Topic modeling and sentiment analysis for Airbnb text data

Sara Sun, Claudia Zhang, Jennifer Mei, Zhuofan Lin, Songhua Liu

Executive Summary:

In this project, our group analyzed text data in the listings and reviews dataset of the Airbnb Boston Area from September 2009 to February 2021. Specifically, we focused on neighborhood description, room description, and customer reviews.

We identified neighborhood themes, room characteristics, and customers' preferences for those themes and characteristics; we also explored customers' considerations when giving reviews and the relationship between numeric rating and text reviews. Based on those findings, we provided three pieces of operation and marketing suggestions for Airbnb.

- Provide coupons if renters write comments of more than 30 words.
- Target customers based on their preferences and provide content-based recommendation
- Exploit more rooms a little bit far from downtown but can give renters a good experience.

We performed sentimental analysis and topic modeling. For sentimental analysis, we chose the VADER model, considering slang, punctuation, capital words, and syntax. We excluded reviews with a foreign language for convenience. For topic modeling, we first normalized and vectorized and then chose the Latent Dirichlet Allocation (LDA) method.

Data Description:

We found publicly available data of Airbnb at the Inside Airbnb website. The data have been cleansed and aggregated by the website. We use Boston listings data scrapped in February 2021 and reviews data from August 2009 to February 2021. The data could be download at http://insideairbnb.com/get-the-data.html

There are two CSV files, "listings.csv" and "reviews.csv" which contain information about rooms listed on Airbnb and customer reviews for those rooms. The listings dataset has 74 columns. For topic modeling, we chose "description" which describes the room, and "neighborhood_overview" which describes the overall environment. The reviews file has six columns; we used the column "comments" which listed each customer's review of the rental experience to conduct topic modeling and sentiment analysis. There are 2960 rows in the listings dataset and 109,721 rows in the reviews dataset.

After viewing the dataset, we had some initial concerns. First of all, we are worried renters' positive reviews might have a potential bias since hosts also rate and review renters. Secondly, we are concerned that most of the data contain similar information since there are limited words used to describe a house or an Airbnb experience. Thirdly, we noticed that a small percentage of the data is in foreign languages. Despite these shortcomings, the dataset fits our objective well.

Methodology:

- **Sentiment Analysis**

The first method we took was the sentiment analysis on reviews to see whether sentiment scores align with users' ratings. The scores can provide us more insights on users' preference for neighborhood and listing selection. Our review data contains punctuation, numbers, and foreign language words. Since the model we chose was VADER, which evaluates a sentence not just based on the meaningful words but also slang, punctuation, capital words, and syntax, we decided to use the original review without cleaning. In order to get rid of the reviews in foreign languages, detect function from Python langdetect package is utilized. If the function returns “en", we kept the review in our database; otherwise, the review would be deleted.

Two difficulties encountered in the process were the lack of labels for the reviews and the bias towards positive reviews. As shown in figure 1, most of the compounding scores are above 0.5. In order to compare the results, we set 5 levels, from “A” to “E”, based on the distribution of all the scores (the cutoff is listed in table 1). Those five buckets can help us compare the relative preference customers have on all the listings. However, as we did not have enough human resources to label the reviews manually, we cannot compare the accuracies of different thresholds and decide what the threshold is for “positive” reviews. The current threshold is set to make the five levels interpretable and meaningful.

The advantage of using VADER is that reviews are usually in casual language, and VADER includes those casual expressions into evaluation. On the other hand, as the data lack labels, the accuracy of the sentiment score remains uncertain.

- **Topic Modeling**

To understand what is behind the text information, we used the Latent Dirichlet Allocation (LDA) topic modeling method, one of the most useful ways to understand the text.

There are 3 columns contains text information - "description" and "neighborhood_overview" in the listings dataset and "comments" in the reviews dataset. Each of these columns is a corpus consisting of up to 2960 documents (the number of listings). The steps we took for LDA topic modeling are summarized below.

1. Normalize Text Data

The first step was to normalize the text data to help the LDA model better discover the topics. We first cleaned up HTML markups, expanded contractions, stemmed and lemmatized to bring the words to their original form, removed special characters such as punctuation marks, get rid of stop words, and removed accents from characters. When processing stop words, besides the common stop words list generated from nltk package, some frequent but meaningless words are added as customized “stop words”. For example, in room descriptions, "license" is a very common word, as many hosts list their license number. "Boston" is another common word as data was collected from Airbnb in Boston. "home" is extremely common because all of the listings are about home. As a result, we added "license", "Boston", "home" to the stop words list when normalizing room descriptions.

2. Vectorize normalized text data

Then, we extracted features from each of the topics by using the Bag-of-Words vectorization method. We chose Bag-of-Words instead of TF_IDF for LDA because LDA is a probabilistic model that tries to estimate probability distributions for topics in documents and words in topics. The weighting of TF-IDF is not necessary for this.

3. Build the model and tune hyperparameters

After the corpus has been normalized and vectorized, we started to build the model. In the LDA model, three

hyperparameters need to be tuned: the number of topics (K), Dirichlet hyperparameter alpha (document-topic density), and Dirichlet hyperparameter beta (word-topic density). So we used the cross-validation method to tune the hyper-parameters and find the best model. In cross-validation, we used log-likelihood as the evaluating criteria to compare the models generated with different hyperparameters.

4. Create a label for each topic

Once we finished feeding the corpus to models generated from step 3, we created labels for each topic by considering the top words and rare & exclusive terms. The visualization for each topic is shown in figure 2-4. The name of topics for “description”, “neighborhood_overview”, and “comments” are listed in the appendix table 2-4.

5. Assign documents to topics

Each document typically contains a set of probabilities corresponding to each possible topic; the last step in topic modeling is assigning a dominant topic to each document in our corpus, and that topic tells us what the document is talking about. For room description and neighborhood overview, we used the topic with the highest probabilities as the topic for a document, but we kept the top three topics for reviews. Review topics show why customers like or dislike the room, and it is common for a customer to list multiple reasons. As a result, we believe keeping the top 3 topics can provide better insights for business decision-makers.

The essence of topic modeling is built around the idea that the semantics of our document is actually governed by some hidden or "latent" variables that we are not observing. It is so meaningful that it makes "dark data" (data collected during regular business operation but never used to derive insights or decision-making) become helpful for decision-makers. For example, we figured out what people like or dislike the rooms by running topic modeling reviews.

However, natural language is messy and ambiguous. Everyone has a different writing style, so there are some limitations of LDA. First, the output of topic LDA is unlabeled, meaning that we have to use our domain expertise or intuition to think of logical names for each topic. To achieve this goal, we read many text data for each corpus to understand the corpus deeply. Second, evaluation is generally complex since LDA is unsupervised, and we do not have absolute criteria to judge how accurate the model is. Another limitation is that LDA assumes topics are based on a multinomial distribution, and words are based on another multinomial distribution trained specifically to that topic. However, the actual distribution could be more complex in reality, and there will be some bias

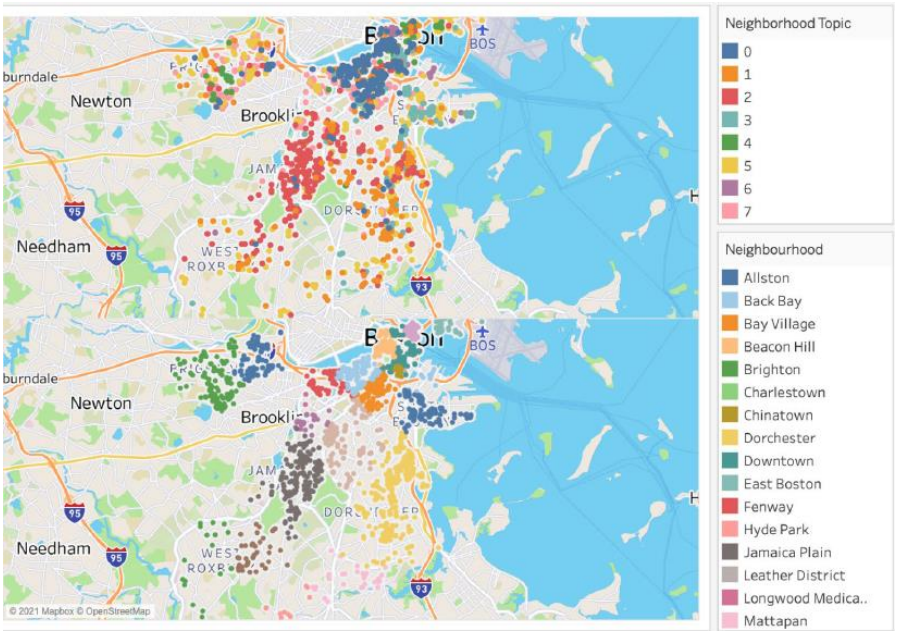
Results and Discussions:

- Neighborhood Analysis

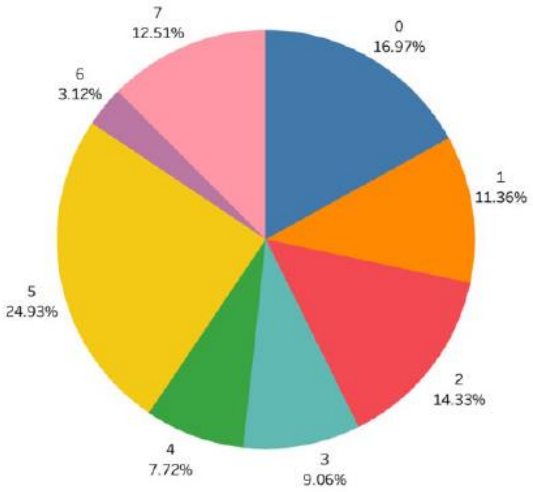
In Boston, there are 25 neighborhoods. In the first step, we were interested in whether those neighborhoods have characters in common or are entirely different. By using the topic modeling, we get eight topics using neighborhood descriptions.

Topic_0	Downtown, finest, high-end lifestyle, historic charm
Topic_1	Racially &culturally& economically diverse, peaceful, quiet, and safe
Topic_2	Racially &culturally& economically diverse, close to parks and ponds
Topic_3	Close to beach and downtown, busy, culturally rich
Topic_4	Close to famous colleges/universities and tourist destinations, the heart of Boston
Topic_5	Quiet, safe, residential, local restaurant
Topic_6	Entertainment, relaxing, new, convenient transportation
Topic_7	Proximity to hospitals and universities, with many restaurants

We can see that most of the listings with topic 0, which is related to downtown and high-quality life, are located in downtown Boston. Topic 5, which represents good restaurants and safe neighborhoods, is more likely to locate in the north of Boston. Houses from Topic 2, which were related to parks and diversity-friendly neighborhoods, tend to locate in the south of Boston and are far from downtown. While downtown, southern and northern Boston seem to vary in the description, as we only extracted eight topics from 25 neighborhoods, some neighborhoods share some characters in common, and some areas, such as the west part of Boston, have a mix of topics. Among the listings that have neighborhood descriptions, 25% are from topic 5, and 16% are from topic 0. The hosts' neighborhood descriptions are written to attract the customers, so it reflects how hosts think of customers. Hence, we could conclude that the hosts believe a quiet environment and safety are important factors in customers' decision process. That is why the hosts often emphasize it in the description. Another critical point is that the host customers are tourists who may desire to live downtown, as it may be easier for tourists to visit the city.

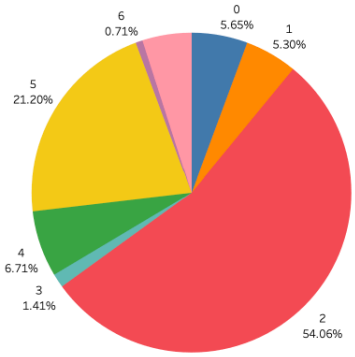


Neighborhood Topic Distribution (Up) and Neighborhood Geographical Distribution (Below)



Percentage of Neighborhood Topics Among All Listings

Whether there is a difference in sentiment scores among neighborhoods?
We then compared the average sentiment scores by neighborhood. The average neighborhood sentiment score is 0.7980, and the minimum score is 0.7042. It demonstrates that all the neighborhoods have a relatively high average sentiment score. As the top 5 neighborhoods of the highest average sentiment score – Longwood, North End, Roslindale, Jamaica Plain, and Chinatown, are not geographically close to each other, we were interested in which topic is the most popular among the top 5 neighborhoods to see whether they are similar in characters.



Percentage of Topics Among Top 5 Neighborhoods

The pie chart shows the percentage of topics in those neighborhoods, and we can observe that topic 2 (54.06%) and topic 5 (21.2%) are the most common. As customers write reviews, the sentiment score and the pie chart reveal what customers value the most when living in an Airbnb. From customers' point of view, they love to live in a beautiful neighborhood (ponds and parks) and is diverse in culture and race. As hosts expected, customers also love quiet and safe places with local restaurants.

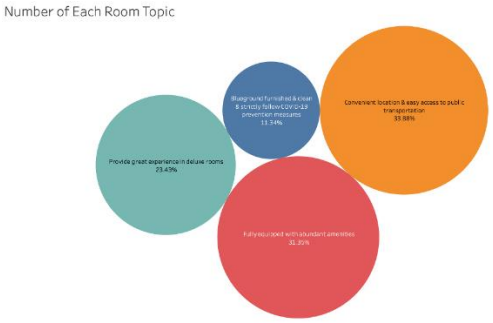
- **Room Analysis**

Which room topics are the most common?

After conducting topic modeling for room descriptions, we have obtained four meaningful topics for the rooms.

Topic_0	Blueground furnished/clean & strictly follow COVID-19 prevention measures
Topic_1	Provide great experience in deluxe rooms
Topic_2	Convenient location &easy access to public transportation
Topic_3	Fully equipped with abundant amenities

First of all, we explored the percentage of each room topic to get insights such as what attributes are most valued by the host when they design and describe their house.

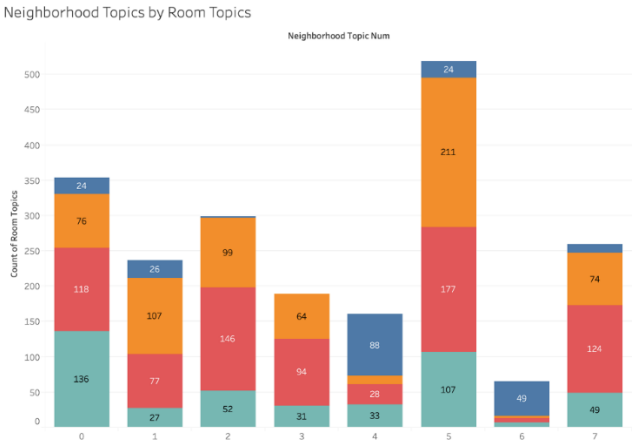


Percentage of Each Room Topic

Topic 2 is the most popular one (33.88%), followed by topic 3 (31.35%). By this means, most hosts believe that ease of access and fully equipped facilities are essential to attract users to choose their houses in Boston.

What is the relationship between neighborhood topics and room topics?

We drew a graph to find out if neighborhood topics are related to room topics.



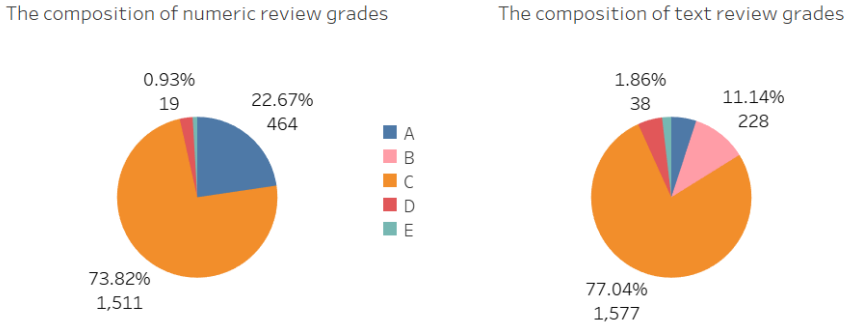
Relationship between Neighborhood Topics and Room topics

Convenient location and fully equipped facilities play important roles in neighborhoods 1, 2, 3, 5, and 7 areas. Neighborhood topic 0 offers the most housings with excellent guest experience and room design, followed by neighborhood topic 5, 2, and 7. While room topic 0, "Blueground furnished & clean & strictly follow COVID-19 prevention measures," is most valued by neighborhood topics 4 and 6, showing these two areas are more likely to follow Covid-19 policy and furnished requirements strictly.

- **Reviews Analysis**

Are numeric ratings and text reviews consistent?

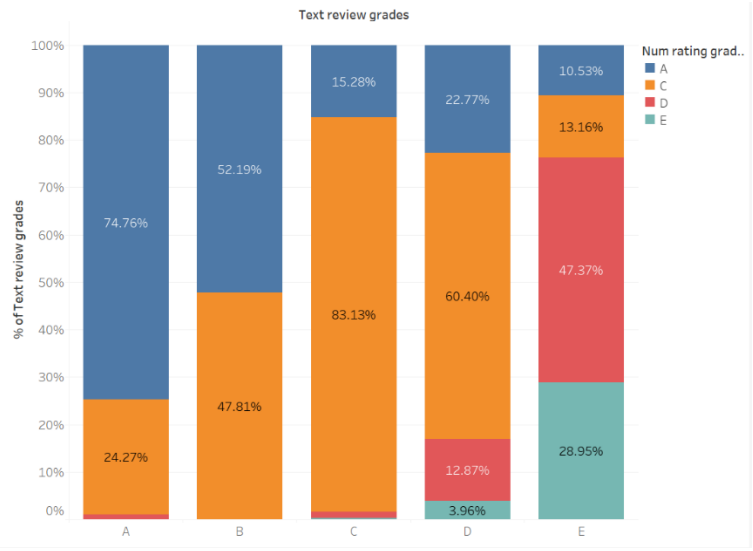
When people review rooms, they usually leave a quantitative evaluation (numeric ratings) and use words to express their feelings (text reviews). We wondered whether the numeric ratings and text reviews are consistent. We categorized numeric ratings into five levels and gave each numeric rating a grade, from "A" to "E". To compare the two evaluation dimensions, we labeled them based on the distribution of the five grades in text reviews. For example, 1.86% of the listings get "E" s in terms of text reviews, so we also wanted to label 1.86% of the listings as "E" in terms of numeric ratings. The results indicate there is some inconsistency between numeric ratings and text reviews.



The distribution of numeric and text review grades

As shown in the figure, the percentage of "A" is much higher in numeric rating; no numeric rating is labeled as "B"; the percentages of "D" and "E" are much higher in numeric rating. The reason is that there is a large number of tie-numeric ratings. For example, 471 listings have a numeric rating of 100! When both numeric ratings and text reviews suffer from positive bias, numeric ratings seem to suffer more.

Besides the overall distribution, we also explored the relationship between numeric ratings and text reviews.



The relationship between numeric ratings and text reviews

Text reviews of “A”, "B", and "C" mainly contain listings with numeric ratings of “A” and “C”; from text review “A” to “C”, the percentage of listings with numeric ratings of “A” decrease while the percentage of listings with numeric ratings of “C” increase, which is reasonable. However, we found that 22.77% of listings with text review “D” has numeric ratings “A” and 10.53% of listings with text review “E” has numeric ratings “A”, indicating a huge discrepancy.

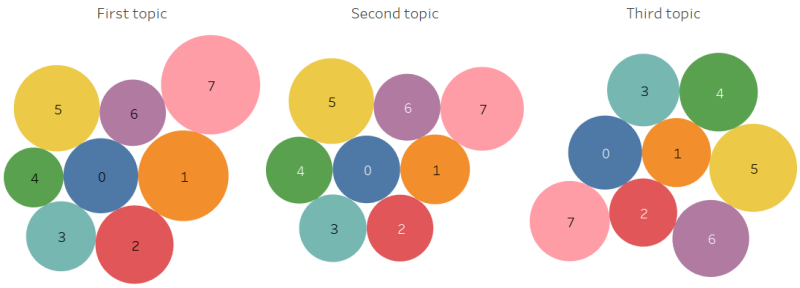
We checked the reviews written for these listings and had three findings. First, some reviewers give 100 numeric ratings even if they pointed out some problems in the room. For example, the reviewer for North End Furnished Condo said "difficult to find parking" while he/she gave a 100 numeric rating. Secondly, the sentiment analyzer performs poorly when the wording is a little bit tricky, causing some fantastic listing to be classified as "E". For example, a reviewer said he/she previously had some bad experience with hotels and then found this fantastic room, but the sentiment analyzer gave a very low sentiment score. Thirdly, this huge gap is more likely to occur when the number of reviews is small. The above findings indicate that text reviews could provide more distinguishable and hierarchical evaluations, which reduced the problem of positive bias a little. However, the grade from text reviews will be misleading if the text is tricky.

What are the reasons for liking or disliking a room?

After topic modeling for customer reviews, we obtained 8 meaningful topics for the customer reviews.

Topic_0	Kitchen, food, bed
Topic_1	Nearby restaurant, quiet environment
Topic_2	Check-in &out, parking
Topic_3	quiet environment
Topic_4	Airport, transportation
Topic_5	Cleanness, communication with host
Topic_6	quick response from host
Topic_7	helpful and responsive host, amenities

As discussed previously, we selected three out of eight topics with the highest probability to reflect people's reason for liking or disliking the room.



The number of reviews under different combination of topics

First topic	Second topic	Third topic 0	1	2	3	4	5	6	7
0	1			223	281	196	287	187	283
	2		214		280	237	314	274	284
	3		246	332		209	316	278	277
	4		179	205	201		269	181	246
	5		194	302	373	589		311	305
	6		173	262	273	194	262		257
	7		315	295	339	256	335	314	
1	0			391	378	374	419	363	524
	2	375			362	314	333	311	384
	3	352		327		319	299	247	385
	4	308		264	263		508	340	581
	5	324		285	268	543		382	625
	6	261		247	267	349	428		519
	7	476		392	409	674	864	538	
2	0		352		442	216	344	258	325
	1	353			375	289	302	243	297
	3	545	365			339	482	310	402
	4	168	164		181		213	148	185
	5	301	226		346	301		259	408
	6	250	221		230	167	280		252
	7	292	260		314	237	454	287	
3	0		248	393		154	291	155	290
	1	231		265		202	203	149	239
	2	471	308			318	383	248	316
	4	139	132	158			184	99	125
	5	264	170	350		228		220	398
	6	152	125	146		98	180		185
	7	270	199	321		193	455	229	
4	0		115	83	94		166	84	109
	1	129		127	142		255	178	269
	2	418	113		112		129	86	119
	3	88	115	104			152	61	120
	5	216	327	159	163			391	396
	6	81	123	77	65		213		189
	7	132	369	121	189		474	171	
5	0		156	183	282	187		278	415
	1	148		166	161	284		320	372
	2	224	140		277	233		222	370
	3	244	178	251		210		182	473
	4	570	314	201	229			295	657
	6	370	234	205	219	245			791
	7	561	832	472	748	967		1,144	
6	0		156	142	129	131	214		243
	1	164		170	144	223	253		284
	2	156	163		143	130	220		187
	3	137	103	136		105	155		207
	4	132	164	125	90		251		244
	5	291	216	244	202	261			520
	7	352	290	257	253	297	612		
7	0		255	247	330	205	674	426	
	1	265		197	240	372	1,976	316	
	2	233	167		348	184	464	362	
	3	320	287	335		244	563	249	
	4	225	307	190	213		849	284	
	5	636	962	535	831	1,894		1,532	
	6	330	206	244	301	290	838		

The combination of the top three topics

If we look at the first, second, and third topics separately, we found no dominating topics. However, if we look at the first, second, and third together, we found that the combination of topics 7&1&5, 7&5&4, 7&5&6, 5&7&6 are most popular. It indicates that guests care most about whether the host is helpful and responsive, whether the amenities work well, whether the room is clean, whether the host is easy to communicate with and can get quick responses from the host, whether the room is accessible to airports and transportation.

Conclusion:

Throughout the process of performing sentimental analysis using VADER model and topic modeling using LDA model, we found that neighborhoods and rooms play critical roles in determining renters' satisfaction degree with Airbnb. Specifically, renters prefer rooms with safe environments, with beautiful landscape, diverse culture, and

good restaurant around; renters like rooms with deluxe design, fully equipped facilities, and those consider COVID-19 prevention; renters prefer rooms with clean environment and easy access to transportations, and rooms with responsive and communicative hosts. We also noticed that renters' numeric ratings are somewhat inconsistent with text reviews, and the numeric rating suffered more from more positive bias.

Therefore, we came up with some recommendations:

- Text reviews suffered less positive bias than numeric ratings, so Airbnb could find ways to encourage renters to leave longer text reviews. For example, Airbnb could provide coupons if renters write comments more than 30 words.
- Airbnb could target customers based on their preferences. For example, Airbnb could use renters' renting history to figure out their favorite topics of neighborhood and room types. Then, certain rooms could be targeted specifically to the customers, such as "Top 10 houses with deluxe-designed rooms" or "Top 15 houses around by best restaurants".
- Even though downtown areas are popular among tourists, most renters preferred rooms with a quiet and safe environment and rich culture. As a result, Airbnb and hosts could exploit more rooms a little bit far from downtown but can give renters a good experience.

Appendix:

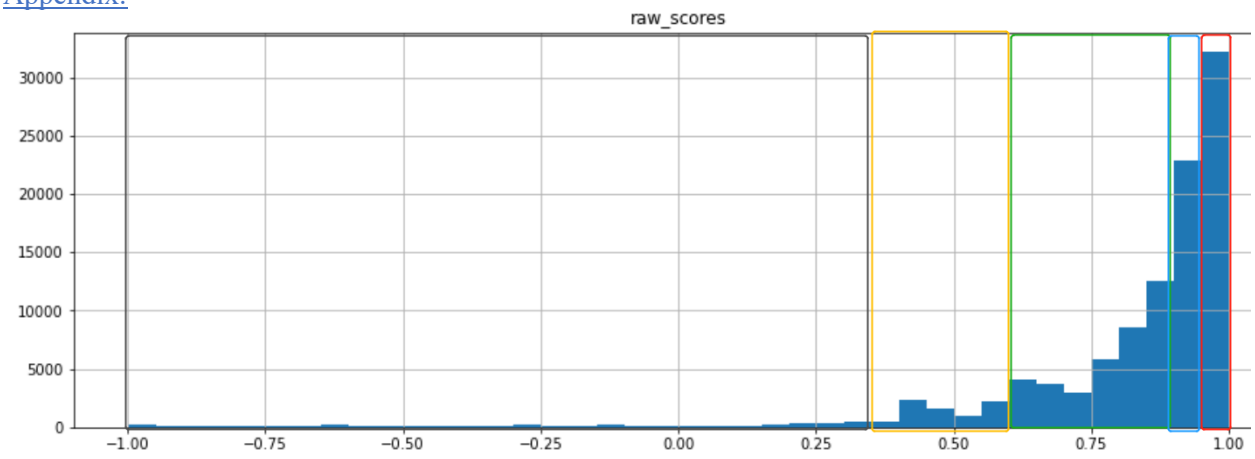


Figure 1: Distribution of sentiment score

Table1: Rules to assign grades to sentiment scores

Sentiment Score	**Text Reviews Grades**
(0.95,1]	A
(0.9,0.95]	B
(0.6,0.9]	C
(0.3,0.6]	D
[-1,0.3]	E

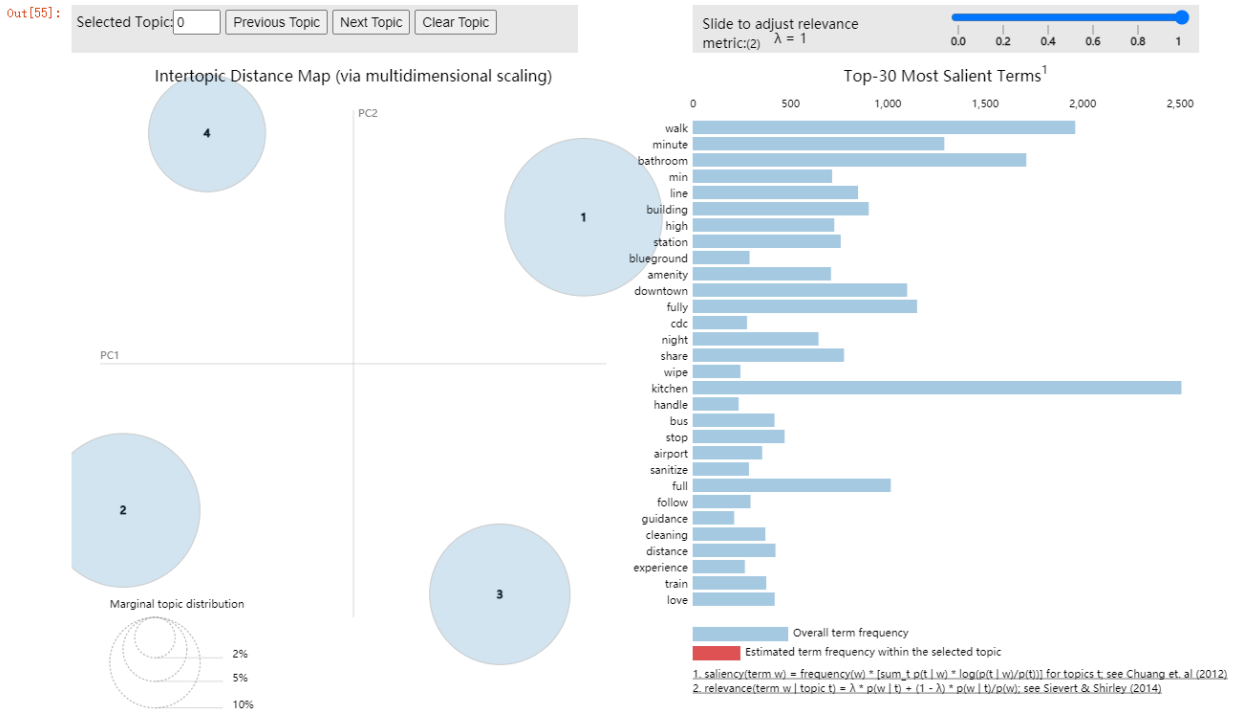


Figure 2: Visualize the topics for room description

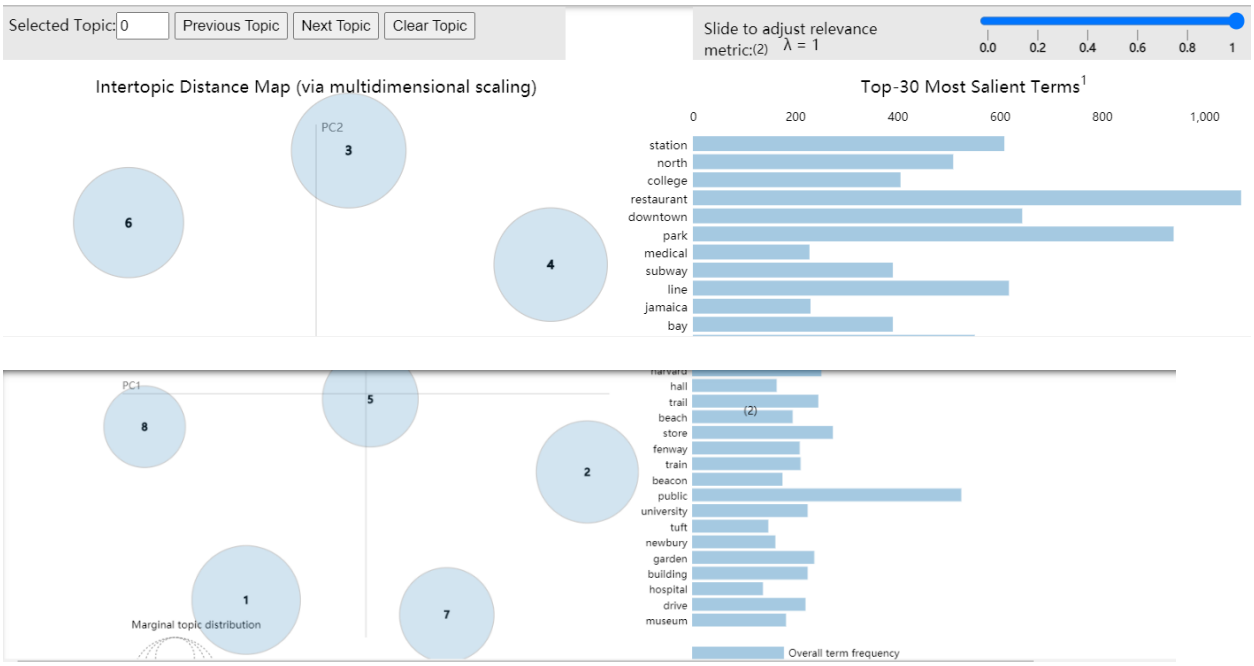


Figure 3: Visualize the topics for the neighborhood overview

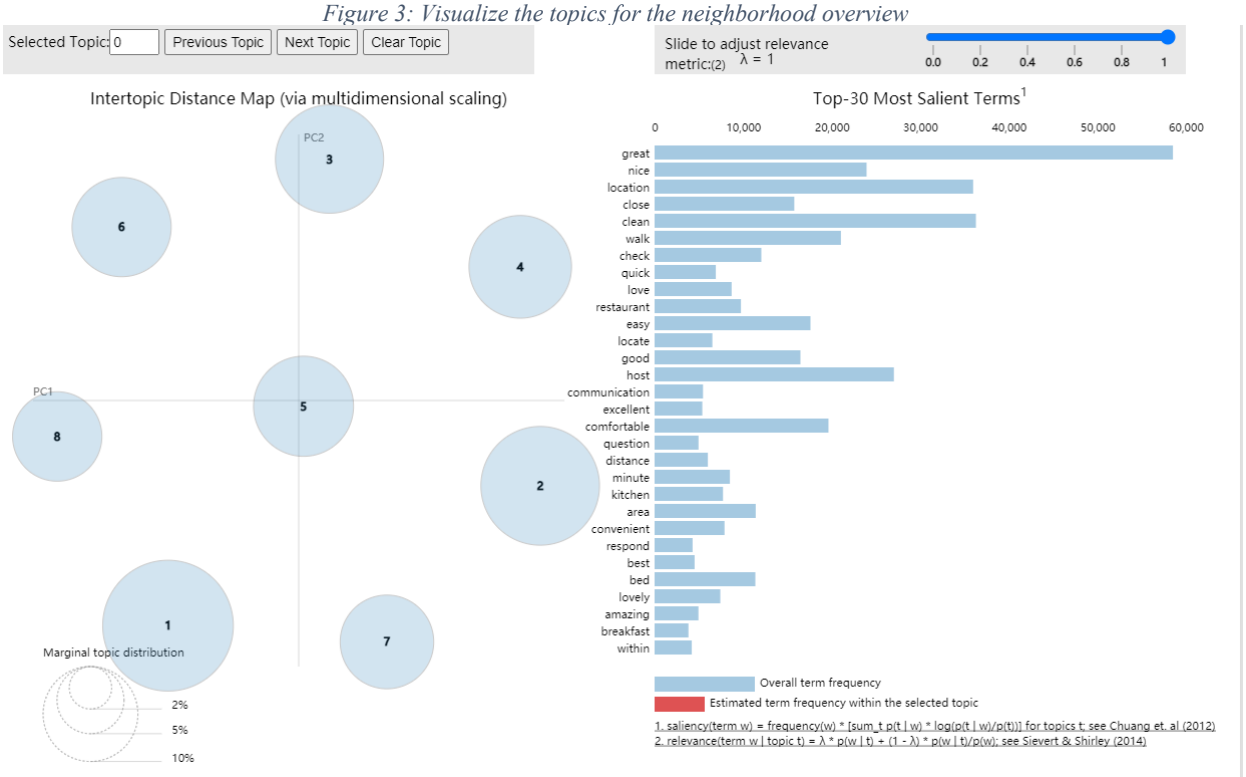


Figure 4: Visualize the topics for review description