# Data

The data used will include postal code data for boroughs and neigbourhoods, the co-ordinate information for the postal codes, and Four Square venue information. In order to use this data it will need to be imported from various sources.

Postal code data for boroughs and neigbourhoods will be scraped from Wikipedia. This is not necessarily the best source of data for many applications as it is unverified, however for basic data such as postal codes I’d consider it accurate

Co-ordinate information will be extracted from Statistics Canada’s Forward Sortation Areas (FSA) boundary files. This is a reputable source with accurate location data.

Venue information in each FSA will be provided by the Four Square API.

All this information will need to be combined into a single data frame from clustering analysis. Once the clustering is done, we’ll be able to paint a semantic map of the types of businesses within an area. This will allow us to identify potentially lucrative locations based on the amount of competition present and the likelihood of getting high foot traffic based on the number and types of venues within the FSA.