Clustering – The battle of Neighborhoods

URL

https://www.kaggle.com/hark99/clustering-the-battle-of-neighborhoods

Problem & Background

Toronto and New York are the famous places in the world. They are diverse in many ways. Both are multicultural as well as the financial hubs of their respective countries. We want to explore how much they are similar or dissimilar in aspects from a tourist point of view regarding food, accommodation, beautiful places, and many more.

Today Tourism is one of the pillars of the economy and the people most often visits those countries who are rich in heritage and developed enough from a foreign prospective, like friendly environment. Every city is unique in their own way and give something new. And now the information is so common regarding location of every place around the world on your fingertips which make it easier to explore. Therefore, tourists always eager to travel to different places on the basis of available information, and the comparison (the part of the information) between the two cities always assist to choose the specific places or according to their choice.

Data Description

For this problem, we will get the services of Foursquare API to explore the data of two cities, in terms of their neighborhoods. The data also include the information about the places around each neighborhood like restaurants, hotels, coffee shops, parks, theaters, art galleries, museums   and many more. We selected one Borough from each city to analyze their neighborhoods. Manhattan from New York and Downtown Toronto from Toronto. We will use machine learning technique, "Clustering" to segment the neighborhoods with similar objects on the basis of each neighborhood data. These objects will be given priority on the basis of foot traffic (activity) in their respective neighborhoods. This will help to locate the tourist's areas and hubs, and then we can judge the similarity or dissimilarity between two cities on that basis.

Methodology

As we have selected two cities Borough to explore their neighborhoods. The data exploration, analysis and visualization for both boroughs are done in the same way but separately.

Exploration

For Downtown Toronto case, we have extracted table of Toronto's Borough from Wikipedia page. Then we arrange the data according to our requirements. In the arrangement phase, which applied multiple steps including but not limited to, eliminating "Not assigned" values, combine neighborhoods which have same geographical coordinates at each borough and sorted against the concerned borough. For data verification and further exploration, we use Foursquare API to get the coordinates of Downtown Toronto and explore its neighborhoods. The neighborhoods are further characterized as venues and venue categories.

For Manhattan, we used a saved data file which is already explored through foursquare API in which we have extracted all the boroughs of New York and then sorted against the concerned borough. Then we explored the Manhattan neighborhoods as venues and venue categories.

Analysis

We analyze both boroughs neighborhoods through one hot encoding (giving '1' if a venue category is there, and '0' in case of venue category is not there). On the basis of one hot encoding, we calculate mean of the frequency of occurrence of each category and picked top ten venues on that basis for each neighborhood. It means the top venues are showing the foot traffic or the more visited places.

Now we applied Machine Learning Technique "Clustering" to segment the neighborhoods in similar objects cluster. This will help to analyze from Tourist perspective and we can easily extract the Tourist places which are present on one of the clusters.

Visualization

We visualize the data many times at different stages. In the beginning, we visualize the selected borough neighborhoods so that we can get an idea or confirmation regarding the coordinates of that Borough. The second time after clustered the neighborhoods, we visualize the clusters to name them. Assigning the names are very important because it can identify the areas or specific places in each cluster.

Results

After clustering the data of the respective neighborhoods, both cities (Boroughs) have venues which can be explored and attract the Tourists. The neighborhoods are much similar in features like Theaters, opera houses, food places, clubs, museums, parks etc. As far as concern to dissimilarity, it differs in terms of some unique places like historical places and monuments.

Observations & Recommendations

When we compare the tourist places, we observe that the historical place is only situated in Downtown Toronto and the Monument or landmark venue is in Manhattan neighborhoods. Similarly, Airport facility, Harbor, Sculpture garden and Boat or ferry services are also available in Downtown Toronto while venues like Nightlife, Climbing gym and Museums are present in Manhattan.

As far as concern to recommendations, we recommend Downtown Toronto Neighborhoods will be considered first to visit. The tourists have an easily travelling access due to Airport facility, which not only saves time but also helps to save money. This saved money can be utilized to explore more, the attracting venues.

Conclusion

The downtown Toronto and Manhattan neighborhoods have more like similar venues. As we know that every place is unique in its own way, so that's argument is present in both neighborhoods. The dissimilarity exists in terms of some different venues and facilities but not on a larger extent.