1. Introduction

As global mobility increases, more and more people are relocating to major cities around the world where there are greater demands for skilled workers. However, when relocating to other cities for work, people might be worried about the quality of life in the new city, especially if the city is unfamiliar to them.

Hence, this project is to do a comparison of major cities located in North America, Europe and Asia-Pacific to find clusters of similar cities. E.g. if one is moving to city A which is found to be similar to city B which the person is familiar with, he/she would have a better idea of how city A is like.

2. Description of data

A total of 30 major cities are selected, with 10 from each region. The cities are as follows:

- North America: New York, Washington DC, Boston, Houston, Chicago, Los Angeles, Seattle, San Francisco, Toronto, Vancouver
- Asia-Pacific: Tokyo, Osaka, Shanghai, Beijing, Seoul, Kuala Lumpur, Jakarta, Melbourne, Singapore, Bangkok
- Europe: London, Paris, Berlin, Frankfurt, Amsterdam, Madrid, Brussels, Copenhagen, Vienna, Zurich.

It is assumed that the expatriates prefer to live downtown, want to have access to wide choice of good restaurants and bars as well as a selection of leisure activities including sports. Hence, factors to consider can be divided into two main groups, namely lifestyle (restaurants & bars, sports/gym facilities, parks, cinemas/theatres) and convenience (supermarket, laundry, public transport).

In this project, the data to be used come from Foursquare's global database which contains data of 105 millions places across 190 countries. Based on the assumptions listed above, the relevant data that can be extracted from Foursquare API are selected as follows:

- Number of venues within 1km radius of the city centre in 8 different categories (restaurant, bar, gym/fitness center, park, movie theatre, supermarket, laundromat, metro station);
- Average number of likes for all the restaurants given in the search for number of venues;
- Variety score for restaurants within 1km radius of the city centre (at least 5 restaurants in each major cuisine type: Asian, French, Italian, Vegetarian, Steakhouse).

3. Methodology

In Section 2, the data to be extracted is described. There are a total of 30 sample points and 10 features. However, each feature has a different range of values as evident in the table below. For example, the number of venues is limited to 50 by the search engine. As such, the values range from 0-50. On the other hand, the average number of likes for the

restaurants does not have a hard bound to it. The last feature, variety score, ranges from 0-5, one point for each cuisine type.

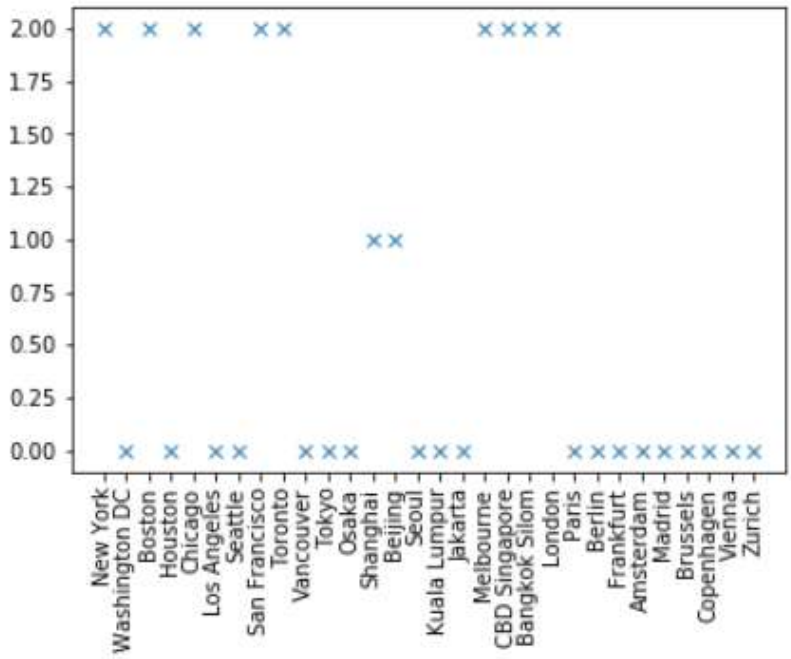
	restaurant	bar	fitness center	park	movie theatre	supermarket	laundromat	metro station	avg likes (restaurants)	variety score (restaurants)
New York	50.0	50.0	50.0	50.0	32.0	22.0	14.0	50.0	117.180000	4.0
Washington DC	35.0	50.0	50.0	50.0	10.0	0.0	0.0	50.0	149.314286	1.0
Boston	50.0	50.0	50.0	50.0	31.0	3.0	5.0	50.0	55.240000	2.0
Houston	37.0	50.0	50.0	50.0	15.0	0.0	0.0	50.0	28.675676	2.0
Chicago	50.0	50.0	50.0	50.0	50.0	0.0	0.0	50.0	60.240000	2.0
Los Angeles	50.0	50.0	50.0	50.0	26.0	1.0	1.0	50.0	42.900000	1.0
Seattle	50.0	50.0	50.0	50.0	24.0	2.0	0.0	50.0	63.020000	2.0
San Francisco	50.0	50.0	50.0	50.0	31.0	7.0	17.0	37.0	56.920000	3.0
Toronto	50.0	50.0	50.0	50.0	40.0	6.0	2.0	50.0	31.440000	4.0
Vancouver	38.0	39.0	35.0	39.0	5.0	0.0	3.0	9.0	18.394737	0.0
Tokyo	50.0	50.0	39.0	25.0	5.0	0.0	0.0	50.0	16.480000	2.0
Osaka	50.0	50.0	28.0	25.0	4.0	1.0	0.0	50.0	2.360000	3.0
Shanghai	44.0	18.0	25.0	3.0	0.0	2.0	0.0	14.0	23.795455	0.0
Beijing	20.0	6.0	4.0	3.0	0.0	1.0	0.0	5.0	3.900000	0.0
Seoul	50.0	50.0	50.0	45.0	6.0	3.0	0.0	50.0	35.340000	2.0
Kuala Lumpur	50.0	50.0	50.0	50.0	10.0	7.0	0.0	50.0	30.860000	3.0
Jakarta	50.0	50.0	50.0	50.0	8.0	2.0	0.0	50.0	19.500000	3.0
Melbourne	50.0	50.0	50.0	50.0	37.0	13.0	0.0	50.0	16.180000	5.0
CBD Singapore	50.0	50.0	50.0	50.0	27.0	5.0	3.0	50.0	55.800000	4.0
Bangkok Silom	50.0	50.0	50.0	50.0	20.0	7.0	0.0	50.0	15.440000	5.0
London	50.0	50.0	50.0	50.0	50.0	13.0	0.0	50.0	357.020000	4.0
Paris	50.0	50.0	23.0	30.0	25.0	2.0	0.0	50.0	42.740000	2.0
Berlin	50.0	50.0	50.0	21.0	4.0	0.0	0.0	31.0	42.680000	1.0
Frankfurt	50.0	50.0	50.0	39.0	2.0	1.0	0.0	26.0	20.440000	0.0
Amsterdam	50.0	50.0	50.0	47.0	10.0	5.0	1.0	50.0	88.560000	3.0
Madrid	50.0	50.0	38.0	50.0	4.0	2.0	0.0	50.0	11.380000	2.0
Brussels	50.0	50.0	50.0	50.0	18.0	3.0	0.0	50.0	8.720000	0.0
Copenhagen	50.0	50.0	50.0	36.0	2.0	0.0	1.0	24.0	14.060000	1.0
Vienna	50.0	50.0	50.0	50.0	7.0	1.0	0.0	50.0	31.840000	0.0
Zurich	50.0	50.0	48.0	37.0	7.0	0.0	0.0	18.0	18.920000	2.0

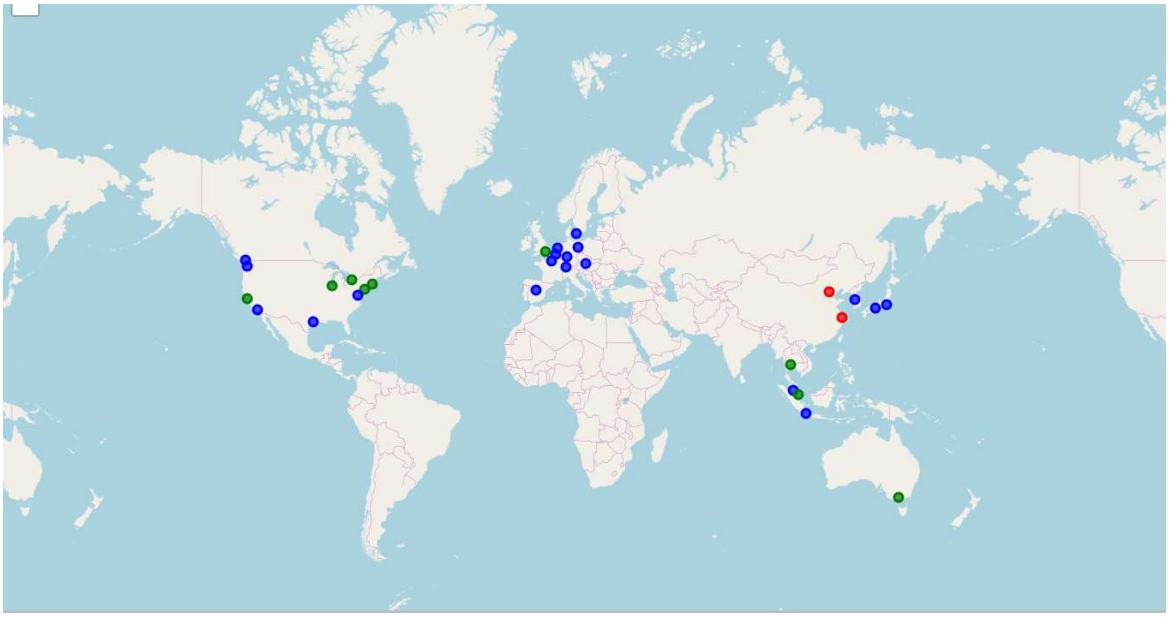
As such, it is necessary to do a pre-processing to condition the feature vectors individually so that their distribution is normalised to zero mean and unit variance. This will ensure that the results of the subsequent clustering algorithm are not skewed due to the differences in distribution of the features.

The k-means algorithm is then used to cluster the cities into 3 groups. The number of clusters is selected to be 3 as the objective is to find broad similarities between major cities and not to find fine-grain differences among the cities.

4. Results

The clustering results can be seen in the plot below. There are 19 cities in cluster 0, 2 in cluster 1 and the remaining 9 in cluster 2. The visualisation on the world map can be seen in the subsequent plot with the blue, red and green circles representing cluster 0, 1 and 2, respectively.





5. Discussion

There are only 2 cities in cluster 1 and these are the only 2 Chinese cities in the list of 30 cities. Checking back on the data, it is obvious that it scores poorly in many categories. It could be due to the relatively small database available in China as the Chinese users may prefer to use their own indigenous version of Foursquare in their native language.

Another observation is that the European cities are very similar, with 9 cities in cluster 0 (only London is in cluster 2). Similarly, North Asian cities (less Chinese ones) fall in the same cluster as the majority of the European cities. On the other hand, USA and Southeast Asian countries have a mix of both cluster 0 and cluster 2 cities. The only Australian city in the list is in cluster 2.

6. Conclusion

While the data used in this study is relatively small due to limitations from the use of Foursquare API and possibly bias in the data for the Chinese cities, this result offers an insight to the broad similarities of the cities studied and as well as the distribution of each category of city in the different regions.