Coursera Capstone

IBM Applied Data science Capstone

**Multicultural Cities Similarity Detection and Visualization**

By: Somnath Khilari

Nov. 2020

# Introduction:

Predict and visualize similarity between multicultural cities like New York/Toronto across the world. If a person from New York migrates to a new city like Toronto and if he wants to start his living in the similar neighbourhood where he was living earlier OR if he wants to start a business in similar area, what place can we recommend?

Basically in this problem we will collect data about New York neighborhoods, and pull location data from Foursquare to understand variety of venues. Same exercise would be done for city of Toronto and both the data sets would be combined and clustered using K-means clustering, and similar neighborhoods would be visualized on the world map

With the growth of smart-phones, and location-based social networks, data is being generated like never before. There are nevertheless many practical questions in urban computing that require the comparison across cities. For example, a job seeker with may wish to focus her search on a single city with jobs that best match her qualifications, rather than dispersing her search efforts across multiple cities. Likewise, a large corporation looking to expand its locations might perhaps select cities it wishes to expand into before considering particular sites or neighborhoods. Additionally, many within-city computations might be aided by modelling a city’s relationship to other cities. For example, a person buying or renting a home in a new city might want to be able to compare the neighborhoods of the city to other neighborhoods in different cities. One issue in comparing spatial regions such as cities is the normalization of absolute data, since often raw data different contexts are incomparable.

# Data:

Newyork and Toronto neighbourhood data would be captured from wikipedia. Also data about longitude and latitude of the data can be extracted from Web.Venue data from Foursquare would be extracted for each neighbourhood's 5km radius. Both the data sets would be merged for comparison and clustering. Number of venues for each neighbourhood would be aggregated using pandas dataframe which can becomes base for similarity between the locations.

Exploring these ideas empirically using real world data re-quires that we gather both the description, or the categorizations, of the venues in a city, as well as information about the city’s municipal neighborhood boundaries. For the venues , we collected data from the widely used location-based Social Network (LBSN) Foursquare. Users of Foursquare “checkin” to their current location on their mobile device by selecting it from a list of nearby named venues. Their check-in is then broadcast to their social connections. Foursquare also specify a hierarchical categorical description to avenue, such as “Restaurant” and “Mexican Restaurant”. We call higher-level (more general) categories the primary category, and we call lower level (more specific) categories the secondary category. An interesting side-benefit of such collaborative tagging is that as the system’s user base increases ,an accurate and up-to-date crowd-sourced representation of the venues types within a city is naturally accumulated. In this work we exploit this natural data for our empirical study. The venue information is easily accessible through a public API, and all venues are annotated with categories of different granularities which represent a natural semantic grouping for venues.