IBM- COURSERA

Data Science Specialization

Capstone project- Final Report

Identify Business Opportunities Through Clustering

Daniel David Ruiz Sanchez

1. Introduction

   Now a days Data analysis is the best tools to see things that we cat see only with observation. For this reason, we have the task to use this tool to give strong answers to a simple problem.

   A restaurant is this type of establishment that you have to think twice to build one, you have to keep on eyes who is living around, who is working around, the frequency that these places is walked. Then Clustering is the key tool to know these things through this way we can give these important advices to our clients.

   - Business Problem

     Thinking about the advantages of clustering and the information about neighborhoods through the use of foursquare, we can view the trade map of the city. Keep on eyes this idea, I'll be able to give an advice about open or not whatever type of commerce in a certain place. Then, the problem that I'm going to solve in this project is: What is the best place to open a restaurant in Toronto? Analyzing the best area, the count of restaurants per area and other important characteristics like the amount of other type of establishments.

2. Data acquisition and cleaning

   - Data

     For this study were necessary to use Foursquare information, this is one of the biggest data sets about places. Foursquare gets the information through its users who all the time are uploading photos, descriptions, opinions about restaurants, parks, and many other types of establishments.

     Wikipedia take its part on this project, this page contents all information about the neighborhoods and boroughs, these were really useful because this information allows to map the sites with the best accuracy.

   - Data cleaning

     About this process we have a lot of work, web scraping is an useful technique to get information of a website make easy the process of getting data although this data brings many undesirable signs.

     The information of foursquare didn’t give us problems, because all of the data were used.

3. Methodology

For this analysis were necessary to build a cluster model, to make it possible I have to split the data in four boroughs East Toronto, Central Toronto, Downtown Toronto and West Toronto, then I sort the neighborhoods by the 10 most common places. Whit this information the code was able to classify the neighborhoods by clusters.

For the first cluster (clus0), I found that the neighborhoods that belongs to this group have the characteristic of public places like parks, trails, dance studios, event spaces and others. But what is important of this cluster is the scarcest are the restaurants.

About cluster two (clus1), this is the cluster where food establishments have the most appearances, like coffees, gastro pubs, restaurants and others. It must be calling the food cluster.

Finally, third cluster (clus2) is the neighborhood where you can go to do physical activities because it has a lot of playgrounds or make an electronic shop and you can find a few varieties of restaurants.

4. Discussion

Although the third cluster wasn't representative, we can see that this location could be a good option to place a restaurant, with its stores and varieties it could be consider a crowded place.

However, I consider that the best option is the neighborhoods that compound the first cluster because this specifics zones of the city has the most crowded places and it´ll be walk by all these of the week, this guarantees that your restaurant won't be alone.

5. Conclusion

Although clustering isn't the only way to analyze this problem, it's an interesting tool to see the advantages or disadvantages of the business. This project doesn't show a completely analysis to give a strong advice because crowed place isn't the most important feature for open or not a restaurant.

For these reasons I consider this project like the begin of an interesting study about establishment consulter, using other more complex techniques between clustering to get better results than the obtain.