CAPSTONE PROJECT (WEEK 3 & 4)
THE BATTLE OF THE NEIGHBORHOODS

BARDIYA B.

1. PROBLEM DESCRIPTION:

We want to have a way to point out neighborhoods with potentially unused business opportunities and suggestions for promising venue types in those neighborhoods, so that a human investor can look into the suggestions and make his decisions. We will try come up with such suggestions by using the information about the neighborhood, in particular information about the number of venues of other types.

The idea is that a large number of venues of some types should support and create business opportunities for other types of venues too. For example in an area with a lot of bars it is very likely that there is potential for food places being profitable too.

Now a simple idea is to make use of the face that in reality the potential is often already exploited and the occurrence of venues types that are supportive of each other should correlate relatively strongly. Therefore we could use some machine learning model (e.g. decision tree or neural network) to predict the number of actual venues of one type by the number of venues of other types in the same area. By then making the reasonable assumption that in the most popular/busy areas the actual number of venues is relatively close to the optimal number (if much larger some businesses would fail and the number should reduce, if much lower the potential is usually exploited until the number grows), we could then interpret a large deviation between predicted values and real values as either an oversaturation of the market, where it would be difficult to open a new venue or a not saturated markets with potential business opportunities.

However, such models would predict for highly correlated venue types A and B, whose occurrence is correlated because they are competing with each other, instead of supporting each other, in a certain area with a large number of venues of type A also a large number of venues of type B. This would most likely be a correct prediction, as business advice not helpful. Opening a large burger joint in an area with a lot of other food places, just because the machine learning algorithm suggests a large number of burger joints in that area too, might be not very wise. In fact we might want to place our burger joint into an area with fewer burger joints than expected and with a large number of venues that are usually highly correlated with burger joints, indicating that they might be supportive of our business and avoid areas with a lot of venues of other types, whose occurrence is usually also highly correlated with burger joints, due to its competing nature.

Furthermore, a full-fledged machine learning model might be also a bit of an overkill, since we are not in building a model that makes business decisions on its own, but only a system that points out possible business opportunities for a human investor to look into more closely. Therefore a simpler model, which allows easy interpretation of the data might be more suitable.

The idea we are going to present it to cluster the correlation matrix into different groups of venues, who are similarly correlated in occurrence with each other and also with venues types of other groups and then let a human decide which venues in a cluster seem to be supportive and which competitive of each other. If we then find an area with a lower number of a certain type of venue, than one would expect from the number of supporting venues we could suggest a business opportunity, especially if the number of competing venue types is low. In a way the clustering the correlation matrix and let a human sub-cluster the groups further would be just a method of feature-engineering for a model suggesting business opportunites

2. DATA:

The data we are going to use is the same as before.

1. *We are going to scrape the following Wikipedia site for information about Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M*
2. *We are going to use the CSV file provided in the previous assignments to get the coordinates of the neighborhoods of Toronto*
3. *We will use the Foursquare API to get information about venues in the different neighborhoods.*