



IS450- Text Mining and Language Processing

**Project Proposal**

**Submitted by:**

Ahmad Saifullah Bin Mustaffa, Gab Min Xuan Darren, Lu Yun Jing Ada,

Muskaan Gupta, Shubham Periwal, Vikram Sanghi

**Submitted to:**

Professor Swapna Gottipati

# **Introduction**

Airbnb is a popular platform for travellers to book their accommodation when they go overseas. All the listing information is provided exclusively by the Airbnb host, with limited ways for the guest to confirm its accuracy. Currently, Airbnb has the following aspects for reviews, along with the overall rating on a scale from 1 to 5 stars:

* Accuracy of listing
* Communication with hosts
* Cleanliness
* Location
* Check-in
* Value

These metrics, while useful, are relatively general and are only shown on an overall scale. Customer review data (in text format) contains hidden information that cannot be shown on a ranking scale. The aim of our project is to use language processing techniques to extract relevant and useful information from the unstructured textual data and to analyse it to aid decision making for users. This analysis, in turn, will also help hotel managers to make better managerial decisions.

Our project can be accessed on: <https://github.com/dagazzg/IS450-project>

# **Data Set**

The data is sourced from the Inside Airbnb website: <http://insideairbnb.com/get-the-data.html>

The dataset to be used for this project is obtained from Inside Airbnb, an independent website that scrapes publicly-available listing information from Airbnb and cleans and aggregates the data.

For the project, we will mainly be using the following files, focusing on Los Angeles, USA as our primary location.

|  |  |
| --- | --- |
| **File** | **Description** |
| listings (unzipped).csv | Detailed “Listings” data for listings. Some of the attributes used in our analysis are “price” (continuous), “longitude” and “latitude” (continuous), “room\_type” (categorical), “neighbourhood” (categorical), “host\_is\_superhost” (categorical), “review\_score\_rating” (continuous), etc. |
| reviews (unzipped).csv | Detailed “Reviews” data on listings with 6 attributes, contains the raw test comments which will be used extensively |
| neighbourhood.csv | A list of all the neighbourhoods in LA |

# **Features & Methodology**

**Feature #1: Aspect-based search for listings / Ranking top listings according to aspects**

1. **Purpose:** We will perform an exploratory analysis of the Airbnb dataset to understand the rental landscape in LA through various static and interactive visualisations.
2. **Methodology:** Exploratory Data Analysis and Visualization of Airbnb Dataset
   1. How do prices of listings vary across the location?
   2. How does the demand for Airbnb rentals fluctuate across the year and over years?
      1. “price” under listings.csv will be manipulated for time-series analysis
      2. “date” under listings.csv and reviews.csv will be transformed to obtain weekly, monthly or yearly insights
   3. Are the demand and prices of the rentals correlated?
   4. What are the different types of properties in LA? Do they vary by neighbourhood?
   5. What localities in LA are rated highly by guests?
   6. What makes a host Super host? **(Appendix: Exhibit A)**
      1. Airbnb has a set of requirements that must be fulfilled in order to become a super host. Maintaining a review rate above 50%, a response rate above 90%, etc. here we investigate our dataset to see how the super hosts perform on two parameters: “Response Rate” and “Ratings”
      2. listings.csv contains “location rating”, “cleanliness rating” and “overall rating” which will be used for analysing reviews **(Appendix: Exhibit B)**
   7. Do regular hosts and super hosts have different cancellation and booking policies?
   8. Common themes that can be identified from the free-text section of the reviews?
   9. What aspects of the rental experience do people like and what aspects do they abhor?

**Feature #2: Summary of Listings- Text Analysis**

1. **Purpose:** To provide summarized results for each listing showing **strengths**, **weaknesses** and the **change in reviews sentiment** across time for a listing.
2. **Methodology:**
3. Data pre-processing (Tokenisation, stopword removal, lemmatisation)- we will use “comments” from the reviews.csv file.
4. Extracting keywords to identify commonly-mentioned items (bag of words, feature selection)
5. Use of bigram model to find words that frequently occur together in reviews **(Appendix: Exhibit E)**
6. Using either Gensim or Scikit to train a topic model to identify a number of topics using Latent Dirichlet Allocation (LDA), a probabilistic model **(Appendix: Exhibit C)**
   * 1. Identifying the keywords of each topic and their weights
     2. Sentiment Analysis:
        + For each listing clean the reviews
        + Provide a sentiment score to each review and add it as a new variable for each row
        + Provide access to each listing for visualizaing how sentiment score has changed over time
        + Identify the most common topics mentioned across reviews, use the topmost associated keywords of those topics to create a summary of the Airbnb listing that showcases its’ most common strength and shortcomings (areas to improve) so that users can easily get information on the listing based on the opinions of others who have stayed there before.

**Feature #3.1: Topic Analysis**

1. **Purpose**: Analyse what features users look for in a hotel and what they comment about. **This product will help host in understanding what travellers are looking for in lodging to assist them in better decision making.**
2. **Methodology:** Pre-define certain categories (bed, service, etc)
   1. Extract topics and clean the data for each review
   2. Remove stopwords
   3. Lemmatise the data
   4. Extract important features
   5. Store sentiment along with the features
   6. Create a frequency diagram of the number of references to each of the pre-defined categories
   7. For visualisation, we can create an n-gram word cloud **(Appendix: Exhibit D)**
   8. Aggregate all topics and sentiments to the hotel level

**Feature #3.2: Cluster Analysis**

1. **Purpose:** Clustering reviews based on certain traits i.e. area with ‘good’ food. **This will help travellers choose their lodging based on what is important to them.**
2. **Methodology:**
   * 1. Potential clusters
     2. Cleanliness (“clean”, “dirty”, “smell”, “hygiene” etc.)
     3. Bedding (“bed”, “pillows”, “mattress” etc.)
     4. Kitchen amenities (“kitchen”, “utensils”, “cook-”, “grease” etc.)
   1. Load all reviews into a single corpus
   2. Pre-process and clean (remove stop-words, lower-case, stemming etc.)
   3. Transform corpus into a vector
   4. Carry out Hierarchical Clustering
      * 1. Gather insights by identifying the most talked-about issue amongst travellers
        2. Could be useful information to hosts

**Feature #4: Chat Box**

1. **Purpose:** To make it easier for our users to specify their requirements instead of using filters
2. **Methodology**
   1. Create a greetings dictionary so bot can at least reply to greetings
   2. Create word vectors with Spacy
   3. Entity Recognition and intent recognition
   4. Use Rasa to create an interpreter model
   5. Connect with SQL
   6. Extract query and respond with answer
   7. Incremental slot filling and negation - if user does not accept the answer given the first time
   8. Add tastefulness

**Feature #5: Dashboard**

1. **Purpose:** To present our analysis results in the form of various visualisations on a single webpage
2. **Methodology:**
   1. Construct visualisations e.g. Tag Cloud, Jigsaw
   2. Consolidate all visualisations on a single platform e.g. Plotly, Node.js
3. **Challenges and Mitigation Plan**
4. **Sheer size of the data:** The size of dataset for Airbnb listings is too large for a complete analysis within the given time frame and resources available. Thus, to mitigate the risk of over scoping we are limiting the project to the city of Los Angeles, California, United States.
5. **User Adoption:** This challenge mainly arises due to the unforeseen results of our model. Our model might not prove to give significant results to aid decision making. In order to mitigate it we have planned a structured approach to each model by first testing it on a smaller dataset based on user questions gathered from our mini survey. This will ensure a greater user value add.
6. **Data Veracity:** Since the entire project is greatly dependent on the data collected from the open source platform on airbnb reviews and listings, any white noise within the dataset might have a significant impact on the model. To mitigate this, we have planned an overall data cleaning process followed by a separate data cleaning unique to each model ensuring as little noise as possible is fed to our project. The data also had null values. To preserve all the information, we imputed or dropped the rows and columns containing null values while conducting exploratory analysis that made use of these features
7. **Appendix**

**References:**

1. <https://machinelearningmastery.com/multi-class-classification-tutorial-keras-deep-learning-library/>
2. <https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff>
3. <https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/>
4. <https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec>

**Must have/ Good to have:**

**Good to have**

**Must Have**

Feature #1: Aspect-based search for listings

Feature #2: Listings Summary

Feature #3.1: Topic Analysis

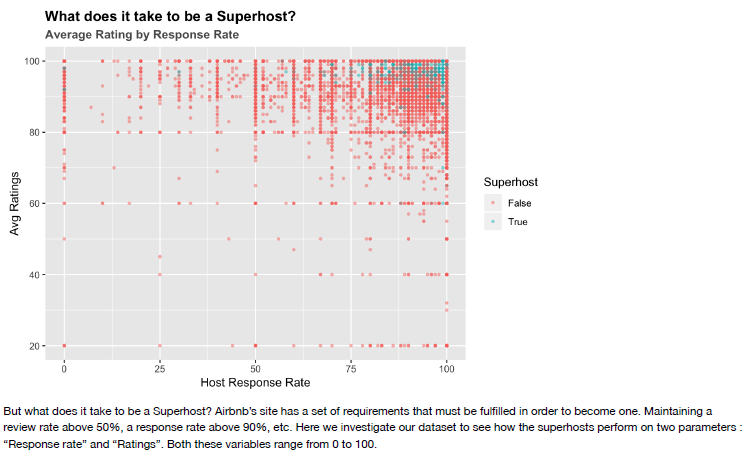
Feature #3.2: Cluster Analysis

Feature #4: Chat Bot

Feature #5: Dashboard

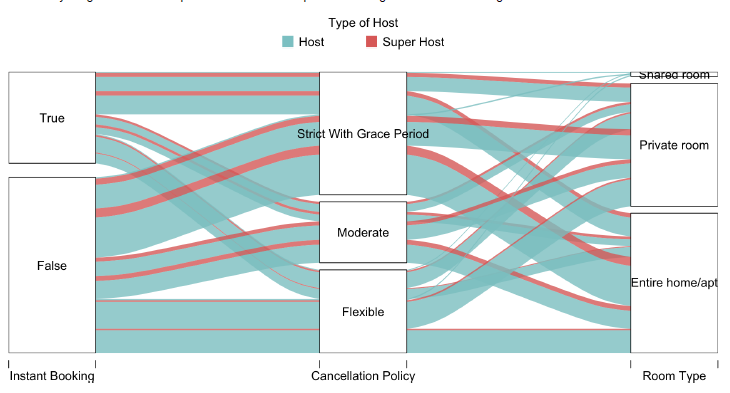
**Exhibit A:**

What does it take to be a super host? Average Rating by Response Rate



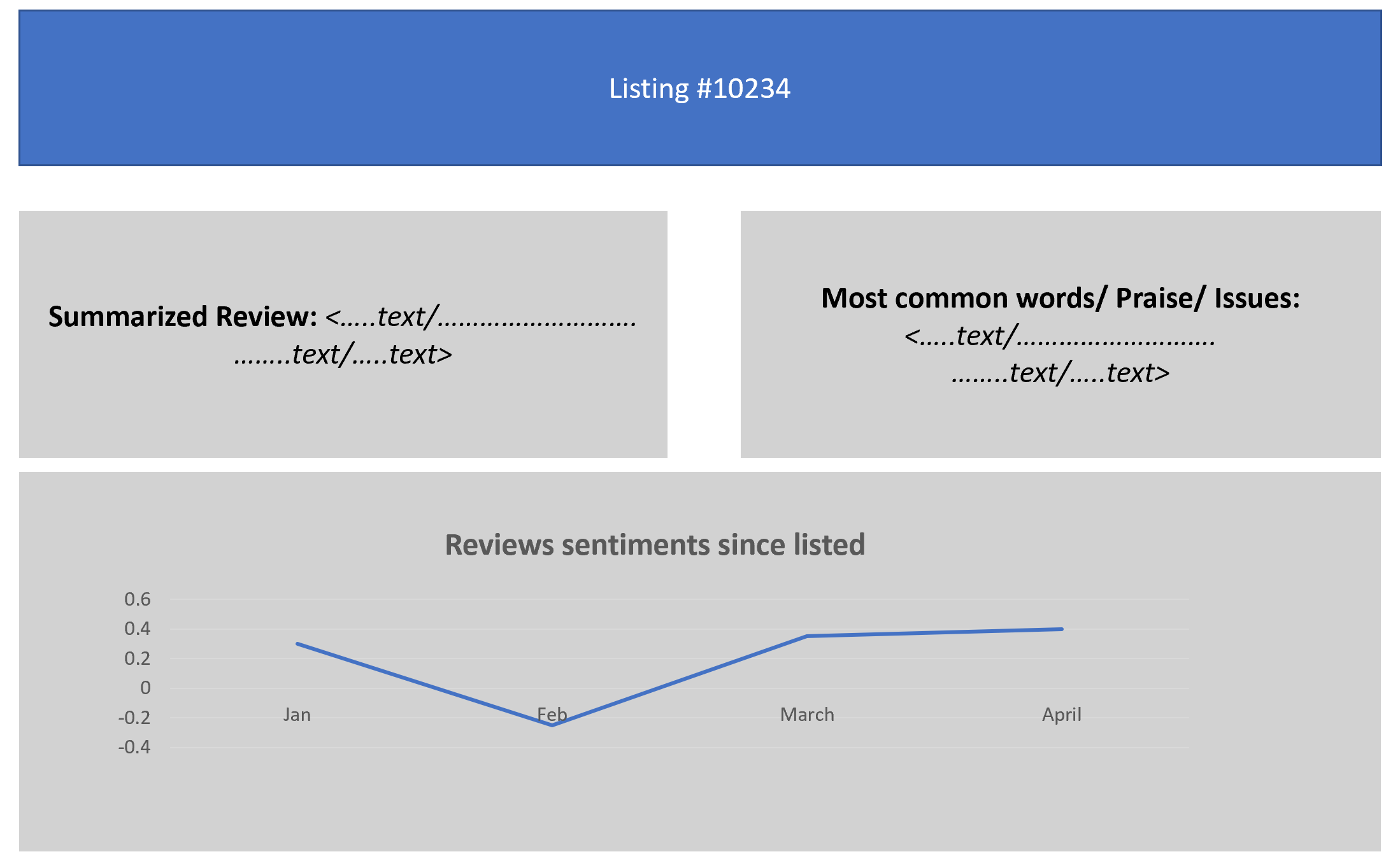
**Exhibit B:**

Type of host based on Instant Booking, Cancellation Policy, Room Type



**Exhibit C:**

Prototype of the visualisation for Feature Listing Summary



**Exhibit D:**

Sample Word Cloud



**Exhibit E:**

Sample Word Network

