Airbnb Reviews Topic Modeling / Recommendation System

Sandy Weng

Abstract

The goal of this project was to use Airbnb reviews for NLP and topic modeling. I worked with data provided by Inside Airbnb. After topic modeling, I built a recommendation system and deployed it to Streamlit.

Design

This project addresses two applications that can be used by Airbnb. With topic modeling, Airbnb can gain insight on what users love and hate. They can work with the hosts and using the findings, they can improve customer experience from multiple angles. The reviews recommendation can help employees find the listings and reviews that are most similar to the one they're working on.

Data

The dataset I downloaded contains 740,667 reviews of listings in NYC from 2009 to 2021, with features including reviews, reviewer ID, review ID, and date of review. The average length of a review document is 48 words.

Algorithms

Natural Language Processing was used to preprocess the Airbnb reviews. CountVectorizer and TfidfVectorizer were used to turn review documents into sparse matrices for further use in modeling.

NMF, LDA, and CoreX were used for dimension reduction and topic modeling before settling on CoreX because it produced the most distinct topics.

Topics:

1. Location
2. Booking Logistics
3. Host Communications
4. Rooms
5. Cleanliness

SVD was used for dimension reduction and content-based recommendation system.

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn, Gensim, and CoreX for modeling
- Matplotlib, Seaborn, and pyLDAvis for data visualizations
- NLTK for NLP
- Streamlit for deploying recommendation system

Communication

- Slides and visuals