Testing The Urban Convergence Hypothesis

Urs Heilbronner

29th of July, 2019

The Urban Convergence Hypothesis

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.

Implications for Toronto and New York City

Toronto and New York City (NYC) are the financial capitals of their respective cities, and can thus be assumed to be relatively comparable in various aspects.

If *The Urban Convergence Hypothesis* holds, and the two cities have become rather alike, **one should not be able to predict whether a given neighborhood, characterized by Foursquare location data on venues, is located in Ney York City or Toronto**

Hypothesis of the Project

My hypothesis is that it is **not possible** to successfully predict, using machine learning, whether a given neighborhood, is located in Toronto or NYC.

Data Acqusision and Cleaning

▶ Neighborhood definitions of New York City and Toronto from the Coursera course are being used

▶ Data from Foursquare queries on venues in the different neighborhoods are being used

▶ All identifying information, such as venue names or geographical locations are removed from data, to remain unbiased

▶ Categorical venue data is recoded to dummy *yes/no* variables using one-hot encoding

Data Acqusisition and Cleaning

The following parameters were used in the Foursquare queries:

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

Support Vector Machines (SVM) Model

The downloaded data was be split 70/30 into a training and a test dataset.

- ▶ A SVM model was trained on the training dataset, and
- ▶ The test dataset was used to predict whether a neighborhood is located in NYC or in Toronto and
- ▶ The accuracy of the model was evaluated

A SVM was used because these models "scale relatively well to high dimensional data" (https://statinfer.com/204-6-8-svm-advantages-disadvantages-applications/).

The model was trained using an radial basis function (RBF) kernel, which is commonly used with SVM.

Evaluation Metric

The **Jaccard Similarity Score** (JSS) was used as evaluation metric.

The *a-priori* set criterion is that with a JSS > 0.7, my hypothesis that it is not possible to predict the city from neighborhood location data, will be rejected.

Results

The resulting JSS was **0.7007168458781362**, thus my hypothesis that I am not able to predict the city from neighborhood venues data **must be rejected**.

Discussion

It appears that there are enough meaningful differences in neighborhood composition, evaluated using Foursquare data, between New York and Toronto.

Conclusion

Both NYC and Toronto have enough characteristic differences to discriminate them from each other using a SVM model.

This may be exploited by the Tourism Industry of the respective cities, stating that the cities have a unique character.

Furthermore, these results are relevant to academic researchers that work on *The Urban Convergence Hypothesis*.