Testing the Urban Convergence Hypothesis

1. Introduction

The hypothesis of urban convergence states that "urbanization creates similar landscapes regardless of the native biome in which cities are built" (https://heffernanlab.weebly.com/urban-homogenization.html), i.e. that cities become more alike over time. For the two cities that have been the subject of exercises during the course, New York City (NYC) and Toronto, I wonder whether this hypothesis holds today, i. e. that neighborhoods in these two cities cannot be distinguished from each other by formal criteria. Specifically, I am interested whether it is possible to predict, using location data on the venues in a neighborhood, whether a given neighborhood is located in NYC or Toronto. In line with The Urban Convergence Hypothesis, my prediction is thus that it is not possible to successfully predict, with high accuracy, whether a given neighborhood, characterized by Foursquare data, is located in Toronto or NYC. The results obtained in this project will be relevant for both academics, that work on the urban convergence hypothesis, and for tourism agencied of NYC and Toronto.

2. Data

To address this problem, I will use Foursquare location data on neighborhood venues that we have used in the previous exercises. For both cities, I'll use the following parameters for the Foursquare search:

VERSION = '20180605'
LIMIT = 100 (the top 100 venues)
Radius= 500 (500 meter radius)

I will use the neighborhood definitions used in the previous exercises. Data will be downloaded by executing Foursquare queries. All identifying information, such as venue names or geographical locations will be removed from data, to remain unbiased.

3. Methodology

The downloaded data will be split 70/30 into a training and a test dataset, and a Support Vector Machine (SVM) model will be trained. Using the test dataset, I'll predict whether a neighborhood is located in NYC or in Toronto and evaluate the accuracy of the model. SVMs "scale relatively well to high dimensional data" (https://statinfer.com/204-6-8-svm-advantages-disadvantages-applications/), and will therefore be used in the present analysis of high-dimensional Foursquare data. I'll train the model using an radial basis function (RBF) kernel, which is commonly used with SVM.

The *a-priori* set criterion for the acceptance or rejection of the hypothesis that the city cannot be predicted from Foursquare neighborhood data will be a Jaccard Similarity Score >0.7 (>0.7: hypothesis rejected, otherwise hypothesis accepted).

4. Results

The Jaccard Similarity index is >0.7 (0.7007168458781362, to be precise), therefore my hypothesis that the city cannot be predicted from Foursquare neighborhood data must be rejected. It appears that there are enough meaningful differences in neighborhood composition, evaluated using Foursquare data, between New York and Toronto. The cities thus have enough characteristic differences to discriminate them from each other using a SVM model.

5. Discussion

The analyses showed that the Urban Convergence Hypothesis does not hold when comparing NYC and Toronto, two metropolitan areas that are also similar in that they are the major financial centers of their respective countries. However, this project compared only NYC and Toronto. To obtain more evidence for or against the Urban Convergence Hypothesis, many other cities should be compared with the methods described herein.

6. Conclusion

The Urban Convergence Hypothesis did not to hold when examined in this project. Toronto and NYC both appear to have unique characteristics, a fact that may be exploited by the tourism agencies of both cities.