Abstract

The goal of this project is to use Natural Language Processing (NLP) tools such as sentiment analysis and topic modeling on Airbnb reviews. By performing the analysis and topic modeling, the hosts can get a better understanding of what the positive or negative reviews left by previous guests and act accordingly. With this type of information, both hosts and guests will greatly benefit and improve the customer experience.

Design

The project was designed to show hosts the needs of guests in northern and southern California. Using data from InsideAirbnb.com, guests reviews left in San Francisco and Los Angeles was used to perform the analysis. The data didn't contain ratings, therefore, sentiment analysis was initially done to distinguish the positive and negative sentiments. Then, topic modeling was applied by region to get the full understanding of the complaints as well as the things guests enjoyed.

Data

This Airbnb dataset can be found at Inside Airbnb. Each observation represents a unique customer and the review they had on a particular stay. There was over 1.5million reviews. To access and view a detailed description of the dataset, click here.

Algorithms & Models

- CountVectorizer and TfidfVectorizer
- Non-Negative Matrix Factorization (NMF)
- Latent Dirichlet Allocation (LDA)
- Correlation Explanation (CorEx)

Tools

The following tools were used in this project:

1. Python, Pandas, NLTK, Spacy to:
    - Clean & Explore
    - Feature Engineering

2. SKLearn & NLTK tools to implement various topic modeling and sentiment analysis

3. Matplotlib and Seabn to visualize the data and model outputs

Communication

The findings and slide deck accompanying this project's presentation are accessible in this GitHub repository.