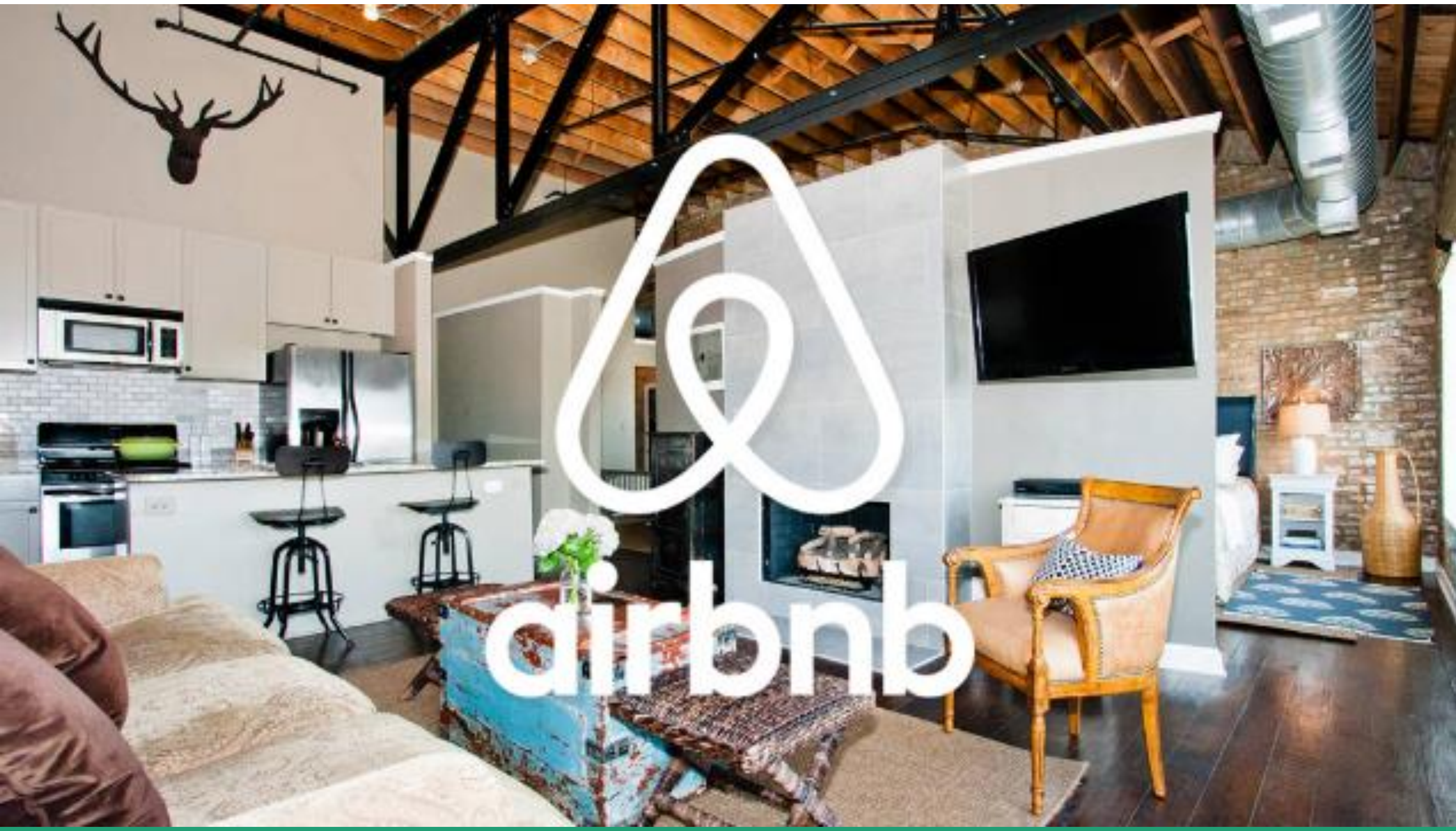
SMU
SINGAPORE MANAGEMENT UNIVERSITY



IS450- Text Mining and Language Processing

Project Proposal

Submitted by:

Ahmad Saifullah Bin Mustaffa, Gab Min Xuan Darren, Lu Yun Jing Ada, Muskaan Gupta, Shubham Periwal, Vikram Sanghi

Submitted to:

Professor Swapna Gottipati

A. Introduction

Airbnb is a popular platform for travellers to book their accommodation when they go overseas. All the listing information is provided exclusively by the Airbnb host, with limited ways for the guest to confirm its accuracy. Currently, Airbnb has the following aspects for reviews, along with the overall rating on a scale from 1 to 5 stars:

- Accuracy of listing
- Communication with hosts
- Cleanliness
- Location
- Check-in
- Value

These metrics, while useful, are relatively general and are only shown on an overall scale. Customer review data (in text format) contains hidden information that cannot be shown on a ranking scale. The aim of our project is to use language processing techniques to extract relevant and useful information from the unstructured textual data and to analyse it to aid decision making for users. This analysis, in turn, will also help hotel managers to make better managerial decisions.

Our project can be accessed on: https://github.com/dagazzg/IS450-project

B. Data Set

The data is sourced from the Inside Airbnb website: http://insideairbnb.com/get-the-data.html

The dataset to be used for this project is obtained from Inside Airbnb, an independent website that scrapes publicly-available listing information from Airbnb and cleans and aggregates the data.

For the project, we will mainly be using the following files, focusing on Los Angeles, USA as our primary location.

File	Description
listings (unzipped).csv	Detailed “Listings” data for listings. Some of the attributes used in our analysis are “price” (continuous), “longitude” and “latitude” (continuous), “room_type” (categorical), “neighbourhood” (categorical), “host_is_superhost” (categorical), “review_score_rating” (continuous), etc.
reviews (unzipped).csv	Detailed “Reviews” data on listings with 6 attributes, contains the raw test comments which will be used extensively
neighbourhood.csv	A list of all the neighbourhoods in LA

C. Features & Methodology

Feature #1: Aspect-based search for listings / Ranking top listings according to aspects

1) **Purpose:** We will perform an exploratory analysis of the Airbnb dataset to understand the rental landscape in LA through various static and interactive visualisations.
2) **Methodology:** Exploratory Data Analysis and Visualization of Airbnb Dataset
   a) How do prices of listings vary across the location?
   b) How does the demand for Airbnb rentals fluctuate across the year and over years?
      i) "price" (listings.csv) will be manipulated for time-series analysis
      ii) "date" (listings.csv, reviews.csv) will be transformed to obtain weekly, monthly or yearly insights
   c) Are the demand and prices of the rentals correlated?
   d) What are the different types of properties in LA? Do they vary by neighbourhood?
   e) What localities in LA are rated highly by guests?
   f) What makes a host Super host? **(Appendix: Exhibit A)**
      i) Airbnb has a set of requirements that must be fulfilled in order to become a super host. Maintaining a review rate above 50%, a response rate above 90%, etc. here we investigate our dataset to see how the super hosts perform on two parameters: "Response Rate" and "Ratings"
      ii) listings.csv contains "location rating", "cleanliness rating" and "overall rating" which will be used for analysing reviews **(Appendix: Exhibit B)**
   g) Do regular hosts and super hosts have different cancellation and booking policies?
   h) Common themes that can be identified from the free-text section of the reviews?
   i) What aspects of the rental experience do people like and what aspects do they abhor?

Feature #2: Summary of Listings- Text Analysis

1) **Purpose:** To provide summarized results for each listing showing **strengths**, **weaknesses** and the **change in reviews sentiment** across time for a listing.
2) **Methodology:**
   a) Data pre-processing (Tokenisation, stopword removal, lemmatisation)- we will use "comments" from the reviews.csv file.
   b) Extracting keywords to identify commonly-mentioned items (bag of words, feature selection)
   c) Use of bigram model to find words that frequently occur together in reviews **(Appendix: Exhibit E)**
   d) Using either Gensim or Scikit to train a topic model to identify a number of topics using Latent Dirichlet Allocation (LDA), a probabilistic model **(Appendix: Exhibit C)**
      i) Identifying the keywords of each topic and their weights
      ii) Sentiment Analysis:
         - For each listing clean the reviews
         - Provide a sentiment score to each review and add it as a new variable for each row
         - Provide access to each listing for visualizaing how sentiment score has changed over time
         - Identify the most common topics mentioned across reviews, use the topmost associated keywords of those topics to create a summary of the Airbnb listing that showcases its' most common strength and shortcomings (areas to improve) so that users can easily get information on the listing based on the opinions of others who have stayed there before.

Feature #3.1: Topic Analysis

1) **Purpose**: Analyse what features users look for in a hotel and what they comment about. **This will help hosts in understanding what travellers are looking for and thus, make better decisions for them.**
2) **Methodology:** Pre-define certain categories (bed, service, etc)
    a) Extract topics and clean the data for each review
        i) Remove stopwords
        ii) Lemmatise the data
        iii) Extract important features
        iv) Store sentiment along with the features
    b) Create a frequency diagram of the number of references to each of the pre-defined categories
    c) For visualisation, we can create an n-gram word cloud **(Appendix: Exhibit D)**
    d) Aggregate all topics and sentiments to the hotel level

Feature #3.2: Cluster Analysis

1) **Purpose:** Clustering reviews based on certain traits i.e. area with ‘good’ food. **This will help travellers choose their lodging based on what is important to them.**
2) **Methodology:**
    a) Potential clusters
        i) Cleanliness (“clean”, “dirty”, “smell”, “hygiene” etc.)
        ii) Bedding (“bed”, “pillows”, “mattress” etc.)
        iii) Kitchen amenities (“kitchen”, “utensils”, “cook-”, “grease” etc.)
    b) Load all reviews into a single corpus
    c) Pre-process and clean (remove stop-words, lower-case, stemming etc.)
    d) Transform corpus into a vector
    e) Carry out Hierarchical Clustering
        i) Gather insights by identifying the most talked-about issue amongst travellers
        ii) Could be useful information to hosts

Feature #4: Chat Box

1) **Purpose:** To make it easier for our users to specify their requirements instead of using filters
2) **Methodology**
    a) Create a greetings dictionary so bot can at least reply to greetings
    b) Create word vectors with Spacy
    c) Entity Recognition and intent recognition
    d) Use Rasa to create an interpreter model
    e) Connect with SQL
    f) Extract query and respond with answer
    g) Incremental slot filling and negation - if user does not accept the answer given the first time
    h) Add tastefulness

Feature #5: Dashboard

1) **Purpose:** To present our analysis results in the form of various visualisations on a single webpage
2) **Methodology:**
    a) Construct visualisations e.g. Tag Cloud, Jigsaw

Consolidate all visualisations on a single platform e.g. Plotly, Node.js

D. Challenges and Mitigation Plan

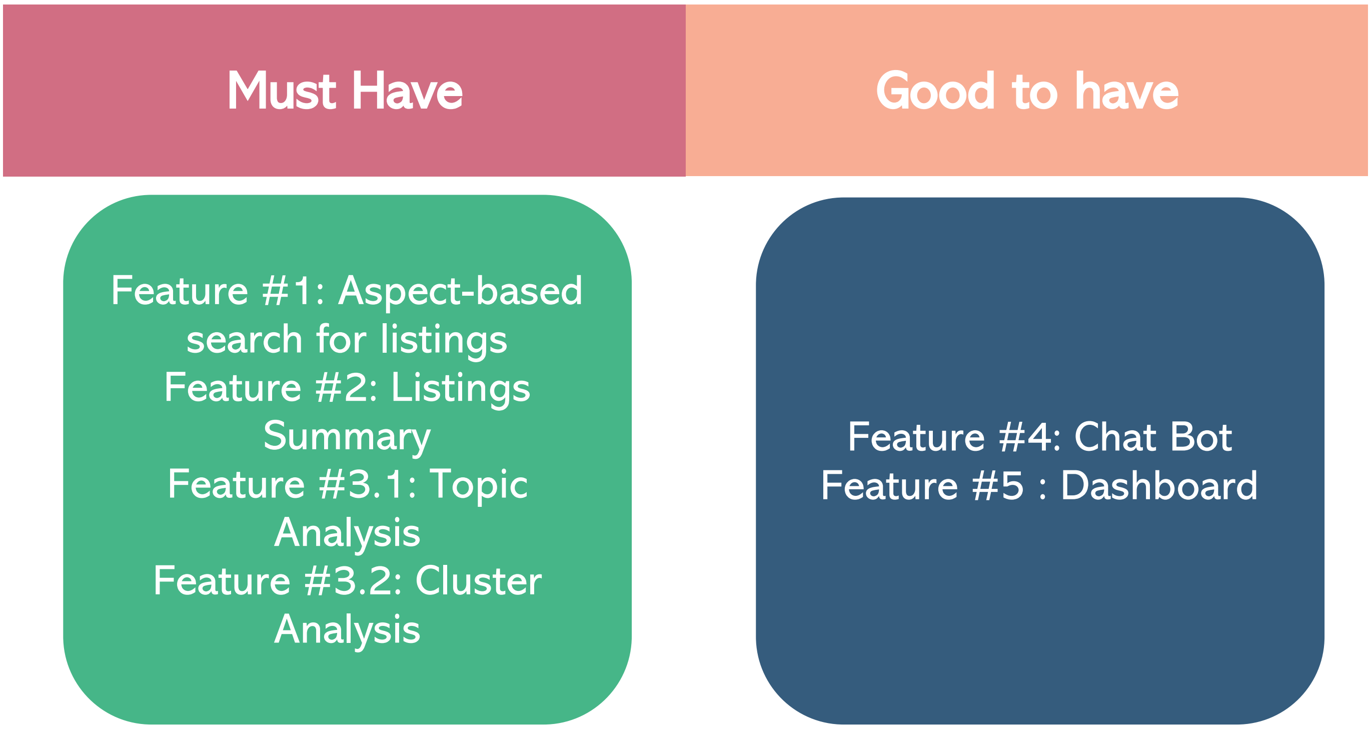
1) **Sheer size of the data:** The size of dataset for Airbnb listings is too large for a complete analysis within the given time frame and resources available. Thus, to mitigate the risk of over scoping we are limiting the project to the city of Los Angeles, California, United States.
2) **User Adoption:** This challenge mainly arises due to the unforeseen results of our model. Our model might not prove to give significant results to aid decision making. In order to mitigate it we have planned a structured approach to each model by first testing it on a smaller dataset based on user questions gathered from our mini survey. This will ensure a greater user value add.
3) **Data Veracity:** Since the entire project is greatly dependent on the data collected from the open source platform on airbnb reviews and listings, any white noise within the dataset might have a significant impact on the model. To mitigate this, we have planned an overall data cleaning process followed by a separate data cleaning unique to each model ensuring as little noise as possible is fed to our project. The data also had null values. To preserve all the information, we imputed or dropped the rows and columns containing null values while conducting exploratory analysis that made use of these features

E. Appendix

References:

1) https://machinelearningmastery.com/multi-class-classification-tutorial-keras-deep-learning-library/
2) https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff
3) https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/
4) https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec

Must have/ Good to have:



Must Have

Good to have

Feature #1: Aspect-based search for listings
Feature #2: Listings Summary
Feature #3.1: Topic Analysis
Feature #3.2: Cluster Analysis

Feature #4: Chat Bot
Feature #5 : Dashboard

Exhibit A:

What does it take to be a super host? Average Rating by Response Rate

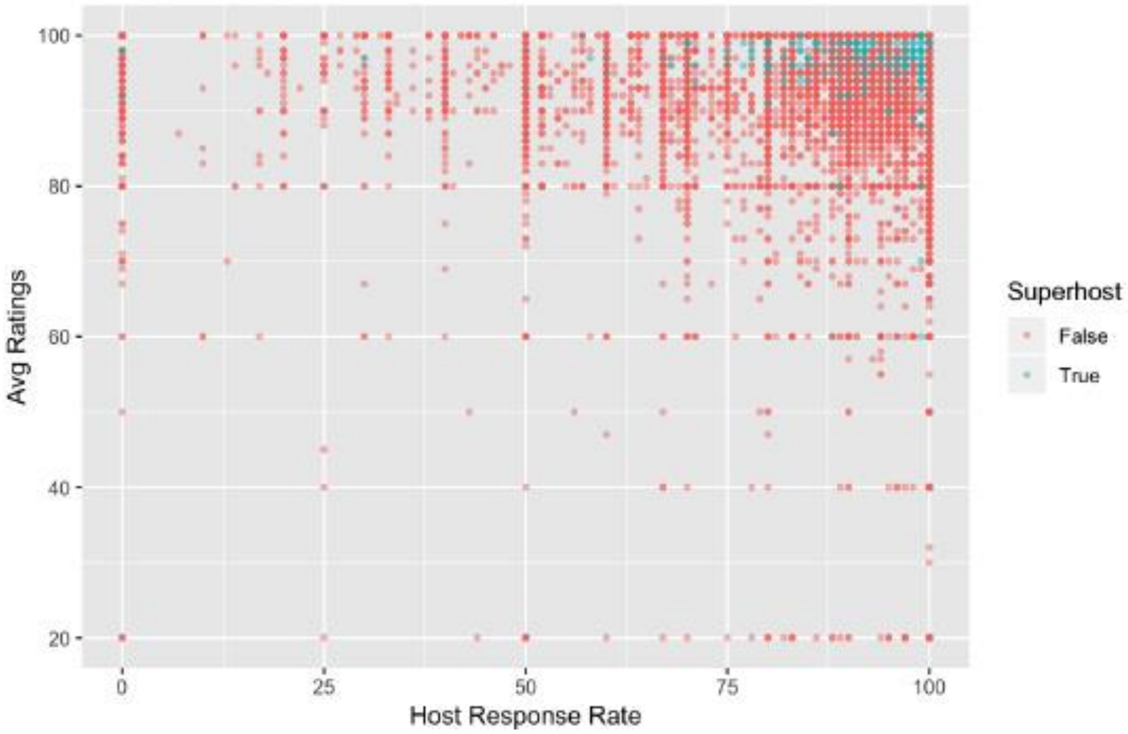


Exhibit B:

Type of host based on Instant Booking, Cancellation Policy, Room Type

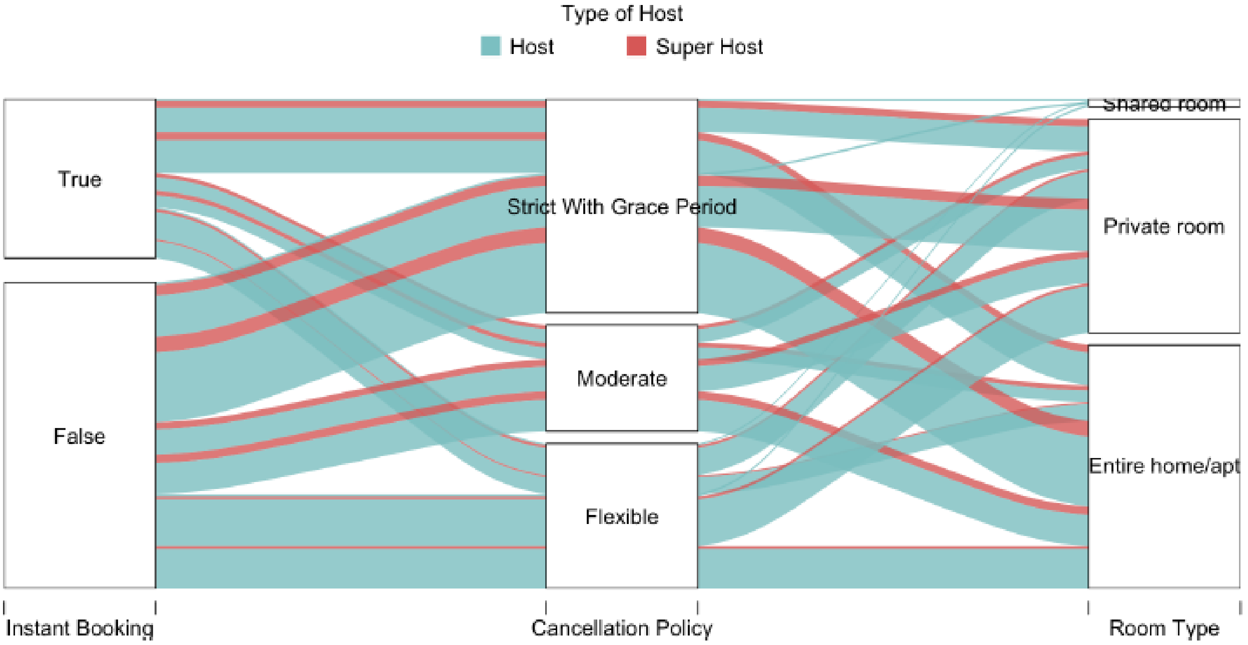
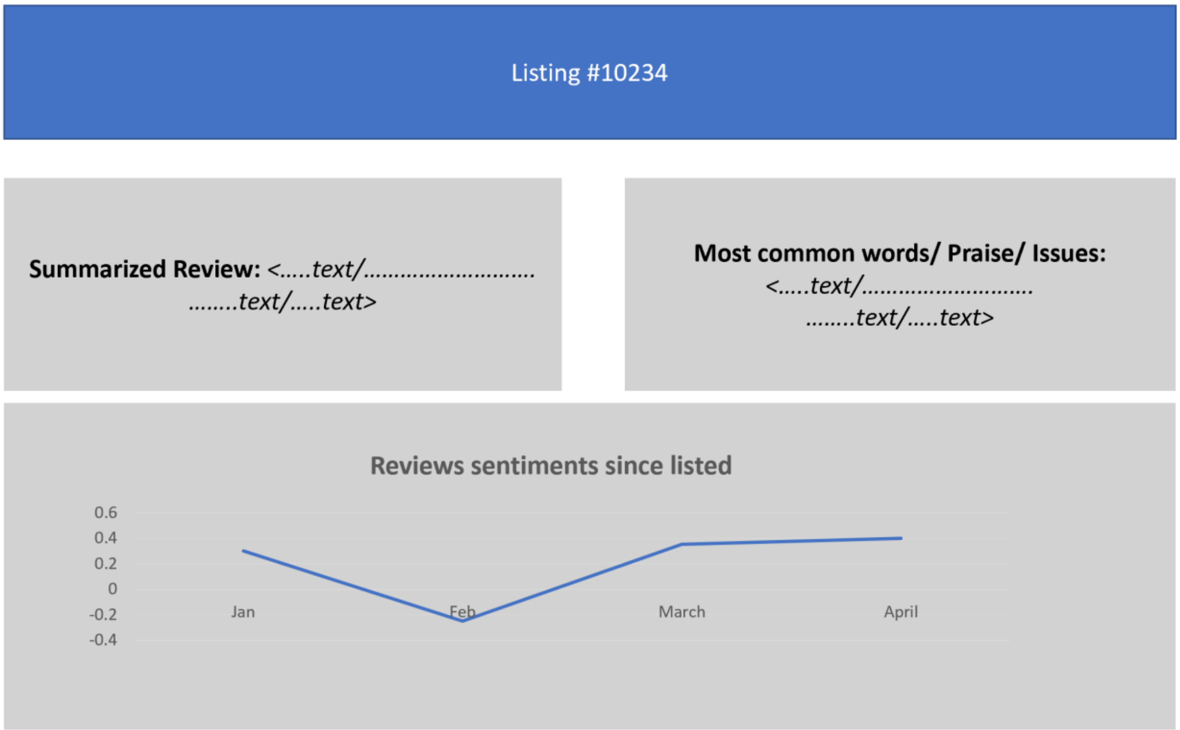


Exhibit C:

Prototype of the visualisation for Feature Listing Summary



Listing #10234

Summarized Review: <.....*text/*............................ ....*text/*.....*text*>

Most common words/ Praise/ Issues: <.....*text/*............................ ....*text/*.....*text*>

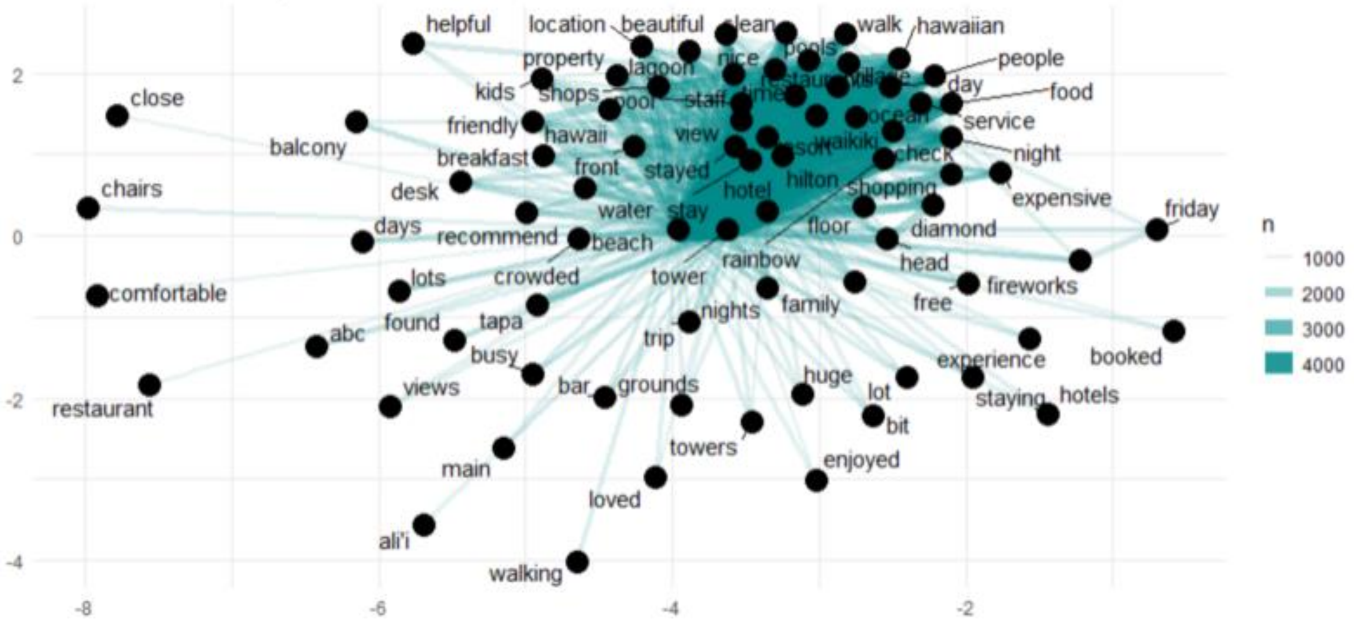
Exhibit D:

Sample Word Cloud

good
location
great
staff
small
ibis
clean
stay
rooms
bugis
breakfast
service
helpful
nice
food
mrt
walk
stayed
friendly
bencoolen
comfortable
street
free
night
excellent
place
located
close
station
business
hotels
bed
easy
price
time
back
budget
shopping
restaurant
bathroom
recommend
airport
expensive
orchard
check
bus
taxi
distance
minutes
money
nearby
cheap
size
internet
modern
reception
problem
wifi
walking
convenient
found
lobby
efficient
area
floor
nights
day
market
road
desk
india
city
local
front
big
amenities
quality
junction
shower
markets
find
hawker
restaurants
store
times
fantastic
worth
mall
bar
coffee
priced
hours
dont
long
visit
taxis
made
busy
mins
morning
put
side
minute
luggage
pool
expect
pretty
included
facilities
hour
extremely
accor
quiet
flight
adequate
paid
simple
reasonable
lim
make
min
happy
quick
family
safe
staying
eat
space
decent
access
door
guests
things
tiny
beds
central
late
arrived
lots
early
perfect
711
checkin
square
standard
selection
fridge
buffet
water
laundry
high
extra
hard
large
travel
couple
fast
view
needed
functional
center
provided
pay
train
pleasant
choice
short
basic
experience
asked
didnt
bad
thing
stopover
card
compact
sim
sleep
shops
rate
options
plenty
star
days
prices
enjoyed
feel
centre
booked
highly
cost
people
court
places
book
lot
trip
bit
fine

Exhibit E:

Sample Word Network


helpful
location
beautiful
clean
walk
hawaiian
property
lagoon
nice
pools
people
kids
shops
pool
staff
time
restaurants
village
day
food
close
friendly
hawaii
view
ocean
service
balcony
breakfast
front
stayed
resort
waikiki
check
night
chairs
desk
hotel
hilton
shopping
expensive
water
stay
friday
days
recommend
beach
floor
diamond
rainbow
head
tower
crowded
lots
comfortable
fireworks
family
free
nights
abc
found
tapa
trip
busy
experience
booked
views
bar
grounds
huge
lot
staying
hotels
restaurant
bit
towers
main
enjoyed
loved
ali'i
walking
n