CLUSTERING THE WORLD CAPITALS

Mudakkar M. Khadim

January, 2020

1. Introduction

1.1 Background

Sister cities or twin towns is a known concept in today's world where two or more cities or countries form a sort of agreement to promote their ties. These cities or countries need not be in the same geographical areas. The concept is mainly aimed at increasing the friendship and understanding between the cities or countries that ultimately helps in increasing trade and tourism. Further details of this concept can be seen on this link.

In this report, we try to find sister cities for world capitals. We select several capital cities across different continents and try to form groups of sister cities using a popular machine learning clustering algorithm and by using the location data of these cities.

1.2 Problem

Currently, the concept of sister cities is more driven by political or social ties between cities or countries. The objective of this report is to form clusters of capital cities that are closer to each other based on some physical attributes.

1.3 Interests

This piece of research can be of interest for many different groups some of which are given below;

a) Tourists or travelers
b) Tourism companies or authorities
c) Traders/Trading companies
d) International students

2. Data

For this research item, we will at least two different data sets. First, country names and their capital cities along with the latitude and longitude information. This information is available on Kaggle and can be accessed via this link. A sample data is shown in below table.

CountryName	CapitalName	CapitalLatitude	CapitalLongitude	CountryCode	ContinentName
Somaliland	Hargeisa	9.55	44.05	NULL	Africa
South Georgia and South Sandwich Islands	King Edward Point	-54.283333	-36.5	GS	Antarctica
French Southern and Antarctic Lands	Port-aux-FranÃ§ais	-49.35	70.216667	TF	Antarctica

Second, location data for the selected cities. This will include information of different venues (like hotels, restaurants, parks, etc..) within a certain radius. This information is accessed using Foursquare API. A sample data is shown in below table.

	Country	Capital	Capital_Latitude	Capital_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	Somaliland	Hargeisa	9.550000	44.050000	Hiddo - Dhawr	9.551411	44.047806	Comfort Food Restaurant
1	South Georgia and South Sandwich Islands	King Edward Point	-54.283333	-36.500000	جزيرة سندويشة	-54.282935	-36.495176	Beer Bar
2	South Georgia and South Sandwich Islands	King Edward Point	-54.283333	-36.500000	Bilinmeyen Yer	-54.281560	-36.506960	Racetrack
3	South Georgia and South Sandwich Islands	King Edward Point	-54.283333	-36.500000	Tang Ke Lek Harbour	-54.280980	-36.508610	Harbor / Marina
4	French Southern and Antarctic Lands	Port-aux-Français	-49.350000	70.216667	Book Time	-49.352470	70.218711	Bookstore

The latitude and longitude information available in the first data set is actually used to get the location data. And then clusters or groups of sister (similar) capital cities will be formed using some clustering algorithm based on the location data.