recommender systems. For example, content-based filtering RS would utilize product properties/attributes (e.g. similar name, category, brand, or a mixture of them) to make relevant recommendations. The system can start making relevant recommendations as soon as a user browses a few items or completes some purchases, without requiring data from other users (Upwork, 2021). Thus, it can significantly improve users' ability to find desired products with just a few clicks, saving time and transaction costs.

1.2. Objectives/Aims

Enhance Customer Retention and Loyalty: The recommendation system aims to improve customer retention and loyalty by providing personalized and relevant product recommendations that cater to customers' preferences. For example, recommending keyboards similar to the keyboard a customer bought recently. With a seamless and personalized shopping experience provided by the recommendation system, we aim to increase user satisfaction and loyalty, leading to repeat purchases, positive reviews, and higher customer retention rates.

Increase Product Discoverability: The second objective of this recommendation system is to help users discover relevant electronic products more easily. By analyzing user review behavior and product metadata, the system can provide personalized recommendations that match users' interests and preferences. This can lead to increased engagement, longer browsing sessions, and higher chances of users discovering and purchasing products they may not have found otherwise.

Boost Cross-Selling and Upselling: The recommendation system aims to increase cross-selling and upselling opportunities by suggesting related or complementary products to users based on their current or past purchase history. By leveraging collaborative filtering, content-based filtering, or hybrid approaches, the system aims to identify relevant products that users are likely to be interested in, thus driving higher average order values and increasing revenue for the e-commerce website. For example, with the content-based RS, if a customer has added a USB pen drive into the shopping cart, the RS will list out recommendations of products with similar features with the products added in the shopping cart.

1.3. Motivation

After implementing the recommendation system in our online electronic shop, we anticipate several benefits that can positively impact our business. Firstly, we expect an increase in website traffic and a longer average browsing time of each user. This can potentially reduce customer churn and improve overall engagement on our platform. Besides, we anticipate a growth in our membership base as the recommendation system can help solve problems customers face in finding suitable products. By recommending personalized products, we can increase customer satisfaction and attract more subscribers to our online electronic shop.

We also expect an increase in the number of orders as the recommendation engine will suggest products that align with customer preferences based on their browsing and purchase history. Customers are more likely to purchase items suggested by the system, leading to a boost in sales. Moreover, it is believed that the implementation of the recommendation system can enhance our online electronic shop's reputation and customer satisfaction. The system's effectiveness in recommending relevant products can lead to positive reviews and ratings, which can attract more customers to our platform and result in increased referrals from satisfied customers.

Furthermore, recommendation systems can enhance the sales of long-tail items, which are unique and niche products that typically have limited demand. Such systems enable customers to explore products beyond their immediate area and discover items that they might not have otherwise known about. On the supplier side, if they stock niche items in a warehouse, a recommendation system can increase the visibility of these under-promoted products, leading to potential profits (Požar, 2021).

In addition, RS encourages the impulse buying phenomenon. This is an added advantage compared to the normal online store. Presenting customers with a variety of products can increase the likelihood of them making an unplanned purchase. However, the potential for impulse buying in online stores without RS is limited due to their configuration, unlike physical retail stores, where major retail chains often guide customers through specific paths to ensure they explore every aisle before exiting the store (Valdata, 2022). The RS can recreate this impulse buying phenomenon by recommending similar products that customers wouldn’t have known before.

From the perspective of a content-based filtering recommender system, one of the advantages of this technique is that it does not require the profiles of other users, meaning that their preferences will not impact the recommendations provided to other users. This makes it an appropriate method when the website has a relatively small number of users, especially during the initial stages of the business when it is still growing. Furthermore, even if a new user joins the system or an existing user's preferences change, this technique has the ability to quickly adapt and update its recommendations (Paschos, 2023).

1.4. Timeline/Milestone

No	Task	Start Date	End Date
1	Start the project with the write up of the Introduction section.	27/3/2023	28/3/2023
2	Conduct study on recommender systems in general as well as content-based filtering recommender systems.	28/3/2023	29/3/2023
3	Find and finalize an appropriate dataset(s) for the product recommendation system together with the rest of the team.	1/4/2023	2/4/2023
4	Load the dataset and display the notebook's fundamental data with the teammate through Google Meet.	6/4/2023	6/4/2023
5	Perform data preprocessing together with team members. ● Extra data/columns removal ● Exploratory data analysis	10/4/2023	10/4/2023
6	Develop the prototype of the content-based filtering recommender system.	15/4/2023	1/5/2023
8	Prepare a Google Form for evaluation purposes.	2/5/2023	2/5/2023
7	Write up Methodology and Result.	25/4/2023	2/5/2023
9	Write up Discussion and Conclusion.	27/4/2023	3/5/2023
10	Submit an individual report to the group leader.	3/5/2023	5/5/2023
11	The group leader collects and finalizes all documentations and prototypes and submits them to the Google Classroom.	5/5/2023	5/5/2023

2. Research Background

2.1. Background of the applications

Recommender systems are a type of information filtering system that provide personalized recommendations to users based on their preferences, behavior, or other relevant data (Ricci et al., 2021). These systems leverage a range of techniques to efficiently filter large amounts of data, ultimately providing a smaller, targeted set of suggestions for users (Barney, 2023). They are widely used across various industries and applications, such as e-commerce, social media, music streaming, movie and TV show recommendations, and more.

Leading platforms such as Netflix and Spotify, along with e-commerce giants like Amazon, have adopted recommendation systems as a critical tool to improve their businesses as they have proven to be an effective tool for improving user engagement, driving sales, and enhancing the overall user experience (Pham, 2022). By providing relevant recommendations, these systems can help users discover new products that they may not have found otherwise, and can help increase customer loyalty and retention. E-commerce companies are well-suited to generate accurate recommendations due to their access to vast amounts of customer data and behavior.

Amazon was one of the pioneers in this area, launching their item-based collaborative filtering method in 1998. This allowed Amazon to provide personalized recommendations on an unprecedented scale for millions of customers and a catalog of millions of items, leading to a significant increase in sales (Smith & Linden, 2017). This is made possible by the recommender systems used in e-commerce that analyze user data such as browsing history, purchase history, search queries, and other relevant data to generate personalized product recommendations. For instance, search result pages often recommend items related to the user's search history.

The recommender systems use a variety of data processing techniques, with some relying on data, information retrieval, and pattern matching, while others incorporate AI and machine learning. Regardless of the methodology, the main filtering approaches used by these systems include content-based, collaborative, and hybrid filtering (Patel, Desai, & Panchal, 2017). Collaborative filtering involves analyzing the behavior and preferences of similar users to generate recommendations, while content-based filtering analyzes item features to generate recommendations similar to a user's previous interests. Hybrid approaches combine multiple techniques for more accurate and personalized recommendations.

2.2. Analysis of selected tool with any other relevant tools

Tools comparison	Remark	Jupyter Notebook	Spyder	Apache Zeppelin
Type of license and open source license	State all types of license	BSD license	MIT license	Apache license 2.0
Year founded	When is this tool being introduced?	2014	2009	2016
Founding company	Owner	Jupyter	Spyder Project Contributor	Apache Foundation
License Pricing	Compare the prices if the license is used for development and business / commercialization	No additional license pricing.	No additional license pricing.	No additional license pricing.
Supported features	What features does it offer?	Supports over 100 programming languages and has a variety of features such as interactive widgets, data visualization tools, and support for LaTeX equations. Allows in-browser editing for code, with indentation, automatic syntax highlighting, and tab completion. Allows displaying computation results using rich media representations, such as HTML, PNG, SVG, etc. Code collaboration available (Google Colab compatible).	Provide horizontal or vertical splitting. Provide integrated documentation browser. Support for multiple IPython consoles. Project support, allowing work on multiple development efforts simultaneously.	Supports common libraries and frameworks, including TensorFlow and PyTorch. Supports a variety of programming languages Supports text, equations, and visualizations in a single document Provides a pluggable notebook storage mechanism (e.g. local Git repository, S3, MongoDB, Azure).
Common applications	In what areas is this tool usually used?	Data analysis Machine learning Data Visualization	Data inspection Interactive testing Data analysis	Data ingestion Data discovery Data analytics Data visualization & Collaboration
Customer support	How the customer support is given, e.g. proprietary, online community, etc.	Online community is provided (forum)	Online customer service is provided	Online community is provided (Slack, GitHub)
Limitations	The drawbacks of the software	• The cells are run by one • No code-style correction.	• Could not load large datasets • Lacks a simple interface for	• Limited visualization capabilities • Limited

		• No IDE integration	cooperative development • Online code collaboration is not direct and requires additional setup.	collaboration features

2.3. Justify why the selected tool is suitable

Jupyter Notebook is a versatile tool that supports over 100 programming languages and offers a variety of features, such as interactive widgets, data visualization tools, and support for LaTeX equations. It is an open-source web application that allows users to create and share documents containing live code, equations, visualizations, and narrative text. These elements are important to aid the process of building a recommender system. The visualization capabilities provided by Jupyter Notebook is easier to use and richer compared to the two other tools. Thus, it is commonly used for data analysis, machine learning, scientific computing, and data visualization. In particular, it can display various types of data visualization, including bar charts, graphs, and pie charts, making it well-suited for projects that require data analysis. Additionally, Jupyter Notebook allows users to test and debug code in real-time, which makes it a great choice for developing systems quickly and efficiently. It is easy to integrate with popular machine learning libraries, such as scikit-learn, TensorFlow, and PyTorch.

For collaboration, Jupyter Notebook can be used with Google Colab to enable collaboration between team members. Meanwhile, Spyder requires additional setup for collaboration and the developer experience is subpar. For Apache Zeppelin, it does have some collaboration features, but it may not be as robust as other tools like Jupyter Notebook.

As it allows us to prototype and test their algorithms quickly and easily, it is a great tool for building recommender systems. With Jupyter Notebook, we can write and execute code in real-time, which allows them to experiment with different approaches and fine-tune the algorithms until the desired results are achieved. Besides, the user base of Jupyter Notebook is much larger than the other two alternatives. Hence, it is easy to access online resources or seek assistance from the community whenever there is a question or problem that requires support.

3. Methodology

3.1. Description of dataset

The dataset used for this project is the **review** and **product metadata** for the **electronics** category adapted from Amazon Review Data 2018 (Jianmo, n.d.). The dataset could be also accessed through Kaggle (Saurav, 2022) or Recommender Systems and Personalization Datasets (Julian, n.d.).

The dataset consists of two files which are user_ratings.csv and electronic_products.json. For my module, I have used the electronic_products.json file.

electronic_products.json

- Consists of the metadata of electronic products.
- Data dictionary:

Variables	Data Types	Description
category	string	List of categories that the product belongs to.
tech1	string	The first technical detail table of the product.
description	string	The description of the product.
fit	string	-
title	string	The name of the product.
also_buy	string	A product list that indicates the user who bought the product also bought the list of products.
tech2	string	The second technical detail table of the product.
brand	string	The brand name for the product.
feature	string	Bullet-point format features of the product.
rank	string	Sales rank information that indicates how the product is selling compared to other items within the same category.
also_view	string	A product list that indicates the user who viewed the product also viewed the list of products.
main_cat	string	The main category assigned for the product.
similar_item	string	HTML code that shows the similar product.

date	string	The date that the product is added.
price	string	The price of the product.
asin	string	The unique identifier that represents the product.
imageURL	string	The URL of the product image.
imageURLHighRes	string	The URL of the product image with a higher resolution.
details	string	The relevant details for the product.

3.2. Applications of the algorithm(s)

The algorithm of the recommendation system implemented as the individual module for this project will be the content-based filtering. This technique generates recommendations using similarities in content features. In this project, I have implemented the recommendations based on the 5 types of attributes listed below:

1. “product_name” only
2. “brand_or_author” only
3. “main_category” only
4. “category_tags” only
5. “ensemble” which is a mixture of all of the 4 attributes above

Feature Engineering

Because of the RAM memory limitations, the selected attributes/columns cannot be particularly long. Moreover, selected attributes should contain valuable and essential information in the form of text we need for the content-based filtering recommender. The description and the feature column could be a potential candidate but the strings are too long and they are not really standardized across the dataset. Thus, in this case, they are not reliable enough to be used for our recommender system. The total number of records are also trimmed down to 30,000 records instead of the original 74,434 records due to RAM limitations of the machine used.

Feature Extraction Techniques & Similarity Measures

For each type of recommendations, they are implemented using 2 types of feature extraction techniques, or text vectorizers respectively, which are TF-IDF Vectorizer and CountVectorizer. There are many other feature extraction techniques available, such as HashingVectorizer, Word2Vec, Doc2Vec, and many more. However, the scope of this module will only focus on TF-IDF Vectorizer and CountVectorizer.

$$\cos(\mathbf{a},\mathbf{b}) = \frac{\vec{\mathbf{a}}\cdot\vec{\mathbf{b}}}{\|\vec{\mathbf{a}}\|\|\vec{\mathbf{b}}\|} = \frac{\sum_{i=1}^{n}\mathbf{a}_i\mathbf{b}_i}{\sqrt{\sum_{i=1}^{n}(\mathbf{a}_i)^2}\sqrt{\sum_{i=1}^{n}(\mathbf{b}_i)^2}}$$

Figure 3.2a: Cosine Similarity

Then, the similarity metric used for determining how similar a vector is to a given vector is the cosine similarity. There are other popular types of similarity measures, including Pearson correlation coefficient, Jaccard similarity coefficient, Euclidean distance, Manhattan distance, and Minkowski distance (Polamuri, 2015). In this module, we will use this measure for both text vectorizers. Cosine similarity is calculated as the cosine angle between the vectors as shown in Figure 3.2a above (Roy, 2020). To be more specific, the dot product of two multidimensional vectors is divided by the product of their magnitudes to yield this result.

While CountVectorizer applies a bag-of-words method to count the frequency of words in a document, TF-IDF Vectorizer considers not only the frequency a word appears in a document, but also how important that word is to the entire corpus (Kaplan, 2022). TF-IDF Vectorizer will penalize (reduce weight) the words that appear commonly across the entire corpus, lowering the count of these words since they tend to have lower significance. Hence, they may produce different recommendations given the same input because of the different approaches to transform text data into numerical representations. In the program, once the dataset is processed and cleansed, the system can start accepting a product name input and generate recommendations using either methods.

Using TF-IDF Vectorizer

For the implementation of TF-IDF Vectorizer, the first step is to construct the required TF-IDF matrix by initializing the TfidfVectorizer, then fitting and transforming the text. Next, the cosine similarity is computed to calculate the numeric value that denotes the similarity between 2 products based on the TF-IDF matrix. For code reusability, the process of calculating the cosine similarity is abstracted into a function that returns the calculated cosine similarity matrix based on the given attribute.

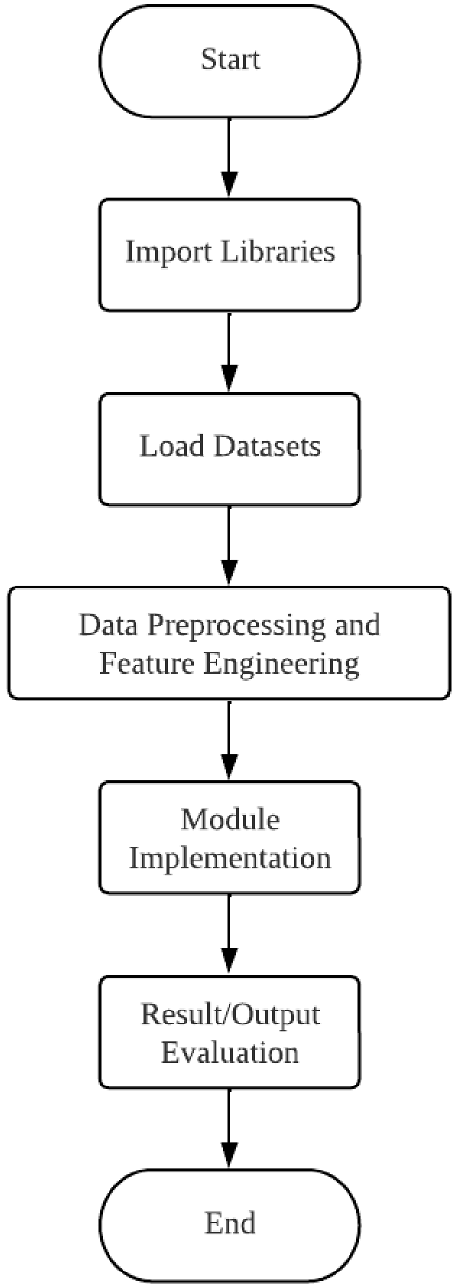
After that, a 1-dimensional array of the product names is built using pd.Series. It is built for getting the index (idx) of the product that matches with the product name inputted by the user. Using this index, we can get the pairwise similarity scores of all products with that product using list(enumerate(cosine_sim[idx])). The list of pairwise similarity scores is then sorted in descending order. It is noteworthy to mention that the user entered product is also included in the list. Therefore, the list item of that product is excluded from the pairwise similarity scores list for better result representation. Next, as the scores are now sorted in descending order, we can extract and get the N most similar products. Similarly, this process of getting the top N product recommendations is abstracted as a function that accepts 3 parameters which are the product name, number of top similar products to get, and the cosine similarity. The function will be reused for different types of attributes. We will utilize the 2 functions mentioned above to get the recommendations for 5 different types of attributes selected with minimum lines of code.

Using CountVectorizer

The second feature extraction technique implemented is the CountVectorizer. The implementation is almost similar to the implementation for TF-IDF Vectorizer. First, the CountVectorizer is initialized and fits into the given attribute. The matrix is then constructed by using the vectorizer's transform method based on the given attribute. With the matrix, we can calculate the cosine similarity. A function called get_cosine_sim_and_matrix_cv(attribute) is defined for this process.

Moving on, by building a 1-dimensional array of indices of product names, we can obtain the index of the product that matches the product name input from the user. We then find out what features have been considered by the vectorizer for a given product name by squeezing activity matrix into an array using np.squeeze(title_matrix[title_id].toarray()), and get the index as long as the feature_array is larger than 0. With that, the cosine similarity of the entered product with other products can be computed by using np.flip(np.argsort(cosine_sim[name_idx,]), axis = 0)[0:n+1] and cosine_sim[name_idx, top_n_idx]. Next, we will find out the top N indices of products with similarity scores larger than 0. The user entered product will also be excluded from the result. Finally, the final recommendation result is computed. A function is also defined for this process for reusability. Again, we will make use of the two aforementioned functions to generate recommendations for 5 different attribute types with minimal lines of code.

3.3. System flowchart/activity diagram



3.4. Proposed test plan/hypothesis

After implementing the content-based filtering recommender system, the results from both the TF-IDF Vectorizer method as well as the CountVectorizer method will be obtained by entering the product name test data.

Test Plan

To evaluate the performance of the prototype of our content-based filtering recommender system, we will conduct a **user feedback survey** using Google Form. By conducting this survey, we can understand the level of relevancy of the recommendations produced by the RS.

- The survey form will be distributed to 10 respondents to answer.
- In this module, I will be using a consistent test data product name input of "Sony CFD-C1000 Compact Stereo System" for the purpose of comparison and evaluation (for the 2 hypotheses stated below).
- For the Google Form survey distributed to 10 different respondents, each respondent will be required to answer how many products are relevant out of the 10 recommended products. There will be a total of 10 questions, which consists of 5 sections for different types of attributes, with each section asking 2 questions for TF-IDF Vectorizer and CountVectorizer respectively (5 x 2 = 10).
- Section 4.2 contains the survey questions and their corresponding responses.

Hypotheses

The specific hypotheses in this module are:

- **Hypothesis 1**: The recommendations based on the mixture of different columns which is the "ensemble" attribute produces better results compared to the individual attributes.
- **Hypothesis 2**: The TfidfVectorizer technique produces better results compared to the CountVectorizer technique.

- For **hypothesis 1**, we postulate that the "ensemble" attribute, which is using a mixture of different columns to generate recommendations, will produce better results than using individual attributes separately. This means that when multiple attributes are combined, they may provide a more comprehensive representation of the item's characteristics and features, resulting in more accurate recommendations.
- For **hypothesis 2**, we postulate that the TF-IDF Vectorizer technique will produce better results than its CountVectorizer counterpart. TF-IDF Vectorizer takes into account the importance of words in a document, which may result in more accurate recommendations since it gives more weight to important words. In contrast, CountVectorizer simply counts the frequency of words in a document, which may result in less accurate recommendations since it does not consider the importance of the words.

4. Result

4.1. Results

4.1.1 Recommendation of Top 10 Products Based on “product_name”

Using TF-IDF Vectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_id	product_name	brand_or_author	main_category	category_tags
10580	B000066IUL	sony cfd-s500 portable am/fm/cd/cassette boombox	sony	home audio & theater	electronics portable audio & video boomboxes
25730	B0001LXQIG	sony cfd-f10 cd / radio / cassette recorder, s...	sony	home audio & theater	electronics portable audio & video boomboxes
10594	B000066JQP	sony cfd-s200 cd radio cassette recorder boombox	sony	home audio & theater	electronics portable audio & video boomboxes
11204	B000068EAO	sony cfd-g55 cd/cassette boombox (black)	sony	home audio & theater	electronics portable audio & video boomboxes
21736	B0000BZL05	avenger c1000 drop-ceiling scissor clamp	avenger	camera & photo	electronics camera & photo lighting & studio l...
25731	B0001LXQIQ	sony cfd-g500 cd radio cassette recorder boomb...	sony	home audio & theater	electronics portable audio & video boomboxes
16896	B00008WTBV	sony cfd-s250 cd/radio/cassette boombox (silver)	sony	home audio & theater	electronics portable audio & video boomboxes
17170	B0000918VW	sony cfd-s550 cd/radio/cassette boombox (silver)	sony	home audio & theater	electronics portable audio & video boomboxes
25729	B0001LXQJ0	sony cfd-s300 cd radio cassette recorder boomb...	sony	home audio & theater	electronics portable audio & video boomboxes
7850	B00005I9S1	sony mhcrg20 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...

Using CountVectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_name	brand_or_author	main_category	category_tags	score
0	sony mhc-rg70av compact stereo system	sony	home audio & theater	electronics home audio compact radios & stereo...	0.600000
1	sony mhcrg20 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.547723
2	sony mhc-mg510av compact stereo system (discon...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.507093
3	sony cmt-md1 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.507093
4	sony mhc-bx5 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.507093
5	sony mhc-m300av compact stereo system (discont...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.507093
6	sony mhc-bx3 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.507093
7	sony mhc-rxd5 compact stereo system (discontin...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.507093
8	jvc mxj300 compact stereo system	jvc	home audio & theater	electronics home audio compact radios & stereo...	0.447214
9	aiwa nsxds11 compact stereo system	aiwa	home audio & theater	electronics home audio compact radios & stereo...	0.447214

4.1.2 Recommendation of Top 10 Products Based on “brand_or_author”

Using TF-IDF Vectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_id	product_name	brand_or_author	main_category	category_tags
863	7505405403	sony vgp-ac19v27 / vgp-ac19v19 19.5v 3.9a ac a...	sony	computers	electronics computers & accessories laptop acc...
923	846130053X	audio component equalizers	sony	home audio & theater	electronics home audio home theater equalizers
2345	B00000DMA3	sony mdr-nc20 noise canceling headphones with ...	sony	home audio & theater	electronics headphones over-ear headphones
2347	B00000DMA4	sony des51 sport discman portable cd player (y...	sony	home audio & theater	electronics portable audio & video portable cd...
2348	B00000DM9W	sony icf-s79v weather band shower radio (disco...	sony	home audio & theater	electronics portable audio & video radios weat...
2596	B00000J48J	sony bcv615 camcorder/mavica battery charger f...	sony	camera & photo	electronics camera & photo accessories batteri...
2664	B00000J4J8	sony 10c90hfr 90 min hifi 10 pack (discontinue...	sony	home audio & theater	electronics accessories & supplies blank media
2668	B00000J4JL	sony 60 minute dvc premium chipless (single)	sony	camera & photo	electronics camera & photo accessories blank v...
2705	B00000J4VS	sony t-160 vhs video cassette (single)	sony	home audio & theater	electronics camera & photo accessories blank v...
2710	B00000J4JI	sony high grade 120 min video cassette	sony	home audio & theater	electronics camera & photo accessories blank v...

Using CountVectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_name	brand_or_author	main_category	category_tags	score
0	sony dsr-pdx10 professional 1/4.7" 16:9 3ccd d...	sony	camera & photo	electronics camera & photo video camcorders	1.0
1	sony cdp-e210 17" crt monitor	sony	computers	electronics computers & accessories monitors	1.0
2	sony sdm-n50 15" lcd monitor	sony	computers	electronics computers & accessories monitors	1.0
3	sony vmc20fr av cable	sony	camera & photo	electronics camera & photo accessories cables ...	1.0
4	sony lcsva1 soft carrying case	sony	camera & photo	electronics camera & photo bags & cases camcor...	1.0
5	sony sdm-m51 15" lcd monitor	sony	computers	electronics computers & accessories monitors	1.0
6	sony vclhg2030 telephoto conversion lens for d...	sony	camera & photo	electronics camera & photo accessories camcord...	1.0
7	sony cdp-e500 21" crt monitor	sony	computers	electronics computers & accessories monitors	1.0
8	sony hvlfdh4 video flash light for the dcrhc40...	sony	camera & photo	electronics camera & photo accessories	1.0
9	sony lcmtrvy semi soft camcorder carrying case	sony	camera & photo	electronics camera & photo bags & cases camcor...	1.0
10	sony ecmmsd1 stereo zoom camcorder microphone ...	sony	camera & photo	electronics camera & photo accessories profess...	1.0

4.1.3 Recommendation of Top 10 Products Based on “main_category”

Using TF-IDF Vectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_id	product_name	brand_or_author	main_category	category_tags
8	0132492776	wireless bluetooth headphones earbuds with mic...	enter the arena	home audio & theater	electronics headphones earbud headphones
15	0303532572	tdk hi8 mp120 premium performance camcorder vi...	tdk electronics corp	home audio & theater	electronics camera & photo accessories blank v...
16	0302643370	the lord of the rings: return of the king - ex...	wb	home audio & theater	electronics accessories & supplies
43	0545811295	i survived 10 book library		home audio & theater	electronics accessories & supplies
70	0594478162	official nook audio ie250 earphones	nook	home audio & theater	electronics headphones earbud headphones
84	0594451647	barnes & noble hdtv adapter kit for nook hd an...	barnes & noble	home audio & theater	electronics computers & accessories tablet acc...
92	0594514851	barnes & noble nook oliver cover, black	barnes and noble	home audio & theater	electronics ebook readers & accessories covers
100	0594514843	barnes & noble fits 6" tablet / e-reader nook ...	barnes & noble	home audio & theater	electronics ebook readers & accessories
113	0594514886	nook simple touch industriell stripe stand cov...	barnes & noble	home audio & theater	electronics ebook readers & accessories covers
130	0684873176	the mutineer: rants, ravings, and missives fro...	enter the arena	home audio & theater	electronics headphones earbud headphones

Using CountVectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_name	brand_or_author	main_category	category_tags	score
0	tdk systems cdrw 80min 700mb 12x hi-speed spin...	tdk systems	home audio & theater	electronics accessories & supplies blank media...	1.0
1	panasonic rphc100 noise cancelling monitor hea...	panasonic	home audio & theater	electronics headphones	1.0
2	juke jam 40gb mp3 player	digital networks	home audio & theater	electronics portable audio & video mp3 & mp4 p...	1.0
3	philips cdr765 audio compact disc recorder	philips	home audio & theater	electronics home audio stereo system component...	1.0
4	audiosource psw100 100 watt 10-inch powered su...	audiosource	home audio & theater	electronics home audio speakers subwoofers	1.0
5	sony 8cm double-sided dvd-r with hangtab - sin...	sony	home audio & theater	electronics accessories & supplies blank media...	1.0
6	atlantic cd storage rack (elf464c47)	atlantic	home audio & theater	electronics accessories & supplies audio & vid...	1.0
7	verbatim digitalmovie dvd-r recordable media 4...	verbatim	home audio & theater	electronics accessories & supplies blank media...	1.0
8	memorex mks2451 karaoke system with echo (disc...	memorex	home audio & theater	electronics portable audio & video cassette pl...	1.0
9	koss ur29 stereo headphone	koss	home audio & theater	electronics headphones	1.0
10	creativelabs nomad muvo 64mb w (73pd028000009)	creative labs	home audio & theater	electronics accessories & supplies	1.0

4.1.4 Recommendation of Top 10 Products Based on "category_tags"

Using TF-IDF Vectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_id	product_name	brand_or_author	main_category	category_tags
2378	B00000J06D	rca rs1249 compact stereo system (discontinued...	rca	home audio & theater	electronics home audio compact radios & stereo...
2537	B00000J3NC	sharp mdm3 mini-disc/cd compact stereo system	sharp	home audio & theater	electronics home audio compact radios & stereo...
2548	B00000J3VK	panasonic sc-ak27 compact stereo system (disco...	panasonic	home audio & theater	electronics home audio compact radios & stereo...
2555	B00000J3TX	sharp cd-c3900 compact stereo system	sharp	home audio & theater	electronics home audio compact radios & stereo...
2564	B00000J3P0	fisher dcs-da300 executive microsystem	sanyo	home audio & theater	electronics home audio compact radios & stereo...
2662	B00000J4G0	fisher slim-1500 executive microsystem (discon...	sanyo	home audio & theater	electronics home audio compact radios & stereo...
2666	B00000J4GD	gpx s7799 3-cd compact stereo system (disconti...	gpx	home audio & theater	electronics home audio compact radios & stereo...
2766	B00000JBSN	jvc fs-7000 executive microsystem (discontinue...	jvc	home audio & theater	electronics home audio compact radios & stereo...
2898	B00000JFJR	jbl harmony cd personal music system	jbl	home audio & theater	electronics home audio compact radios & stereo...
2912	B00000JFEK	sony dhcmd-500 minidisc executive microsystem	sony	home audio & theater	electronics home audio compact radios & stereo...

Using CountVectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_name	brand_or_author	main_category	category_tags	score
0	panasonic sc-ak320s 300-watt 5-cd shelf system...	panasonic	home audio & theater	electronics home audio compact radios & stereo...	1.0
1	panasonic sc-dk2 5-dvd home theater compact st...	panasonic	home audio & theater	electronics home audio compact radios & stereo...	1.0
2	bose lifestyle 50 home theater system (black) ...	bose	home audio & theater	electronics home audio compact radios & stereo...	1.0
3	philips fw-r8 cd recorder executive microsystem	philips	home audio & theater	electronics home audio compact radios & stereo...	1.0
4	philips fwr7 compact stereo system with integr...	philips	home audio & theater	electronics home audio compact radios & stereo...	1.0
5	sony cmt-ep707 3-disc cd desktop microsystem (...	sony	home audio & theater	electronics home audio compact radios & stereo...	1.0
6	philips fw-d5 dvd home theater compact stereo ...	philips	home audio & theater	electronics home audio compact radios & stereo...	1.0
7	philips mz3 executive microsystem (discontinue...	philips	home audio & theater	electronics home audio compact radios & stereo...	1.0
8	panasonic sc-pm22 executive microsystem (disco...	panasonic	home audio & theater	electronics home audio compact radios & stereo...	1.0
9	jvc fs-h100 micro audio system	jvc	home audio & theater	electronics home audio compact radios & stereo...	1.0

4.1.5 Recommendation of Top 10 Products Based on “ensemble” (mixture of 4 individual attributes)

Using TF-IDF Vectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_id	product_name	brand_or_author	main_category	category_tags
7850	B00005I9S1	sony mhcrg20 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...
9069	B00005T387	sony mhc-rg70av compact stereo system	sony	home audio & theater	electronics home audio compact radios & stereo...
3412	B00001XDXK	sony mhc-rxd10av home theater compact stereo s...	sony	home audio & theater	electronics home audio compact radios & stereo...
3504	B00001ZWT0	sony mhc-rxd5 compact stereo system (discontin...	sony	home audio & theater	electronics home audio compact radios & stereo...
5036	B00004VVUH	sony mhc-m300av compact stereo system (discont...	sony	home audio & theater	electronics home audio compact radios & stereo...
5040	B00004VVUF	sony mhc-bx5 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...
6561	B000050YFN	sony mhc-bx3 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...
7860	B00005I9RY	sony mhc-mg510av compact stereo system (discon...	sony	home audio & theater	electronics home audio compact radios & stereo...
3413	B00001XDXR	sony cmt-md1 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...
26984	B00022RV8A	sony cmt-cpx22 compact stereo micro audio syst...	sony	home audio & theater	electronics home audio compact radios & stereo...

Using CountVectorizer

User input of product name: Sony CFD-C1000 Compact Stereo System

	product_name	brand_or_author	main_category	category_tags	score
0	sony mhc-rg70av compact stereo system	sony	home audio & theater	electronics home audio compact radios & stereo...	0.928571
1	sony mhcrg20 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.912421
2	sony mhc-m300av compact stereo system (discont...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.897085
3	sony mhc-mg510av compact stereo system (discon...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.897085
4	sony cmt-md1 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.897085
5	sony mhc-bx5 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.897085
6	sony mhc-bx3 compact stereo system (discontinu...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.897085
7	sony mhc-rxd5 compact stereo system (discontin...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.897085
8	sony mhc-rxd10av home theater compact stereo s...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.889052
9	sony cmt-cpx22 compact stereo micro audio syst...	sony	home audio & theater	electronics home audio compact radios & stereo...	0.881917

4.2. Discussion/Interpretation

4.2.1 Results of the Google Form survey

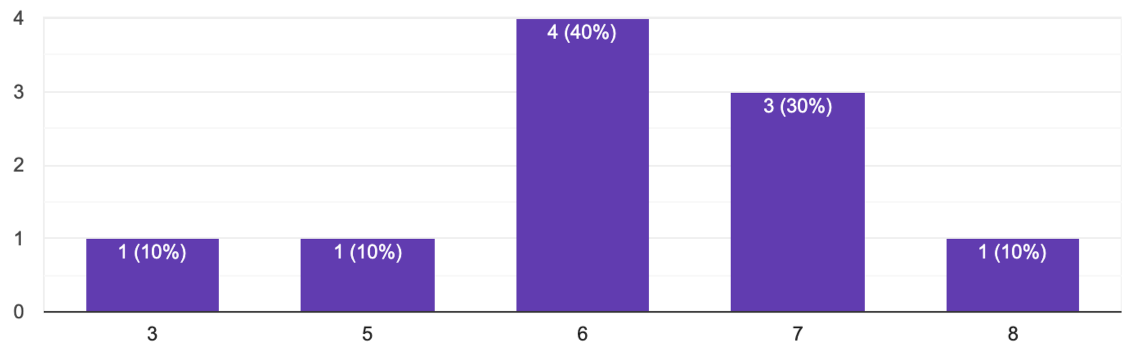
Recommendation of Products Based on Product Name.

Copy

Below is the recommendation result of top 10 products with similar **product name** with **"Sony CFD-C1000 Compact Stereo System"** based on **TF-IDF Vectorizer**.

How many products are relevant to you?

10 responses

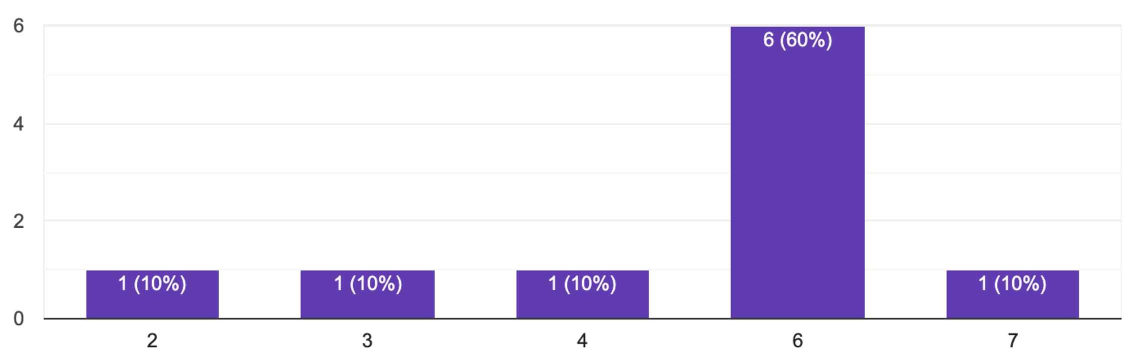


Copy

Below is the recommendation result of top 10 products with similar **product name** with **"Sony CFD-C1000 Compact Stereo System"** based on **CountVectorizer**.

How many products are relevant to you?

10 responses



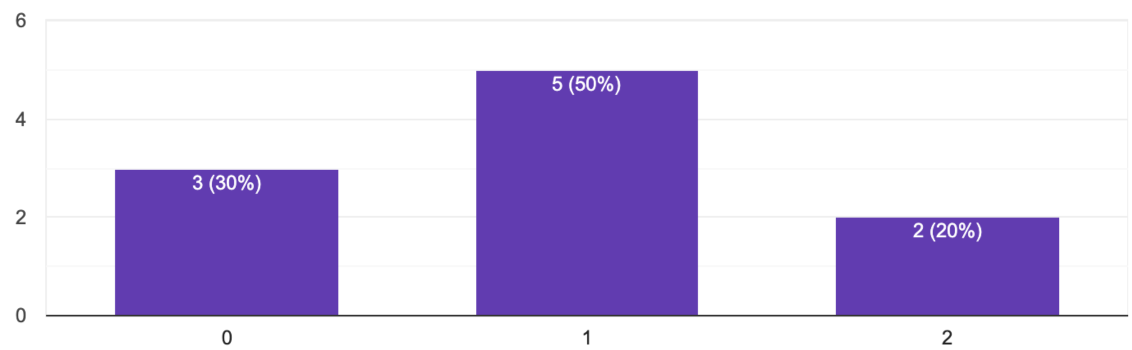
Recommendation of Products Based on Brand or Author.

Copy

Below is the recommendation result of top 10 products with similar **brand or author** with **"Sony CFD-C1000 Compact Stereo System"** based on **TF-IDF Vectorizer**.

How many products are relevant to you?

10 responses

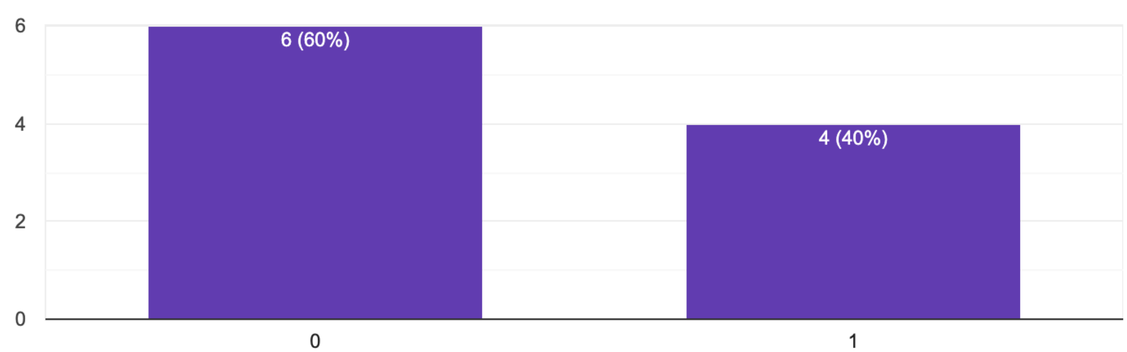


Copy

Below is the recommendation result of top 10 products with similar **brand or author** with **"Sony CFD-C1000 Compact Stereo System"** based on **CountVectorizer**.

How many products are relevant to you?

10 responses



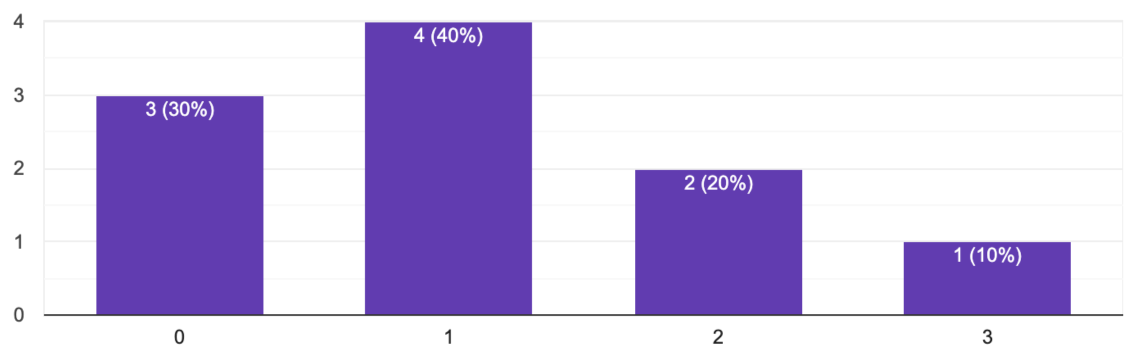
Recommendation of Products Based on Main Category.

Copy

Below is the recommendation result of top 10 products with similar **main category** with **"Sony CFD-C1000 Compact Stereo System"** based on **TF-IDF Vectorizer**.

How many products are relevant to you?

10 responses

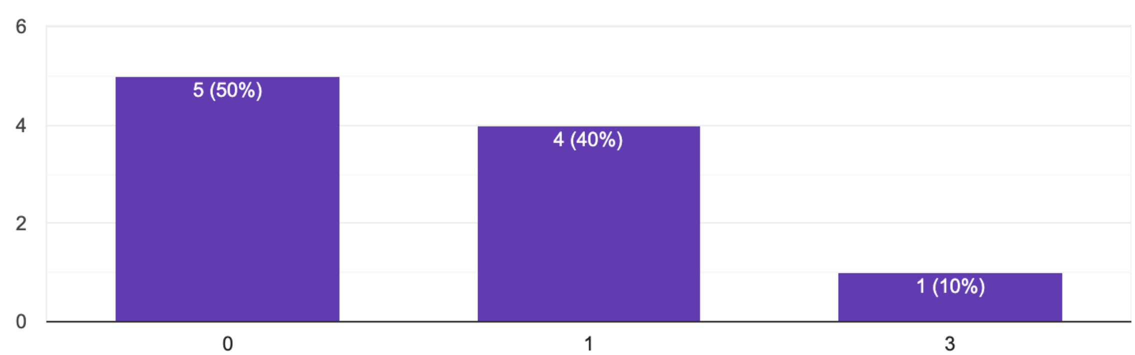


Copy

Below is the recommendation result of top 10 products with similar **main category** with **"Sony CFD-C1000 Compact Stereo System"** based on **CountVectorizer**.

How many products are relevant to you?

10 responses



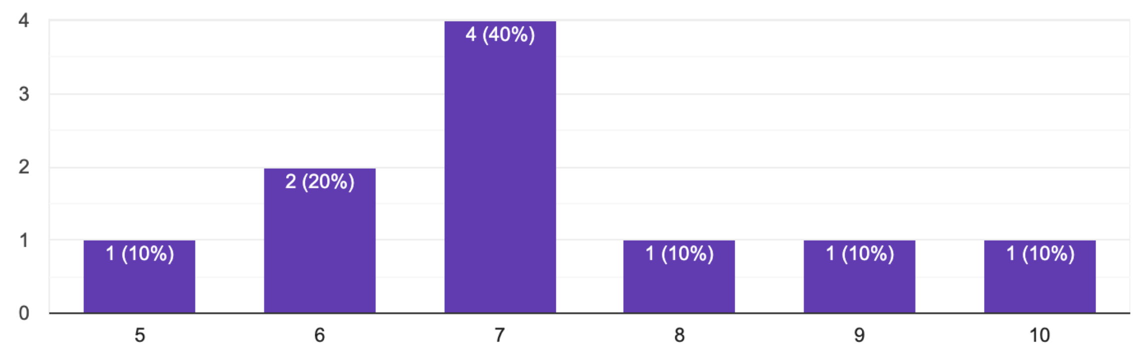
Recommendation of Products Based on Category Tags.

Copy

Below is the recommendation result of top 10 products with similar **category tags** with **"Sony CFD-C1000 Compact Stereo System"** based on **TF-IDF Vectorizer**.

How many products are relevant to you?

10 responses

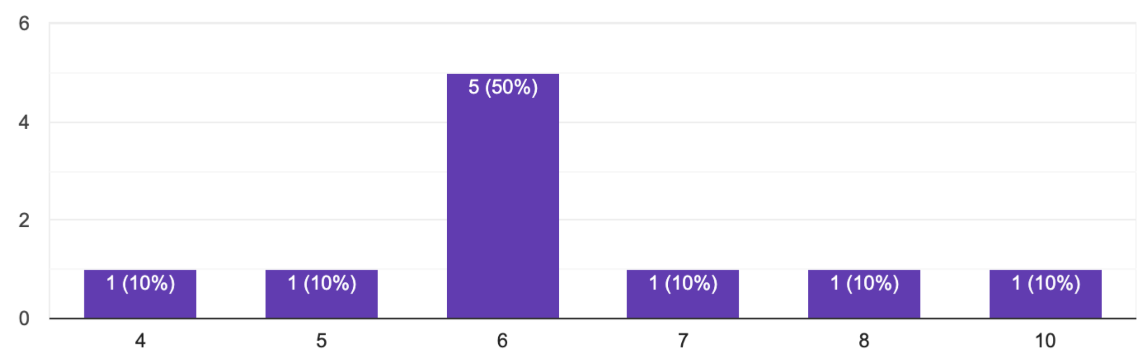


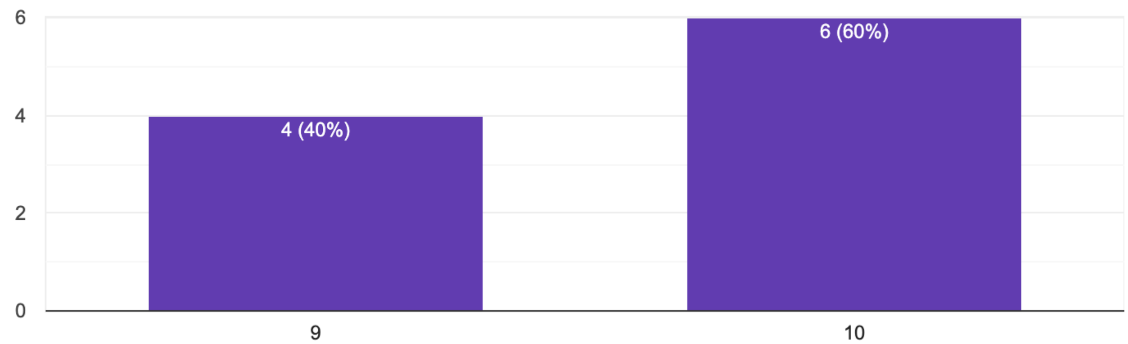
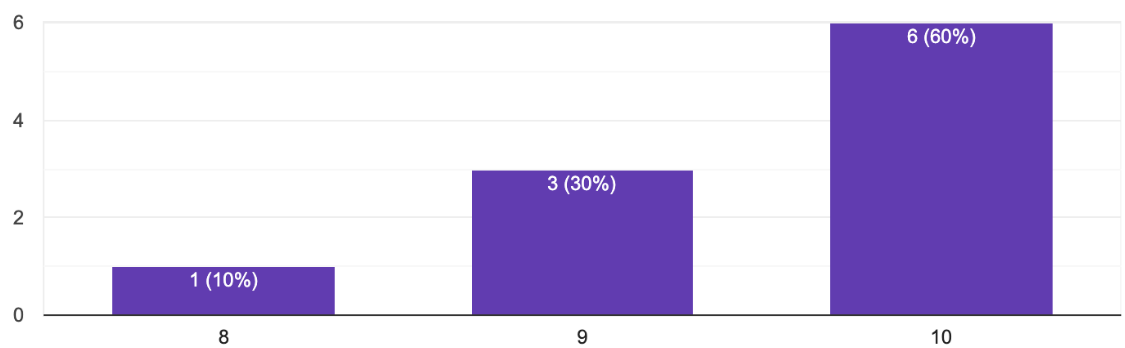
Copy

Below is the recommendation result of top 10 products with similar **category tags** with **"Sony CFD-C1000 Compact Stereo System"** based on **CountVectorizer**.

How many products are relevant to you?

10 responses



Recommendation of Products Based on the Mixture of the 4 Attributes.
Copy Below is the recommendation result of top 10 products with similar **mixture of the 4 attributes** with **"Sony CFD-C1000 Compact Stereo System"** based on **TF-IDF Vectorizer**. How many products are relevant to you? 10 responses 
Copy Below is the recommendation result of top 10 products with similar **mixture of the 4 attributes** with **"Sony CFD-C1000 Compact Stereo System"** based on **CountVectorizer**. How many products are relevant to you? 10 responses 

4.2.2 Summary of the results

The table below lists the average number of recommended items that are considered as relevant based on the survey responses for the 5 types of attributes.

Table 4.2.2: Precisions

	TF-IDF Vectorizer	**CountVectorizer**
Product Name	(3+5+24+21+8)/100 = 0.61 **= 61%**	(2+3+4+6*6+7)/100 = 0.52 **= 52%**
Brand or Author	(0*3+5+2*2)/100 = 0.09 **= 9%**	(0*6+1*4)/100 = 0.04 **= 4%**
Main Category	(0*3+4+2*2+1)/100 = 0.09 **= 9%**	(0*5+4+3*1)/100 = 0.07 **= 7%**
Category Tags	(5+6*2+4*7+8+9+10)/100 = 0.72 **= 72%**	(4+5+5*6+7+8+10)/100 = 0.64 **= 64%**
Mixture of Product Name + Brand or Author + Main Category + Category Tags	(4*9+6*10)/100 = 0.96 **= 96%**	(8+3*9+6*10)/100 = 0.95 **= 95%**

4.2.3 Interpretation of the results

- **Interpretation 1**: It can be observed that the precision for the recommendations based on "ensemble" for both TF-IDF Vectorizer and CountVectorizer are almost the same. This may be due to the result produced by both TF-IDF Vectorizer and CountVectorizer being almost the same.
  - It can also be indicated that combining the several attributes produces very stable and relevant recommendations for the users.
- **Interpretation 2**: It is observed that the precision for TF-IDF Vectorizer for all 5 types of attributes is higher than the CountVectorizer counterpart.
  - It can be indicated that the approach used by TF-IDF Vectorizer to analyze the text produces better results for recommendation.

Review Hypotheses

- *Hypothesis 1: The recommendations based on the mixture of different columns which is the "ensemble" attribute produces better results compared to the individual attributes.*
- *Hypothesis 2: The TfidfVectorizer technique produces better results compared to the CountVectorizer technique.*

Based on the results, it can be said that hypothesis 1 and hypothesis 2 are true.

5. Discussion and Conclusion

5.1. Achievements

For this module of content-based filtering recommender system, it has successfully produced a functioning RS which will recommend **top relevant N products** similar to a particular product based on various attributes of that product. It can be useful especially to the new users who have not yet made any purchases on the platform. In addition, the content-based filtering approach used in this RS allows it to provide relevant product recommendations **even without any explicit user data**. By analyzing the attributes of each product, such as product title, description, and category, the RS can identify products that are similar to the one being viewed or searched for. This can be especially helpful when the website is new without much data on the user and their profile/preferences. In other words, the content-based approach can solve the cold-start problem that may occur in the collaborative filtering approach.

Besides, this module has implemented the recommendations using different types of feature extraction techniques which are **TF-IDF Vectorizer and CountVectorizer** for comparison and informing the decisions on choosing which method to use for the final deployment on the e-commerce website.

For the objectives, firstly, it has proved to enhance the **customer retention and loyalty** by providing relevant product recommendations that cater to customers' preferences based on the content of the products. For example, recommending keyboards similar to the keyboard a customer bought recently. When the RS can be integrated with the website, it can expand the user base of MyElectronic by recommending electronic products that are of interest to the customers.

Next, the second objective of our recommendation system, which was to **increase product discoverability** for customers, is proved to be achieved. By analyzing product metadata such as product name, category tags, brand, and main category, the RS has been able to provide personalized and relevant recommendations that match the interests and preferences of users. As a result, we can expect longer browsing sessions, and higher chances of users discovering and purchasing products that they may not have found otherwise.

The developed prototype of the content-based filtering recommender system has also successfully achieved the objective of **boosting cross-selling and upselling** opportunities for the e-commerce website. By identifying relevant products that users are likely to be interested in, the system effectively increased the average order value and revenue for the website. For instance, the content-based RS can later be implemented on the website to list out recommendations of products with similar features as the products added to the shopping cart, thus enhancing the user's shopping experience and increasing the likelihood of purchasing additional items.

5.2. Limitations and Future Works

The limitations of this recommendation system is that the prototype is contrived and is **not able to fully simulate real-life explicit and implicit data collection** to make recommendations in an e-commerce website. For example, observing the items that a user views in an online store and keeping a record of the items that a user purchases online. For future improvement, we can consider integrating our recommender system onto the website of MyEletronic to better demonstrate how a recommender system works in a real website. For example, when a user views the products of earphones and laptops frequently, we can use this content-based recommender system to get relevant products of earphones and phones that are similar to the ones viewed by the user.

Besides, the datasets that we used only have about **70k product records which is relatively small** in the context of e-commerce industry where there is generally an enormous and wide variety of data to work with and analyze. This limited amount of data can lead to various disadvantages in developing a recommender system, such as insufficient representation of item attributes, which may result in inaccurate recommendations and poor performance. In addition, the datasets used in this project are based on Amazon's data from 2018. For future improvement, the recommender system's performance can be improved by incorporating more recent and diverse datasets from different sources to increase the variety and quality of data. Additionally, using more advanced techniques, such as deep learning, can enhance the system's ability to extract meaningful features from the data and improve the accuracy and relevancy of the recommendations.

Moreover, another limitation is that a **user's preference could change** over time. This can affect the accuracy of predictions especially if the user did not browse our websites for a long time. For future improvement, we can use machine learning techniques to continuously learn, update, and predict the user's preferences based on their interactions with the system, such as their feedback on recommended products.

Another limitation is that content-based recommender systems can become **less effective when dealing with new or unfamiliar items**. This is known as the cold-start problem, where the system may not have enough information about a new item to make accurate recommendations. Defining and tagging the attributes of a new product, service, or content is necessary for incorporating it into the recommender system. However, this process is time-consuming and can make scalability challenging as it requires continuous effort to assign attributes to new items. This issue can be mitigated by using a hybrid approach that combines both content-based and collaborative filtering methods to provide better recommendations, even for new or underrepresented items. Hence, it is worth developing a hybrid recommender system in the future to further improve our recommender system.

Last but not least, the current content-based recommender system only recommends products based on the attributes of product name, brand, main category, category tags, and a mixture of these attributes. For future improvement, it is imperative for our content-based filtering recommender system to recommend products with **product images using image processing**. Image processing can also help identify attributes of a product that may not be captured in its textual description, such as color or pattern. With image-based recommendations, the system can suggest products based on visual similarities of the products, which can be particularly useful for items with complex attributes that are difficult to describe textually. For example, a user can upload an image of a foldable tablet, and the system is able to analyze the image attributes of the tablets and generate recommendations of similar foldable tablets. By including image processing in the recommendation process, the system can provide a more comprehensive and personalized user experience.

Reference & Source

Sources of Datasets

- https://www.kaggle.com/datasets/saurav9786/amazon-product-reviews
- https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

Development Tools

- Jupyter Notebook - web-based interactive computing platform
- Flask - backend web framework for Python
- Next.js - frontend framework based on React

References

Barney, N. (2023) *What is a recommendation engine?, TechTarget*. Available at: https://www.techtarget.com/whatis/definition/recommendation-engine (Accessed: May 3, 2023).

Jianmo, N. (n.d.) 'Amazon review data (2018)', *cseweb.ucsd.edu*. Available at: https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/ (Accessed: 23 April 2023).

Julian, M. (n.d.) 'Recommender Systems and Personalization Datasets', *cseweb.ucsd.edu*. Available at: https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews (Accessed: 23 April 2023).

Kaplan, D. (2022) *Machine Learning 101: CountVectorizer Vs TFIDFVectorizer, EML.* Available at: https://enjoymachinelearning.com/blog/countvectorizer-vs-tfidfvectorizer/ (Accessed: May 5, 2023).

Paschos, A. (2023) *The importance of recommendation systems nowadays, Qymatix*. Available at: https://qymatix.de/en/recommendation-systems-nowadays/ (Accessed: May 4, 2023).

Patel, B., Desai, P. and Panchal, U. (2017) 'Methods of recommender system: A Review', *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* [Preprint]. doi:10.1109/iciiecs.2017.8275856.

Pham, K. (2022) *What are Recommendation Systems?, Medium*. Available at: https://medium.com/@khang.pham.exxact/what-are-recommendation-systems-6bb5036042db (Accessed: May 3, 2023).

Polamuri, S. (2015) *Five most popular similarity measures implementation in Python, Dataaspirant.* Available at: https://dataaspirant.com/five-most-popular-similarity-measures-implementation-in-python (Accessed: May 5, 2023).

Požar, N. (2021) *Recommendation systems in ecommerce: How it works?, BE-terna.* Available at:

https://www.be-terna.com/insights/recommendation-systems-in-e-commerce-whats-the-thing-youve-never-known-but-always-wanted-to (Accessed: May 4, 2023).

Ricci, F., Rokach, L. and Shapira, B. (2021) 'Recommender systems: Techniques, applications, and challenges', *Recommender Systems Handbook*, pp. 1–35. doi:10.1007/978-1-0716-2197-4_1.

Roy, A. (2020) *Introduction To Recommender Systems- 1: Content-Based Filtering And Collaborative Filtering, Medium.* Towards Data Science. Available at: https://towardsdatascience.com/introduction-to-recommender-systems-1-971bd274f421 (Accessed: May 5, 2023).

Saurav, A. (2022) 'Amazon Product Reviews', kaggle. Available at: https://www.kaggle.com/datasets/saurav9786/amazon-product-reviews (Accessed: 23 April 2023).

Smith, B. and Linden, G. (2017) "Two Decades of Recommender Systems at Amazon.com," *IEEE INTERNET COMPUTING* [Preprint].

Upwork (2021) *What Content-Based Filtering is and Why You Should Use It*. Available at: https://www.upwork.com/resources/what-is-content-based-filtering (Accessed: May 3, 2023).

Vaidya, N. and Khachane, A.R. (2017) "Recommender Systems-the need of the ecommerce era," *2017 International Conference on Computing Methodologies and Communication (ICCMC)* [Preprint]. Available at: https://doi.org/10.1109/iccmc.2017.8282616.

Valdata, G. (2022) *Building a recommender system for Amazon products with python, Medium.* Towards Data Science. Available at: https://towardsdatascience.com/building-a-recommender-system-for-amazon-products-with-python-8e0010ec772c (Accessed: May 5, 2023).