Neighborhood and Score Prediction for Airbnb Listings

Kunal Sangani and Emily Tang

Research Question

We explore an Airbnb dataset that includes all listings in San Francisco. We seek to predict three outputs using text and feature data:

(1) Neighborhood: Predicting neighborhood from listing data provides insight into the diversity of neighborhoods and may pave the way for future Airbnb recommendation systems (e.g. "If you enjoyed your stay in the Haight-Ashbury, maybe you should check out the Mission!")

(2) Cumulative review score: Detecting listings that may have disingenuous information or hosts that may not be invested in creating a good experience for visitors

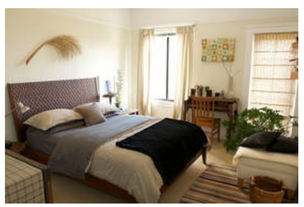
(3) Price: Providing recommendations to Airbnb property owners about appropriate pricing for their listings

Data

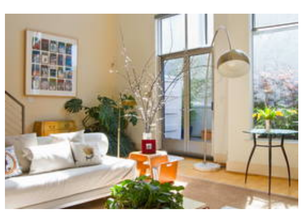
The "Inside Airbnb" project includes detailed information about all listings available in San Francisco as of November 2, 2015. Each listing includes:

- Text descriptions (name of the listing, summary, space, description, experiences offered, neighborhood overview, transit, host bio, and other notes)
- A thumbnail image of the listing
- Data about the accommodation (number of people accommodated, amenities available, number of bathrooms, bedrooms, and beds)
- Extra information about the listing (shared or private room, type of bed, the host's cancellation policy, etc.)
- The neighborhood of the listing
- The price of the listing
- A cumulative review score of the listing

ID 8339: ***"Historic Alamo Square Victorian"***
"European feel, reminiscent of the Boutique Hotels in Paris, Berlin, Antwerp and London. This spacious sunny apartment is well known by local photographers for its ambiance and beautiful natural light Safe walkable neighborhood. About the host: Always searching for a perfect piece at European flee markets and design shops!"

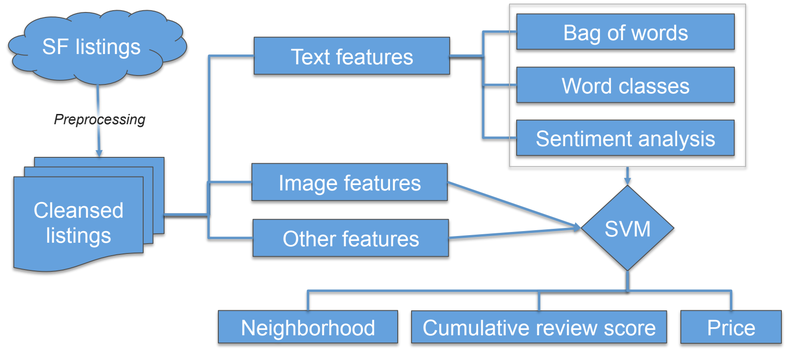


ID 25463: ***"Modern Zen in the Lower Haight"***
"Located in one of San Francisco's most bohemian and central neighborhoods, I offer a large sunny room with a very comfortable queen bed and shared bathroom. My home is one block from the trendy shops, restaurants, cafes, art galleries and salons of the Lower Haight."



ID 25827: "***Chic Loft in SOMA***"
"Walking distance to everything; night clubs, union square, Bart, SFMOMA, etc. Professionally decorated condo in modern building that is a rear unit with a large private patio with BBQ and Heat Lamp for late night star musing. Amazingly quiet day & night in spite of being just blocks from nightlife legends. Two bikes available for longer day trips around town."

Feature Extraction



Bag of Words features
- Removed all stop words and all words associated with neighborhoods
- Used NLTK PorterStemmer package to reduce all words to stem
- Chose 1000 words that appear in the most listings
- Counts of each word in all text associated with each listing, normalized to sum to 1

Bucketed text features
- Handpicked 9 word classes (people, nightlife, day-time activities, style, accessibility, culture, nature, amenities, comfort) with pre-defined lists of associated words
- Number of words from each bucket in listings' text descriptions

Miscellaneous features of listing
- Property type (e.g. apartment), listing type (e.g. entire house, private room), bed type (e.g. full bed, airbed, couch, futon), cancellation policy
- # of bedrooms, # of bathrooms, # of baths, square-footage, and # accommodated

Sentiment analysis
- Calculated the polarity of various descriptions using the TextBlob Python package
- Extracted sentiment from the following: Summary, Space, Description, Experiences offered, Neighborhood overview, Host bio, Notes

Image features
- Created a visual word dictionary of 100 words using OpenCV SURF descriptors and K-means clustering
- Extracted SURF descriptors from each image, and built visual feature vectors using the dictionary and each descriptor

Methods

We split the data into train, dev, and test sets (80/10/10). For review score and price prediction, we use buckets, determined by the median and quartiles, for our labels. We build two classifiers, Support Vector Machines (SVM) and Logistic Regression, using the balanced class weight setting. For SVM, we tune the C parameter using the dev set.

To better understand our data, we also use K-means clustering to cluster the bag of words features and create word clouds of some of the most common words used.

Results