APPLIED DATA SCIENCE CAPSTONE

Capstone Project: The Battle of Neighborhoods

Peer-graded Assignment:

Capstone Project - The Battle of Neighborhoods

Tharusha Morgan
Business Intelligence Developer

Now that you have been equipped with the skills and the tools to use location data to explore a geographical location, over the course of two weeks, you will have the opportunity to be as creative as you want and come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

Content:

1) Introduction/Business Problem

Clearly define a problem or an idea of your choice, where you would need to leverage the Foursquare location data to solve or execute. Remember that data science problems always target an audience and are meant to help a group of stakeholders solve a problem, so make sure that you explicitly describe your audience and why they would care about your problem.

The idea of this study is to identify the difference between food in Toronto, Canada and New York City, USA.

The questions that are been asked are, how does the different types of cuisines in Toronto differ from New York? Are there any regions that are dominant?

These are the kinds of questions that is going to be investigated in the Notebook and its findings. The data insights will let us know.

We will be diving in the location data using Foursquare.

2) Data Description

Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So, make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.

Dataset 1: New York City Location

The data we will be looking into is taken from the NYU Spatial Data Repository. The data is then going to be converted in csv. The csv features that we will be looking at are:

- Borough
- Neighborhoods
- Latitude
- Longitude

Dataset 2: Toronto Location

The csv features will we will be looking at are:

- Postal Code
- Borough
- Neighborhood
- Latitude
- Longitude

These two datasets will be working from the Foursquare API.

3) Methodology

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why.

The Notebook will be looking at the following Methodology:

- Collecting Data
- Cleansing Data
- The Visuals
- Machine Learning

1. **Collecting Data**

The datasets used in this notebook are collected from Wikipedia and NYU Spatial Data Repository.
The Foursquare data will be added within the notebook and no external file will be created for it. Once the necessary data is collected, cleaning the data is done.

2. **Cleansing Data**

This step takes much time. The cleaning of data in this notebook is done in two phases, one for the Toronto Data and other for the New York City Data. Removal of unnecessary columns and filtering data on what is needed to investigate.

3. **Visuals**

Few and simple Visualization will be added to the Notebook for further investigating from a visual point of view, which makes it easier to interpret the kind of data we are investigating to make the decision easier to understand.

Folium is used for this purpose. Folium generates interactive and beautiful maps, using the Latitude and Longitude we provide. Further to improve the understanding of the data, count plots are used to comprehend the result.

4. **Machine Learning**

K-Means Clustering is used on common venues to identify the probable clusters for the different kinds of restaurants. We will delve on 5 Clusters.

4) Results

Results section where you discuss the results.

1. As anticipated, the Italian Restaurant has highest count in Manhattan as **1st Most Common Restaurant**. But Vegetarian/Vegan Restaurant has highest count in Downtown Toronto as **2nd Most Common Restaurant**. However, we can conclude that the Italian and the Vietnamese Restaurants control both Manhattan and Downtown Toronto.
2. In Manhattan, **1st Most Common Restaurant** have Italian Restaurant as top followed by American and a tie with Mexican and Japanese Restaurant. For **2nd Most Common Restaurant**, we have rather a different picture. It is topped by a Sushi Restaurant followed by Mexican, Japanese and French Restaurants.
3. In Downtown Toronto, **1st Most Common Restaurant** we have 'Restaurant' as second common. But what type of cuisine they serve is unknown, which is an inconsistency. It shares top with Vietnamese Restaurant. For **2nd Most Common Restaurant**, Vegetarian/Vegan and American Restaurants take top spot.

5) Discussion

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

What is discovered:

There is a small inconsistency in the datasets. There are some entries in the dataset that just say 'Restaurant'. We can consider this category as a restaurant which serves different cuisines. There are a few methods by which we can solve this inconsistency. One such method is to consider this restaurant as a mainstream cuisine in that area. Example: If in an area, Italian Cuisine has highest count, we can consider this 'Restaurant' as an Italian Restaurant.
The problem with this method is we might not always be correct. Another concrete way of solving this is rather tiresome approach, which is not feasible. We look for the corresponding coordinates and use google to find what type of restaurant is that.

6) Conclusion

Conclusion section where you conclude the report.

This report may be helpful for someone who can identify the relations between the food cuisines in either Toronto - Downtown or New York City - Manhattan. The Notebook may also assist in data related to other aspects such as entertainment venues or even sport sites, etc.