Battle of neighbourhoods: an Overview

# Business Problem:

One of the main problems faced by retailers is knowing when, what kind of a customer might come in. If the retailers have a good understanding of their customer base, they can better cater to their needs, resulting in higher customer satisfaction and higher customer retention.

If the check-ins of customers to particular types of venues is analysed, one can predict, or atleast get a good idea on when what kind of customers go to which venue, a retail store, or a café, or a bookstore, or any venue, can expect them, and then do the necessary actions (repairing ramps for the elderly, stocking the item more popular amongst the youngsters at the front of the store to attract them etc.)

# Data

As the foursquare API is not working properly for me, I’ll be using the Dataset Global-scale Check-in Dataset by Dingqi Yang [1]

This dataset includes long-term (about 18 months from April 2012 to September 2013) global-scale check-in data collected from Foursquare. It contains 33,278,683 checkins by 266,909 users on 3,680,126 venues (in 415 cities in 77 countries). Those 415 cities are the most checked 415 cities by Foursquare users in the world, each of which contains at least 10K check-ins. Please see the references for more details about data collection and processing.

The dataset can be found [here](https://drive.google.com/file/d/0BwrgZ-IdrTotTXRiYXJkX2N3a0U/view?usp=sharing) (about 775MB zipped)

It contains three files in tsv format.

- **File dataset\_TIST2015\_Checkins.txt** contains all check-ins with 4 columns, which are:

1. User ID (anonymized)

2. Venue ID (Foursquare)

3. UTC time

4. Timezone offset in minutes (The offset in minutes between when this check-sin occurred and the same time in UTC, i.e., UTC time + offset is the local time)

- **File dataset\_TIST2015\_POIs.txt** contains all venue data with 5 columns, which are:

1. Venue ID (Foursquare)

2. Latitude

3. Longitude

4. Venue category name (Foursquare)

5. Country code (ISO 3166-1 alpha-2 two-letter country codes)

**- File dataset\_TIST2015\_Cities.txt** contains all 415 cities data with 6 columns, which are:

Venue category ID (Foursquare)

1. City name

2. Latitude (of City center)

3. Longitude (of City center)

4. Country code (ISO 3166-1 alpha-2 two-letter country codes)

5. Country name

6. City type (e.g., national capital, provincial capital)

Along with this, I will be using the User profile dataset by Dingqi Yang

This dataset includes some user profile data for privacy study (i.e., gender, #friends, #followers). It contains 18,201 and 11,874 users who have checked in New York City and Tokyo, respectively. The corresponding user check-in data can be found in the global-scale check-in dataset I published. The two dataset can be linked by the anonymized user ID (the unique key).

By linking the two datasets, we can find more insights on the customer base of any venue.