2. Data acquisition and cleaning

2.1 Data sources

To assess the best neighbourhood for a Mexican restaurant, we need data about the communities of Toronto and data about venues in the areas.

We will fetch data about Toronto' neighbourhoods from Wikipedia page with postal codes, boroughs, and corresponding neighbourhoods. Coordinates of areas we will get from CSV file kindly provided by this course.

Information about venues in each area will be obtained via the Foursquare API. We will take 200 places at max in a radius of 1 km from the centre of each neighbourhood.

After cleaning, it will be used to classify areas of Toronto and get the most suitable cluster for building a new Mexican restaurant.

2.2 Data cleaning

Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.

More than one neighbourhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighbourhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the communities separated with a comma, as shown in row 11 in the above table.

If a cell has a borough but a Not assigned neighbourhood, then the area will be the same as the borough.