

1. Introduction

The city of San Francisco is one of the most famous people in the world. It attracts millions of tourists each year by its numerous landmarks, restaurants, cultural sites and so on. It might just be the starting point of a traveler on his/her journey to travel around the world.

However travelling abroad is always very expensive and time consuming for most people. The flights and travel expenses are themselves huge amount of money, let alone the planning phase of which landmark or restaurant to visit. Additionally there are so many cities around the world, which can serve as the destinations of journey. Which city should a traveler choose to be his/her next wonderland? The problem can be solved by comparing the city (cities) he/she has explored with the potential options.

In this project, the city of San Francisco is considered as a city Jamie (or a group of friends) has traveled. From that point, Jamie is deciding which city to explore next. London and Moscow the capitals of the United Kingdom and Russia respectively, and they are the options in Jamie's mind. By comparing how similar or dissimilar London and Moscow are to San Francisco, we can recommend Jamie to travel to either one of them or both. The same process can theoretically apply to any other different cities in the world.

2. Data

The raw data of the project is scraped using BeautifulSoup from the following three Wikipedia pages.

https://en.wikipedia.org/wiki/List_of_areas_of_London

https://en.wikipedia.org/wiki/Administrative_divisions_of_Moscow

https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Francisco

The neighbourhoods, areas and boroughs (in Moscow they are named differently) are scraped and formatted. The areas include their latitude and longitude information in their respective pages. The format of geographical coordinates are in different styles, such as '1'37'45'16"N 122°27'33" or "37.76424°N 122.42366°W". A cleansing of the formats are done so that the latitudes and longitudes are all in decimal formats with positive signs (N, E) and negative signs (S, W).

The formatted dataframe looks like the following:

	Link	Location	London_borough	Post_town	Postcode_district	Dial_code	OS_grid_ref	Latitude	Longitude
0	/wiki/Abbey_Wood	Abbey Wood	Bexley, Greenwich [1]	LONDON	SE2	020	TQ465785	51.486400	0.110900
1	/wiki/Acton_London	Acton	Ealing, Hammersmith and Fulham[2]	LONDON	W3, W4	020	TQ205805	51.513519	-0.270661
2	/wiki/Addington_London	Addington	Croydon[2]	CROYDON	CR0	020	TQ375645	51.358300	-0.030500
3	/wiki/Addiscombe	Addiscombe	Croydon[2]	CROYDON	CR0	020	TQ345665	51.381000	-0.066300
4	/wiki/Albany_Park_Bexley	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728	51.426400	0.102600

The formatted geometrical coordinates are fed into the foursquare api to generate recommended venues. All the areas/neighbourhoods/locations form a web and a Delaunay analysis is performed to find out the average distance between the vortices. This distance is used as the radius parameter in the foursquare api, so that as many as venues can be generated/captured. The venues information is similar to what we did in the lab, as following

	name	categories	lat	lng
39	Brill	Coffee Shop	51.525767	-0.109477
18	PizzaExpress	Pizza Place	51.406065	0.016451
73	Tariro Fairtrade Coffee House	Coffee Shop	51.401712	-0.195487
38	Jun Ming Xuan	Chinese Restaurant	51.595409	-0.242935
24	The London Borough of Barking & Dagenham Stadium	Soccer Stadium	51.547615	0.160125

Altogether there are ~36k, ~12k, ~7k venues generated respectively in London, Moscow and San Francisco. Then the venues are used to analyze the similarities between the cities.