Airbnb Data Analysis

Arjun Uday Prabhu
NYU Tandon School of Engineering
Computer Science
Brooklyn,New York
Email: aup211@nyu.edu

Garima Negi
NYU Tandon School of Engineering
Computer Science
Brooklyn,New York
Email: gn647@nyu.edu

Shivaraj S Nesaragi
NYU Tandon School of Engineering
Computer Science
Brooklyn,New York
Email: ssn314@nyu.edu

***Abstract*—Airbnb is an online community marketplace that connects people looking to rent their homes with people who are looking for accommodations. Airbnb users include hosts and travelers: hosts list and rent out their unused spaces, and travelers search for and book accommodations in 192 countries worldwide. Disruptive, brazen, and overall brilliant, the home-sharing empire has become the biggest lodging provider and earned many titles.**

This project examines Airbnb data from 2015 to 2017. Reviews are the richest source of data about how a trip went with cleanlinessŁ, check in, communication, host, etc. in focus. With increasing number of hosts listing their rental on Airbnb and more and more travellers preferring it over traditional accommodations it is important to understand which factors drive the reviews and ratings. In this project we have analyzed and inferred the top factors leading to negative reviews in US cities to help hosts improve in those aspects. We have also analyzed combined ratings for different factors to plot on map the best-rated listings in these cities. The technologies used for data analysis in this project are SparkML, Python, Pyspark, NLTK and Tableau, Plotly for its visualisation. This report elaborates on all concepts, tech stack, steps taken and provides visualisations for better understanding.

***Keywords*—Big Data,Machine Learning, AirBnb, Spark, Naive-Bayes,NLTK**

I. Introduction

Airbnb is an online marketplace and hospitality service, enabling people to lease or rent short-term lodging including vacation rentals, apartment rentals, hostel beds or hotel rooms. The company does not own any lodging but is merely a broker and receives percentage service fees from both guests and hosts in conjunction with every booking. It has over 3,000,000 lodging listings in 65,000 cities and 191 countries, and the cost of lodging is set by the host. Like all hospitality services, Airbnb is a form of collaborative consumption and sharing. The site's content had expanded from air beds and shared spaces to a variety of properties including entire homes and apartments, private rooms,castles, boats, manors, tree houses, tipis, igloos, private islands and other properties. With growing popularity and market presence, it has current and future impacts on the traditional accommodation sector. The data generated by Airbnb consisting of listings, ratings and reviews has increased and can be analyzed to gain more insights into customer reviews.

II. Motivation

One interesting feature of these online sharing economies is the review system.In case of Airbnb, hosts and guests may review their experience in 500 words or less at the end of each stay. The reviews contain strong recommendations for e.g. George was a great host, we stayed an extra day! In addition to creating a sense of accountability, reviews provide useful information to travelers about what to expect of their rental experience. Qualitative descriptions such as small but charming, noisy at night, free coffee are often far more insightful than numerical ratings. Our motivation behind this project was to help hosts get better understanding of what factors majorly affect ratings and reviews in their city and what characteristics are associated with the negative rental experiences. This would let the hosts improve in certain areas of their listing to improve their ratings and experiences for their guests. Secondarily, we wanted to aid customers by providing visual information of best listing based on consolidated ratings

III. Objectives

- To perform sentiment analysis on Airbnb review data.
- Classify reviews as positive and negative based on sentiment analysis.
- To determine what makes a rental experience good or bad.
- Analyze data to obtain keywords that frequently occur in negative reviews of listings in a US city.
- Plot top-rated listings in each category (apt, BnB, house, etc.) in US cities.

IV. Data

In this section, we discuss on the dataset that will be used in the project.The dataset was obtained from http://insideairbnb.com/get-the-data.html. The dataset consisted of 3 different CSV files namely Reviews dataset, Listings Dataset, Calendar dataset. The dataset and the attributes are as shown below

- **Reviews Data:** ID, Listing_ID, Date, Reviewer_ID, Reviewer_name, Comments
- **Listings Data:** ID, Host_name, city, state, Latitude, Longitude, Property_type, Review_score_consolidated
- **Calendar Data:** ID, Dates, Price

The dataset consits of various attributes including numerical,date and text attributes. The dataset is used accordingly for different analysis.

V. Technologies Used

For Data Preprocessing we made use of Python. To get properly labelled training dataset for our model, NLTK was used. We initially began using H2O.ai for our machine learning model. However, the model created was not efficient and gave incorrect results. We then created a model in SparkML.

After getting classified data as positive and negative reviews, for Data Analysis we used Scala in Apache Spark. The results received from the analysis were given a presentable form using Tableau. A web app was created to display results on map using Plotly.

Experiment Setup: The dataset was collected in the form of csv files for various cities (New York, San Francisco, Boston etc.). The csv files were reviews.csv, listings.csv and calendar.csv. Jupyter notebook environment was setup and all three CSV files containing 30 attributes in total were loaded. The environment setup for the Jupyter notebook was done during the Project proposal phase. The dataset was cleaned and then was run through various steps to obtain visualizations using Plotly and Tableau.

Issues with data and data cleaning: The dataset had numerous junk in it and some of the comments were not even in English language we had to preprocess the data so that it could be fed to Machine Learning model.The data had to be cleaned since it had NaN,None and infinite values in various numeric attributes and the cleaning was done using pandas package. Newline characters from review comments were removed which otherwise wouldn't parse as a csv file in Spark or H2O. Our model was built considering only review comments of English language. So, we filtered out comments of other languages using the langdetect python package.Empty prices in calender file was replaced by the mean price.

VI. Architecure Design:

The architecture design consists of two models out of which one is Machine Learning Model and the other is the Spark model for Big Data.

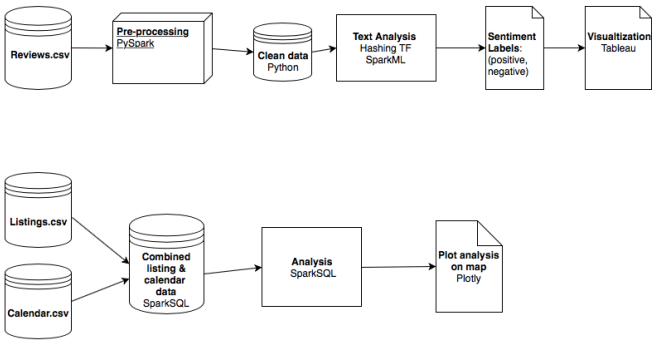


Fig. 1. Architecture Diagram

A. Machine Learning Model:

The machine learning model was built using Multinomial Naive Bayes model. The initial training data was categorized into positive and negative comments and made supervised data by using NLTK Vader. For this we made use of NYC data which was comparitively large compared to datasets of other cities. The machine learning model was then built in Spark Ml and was composed of couple of main elements

Hashing TF:

HashingTF is a Transformer which takes sets of terms and converts those sets into fixed-length feature vectors. In text processing, a set of terms might be a bag of words. HashingTF utilizes the hashing trick. A raw feature is mapped into an index (term) by applying a hash function. The hash function used here is MurmurHash 3. Then term frequencies are calculated based on the mapped indices. In our case we used comments which was text data and was converted to sparse matrix of numbers which could be fed to Naive Bayes Model.

Naive Bayes:

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes theorem with the naive assumption of independence between every pair of features. The Naive Bayes model we used was Multinomial Naive Bayes Model and tuning parameters were set to particular set of values to get better accuracy.

The comments were sentimentally analyzed to categorize as 'negative' and 'positive' comments.

B. Spark Analysis Model:

Zeppelin Notebook was started from docker. listings.csv and calendar.csv for one US city was loaded into spark. After loading them, using sparkSQL, those files were loaded as tables. Both tables had listing_ID in common. So a natural join was performed on both of them. The filtered tuples had many rating columns. A consolidated rating column was formed by taking an average of all the rating columns. After this step, the rows were ordered in the descending order of ratings, per Property type. A dataframe of results obtained, was saved into a file.

A web application was developed which took input from the user regarding selection of a US city and Property type from the dropdown menu. With this combination, a map showing the top ten listings for each property type was depicted using plotly.

VII. Code

The code for the final report has been attached along with the report and uploaded in the NYU Classes

VIII. Visualizations

The visualizations were done using Tableau and are shown below. The web app was built using HTML,CSS and Javascript and the related images of web application are as shown below

IX. Result and Evaluation

We analyzed the combination of listings data as well as calendar data to group rows based on Property type and decreasing order of ratings. The accuracy of classification of Naive Bayes model was calculated for initial training data by test train split and it gave a better accuracy of around 98%.

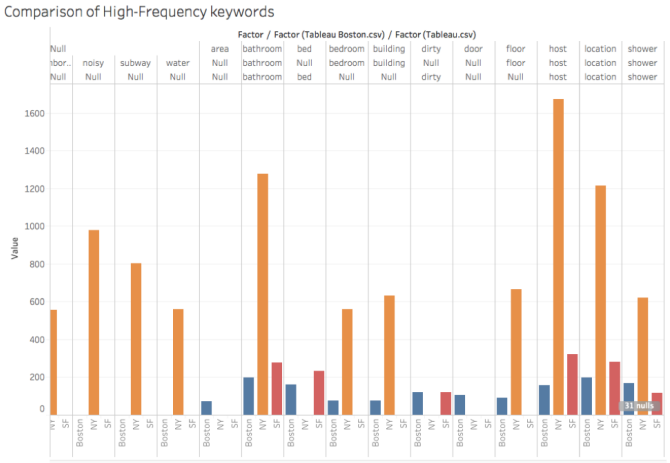


Fig. 2. Clustering of common characteristics in different cities

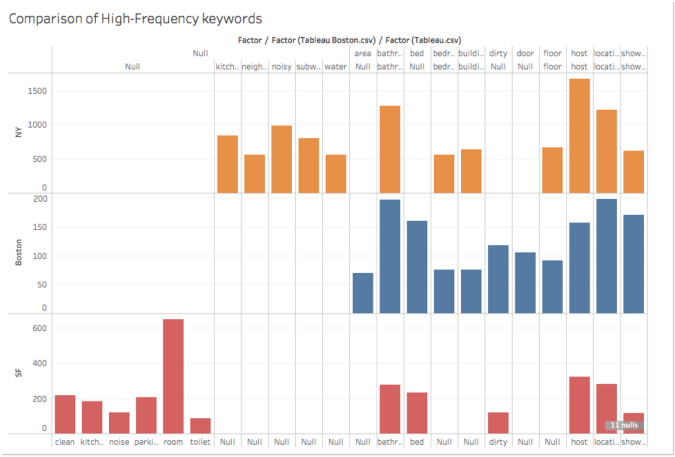


Fig. 3. Depiction of factors leading to negative reviews specific to city

The Spark Model was able to categorize top 10 listings for each cities accurately using the consolidated ratings.

The things that didn't go as expected was we were unable to use H20 Machine Learning model because the model built was split into 4 parts and we were unable to proceed with it as H20 had limited documentation and we were unsure of how to use it.

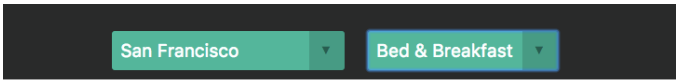
X. Conclusion And Future Work

Labelling of positive and negative reviews was carried out by the Machine Learning model successfully.We conclude from our analysis that negative reviews in different cities focus on different keywords.Also, we successfully suggested top 10 listings to the users based on consolidated ratings.

For Future work we will be incorporating Spark Streaming to stream static files.Establish a relation between listing availability and rating and analysis of trends in price change by hosts during holiday season, festivals, etc.

References

[1] http://insideairbnb.com/get-the-data.html.



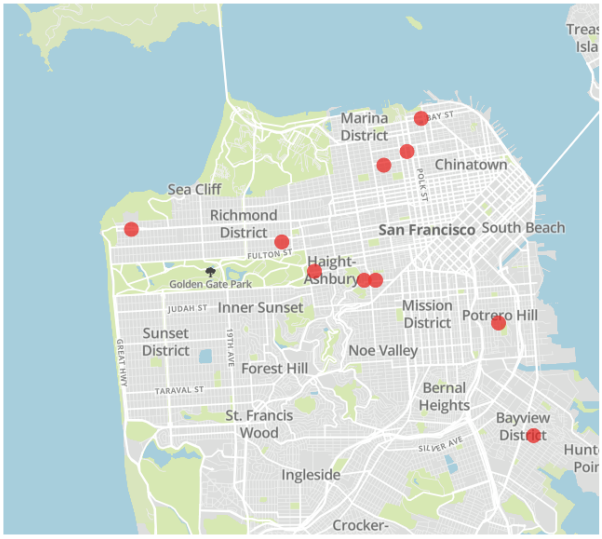


Fig. 4. Top 10 listings of SanFransico

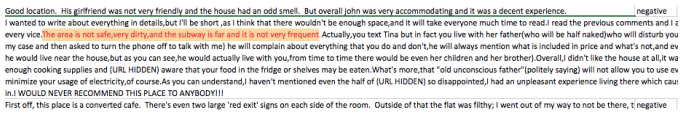

Good location. His girlfriend was not very friendly and the house had an odd smell. But overall john was very accommodating and it was a decent experience. negative
I wanted to write about everything in details,but I'll be short ,as I think that there wouldn't be enough space,and it will take everyone much time to read.I read the previous comments and I a
every vice.The area is not safe,very dirty,and the subway is far and it is not very frequent. Actually,you text Tina but in fact you live with her father(who will be half naked)who will disturb you
my case and then asked to turn the phone off to talk with me) he will complain about everything that you do and don't,he will always mention what is included in price and what's not,and ev
he would live near the house,but as you can see,he would actually live with you,from time to time there would be even her children and her brother).Overall,I didn't like the house at all,it wa
enough cooking supplies and {URL HIDDEN} aware that your food in the fridge or shelves may be eaten.What's more,that "old unconscious father"(politely saying) will not allow you to use ev
minimize your usage of electricity,of course.As you can understand,I haven't mentioned even the half of {URL HIDDEN} so disappointed,I had an unpleasant experience living there which cau
in.I WOULD NEVER RECOMMEND THIS PLACE TO ANYBODY!!!
First off, this place is a converted cafe. There's even two large 'red exit' signs on each side of the room. Outside of that the flat was filthy; I went out of my way to not be there, t negative

Fig. 5. Comment showing negative review in NYC

[2] https://venukanaparthy.wordpress.com/2015/07/04/spam-classification-with-naive-bayes-using-spark-mllib/
[3] http://spark.apache.org/docs/latest/programming-guide.html/resilient-distributed-datasets-rdds
[4] https://github.com/h2oai/sparkling-water/blob/rel-2.0/py/README.rst
[5] https://www.kaggle.com/residentmario/sentiment-analysis-and-collocation-of-reviews
[6] https://pypi.python.org/pypi/langdetect?