Name: Haritha Poornachandran
Roll No: 2017103057

Web Programming CS7604

Restaurant Information Collection & Analysis Using Python Crawler

Objective:

To collect information such as name, cuisine, cost, rating of restaurants from popular food review apps such as **Zomato using Web Scraping in Python and apply K-Means clustering** algorithm on this information to categorize restaurants by **Value for money**.

Algorithm:

1. Import BeautifulSoup for web crawling.
2. Construct request.get () by specifying URL for the HTML Parser.
3. Declare class names and ID to pull information from and staore the information as key-value pairs in a DataFrame.
4. Use Pandas to store the DataFrame contents into a .csv file. (df.to_csv("zomato_restaurants.csv")
5. Now that Web crawling is complete proceed to use this information to perform clustering.
6. Import numpy, matplotlib, Kmeans. Take only cost and rating from CSV file and plot.
7. Construct WCSS with scaling the input and find number of clusters using elbow method.
8. Now apply KMeans function for number of clusters on scaled data.
9. Plot this new data to see grouping of data based on the Value for money ratio.

Dataset:

The data set is constructed by performing web crawling using BeautifulSoup on a food website. It is stored in CSV file. We can get information such as Average Cost, Rating, Number of Votes which are Numeric values. We also get Name, Locality, Cuisine which are strings.

The Dataset used for this project has details of 832 restaurants located in Bangalore, each containing all the attributes.

*Note: The attributes used in the IEEE paper differ from the values in this project due to **unavailability of banking transaction information** of customers for any business. Therefore, a similar method is applied to a different dataset and different output is computed.*

Sample Code:

```
#For Web Scraping
response = requests.get("https://www.zomato.com/bangalore/restaurants)
content = response.content
soup = BeautifulSoup(content, "html.parser")
search_list = soup.find_all("div", {'id': 'orig-search-list'})
list_content = search_list[0].find_all("div", {'class': 'content'})
for i in range(0,15):
    res_name = list_content[i].find("a", {'data-result-type': 'ResCard_Name'})
    locality = list_content[i].find("b")

dataframe ={}
dataframe["rest_name"] = res_name.string.replace('\n', ' ')
dataframe["locality"] = locality.string.replace('\n', ' ')
dataframe["rating"] = ratings.string.replace('\n', ' ')

df = pandas.DataFrame(list_restaurants)
df.to_csv("zomato_restaurants.csv")
```

Name: Haritha Poornachandran
Roll No: 2017103057

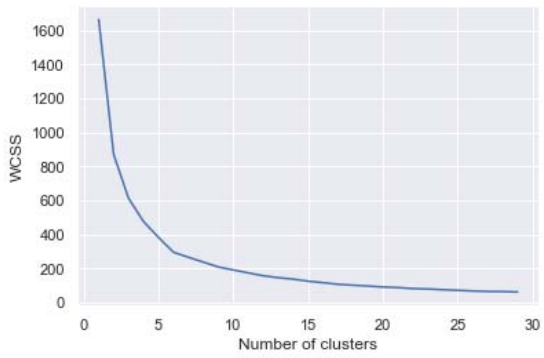
```
#FOR KMEANS
from sklearn import preprocessing
x_scaled=preprocessing.scale(x)
wcss=[]
for i in range(1, 30):
    kmeans = KMeans(i)
    kmeans.fit(x_scaled)
    wcss.append(kmeans.inertia_)
plt.plot(range(1,30),wcss)
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

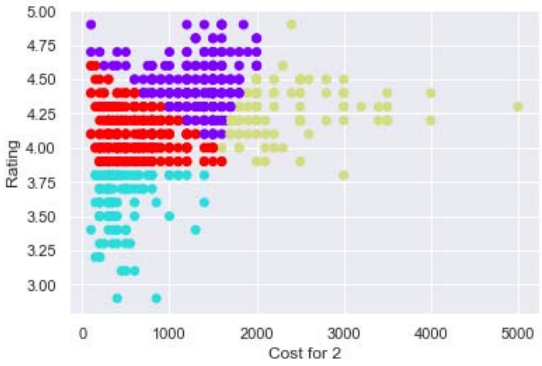
```
kmeans_new=KMeans(4)
kmeans.fit(x_scaled)
cluster_new=x.copy()
cluster_new['cluster_pred']=kmeans_new.fit_predict(x_scaled)

plt.scatter(cluster_new['cost_for_two'],cluster_new['rating'],c=cluster_new['cluster_pred'],cmap='rainbow')
plt.xlabel('Cost for 2')
plt.ylabel('Rating')
plt.show()
```

Output:

The output of this will be the optimal number of clusters (K=4 by elbow method) as well as the clustering of restaurants with most similar Cost Satisfaction ratio. These grouping can be used to label restaurants by Value for Money.

Plot:





182	800	4.1	3
183	1200	4.9	0
184	2000	4.5	2

1: sample labelling by groups

Inference:

This information can be used to label the restaurants as Valuable, Average to Low recommendation while recommending restaurants to new customers. This Information can be used to streamline and finetune the apps recommendation algorithms to ensure customer satisfaction as well as identifying and improving low performing restaurants.

Reference:

https://medium.com/code-to-express/k-means-clustering-for-beginners-using-python-from-scratch-f20e79c8ad00

Name: Haritha Poornachandran
Roll No: 2017103057

Customer Value Analysis Based on Python Crawler

Ming Liu, Yurui Du, Xifen Xu

School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China
E-mail: liuming@njust.edu.cn

Abstract: Customer value analysis is an important work in customer relationship management. In this paper, we use ABC classification, RFM (Recency, Frequency, Monetary) model and K-means clustering method to analyze the customer value. First, we use python language to compile a crawler program to collect the data of transaction records from an enterprise's customer information management system. Then, by using different methods, we obtain different classifications of the customers. The test results demonstrate that the proposed methods can provide constructive suggestions to manage the customer value in practice, although they are easy and basic.

Key Words: Python, Customer Value, ABC Classification, RFM Analysis, Clustering

1 INTRODUCTION

Under the background of big data era, the focus of corporate marketing has shifted from product-centered to customer-centered, and companies have gradually recognized the importance of customers and services [1]. Customer relationship management has become the core issue of the enterprise. Customer value classification is one of the key issues in customer relationship management, and it is an important basis for decision makers to optimize marketing resource allocation [2]. Through the results of customer classification, the managers can understand the different value attributes of the customer. Based on the analysis of the current value and potential value of the customer, we can divide the customers into different groups according to their different value. It is a critical way to achieve the ultimate goal of maximizing corporate profits by analyzing the difference value of different groups of customers, formulating personalized marketing plans, and applying customer classification results to enterprise customer relationship management practice [3].

Recently, many scholars have used different classification methods to mine and subdivide customer data sets, and provided several constructive strategies to guide enterprise practice. For example, Mosavi and Afsar (2018) focused on the Tejarat Bank branches in Iran and systematically integrated several data mining techniques and management issues to analyze customer value [4]. Daoud et al. (2015) conducted a case study of applying RFM model and clustering techniques in the sector of electronic commerce with a view to evaluating customers' values so as to achieve maximum benefit and a win-win situation [5]. Lu et al. (2018) used SPSS statistical software direct RFM point model for the customer segmentation and then used the nested classification to distinguish the customer value by RFM score to design different marketing strategies for online store customers [6]. Ge and Chen (2016) developed an operational indicator system to measure customer value and proposed several suggestions to effectively manage customer relationship [7]. Sarvari et al. (2015) determined the best approach to customer segmentation and extrapolated associated rules based on recency, frequency and monetary (RFM) considerations as well as demographic factors [8]. Chang and Ho (2017) built a two-layer clustering model for mobile telecom customer analysis and used it to combine with the product to assist staff to implement effective marketing [9]. Singh (2015) created a risk-adjusted recency, frequency, monetary value (RARFM) score for each customer to identify the under- lying demographics and behavioral characteristics [10]. Although these scholars have used relevant algorithms to classify and calculate the value of customers, they have not integrated different customer values. Therefore, this paper combines python crawler to obtain relevant data, then applies three customer segmentation algorithms to offer more instructive management strategies.

With the explosive growth of network information, it is difficult for us to get the required information quickly and accurately from the massive information. Under this background, we use python web crawler technology with its powerful ability of automatically extracting web information to program and crawl the required data of customer transaction records. After cleaning the customer transaction records, we classify the customers by using ABC classification method and RFM analysis model, and then apply K-means clustering based on the analysis results of RFM model, which is the main basis for evaluating the value of all kinds of customers. Finally, according to the test results, we propose several directory marketing strategies of customers with different value to realize the precise marketing of enterprises. As far as we know, most of the existing literature only use one related classification method for customer value segmentation. In this study, we use ABC classification, RFM model and K-means clustering to subdivide customer value on the basis of python crawler data. Based on the classification results of RFM model,

This work is supported by National Nature Science Foundation of China under Grant (No.71771120) and Ministry of Education Project of Humanities and Social Sciences under Grant (No. 17YJA630058).

978-1-7281-0106-4/19/$31.00 ©2019 IEEE