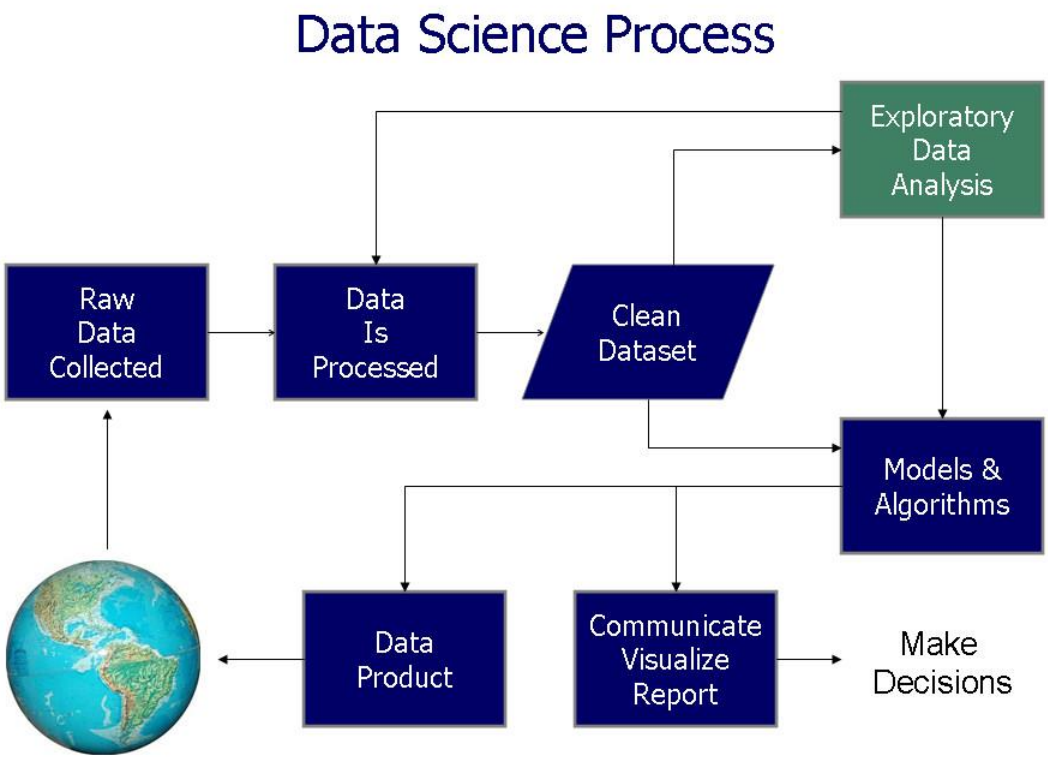


Using Foursquare Results – Relevancy in Local Context?

CAPSTONE PROJECT

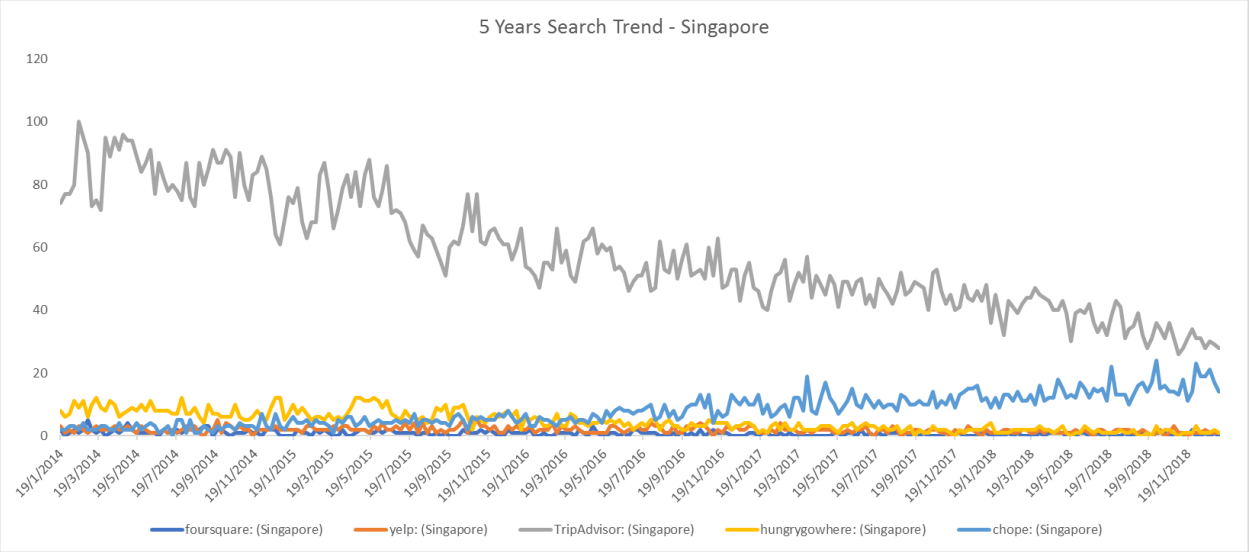
IBM DATA SCIENCE | Jan 2019

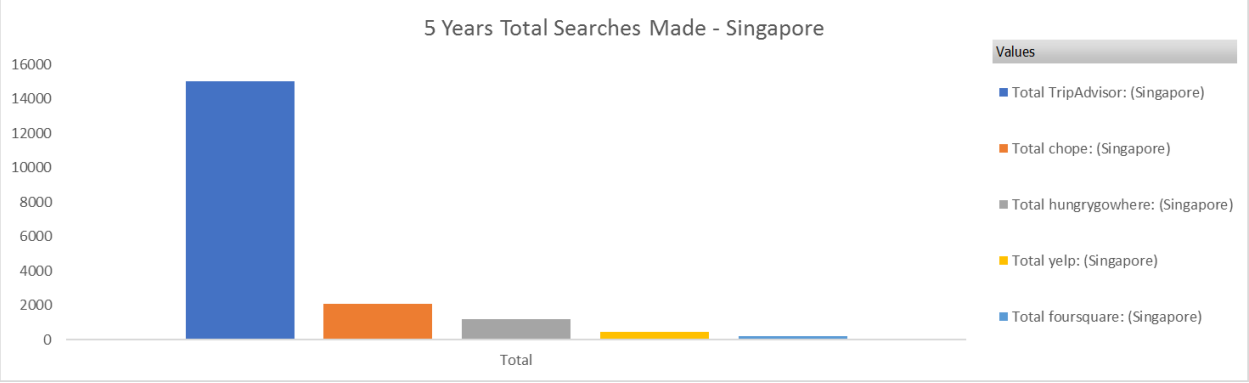
Introduction

As stated in its website, Foursquare is a technology company that uses location intelligence to build meaningful consumer experience and business solutions. Foursquare also provides developers free and premium access to its huge database to develop innovative business solutions or to check out nearby venues for making personal (e.g. where to dine) or business (e.g. where to set up shop) decisions.

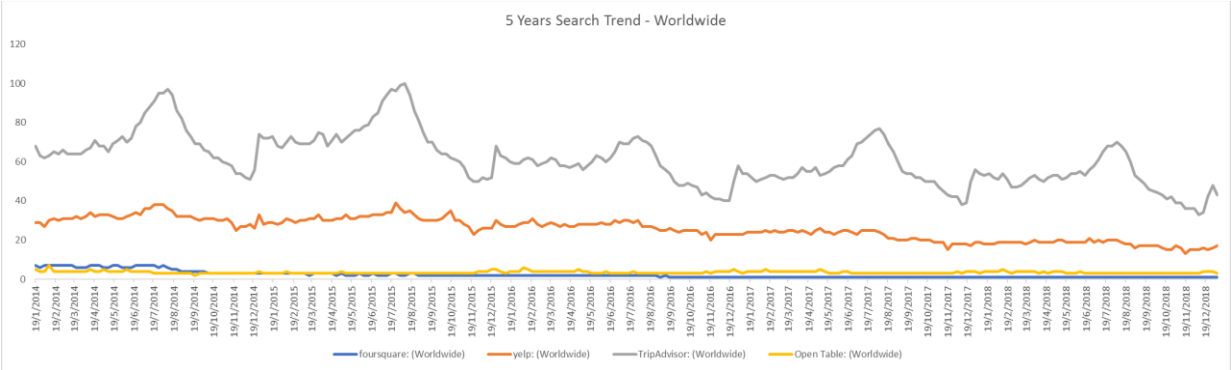
Other than Foursquare, there are also other international (e.g. Yelp, Trip Advisor) or local (e.g. Hungry Go Where, Chope) companies or apps which provide location-based advice / suggestion.

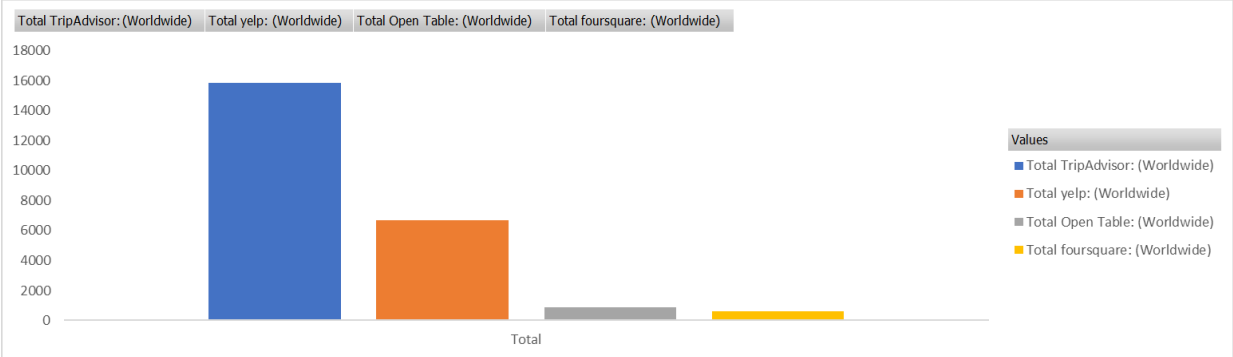
Performing a Google trend on the above search terms for the past 5 years, it appears that Foursquare is the least searched term in local context, while Trip Advisor is a highly searched term but on a downward trend. Local location-based app, Chope, has seen a rise in the number of searches over the past 5 years.





Given this background, we performed another Google trend analysis for the search terms: Foursquare, Yelp, Trip Advisor and Open Table on a worldwide basis.





The information from the worldwide trend analysis collaborated with that of the local analysis and Foursquare is the least searched term amongst the selected or targeted terms. Though this does not necessarily mean lowest usage nor market penetration, but it does give a hint to its popularity or market awareness. If it is not a frequently searched term, could this have an impact on usage and thus affecting the location-based intelligence it provides since the algorithm works best when there are many users providing feedback, comments, etc. on the platform? Can we still place relevance on the advice / suggestion it provides given there are other local-based apps or services available?

With the above as the back drop, we will attempt to explore:

- Does Foursquare return more venues in neighbourhoods where there is a higher portion of tourists compared to locals on the basis that locals are more familiar with homegrown apps?

- Do the results reflect the nature or characteristics of the neighbourhood?

The targeted audience is those who are interested to find out whether results from Foursquare is relevant in local context for decision-making purposes (personal and/or business) given Foursquare low searches on Google.

DATA

Based on domain / business knowledge, 5 neighbourhoods have been selected for analysis on the relevancy of the returned results provided by Foursquare:

1. Marina Bay: waterfront / Marina Bay Sands (tourist spot)
2. Sentosa: tourist attraction / themed park (tourist spot)
3. Jurong East: transport interchange / shopping mall (not known tourist spot)
4. Yishun: local neighbourhood (not known tourist spot)
5. Orchard: shopping belt / hotels (tourist spot)

After selecting these 5 neighbourhoods, performed a google search to obtain the latitudes and longitudes of these locations and transcribed them into an Excel spreadsheet. Also, randomly select 5 postal codes and coordinates from the geospatial_data.csv file for Toronto, Canada and include in the same spreadsheet. The inclusion of these postal codes is to demonstrate that data received is not always what we desired, and much time is spent on cleaning and getting the data ready (pre-processing work).

The fields in the Excel spreadsheet consist of:

- Neighbourhood: name of the neighbourhood
- Country: name of country which the neighbourhood is in
- Latitude: latitude of the neighbourhood
- Longitude: longitude of the neighbourhood
- Remarks: whether neighbourhood is more of a tourist or local spot

We will use the Foursquare API (Application Programming Interface) to extract nearby venues for the 5 neighbourhoods with a radius of 800 meters and limit the returned results to 1,000 through python programming. The returned results will show the name of establishment, the category which the establishment is in (e.g. hotel, Chinese restaurant, Korean restaurant, etc.), its latitude and longitude. Thereafter, the number of establishments / venues will be grouped by neighbourhood to provide an indication whether those neighbourhoods which are known to be tourist spots have higher returned results.

We will then perform one-hot encoding on the unique categories returned to obtain the frequency a category / venue appears and sort this frequency in descending order (from highest to lowest) for each neighbourhood to have a feel whether the information corresponds to the characteristics of the neighbourhood.

Finally, we will apply clustering, an unsupervised machine learning algorithm, to the data obtained.

METHODOLOGY

Based on domain / business knowledge, 5 neighbourhoods (Marina Bay, Sentosa, Jurong East, Yishun and Orchard) have been selected to analyse the relevancy of the results provided by Foursquare. Searched the web for a centralized file for Singapore postal codes and location coordinates but to no avail. Thus, Google search each location to find out its latitude and longitude and copy the information into an Excel spreadsheet, including additional information like country and remarks (whether the location is frequented more by tourist or local or mixture of both). Also, include 5 random postal codes from the Toronto geospatial file in the same Excel spreadsheet to perform pre-processing work to demonstrate data obtained from real world comes with noise. The information is not saved as a csv file as we take this opportunity to demonstrate Python (a programming language) is capable of reading Excel as well.

Import relevant modules and read the Excel file into Python using Jupyter Notebook. Personal preference is to make a copy of the original dataset prior to working on it so that initial dataset remains intact and could be recalled where necessary.

Examine the data and drop information not relevant for the analysis (e.g. column for Country and rows for Toronto locations). With the 'cleaned' dataset, generate a map of Singapore as well as the 5 locations for visual presentation.

Using the Foursquare API, extract venues and their associated categories for a randomly selected location (Sentosa, in this case) to have a feel of the results returned. Print out first 5 entries of the returned dataset for review:

	name	categories	lat	lng
0	Capella Singapore	Resort	1.249690	103.824323
1	Palawan Beach	Beach	1.248738	103.822002
2	The Knolls	Restaurant	1.249281	103.824356
3	So SPA by Sofitel	Spa	1.248546	103.826667
4	Auriga Spa	Spa	1.249532	103.824008

From these 5 entries, it appears that the location is properly matched with the categorization. We will extend the extraction to the rest of the locations. To find the number of venues returned for each location, we will group the returned results. Thereafter, we will use the one-hot encoding technique on the returned dataset for determination of frequency to rank the top 10 categories for each location.

Finally, we apply an unsupervised machine algorithm, clustering the returned results into the number of clusters we specified. "Clustering" is grouping similar entities together and the goal to find similarities in the data point and group similar data points together. The 2 most popular and widely used clustering techniques are K-mean Clustering and Hierarchical Clustering. K-mean clustering starts with K as the input for the number of clusters we want to find. Hierarchical clustering starts by assigning all data points as their

own cluster, builds the hierarchy, combines the two nearest data point, merges it together to one cluster and repeat the process.

For the analysis, we will use K-mean clustering and the number of clusters (K) will be 4. Due to the nature of our dataset, if we use K = 5 as the number of clusters, the dataset will be clustered back to the 5 neighbourhoods which we started off.

RESULTS

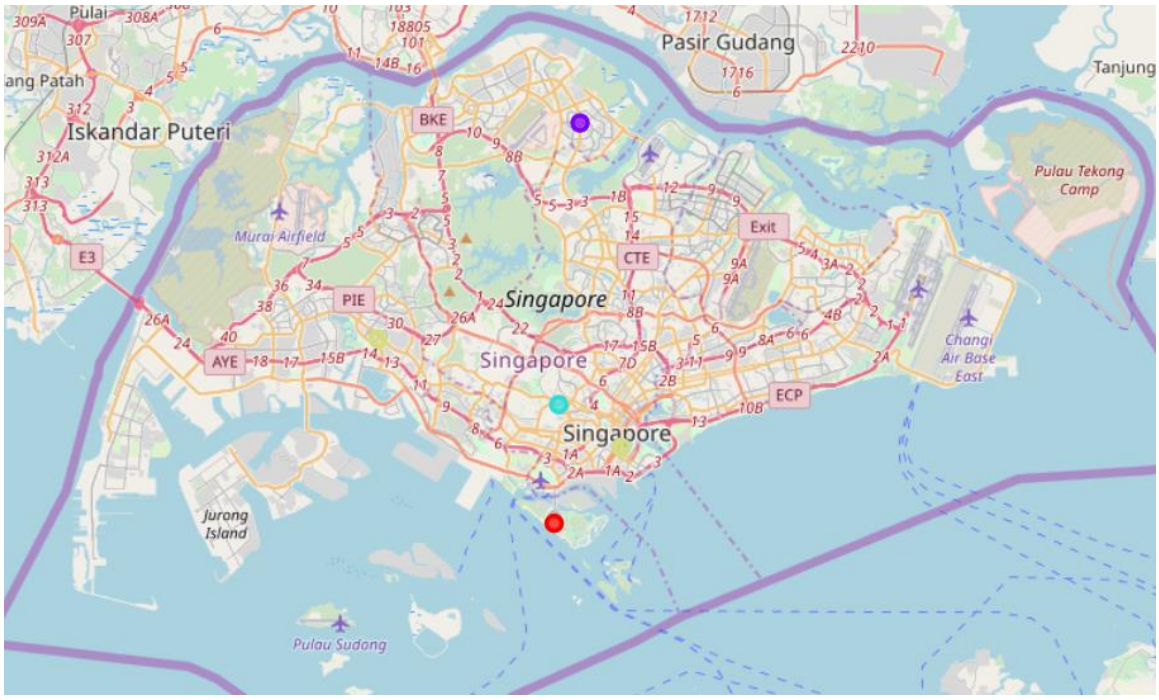
If we grouped the returned venues by each location, it is observed as follows (in descending order):

Neighbourhood	Number of Venues
Marina Bay	100
Sentosa	96
Orchard	84
Jurong East	80
Yishun	57

And the following is the top 10 venues for each neighbourhood:

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Jurong East	Japanese Restaurant	Shopping Mall	Chinese Restaurant	Food Court	Café	Department Store	Sandwich Place	Bubble Tea Shop	Bakery	Clothing Store
1	Marina Bay	Waterfront	Japanese Restaurant	Hotel	Salad Place	Cocktail Bar	Italian Restaurant	Sandwich Place	Gym / Fitness Center	Gym	Coffee Shop
2	Orchard	Hotel	Chinese Restaurant	Japanese Restaurant	Café	Boutique	Supermarket	Miscellaneous Shop	Cosmetics Shop	Coffee Shop	Cocktail Bar
3	Sentosa	Theme Park Ride / Attraction	Theme Park	Fast Food Restaurant	Café	Restaurant	Hotel	Beach	Food Court	Italian Restaurant	Resort
4	Yishun	Coffee Shop	Fast Food Restaurant	Indian Restaurant	Food Court	Grocery Store	Café	Supermarket	Shopping Mall	Chinese Restaurant	Ramen Restaurant

This is what we get if we cluster the neighbourhoods into 4 (rather than original 5):



	Neighbourhood	Country	Latitude	Longitude	Cluster Labels
0	Marina Bay	Singapore	1.2840	103.8535	3
1	Sentosa	Singapore	1.2494	103.8238	0
2	Jurong East	Singapore	1.3329	103.7436	3
3	Yishun	Singapore	1.4304	103.8354	1
4	Orchard	Singapore	1.3030	103.8258	2

The "after-clustering" result is that Jurong East is grouped together with Marina Bay as Cluster Labels 3. Please note that in Python's term, cluster labels 0 is our usual 1. Thus, though it is shown the maximum Cluster Labels is 3, there is actually 4 clusters.

DISCUSSION

A recap: the targeted audience for this analysis are those who are interested to know whether the returned Foursquare results are relevant in local context given Foursquare is the least searched term amongst other location intelligence apps / services.

To provide some assurance that the returned Foursquare results have some relevance in local context, we re-visit the following queries:

(A) Does Foursquare return more venues in neighbourhoods where there is a higher portion of tourists compared to locals, on the basis that locals are more familiar with homegrown apps?

Neighbourhood	Number of Venues
Marina Bay	100
Sentosa	96
Orchard	84
Jurong East	80
Yishun	57

From the results shown, after we grouped the returned venues by neighbourhood, Yishun, which is predominantly a local spot, has the lowest number of returned venues while Marina Bay has the highest as it is one of the more known tourist spots as well as locals visiting to soak in the atmosphere as well as trying their hands in the casino.

However, one surprise stood out from the results:

- Orchard is a shopping / hotel belt with a good mix of locals and tourists, has 84 returned venues while Jurong East, which is predominantly visited by locals though this is an MRT (subway) hub, registered 80 venues. Only a small difference separates the two.

To further examine this anomaly, we can extend the radius / coverage (currently sets at 800m) as well as getting Fourquare's Premium Plan (instead of the free plan used for this analysis) as there appears to be a limit of 100 returned venues per neighbourhood under the free plan.

Considering the constraints of the free plan, it appears that tourist spots (Marina Bay, Sentosa and Orchard) do have more venues returned.

(B) Do the results reflect the nature or characteristics of the neighbourhood?

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Jurong East	Japanese Restaurant	Shopping Mall	Chinese Restaurant	Food Court	Café	Department Store	Sandwich Place	Bubble Tea Shop	Bakery	Clothing Store
1	Marina Bay	Waterfront	Japanese Restaurant	Hotel	Salad Place	Cocktail Bar	Italian Restaurant	Sandwich Place	Gym / Fitness Center	Gym	Coffee Shop
2	Orchard	Hotel	Chinese Restaurant	Japanese Restaurant	Café	Boutique	Supermarket	Miscellaneous Shop	Cosmetics Shop	Coffee Shop	Cocktail Bar
3	Sentosa	Theme Park Ride / Attraction	Theme Park	Fast Food Restaurant	Café	Restaurant	Hotel	Beach	Food Court	Italian Restaurant	Resort
4	Yishun	Coffee Shop	Fast Food Restaurant	Indian Restaurant	Food Court	Grocery Store	Café	Supermarket	Shopping Mall	Chinese Restaurant	Ramen Restaurant

The top 10 venues / categories returned shows that the results are in line with the nature / characteristics for each neighbourhood:

- Marina Bay: waterfront
- Sentosa: theme park ride / theme park
- Orchard: hotel, boutiques
- Jurong: shopping mall
- Yishun: coffee shop

CONCLUSION

Though Foursquare is the lowest searched term among the selected location intelligence services, through the analysis, the results returned from the API search call are fairly consistent with the visitors' makeup (tourist / local) of that neighbourhood and highlighted its nature / characteristics. For example, Marina Bay and Sentosa ranked first and second in venues returned, are more of a tourist spot and the top venues for Marina Bay reflects its waterfront nature and as Sentosa has Universal Studio, it reflected theme park. Thus, it can be gathered that Foursquare results are fairly relevant for local context, with an inclination towards a tourist visitor base. Whether there are better alternative compared to Foursquare will have to be further explored.

It should be noted that this analysis is done with the free plan, which placed some constraints on the results returned.

Next steps could possibly be:

- Subscribe to the Premium plan offered by Foursquare to validate the results
- As Yelp also provides API calls, we could also make use of its services to do a further analysis.
- Follow up on local location intelligence services / apps for access to database to conduct same analysis

The python codes used for this analysis could be found in this link:

https://github.com/fongwc/Foursquare-Results-Relevancy-in-Local-Context-/blob/master/Foursquare%20Results%20-%20Relevancy%20in%20Local%20Context%20(Submission).ipynb