Uncovering Motivators for Solo Travel

Author: Dhruval Bhatt

Master's in Computational Social Science: Perspectives in Research

Contents

Abstract

Solo travel is a highly popular trend in travel, a multibillion-dollar industry, that is continuing to grow. Yet not enough research has been conducted to understand why people travel alone for leisure. This paper uses computational methods Trip Advisor reviews for three San Francisco Area hostels to explore the question: what are the key motivators for solo travel? An exploratory analysis using top ten most frequently mentioned words for each traveler type was conducted to determine if hostel reviews reveal motivations and if there is any indication of solo travelers' quest for solitude or company. Additionally, a decision tree and random forest classification model is made using the text content of the reviews to predict the traveler type. Based on the analysis, it is evident that appropriate analysis on accommodation reviews can uncover travel motivation, solo travelers do have distinct motivations than other traveler types and solo travelers have a higher emphasis on human interaction than other types.

Introduction

In US alone, leisure travel cost $761.7 billion dollars in 2018 ("U.S. Travel Answer Sheet." 2019). Of which, there is an increasing share of solo travelers for both short term and long-term travel. According to a survey quoted in a Telegraph article, *Why are so many of us now choosing to travel alone?*, "one in nine holidaymakers reported that they took a holiday on their own in the previous 12 months - double the number compared to six years previous" (Elliot). What attributes to the growing interest in solo travel? According to Solo Traveler World statistics, "59% said they want to see more of the world and they're not willing to wait for others yet 54% of respondents like group travel at least some of the time" ("Solo Travel Statistics", 2019). Statistics like this are made available by travel marketing sites but little work has been done in academia to dig deeper into why someone travels alone for leisure and what is propelling the growth in this trend. In this paper, I will study the question: what motivates solo travelers to travel alone for leisure?

Based on small scale, qualitative research by Constanza Bianchi, top motivators include, personal factors, destination factor and human interaction with locals and service providers (Bianci 202). For any type of traveler, destination factor would hold importance and personal factors may vary significantly from person to person. Therefore, this study focuses on any difference in the need for human interaction for solo travelers. To answer this question, computational methods, such as top word count and classification models, will be applied on San Francisco area hostel reviews curated from Trip Advisor. The three main objectives of this study are:

- Can online hostel reviews reveal patterns in motivation?
- Is there a distinction in travel expectation between solo travelers and other travelers?
- Do solo travelers' value human interaction more than, less than or same as other traveler types?

Based on the analysis conducted, it can be established that traveler reviews on accommodation can be used to extract motivation insights, solo travelers do have distinct expectations and have higher value on human interaction than business or couple travelers. This paper provides an interesting use of computational methods on online reviews to understand a growing trend in a field that is dominated by qualitative research. It contributes to the still maturing, niche intersection of sociology, travel research and computational methods. The results of this analysis can add value in corroborating the findings of past work and build a basis for future use of computational methods in travel research. In this paper, I start off by expanding on the theoretical basis for this study, followed by a detailed description of data, an overview of analysis methods used, and results obtained and ends with some concluding remarks.

Theory

As solo travel is becoming one of the most popular travel trends, it is important to clearly understand why people travel alone for leisure. Currently, the most common reason for such research is to give insight to the tourism management and service personnel to cater to the needs of the changing demographic of travelers. However, we can go beyond marketing research benefits and use the understanding of this major trend to glean sociological implications. As Eric and Scott Cohen's paper, "New Directions in the Sociology of Tourism" indicates, novel theories from other fields are being incorporated in tourism research which creates an imperative to expand solo travel research from "solely in terms of economic activity" to a framework that includes emotions of belonging and longing to add a new dynamic to how this topic has been typically addressed (Cohen et al, 6). Knowing why people solo travel can be utilized in sociological and anthropological studies to answer questions such as: do more people choose to live in single person households and why? Are people getting lonelier and seeking a sense of community or are people overwhelmed in a highly connected society and seeking means to escape in solitude? Answer to why people solo travel could give some initial insights to such questions.

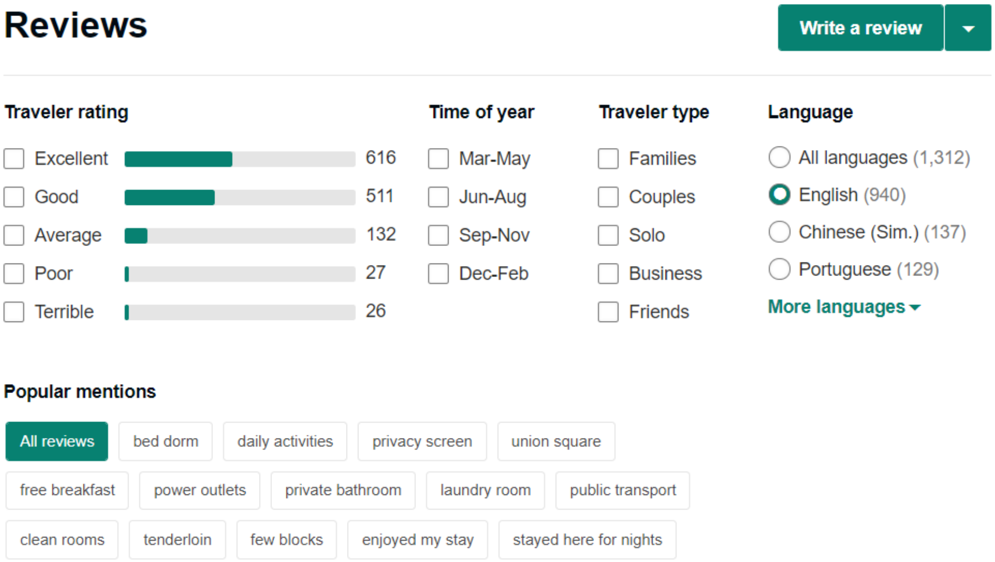
Typically, travel research utilizes qualitative methodology such as surveys and interviews to answer questions of motivations and expectation. While qualitative methods are successful in getting subtle insights from a small group of people, it is difficult to scale the study for a larger, diverse group of people. Additionally, while answering direct questions survey responders may be biased to answer based a commonly accepted response versus the truth. Computational methods such as text analysis on social media, blog posts and review content could be one way to address this issue. However, such online content is not generated to answer specific research query, so it is important to explore the data available and gauge if it reveals trends valuable for the question being asked. While travel research utilizes computational methods such as sentiment analysis on data available on twitter, travel review sites and blogs, a survey of travel research indicates that no prior work utilizes travel reviews for understanding traveler motivation. Therefore, it is imperative to conduct an exploratory analysis to confirm that the choice of trip advisor data is appropriate. To use in analysis, it is important to answer if the traveler provided accommodation reviews, given to assess whether they were satisfied or not with the place, is useful in revealing motivation? The initial hypothesis would be yes as framework of expectation is built upon what one wants out of the experience. For instance, if one wants to meet other people, then they will regard socials as a positive experience that is worth mentioning.

Furthermore, to build an initial hypothesis and theoretical framework of why people solo travel, the small size qualitative analysis conducted by Constanza Bianchi is used. In this study, Bianchi uses qualitative research methodology, Critical Incident Technique (CIT) on information gathered from face to face conversation with 24 participants from Queensland, Australia to learn about the experiences of solo travelers. The study found that top three main drivers were "personal factors (37% of incidents), followed by destination factors (35% of incidents) and human interaction factors (28% of incidents)" (Bianchi 202). Where personal factors include feeling of bravery, sense of achievement etc., destination factor includes historical sites, landscapes, restaurants, etc. and human interaction includes both service providers and local people or travelers. The authors acknowledge that a major limitation in this study is the small number of participants from one region of the world. This limitation can be overcome by using computational methodology and using data from a wide population sample. Using these findings, we hypothesize that most solo travelers are not seeking solitude and lone relaxation time but are predominantly seeking to meet new people and value interactive service personnel. While this may not be the case for all traveler types.

Data

Data Overview

The data for this study is collected from the leading travel review website, Trip Advisor. The website offers reviews for restaurants, hotels, tour packages and tourist attractions. For this research, I focused on the accommodation (hotels/hostels) reviews as all travelers will need a place to stay and many times these reviews are telling of their top priorities and non-negotiable needs. For all hotel listings information such as overall rating, average rating for different categories (cleanliness, service, value, location), property amenities, room features, location details, and guest reviews is available. Within reviews, one can filter by traveler rating, time of year, traveler type, select language and words that were popular mentions. The screen shot below shows an example of the available filters on reviews and typical format of a traveler review.



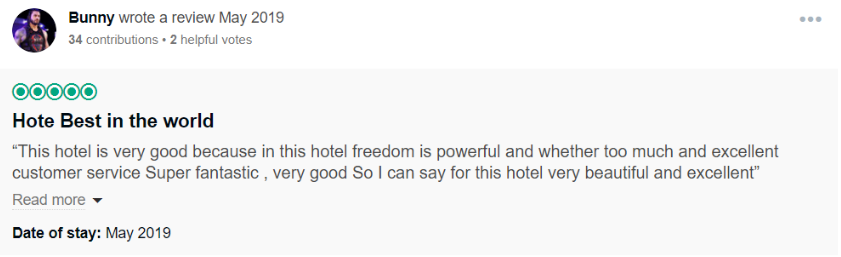
Reviews

Traveler rating		Time of year	Traveler type	Language
□ Excellent	616	□ Mar-May	□ Families	○ All languages (1,312)
□ Good	511	□ Jun-Aug	□ Couples	● English (940)
□ Average	132	□ Sep-Nov	□ Solo	○ Chinese (Sim.) (137)
□ Poor	27	□ Dec-Feb	□ Business	○ Portuguese (129)
□ Terrible	26		□ Friends	More languages

Popular mentions

All reviews · bed dorm · daily activities · privacy screen · union square · free breakfast · power outlets · private bathroom · laundry room · public transport · clean rooms · tenderloin · few blocks · enjoyed my stay · stayed here for nights

Figure 1 Filters available on Reviews



Bunny wrote a review May 2019

34 contributions • **2** helpful votes

●●●●●

Hote Best in the world

"This hotel is very good because in this hotel freedom is powerful and whether too much and excellent customer service Super fantastic , very good So I can say for this hotel very beautiful and excellent"

Read more

Date of stay: May 2019

Figure 2 Review Example

Data Collection Method

Though Trip Advisor is a rich source of data, it is not easy to get data from the site. The website does not allow API access for research and academic use, so the best option was to create a web scraping script in python to get the required data. The python program utilized packages such as Selenium (to connect with the website) and Beautiful Soup (to parse the data). The results are stored in Pandas data frame and output to csv file. The data collection was quite challenging with the dynamic webpage. The major hurdle encountered was not being able to successfully click on checkboxes for different "Travel Type" filters and "Read More" to get the expanded, complete reviews. In order to create real browser like behavior, multiple methods from selenium package are used. These include: find the element by path name or by input name and then clicking on it is tried; timeout is added; action chains with expected conditions is used; and finally, though the initial attempt was made on Chrome browser, an attempt with PhantomJS is also made. Despite making the efforts to rectify the inability of scraper to click on the checkbox, the scraper did not work as expected. In the interest of time, much of the needed for this analysis data is prepared manually.

Other Academic Work

Most other academic work found on this data source also use web scraping but since they do not impose traveler type constraint, the data collected or even data collection methodology is not easy to use. Also, web pages change frequently so it is not possibly to leverage any reproducible code available. Studies that utilized Trip Advisor data include, “Reliability of Reviews on the Internet: The Case of TripAdvisor” by Chau and Banerjee, “Sentiment Analysis in TripAdvisor” by Ana Valdivia, M. Victoria Luzón, and Francisco Herrera and “Applicability of Demographic Recommender System to Tourist Attractions: A Case Study on TripAdvisor” by Yuanyuan Wang, Stephen Chi-Fai Chan and Grace Ngai.

Data Description

Due to the challenge of data collection, two sets of data are collected and the description of both is included here. The information on where to access data collection script and all dataset curated is included in the appendix.

Preliminary Dataset

First a preliminary dataset is collected using the web scraping script on one hostel and one hotel in San Francisco area. In this data, the review text is not complete, and it is not filtered by all traveler types. Each of the entry in this data is manually sorted into solo or non-solo. The following table provides an overview of the data collected.

Table 1 Summary of Preliminary 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

Final Dataset

After conducting an exploratory analysis on the preliminary data, it is determined that further granularity on traveler types and extended text for reviews would be helpful in making meaningful inferences. Additionally, as seen in the table above, budget hotels have significantly fewer solo travelers compared to non-solo traveler type so to get a more balanced data, only hostel is data is ultimately considered.

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, only hostels in San Francisco are considered. Over 500 Reviews from the year 2013 to 2019, from three different hostels are curated: San Francisco City Center Hostel, Orange Village Hostel and Green Tortoise Hostel - San Francisco. For each review the following information is collected: Review Title, Review Text, Date, Overall Rating and Traveler Type. If any review is missing the traveler type specification, that data is not included. Of the information collected, the review text and the traveler type are the main items used in the analysis, but additional information is collected to future possible use. Additionally, the data has 5 types of travelers – Business, Couple, Friends, Solo and Family but all analysis is conducted on just the first four. The figure below shows the distribution of traveler types for each of the year.

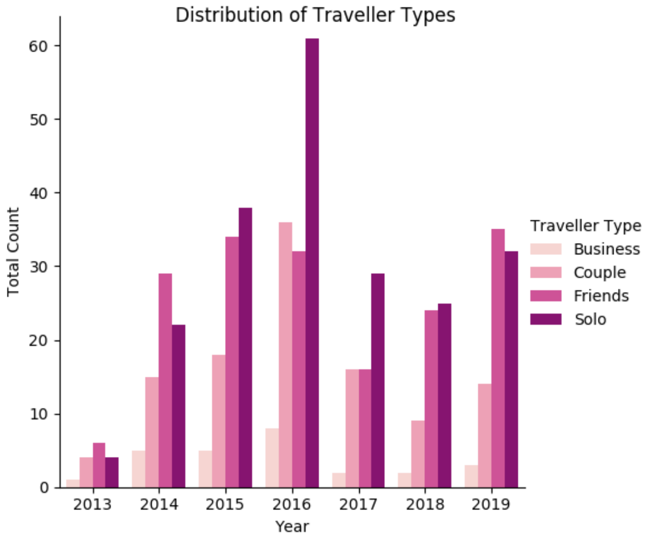


Figure 3 Distribution of Traveler Types in Final Data

Exploratory Analysis

Method

As discussed in the theory section, there are many instances where travel accommodation data has been analyzed for sentiment, but researchers do not often use those reviews to infer motivation for travel. In this study, an initial exploration of data is critical to gauge what the data reveals and decide if the content is suitable to answer the question of solo travel motivation. The first step in the research is to explore the hostel reviews collected for each of traveler type and see if there are any trends in what is being said and if there is any distinction between traveler types. To find out the top words used by each of the traveler types, we count the most frequently used words in all reviews of a given traveler type and plot it in a bar chart.

To accomplish this task, python script and packages such as NLTK, Collections and Pandas are utilized. The first step that is applicable to both exploratory analysis and predictive model – the data is processed to be standardized. For every review in the dataset, all the text is converted to lower case, punctuation is stripped, and numbers are removed. Furthermore, the sentence string is split into a list of words. From these words two types of stop words are removed – commonly occurring words such as “a, the, is, …, etc.” using NLTK stop words as well as domain specific stop words such as “hostels, name of the hostels, San Francisco”. This processed data is stored in a new column in pandas data frame for each of the review.

Next a function to create n-grams is applied to the word collections to create pairs of words using the Collections package. This function is written such that varying lengths of n-grams can be created. This allowed me to explore the optimal number of words that should be paired. Using just one word was not useful as it is difficult to qualify anything meaningful about the word. For instance, knowing “room” is a top word is not as useful as knowing “clean room” is a frequently used word pair. Still, there are many common word pairs that are likely to repeated in hotel reviews, so the top ten word-pairs is still not very revealing. However, using trigrams, differing priorities among different types of traveler starts to become evident. Any higher number is not considered as it would start to become more specific with very few counts of each. The results from trigrams is included in the results and one word and two words pairs is shown in the appendix for reference.

Additionally, Lancester stemming is applied to the words to avoid over counting of similar word pairs. Newer top words emerge through this technique and the over results are further refined. The un-stemmed words are still helpful in deciphering some of the less obvious stems.

Results

The resulting plots from exploratory analysis is shown below. Interesting and distinct trends for each traveler type is noticeable. For instance, couples are staying in and reviewing private rooms and appreciate the fact that people “didn’t bother us”. On the other business travelers use terms to indicate comfortable bed, quick answer and walkability to landmarks. For friends, the proximity to tourist attractions is a highly rated one but there is also considerable mention for breakfast and front desk staff. Finally, for solo travelers, friendly staff is a dominant one and other terms that indicate types of room. All reviews mention location specific issue - lots of homeless people! Even this simple analysis is powerful in showing some meaningful distinctions in what different travelers may be seeking.

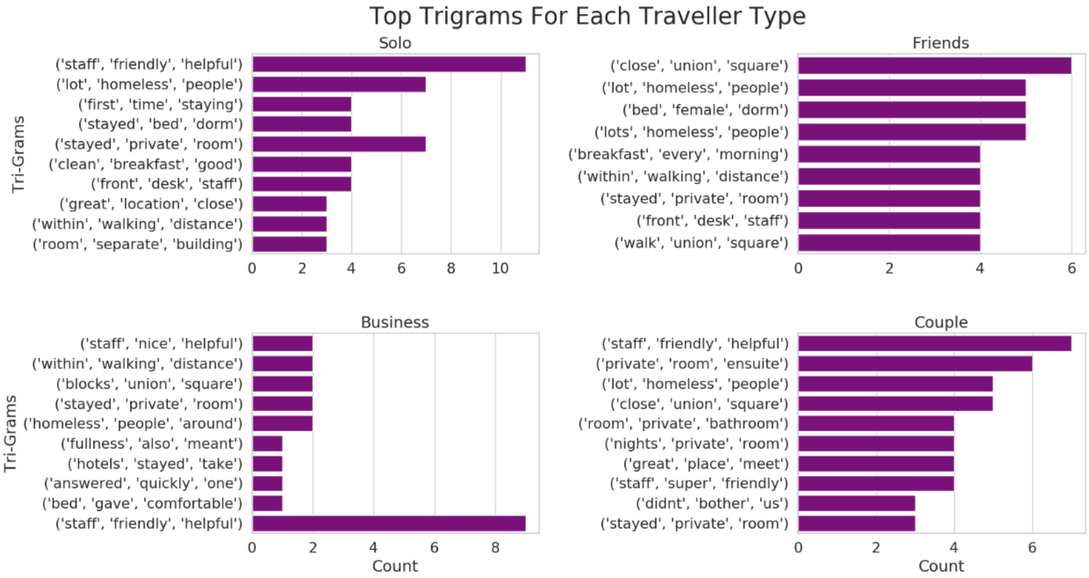


Figure 4 Top Word Count for Each Traveler Type

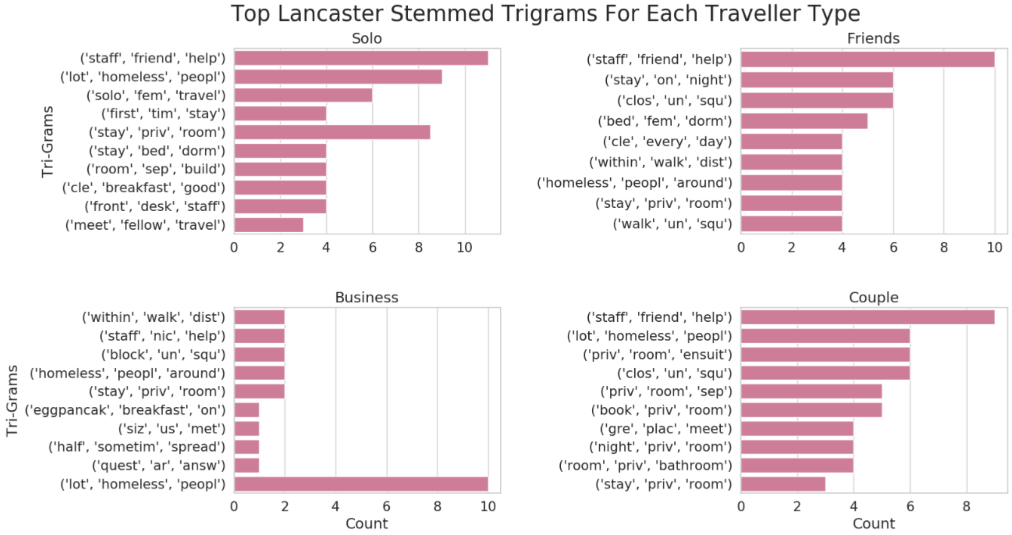


Figure 5 Top Stemmed Words for Each Traveler Type

By using complete words, there is some indication that solo travelers have higher emphasis on friendly staff. However, when Lancaster stem is applied the results are even more pronounced. Trends such as “solo female travel” and the need to “meet fellow travel” also make it to the top ten words. That combined with the previous top result of friendly staff, the hypothesis that solo travelers do in fact value human interaction with fellow travelers and service personnel seems to hold true.

Predictive Model

Method

The results of the exploratory analysis confirm that meaningful information to answer the question of solo travel motivation can be gained from just looking at top word count. From the results, it is evident that solo travel is distinctly different from travelers like business and couples and slightly different from those travelling with friends. If the theory that reviews from solo travelers differ perceptibly from the other types of travelers holds true, a simple supervised classification model should be able to accurately predict traveler type based on the review text.

To do this, the data is initially pre-processed in the same manner as described above. The processed data is tokenized using sklearn’s count vectorizer. By applying the method *fit_transform(),* the text is converted to a matrix with each word as a feature and for each row, a count of number of times a word appears in a review is marked. The feature names are extracted using *get_feature_names()* method. A data frame with the matrix of tokens, feature names, and column indicating the type of traveler is created. A final step in this data preparation is to convert the traveler type to numerical values. That is, it is labelled 1 if the traveler is solo and 0 if it is not. In this study, we only use binary classification as the objective is to distinguish solo travelers from others and having multiple labels is not necessary.

Once this data frame is prepared, steps to create and optimize classification algorithms are employed. In this paper, I focused on using an optimized decision tree classifier and comparing it with results from optimized random forest classifier. First step is to randomly split into training and testing data with all random state is set to 0 for repeatability. For each model, the hyperparameters are tuned using randomized grid search with 10-fold cross validation and the best model is used to predict the test set. The results of the hyperparameters chosen are shown in the table below.

Table 2 Hyperparameter Values for Classification Models

Hyperparameter	Decision Tree Classifier	Random Forest Classifier
Criterion	Entropy	Gini
Min weight fraction leaf	0.05	0.38
Max depth	8	8
n estimators	NA	400
bootstrap	NA	False

Results

The following plots show the variation in accuracy for over different trials of 10-fold cross validation for hyper parameter tuning.

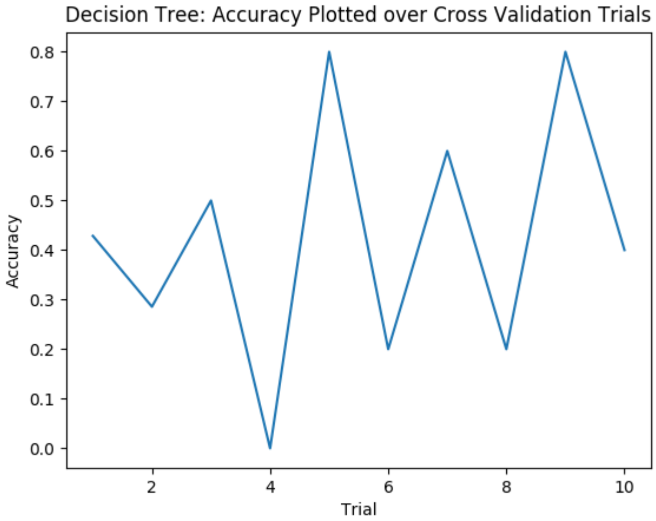


Figure 6 Accuracy from 10 - Fold Cross Validation

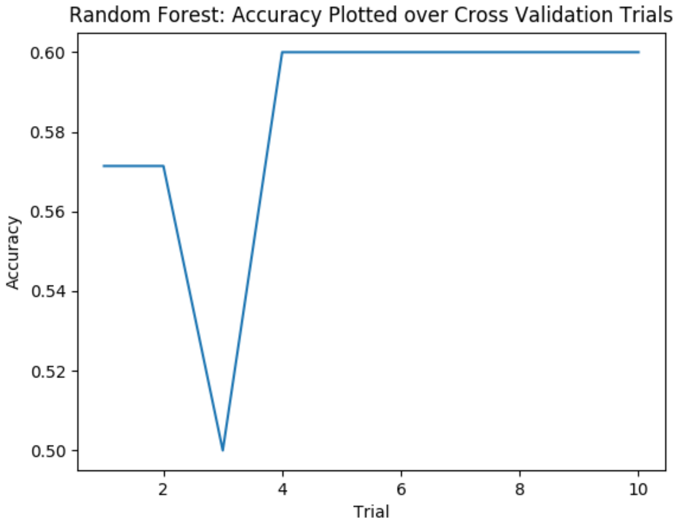


Figure 7 Accuracy from 10- Fold Cross Validation

The best model based on the tuned hyper parameters is selected and accuracy and time from the final model predictions are summarized below.

Table 3 Summary of Results from Classification Model

Model Type	Test Accuracy	Time Taken
Decision Tree Classifier	51.8%	7.04 [s]
Random Forest Classifier	62.1%	67.24 [s]

If we assumed the base line to be 50%, a tuned decision tree classifier just does marginally better. The random forest classifier's performance does improve significantly but the time taken is also longer. From these results, it appears that there is a perceptible difference between reviews of solo travelers and non-solo travelers that can be detected by a predictive model. By using a more robust model, the results improve significantly so further tuning to selecting impactful features would make the predictive ability even stronger.

Conclusion

Though most travel research does not prominently utilize computational methods to understand traveler motivation, the findings of this study show that it has a lot of potential. By conducting text analysis on Trip Advisor reviews for accommodation, it is evident that hotel or hostel reviews can be an important source to uncover travel motivation. The exploratory analysis indicates that different traveler type's motivation can be inferred from expectations as presented in the reviews. While everyone has opinions on basic requirements of a place of accommodations, different traveler types also have distinct expectations. It appears that solo travelers' value friendly staff and fellow travelers, friends seek proximity to tourist attraction, business travelers seek comfort and promptness and couples appreciate privacy. Additionally, a supervised learning model to classify whether a traveler is solo based on the review text reveals that is possible to distinguish solo travelers from reviews alone with an accuracy of ~60%.

The major limitation of this was certainly the lack of a larger data set due to difficulty in data collection. The insights from this study would be more meaningful and generalizable if a bigger corpus of data was available. The set back in data collection also impacted the variety and robustness of data analysis. While the current analysis does provide a good method to infer conclusions about what different travelers find important, additional text analysis techniques such as topic modelling or LDA could be utilized to computationally gauge important themes for each traveler type and similarity between types.

Nevertheless, this study provides and good foundation to appreciate the need to further investigate the travel motivations of solo travelers using computational methods. The work should be expanded to additional cities across the globe to find if the trends are generalizable. Additionally, travel reviews for attractions could also be a good resource to get insights on traveler differences. Also, since solo travel is a growing trend and must have seen transformation in the demographics as well. Future work with time in consideration would potentially reveal interesting insights.

References

Banerjee, Snehasish, and Alton Y.k. Chua. “In Search of Patterns among Travellers Hotel Ratings in TripAdvisor.” *Tourism Management*, vol. 53, 2016, pp. 125–131., doi:10.1016/j.tourman.2015.09.020.

Bianchi, Constanza. “Solo Holiday Travellers: Motivators and Drivers of Satisfaction and Dissatisfaction.” *International Journal of Tourism Research*, vol. 18, no. 2, 2015, pp. 197–208., doi:10.1002/jtr.2049

Cohen, Scott A., and Erik Cohen. “New Directions in the Sociology of Tourism.” *Current Issues in Tourism*, vol. 22, no. 2, 2017, pp. 153–172., doi:10.1080/13683500.2017.1347151.

Elliott, Annabel Fenwick. “Why Are so Many of Us Now Choosing to Travel Alone?” *The Telegraph*, Telegraph Media Group, 14 May 2018, www.telegraph.co.uk/travel/comment/whats-behind-the-rise-in-solo-travel/.

Valdivia, Ana, et al. “Sentiment Analysis in TripAdvisor.” *IEEE Intelligent Systems*, vol. 32, no. 4, 2017, pp. 72–77., doi:10.1109/mis.2017.3121555.

Wang, Yuanyuan, et al. “Applicability of Demographic Recommender System to Tourist Attractions: A Case Study on Trip Advisor.” *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2012, doi:10.1109/wi-iat.2012.133.

Muttecht@ustravel.org. “U.S. Travel Answer Sheet.” U.S. Travel Association, 3 Apr. 2019, www.ustravel.org/answersheet.

“Solo Travel Statistics and Data : 2018 - 2019.” Solo Traveler, solotravelerworld.com/about/solo-travel-statistics-data/.

Appendix

Data

All datasets used are available HERE

Methods

All code used to conduct data collection and analysis is available HERE

Results

Frequency of Words

Graphs 1 and 2 n-grams are shown below.

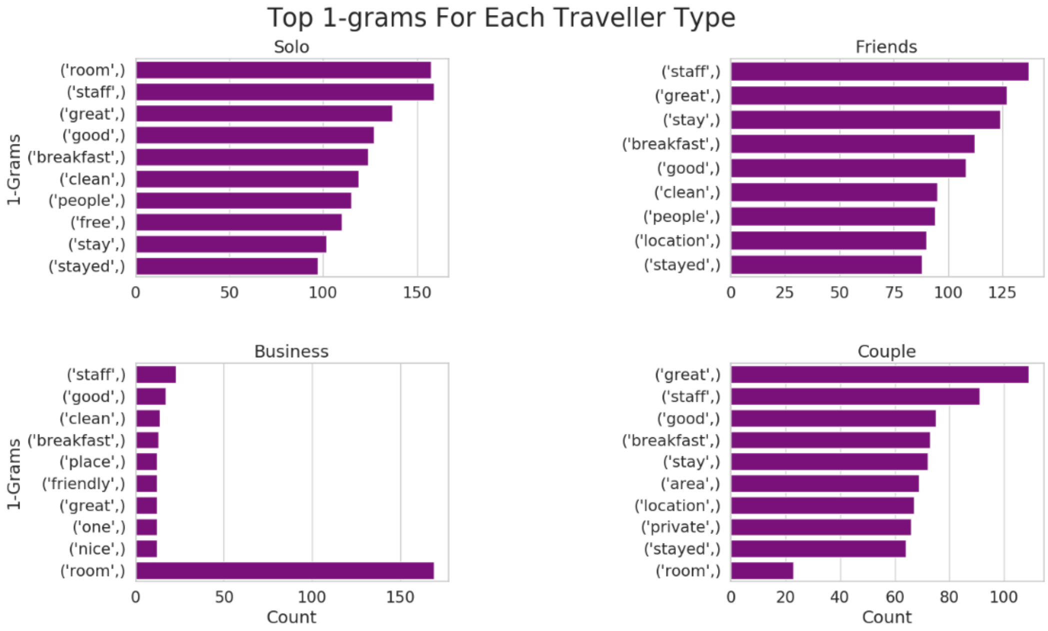


Figure 8 Top 10 words for each traveler type

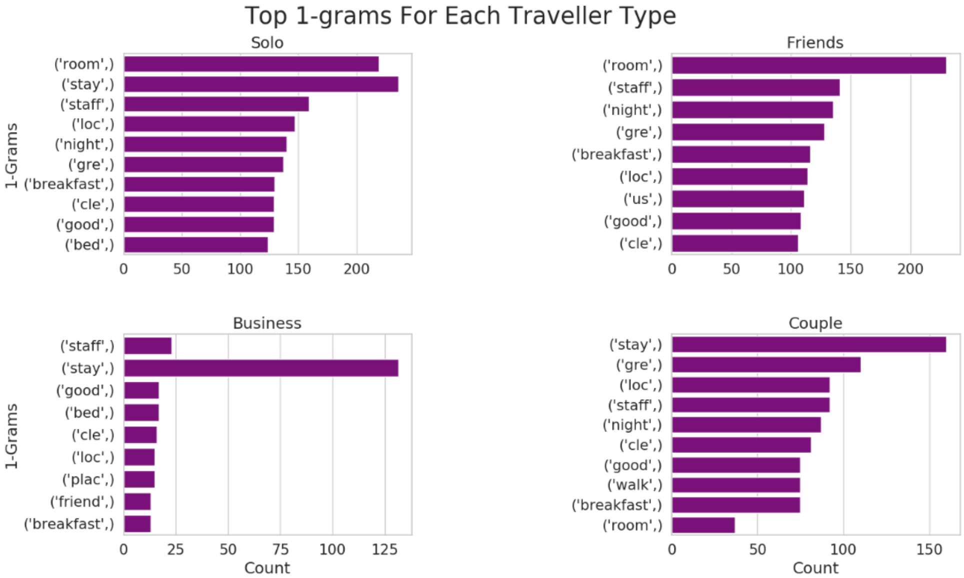


Figure 9 Top 10 words for Each Traveler Type – Stemmed

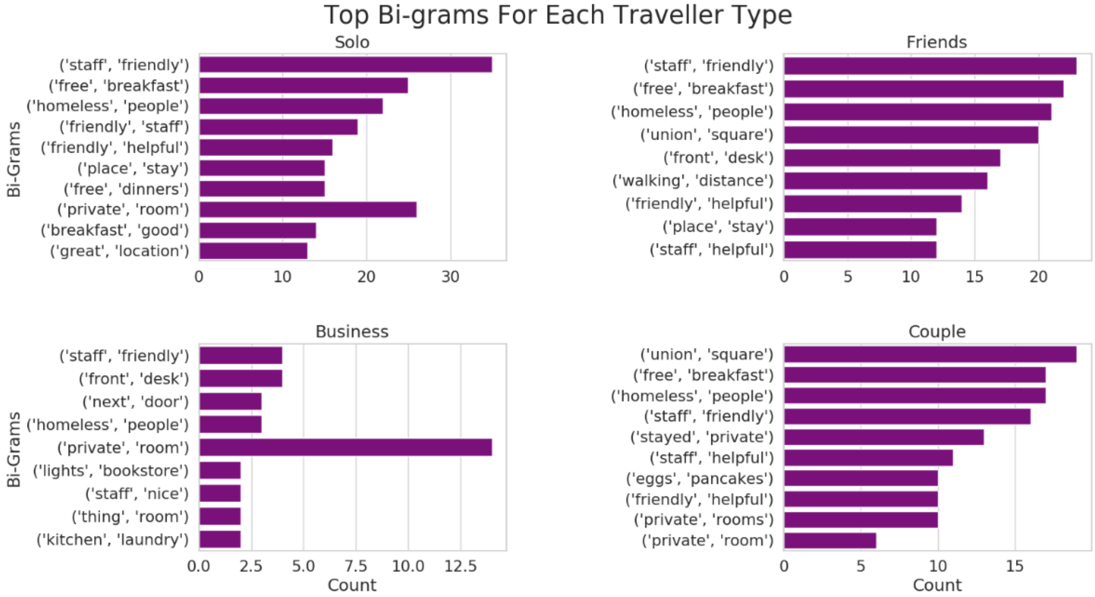


Figure 11 Bigrams for Each Traveler Type

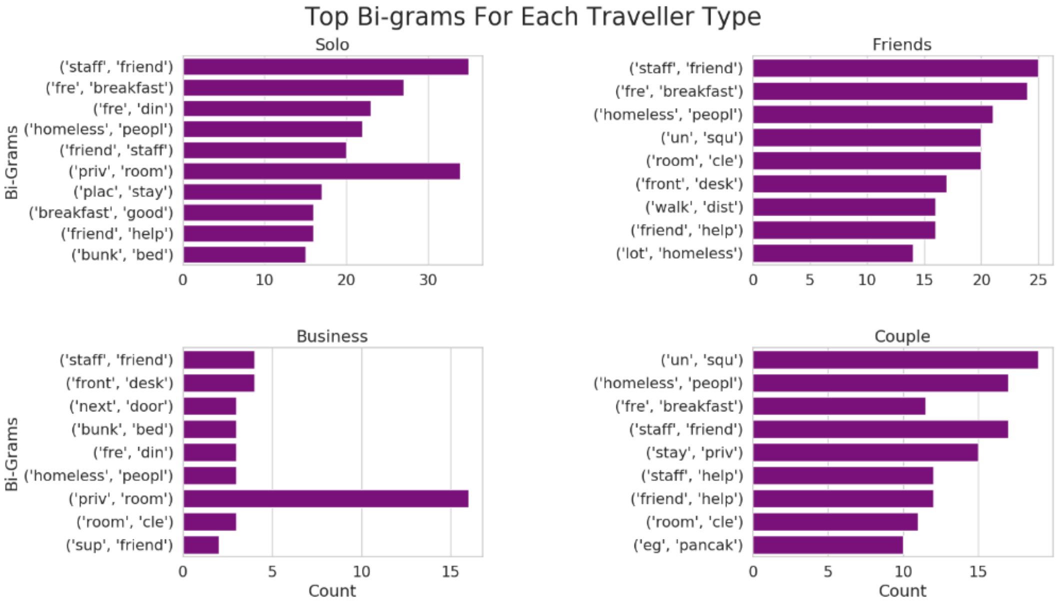
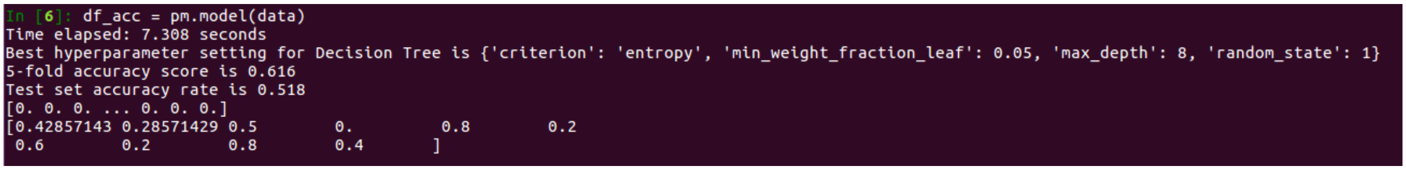


Figure 12 Bigrams for Each Traveler Type - Stemmed

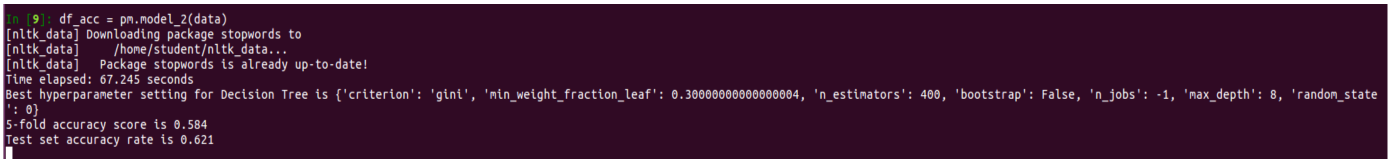
Predictive Model

The screenshots of the results of the predictive model show which settings are used and what was the resulting accuracy.



```
In [6]: df_acc = pm.model(data)
Time elapsed: 7.308 seconds
Best hyperparameter setting for Decision Tree is {'criterion': 'entropy', 'min_weight_fraction_leaf': 0.05, 'max_depth': 8, 'random_state': 1}
5-fold accuracy score is 0.616
Test set accuracy rate is 0.518
[0. 0. 0. ... 0. 0. 0.]
[0.42857143 0.28571429 0.5        0.         0.8        0.2
 0.6        0.2        0.8        0.4       ]
```

Figure 10 Screenshot of Python Terminal Results of Decision Tree Classifier



```
In [9]: df_acc = pm.model_2(data)
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/student/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
Time elapsed: 67.245 seconds
Best hyperparameter setting for Decision Tree is {'criterion': 'gini', 'min_weight_fraction_leaf': 0.30000000000000004, 'n_estimators': 400, 'bootstrap': False, 'n_jobs': -1, 'max_depth': 8, 'random_state
': 0}
5-fold accuracy score is 0.584
Test set accuracy rate is 0.621
```

Figure 11 Screenshot of Python Terminal Results of Random Forest Classifier