Solo Travel Research

Data, Methods and Initial Results

The objective of this paper is to better understand the motivations of solo travelers using computational methods. Despite being a growing trend in a multibillion-dollar industry, no large-scale study has been done on solo traveler motivations. A solo travel focused work in the field is the qualitative research conducted by Constanza Bianchi, "Solo Holiday Travellers: Motivators and Drivers of Satisfaction and Dissatisfaction". In his research, he was able to identify that the top motivators include, personal factors such as independence or bravery, destination factors such as sites or landscapes and human interaction with locals and service providers. The major limitation of his research was the small sample size of 24 people from Queensland, Australia. To expand his method to other places and larger sample size would be extremely expensive. A proposed methodology for this paper is to extract the motivations for solo travel by conducting text analysis on traveler reviews.

Data

To start off this research, the location of travel is narrowed down. Considering that the top solo travel destination is the United States and San Francisco is a highly regarded tourist spot, hotels in San Francisco are considered. Hotel and hostel reviews are chosen as the focus as all travelers will need a place to stay and many times these reviews are telling of their top priorities and non-negotiable needs. Hostels are usually budget friendly and known to be great place to meet people which serves a good place to meet other travelers while hotels provide a more standard experience and better amenities. To study any difference between solo travelers staying in budget hotels and hostels, both are considered.

All data was gathered from Trip Advisor. Trip Advisor is one of the largest travel review sites with reviews from people around the world sharing their opinions and rating on hotels, attractions, restaurants, etc. The first dataset is a hostel in San Francisco, "The Green Tortoise Hostel" and other one is a hotel called "The Metro Hotel". For both datasets, all reviews are scraped, and a subset of solo travel specific and all but solo travel dataset is created. Therefore, in total there are six datasets prepared.

Collection Procedure

Trip Advisor does not grant API access for academic research so the content was gained by scraping pertinent information using a Python program. The python program utilized packages such as Selenium (to connect with the website) and Beautiful Soup (to parse the data). The results were stored in Pandas data frame and output to csv file. The data collection was quite challenging with the dynamic webpage. The two major issues encountered were timing out before creating the soup object and not being able to successfully click on checkboxes for different "Travel Type" filter. For the first issue, if the data from a given page (5 reviews) was missing, those reviews are tentatively not included. For the second issue, as a last resort, the solo traveler reviews were manually categorized. Both these issues will be readdressed in a second pass at data collection if time permits, but this gave good set of tentative data to work with. The final datasets are as follows:

Table 1 Summary of Data Collected

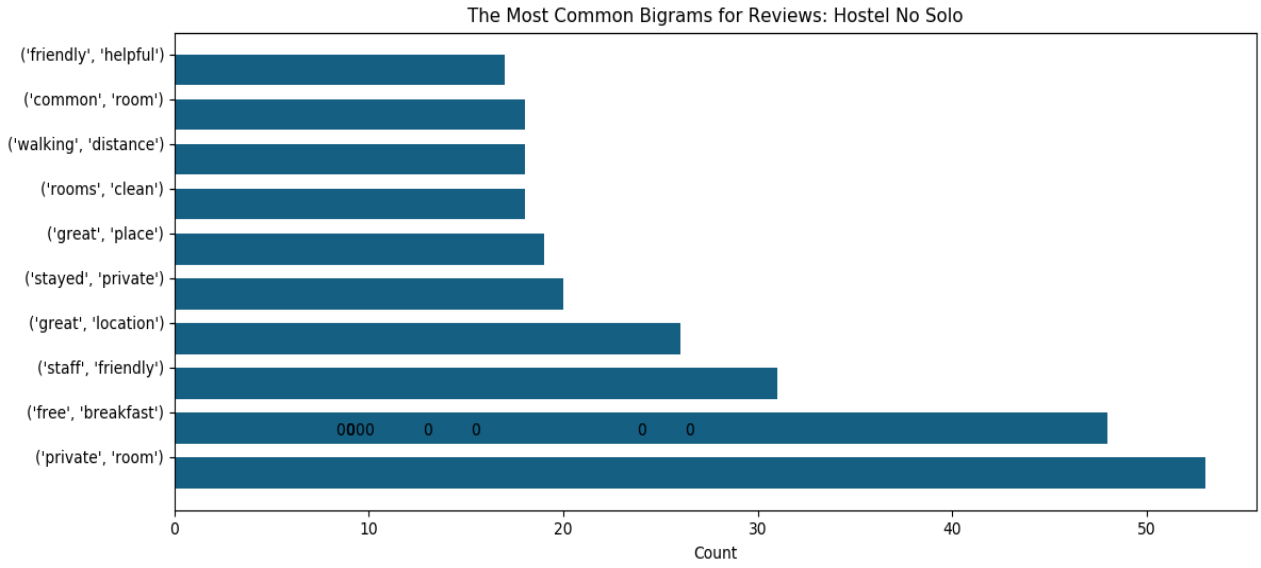
Hotel/Hostel Name	Dataset Name (.csv files)	Dataset Description	Total Number of Entries
Green Tortoise	hostel_all_reviews	All	663
Green Tortoise	hostel_solo_reviews	Filtered to only Solo	236
Green Tortoise	hostel_all_but_solo_reviews	Filtered out Solo	427
Metro	hotel_all_reviews	All	518
Metro	hotel_solo_reviews	Filtered to only Solo	73
Metro	hotel_all_but_solo_reviews	Filtered out Solo	445

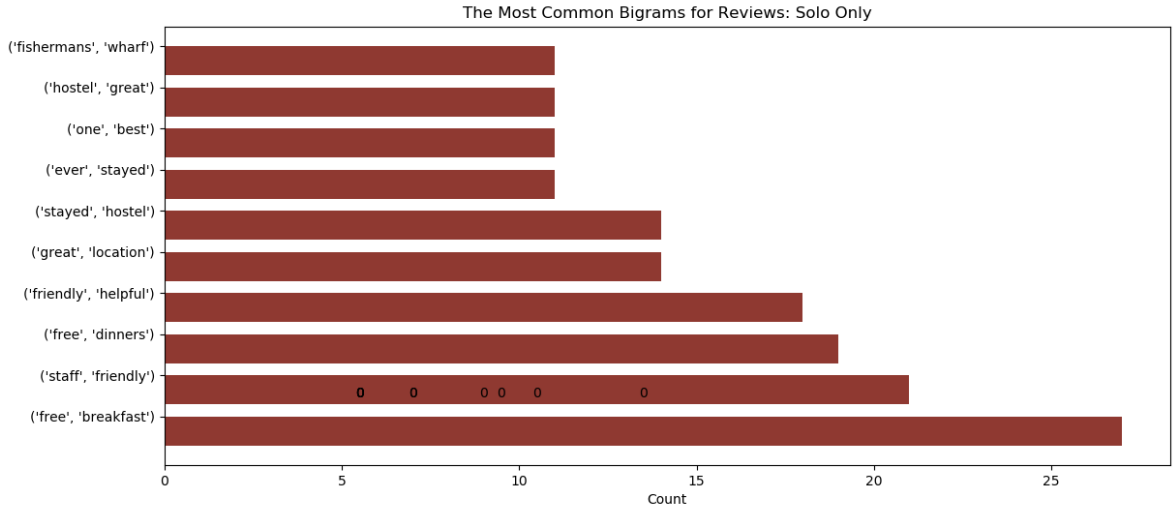
Methods

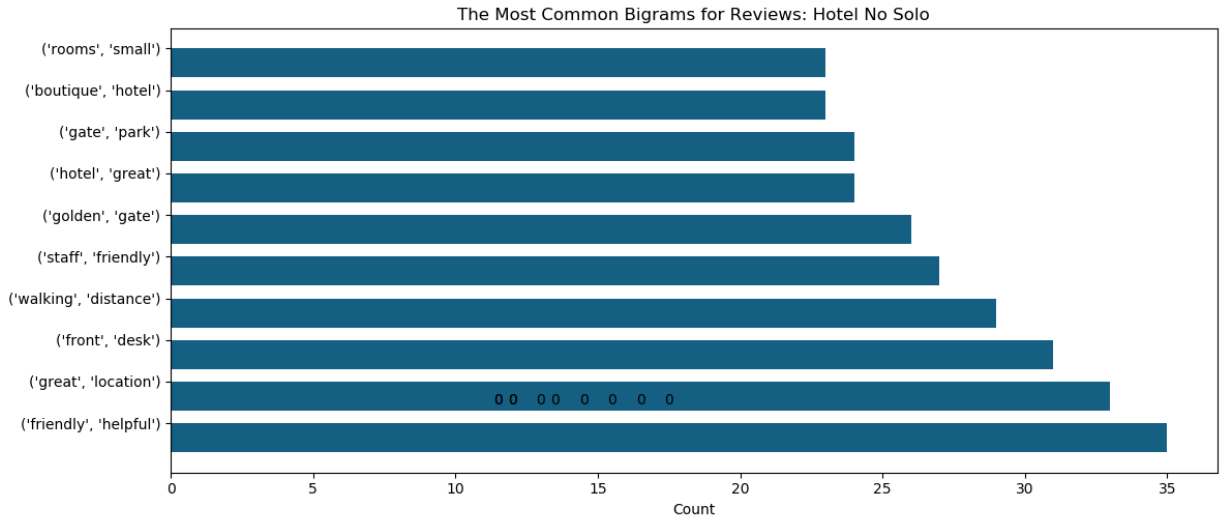
Having prepared a corpus of digital data consisting of reviews, a simple text analysis method is first applied. For each of the datasets, the most frequent bigrams are evaluated. The text data is cleaned by removing punctuations, turning it all to lower case, removing commonly occurring stop words from nltk package and removing domain specific common words such as San Francisco or the hotel names. The text is tokenized, and count of each bigram is noted. Having this information for all the data, the most common values are found by applying most_common from collections package. The results are then plotted on a bar plot.

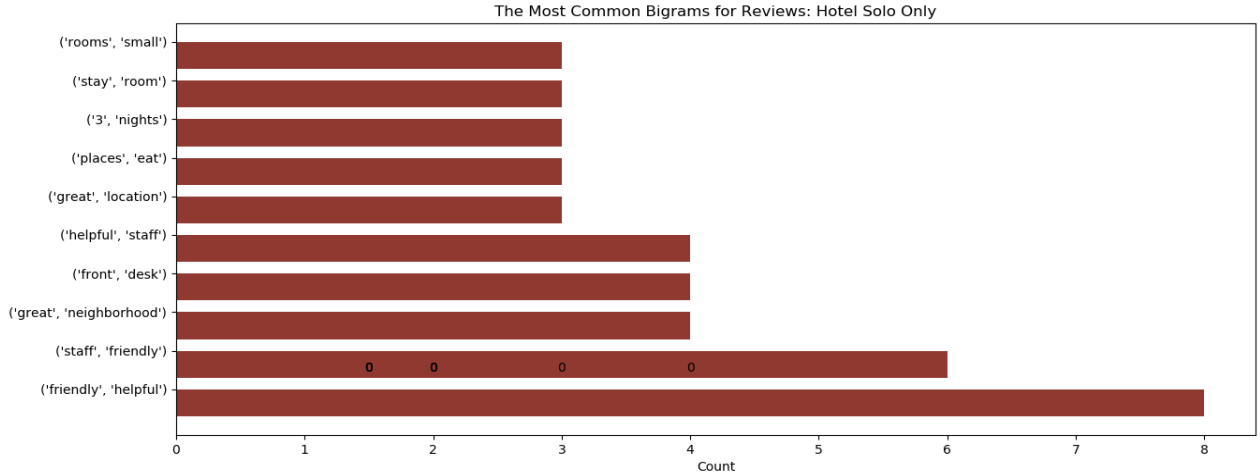
Preliminary Results

The graphs below show the result for most commonly occurring bigrams in each on the dataset. The first two plots are for the hostel data and followed by hotel data.




The Most Common Bigrams for Reviews: Solo Only
('fishermans', 'wharf')
('hostel', 'great')
('one', 'best')
('ever', 'stayed')
('stayed', 'hostel')
('great', 'location')
('friendly', 'helpful')
('free', 'dinners')
('staff', 'friendly')
('free', 'breakfast')
0
5
10
15
20
25
Count


The Most Common Bigrams for Reviews: Hotel No Solo
('rooms', 'small')
('boutique', 'hotel')
('gate', 'park')
('hotel', 'great')
('golden', 'gate')
('staff', 'friendly')
('walking', 'distance')
('front', 'desk')
('great', 'location')
('friendly', 'helpful')
0
5
10
15
20
25
30
35
Count


The Most Common Bigrams for Reviews: Hotel Solo Only
('rooms', 'small')
('stay', 'room')
('3', 'nights')
('places', 'eat')
('great', 'location')
('helpful', 'staff')
('front', 'desk')
('great', 'neighborhood')
('staff', 'friendly')
('friendly', 'helpful')
0
1
2
3
4
5
6
7
8
Count

From these initial results, it is evident that the reviews are indeed able to show what is important to travelers staying there. However, the method certainly needs to be refined to further remove commonly occurring word pairs and pairs that essentially convey the same meaning.

Additional Methods

In addition to refining the bigram analyses, it would be valuable to review the sentiment of reviews and find a correlation to what leads to a more positive experience. This too can be visualized in a plot and compared with the other datasets to see if solo travelers are seeking something different.