**Jim Keenan, Capstone Week 1, Battle of the Neighborhoods**

**Description of Problem**

As a Data Scientist, I would like to make use of the methods and tools we have studied to compare the similarities and dissimilarities of New York City, New York, and Toronto, Canada, which will then be compared to a new City, London, England. The reason for doing this is many people from the United Kingdom, including myself, move from Europe to The United States of America, and it would be useful to have a meaningful scientific study to show similarities and dissimilarities between the places they are moving from, in this case London, England, to either New York City, New York, or Toronto, Canada.

In that all three cities are the financial and cultural capitals of their respective countries, this information should prove useful, the similarities being a good recommendation factor for people who would like to move to North America ( USA and Canada ), but would like somewhere similar to where they had come from ( London ). Equally, there are some immigrants who may like to move to somewhere entirely different – the dissimilarities.

**Description of Data**

The data used to analyze similarities and dissimilarities will initially be found by looking at the Zip code (in the UK postcodes) in the central 5 boroughs of the given city. The data source for New York City which will be used is: <https://geo.nyu.edu/catalog/nyu_2451_34572>

The Data used for Toronto will be:

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

The Data used for London, England, UK will be:

<https://www.doogal.co.uk/UKPostcodes.php?Search=EC>

The three data sources will be used primarily to derive zip or post codes, Boroughs and Neighborhoods.

The fourstone.com Rest API will then be used to return three JSON dictionaries, the main key for which we will be concerned is Item, the contents of which are lists for which are broken down by locations in the boroughs by Venue name, Venue Category, Address, Zipcode and other interesting information, including web sites, photos and reviews. The Venue Category and lat/longs will then be used as features with quantitative measures to establish a model.

The Data will be grouped by venue type and count, by Borough, for all three cities. A key will be assigned for each city, NY, TC and LE and data merged to allow comparative studies on the summary data.

Based on the data for all three cities, a common Python function will be created to compare the summary information grouped by venue type and quantity by City.

A Python dataframe will be created for each city, which will find the 10 most popular venue categories by quantity and ranked for each neighborhood by city.

Using the 3 common dataframes, using K-Means Clustering to look at inter-clustering distances and inter-clustering, with an arbitrary K value of 5, that is, 5 clusters. Since K-Means relies on non-categorical data, all data will be normalized into numeric values. (contd. On next page)

Each of the 5 neighborhoods by city will then be mapped, and relative comparisons provided, based on similarities and dissimilarities.