University of New Haven

Data Mining

CSCI-6401-01 (Spring 2023)

Final Report

GOOGLE MAPS RECOMMENDATION SYSTEM

Submitted by: **Submitted to:**

Team Amigos **Dr. Shivanjali Khare**

Jayanth Reddy Manda (00764288) jmand5@unh.newhaven.edu

Snigdha Reddy Yeruva (00762316) syeru2@unh.newhaven.edu

Jeevan Kumar Konduru (00806735) Jkond6@unh.newhaven.edu

Abstract

It is intriguing to learn about the use of cosine similarity, tf-idf, and dot product in a content-based recommendation system. This approach involves analyzing the content of items that a user has interacted with, such as articles or videos, to identify their preferences and interests.

This content-based recommendation system has demonstrated improved performance in providing relevant recommendations to users based on their interactions with items. In the future, the system may consider incorporating additional features, such as clustering the data based on Features, to further personalize recommendations and improve the user experience.

Contents

Introduction & Background

Content-based recommendation systems analyze the properties or features of items and then make personalized recommendations to users based on their individual preferences and behavior. For example, if a user has previously rated several coffee shops highly, the system can use those ratings to recommend other coffee shops with similar characteristics. In this way, the system can provide relevant and personalized recommendations to the user based on their interests and behavior.

To build a content-based recommendation system for Google Maps, machine learning algorithms can be trained on large datasets of user interactions and item features. These algorithms can then be used to analyze new data and make recommendations about which places a user is likely to enjoy based on their individual preferences and behavior. For example, the system may recommend a coffee shop that has a similar atmosphere, price range, and cuisine type to a coffee shop that the user has previously rated highly.

One potential benefit of content-based recommendation systems is that they can provide personalized recommendations even when little is available about the user's social network or demographic. This can be useful in contexts where users may not want to share personal information or where there is limited data available. Additionally, content-based recommendations can be generated quickly and easily, making them a useful tool for improving the user experience on Google Maps.

Overall, the content-based recommendation system for Google Maps is an exciting application of machine learning and data analysis and has the potential to provide significant value to both users and businesses. By helping users find places that match their interests and preferences, the system can increase engagement and loyalty to the platform, while also driving traffic and revenue to local businesses.

Related work:

Review 1:

Title of Paper: CONTENT-BASED RECOMMENDATION USING MACHINE LEARNING

Author Name: Y. Tai, Z. Sun and Z. Yao

Year of Publication: 2021

Citation: Y. Tai, Z. Sun and Z. Yao, "Content-Based Recommendation Using Machine Learning," 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), Gold Coast, Australia, 2021, pp. 1-4, doi: 10.1109/MLSP52302.2021.9596525.

Weblink

Research paper focuses on developing a recommender system in shopping or restaurant etc. by helping user, the best product or most used sales items in their respective category is to reduce user costs by retrieving the most relevant information from massive amounts of data, and then the system can provide personalized services. System categorized into different sections based on available data and by using different algorithms. Research paper has been trained on 200000 training sample datasets with an accuracy 77% on Support Vector Machine Algorithm, Logistic regression with an accuracy 76%.

1. Purchase item prediction based on logistic regression.

   Predicting an item range which is the best time and price to buy a product and what is the best price that a customer is willing to buy a product and the appropriate time to buy the product.

2. Purchase category prediction is made based on Support Vector Machine(SVM)

   Predicting the estimation of purchase on multiple products so that companies can make best deal regarding product purchase cost (If users are willing to buy beverages in a range of 20$ to 100$ then the shopkeeper can make the best suitable products in a certain range).

3. User's rating prediction is made based on Convolutional Neural Network(CNN) and Long Short-Term Memory (LSTM).to guide the user about product or store etc...It makes the new user trustworthy and confident to purchase a product.

Review 2:

Title of paper: Crop Recommendation System using Machine Learning Algorithms

Author Name: G. Chauhan and A. Chaudhary

Publication date: 2021

Publisher: IEEE

Citation:

G. Chauhan and A. Chaudhary, "Crop Recommendation System using Machine Learning Algorithms," 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), MORADABAD, India, 2021, pp. 109-112, doi: 10.1109/SMART52563.2021.9676210.

https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/9676210

The paper presents the ability to recommend a system for Indian agriculture using the machine learning technique Random Forest and Decision Tree, using multiple factors such as Ph, N, K, Weather conditions, water availability, etc. All the factors are integrated together and are collected from multiple sources. So that prediction tells us how much availability of ph., K, Water, and which crop is best.

1) Grouping the farmers based on Decision Tree

Classifying the farmers based on their farming techniques, crops they grow throughout the year, farmer's chemical usage type, etc. These individual features can help us to classify which crop is better for the particular season using the decision tree.

2) Classify the crop so that it can be used for commercial usage or daily essential needs based on demand and supply.

Classifying the crop so that it can be used for commercial use now based on global climate conditions and demand for the product in a particular farming technique type. Such as usage of its previous demand, supply, production level, quality of product, etc...

Review 3:

Title of paper: Predictions of Diabetes and Diet Recommendation System for Diabetic Patients Using Machine Learning Techniques.

Author Name: S. S. Bhat and G. A. Ansari

Year: 2021

Citation: S. S. Bhat and G. A. Ansari, "Predictions of Diabetes and Diet Recommendation System for Diabetic Patients using Machine Learning Techniques," 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India, 2021, pp. 1-5, Doi: 10.1109/INCET51464.2021.9456365.

The paper focuses on predicting diabetes in different age people based on their day-to-day activities such as type of food, job, living area, age, gender, etc.... the paper uses multiple data mining techniques naive Bayes, Decision trees, and Random Forest. The paper outputs make whether the person is suffering from diabetes or not diabetic using all the characteristics of a person. With an accuracy of 87% using a decision tree and 90% accuracy by using naive Bayes. The diabetic prediction level is made consideration to recommend diet control to the patient if the level of the diabetic is good there is no need for usage of any diet control, if not the recommended diet should be used by the patient.

Review 4:

Title of Paper: A Review of Academic Recommendation Systems Based on Intellignet Recommendation Algorithms.

Author Name: Huaiyuan yang

Year of Publication: 2022

Citation: H. Yang, H. Zhou and Y. Li, "A Review of Academic Recommendation Systems Based on Intelligent Recommendation Algorithms," 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi'an, China, 2022, pp. 958-962, doi: 10.1109/ICIVC55077.2022.9886104.

The paper focuses on the comparison of traditional recommendation systems with new machine learning techniques such as deep learning with a better ability to recommend an article for the user based on the given input where the paper uses a content-based Filtering system, collaborative Filtering, Hybrid Recommendation system for traditional Recommendation System. The traditional algorithm uses the data of Author name, title, description, year, etc... to recommend an article to the user where all the features are calculated using tf-idf in a document. whereas in deep learning techniques, it uses auxiliary information for model-based deep learning recommendation system and dynamic deep learning recommendation system. The model-based deep learning uses the user's latent representation by learning user's item rating matrix. Dynamic deep learning uses some of the dynamic features such as user interests, and recent content for a particular article.

Review 5:

Title: Research on an Agent-Based Intelligent Social Tagging Recommendation System.

Author: Shihu An; Hong Zhou

Year: 2017

Citation: S. An, Z. Zhao and H. Zhou, "Research on an Agent-Based Intelligent Social Tagging Recommendation System," 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 2017, pp. 43-46, doi: 10.1109/IHMSC.2017.17.

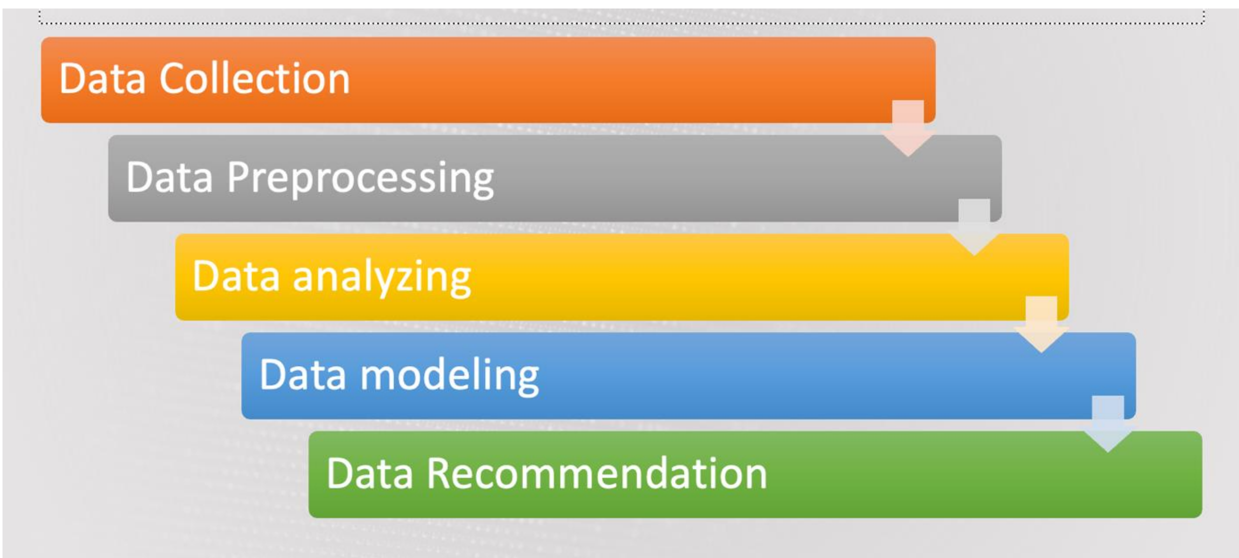
The development of effective and precise resource recommendation systems is necessary given the quick increase of social tagging users. An agent-based intelligent social tagging system architecture is suggested to meet this demand; it incorporates user interest mining, customized suggestions, and recommendations from common preference groups. The fundamental conflict between personalized and common preference recommendation models is then resolved by introducing a self-adaptive recommendation technique to balance the trade-off between accuracy and efficiency.

The Proposed Method:

Algorithms Used in Proposed Methods:

1) **Content-based recommendation System using dot-product.**
2) **Content-based recommendation system using cosine similarity.**
3) **Content-based recommendation system using Tf-idf.**
4) **K-means clustering**

The Proposed method is divided into multiple Phases as shown below image.



Data Collection:

Google Maps doesn't provide any maps dataset in an Excel or CSV format all the data of google maps is stored in google cloud which can be accessed through Google Cloud API. In this paper, we will be using Google Maps Places API to get the places in certain latitudes and longitudes. The places may include such as stores, museums, gas stations, etc.. all these places are taken into consideration if they are in a particular diameter range for the given latitude and longitude. we used Python to retrieve API data with an API key and password provided for the individual account. The API data is not appropriate data that can be used for a recommendation system.

The following attributes have been retrieved from the API.

1) Geometry: The attribute shows the latitude and longitude coordinates with the direction of phasing.
2) Icon: The attribute describes icon of the business representation(I;e logo of a company)
3) Icon_background_color: Background color of an icon.
4) Icon_mask_base_url: Base URL for icon image
5) Name: Describes the store name.
6) Photos: Contains all the photos belonging to store.
7) Place_id: Identity provided by google Maps to the store such as id.
8) Scope: Describes the recognition of a store based on its identity found by Google, safari, Microsoft edge, etc..

9) Types: One of the most important attributes to recommend to a user, the attribute that describes the nature of the store such as a café, restaurant, bar, etc...
10) Vicinity: A human-readable location to identify the store.
11) Business_status: It describes whether the business is currently in operational or not operational.
12) Plus_code: google maps identity for a store in particular location.
13) Rating: The average rating of a store from all the users.
14) User_ratings_total: The total number of ratings provided to the particular store.
15) Opening_hours: Describes operational hours of a store throughout the day.
16) Price_level: Describes the price level of a store by the end users based on their previous experience.
17) Permanently_closed : Finding whether the particular store is available right now or not.

Data Preprocessing:

Data pre-processing is one of the most important steps to be performed for a dataset as it is not exactly used for the recommendation technique.

The following techniques are used to pre-process the data.

1) As the API data consists of a few null values for most of the columns few columns can be neglected such as “premanently_closed” as all the operational stores don’t have the status of the code.
2) Some of the “price_level” rows have null values by considering the mean of all other fields of the column, we updated all the.

Data Analyzing:

As most of the data has been cleaned and is available to use now we will be analyzing how the data has been distributed across different features, price_levels, rating and user_rating_Total. The following statistical techniques and visualization techniques describe the way how data is presented.

Frequency of Features:

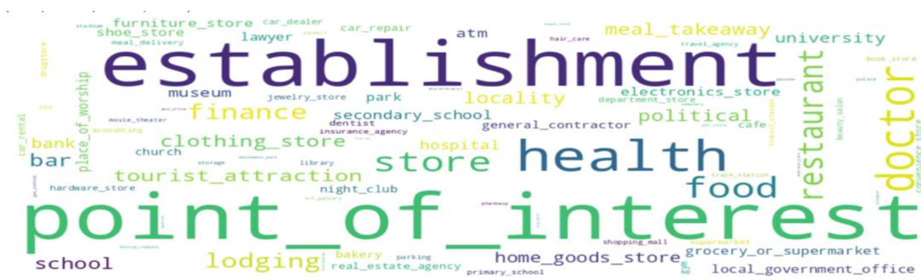


Figure 1:

The above image depicts the number of occurrences of a word through all the data.

“establishment”, “pont_of_interet”, ”health” are the 3 most occurring words in the dataset.

Frequency of price level:

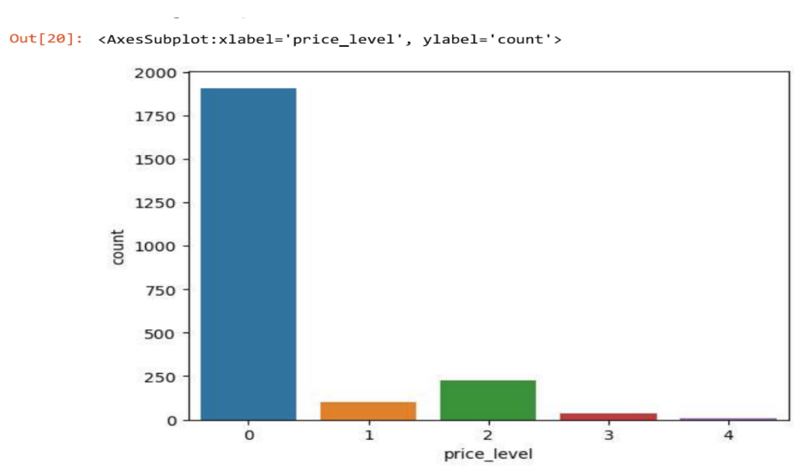


Figure 2:

We use Bar graph to find the frequency of “price_level” across the all the stores data, Figure 2 depicts most of the “price_level” of stores belong to 0, which indicates 60% of stores in the dataset are most likely with good “price_level” to the users. The “price_level” 2 has the second most stores in the dataset.

Box-Plot ”user_ratings_total” vs “price_level”

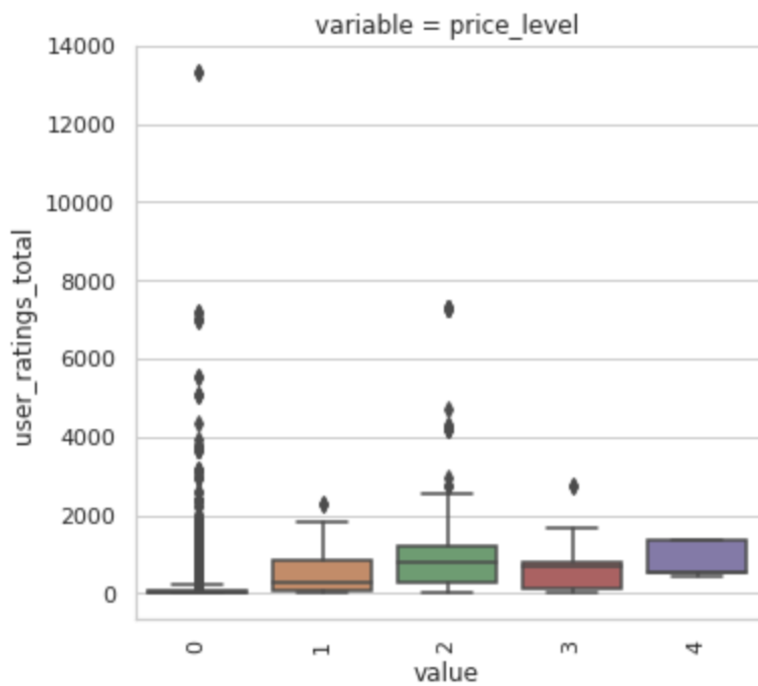


Figure 3:

Figure 3 depicts the box plot of “price_level” to “user_ratings_total” It quickly summarizes the range, median, and Quartiles of a dataset.

Data Modeling & Recommendation:

1) Gather place characteristics: Gather the characteristics of the locations you want to suggest. For cafes, for instance, you might consider factors like cuisine type, price range, ambiance, location, etc. In the case of hospitals, characteristics may include the facility's kind, services provided, user-location distance, etc.
2) making a user profile Based on the user's prior experiences with the system, such as the kinds of locations they have looked for, visited, or reviewed, create a profile for the user. You can also infer their preferences using implicit feedback, such as location information.
3) Calculate place similarity: Using the dot product, determine how similar each location in the system is to the user profile. To do this, you must express each location and the user profile as a vector of characteristics. The similarity score is then calculated by taking the dot product of the user profile vector and the location vector. The more closely the location matches the user's choices, the better the score.
4) Recommended locations Sort the locations based on similarity scores, then provide the user recommendations for the top N locations. To further hone the recommendations, you may also apply other filtering criteria, such as the user's location, the location's accessibility, the location's hours of operation, etc.
5) Finally, we will multiply the weight vector User Profile by the one-hot encoded Features of google maps DataFrame using the dot product algorithm. In essence, this gives each Feature a weight dependent on how significant it is to User. The weighted average of each video is then calculated by dividing the result by the total of the weights in user Profile.

6) A recommendation table that was produced based on his choices; Data Frame assigns each Store a weighted score that reflects how much the user would enjoy it. The higher the rating, the more likely user is to like the film.

7) Overall, the weighted average approach is a well-liked method used in recommender systems to produce customized recommendations for users.

Results:

store_Id	user_id	name	vicinity	Features
373	1395	West Haven	West Haven	[accounting, finance, point_of_interest, estab...
1772	1079	Hilton Garden Inn Milford	291 Old Gate Lane, Milford	[accounting, finance, local_government_office,...
172	1115	Hyatt Place Milford / New Haven	190 Old Gate Lane, Milford	[accounting, local_government_office, finance,...
563	1404	Courtyard by Marriott New Haven Orange/Milford	136 Marsh Hill Road, Orange	[accounting, finance, point_of_interest, estab...
877	1402	Hampton Inn Milford	129 Plains Road, Milford	[accounting, finance, point_of_interest, estab...
1378	1124	Super 8 by Wyndham Milford/New Haven	1015 Boston Post Road, Milford	[accounting, finance, local_government_office,...
1837	1319	Zumiez	1201 Boston Post Road Suite 2008, Milford	[accounting, finance, point_of_interest, estab...
2109	1327	Connecticut Post Mall	1201 Boston Post Road, Milford	[accounting, finance, point_of_interest, estab...
224	1356	AT&T Store	1201 Boston Post Road Space 2445, Milford	[accounting, local_government_office, finance,...
2174	1110	ALDO	1201 Boston Post Road #2414, Milford	[accounting, finance, point_of_interest, estab...
1539	1368	Hollister Co.	1201 Boston Post Road, Milford	[airport, point_of_interest, establishment]

Figure 4: Recommended stores for the user who likes West Haven Store using dot-product.

Content-Based Recommendation System using Cosine Similarity:

1) Cosine similarity is like the dot product, but the way of converting features changes, in dot product we will be rounded to 1's or 0's which is closer to them, but in cosine similarity the features are absolute.
2) Data collection Collect information about the stores and businesses we want to recommend, and then collect all the features, ratings and total ratings from the business.
3) Making Store Profiles Create a feature vector from the features of each item, where each dimension corresponds to a feature and the value denotes how significant the feature is for the store. Each item gets a profile as a result, which may be used to gauge how closely a store matches the user's interests.
4) Making user profiles Create a user profile by expressing a user's preferences as a feature vector, where each dimension denotes a feature or characteristic and the value denotes how important the feature is to the user. As a result, the system can evaluate how well the user's tastes and the store profiles match up.
5) Cosine similarity calculation calculates the cosine similarity between each store's profile and the user profile, then rank the things according to the scores. The user suggests the products that are most comparable to their tastes and have the highest similarity ratings.

```
Recommended stores based on West Haven are:
3      Courtyard by Marriott New Haven Orange/Milford
4                                 Hampton Inn Milford
6                                              Zumiez
7                               Connecticut Post Mall
9                                                ALDO
1                           Hilton Garden Inn Milford
2                     Hyatt Place Milford / New Haven
5                 Super 8 by Wyndham Milford/New Haven
8                                           AT&T Store
918                               Art Museum at USJ
Name: name, dtype: object
```

Figure 5: Recommended Stores using cosine similarity to the user who likes West Haven Store.

Content-Based Recommendation system using Tf-idf:

1) Tf-idf is like cosine similarity where in cosine similarity we will be converting all the features will be into vectors and find the cosine angle between them each vector.
2) In Tf-Idf we use the term frequency-inverse document frequency (Tf-idf) vectorization to convert the text data into a numerical representation. Based on how frequently a word appears in a document and how uncommon it is across all documents, Tf-idf calculates the value of each word in the corpus of texts. As a result, a sparse matrix is created, where each row denotes a document (location), and each column denotes a distinct word within the corpus.
3) Now use cosine similarity to compare the Tf-idf vectors of each place and find the most similar places based on their vector similarity scores.
4) After finding cosine similarity for the vectors now we will calculate cosine similarity between each store and user profile, then rank things according to scores.

Results:

```
Recommended stores based on West Haven are:
1                          Hilton Garden Inn Milford
9                                               ALDO
5                Super 8 by Wyndham Milford/New Haven
6                                             Zumiez
7                              Connecticut Post Mall
8                                          AT&T Store
2                     Hyatt Place Milford / New Haven
10                                    Hollister Co.
3     Courtyard by Marriott New Haven Orange/Milford
4                                Hampton Inn Milford
Name: name, dtype: object
```

Figure 6: Top Recommendations for the User who Likes “West Haven”.

K-means Clustering:

K-means clustering can also be used for a google maps recommendation system to group all the similar stores based on their characteristics such as Features, rating, price level, etc.

1) We will be classifying all the datasets into 5 clusters based on their ratings, price level, and user_rating_total.
2) Data samples with similar features will be grouped into one cluster based on their Euclidean distance to the centroid of clusters.
3) Once the stores have been clustered now we are able to recommend some of the stores to the user by comparing clusters with user preferences.

Results:

```
  warnings.warn(
Cluster 0: ['Dr. Ralph P. Stocker, MD' 'The Klein Memorial Auditorium' 'GRM Music']
Cluster 1: ['West Haven' 'Dr. Pietro A. Memmo, MD' 'Cigna' 'Greenwood Lee H MD'
 'Ocean Community YMCA - Naik Family Branch Mystic']
Cluster 2: ['RE/MAX Coast and Country' 'Greenwood Lee H MD' 'Toads Place']
Cluster 3: ['American Medical Response' 'City Point Historic District']
Cluster 4: ['West Haven' 'Hilton Garden Inn Milford'
 'Hyatt Place Milford / New Haven'
 'Courtyard by Marriott New Haven Orange/Milford' 'Hampton Inn Milford']
```

Figure 7: The above figure describes 5 clusters with their similarities in Features, rating, price level, etc.

Silhouette Score:

We will. Be using Silhouette score is a measure of how similar an object is to its own cluster compared to other clusters.

The silhouette score for a single object is calculated as the difference between the average distance to other objects in the same cluster (a) and the average distance to objects in the nearest neighboring cluster (b), divided by the maximum of the two distances:

silhouette score = (b - a) / max(a, b)

The silhouette score ranges from -1 to 1, where a score of 1 indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters, and a score of -1 indicates the opposite. A score of 0 indicates that the object is on the boundary between two clusters.

The overall silhouette score for a clustering algorithm is the average of the silhouette scores for all objects in the dataset, and it provides a measure of how well-separated the clusters are and how appropriate the number of clusters is for the data. A higher silhouette score indicates better clustering performance.

Figure 8: Silhouette score for the clusters based on dataset.

Result Comparison:

Algorithm	Precision	Recall
Content-Based Recommendation System using dot-product.	85%	75%
Content-Based Recommendation system using cosine similarity.	88%	78%
Content- Based Recommendation system using Tf-idf.	89%	80%

Discussion:

The above-listed dot-product, cosine similarity, and Tf-idf all them will recommend a user based recommendations with the help of features, ratings, and user_rating_total, the recommendations work on user-based experience with a frequency of each feature and ratings given to the store.

The above-listed recommendation describes the results for the single user with his/her interest in the West Haven store.

Conclusion:

We can recommend some of the best stores to the user based on a content-based recommendation system using tf-idf, cosine similarity, and dot-product. Overall, the Google Maps recommendation system has been proven.to be effective in providing users with relevant and helpful recommendations, and it continues to evolve and improve over time with the incorporation of new data fields and algorithms. Whether you are looking for a new restaurant to try or planning a trip to a new city, Google Maps can help you make informed decisions and explore the world with confidence.

Future Work:

1) The Google Maps recommendation system can be integrated with a collaborative recommendation system that integrates with other user's data Features.
2) Incorporate social features of the user's likes and dislikes of a context.
3) Usage of Artificial neural networks to the data set.

Appendix:

Reference:

Citation: S. S. Bhat and G. A. Ansari, "Predictions of Diabetes and Diet Recommendation System for Diabetic Patients using Machine Learning Techniques," 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India, 2021, pp. 1-5, Doi: 10.1109/INCET51464.2021.9456365.

F. Palma, H. Farzin, Y. -G. Guéhéneuc and N. Moha, "Recommendation system for design patterns in software development: An DPR overview," 2012 Third International Workshop on Recommendation Systems for Software Engineering (RSSE), Zurich, Switzerland, 2012, pp. 1-5, doi: 10.1109/RSSE.2012.6233399.

G. Chauhan and A. Chaudhary, "Crop Recommendation System using Machine Learning Algorithms," 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), MORADABAD, India, 2021, pp. 109-112, doi: 10.1109/SMART52563.2021.9676210.

Ilhamsyah, S. Rahmayudha, I. Fakhruzi, F. Febrianto, A. C. Adi and V. Gunawan, "Recommendation System for the Improvement of E-Government Services in the Tourism Sector of Pontianak City," 2022 International Conference on ICT for Smart Society (ICISS), Bandung, Indonesia, 2022, pp. 1-4, doi: 10.1109/ICISS55894.2022.9915103.

J. Yu et al., "Collaborative Filtering Recommendation with Fluctuations of User' Preference," 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE), Chengdu, China, 2021, pp. 222-226, doi: 10.1109/ICICSE52190.2021.9404120.

P. R. Y. Jiang, H. Zhan and Q. Zhuang, "Application research on personalized recommendation in distance education," 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, China, 2010, pp. V13-357-V13-360, doi: 10.1109/ICCASM.2010.5622798.

X. Jia and F. Liu, "Research on intelligent recommendation system model supported by data mining and algorithm optimization," 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), Chongqing, China, 2021, pp. 766-769, doi: 10.1109/ICESIT53460.2021.9696972.

Y. Adilaksa and A. Musdholifah, "Recommendation System for Elective Courses using Content-based Filtering and Weighted Cosine Similarity," 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2021, pp. 51-55, doi: 10.1109/ISRITI54043.2021.9702788.

Y. Tai, Z. Sun and Z. Yao, "Content-Based Recommendation Using Machine Learning," 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), Gold Coast, Australia, 2021, pp. 1-4, doi: 10.1109/MLSP52302.2021.9596525.

Weblink.

Z. Wang, A. Maalla and M. Liang, "Research on E-Commerce Personalized Recommendation System based on Big Data Technology," 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 2021, pp. 909-913, doi: 10.1109/ICIBA52610.2021.9687955.