Applied Data Science - Capstone Project

Author: Ramajeyam Gopalraj

Introduction / Business Problem

I am a business man who is looking to setup a bar in Toronto area. I want to pick a place where there are not many existing bars. Hence, I want to apply data science using FourSquare data on Toronto area and locate area where there are other amenities such as Pizza Place and Diners, but not many Bars. We will use K-Means clustering to group the data by number of bars and pick the area with least number of bars and many Pizza and Diner places.

Data

We will scrape Toronto neighborhood information from this WikiPedia page. <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

After we collect the longitude and latitude of all Toronto area neighborhoods from the above source, we will then collect the top venues from each of these neighborhoods using FourSquare API.

Finally, we will synthesize the data to come up with a group of neighborhoods with least number of bars.

We will use the following data attributes in each cluster.

1. Neighborhood
2. Venue
3. Venue Category
4. Latitude
5. Longitude

Methodology

After obtaining the neighborhood data on top venues, we use K-means clustering to group the neighborhoods into top 5 groups with the top venues sorted descending. This will give us an idea of what neighborhoods already have a heavy concentration of bars. Thereby we can pick the one that does not have bar as their most common venues.

Results

After running K-means clustering on the neighborhood data, we have arrived at 5 clusters of neighborhoods with most common venues. Out of these 5 clusters we can pick the one that has Bar as the 6th or more most common venue.

Discussion

After reviewing the cluster data, we can conclude that Downtown Toronto has the least number of Bars as their top venues. Hence this neighborhood can be recommended for setting up a new bar.

Conclusion

To setup a new bar, though many locations are available, we conclude that Downtown Toronto is the best location. Because it has high concentration of other related venues such as Café, Airport, etc. But Bar there are not many bars.