Project Fletcher - Using NLP and unsupervised classification to understand Airbnb reviews

Background

The website insideairbnb.com has an extensive dataset of Airbnb reviews, along with corresponding listing information scraped from Airbnb.com. It includes multiple reviews per listing, information about the price to rent each property, and more. The data can be found at http://insideairbnb.com/get-the-data.html. For my project, since my project #2 focused on Seattle's real estate market, I'd focus on a city outside of Seattle. I am planning to do Portland for the time being.

Approach

I'd like to look at the overall "feature space" of each review and see if they are well categorized into several different types of reviews - for example, "short and sweet", "explaining amenities in detail", "gripes with the property", "gripes with the owner", and "customer disasters", for example. I think that these reviews will have very different characteristics - amount of words, words used, length of the review, emotional polarity of the review, etc. With sample size permitting, I'd like to use this information to predict the cost of the rental, and determine what characteristics cause a renter to stay in a property again.

I'd also like to use analysis of each property to create a tool (and possibly associated Flask app?) to highlight pros and cons of each property using machine learning. In essence, this would highlight issues with current properties for current owners in a quantitative way that would be easily digestible.

Additional Sources of Data

I may bring in Portland hotel review data from Kaggle (located at https://www.kaggle.com/datafiniti/hotel-reviews). This would provide an interesting comparison between Airbnb and hotel reviews.

Concerns/Known Unknowns

I'd like to be able to predict the star rating of each property, but we don't have that data. I'd also like to classify each property as commercial owner vs. mom-and-pop owner, but I can't think of a way to classify this.

My major concern is that doing another project in the real estate space might feel like a repeat of Project Luther, and I could better utiliize this project to focus on a different domain. However, the scope and purpose of the project is totally different (linear regression/supervised learning vs. NLP/unsupervised learning), so I think that'll work out nicely. If anything, I like the real estate space because it allows me to do a lot of additional feature engineering and add some domain knowledge.