Housing Rental Text Analysis

Yunhan Xu
Project Proposal
CS 109: Data Science
Fall 2016

Motivation

The success of Airbnb, an online marketplace that enables travelers to rent accommodations from hosts, represents the rise of "peer-to-peer Internet-enabled … markets for durable goods in which consumers … trade their durable assets in (traditional) secondary markets."[1]

One interesting feature of these online sharing economies is the **review system**. In the case of Airbnb, hosts and guests may review their experience in 500 words or less at the end of each stay. One Airbnb host describes this review process as "absolutely vital," as it "creates incentives for hosts to go the extra mile, and for guests to follow the campsite rule."[2]

In addition to creating a sense of accountability, reviews provide useful information to travelers about what to expect of their rental experiences. Qualitative descriptions ("small, but charming," "noisy at night," "free coffee") are also often far more insightful than numerical ratings.

Objective

The objective of this project will be to perform **topic modeling** and **sentiment analysis** on Airbnb review data to determine what makes a rental experience good or bad. Specifically, we will answer the following questions.

1. What characteristics do travelers care about?
2. What characteristics are associated with positive rental experiences?
3. What characteristics are associated with negative rental experiences?

[1] S. Fraiberger and A. Sundararajan. Peer-to-Peer Rental Markets in the Sharing Economy. SSRN. March 2015.

[2] S. Porges. "The Strange Game Theory Of Airbnb Reviews." Forbes. October 2014.

Secondarily, we will examine how these text-based signals correspond to quantitative signals (i.e. price of listing, minimum nights per stay) across listings.

Data

We will be using review and listing data from Inside Airbnb, an independent data repository with coverage across 40 cities in 15 different countries. Each city has order of magnitude 1,000 - 10,000 listings (rows). Links are provided below.

Inside Airbnb Home
Inside Airbnb Data

Milestones

---

1. Project selection

Form a team of 3-4 students and select a project.

2. Literature survey

Find three or more papers related to the methods used in this project and summarize them in a one-page literature review. Some resources, which you may use for background reading or formal literature review, are included below (you may also find other papers).

Text Analysis

Tips for Computational Text Analysis

Topic Modeling

Probabilistic Topic Models
Modeling Online Reviews with Multi-grain Topic Models
A Topic Model for Movie Reviews

Sentiment Analysis

A Study and Comparison of Sentiment Analysis Methods
Sentiment analysis using product review data

Stanford CS 229 Final Projects

Sentiment Analysis for Hotel Reviews
Sentiment Analysis of Users Reviews and Comments
Yelp Personalized Reviews
Sentiment Analysis on Movie Reviews
Prediction of Yelp Review Star Rating Using Sentiment Analysis
Vector Based Sentiment Analysis of Moview Reviews
Predicting Helpfulness Ratings of Amazon Product Reviews
Multiclass Sentiment Analysis of Movie Reviews

Kaggle

Sentiment Analysis on Movie Reviews

Airbnb

A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average

3. Data exploration

The purpose of the exploration phase is to both prepare the data for analysis and to observe (and visualize and interpret) interesting trends in the data that are worth exploring in greater depth later. Specifically, the following steps are required.

- Obtain review and listing data for one city. Make sure that review coverage is robust enough to support analysis for this city.
- Choose a database to store the data and a computing resource to process the data. Given the size of the review data (> 50 MB) and the heavy computational lifting required of text analysis, Spark is recommended.
- Clean the listing data. Make sure that units for each column are standardized (i.e. rental prices are for the same duration of time).
- Clean the review data using natural language processing techniques (remove unnecessary characters and words from the corpus, convert reviews to “bag of words” format for Naive Bayes classification, extract different parts of speech, etc.). The quality of your data cleaning will affect the quality of your analysis and results (GIGO), so proceed with care and justify your decisions and assumptions!
- Create a vocabulary to prepare the text data for analysis (transform the reviews so that each review stores both the vocabulary size and word frequencies as a sparse vector).
- Explore and visualize the listing data. How are rental prices distributed? Availability? Minimum nights per stay? Interpret these visualizations and check for correlations among these variables.

- Explore and visualize the review data. How are review lengths distributed? Check for correlations between review and listing variables (do higher-rated rentals have longer reviews?).

4. Proposal and implementation

The final step is to propose and implement methods to answer the motivating questions outlined in the Objective section. Specifically, the following steps are required.

- Decide on the metric(s) you will use to evaluate performance.

- Implement Latent Dirichlet Allocation (LDA) as a baseline for topic modeling. LDA is an unsupervised algorithm, so there is no “performance” to measure - however, you should print and interpret random samples of the topic clusters you generate.

- Implement Naive Bayes as a baseline for sentiment analysis. Evaluate performance through the metrics you defined.

- Implement a non-baseline method (SVM, random forest, etc.) for sentiment analysis. Evaluate how this method performs against Naive Bayes.

- Incorporate quantitative variables from the listings data into your analysis by joining these rows based on listing ID. Interpret any correlations you observe.