*Your proposal should describe at a high level what you’re seeking to accomplish, and your motivation for performing this analysis.  A guiding question or hypothesis to test is one good way to start.  If you are going to work in a small group (encouraged!), you should also list your partners’ names.*

*You should briefly describe your data sources, plan for doing the work, and up-front concerns.  If you are working in a group, please describe the roles and responsibilities of each group member.*

*We’ll treat this proposal as a planning document, not a blueprint containing “firm, fixed requirements.”*

**Team Members:**

Stefano Biguzzi, Ian Costello, Dennis Pong

**Goal and Motivation**

We wanted to show that with the help of network analysis and text mining we can identify a group of hosts that comprise of fake a certain % of fake positive reviews, and they are a group of hosts that are likely engaging in activities such as renting out their units for over 90 days, in clear violation of recently introduced short-term rental laws effective since Feb 1st, 2015.

The criteria to be eligible to rent out a residential property for less than 30 days on platforms like Airbnb are:

* Permanent Residents (owners and tenants) must place their residential unit on the Planning Department’s Short-Term Residential Rental Registry
* If the resident is present, there are no limits to the number of nights per year a unit can be rented
* If the resident is not present, the unit may not be rented more than 90 nights per year

**Hypothesis**

Our hypothesis here is plain and simple. We can see there are more fake reviewers on Airbnb listings rented out with an average over 90 days a year than listings rented out less than 90 days a year.

**Data Sources**

|  |  |  |  |
| --- | --- | --- | --- |
| Date Compiled | Country/City | File Name | Description |
| 07 April, 2021 | San Francisco | [listings.csv.gz](http://data.insideairbnb.com/united-states/ca/san-francisco/2021-04-07/data/listings.csv.gz) | Detailed Listings data for San Francisco |
| 07 April, 2021 | San Francisco | [reviews.csv.gz](http://data.insideairbnb.com/united-states/ca/san-francisco/2021-04-07/data/reviews.csv.gz) | Detailed Review Data for listings in San Francisco |
| 07 April, 2021 | San Francisco | [listings.csv](http://data.insideairbnb.com/united-states/ca/san-francisco/2021-04-07/visualisations/listings.csv) | Summary information and metrics for listings in San Francisco (good for visualisations). |
| 07 April, 2021 | San Francisco | [reviews.csv](http://data.insideairbnb.com/united-states/ca/san-francisco/2021-04-07/visualisations/reviews.csv) | Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing). |

Metrics that we’re laser-focused on:

1. High availability (more than 90 days / year)
2. % of hosts w/ multi-listings

**Work Plan / Roles and Responsibilities for Group Members (weights)**

* **Network Analysis (Dennis Pong) (60%)**
* **Text mining and sentiment analysis (Stefano Biguzzi) (35%)**
* **Manually building labels for a set of 100 randomly selected reviews (Ian Costello and Dennis Pong - 50/50) (5%)**

**Concerns / Caveats**

* There are reviews that are going to be non-English. We need to figure out a way to effectively sieve out reviews in English whether to use a python package or a segmentation approach enabled by the data sets.
* There can be fraudsters that are in the making so the groupings of identified hosts with fraudulent positive reviews can very well be living the <90 days / year category.
* Besides positive and negative labels for reviews, we can always have cases that are neutral. So in those cases we can’t use them as denominators in calculating the TPR and FPR in the calculation of the precision.
* While we have in mind that we wanted to figure out which hosts are in clear violation of the laws by listing
  + The entire home, and,
  + Renting out for more than 90 days / year

, the real result that can be brought to light in the lens of network analysis is really to expose hosts with multiple listings on the same platform, namely Airbnb. But there is a limitation of not being able to account for hosts listing across multiple platforms in addition to Airbnb.

Meeting Notes 06.27.2021:

* Subset data split by year; 2018 (50%) – 2019 (50%) …sampling data and constructing training set.
  + Show multi-year tracking based on address to get picture
* SF: Resident is not present, the unit may not be rented more than 90 nights per year
  + Hypothesis: Sentiment analysis after figuring out or flagging which are possibly fraud reviews…compare side by side fraud reviews more likely to review listing that
  + Two viz 1) visualization as-is…color for units more than 90 days vs less than 90 days; 2) with labeling for the potential fraudulent units … of those how many are 90 days and over?
  + More units that are rented 90 days and over are flagged by fraudulent; motive of landlords to post fraudulent reviews… because they rent more than 90 days; sentiment analysis of reviews
  + Minimum and maximum nights within a location
* Language detects … for the import
* Can take a sample of 100 reviews… manually review positive or negative as a training set…this is what the true and false positive rates … we believe the error rate is x%.
* Landlord abuse stems from understanding whether the “entire home” is available or just a part
* In London, maximum rentals are ~90 days in a year – limiting an owner’s ability to rent space more
  + Landlords change and vary postings to get around law
* Entire home listed and duration of rental by unit during the year (how many days in the year was unit rented out); listing more than one location
* Multiple Dwelling Law
  + Can we connect names of multiple hosts of multiple units
  + Connect hosts perhaps through deep learning (e.g., pictures)
  + Can we construct some sort of probability to guess that someone with different spellings of names are in fact the same person to illegally rent out more spots
* Does the final project need both aspects of the class (Network and Text analysis)
  + Suggestion if both, connect reviews of same person with different users… predicting which reviews may be real or fake, positive, or negative reviews
  + Reviews that have single users review multiple units >5, of everyone else 75% are negative reviews… with the analysis of the network of that host – this host should be flagged based on listing network and of the reviews.
* Find a corpus of words with negative or positive sentiments; what is the method to determine sentiment? NLP sentence tokenization
* Once we identify a potential fraud, can we further support our conclusion by looking at the type of reviews… excluding fake reviews via analysis and review sentiment
* Policy environment of the location
* Longevity of hosts and prolific posting number of posts per year?

[Large-Scale Sentiment Analysis on Airbnb Reviews from 15 Cities](https://www.researchgate.net/profile/Abdulkareem_Alsudais/publication/333114485_Large-S%5b…%5dle-Sentiment-Analysis-on-Airbnb-Reviews-from-15-Cities.pdf)

<http://insideairbnb.com/>

[**Uncovering Hidden Trends in AirBnB Reviews using Data Science**](https://towardsdatascience.com/uncovering-hidden-trends-in-airbnb-reviews-11eb924f2fec)