IBM Data Science Professional Certificate - Capstone Project

The Battle of Neighborhoods

Where to open my Tea Room?

Introduction

For this assignment of the Capstone Project we are required to prepare a business plan using location data queried from Foursquare using the developer's API. The problem selected to be solved is the search for the most preferable country capital to open a Tea Room, considering that most countries have a definite preference between coffee and tea consumption. Though this problem may sound very specific, small tweaks in the algorithm or the initial datasets could enable the solution to be used for internal markets, or different product categories.

Data

The specific problem is highly dependent on location data. The data to be used will come from four datasets. The first dataset used contains the all world capitals, with their respective coordinates, scraped from Wikipedia. The second dataset is the largest one and contains all venues returned from the Foursquare API. The API returns data for the “100 best” venues including coordinates, tips, category, rating, opening hours, etc. The problem is the algorithm used to select which 100 venues are the best response to the user’s query is unclear and so cannot be used to draw exact conclusions or give quantitative answers, since there is a degree of unknown or randomness. This restricted many different algorithms and problems that were deliberated before this one, which did not require absolute numbers and could work with relative ratios between quantities. The third dataset used contains the annual per capita tea consumption per country for 2016, as measured by Statista, and was scraped from Wikipedia. The fourth dataset used contains the annual per capita coffee consumption per country for 2013 (for the 49 most consuming countries), as provided by Euromonitor.com, and was scraped from an online article.

Methodology

The initial business question set is “Where should I open my Tea Room?” The answer will be based on the premise that most countries have a definite preference in their coffee or tea consumption, primarily based on their history and tradition. The location recommendations will not be based just on national tea consumption and Tea Room availability in an area, but we will also take into account national coffee consumption as coffee is more widely consumed worldwide and can give a better comparison between the inversely correlated values of consumption and venue sparsity.

The first step was to import all the required Python libraries that would be used to import, process and visualize the data. The API connection with Foursquare is initialized with the user’s credentials and tested with a simple query to ensure it is returning correct results.

Secondly, Wikipedia is scraped for a list of the world capitals with their corresponding coordinates. Also, information on the annual per capita consumption of coffee and tea for as many countries as is available is scraped from two web pages and appended to pandas dataframes.

Initially, the API is queried for all the venues that contain the strings 'coffee' or 'tea' within 5 km from the center of each capital. Two dataframes are returned for the two strings containing the 100 "best" venues related to our query, as selected by Foursquare. Looking at the type of venues returned in each dataframe, we find many venue categories that may indeed sell coffee or tea but it is not a main part of their brand. Consequently we filter the coffee dataframe for venues whose category is 'Cafe', 'Cafeteria' or 'Coffee Shop', and the tea dataframe for 'Tea Room'. Through this filtering we get a better idea of the relative numbers between shops specializing in selling just coffee and tea. By calculating and sorting by a ratio between the shops and their categories (RatioCT) we get a dataframe containing capitals by ascending density of coffee shops, versus tea rooms.

However, that data would not give us the full picture as different countries have a tendency towards consuming more coffee or tea, which would lead to a higher percentage of e.g. coffee shops in Brazil or tea rooms in China. Using the national per capita consumption for the countries with available data, we can extract the ratio of coffee vs tea consumption (annually per capita, even though they are from slightly different years). As the best location would obviously be a country with high tea consumption, but low coffeeshop-to-tea room ratio, we can produce a new metric by dividing the two ratios, sorting by descending order, and so a list would appear with the most preferable capitals to open a tea room.

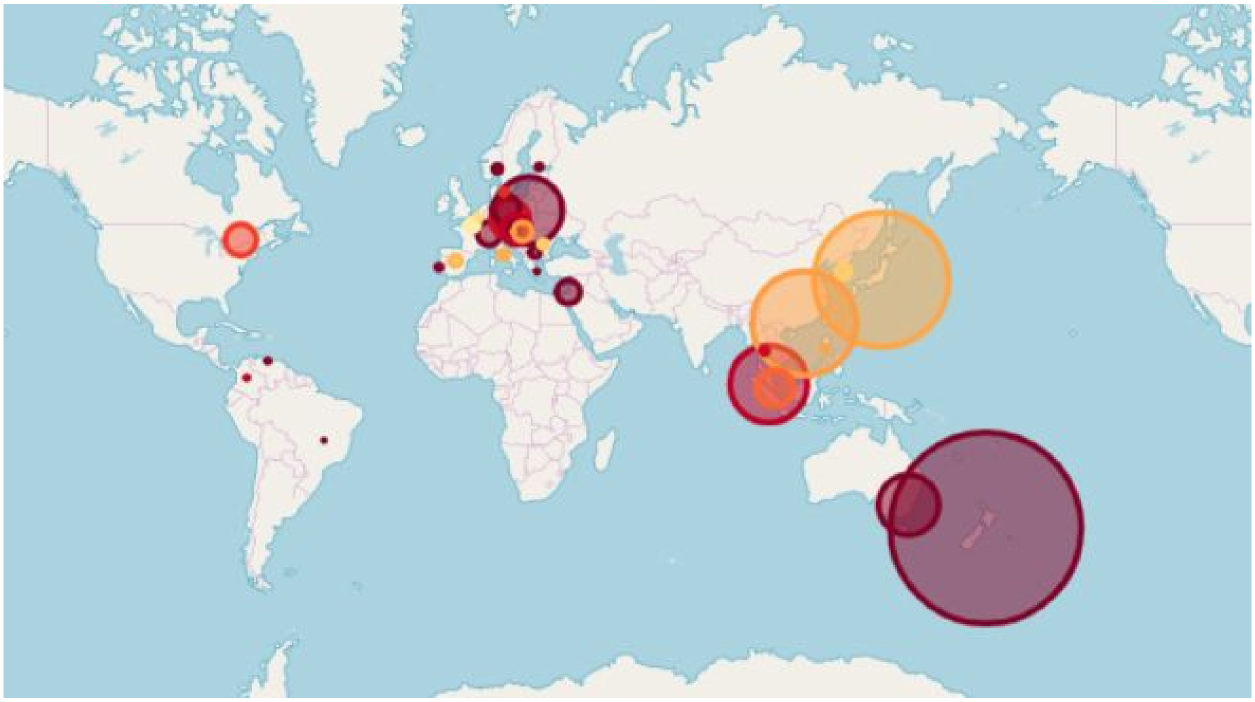
The above data was also used as input for two Unsupervised Learning algorithms, k-Means and Hierarchical clustering, to check if any useful relationships or clusters could be discovered between the country capitals, their locations, the preferrences, etc. The algorithms were rerun with different parameters (number of clusters, tree depth, linkage, etc.) to explore if they would have any bearing on the results. Unfortunately, it was found that no useful results could be reported, taking into account neighboring countries, language, or common history.

Results

Below we present the 5 most preferable capitals where one could open a tea room, and the 5 least preferable ones beneath that:

Capital	Coffee v Tea Shops	Coffee v Tea Consumption	Preference Metric (Shops/Consumption)
New Zealand, Wellington	10.000	1.092	9.153
Israel, Tel Aviv	81.000	9.000	9.000
Australia, Canberra	22.500	3.466	6.490
Poland, Warsaw	13.000	3.100	4.193
Malaysia, Kuala Lumpur	8.750	2.708	3.230
...	...	...	...
Colombia, Bogotá	8.750	77.777	0.112
Philippines, Manila	4.250	44.444	0.095
France, Paris	1.268	16.000	0.079
Brazil, Brasília	17.000	266.666	0.063
Belgium, Brussels	2.280	37.692	0.060

As a next step, the above table of results has been joined with the initial country capital coordinates dataframe so that the results can be visualised on a world map. The radius of the markers represents the annual per capita tea consumption of the specific country, whereas the red color scale represents the scarcity of tea rooms in the specified capital. Therefore, the most preferred locations are visualized by the larger purple-colored circles, as seen in New Zealand, Australia, Poland, etc. On the other hand, smaller (Europe, latin America) or whiter (China, Japan) circles represent capitals where coffee is preferred instead of tea, or there is a significant amount of tea rooms in the area, respectively.



Discussion

From the above results we get a clear answer for the best candidate capitals where we could open a successful tea room. Of course, this whole method is based on some hypotheses, such as: the API returns a representative ratio of coffee and tea shops without bias, the consumption values for each country are considered unchanged throughout the last few years, there is an inverse correlation between coffee and tea consumption worldwide, and the capitals adequately represent the national urban trends. Also, there are many more parameters that would have a significant or critical effect on the success of a business, related to the financial situation of each country, or the available support for new businesses. This solution only offers insight more closely related to the cultural and historic trends prevalent in each country.

Another observation is that there is no discernible geographical or cultural connection between the most or least preferred results, apart from Australia and New Zealand. For example, Poland, Israel and Malaysia are not popular tea drinking countries, contrary to what our data suggests. On the other hand, Brazil and Colombia would be expected to have a much higher preference towards coffee, but Belgium and Paris, two West-European countries would not be expected to show such a trend.

Unfortunately, trials with Machine Learning algorithms did not return any useful results. This would mainly be attributed to the small range of studied attributes, which in turn could be attributed to the restricted amount of data returned from the Foursquare API. Were we to have less restricted access to Foursquare's database, it would be easier to create more useful queries with more and better represented data to be used for more complex problems and more interesting insight.

Conclusion

Concluding, we can report that a new tea room would be a most welcome addition in Melbourne, Australia and Wellington, New Zealand. Malaysia and Poland seem to be interesting locations to make such a venture, whereas Latin America with its rich coffeebean production is expected to be more interested in coffee shops.

Due to the restricted amount of available or queriable data, more complex or interesting problems could not be tackled in a substantial way, with or without the help of Machine Learning algorithms.