HostelWorld

Caitlin Mowdy

Hostelworld is a website that connects independent, budget, and youthful travelers to hostels. For the past 10 years Hostelworld has lead the market in online reservations for this demographic. The website lets users search for hostels given a location, group size, and dates. It also lets the user filter for prefered features and prices.

Project Goal

The goal of this project is to make hostel recommendation for a users based on other users that are similar to them, and hostels that are similar to those they have reviewed highly.

However most hostels have very similar features, and given that hostels are budget accommodation they usually fall in the same price range. For this reason the text of hostel descriptions and reviews needs to be taken into account when making hostel

recommendations. Below are two hostels in Budapest that can help demonstrate how different hostels and their guests can be even with similar price, score, and features.

Score: 9.1 | 1710 reviews | $11

Perfect Location, Superb Staff, Fantastic Cleanliness

Linen Included, Free WiFi, Security Lockers, Hot showers

Maverick Hostel

Located in the heart of the city in a fully renovated royal mansion built by the Hapsburg dynasty, the Maverick Hostel is waiting for you in its original magnificence. Because of our central location, unnecessary travel can be spared.

Most of the city's attractions can be reached within 15 minutes. Furthermore, the famous ruin bars and other exciting night programmes are just 10 minutes away on foot.

USA, Male, 41+

9.7   *A wonderful, ornate hostel! There's both a common area by the reception, and another common area/kitchen on each individual floor of the hostel. Wash and dry laundry service is only 7€, the coffee and tea are complimentary, and the rooms are spacious with no bunk beds. You can both wander around the tourist Pest side of town, or easily cross the bridge to Buda castle and the Buda side (20min walk/ 5 min bus). My dorm mates were boring, otherwise it was top notch.*

USA, Female, 18-24

9.4   *Quiet hostel right off of the metro line. Centrally located. Staff was helpful and the place was clean. Cool building and felt safe. Would stay again.*

Score: 9.1 | 2,314 reviews | $12
Perfect Location, Superb Staff, Awesome Cleanliness
Linen Included, Free WiFi, Security Lockers, Hot showers

Retox Party Hostel

This is not a place where people come to cleanse, rejuvenate or detox. It's a place where people come to drink, get wild and RETOX. It is a close family of drinkers who like to drink at the bar during the day and head out to party at night. If you want to get a good night's sleep and achieve a lot during the day, this is not the hostel for you, but if you want to see what Budapest's nightlife has to offer, not be judged and maybe squeeze some sightseeing in between, you have found the right place.

USA, Male, 18-24

9.4   *This is probably one of the greatest hostels in the world to get drunk at and then recover from a hangover. If you have other aspirations, try somewhere else.*

Australia, Female, 18-24

10.0   *One of the best hostels i've stayed at! As soon as I walked in I was looked after by all the staff, as I was by myself I got introduced to everyone and felt part of the family! There is events*

happening every night, drinks are cheap and the party lasts all night!! A big thank you to Popeye, Luke, Spanky & Vinny. I'll be back soon to see you all again!!

Collecting Data

To gather data for this project I randomly generated codes to insert in the following path
http://www.hostelworld.com/profile/16255753/reviews
After checking if the codes actually belonged to users' profiles I saved the codes as user ids. Using the ids and the path above I scraped user information, review text, and links to the hostels reviewed.

Using the hostel links I scraped for hostel details and descriptions, and checked that the links belonged to hostels that were still on hostelworld. Here are links to the code I used to test user IDs and the code I used for collecting user, review, and hostel details.

Cleaning Data

After scraping my data it contained a lot of html tags and and odd spacing. Also many of the reviews had belonged to hostels that didn't exist anymore.
  ● I deleted the reviews that had belonged to the hostels that were no longer on hostelworld
  ● Deleted the reviews that weren't in english
  ● Deleted the users that were then left with no reviews
  ● Removed html tags, unicode, and odd spacing
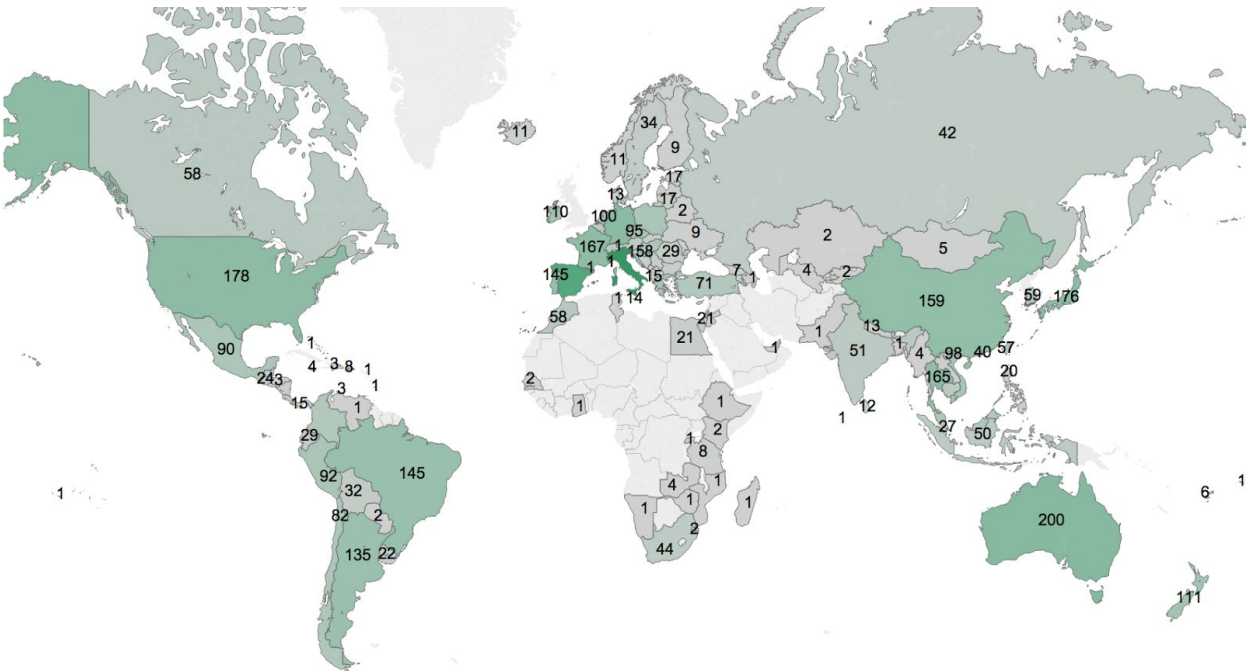 ● Added country and hostel ids
Here is the link to the code I used to clean my data and add hostel and country ids.

EDA

Hostels Data

In the hostels data I found that

- Most hostels have a score between 8 and 9.5
- Italy, Spain, and Australia have the most hostels


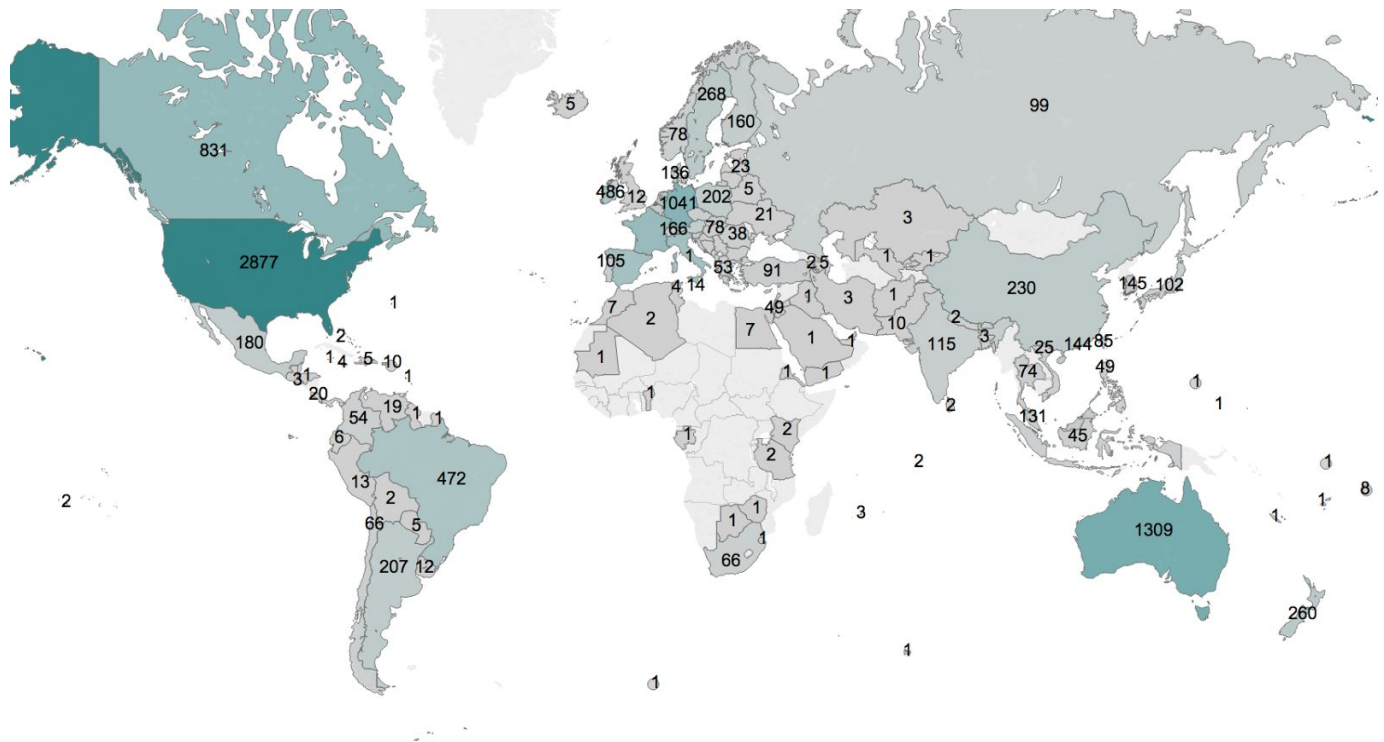<div align="right">Map of hostels by country</div>

Review Data

- In the review data I found that Italy, Spain and England are receiving the most reviews. This corresponds with the information above showing that Italy and Spain have the most hostels in the data and England is in the top 5.
- The Users that leave the highest reviews are in Tunisia, Antigua, and San Marino

- The months with the most reviews are Jan, Dec, and July. This shows that most people review over the winter and summer holidays. Which means we can infer that most people are staying in hostels during these months.
- I also looked at the distribution of review scores. This showed that most reviews are given high scores.
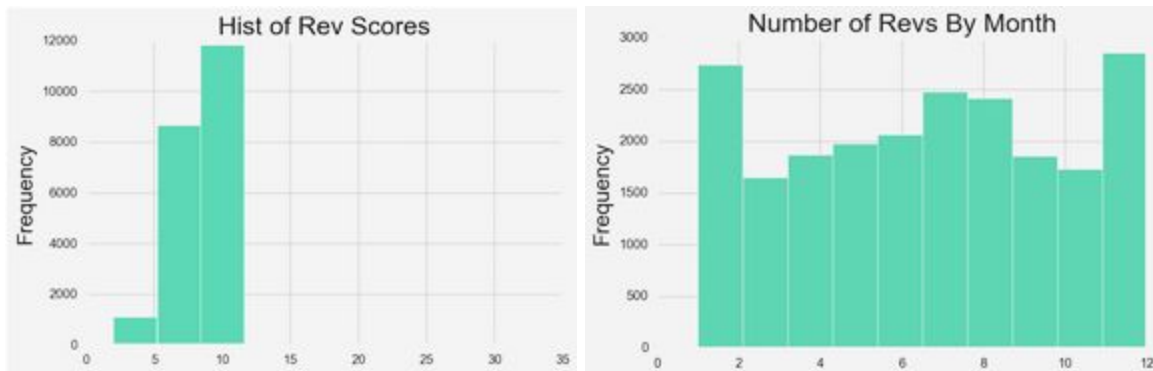
Users Data

- Most reviewers are from USA, England, and Australia
- The most common travel group types are male, female, and couple
- Most reviews did not specify an age group, but out of the ones that did 25-30 was the most common and 18-24 was second

Map of users by country

This is the link for my **eda and graphs**.

Topic Modeling

I decided to use topic modeling on both my reviews and hostel descriptions. For both of them I created 10 topics and found the topic probabilities for every review and hostel description. Here are the links to the topic modeling for **hostels** and **reviews**.

Looking into User and Hostel data

Once I had the topic probabilities I edited some of my hostel features to fit a regression for hostel scores. I eventually gave up on fitting a regression model for hostel ratings and review scores. Below are correlation matrix that help explain why my regression scores were so low.
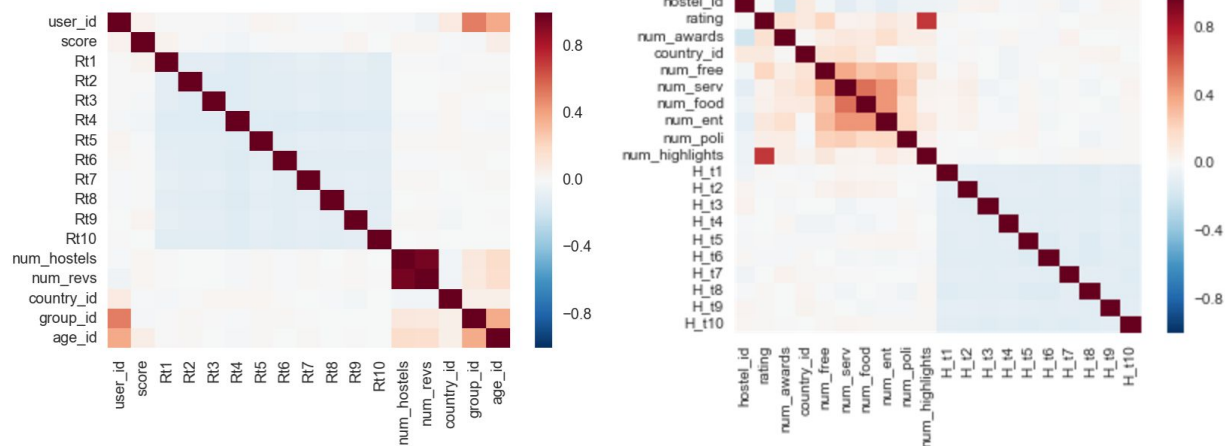
User and Hostel Distance

I used jaccard distance for finding the distances between hostels and users.

- len(intersection of setA and setB) divided by len(union of setA and setA)

To use jaccard distance I first had to change my user and hostel information into sets.
Below is an example for a hostel and a user

```
3637809
[The Naughty Squirrel Backpackers Hostel, Latvia, >8.5, Hs:1, nationality: Australia, Male, 31-40]
```

```
The Naughty Squirrel Backpackers Hostel
[>8.5, Awards:16, Latvia, Riga, Linen Included, Free City Maps, Towels Included, Free WiFi, Free Internet Access , Se
curity Lockers, Key Card Access, Common Room, Adaptors, Book Exchange, Hot Showers, Self-Catering Facilities, Hair Dr
yers, Safe Deposit Box, Steam Room, Internet Access, Laundry Facilities, Luggage Storage, 24 Hour Reception, Tours/Tr
avel Desk, Postal Service, 24 Hour Security, Bar, Tea & Coffee Making Facilities, DVD's, Board Games, Foosball, Credi
t Cards Accepted, No Curfew, Non Smoking, Taxes Included, H_t6]
```

I created a distance function for users that would take a given user and find the distance between that user and every other user. It returns the users closest to the given users and lists of hostels those users have rated highly. Here is the link to my **user distance code.**

```
j_user_dist(3298399)
```

	users	distances	hostels
6563	8144810	0.312500	[Iva Apartments]
3767	9802505	0.300000	[El Jardin Hostal, Way Hostel, Hostel Rosemary]
9161	9936860	0.277778	[Boipeba Bed and Breakfast, Hostal Andalucia]

The function for finding hostel distance is very similar to the function for user distance. It takes in a given hostel and a list of hostels, and finds the distance between the first hostel and every hostel in that list. It returns the hotels closest to the first hostel and links to them. Here is the link to my **hostel distance code**.

```
j_hostel_dist('Bed & Bike Barcelona',h_list)
```

	hostel	distances	link
2	Way Hostel	0.530612	http://www.hostelworld.com/hosteldetails.php/W...

Using the hostel distance and user distance functions I made a function that would give hostel recommendations for a given user. The function first uses the user distance function to find the users closest to the given users. It collects the hostels the given user has rated highly, and the hostels the closest users have rated. The function then takes every hostel the given user has rated highly and compares it to the list of hostels the best matched users have rated highly using the hostel distance function. It then returns a list of recommended hotels and links to those hostels. Here is the link to my **hostel recommendation code**.

Using the above distance codes my recommendation function made these suggestions for user 3298399.

```
hostel_rec(3298399)
```

	user_hostel	hostel	distances	link
0	Bed & Bike Barcelona	Way Hostel	0.530612	http://www.hostelworld.com/hosteldetails.php/W...
1	B&B Giovy	Hostal Andalucia	0.375000	http://www.hostelworld.com/hosteldetails.php/H...
2	CroParadise Green Hostel	Way Hostel	0.418182	http://www.hostelworld.com/hosteldetails.php/W...
3	B&B Giovy	Hostal Andalucia	0.375000	http://www.hostelworld.com/hosteldetails.php/H...

It found that user was user 8144810 the closest user with a jaccard score of 0.312500.

And out of the best matched users it recommended Way hostel with a jaccard distance score of 0.530612 to Bed & Bike Barcelona.

Conclusion

My distance function is not perfect, but the outcomes I've observed are decent recommendations. If I had more time for this project I would like to improve my review and hostel description topics as well as test other methods for making recommendations.