CS 4300 Final Project: Airbnb Similar Listings

Ambar Soni (as2495), Yunjie Bi(yb89), Liam Bui (lb598), Laura Wade (ljw96)

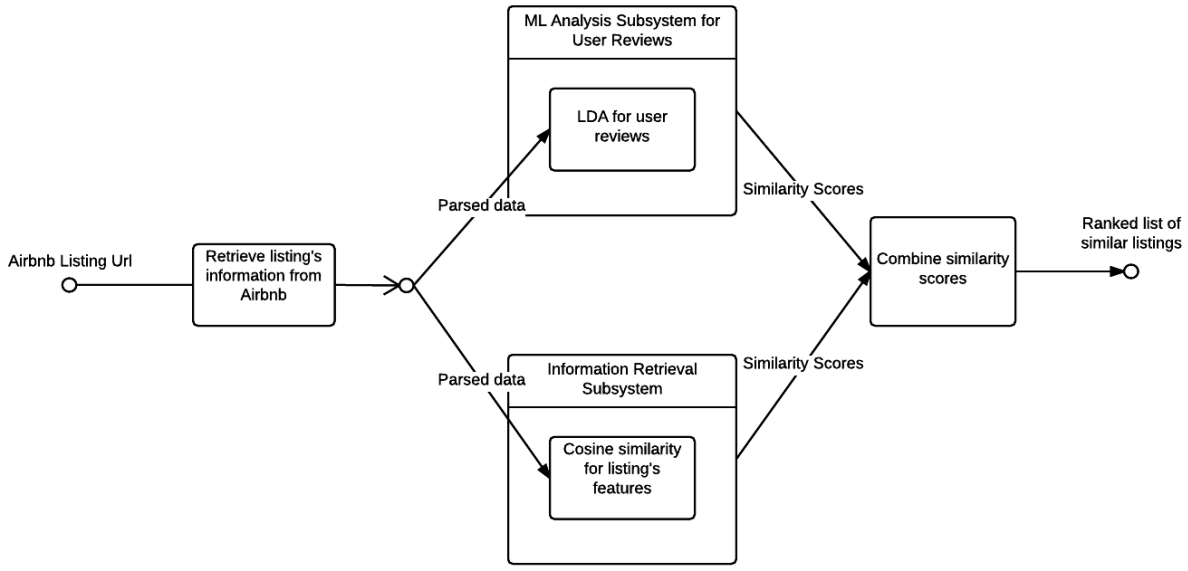
April 22, 2016

1 Description

1.1 Goals

Airbnb has become one of the top housing choices people consider when they travel, especially when the destination is known to have high hotel rates such as New York City and San Francisco. However, you are not always guaranteed to have the best listing you find since the host needs to approve before you can make a reservation. Airbnb users who have a good experience with one listing might want to find similar listings in other cities for future travel. We want to build a system that can give you a ranked list of similar listings in the user's desired city by using both the set of features of the original listing and the written reviews from users.

1.2 Workflow Diagram



Given the input URL of the listing, the system first validates that the listing is valid and is currently active. Next the system compares the features of the input listing to those of the ones in the dataset to generate a ranked list of similar listings in the New York City or San Francisco area. Our dataset for both cities is small enough to fit in memory and we can create a TF-IDF matrix in resonable time. When analyzing reviews for the listings, we will be utiltizing the Latent Dirichlet Allocation (LDA) to create topic models for the content. Lastly, we will output a list of similar listings for both SF and NYC to the user.

1.3 Machine Learning

We plan on using Latent Dirichlet Allocation (LDA) to retrieve common topics from user reviews of the original listing and of a candidate listing. LDA is a generative statistical model that doesnt require labeled training

data, so our data would suffice. The set of the features we are considering are: price, occupancy, amenities, room type, and reviews.

1.4 Information Retrieval

We can use cosine similarity with tf-idf scoring to determine how similar two listings are based on their respective sets of features. The final ranking of listings would be the result of combining user reviews similarity and the features similarity between the original listing and the candidate listings.

1.5 Social

We will be utilizing user reviews from Airbnb as a measurement of similarity.

2 Input and Output

The input of our system will be an url of any Airbnb listing, and the output will be a ranked list of similar listings in NYC as well as a ranked list of similar listings in SF. For now we are using only data from San Francisco and New York City, but our hope is that this system will be extended to accommodate more cities. This restricts the user's desired city to either New York City or San Francisco.

3 Use Cases

Potential use cases include the following scenarios. When a user, who plans to visit NYC, is rejected by the host after requesting to book a certain listing, the user can enter the url of the original listing into our application and receive a list of similar listings in NYC so the user can still find homes that are not too far off from his/her first choice. A slightly different use case would be if the user accidentally forgets to put in exact start and end dates of the intended stay and finds a great listing, but only to find out later on that this listing is already booked for the time frame the user is aiming for, a ranked list of similar listings in NYC would be very helpful. Or when an user, who has already visited NYC, loved his/her stay there and plans on traveling to SF soon, wants to find a listing thats similar to the past NYC listing, the user can also get a ranked list of similar listings in SF to choose from.

4 Data

4.1 Data Sources

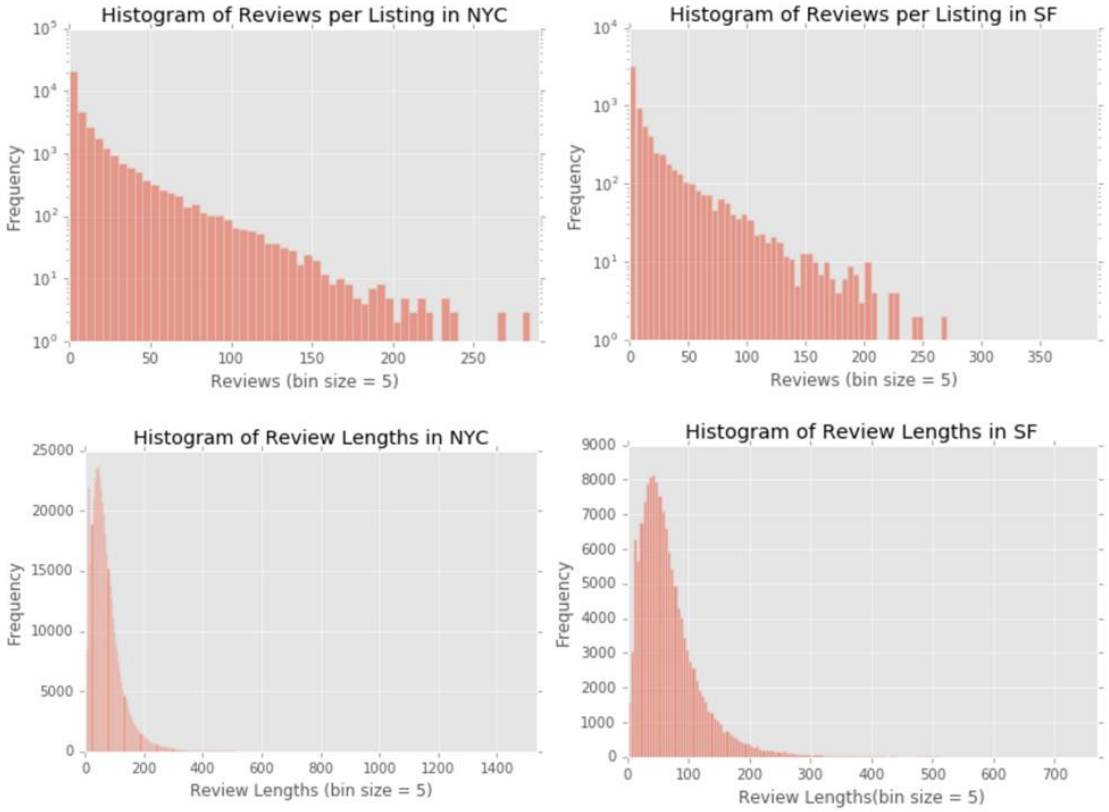
- http://airbnb.io/airpal/
- http://insideairbnb.com/get-the-data.html

4.2 Data Exploration

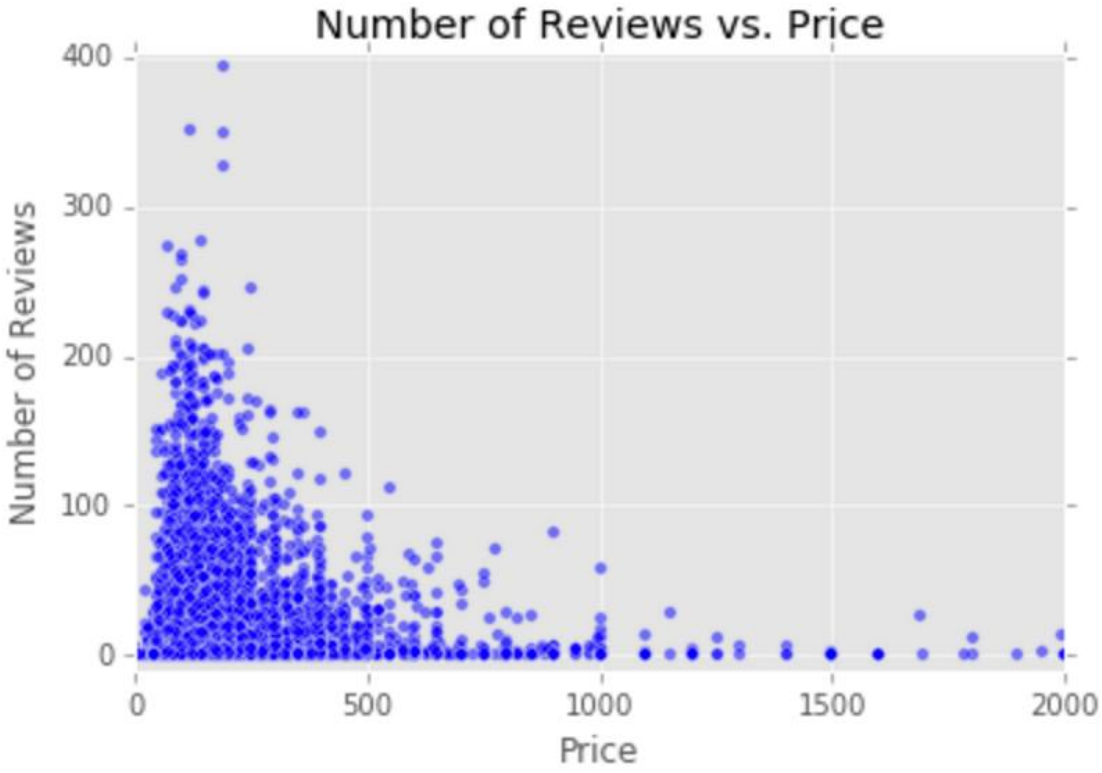
- **Data Description** The Inside Airbnb dataset includes separate reviews and listing information for major international cities. From the listings data, several key features are included, such as the name, URL, descriptions, host information, location, room types, number of bedrooms, number of bathrooms, bed types, amenities, price, space, and availability. From this data, there are numerous ways to extrapolate statistics, such as the number of reviews versus price, number of reviews per listing in a given city, as well as more specific, interesting statistics, such as number of houses that are more likely to be rented out to tourists rather than long-term residents. Some statistics are already built into the data, as well, including number of reviews per month, host acceptance rate, and availability per year. Many filters and other metrics are provided from the listings data, which are non-essential, such as host name and last update of the listing itself. From the reviews data, however, we have only the key features, including the listing, date, reviewer, and comment raw text.

From the data, we do not expect to use all features provided, because there are numerous non-essential features, such as the host name or validity. From the data, an intuitive idea of similarity metric is based on price, location relevance, amenities, number of rooms, and amount of space. Many non-listed features can also be extrapolated from reviews, including general sentiment, though this may be supplemental. These features will be weighted to create a similarity metric between two listings in different cities. There are approximately thirty-five thousand data points from New York, New York, and approximately seven thousand data points from San Francisco, California. These will provide ample data to extrapolate statistics and create a reliable similarity metric.

- **Basic Statistics**:
  - Number of reviews: San Francisco 143237, NYC 439092
  - Number of listings: San Francisco 7029, NYC 35957
- **Plots**
  - The histograms depicting reviews per listing are on a log scale. We can see that most listings get around 80 reviews in New York City and around 70 in San Francisco. The review lengths histograms show that they are approximately lognormally distributed. The NYC distribution tails out faster than the San Francisco plot.



  - From this plot we see that the lower the price of an Airbnb listing the more likely it is to have a lot of reviews. This indicates that price is an important factor for Airbnb users when determining where to stay. Thus price will serve as a significant weight when determining if a listing is similar to the given listing.



- **Pros**:
  - Contain all features needed for project
  - Precompiled CSV files for each city (listings and reviews)
- **Cons**:
  - Highlight data from Airbnb requires scraping
  - Access to pre-loaded similar listings, which may be used for evaluation of built ranked list

5 Evaluation

Our primary evaluation metric will be to examine the highlight section of the AirBnb post. On each post this section presents a set of keywords from the entire listing that users have found most relevant. When comparing the initial listing with a similar listing that we propose, we can verify with a count of overlapping keywords from the highlights section. We will obtain this information for each posting by scraping the listing when the search is run.

There are really no other direct measures to evaluate our systems performance since theres no numerical value that can represent the accuracy of our output. Similarity between two listings is very subjective. Airbnb itself does already provide similar listings when a user is browsing, so we could compare our results to theirs, but this wouldnt be very optimal either since we could be using a set of different criterias when evaluating two listings. Therefore, the most practical way to evaluate our system would be using the highlights section as proposed above and human judgements.

6 Weekly Schedule

1. Week of 4/11:
   - 4/15 Submission of updated project proposal

- Data preprocessing - e.g. only saving the fields that are related to the project

2. Week of 4/18:
   - Build the ML part of the system
   - Build the IR part of the system
3. Week of 4/25:
   - Build the UI of the web app
   - 4/26 Project Madness 3-5min presentation
   - Run tests
   - Tune the system - e.g. increasing performance
4. Week of 5/2:
   - More testing before final presentation
   - 5/3 and 5/5 final presentation
   - Start working on the final writeup
5. Week of 5/9:
   - Wrap up the final writeup

6.1 Division of Labor

As we proceed in our project, we have come up with four essential roles that we need to deliver on time efficiently:

1. Analysis of reviews data with LDA topic modeling (Frontend) → Liam
2. Intgration of system with frontend via heroku (ML + Reviews) → Maggie
3. Generate TF-IDFs for all features and create final ranked list (Backend) → Ambar, Laura

7 Literature Comparison

There have not been any similar projects created for finding similar listings in different cities; however, Airbnb provides listings that are similar to any given room from the same city. This, however, does not provide an accurate validation of the ranked list that we will create for any given url. The Airbnb provided similar listings is based dissimilar metrics, as well, including the listing quality (review quality and quantity, and frequency of guest clicks), ease of booking (response rate, response time, and reliability), and particular guest preferences (location relevance, social connections, and preferences for specific amenities). The metrics we intend to use for creating ranked lists of similar listings based on any given url include specific features that we determine to be relevant in providing similar listings. This, while is not specific to the personalized user preferences, does take into account additional features that are not considered in the Airbnb similarity metric. For example, our generated ranked lists may take into account costs and type of room (e.g. apartment or house), which may provide a more preferable metric to the user. The Airbnb similarity metric only personalizes user preferences for amenities and location relevance, but not features such as cost, room type, property type, number of beds, bed type, availability, and maximum length of stay.

Thus, in some ways, the Airbnb similarity metric for listings does not take into account several key features that our ranked list will be taking into account. As a result, the validation of similarity between our constructed ranked lists and Airbnb similar listings may not necessarily be analogous, but nonetheless, the Airbnb provided similar listings do provide a generalized similarity baseline for same-city similar listings.