MY459 - Quantitative Text Analysis

What Airbnb booking data tell us?

London School of Economics and Political Science

Candidate Number: 23267

Date: 05.04.2021

Abstract

Use the scraped Airbnb Beijing historical booking dataset to build the corpus. Then conducted exploratory data analysis (EDA) and using NLP methods like sentiment analysis and topic modelling to discover: 1) What are the specific aspects of the gap in Airbnb listings? 2) How is the pricy apartment titled? -How can hosts phrase their listing in the best way, and 3) What do visitors like and dislike? The text analysis including both Airbnb listing title (short sentence) and comments (long sentence).

For Airbnb listings title analysis, explored the most frequent wordings, bigrams, and trigrams in describing the expensive places in Beijing. Find out that the ideal listing's title including Fancy Location (CBD) + Local characters ("Hutong") + Cozy Environment (hazy hill view), in hopes to maximize profits.

For Airbnb listings comments analysis: classifies the overall sentiment of the text as Positive, Negative, Neutral, and calculate the compound score. Then comparing Negative and Positive Comments respectively.

Keywords: Airbnb historical booking data, NLP, Sentiment analysis, Topic modelling

1. Introduction

This project is focused on a homestay reservation company called Airbnb. Airbnb is a website for people to rent out their accommodation, offering a service of renting out houses or rooms on a short-term basis. The company has more than 3 million homes in 65,000 cities in 191 countries. Using the historic booking data and details about homes and customer behaviours to discover:

1. What are the specific aspects of the gap in Airbnb listings?
2. How is the pricy apartment titled? - How can hosts phrase their listing in the best way
3. What do visitors like and dislike?

The corpus is the scraped Airbnb Beijing historical booking dataset, we main focus on 1) the description of all listings; 2) comments of all Airbnb listings in Beijing before 21 December 2020. The row dataset was scraped from [About Inside Airbnb]-- http://insideairbnb.com/about.html using R. Moreover, due to the large dataset, several processes took hours to run. In these cases, I have saved the required objects into .csv file and reloaded them when required. If you want to reproduce it, the original code is included (see appendix Github link) but has been placed in docstrings so that it will not be executed.

For Airbnb listings title analysis, explored the most frequent wordings, bigrams, and trigrams in describing the expensive places in Beijing. Find out that the ideal listing's title, including Fancy Location (near CBD) + Local characters ("Hutong") + Cozy environment(hazy hill view), in hopes to maximize profits.

For Airbnb listings comments analysis: classifies the overall sentiment of the text as Positive, Negative, Neutral, and calculate the compound score -the sum of all of the lexicon ratings which have been standardized to range between -1 and 1.

Findings positive sentiment contains if the apartment is clean. The apartment contains local characters like provide “Hutong” style house or near some famous local attractions like the great wall. Moreover, The area is centrally located with short walking distances, good public transport connections, and has cafes and restaurants nearby.

The negative sentiment contains if the apartment dirty, The Location of the apartment is too noisy. Furthermore, The area is centrally not located short walking distances, without good public transport connections like the subway.

2. Motivation

In Beijing alone, there are over 13,768 reviews on Airbnb per season. Reviews are typically about areas of the stay that were most memorable for guests and provide much insight on what a listing actually has to offer. By understanding what guests are saying about different listings and finding trends in the way people review, we can better understand the relationship between guest and host.

Moreover, in China, many guests choose Airbnb to write reviews in Chinese, so as the hosts willing to choose Chinese to describe their house. Nevertheless, more than 60% of the description and reviews are presented in English through the data collected. Thus I am curious about such types of comments/descriptions written in English in China and want to analyze this part of the text.

The goal is to use a series of text analysis methods to find out what reviews are talking about and then develop insightful improvement recommendations.

3. Construction of Corpus – Airbnb reviews

3.1 Data collection

Data collection: The datasets mainly use from [About Inside Airbnb]--http://insideairbnb.com/about.html.

Then Extracting, Transforming, and Loading data to a relational schema.

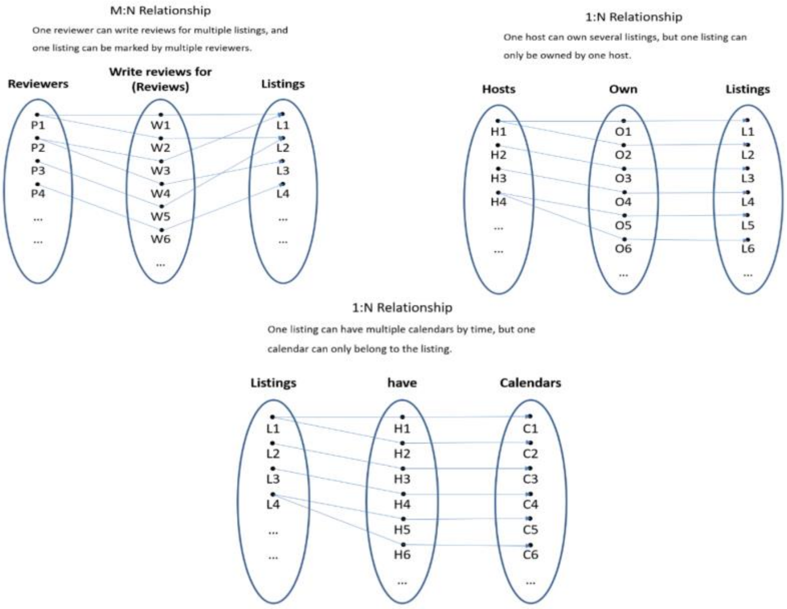

M:N Relationship
One reviewer can write reviews for multiple listings, and one listing can be marked by multiple reviewers.
Reviewers
Write reviews for (Reviews)
Listings
P1
P2
P3
P4
W1
W2
W3
W4
W5
W6
L1
L2
L3
L4
1:N Relationship
One host can own several listings, but one listing can only be owned by one host.
Hosts
Own
Listings
H1
H2
H3
H4
O1
O2
O3
O4
O5
O6
L1
L2
L3
L4
L5
L6
1:N Relationship
One listing can have multiple calendars by time, but one calendar can only belong to the listing.
Listings
have
Calendars
L1
L2
L3
L4
H1
H2
H3
H4
H5
H6
C1
C2
C3
C4
C5
C6

Figure 1 Relational Schemas

3.2 Data description

Description of Data: The dataset comprises three main tables:

* `listings` - Detailed listings data showing 106 atttributes for each of the listings. Some of the attributes used in the analysis are `price` (continuous), `longitude` (continuous), `latitude` (continuous), `room_type` (categorical), `is_superhost` (categorical), `neighbourhood_cleanse` (categorical), `bedrooms` (categorical) among others.

* `reviews` - Detailed reviews given by the guests with 6 attributes. Key attributes include `date` (datetime), `listing_id` (discrete), `reviewer_id` (discrete) and `comment` (textual).

* `calendar` - Provides details about booking for the next year by listing. Seven attributes in total including `listing_id` (discrete), `date` (datetime), `available` (categorical) and `price` (continuous).

3.3 Data pre-processing

The data quality of the data is not perfect. There are more than half missing data in columns'square_feet', 'weekly_price,' 'monthly_price,' 'security_deposit,' 'cleaning_fee' and about 40% of the data in the feature about reviews. We had to drop this column. For other columns that are relatively complete, we need to perform a few imputations and transformations on our dataset to create the desired visualizations. Some of the columns in the file 'Listings' miss many data. Some of them contain outliers, and most of the columns/features we were interested in did not contain data in the required format and hence were manipulated so that their meanings are retained.

Key Feature Engineering:

1. ```comment (reviews)```: We extensively used this feature in our analysis. The dataset contained reviews in multiple languages such as Chinese, Spanish, and English which made it difficult for it to be analyzed. We subsetted the data to include only the reviews in English and Chinese and performed text filtering to remove common stop words and phrases that do not significantly contribute to the meaning of the review. And then conduct word cloud to discover what kinds of features or services about the property matter to customers.

2.``` price,extra_people(listings, calendar)```: The price column contained data in string format with the currency symbol '$' and comma separator ',' attached to it. This column was manipulated to contain integer values for analysis. Moreover, the feature extra_people needs to be transformed in the same way. Moreover, there are the targeted variables of the predictive model.

3. ``` date (calendar, listings, reviews)```: The date was contained in mm-dd-yyyy format. They are displayed as character, so they need to be transferred to date format. Furthermore, it was transformed multiple times during the analysis to obtain weekly, monthly or yearly insights.

Data cleaning:

Including drop unuseful variables, check missing values, deal with single variables, Draw histogram and boxplot for the cleaned numeric data to view their distribution and outliers to deal with outliers.

In the beginning, in order to discover the relation between variables, draw histograms and boxplots for every numerical variable. Found that the distributions of several variables, including "host_listings_count"and "price, " are positively skewed, consistent with common sense because most rental housing is cost-effective, while some luxury houses are well above the average market price also exist.

3.4 Summary statistics about the data:

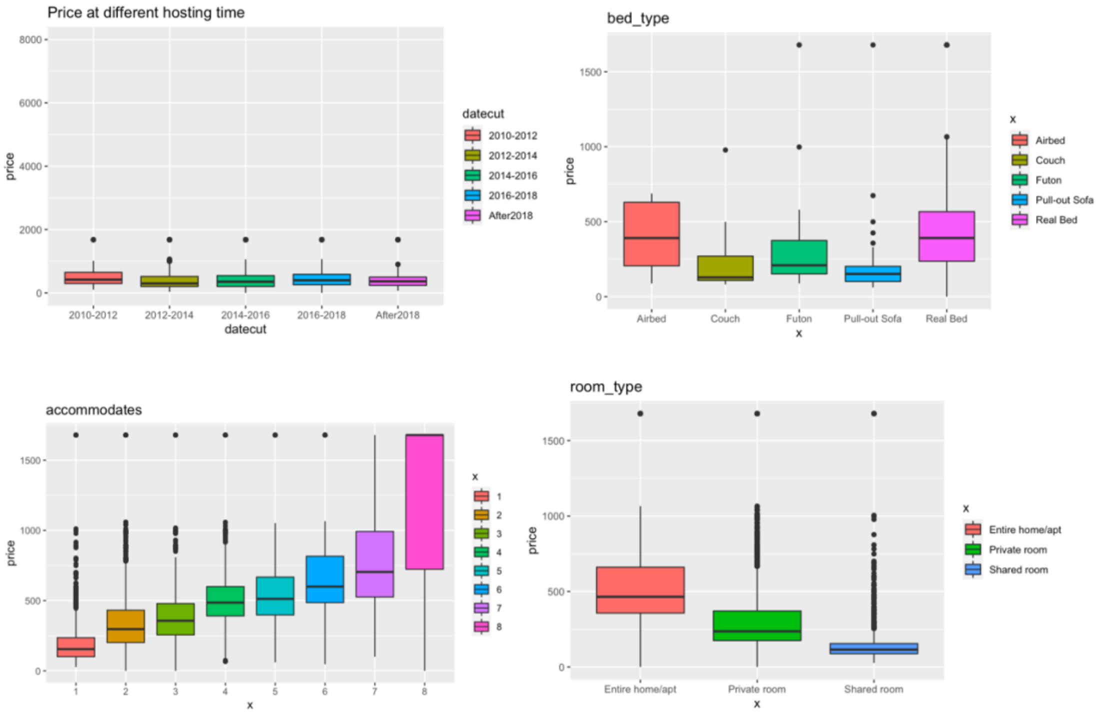


Figure 2: Relation between variables

Then tried to logarithm these features, while the result of the linear regression model is even worse, so we keep them in original format. Through the boxplot, we found some columns that contain outliers, which may affect the accuracy of the model. So we use the highest and lowest 5% values to replace the outliers. Furthermore, we can see from the boxplot that the transformed data is more centralized. Some special variables need to be deal with individually.

3.4 Review Comment:

The comment for each Airbnb listing is accessible through a given URL link through web-scraping. This raw data will be used to conduct exploratory and sentiment analysis, introduced in the data analysis section.

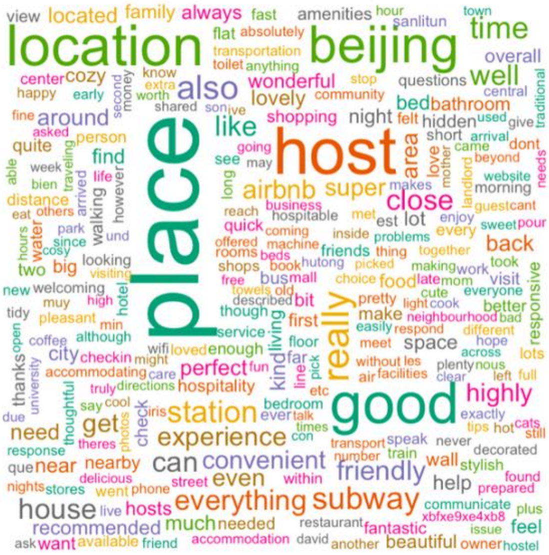


Figure 3: Word Cloud of Airbnb comment

4. Data Analysis Methodology

Natural Language Processing is usually the most automated form of text analysis (Manning et al., 2008). This method simulates how humans understand and process language (Chowdhury, 2003; Collobert et al., 2011; Joshi, 1991). For example, NLP technology can mark the parts of speech of words in sentences (e.g., nouns, adjectives, etc.), translate documents from one language to another, and even use sentence context to clarify the meaning of words (Buntine &and Jakulin, 2004).

Therefore, NLP considers the order of words to be necessary. Using training sets, sentiment analysis using cutting-edge techniques such as deep learning and multi-modality (combining text and images) is a popular form of NLP (Kouloumpis, Wilson and Moore, 2011). This particular analysis classifies the overall attitude, sentiment, or opinion of the text as positive, negative, or neutral.

This paper produces four sentiment metrics from these word ratings, positive, neutral, negative, and compound. The first three - positive, neutral and negative - represent the proportion of the text that falls into those categories. The final metric, the compound score, is the sum of all of the lexicon ratings standardized to range between -1 and 1. For example, if a sentence has a rating of 0.5, which is considered pretty neutral.

This paper aims to calculate the Sentiment Scores for only the English language. First, We got the negative sentiment score, neutral sentiment score, positive sentiment, and compound. Then discuss the overall sentiment in all Beijing Airbnb reviews, comparing negative and positive comments. Further, discover what the negative and positive comments are about. Finally, compare the length of both positive and negative comments:

Topic Modeling on Airbnb Comments [each document contains all sentences about host per each listing per year]. The model is used to extract trends of topics over time.

After document preparation, we can execute the topic modelling to find optimal topics. However, the number of topics (K) first needs to be decided.

5. Results:

5.1 Exploratory data analysis

Try to get ideas on the summarized data grouped by each district on the map. Consider drawing the map of average price, average rating and house counts in each district, and get the figure as follows:

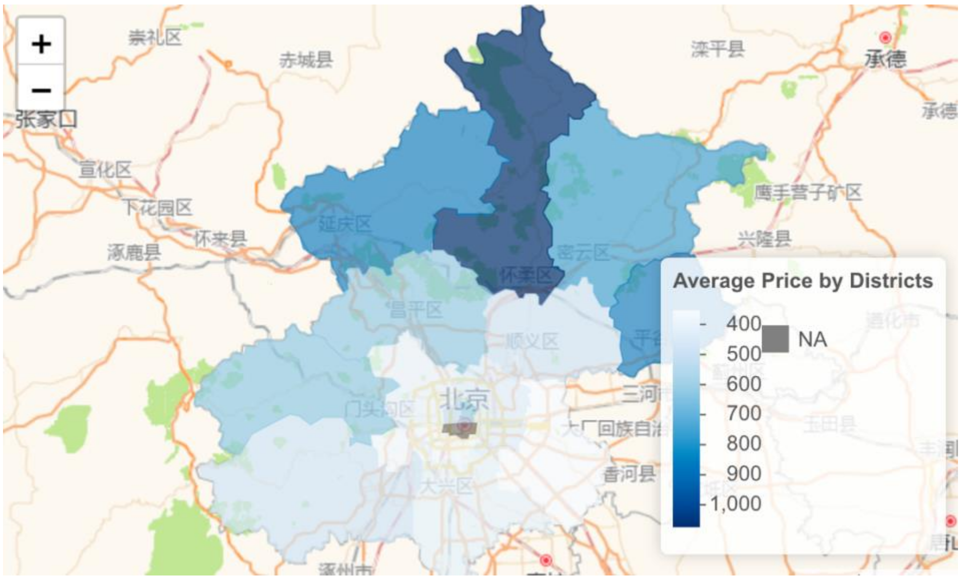


Figure 4: Distribution of Average Price by Districts

From the figure above, we find that the average price is lower in the city center and higher in the suburb. We can also get that the average rating is approximately correlated with the average price of the house rented. We see higher average prices in the suburbs of Beijing, located in the east and north Beijing, and an obvious higher rating among them than the center of the city. In the last figure above, we can find a significant difference in the number of houses listed on Airbnb between the center of the city and suburbs. The number of houses listed in the city center is many times that in other districts.

5.2 Sentiment analysis

This part aims to analyze short sentences. The 'Title' feature is used to describe the Airbnb property in a short sentence. To discover how the householders can describe their house better, I decided to use natural language processing (NLP) to break out the Titles of the Listings into n-grams. At first, I will use grams and bigrams.

Then merging the top 20 most frequently used bigrams and the top 40 most frequently used grams to my dataset. This way can help measure the relationship of title language on the net revenue. Thus, allowing hosts to phrase their listing in the best way possible in hopes of maximizing profits!

Most frequent wordings for expensive places

The 'title' feature is used to describe the Airbnb property in a short sentence. Let us explore the titles of the apartments that cost more than 1,000 for a night.

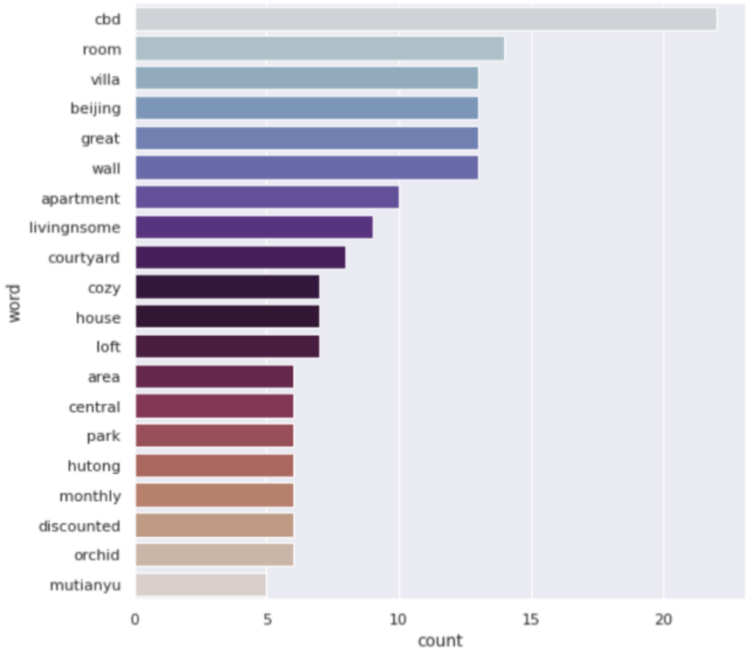


Figure 5: Top 20 words in Beijing pricy apartments' title

It can be seen that the top 20 words for expensive places of Beijing show a similar feature. Location and environment are the most important word, and many words are used to describe the price of the apartment. That means a majority of customers concern about the Location and the actual living quality.

Furthermore, the "recommend" represents that these neighbourhoods are popular and customers' needs are satisfied well. There also are some words in the top 20 like "discouted," "park," "area," showing the importance of transportation and Location.

Taking the "Hutong" and "Great wall" into account, it can be seen that the unique and local characteristics of the apartment are also essential factors in operation.

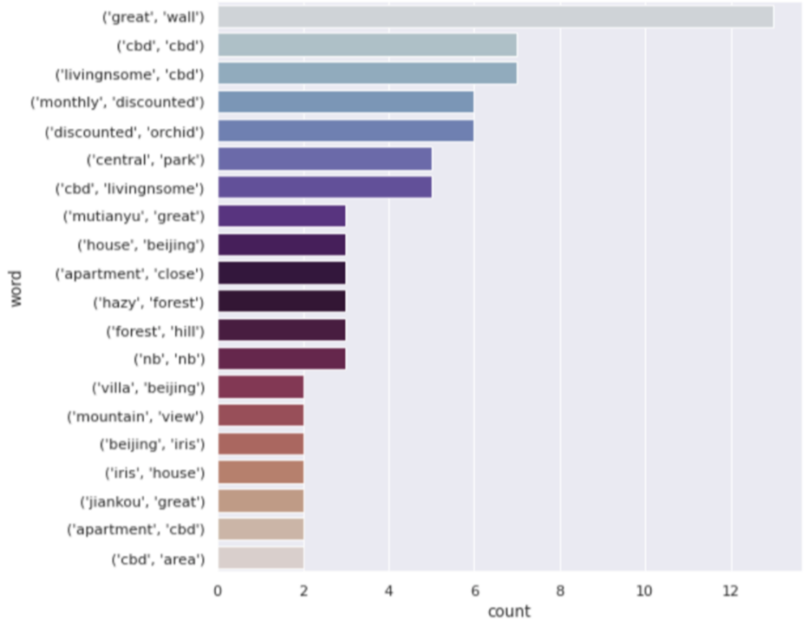


Figure 6: Top 20 bigrams in Beijing pricy apartments' title

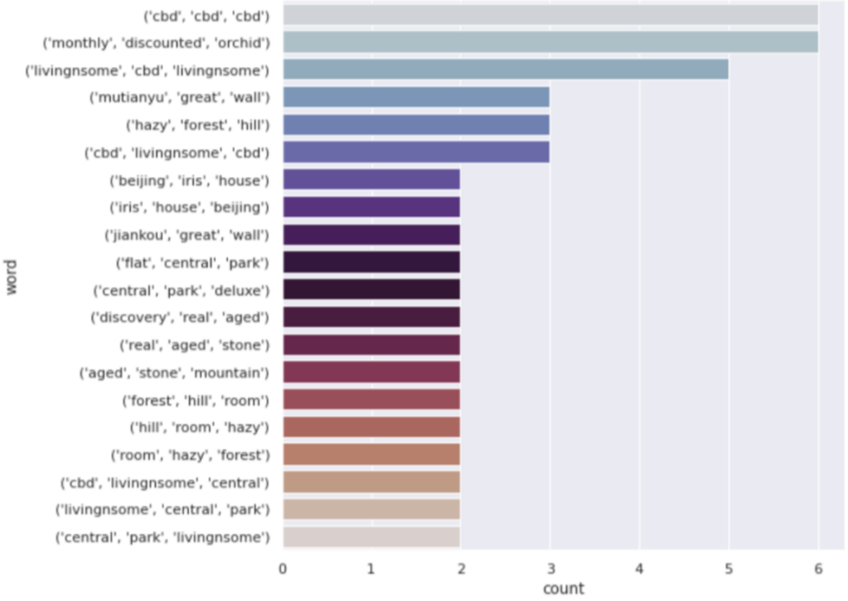


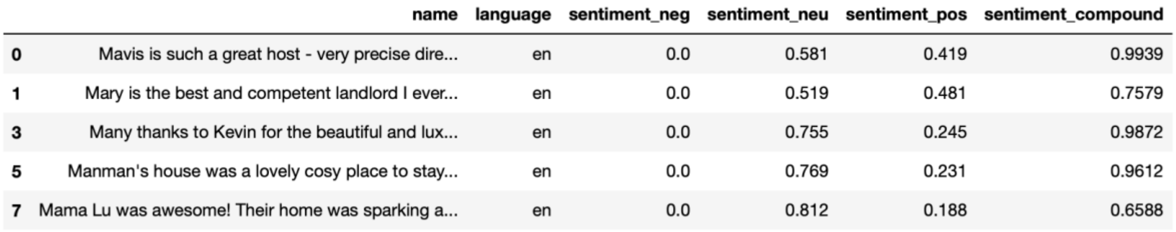
Figure 7: Top 20 trigrams in Beijing pricy apartments' title

We can see from the plot that such bigrams look much more specific for Beijing than single wording; by looking at the plots above, it is easy to identify some pattern. We encounter some terms that were not present in the frequently used words when we were looking at all the data.

Titles for apartments with high-end prices uses such words as:

- CBD
- Great Wall
- Monthly discount
- Hutong
- Hazy forest hill
- Central Park

The results seem in line with common sense and expectation. You do expect an apartment with a hazy hill view or near the Great wall or even just live in a Hutong style house (even in China, there is only in Beijing, that have "Hutong" style of housing) to be more expensive than a median price tag. The same for the term "CBD."



	name	language	sentiment_neg	sentiment_neu	sentiment_pos	sentiment_compound
0	Mavis is such a great host - very precise dire...	en	0.0	0.581	0.419	0.9939
1	Mary is the best and competent landlord I ever...	en	0.0	0.519	0.481	0.7579
3	Many thanks to Kevin for the beautiful and lux...	en	0.0	0.755	0.245	0.9872
5	Manman's house was a lovely cosy place to stay...	en	0.0	0.769	0.231	0.9612
7	Mama Lu was awesome! Their home was sparking a...	en	0.0	0.812	0.188	0.6588

As introduced before, this method divided four sentiment metrics from these word ratings, positive, neutral, negative and compound. The compound score is the sum of all lexicon ratings, which have been standardized to range between -1 and 1.

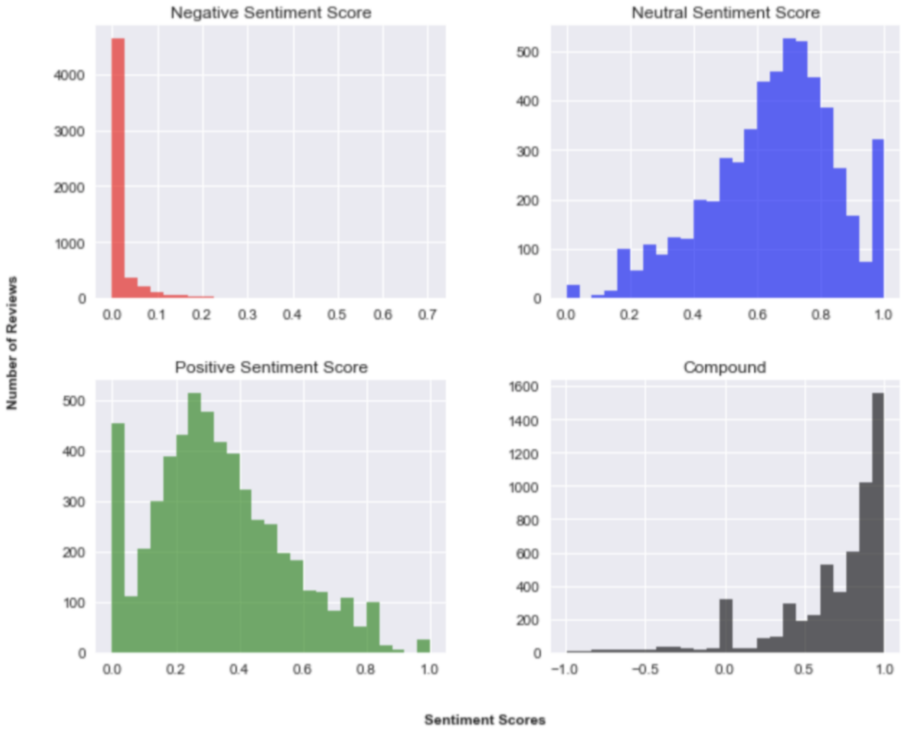


Figure 8: Sentiment Analysis of Airbnb Reviews for Beijing

Comparing Negative and Positive Comments

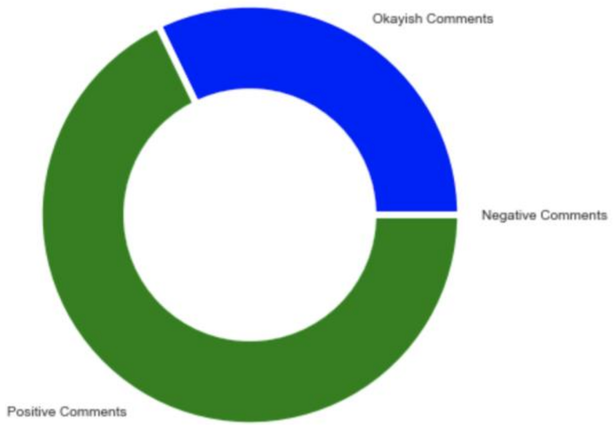


Figure 9: Sentiment proportion of Airbnb Reviews for Beijing

The bulk of the reviews are tremendously positive. Let us see what the negative and positive comments are about.

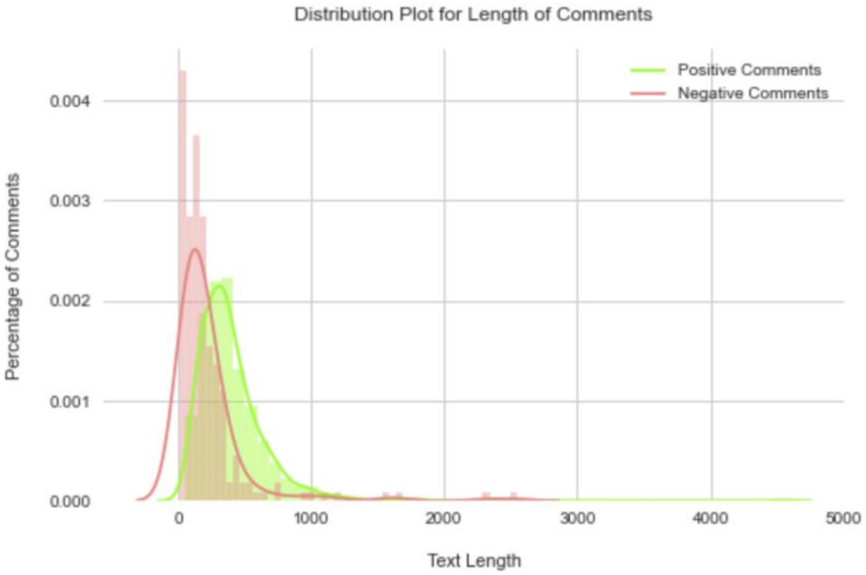


Figure 9: Distribution Plot for Length of Comments

The text length of negative comments can be found more to the left than for the positive comments, which means most of the negative are shorter than most of the positive comments. However, the tail for negative comments is thicker.

Frequency Distribution

Another method for visually exploring text is with frequency distributions.

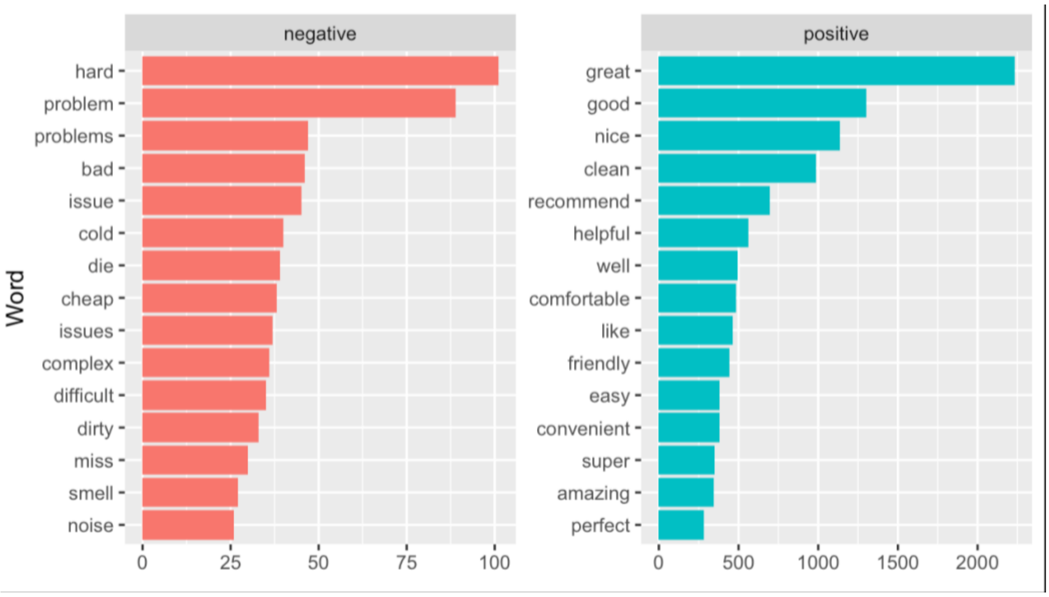


Figure 10: Most Frequently Used Words

Consider a new host who plans to rent his place to Airbnb; what are the facilities and things the host needs to provide according to the Airbnb reviews around Beijing. Moreover, what are the things people do not like and better avoid them as a new host? First, we create a Gensim dictionary from the normalized data, then we convert this to a bag-of-words corpus and save both dictionary and corpus for future use.

Instigate positive comments:

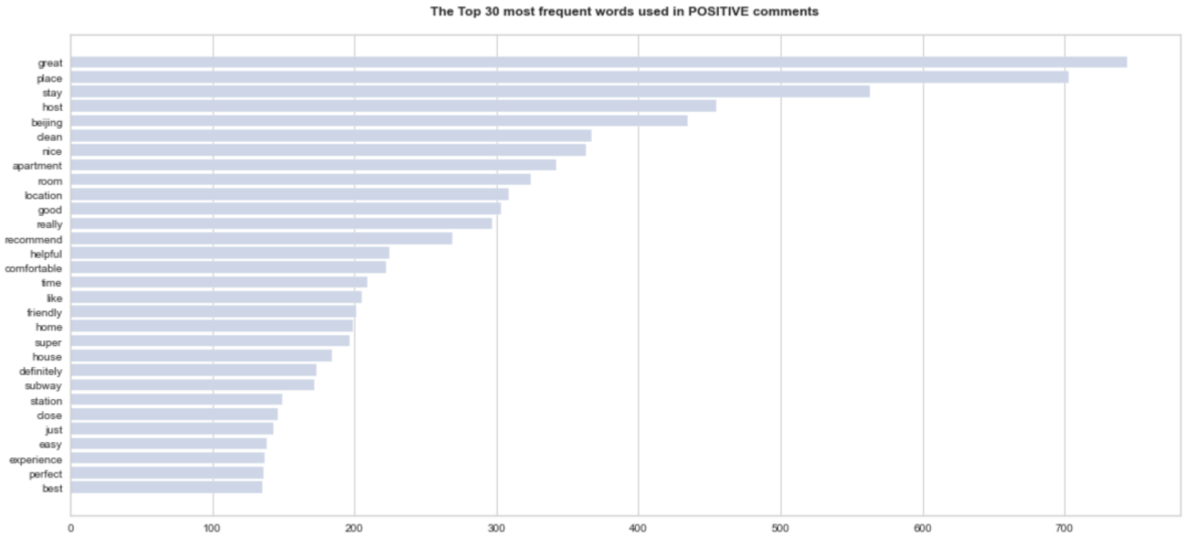
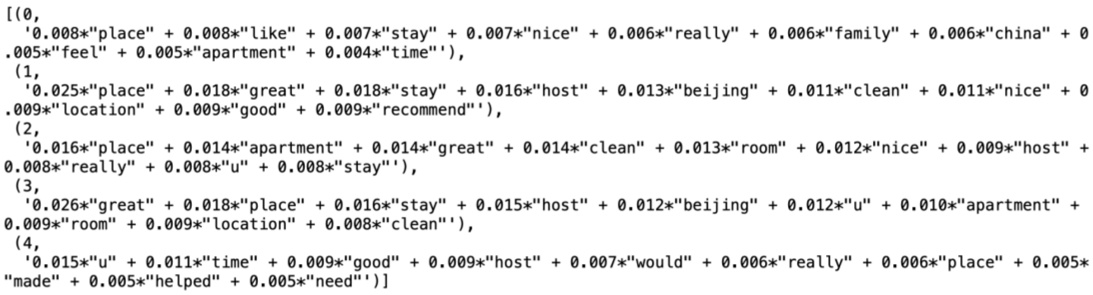


Figure 11: The Top 30 most frequent words used in positive comments



```
[(0,
  '0.008*"place" + 0.008*"like" + 0.007*"stay" + 0.007*"nice" + 0.006*"really" + 0.006*"family" + 0.006*"china" + 0
.005*"feel" + 0.005*"apartment" + 0.004*"time"'),
 (1,
  '0.025*"place" + 0.018*"great" + 0.018*"stay" + 0.016*"host" + 0.013*"beijing" + 0.011*"clean" + 0.011*"nice" + 0
.009*"location" + 0.009*"good" + 0.009*"recommend"'),
 (2,
  '0.016*"place" + 0.014*"apartment" + 0.014*"great" + 0.014*"clean" + 0.013*"room" + 0.012*"nice" + 0.009*"host" +
0.008*"really" + 0.008*"u" + 0.008*"stay"'),
 (3,
  '0.026*"great" + 0.018*"place" + 0.016*"stay" + 0.015*"host" + 0.012*"beijing" + 0.012*"u" + 0.010*"apartment" +
0.009*"room" + 0.009*"location" + 0.008*"clean"'),
 (4,
  '0.015*"u" + 0.011*"time" + 0.009*"good" + 0.009*"host" + 0.007*"would" + 0.006*"really" + 0.006*"place" + 0.005*
"made" + 0.005*"helped" + 0.005*"need"')]
```

The first topic includes words like place, clean, host, Location. This sounds like the topic related to the place as a nice and a good host, and they would recommend that pace.

The second topic includes words like room, host, apartment, clean. This sounds like a topic related to convenient distances from the accommodation to wherever something interesting was to go to and the environment of the apartment.

The third topic includes words like Location, place. This sounds like the topic related to the transport facility from the place they stay. Roughly "one minute walk to bus or train station."

The fourth topic seems like a topic about describing the place. Like how many apartments and the room, etc.

The fifth topic telling about the tidiness of the place and also the "helped" they had and "need" as well as "great host.

In summary, what the guests are mostly talking about their stay:

- Topic 1: Location
- Topic 2: Overall - Stay, Location, place, comfort, space, host
- Topic 3: space and comfort
- Topic 4: Particularly about the room and stay
- Topic 5: Tidiness and service

Investigating Negative Comments

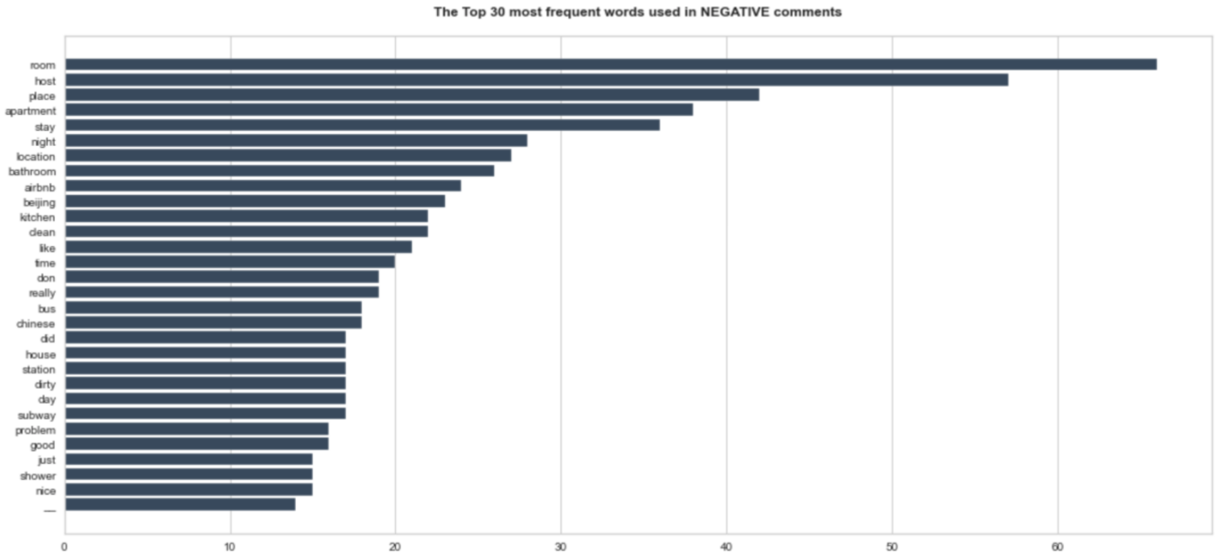
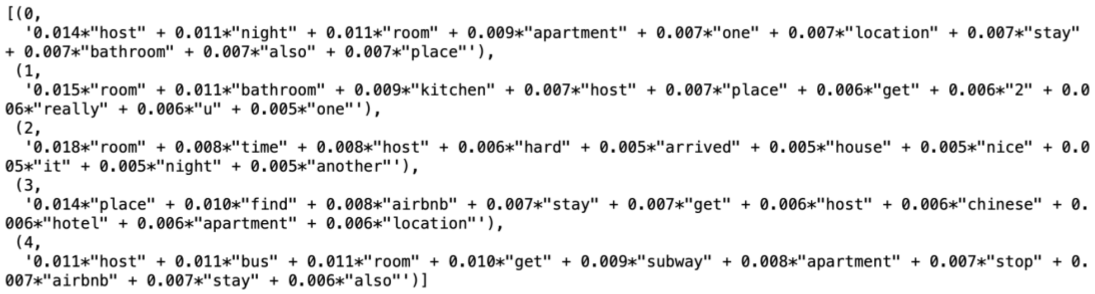


Figure 12: The Top 30 most frequent words used in negative comments

Topic modelling for negative reviews:



```
[(0,
  '0.014*"host" + 0.011*"night" + 0.011*"room" + 0.009*"apartment" + 0.007*"one" + 0.007*"location" + 0.007*"stay"
+ 0.007*"bathroom" + 0.007*"also" + 0.007*"place"'),
 (1,
  '0.015*"room" + 0.011*"bathroom" + 0.009*"kitchen" + 0.007*"host" + 0.007*"place" + 0.006*"get" + 0.006*"2" + 0.0
06*"really" + 0.006*"u" + 0.005*"one"'),
 (2,
  '0.018*"room" + 0.008*"time" + 0.008*"host" + 0.006*"hard" + 0.005*"arrived" + 0.005*"house" + 0.005*"nice" + 0.0
05*"it" + 0.005*"night" + 0.005*"another"'),
 (3,
  '0.014*"place" + 0.010*"find" + 0.008*"airbnb" + 0.007*"stay" + 0.007*"get" + 0.006*"host" + 0.006*"chinese" + 0.
006*"hotel" + 0.006*"apartment" + 0.006*"location"'),
 (4,
  '0.011*"host" + 0.011*"bus" + 0.011*"room" + 0.010*"get" + 0.009*"subway" + 0.008*"apartment" + 0.007*"stop" + 0.
007*"airbnb" + 0.007*"stay" + 0.006*"also"')]
```

Some of the topics we observed:

1. Dirty Bathrooms, Bad towels and a lousy kitchen.
2. The Location of the apartment is too noisy.
3. Not enough space and the damaged items, not functioning amenities in the room.
4. Provided only one key for the flat.
5. The area is centrally not located with short walking distances, good public transport connections.

6. Conclusion

1. What are the specific aspects of the gap in Airbnb listings?

Location: China has a clear concept of rural and urban areas, as well as regional divisions. The average prices of all kinds of property are generally higher in the suburb. In urban areas, the Chaoyang district contains the most significant part of the property, much larger than other districts. The complete apartment-style listings are located in Chaoyang district, constituting more than 50% of all properties in that neighbourhood. Haidian district is in second place. That might relate to the Location itself - they are all economically developed regions. So that has the most abundant listings.

Property types: There is a significant difference in composing properties between the two kinds of Chaoyang district and Haidian district. In the center of the city, we can find that apartments are much more than other types of properties, at the same time, a whole set of houses and loft houses are provided with more ratio in the suburb areas. Consider the price elements. Houses and loft properties are higher than other types.

Possible reason: The counts are high, and the average price is lower in the center of the city because there are more needs there, so the property is much higher, and the property comes with more standard service and facilities. Therefore, they can provide a more affordable price for most people, which is just why the rating can be lower. However, there are more minor needs in the suburb, and people are more likely to rent a house for their vacation to provide more convenient services; at the same time, higher prices and higher ratings come due to that.

2. How is the pricy apartment titled? -How can hosts phrase their listing in the best way

Before discovering what owners usually use words to describe their places, we discuss the most frequent trigrams to describe Beijing apartments in advance, find a very interesting trigram indeed – “Huaxia Sunshine serviced"- the name of a specific and popular Beijing hotel-style name.

Then merging the top 20 most frequently used bigrams and the top 40 most frequently to measure the relationship of title language on the net revenue.

In this section, we have identified exciting patterns.

- It is possible to guess the price range of an Airbnb apartment by simply looking at its title.
- More expensive places use such words as CBD, Villa, Hutong, livingsome, courtyard, cozy, orchid.
- Flats which titles indicate the vicinity of city attractions also tend to cost more.

It looks like it is essential what to write in the title of the Airbnb place after all. Thus, hosts should carefully and creatively phrase their listing in the best way in hopes of maximizing profits!

3. What do visitors like and dislike?

Putting the WordCloud, the Frequency Distribution and the Topic Modelling all together- it is often the following criteria that make someone rate an apartment positively:

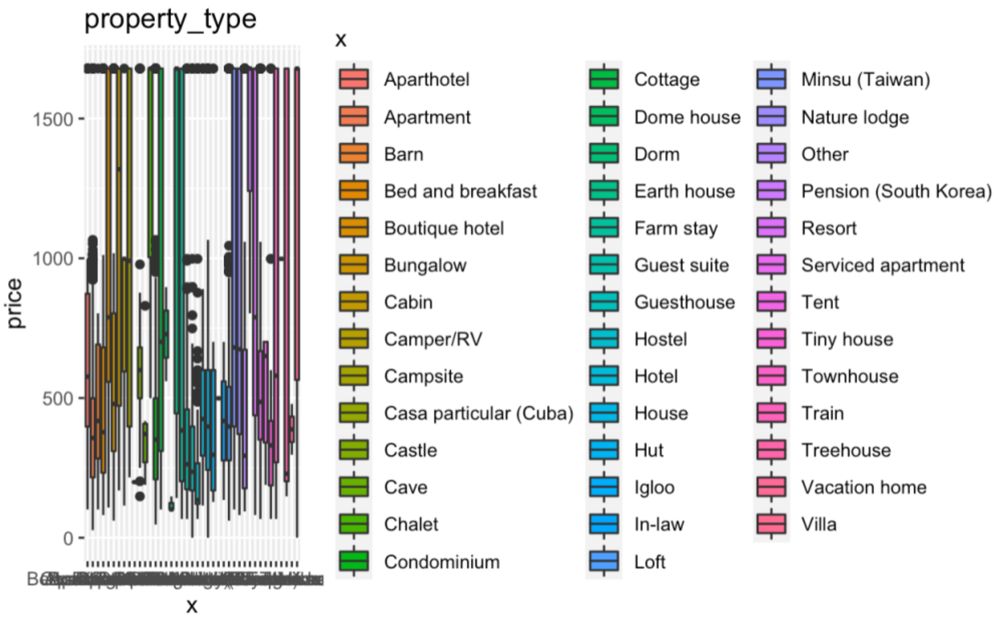
1. The apartment is clean; the bathroom is clean, the bed is comfortable.
2. The apartment contains local characters like provide “Hutong” style house or near some famous local attractions like the great wall.
3. The area is centrally located with short walking distances, good public transport connections, and has cafes and restaurants nearby.
4. The check-in formalities and the process of booking should be straightforward.

Reviewers care about four main aspects when they rate: the Location, view, comfort(feel like home) and interior.

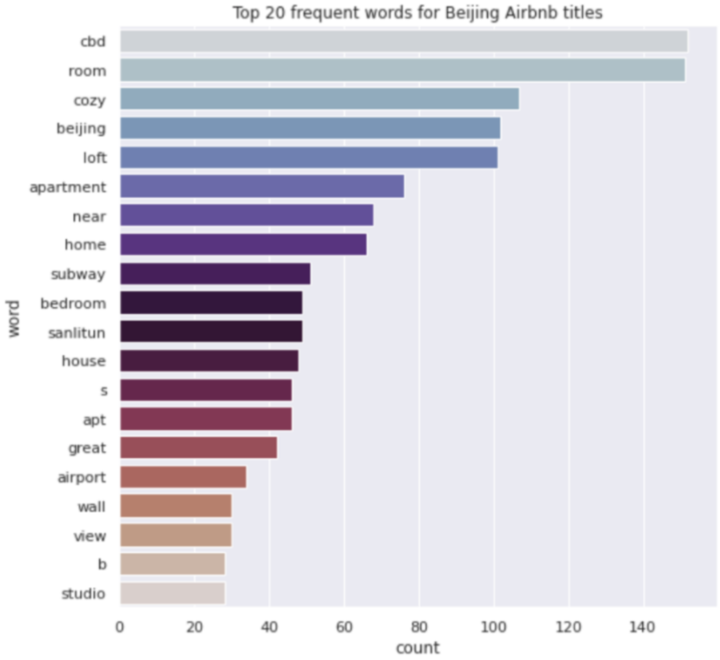
On the contrary, guests will complain in particular:

5. Dirty Bathrooms, Bad towels, and bad kitchen.
6. The Location of the apartment is too noisy.
7. Not enough space and the damaged items, not functioning amenities in the room.
8. Provided only one key for the flat.
9. The area is centrally not located short walking distances, without good public transport connections like the subway.

Appendix:



Owners usually use words to describe their places.



Most frequent bigrams to describe Beijing apartments.

Most of the top 20 words obtained in the previous section are not specific to Beijing. Now I will find top bigrams, i.e. the sequences of two neighbouring words, used in Airbnb titles for Beijing flats.

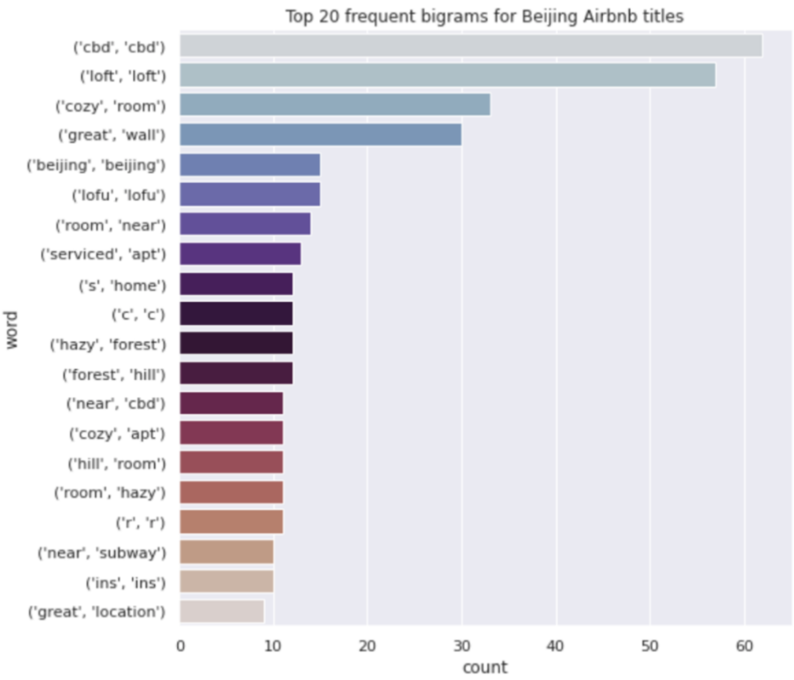


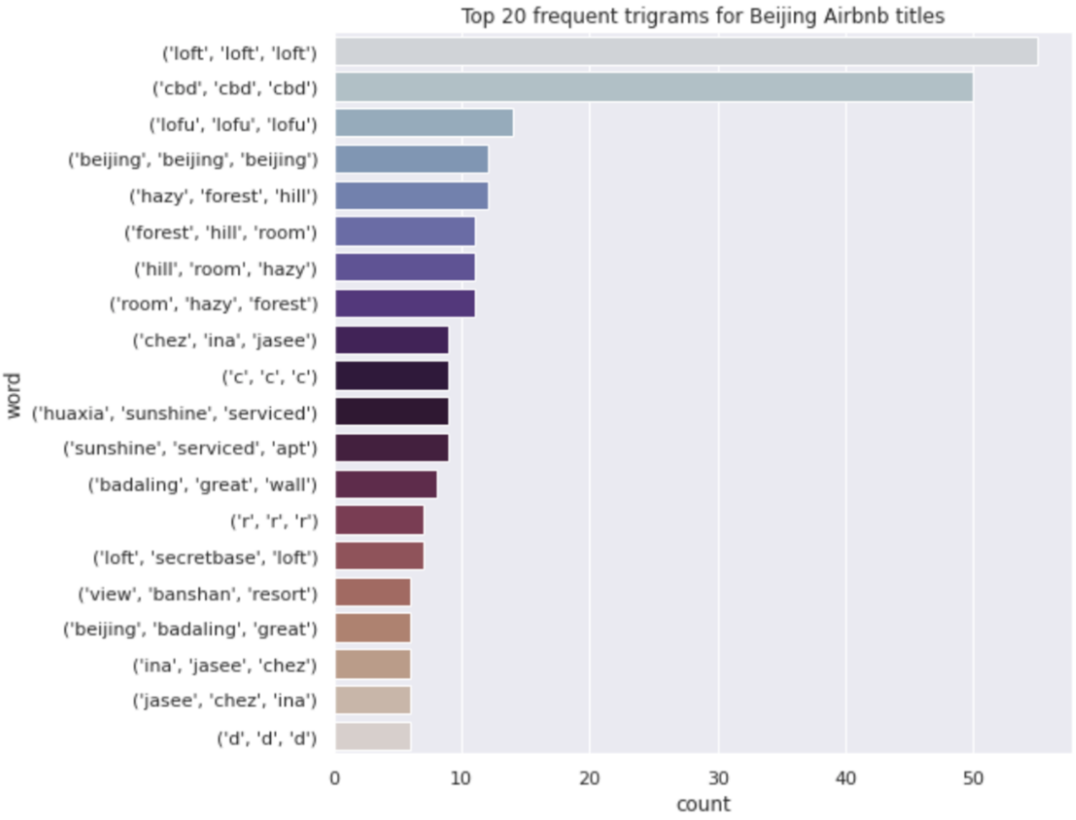
Figure: Most frequent bigrams to describe Beijing apartments.

We can see from the plot that such bigrams look much more specific for Beijing

- Cozy zoom
- Great wall
- Hazy forest
- forest hill
- Near CBD

Most frequent trigrams to describe Beijing apartments.

As long as we are here and looking at n-grams used in Airbnb titles, let us create a plot for top trigrams. We find a very interesting trigram indeed – "Huaxia Sunshine serviced"- the name of a familiar and popular Beijing hotel-style name.



Links to the repository that contain the codes and dataset required to replicate this research: https://github.com/lllllxnnnnn/MY459-FinalProject