Comparing Neighborhoods in Dubai and Doha

Introduction

This article is the first part of my capstone project for Coursera IBM Data Science Professional Certificate, a 10-course program offerd by IBM that explores several disciplines of the Data Science field. Some of those disciplines will be applied to solve the Business Problem described below.

Business Problem

Dubai and Doha are Middle East cities very similar to each other, not only because of the climate, geographic location and futuristic skyscrapers. Both are ranked among the most high-tech and safest cities in the world. The life cost and average salary are also similar in both cities.

Dubai and Doha are also very similar in terms of opportunities for work, especially in IT. Based on that, we may face the following situation: an IT professional, based in Dubai, who got a great job offer from a big company in Doha, decides to move to Qatar's capital.

When moving to the new city, the professional would like not only to live near the new job but also to settle in somewhere similar to where he or she is currently based. When people are used to live near places that make their lives easier and more confortable (such as gyms, restaurants, supermarkets...), it is natural that they look for neighborhoods with the same characteristics in a new city.

Using Data Science to Solve the Problem

The project goal is to cluster similar neighborhoods in Dubai and Doha leveraging Foursquare location data. With the final report, people will be able to easily find neighborhoods with the characteristics they are looking for.

Using Foursquare API and some Data Science tools and techniques, we can segment and compare neighborhoods of the two cities.

To get a detailed final report, we will follow the steps below in a Jupyter Notebook:

1) Using Requests and BeautifulSoup packages to web scrape the wikipedia pages of Dubai and Doha, we get a list of their neighborhoods and store it into a Pandas dataframe.

2) With Geocoder Python, we get the latitude and the longitude coordinates of each neighborhood.

3) Once we have the latitude and the longitude coordinates, we call Foursquare API to explore each neighborhood and list the most common venues of them and group the information into a new Pandas dataframe.

4) Having the Foursquare information of the neighborhoods, we use K-Means method, from scikit-learn library, to cluster neighborhoods based on their similarities.

The final result will help the IT professional to analyze the neighborhoods and find the one that best suits his or her needs and wishes.