ISSN No:-2456-2165

Precise Emplacement

Shruti Sangal [1], Muskan Mittal[2], Nitin Singh [3], Shweta Gupta[4], Pooja Vajpayee[5]

[1]Student, Dept. of Computer Science & Engineering, RKGIT, UP, India
[2]Student, Dept. of Computer Science & Engineering, RKGIT, UP, India
[3]Student, Dept. of Computer Science & Engineering, RKGIT, UP, India
[4]Student, Dept. of Computer Science & Engineering, RKGIT, UP, India
[5]Asst. Professor, Dept. of Computer Science & Engineering RKGIT, UP, India

Abstract:- The project synopsis deals with the technique of Data-Analytics which is becoming a very influential tool for decision-making today.

Software data analytics is key for helping stakeholders make decisions, and thus establishing a measurement and data analysis program is a recognized best practice within the software industry. However, practical implementation of measurement programs and analytics in industry is challenging. In this chapter, we discuss real-world challenges that arise during the implementation of a software measurement and analytics program. We also report lessons learned for overcoming these challenges and best practices for practical, effective data analysis in industry.

The main objective of this work is to understand data for decision making. The author here tries to select those locations that are not already crowded with restaurants within the region and have a greater population by using various data science and analysis techniques to reach their goal of selecting optimal locations. The advantages of each area will be clearly expressed so that the best possible final location can be chosen by the stakeholders.

The Author initialized a crawler to scrape the data about the areas of cities using Wikipedia web page and the real-time data set. Python's geocoder library with ArcGIS as a geocode provider, to get the coordinates of the neighborhoods. After which the author applies the clustering algorithm, K-Means, on the data to cluster the neighborhood based on general venue density and analyze & compare the sets in each cluster to conclude the most promising and optimal locations for each restaurant type. Which is then filter out only the target restaurant of interest in each neighborhood, to analyze within the clusters.

Keywords:- *Integrated Development Environment(IDE), Machine Learning (ML), Application Programming Interface (API), Domain Specific Language (DSL), Representational State Transfer (REST).*

I. INTRODUCTION

New York's statistical data shows that it is a large and diverse metropolis. It is the 27th largest state geographically. New York has a diverse culture and with diverse culture comes with a diverse food items.

New York has a long history of international immigration. It was a home to nearly 8.5 million people in 2014 having over 40% of the population of New York State and a slightly lower percentage of the New York metropolitan area, home to approximately 23.6 million. New York has also emerged as a global node of creativity, entrepreneurship and environmental sustainability, and a symbol of freedom and cultural diversity.

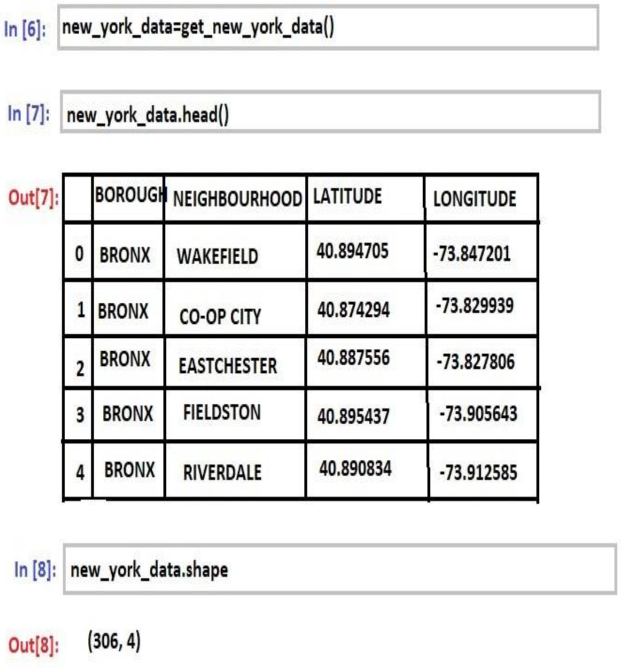
In 2019, New York was voted the greatest city in the world as per a survey of over 30,000 people from 48 cities worldwide, citing its cultural diversity. There are many districts and monuments in New York City that are the major tourist attractions .There has been a record of 66.6 million tourists visited in New York city. Since it is such a famous place, therefore it has a large and diverse cuisines.

New York being a city full of diversity there are many restaurants out their each belonging to different category such as Indian, Chinese, Continental, Korean and French etc.

And our project basically selects that locations for the entrepreneurs and stakeholders where there are less number of restaurants of a particular cuisine and gives the coordinates of that particular location which helps make the entrepreneurs and the stakeholders take decision regarding the location to be selected for the opening and the place where the restaurant opening can turn out as a success.

II. DESIGN METHODOLOGY

- We start off by assembling the New York city data from the following link "https://cocl.us/new_york_dataset"
- We will observe all venues for each neighbourhood using Foursquare Api. The above result shows that there are 306 different neighborhoods in New York.



```
In [6]: new_york_data=get_new_york_data()

In [7]: new_york_data.head()

Out[7]:
     BOROUGH  NEIGHBOURHOOD  LATITUDE   LONGITUDE
  0  BRONX    WAKEFIELD      40.894705  -73.847201
  1  BRONX    CO-OP CITY     40.874294  -73.829939
  2  BRONX    EASTCHESTER    40.887556  -73.827806
  3  BRONX    FIELDSTON      40.895437  -73.905643
  4  BRONX    RIVERDALE      40.890834  -73.912585

In [8]: new_york_data.shape

Out[8]: (306, 4)
```

Fig 1: Output showing no of neighbourhood

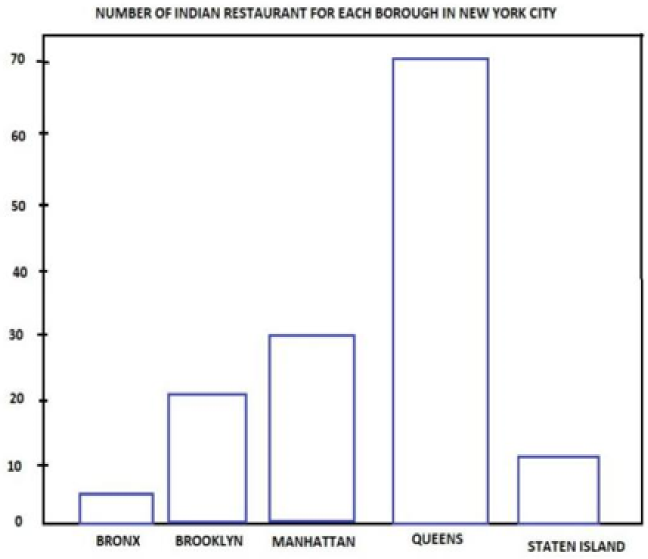
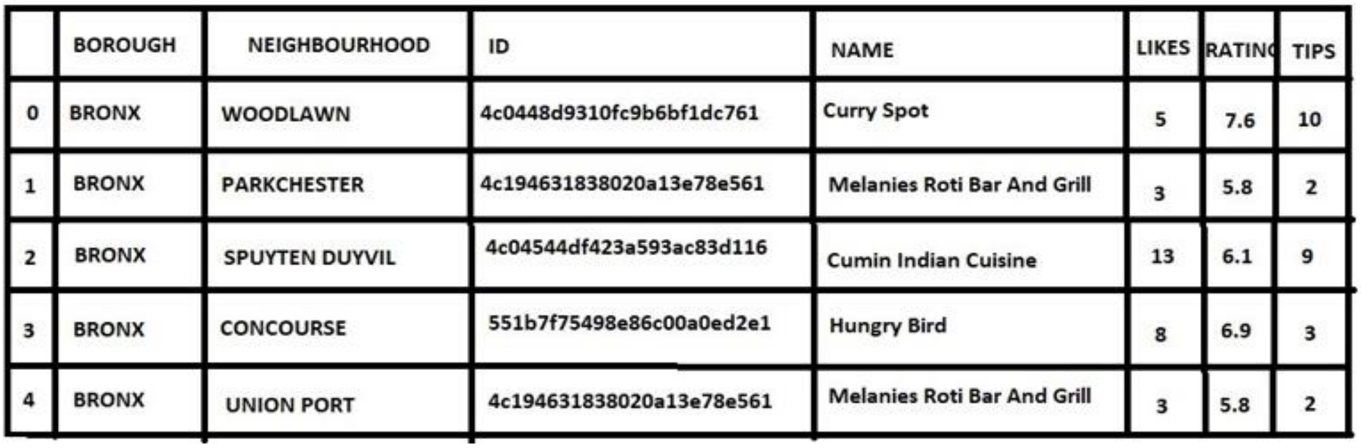


Fig 2: No of Indian restaurant in each borough

- ➢ We will then clear out all venues with Indian restaurant for further analysis.

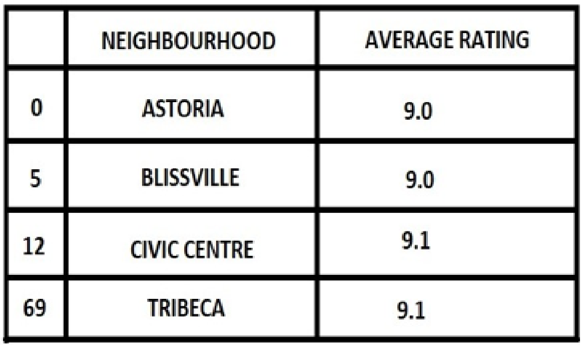


	BOROUGH	NEIGHBOURHOOD	ID	NAME	LIKES	RATING	TIPS
0	BRONX	WOODLAWN	4c0448d9310fc9b6bf1dc761	Curry Spot	5	7.6	10
1	BRONX	PARKCHESTER	4c194631838020a13e78e561	Melanies Roti Bar And Grill	3	5.8	2
2	BRONX	SPUYTEN DUYVIL	4c04544df423a593ac83d116	Cumin Indian Cuisine	13	6.1	9
3	BRONX	CONCOURSE	551b7f75498e86c00a0ed2e1	Hungry Bird	8	6.9	3
4	BRONX	UNION PORT	4c194631838020a13e78e561	Melanies Roti Bar And Grill	3	5.8	2

Table 1: Venues with Indian Restaurant

- ➢ In the Next step, using Foursquare API, we will figure out the Ratings, Tips, and Number of Likes for all the Indian Restaurants.
- ➢ We will then sort Neighbourhood and Borough the data keeping Ratings as the constraint.
- ➢ In the Next step, we will consider all the neighbourhood with average rating greater or equal 9.0 to visualize on map.

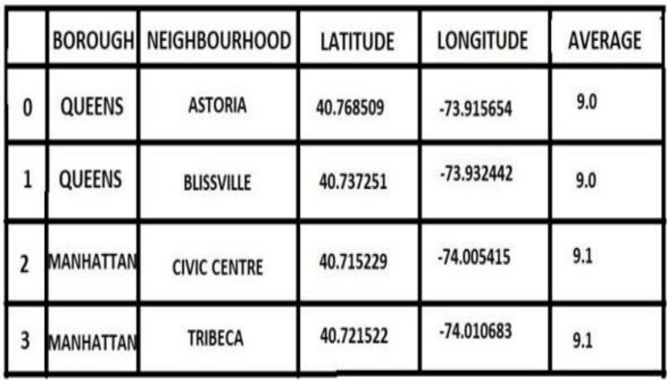
For this step the author gathered the rating of the neighbourhood.



	NEIGHBOURHOOD	AVERAGE RATING
0	ASTORIA	9.0
5	BLISSVILLE	9.0
12	CIVIC CENTRE	9.1
69	TRIBECA	9.1

Table 2: Average Restaurant Rating by area

- ➢ We will join this data set to previous New York data to get longitude and latitude.



	BOROUGH	NEIGHBOURHOOD	LATITUDE	LONGITUDE	AVERAGE
0	QUEENS	ASTORIA	40.768509	-73.915654	9.0
1	QUEENS	BLISSVILLE	40.737251	-73.932442	9.0
2	MANHATTAN	CIVIC CENTRE	40.715229	-74.005415	9.1
3	MANHATTAN	TRIBECA	40.721522	-74.010683	9.1

Table 3: Applying K-means Clustering

- Finally, we will visualize the Neighbourhood and Borough based on average Rating using python’s Folium library.

III. FLOW CHART

Once we get the data resource we initialize a crawler to scrape the data then by using the python’s geocoder with ArcGIS as a geocoder provider to get the coordinates of neighbourhood. The neighbourhood are then visualized using python’s folium package. The coordinates data is used by the Foursquare API to fetch information and then by applying the k-means clustering algorithm the data is clustered based on general venue density and analyze & compare the set in each cluster to conclude the most promising and optimal locations for each restaurant type.

Then the filtration is done accordingly and the result comes out as boon for the stakeholders and the entrepreneur the process is then repeated for different cities with the data available for scrapping and the result is updated whenever their is change or upgradation in the data.

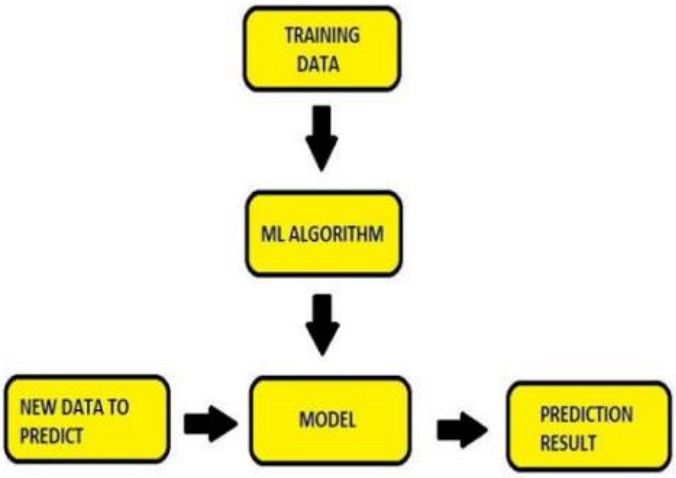


Fig 3: Mechanism of predicting data

IV. SOFTWARE DETAILS

- *Jupyter Notebook*

Jupyter Notebook gives you an convenient, interactive data science environment across many programming languages that doesn't only perform as an (IDE), but also as a presentation or education tool.

- *Machine Learning*

ML vision of Artificial Intelligence which permit software products to become more precise at forecasting consequences without being directly programmed to do so. (ML) is considered to be one of the most fascinating technology as it's clear by name that it will provide learning capability to computers.

- *Python*

Python is general- purpose, object oriented programming language. It is used for conducting data analysis , automating tasks , software and websites. General purpose language is opposite to a (DSL). GPL is programming language that is efficient of creating all types of programs. It include exceptions, dynamic typing, high level dynamic data types, modules, and classes.

- *Pandas*

Pandas is an open source, three-clause Berkeley Source Distribution (BSD) licensed library written in Python Programming Language. It is extensively used for Data Analysis and ML tasks. It mainly works with the tabular data and provide quick, easy to use data structures , high performance and data analysis tools.

- *Numpy*

NumPy (Numerical python) was created in 2005 by Travis Oliphant. It offers a powerful object called Array to perform huge variety of mathematical calculation. It mainly works with numerical data and matrices. Matplotlib, pandas and Scikit-learn are built on top of numpy.

- *Geocoder library*

Geocoding is process of translating street addresses into geographic co-ordinates (longitude/latitude) for exactness. It accept data from the user and convert it into map view. It is licensed by Apache2.

Apache2 is a web server software.

- *Foursquare API*

It is link between Client and Server.

Request: REST API endpoint URL + API method .

Response: Representation of resource (REST- XML -RPC - SOAP - JSON - Serialized PHP).

V. CONCLUSION

Nowadays Worldwide, Data Analytics is broadly used . This project can be very helpful to data mining and analysis department for continuously monitoring and upgrading divisions in the field of data scrapping. This project can serve mankind for stakeholders and entrepreneurs for the New York city to search for optimal locations very effectively. It is very cost effective and flexible in terms of usage for the entrepreneur.

There is always room for improvement and hence the search result obtained can be improved for better results depending upon the data.

FUTURE SCOPE

The author beleives that this briefing proposes an application that simplifies the process of web scrapping meaning extracting data from large documents and websites, by making the use of data analysis, i.e., extracting only meaningful excerpts from a large document or website. There are websites that enforce their security by the use of DRM and scrapping such websites yield no results Therefore, the next version of this application may work on such websites for data analysis. Further the author making versions of it as an android or web application which may also include the upload of different kinds of reviews of restaurant for much more efficient summarization, since most of the people nowadays take decisions based on reviews.

REFRENCES

[1]. Paula Fernandez Costa, Irving Badolato, Rogerio Borba, Julia Strauch.”Strategy for extraction of foursquare’s social media geographic information through data mining” April 2019.

[2]. Shi Na, Liu Ximin and Guan Yong, “Research on k-means clustering algorithm” 2010 Third International Symposium on Intelligent Information Technology and Security Informatics tech.Rep.20-04-2010.

[3]. Roberto Catini, Dmytro Karamshuk, Orion Penner and Massimo Riccaboni” Identifying Geographic Clusters: A network analytical approach,” DOI:10.1016/j.respol.2015.01.011

[4]. Web Scraping https://www.unescap.org/sites/default/files/Leveraging_online_price_data _from_web_crawling_Malaysia_Stats_Cafe_30 Nov2020.pdf