Investigating Effective Techniques on Shared Economy Lodging in NYC

Trang Nguyen & Tristan De Alwis

Problem Statement:

For our project, we will be looking at AirBnb data[1] within the geographic region of NYC and it's boroughs. The dataset we have sourced has over 48k listings with their associated attributes such as price, location, title, and reviews. Our focuses will be finding housing preferences such as location, type of listing (private/shared), the average review and how it may be correlated to price, what keywords in title appear the most often with higher booked rooms, seasonal trends, and, finally, the impact AirBnb has on the overall NYC lodging industry.

To find housing preferences we will be looking at the location attribute and determine the most popular locations to book.[2]

We will also look to see which type of listings are most popular to determine if customers prefer to book a private place to themselves or if a room within a shared apartment is prefered.

We will also look at the average score customers leave on a listing and determine the effect it has on the availability and price. To do so, we will first remove stop words (the, a, of, etc.). Next we will determine the topic of the review by using Latent Dirichlet Allocation to determine frequent n-grams. Lastly, we will use FP-growth to determine the frequent topics found in each review.

Also, we will be looking at the title of the listing to determine popular keywords[4]. We will use FP-Growth to determine what keywords stringed together make the listing more attractive to potential customers.

Lastly, with limited resources, we will try to determine if AirBnb has made a significant impact on the lodging industry[5] [6]. Questions we have include has the cost of lodging reduced? Is tourism more or less frequent?

From what we have read in current literature some of these topics have been looked into, but we will attempt to find improvements or compare methods.

Data Acquisition

We collected the dataset from Kaggle website (New York City Airbnb Open Data)
The data includes 16 attributes (title, neighborhood, price, reviews, and etc.) with over 760k data points.

Algorithm Choices

We are thinking about using few of these algorithms. It will be decided as we go further with the project.

- Principal Component Analysis to reduce the dimension of the data
- Topic modeling (Latent Dirichlet Allocation)

- Regression
- Clustering
- Apriori vs. FP Growth
- Visualization: Using Python library to illustrate descriptive characteristics of each Borough of NYC, correlation plot.

Reference

[1] “Get the Data - Inside Airbnb. Adding data to the debate.” 12 September 2019. Inside Airbnb. 05 November 2019. http://insideairbnb.com/get-the-data.html

[2] Quattrone, G., Greatorex, A., Quercia, D. et al.”Analyzing and predicting the spatial penetration of Airbnb in U.S. cities” EPJ Data Sci. (2018) 7: 31. https://doi.org/10.1140/epjds/s13688-018-0156-6

[3] Quattrone, G; Nicolazzo, S; Nocera, A; Quercia, D; Capra, L; (2018) Is the Sharing Economy About Sharing at All? A Linguistic Analysis of Airbnb Reviews. In: Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018). (pp. pp. 668-671). Association for the Advancement of Artificial Intelligence (AAAI): Stanford, CA, USA.

[4] Grbovic, Mihajlo, and Haibin Cheng. Real-Time Personalization using Embeddings for Search Ranking at Airbnb, ACM, 2018, doi:10.1145/3219819.3219885.

[5] ZERVAS, GEORGIOS, et al. “The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry.” Journal of Marketing Research (JMR), vol. 54, no. 5, Oct. 2017, pp. 687–705. EBSCOhost, doi:10.1509/jmr.15.0204.

[6] Wachsmuth, David, and Alexander Weisler. "Airbnb and the Rent Gap: Gentrification through the Sharing Economy." Environment and Planning A, vol. 50, no. 6, 2018, pp. 1147-1170.