APRIL 5, 2020

APPLIED DATA SCIENCE CAPSTONE

FINAL PROJECT

TABLE OF CONTENTS

Problem Statement

Many people throughout the united states of America internally emigrate to Florida. In fact Florida is the number one state on receiving internal emigrants every year. Also, a significant percentage of the population of the new emigrants (emigrants from another countries) prefer to land in Florida in hope to either find a job or start a new business. Unfortunately, most people had a business idea but did not know which county in Florida had real need to their business and thus their business had better potential to flourish. Another fact is, most of these emigrants target Central Florida because its less crowded than other metropolitan areas which makes the housing market reasonable and nice lifestyle is achievable. For this reason, other study will focus only on Central Florida.

In our project we attempt to help the new emigrants (both internal and external) by clustering the venue categories together. We will supply the model with all venue categories and all the cities of central Florida. After examining the clustering, we should be able to generate a set of recommendations to the new emigrants which should increase their changes to succeed with their business ideas.

Data Description

In our project we will need to get a list of all the cities in central Florida which we are going to get from Wikipedia, and specifically from : https://en.wikipedia.org/wiki/List_of_municipalities_in_Florida. We are going to retrieve the data and then perform scraping using BeautifulSoup library to extract the information we need. We will then get the co-ordinate of each of Central Florida cities from World map and specifically: https://www.mapsofworld.com/usa/states/florida/lat-long.html. We will join the two data sets together and then prepare a query (or set of queries) to get Venue information from FourSquare. After getting the information from FourSquare, we will prepare the data for clustering. Please keep into consideration that although the model looks similar to the exercises we had for Toronto and New York, our project is different, because, the data supplied to the cluster is Venue Category revenue and not the cities.

Data required for the model to Run

Data	Data Source
Latitude and longitude of Florida	https://www.mapsofworld.com/usa/states/florida/lat-long.html
Venue and frequency per FL county	Foursquare
Cities list	https://en.wikipedia.org/wiki/List_of_municipalities_in_Florida.

Output/Cluster data

After the model generates the cluster, we will merge it and visualize the output to present the best neighborhood for each venue category. The data should look similar to this:

Venue Category	Mostly exist 1								Neighborhood n
	- NUM - Freq - Num o venue/person								