International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217284

Anatomization of Market

Anurag Dave[1], Aditya Kumar Singh[2], Ashish Kumar Singh[3], Hemant Pandey[4] , Nandeep Singh Hada[5]

[*1]Computer Science and Engineering, Lovely Professional University, Kanpur, Uttar Pradesh, India
[2]Computer Science and Engineering, Lovely Professional University, Vadodara, Gujarat, India
[3]Computer Science and Engineering, Lovely Professional University, Jalandhar, Punjab, India
[4]Computer Science and Engineering, Lovely Professional University, Gwalior, MP, India
[5]Computer Science and Engineering, Lovely Professional University, Port Blair, Andaman and Nicobar, India

Article Info
Volume 7, Issue 2
Page Number: 396-400

Publication Issue :
March-April-2021

Article History
Accepted : 15 April 2021
Published : 20 April 2021

ABSTRACT

Data mining and machine learning have become crucial and inspiring fields of work for today's researchers. Mostly in every field, there is considerable use of machine learning to make micro to massive operations more feasible and possible. And with the world witnessing the worst ever Pandemic that has affected all section of society but specifically the economic sector i.e., businesses sector and with now people relying less on human-human physical interaction and more on technological means because of obvious reasons, our project has provided a better and more reliable platform for providing the upcoming entrepreneur as well as big investors such as government and private sector to fulfill their shattered dreams by knowing best location they can start/ restart their businesses. The model is majorly based on K-mean clustering. The users can provide the details of the individual sector they are interested in to begin a business and find the best opportunity around them.

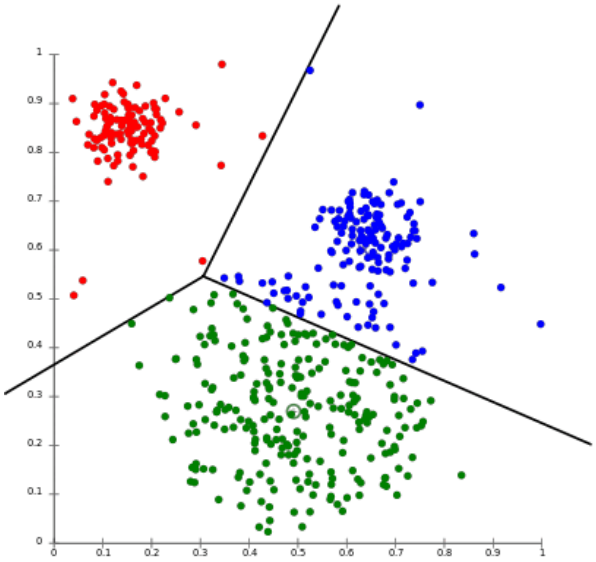
Keywords : Clustering, K-Means Clustering, API

I. INTRODUCTION

As we all know the world witnessed its nightmare at Beginning of 2020 in the name of the Coronavirus pandemic and the world was in a virtual shut down for the next few months. The government almost all over the world closed all shops, factories, and trade. This shutdown/ lockdown intensively dropped the business that was not considered online. And resulted in massive loss of job. With the lost job and reopening of all commercial sector now people are considering to be self-sufficient and in need of a platform that will provide them with the accurate data to start up their own business. Seeing the upsurge in the demand for new businesses we came up with the idea of an automated system that will analyze the data worldwide, detect, update and prompt exact location that will fit the user's demand for establishing a business.

The project is done using the python programming language. In the project, we have used the K-mean clustering algorithm intensively for grouping the

dataset. As we know online data are dynamic i.e., it keeps on changing hence we will store the dataset in a variable. As we need the data categorically, we have used the soup code package present in python. The data present in websites and the internet are mostly in HTML format and to process this data one needs to convert it into an XML type file. Therefore, soup code will also help in converting the dataset into XML format from HTML format. In the project, we have opted for 5 clusters i.e. K=5. We have taken the accurate longitude and latitude of the postal code of Toronto city and plotted it on the present-day map to provide a user-friendly and precise location of a particular business (cluster).



II. METHODS AND MATERIAL

To understand what K-mean clustering we must know first what clustering means and how the clustering algorithm work.

Clustering in simple terms is the process of dividing the datasets into groups consisting

of similar data points. An unsupervised machine learning method. Unlike supervised learning, unsupervised learning is a method of drawing references for a dataset of collection of datasets. The dataset consists of input data without labelled responses. It is used for grouping the dataset with similar characteristics, but it can be also used for finding useful structure, generative features, etc. The grouping can be done with d-dimensional i.e., 2-dimensional, 3- dimensional, etc., objects. It is a distance-based clustering.

Operation and is useful for many applications. The clustering algorithm is also sometimes called pattern discovery or knowledge because of its ability to help in the data analysis process

III. WHY CLUSTERING

Clustering plays an important role when it comes to grouping similar data that is unlabeled. It is depending on the user that what criteria he wants to use to satisfy their need that automatically widen the scope for the international market in large. A different method that one can use to form the clusters:

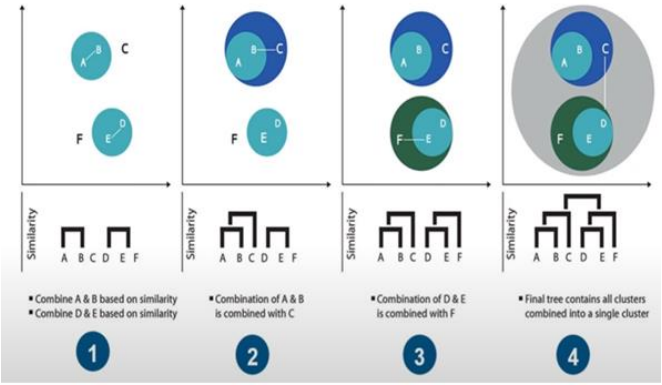
A. Density-based methods

The identification of a cluster or a group is based on the idea that the region of high point density with similar data space is separated by a dataset with low point density are considered noise or outliers. Example: DBSCAN and OPTICS.

B. Hierarchical based methods

In this the clustering is done in a tree-type structure i.e., the formation of a new cluster is dependent on the previously formed cluster.

Example: CURE, BIRCH.



C. Grid-based methods

The dataset is formed infinite no of cells that form a grid-like structure. An advantage of this methods is all the operations done is independent cell is fast.

Example: STING, CLIQUE.

D. Partitioning methods

The cluster used in our project follows this approach. The partition of the objects is

done using K-cluster. The data point in this exclusively belongs to one cluster. Example: k-means and CLARANS

E. K-mean clustering algorithms:

It is a data mining algorithm that uses a distanced-based clustering algorithm. K is the number of clusters and consists of all the input data that belong to the cluster with the nearest mean. If we apply K-Means to a set of N objects, then the result will be K disjoint groups.

How is the value of K specified?.

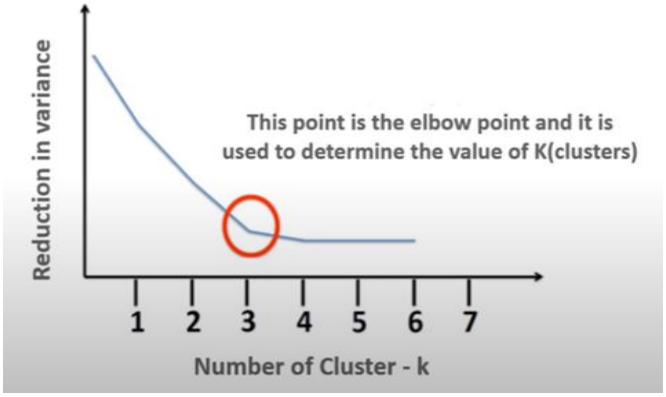
The selection of K is depended on the individual. K vale to specific to the application it is in use for. But If one is keen on knowing the optimal value of K, he can find it by hit and trial method i.e., try a different value of K. Start from smallest value, i.e., K=1 which means all the data value lies in the same cluster. It is the worst-case scenario.

The best value of K will be the one with the least variation. For instance, we take K=2, now we have to compare the variation of K=1 to the variation of k=2. We will notice the variation of the 2 clusters will be smaller than the cluster of 1. Hence it can be summarised as the cluster increase the variation decrease.

The value of the elbow point from the elbow plot will decide the value of K for the application. We can see in the graph below that there is a steep reduction in the variance up till k=3 and then there is no significant downfall. This point of change is called the elbow point.

How is the grouping done?

The simplest answer to this is it done by



minimizing the sum of the square of the distance between the data point and centroid.

Formalizing k-mean clustering:

Let, X1 = (x11 , x12 , x13 , … , x1d)

X2 = (x21 , x22 , … , x2d)

.

.

Xn = (xn1 , xn2 , … , xnd) Now we have to partition these (X1, X2, …, Xn) into K cluster (C1, C2, …, Ck)μi is the mean of points in Ci.

F. Algorithm

Step 1: divide the complete N object from the dataset into K non-empty clusters.

Step 2: calculate the centroid i.e., mean point or centre of the cluster (μi) for the partition performed in step1.

$$\boldsymbol{\mu}_i = \frac{1}{|c_i|} \sum_{j \in c_i} \mathbf{x}_j, \forall i$$

Step 3: Assign each object to the cluster with the nearest centroid.

$$c_i = \{j : d(x_j, \mu_i) \leq d(x_j, \mu_l), l \neq i, j = 1, \ldots, n\}$$

Step 4: Repeat steps 2 and 3 until no new assignment is required. (i.e., all data points are stored in the clusters).

Step 5: repeat all the step until desired output (no of cluster) is achieved.

This algorithm is an iterative algorithm and repeat until there is no change in the centroid. The membership odd each cluster will be re calculated based on the present centroid assigned to the cluster.

WHY CLUSTERING OVER CLASSIFICATION?

Classification, which is a supervised learning algorithm, is the process in which input instances are classified based on labels. Hence, there is a need of training the data before mining it. Whereas clustering is groping of the dataset according to the similarity character of it there means there is no use of labels required hence no training is required.

The next reason to use clustering over-classification is the complexity. There are many levels of classiication of dataset whereas only grouping is done in clustering that makes it easy to process and work on the data.

V. API

In layman's terms, API is software that exchanges data and functionalities using a machine-readable platform. API stands for an application programming interface.

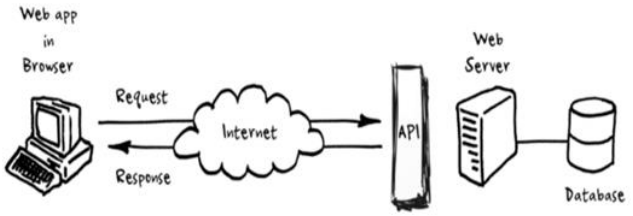
API isn't a database or server; it is a set of codes that enable the interaction of data from one system software to another.

A common example of API is e-commerce websites, online ticket booking, food

ordering, etc. one advantage of API is that there is no need for actual system access. The parent system can assign an API secret key to the child system and this secret key will help to authenticate the child system every time they attempt to interact with the parent system. API system also enables a secure interaction with the third-party system that keep the actual data safe and only share the data that is required to the third-party system. (some APIs are free from secret keys too.

Foursquare API:

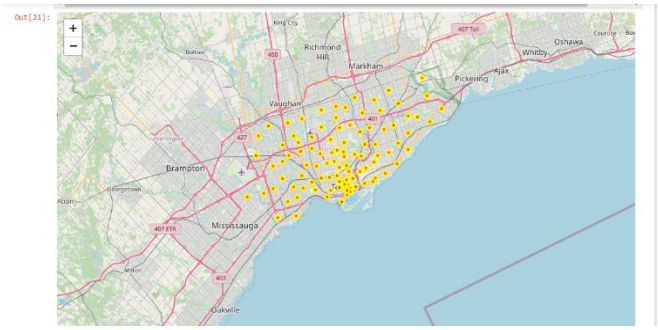
Foursquare API is a company that has built a huge dataset of accurate location data. This API is being used all around the world with more than 100000 developers. This tool has been used by many developers to develop searches, augmented reality, etc. The main objective of foursquare API is to provide the latest satellite view of earth and connect it with the developer's project to develop a new project. In our project, we have used foursquare API for accessing the latest map and overlapping it with the geolocation of postal codes of Toronto city and thereafter, with clusters of suitable business opportunity.



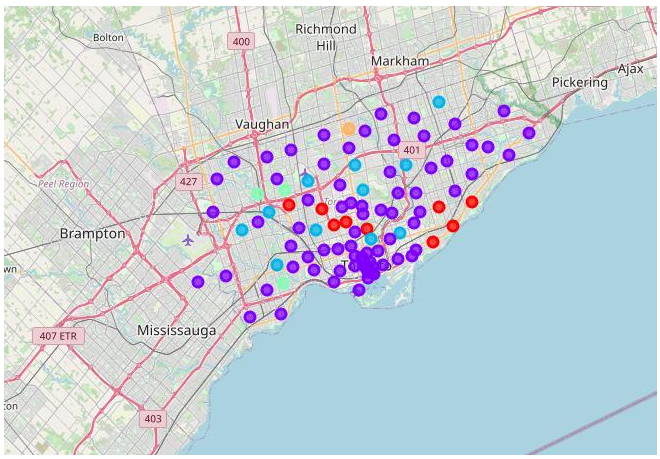
A. Figures and Tables

SNAPSHOTS OF TORONTO CITY USING FOURSQUARE API FROM THE PROJECT:

Map of Toronto:



After clustering:



IV. CONCLUSION

We have developed a futuristic approach to deal with upcoming entrepreneurs or investors who need to identify a fruitful location for their business. This project will provide a platform to the interested mind who acknowledge and admire the use of machine learning and data mining for the upcoming world. Our project has unlimited boundaries as it can be used by anyone from private players to governmental players and also have no border constrain that means can be used in any country.

We in this project were also able to explore the real-time use of K-mean clustering to make human life simpler. We have demonstrated how background information present on the internet can be used in a well-constructed manner to provide people with a great opportunity in the field of business and market. Moreover, in a nutshell, this project can fulfill India's "Atmanirbhar Bharat" (self-reliant India) program as now people can set up business in our market and boost the country's economy. We believe this project has the sky as its limit.

V. REFERENCES

[1]. Data Algorithms: Recipes for Scaling Up with Hadoop and Spark, Book by Mahmoud Parsian. Publisher: O'Reilly Media, Inc. Release Date: 15, 2015 ISBN: 9781491906187

[2]. Ramanpreet Kaur and Amandeep Kaur, Text Document Clustering and Classification using K-Means Algorithm and Neural Networks, Indian Journal of Science and Technology, Vol 9(40), DOI: 10.17485/ijst/2016/v9i40/97722, October 2016.

[3]. https://machinelearningmastery.com/clustering-algorithms-with-python/#:~:text=Clustering%20or%20cluster%20analysis%20is,customers%20based%20on%20their%20behavior.&text=Clustering%20is%20an%20unsupervised%20problem,feature%20space%20of%20input%20data.

[4]. https://www.geeksforgeeks.org/clustering-in-machine-learning/

[5]. https://www.geeksforgeeks.org/ml-classification-vs-clustering/#:~:text=Classification%20is%20used%20for%20supervised,is%20used%20for%20unsupervised%20learning.&text=As%20Classification%20have%20labels%20so,and%20testing%20dataset%20in%20clustering.

[6]. https://developer.foursquare.com/article/how-our-intern-led-pants-migration-to-python-3/

[7]. https://medium.com/@aboutiana/a-brief-guide-to-using-foursquare-api-with-a-hands-on-example-on-python-6fc4d5451203

[8]. https://www.youtube.com/watch?v=1XqG0kaJVHY&t=845s

Cite this article as : Anurag Dave, Aditya Kumar Singh, Ashish Kumar Singh, Hemant Pandey , Nandeep Singh Hada, "Anatomization of Market", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 2, pp. 396-400, March-April 2021. Available at
doi : https://doi.org/10.32628/CSEIT217284
Journal URL : https://ijsrcseit.com/CSEIT217284