**Capstone Project Report**

**Clustering neighborhoods of New York, Toronto and London**

**Serdar MAMIYEV**

***Introduction:****Previously, we have examined the similarity of neighborhoods of a city according to the famous places retrieved from*[*www.foursquare.com*](http://www.foursquare.com/)*. This time we are gonna do similar analysis, but with comparison among cities, instead.*

*Imagine, you work in a city and have been proposed a position in a company that has branches in different cities. Moreover, you were given a chance to choose among them. Definitely, you'd choose the most similar one to your city. This project is aimed to assist those people in given situation.*

*Subject cities,****Toronto, New York and London****, are the biggest and the most diverse, multicultural cities of the world. People living in any of these cities can surely refer to this project, in case they are to choose where to move among these cities.*

*To solve this problem, I will use data, retrieved from****Foursquare****website, about the famous places in each neighborhood of these cities.*

***Data:****For the data of this project, I will refer to number of sources.*

**1. New York neighborhoods:** I will convert addresses, retrieved from IBM server, <https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json> (so you can simply run a **wget** command and access the data), into their equivalent latitude and longitude values.

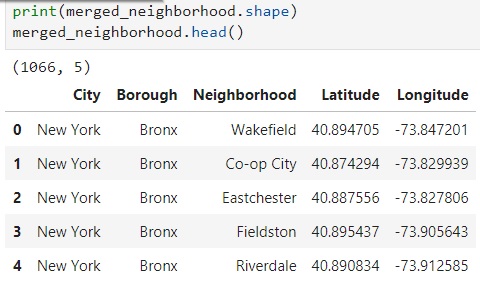
**2. Toronto neighborhoods:** I will scrape the following Wikipedia page, <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>, using the BeautifulSoup package to obtain and transform the data in the table of postal codes on the page and transform it into a pandas dataframe.

But this data lacks the coordinates, I will obtain them from a link to a csv file that has the geographical coordinates of each postal code in Toronto: <http://cocl.us/Geospatial_data>.

Later I will concatenate two dataset using methods provided in pandas library.

**3. London neighborhoods:** I will get the raw data about London neighborhoods from a link to a csv file: <https://www.doogal.co.uk/UKPostcodesCSV.ashx?area=London>

I will use number of techniques of python to transform and join the retrieved data into single dataframe of all ***neighborhoods*** of each city, their ***latitude*** and ***longitude*** and ***boroughs*** they are in.



Also, I will use the Foursquare API to explore neighborhoods in the cities. I will use the **explore** endpoint to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters.

I will use the ***k*-means clustering** algorithm to complete this task. Finally, you will use the **Folium** library to visualize the neighborhoods in given three cities and their emerging clusters.

**4. Foursquare:** I will refer to [www.foursquare.com](http://www.foursquare.com/) for data about venues in neighborhoods.

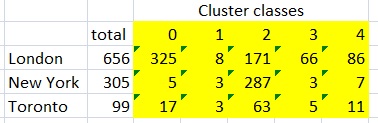
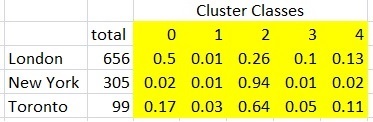
**Methodology:**

In order to address the problem set above, after collecting the data from given sources, especially data from [www.foursquare.com](http://www.foursquare.com/) using ***/explore?*** endpoint, I find out that there are a number of recommended places that these cities share.

That’s why I decided to utilize that information and divide neighborhoods into clusters according to recommended places in them. In order to do this, I had to transform my data into numeric. Initially, I had lists of categories of venues in neighborhoods. So, I decided to get dummy values for each category. Now my data is numeric. Finally, to prepare our data for machine learning process I get the mean values of occurrences of each category in neighborhoods.

Next step was to use ***k*-means clustering**, and let computer to decide neighborhoods' reference to clusters. We will get a table of *Distribution of Neighborhoods among Clusters* as follows (right) (numbers may change in each run of the code):

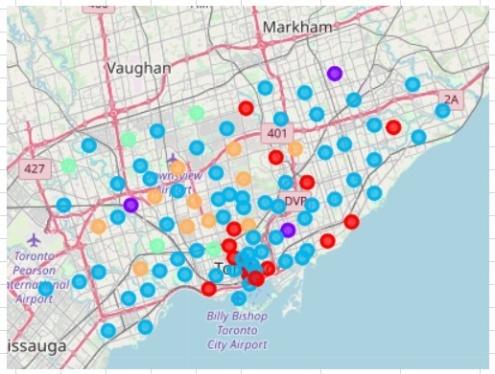
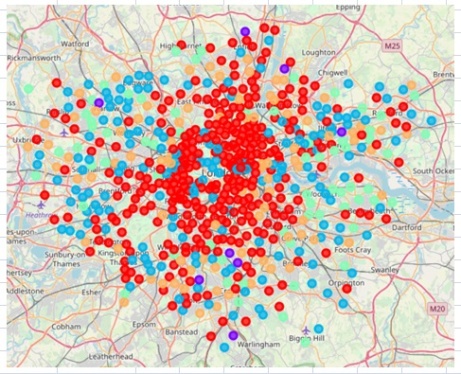
**Number of Neighborhoods in Clusters Percentage of Each Cluster Per Total**

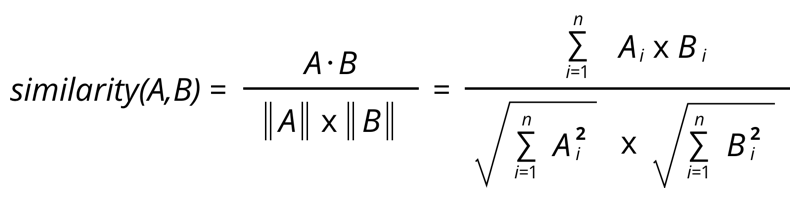
On the left table I calculated the percentage of each cluster in a given city neighborhoods, i.e., half of the neighborhoods of London fall into cluster 0.

Clusters are displayed below:

**London Neighborhoods Toronto Neighborhoods New York Neighborhoods**



And finally, I used **Cosine similarity** for comparison among the cities.

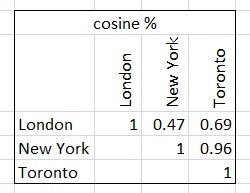


Here,  and  refer to vectors in , where each component representing a cluster. The vector represents the city's cluster distribution.  stands for scalar product of the vectors, and  and  stand for magnitudes of those vectors. Normally, values range between -1 and 1, where -1 is perfectly dissimilar and 1 is perfectly similar. But as our vectors are all positives, the range of our results will be 0 to 1.

Cosine similarity test was adapted rather than Euclidean similarity test because in some cases, like zero value for some clusters, the result was misleading in later method.

**Report:**

I could achieve my goal of comparing three cities, London, Toronto and New York, by putting into practice machine learning technique, specifically, ***K*-means clustering** and ***cosine similarity test.*** And according to my findings, London happens to be more similar to Toronto, while New York and London seem to be less similar.



As illustrated on the table above, **Toronto and New York** cities have the highest Cosine similarity of **0.96**, while the least similar cities among three are **London and New York** with Cosine similarity of **0.47**.

By the way, it worth mentioning that Cosine similarity of cities to themselves respectively are equal to **1**, which confirms correctness of our outcome.

**Discussion:**

A notable point is that, indeed, New York and Toronto were established by totally by immigrants, hence, the likelihood that people living there have similar interests is very high in contrast to London, where conservative population's rate is much higher, despite of increasing number of immigration. This may be one of the explanations of my findings.

For further researches I recommend that the number of cities were increased, and cities like Paris, Tokyo, Dubai and Singapore included.

**Conclusion:**

As a conclusion, in my project, I tried to compare different cities around the world with respect to their neighborhoods. For this, I gathered data of venues in these cities from [www.foursquare.com](http://www.foursquare.com/).

The data is processed and filtered before it is sent for machine learning.

As a machine learning technique ***K*-means clustering** algorithm is used, where number of clusters is decided to be **5**. Later, outcome is subjected to **Cosine similarity test** to compare the similarities between the cities.

Findings were decided to be precise and close to the reality as New York and Toronto are cities composed of **100%** of immigrants and have history of few centuries only, their inhabitants would have similar interests. That is affirmed by the Cosine similarity result which is **0.96**. While Cosine similarities between London and Toronto and New York are **0.69** and **0.47**, respectively.

According to the findings, if anyone wants to move from Toronto to London or New York, he or she must opt for New York as it is very much alike to his or her city with respect to its famous places.