Final Project

# Text Mining of Hotel Reviews From TripAdvisor

*Prepared by:*

**Team 3**

**UCONN OPIM-5671**

**Summer 2016**

## Table of Contents

***OPIM 5671 – Summer 2016 Team 3 Members:***

- Vidya Minukuri
- Ramya Mupparaju
- Naresh Vemula
- Youngeun Park

“***The work presented in this report, project and data files is our work and our work alone.***” – Team 3

# 1 EXECUTIVE SUMMARY

## 1.1 Business Problem

Hospitality industry thrives on customer feedback and customer online reviews makes a huge impact on the brand value of a service offered by a hospitality service provider. In the recent era of globalization, TripAdvisor is the popular website that most customers use for booking hotels online and also provide their reviews. It is imperative for the hotel service providers to understand and analyze the customer sentiment about a service in the hotel and improve their services to increase their occupancy and withstand the competition. In addition, the increasing amount of reviews pose a challenge on hotel management on what reviews to consider that are worth improving their reputation.

As a result, Hospitality service providers are in search of finding answers to some of the key questions impacting their business as listed below.

- Strategy to run chain of hotels across many cities?
- What drives customer to reserve a hotel booking?
- What turns customer away from choosing a hotel?
- How do I balance occupancy and price to maximize revenue?
- Which segment do I operate in?
- Which segment should I compete with?
- How is my performance over a period of time?

Our objective is to do text mining on TripAdvisor customer Reviews and create business value for hotel management. Our Project aims to analyze and provide useful insights for hotels at different levels and answer the above questions.

- ➢ Prominent factors for Hotels located in different demographics
- ➢ In-depth Analysis of hotels in a particular city
- ➢ Analysis for individual hotel

. Our insights will help them improve and further innovate the services offered.

### 1.1.1 Data Source

Our dataset is collected from the below link

http://kavita-ganesan.com/entity-ranking-data

## 1.2 Key Takeaways

Text mining of Customer reviews revealed a number of different dimensions that hospitality service providers should focus on. The below set of recommendations can be part of their core strategy to sustain in the industry for long time.

- Consider city specific customizations and branding for hotel chains operating in multiple cities.
- Develop competing strategy with clusters of higher average rating.
- Consider Seasonal variations of customer services
- Consider measuring impact of renovation that leads to value maximization on investment made on renovation.
- Usage of tools and Categorizing customer sentiment for improvements.

# 2 DATA PREPROCESSING

Our Dataset initially consists of **237303** reviews of hotels in 10 different cities (Dubai, Beijing, London, New York city, New Delhi, San Francisco, Shanghai, Montreal, and Chicago) and their aspect ratings. Each city has about 80-700 hotels.

## 2.1 Variable Analysis

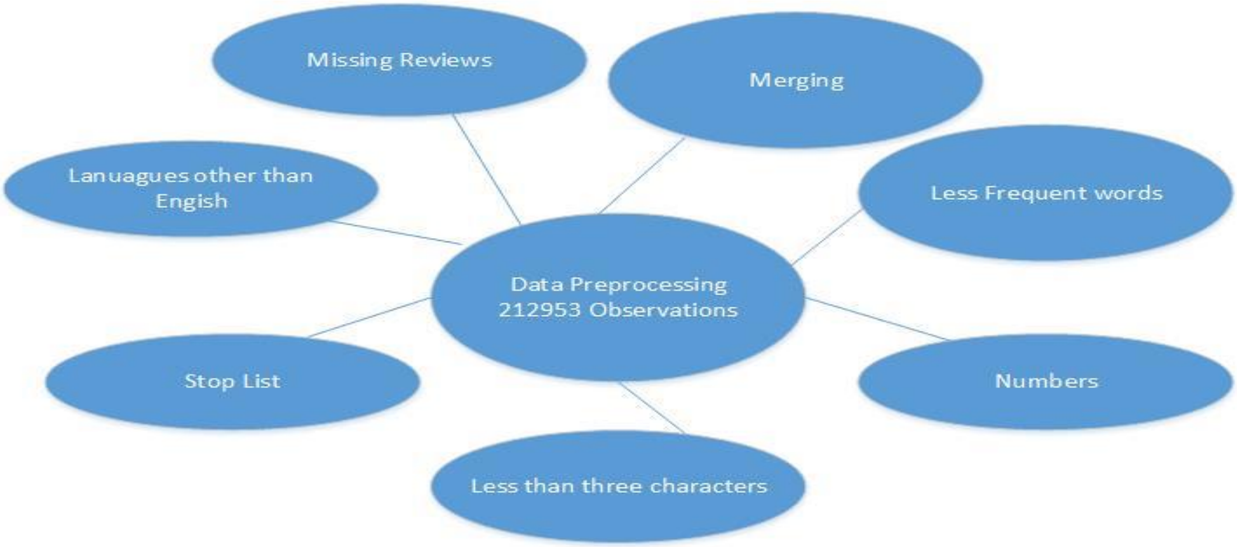
We have 16 variables in our data:

doc_id, hotel_name, hotel_url, street, city, country, zip, num_reviews, cleanliness, Room, Service, Location, Value, Overall_rating source, Review_Title,Full_Reviews

However, we used Review title, Full reviews, doc_id and hotel_name for our Text Mining and explored the aspect ratings depending on country, pin code.

## 2.2 Data Cleaning

We have implemented the following steps for data cleaning which resulted in 212953 observations.


Missing Reviews
Merging
Lanuagues other than Engish
Less Frequent words
Data Preprocessing 212953 Observations
Stop List
Numbers
Less than three characters

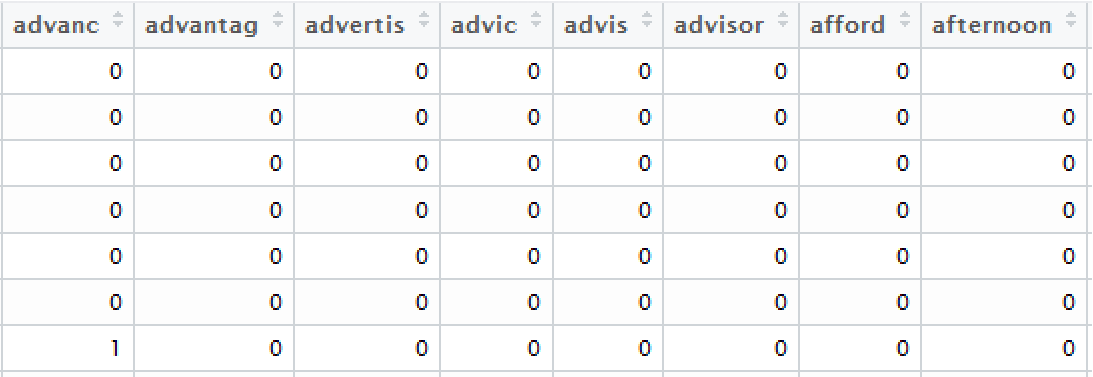
1. Generated the corpus using the textual data we have in the form of customer reviews for the hotels.
2. Converted all the textual data to lower case which helps the document to be plain and useful in searching for the text.
3. Converted corpus to plain text document to make the corpus stable.

4. As we are using hotel reviews, there is a high possibility that customers use numbers such as bill amount, room number in the reviews which might lead to might be the noise in our data. Hence, removed all the numbers in the reviews.
5. We have used default English stop words and customized stop words i.e. typically the words which are used quite often but they were not useful to classify the document.
6. we have performed stemming on the words in corpus to provide better classification rate of similar words.
7. We have deleted the Observations which has only review title and did not have the Full Review text
8. We have deleted words with less than three characters. These can be type errors or prepositions which don't help in analyzing the data.

# 3 Customer Review Analysis Using R

Pic 1.1 shows a typical document term matrix(DTM), in our case we used all the reviews by customers to the hotels across 8 different cities in the world to generate the matrix. DTM has been the founding step to perform multivariate analysis such as clustering, Linear Discriminant Analysis (LDA) on textual data, this will be explained in the coming sections. We have selected the words which are at least repeated in 1% of the data. This helps us to capture better classifiers and we can reduce the unwanted terms from the data. This process is called removing sparse terms.

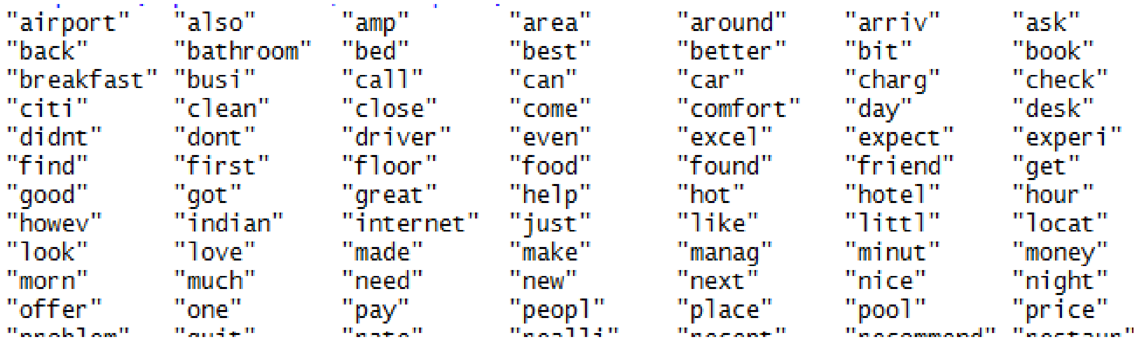
1.1 Typical document Term Matrix



| advanc | advantag | advertis | advic | advis | advisor | afford | afternoon |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

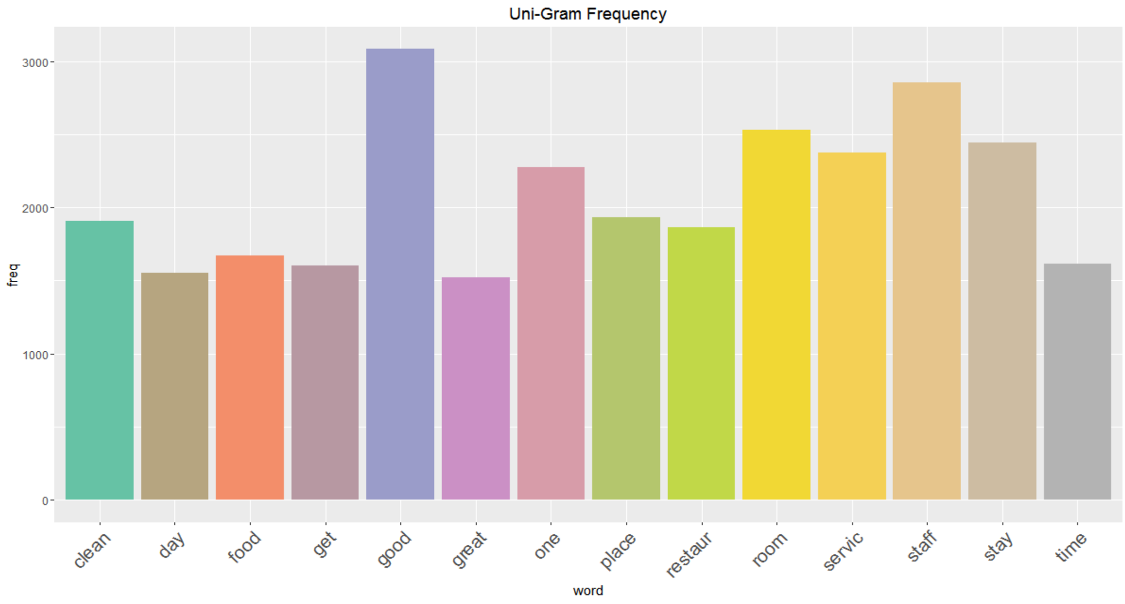
This helps us to determine what words are prominent in reviews, probably most weighted terms or words used the most number of times in the user reviews. A better example in the hotel reviews is shown in the below diagram, these are the words which are repeated at least 500 times in the reviews for Delhi hotels. If we observe words like "area, clean, breakfast, charge, comfort, great, friend' shows positive attitude towards the hotel, where as words like "hot, didn't, better, little' explains negative impact. No wonder a hotel which is liked by someone is not good for another person.

Words used at least for 500 times in all Delhi reviews.



```
"airport"   "also"      "amp"       "area"      "around"    "arriv"     "ask"
"back"      "bathroom"  "bed"       "best"      "better"    "bit"       "book"
"breakfast" "busi"      "call"      "can"       "car"       "charg"     "check"
"citi"      "clean"     "close"     "come"      "comfort"   "day"       "desk"
"didnt"     "dont"      "driver"    "even"      "excel"     "expect"    "experi"
"find"      "first"     "floor"     "food"      "found"     "friend"    "get"
"good"      "got"       "great"     "help"      "hot"       "hotel"     "hour"
"howev"     "indian"    "internet"  "just"      "like"      "littl"     "locat"
"look"      "love"      "made"      "make"      "manag"     "minut"     "money"
"morn"      "much"      "need"      "new"       "next"      "nice"      "night"
"offer"     "one"       "pay"       "peopl"     "place"     "pool"      "price"
"problem"   "quit"      "rate"      "realli"    "recent"    "recommend" "restaur"
```

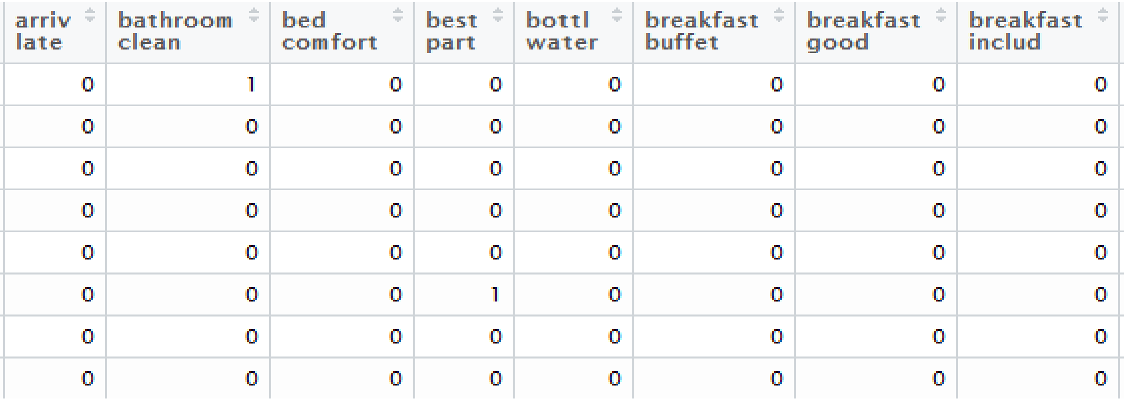
The same is explained using the UNI-gram frequencies, where we have calculated the number of times a word is repeated in the reviews, the words in the picture below are at least used more than 1500 times by the customers in their reviews. We have used 4300 Delhi customer reviews to analyze the patterns. We observed most of the time people used words like “clean, good, food, service, staff, stay”, these are the basic things everyone will look when they want to book a hotel. For a beginner in hotel industry this information is crucial to come up with the quality of services they want to provide to the customers.



However, by using single words we will not be able to tell whether these are appreciating the services offered or complaining about services offered, to study this in detail we have used a theory called N-Gram-Based Text Categorization, where N is combination of words required to be formed.

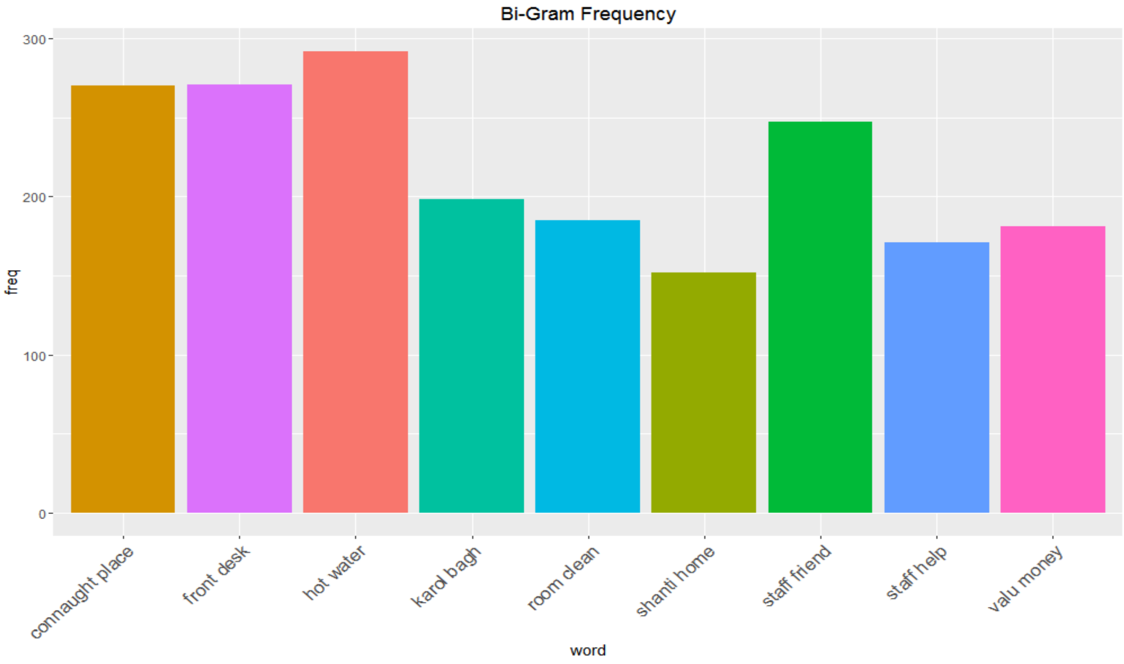
Typically, after cleaning the data in the corpus with the above stated text pre-processing steps, we create a DTM using N-grams where N (Integer). This iteratively computes the distances between words among the documents and selects the word combinations the customer used the most based the N.

We have computed the N-gram frequencies on Delhi reviews data with N values (1,2). Picture below shows two-word combination document term matrix that is generated. This has more insights because two words tend to give a meaningful sentence, for example if we take breakfast good variable, when this is 1 for a document, the customer is saying they liked breakfast in here. This type of information if useful for the future customers.

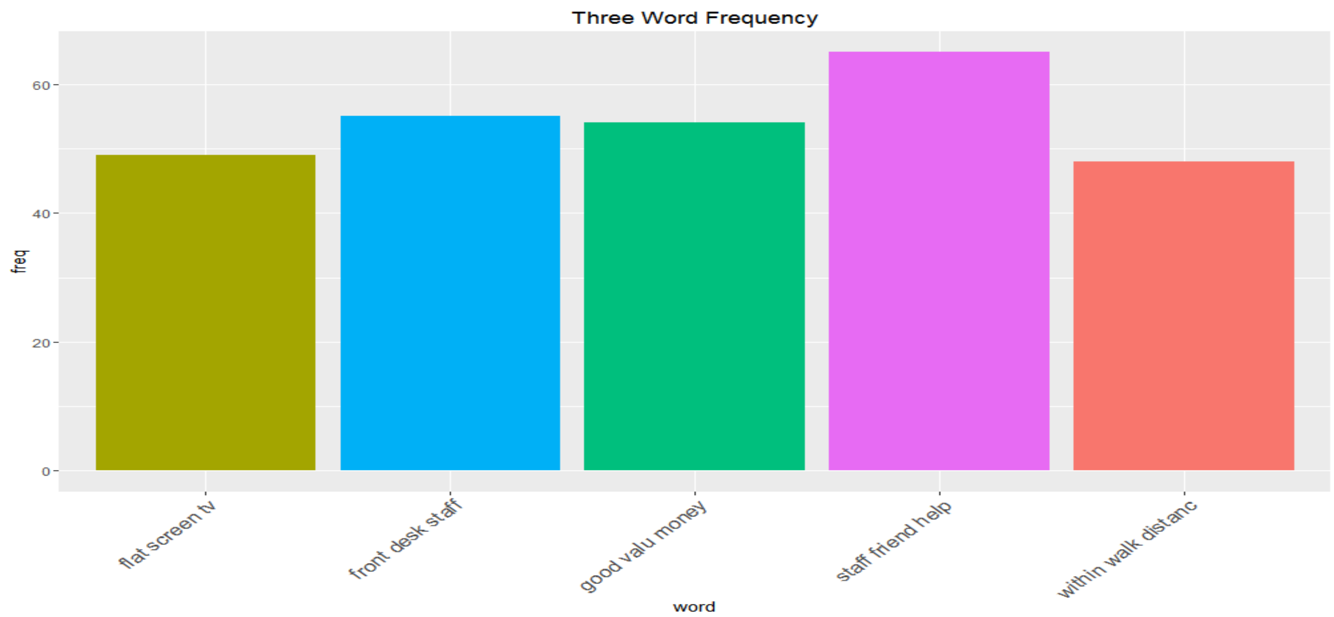


| arriv late | bathroom clean | bed comfort | best part | bottl water | breakfast buffet | breakfast good | breakfast includ |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

A more detailed visual is shown below, these are the word combinations which N-gram computed and has a frequency of more than 150 times in the documents, Words like ‘carol bagh, shanti home’ are place names, this helps us removing these words as stop words from the data. More over words like ‘room clean, staff help, value money’ shows positive attitude towards the hotel.



We have also used TRI-gram computations to find more trends among the user reviews, these categories helped us to initially categorize among hotels, if we can observe few hotels review were having comments “good value money, staff friend help, within walking distance”, these are positive words which are helps customers know more about the hotels.

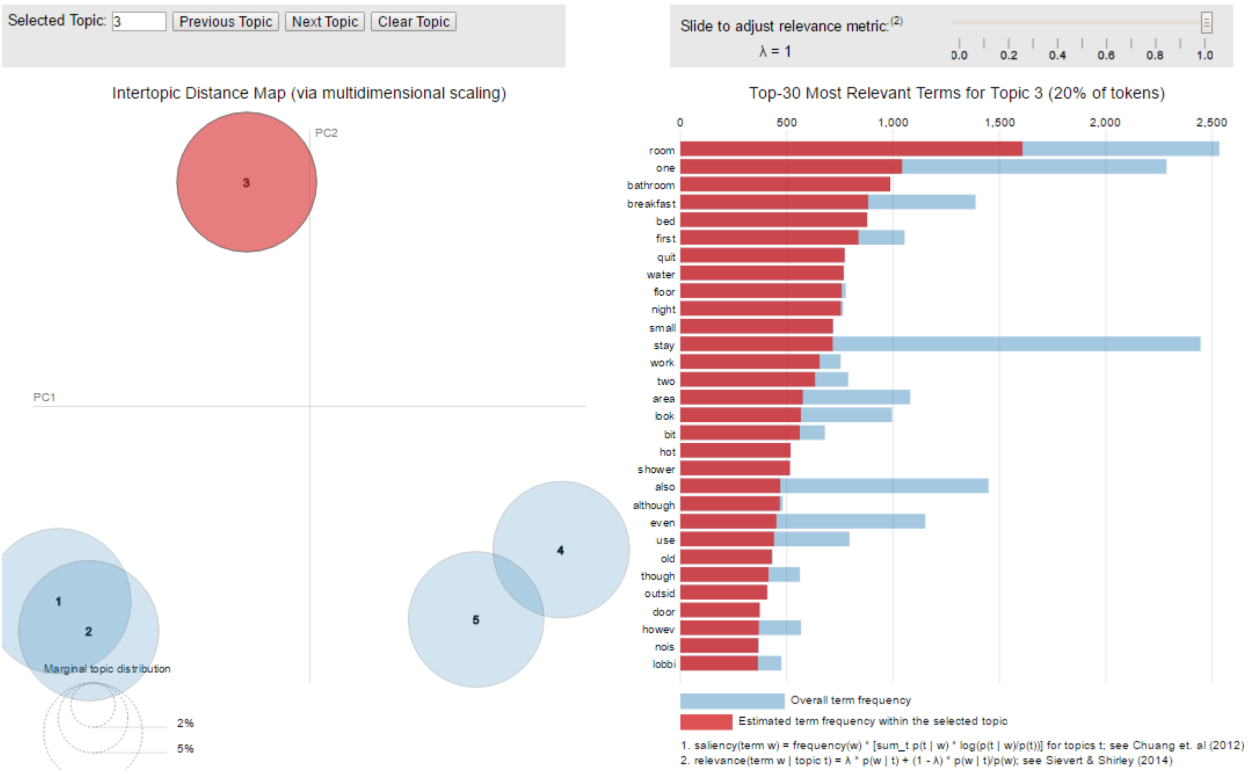


## Topic Models using LDA

Latent Dirichlet Allocation is the process of topic modelling technique we used to segregate between the Reviews, with help of Gibbs sampling we created five different topics for the hotels in city Delhi. I have incorporated a folder of the web application, please use it in a Mozilla browser to get more insights. Below are the text topics we generated for the hotels in Delhi, From the LDA visualization and the normal tabular forms.

We can clearly see that distances between the topics 3 and rest is very large, as we assumed to our surprise The topic 3 has few negative words such as ‘noise, tired, lack, dirty’ and many more which you can effectively observe using the web application we provided. Using this data combining with individual hotel rating we determined that reviews in the topic three are more towards negative sentiment. Moreover, Topics 4 and 5 have words such as ‘service, good, clean ‘stating positive attitude towards the hotels.

Latent Dirchelet Allocation of Topic Models Visulization.

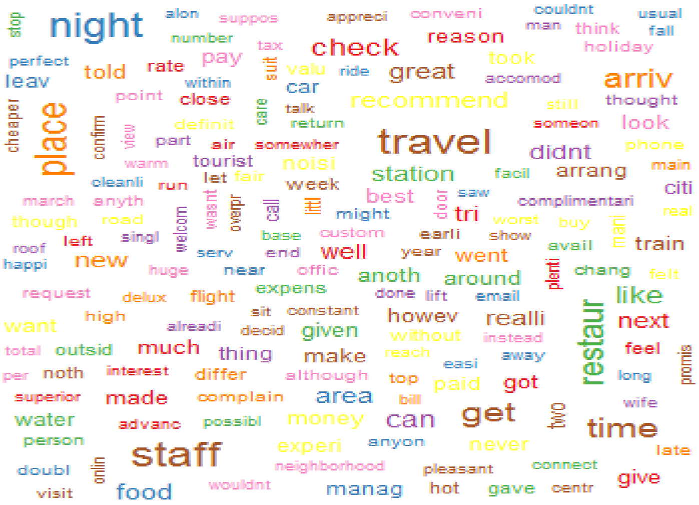


1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

## Word Cloud

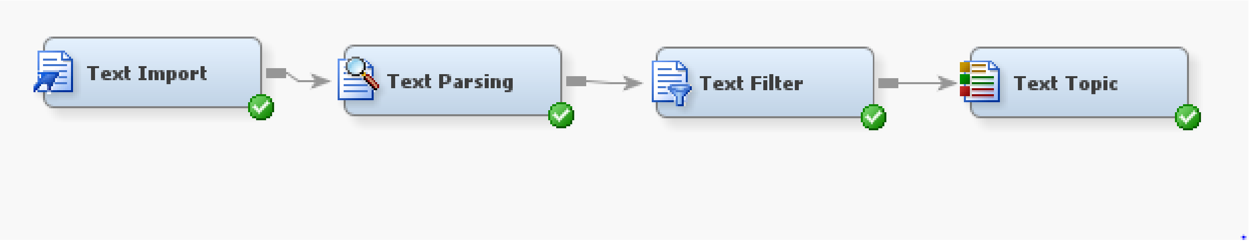
We have created a dynamic function to create word clouds  based on the user's input city and the hotel name. Typically a future customer or a hotel management department would be using these word clouds for better information on  services.

Hotel Ajanta word cloud


stop
night
alon
suppos
appreci
conveni
couldnt
usual
man
think
fall
number
tax
check
reason
holiday
perfect
told
rate
pay
suit
valu
ride
great
took
leav
within
car
accomod
arriv
point
close
recommend
still
thought
cheaper
place
confirm
view
definit
care
talk
return
someon
look
part
air
somewher
travel
didnt
phone
warm
tourist
noisi
station
facil
arrang
main
cleanli
run
let
fair
week
citi
march
anyth
wasnt
overpr
call
ital
best
door
saw
complimentari
real
though
road
welcom
might
tri
worst
buy
mani
singl
base
custom
earli
show
roof
left
train
happi
new
serv
end
well
year
went
avail
huge
near
offic
anoth
around
plenti
chang
felt
request
delux
flight
expens
done
lift
email
like
want
high
sit
constant
howev
realli
restaur
next
alreadi
decid
given
without
instead
total
outsid
much
thing
make
reach
easi
away
feel
per
noth
interest
differ
although
top
paid
got
long
promis
superior
made
complain
area
bill
wife
water
advanc
possibl
money
can
get
two
time
person
staff
experi
anyon
never
late
doubl
onlin
neighborhood
pleasant
connect
give
visit
food
wouldnt
manag
hot
gave
centr

# 4 Topic Modelling and Clustering Using SAS

We have used the process of text mining nodes as shown below and analyzed the text filter,text topics node results for different cities.



**Dubai**

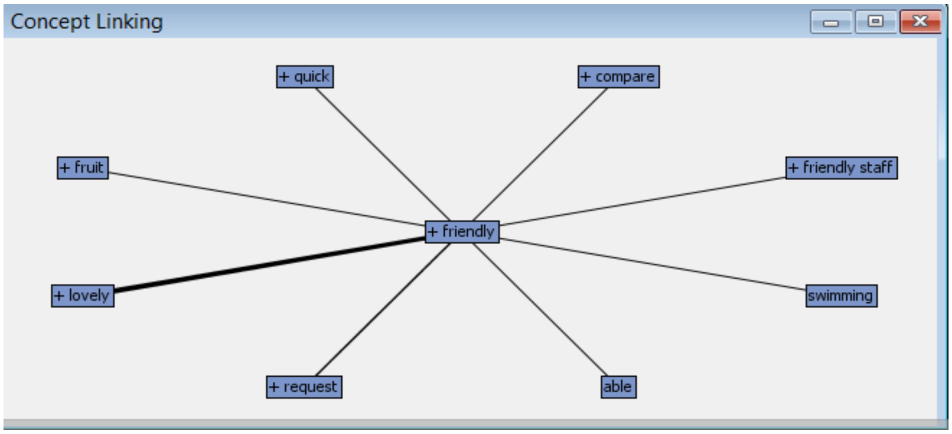
We have imported the reviews of all the hotels in Dubai and created 5 Multi term topics.

Analyzing the documents under topics shows that the classification is consistent.

| topic_id | topic_name | Number of Terms | #Docs |
|---|---|---|---|
| 1 | +room+stay+staff+pool+good | 291 | 44 |
| 2 | +villa+resort+sea+mina+buggy | 441 | 32 |
| 3 | +club+nightclub+creek+disco+music | 465 | 39 |
| 4 | executive,sheraton,zayed+lounge,hyatt | 490 | 50 |
| 5 | +apartment+kitchen+washmachine+studio+wash | 292 | 34 |

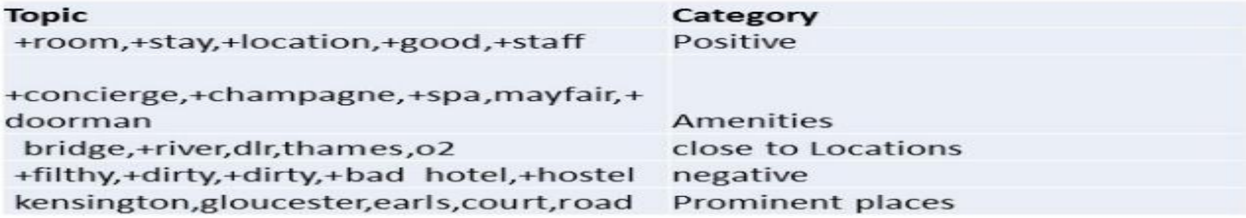
- Staff in Dubai are friendly.

Term friendly has a frequency of 3134 in 212 documents. Concept link also suggests that friendly is thickly associated with lovely and also related with friendly staff. Examination of documents with words friendly suggested that the staff are friendly.



We also observed that most customers are speaking about beach fun and shopping.So, new players entering the hotel industry in Dubai must ensure that their hotel is located near to Shopping Malls, beach and staff is properly trained.

### London



| Topic | Category |
|---|---|
| +room,+stay,+location,+good,+staff | Positive |
| +concierge,+champagne,+spa,mayfair,+ doorman | Amenities |
| bridge,+river,dlr,thames,o2 | close to Locations |
| +filthy,+dirty,+dirty,+bad hotel,+hostel | negative |
| kensington,gloucester,earls,court,road | Prominent places |

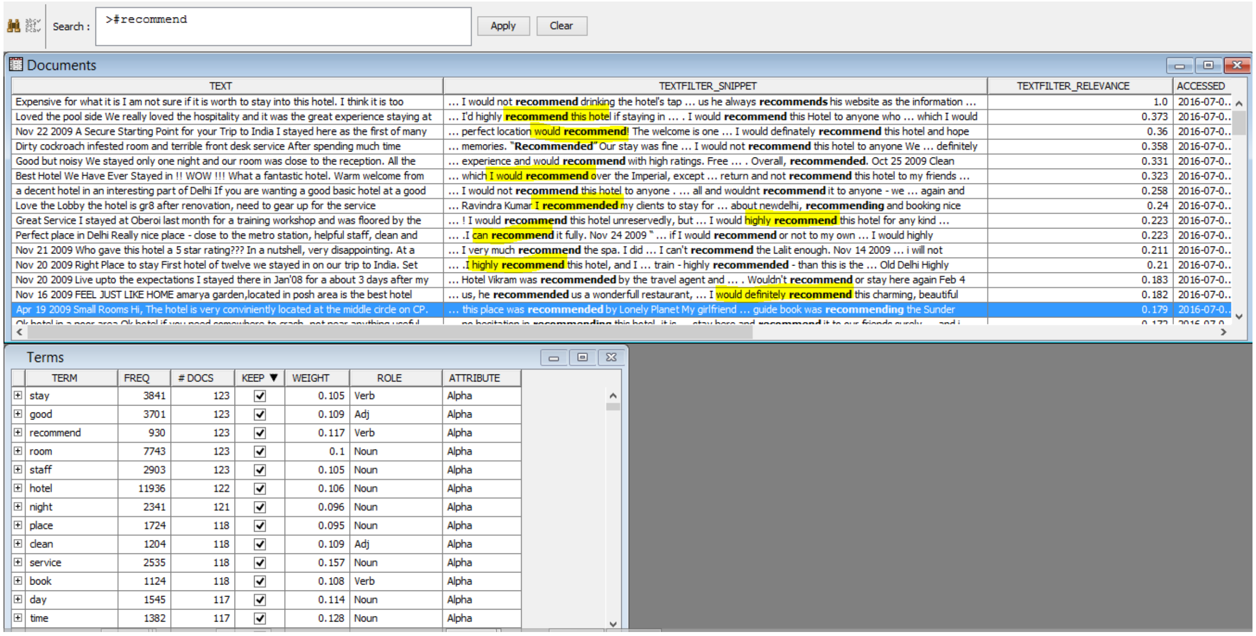
We observe that customers are mainly speaking about tourist attractions like London bridge, Thames river, o2 state of the art Athena

### Montreal

We observe that there is bedbug problem in Montreal and customers are also inclined towards Airport shuttle.

### Delhi

Some of the hotels are highly recommended.

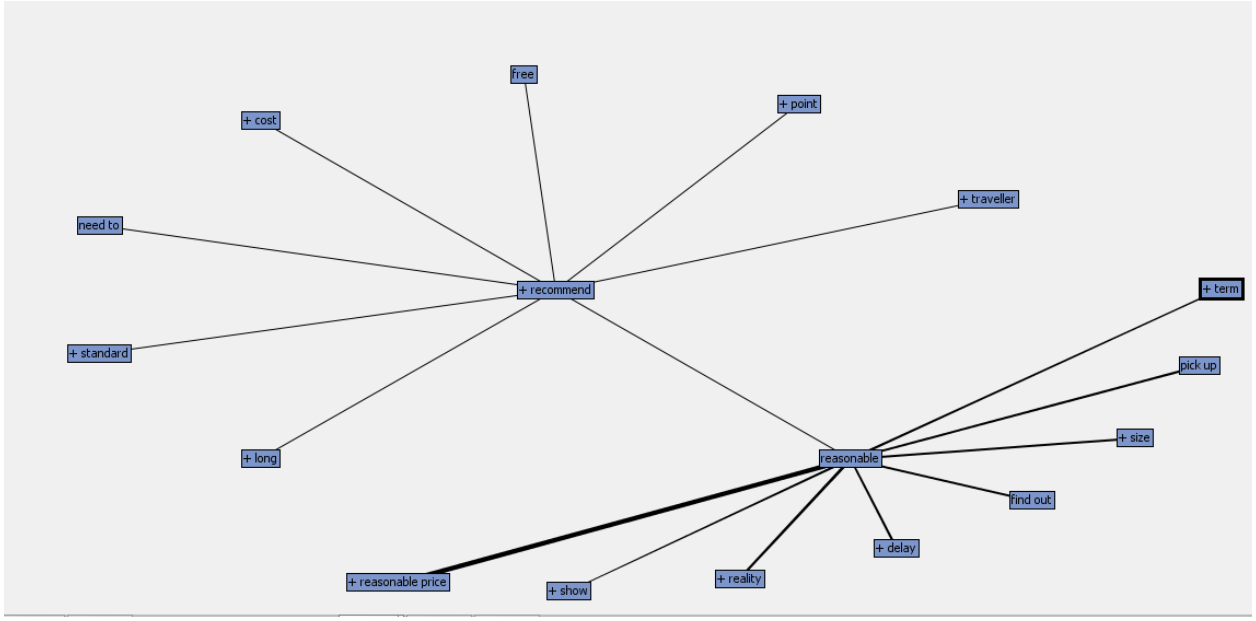


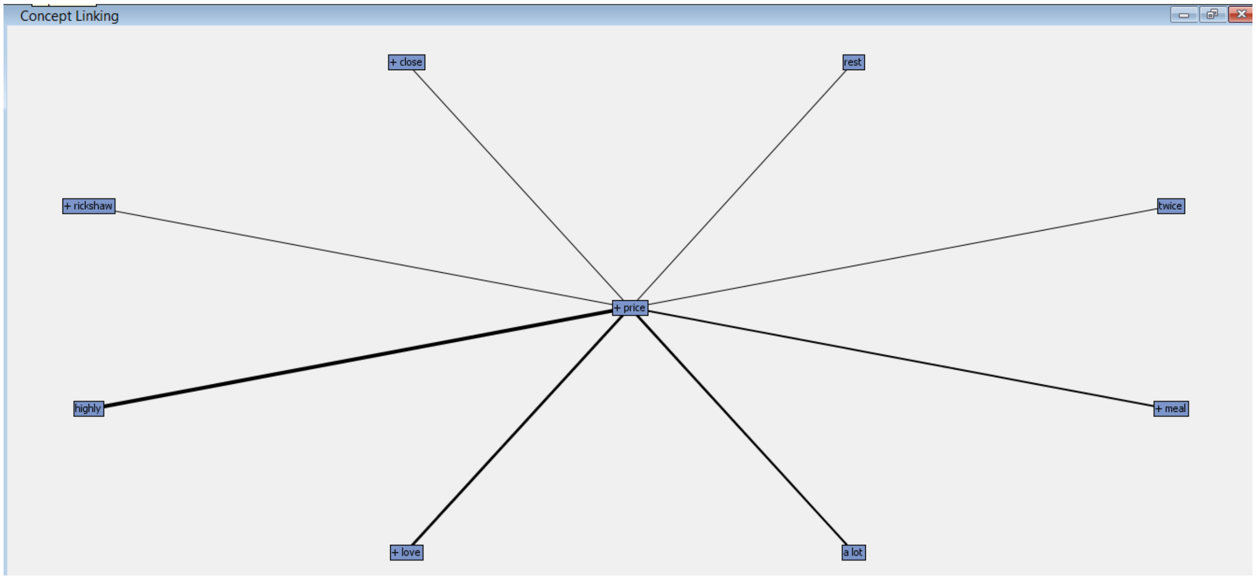
>#recommend

| TEXT | TEXTFILTER_SNIPPET | TEXTFILTER_RELEVANCE | ACCESSED |
|---|---|---|---|
| Expensive for what it is I am not sure if it is worth to stay into this hotel. I think it is too | … I would not **recommend** drinking the hotel's tap … us he always **recommends** his website as the information … | 1.0 | 2016-07-0.. |
| Loved the pool side We really loved the hospitality and it was the great experience staying at | … I'd highly **recommend** this hotel if staying in … . I would **recommend** this Hotel to anyone who … which I would | 0.373 | 2016-07-0.. |
| Nov 22 2009 A Secure Starting Point for your Trip to India I stayed here as the first of many | … perfect location would **recommend**! The welcome is one … I would definately **recommend** this hotel and hope | 0.36 | 2016-07-0.. |
| Dirty cockroach infested room and terrible front desk service After spending much time | … memories. "**Recommended**" Our stay was fine … I would not **recommend** this hotel to anyone We … definitely | 0.358 | 2016-07-0.. |
| Good but noisy We stayed only one night and our room was close to the reception. All the | … experience and would **recommend** with high ratings. Free … . Overall, **recommended.** Oct 25 2009 Clean | 0.331 | 2016-07-0.. |
| Best Hotel We Have Ever Stayed in !! WOW !!! What a fantastic hotel. Warm welcome from | … which I would **recommend** over the Imperial, except … return and not **recommend** this hotel to my friends … | 0.323 | 2016-07-0.. |
| a decent hotel in an interesting part of Delhi If you are wanting a good basic hotel at a good | … I would not **recommend** this hotel to anyone . … all and wouldnt **recommend** it to anyone - we … again and | 0.258 | 2016-07-0.. |
| Love the Lobby the hotel is gr8 after renovation, need to gear up for the service | … Ravindra Kumar I **recommended** my clients to stay for … about newdelhi, **recommending** and booking nice | 0.24 | 2016-07-0.. |
| Great Service I stayed at Oberoi last month for a training workshop and was floored by the | … ! I would **recommend** this hotel unreservedly, but … I would highly **recommend** this hotel for any kind … | 0.223 | 2016-07-0.. |
| Perfect place in Delhi Really nice place - close to the metro station, helpful staff, clean and | … .I can **recommend** it fully. Nov 24 2009 " … if I would **recommend** or not to my own … I would highly | 0.223 | 2016-07-0.. |
| Nov 21 2009 Who gave this hotel a 5 star rating??? In a nutshell, very disappointing. At a | … I very much **recommend** the spa. I did … I can't **recommend** the Lalit enough. Nov 14 2009 … i will not | 0.211 | 2016-07-0.. |
| Nov 20 2009 Right Place to stay First hotel of twelve we stayed in on our trip to India. Set | … .I highly **recommend** this hotel, and I … train - highly **recommended** - than this is the … Old Delhi Highly | 0.21 | 2016-07-0.. |
| Nov 20 2009 Live upto the expectations I stayed there in Jan'08 for a about 3 days after my | … Hotel Vikram was **recommended** by the travel agent and … . Wouldn't **recommend** or stay here again Feb 4 | 0.183 | 2016-07-0.. |
| Nov 16 2009 FEEL JUST LIKE HOME amarya garden,located in posh area is the best hotel | … us, he **recommended** us a wonderfull restaurant, … I would definitely **recommend** this charming, beautiful | 0.182 | 2016-07-0.. |
| Apr 19 2009 Small Rooms Hi, The hotel is very conviniently located at the middle circle on CP. | … this place was **recommended** by Lonely Planet My girlfriend … guide book was **recommending** the Sunder | 0.179 | 2016-07-0.. |
| Ok hotel in a poor area Ok hotel if you need somewhere to crash, not near anything useful | … no hesitation in **recommending** this hotel, it is … stay here and **recommend** it to our friends surely … and i | 0.172 | 2016-07-0.. |

| | TERM | FREQ | # DOCS | KEEP ▼ | WEIGHT | ROLE | ATTRIBUTE |
|---|---|---|---|---|---|---|---|
| ⊞ | stay | 3841 | 123 | ☑ | 0.105 | Verb | Alpha |
| ⊞ | good | 3701 | 123 | ☑ | 0.109 | Adj | Alpha |
| ⊞ | recommend | 930 | 123 | ☑ | 0.117 | Verb | Alpha |
| ⊞ | room | 7743 | 123 | ☑ | 0.1 | Noun | Alpha |
| ⊞ | staff | 2903 | 123 | ☑ | 0.105 | Noun | Alpha |
| ⊞ | hotel | 11936 | 122 | ☑ | 0.106 | Noun | Alpha |
| ⊞ | night | 2341 | 121 | ☑ | 0.096 | Noun | Alpha |
| ⊞ | place | 1724 | 118 | ☑ | 0.095 | Noun | Alpha |
| ⊞ | clean | 1204 | 118 | ☑ | 0.109 | Adj | Alpha |
| ⊞ | service | 2535 | 118 | ☑ | 0.157 | Noun | Alpha |
| ⊞ | book | 1124 | 118 | ☑ | 0.108 | Verb | Alpha |
| ⊞ | day | 1545 | 117 | ☑ | 0.114 | Noun | Alpha |
| ⊞ | time | 1382 | 117 | ☑ | 0.128 | Noun | Alpha |

Analysis of concept links shows that hotels with reasonable price and airport pick up are mostly recommended.





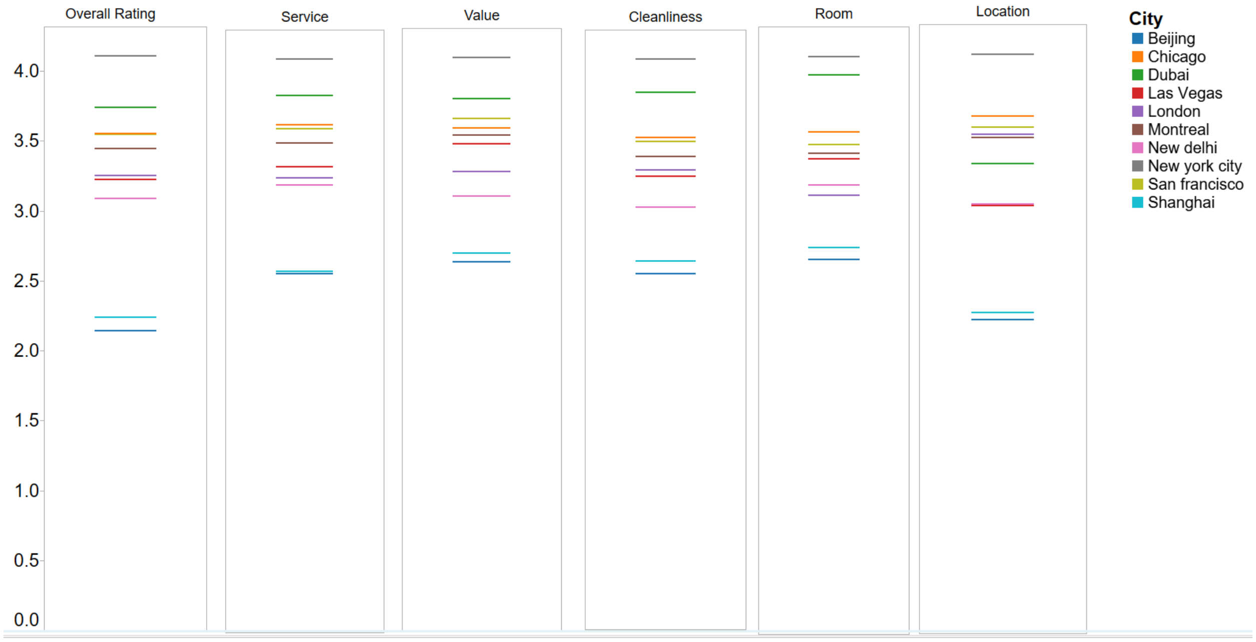
Price pays a major role for hotels in Delhi and customers think that some hotels are overpriced. Metro station access and airport pick up are the prominent factors.

### Shanghai

Our Analysis shows that breakfast buffet in shanghai hotels is very good and staff must be well trained in English. Some of the customers faced problems as hotel staff understands only Chinese language.

## 4.1 Sentiment Analysis

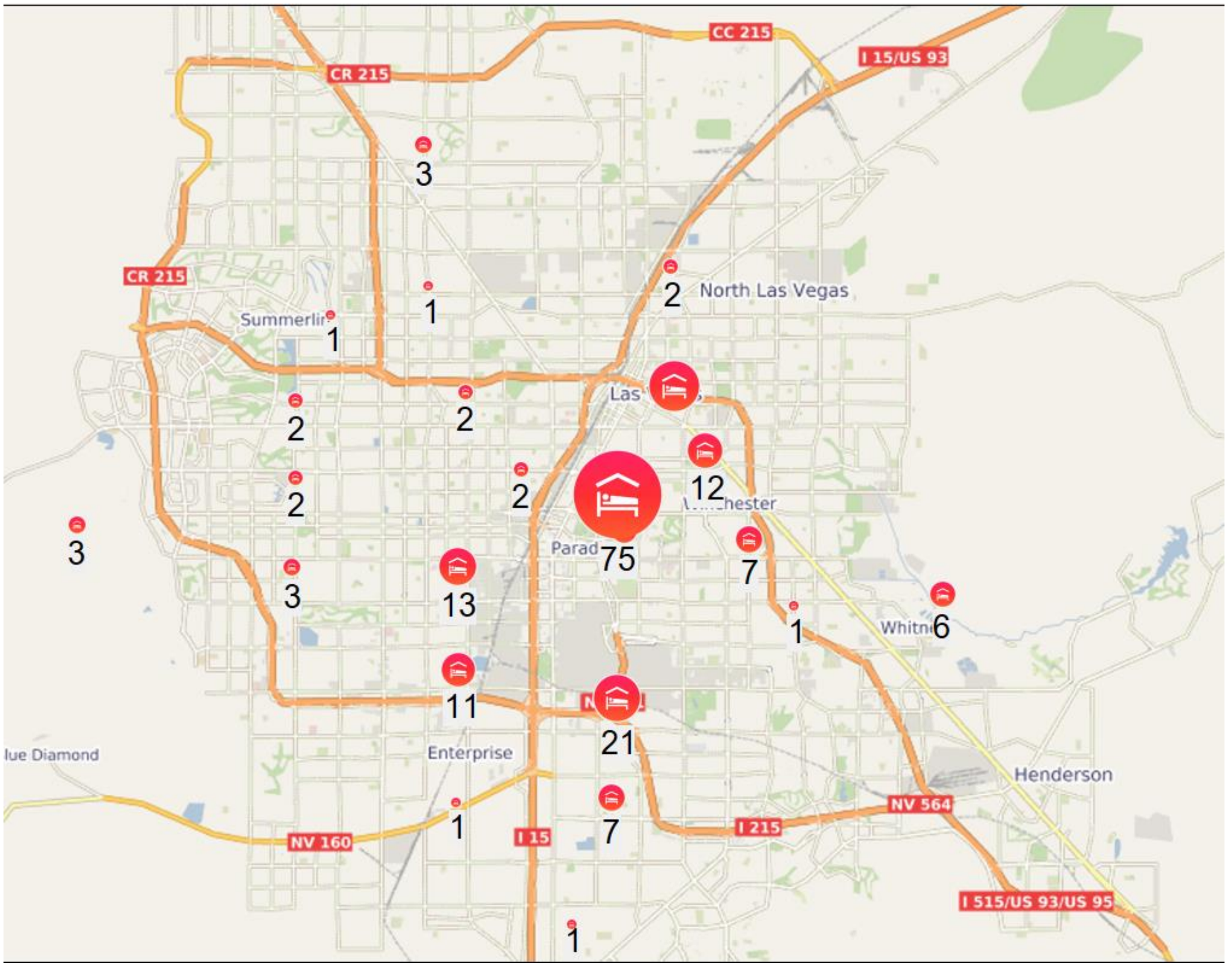
The data contains judgement related ranking for reviews on hotels across all cities. The plot below indicates New York excelled in many factors whereas Beijing, Shanghai should improve their relative ranking. From this , we have considered Las Vegas as a city of interest for deep dive analysis.



### Las Vegas Hotel review as a case study:

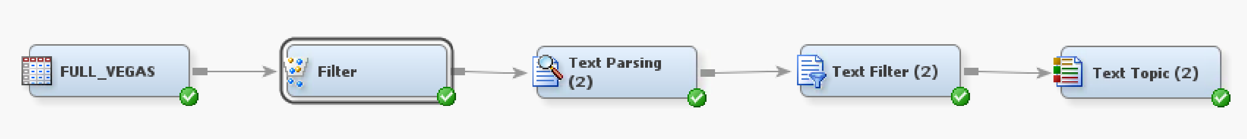
Las Vegas city in United States is one of the most sought after holiday destination for thousands of tourists every day. For hospitality service providers, doing business in Los Vegas is a significant challenge considering presence of several players in the industry. For this study, we considered to do deep dive analysis into Las Vegas hotels in order to bring useful insights to the service providers.

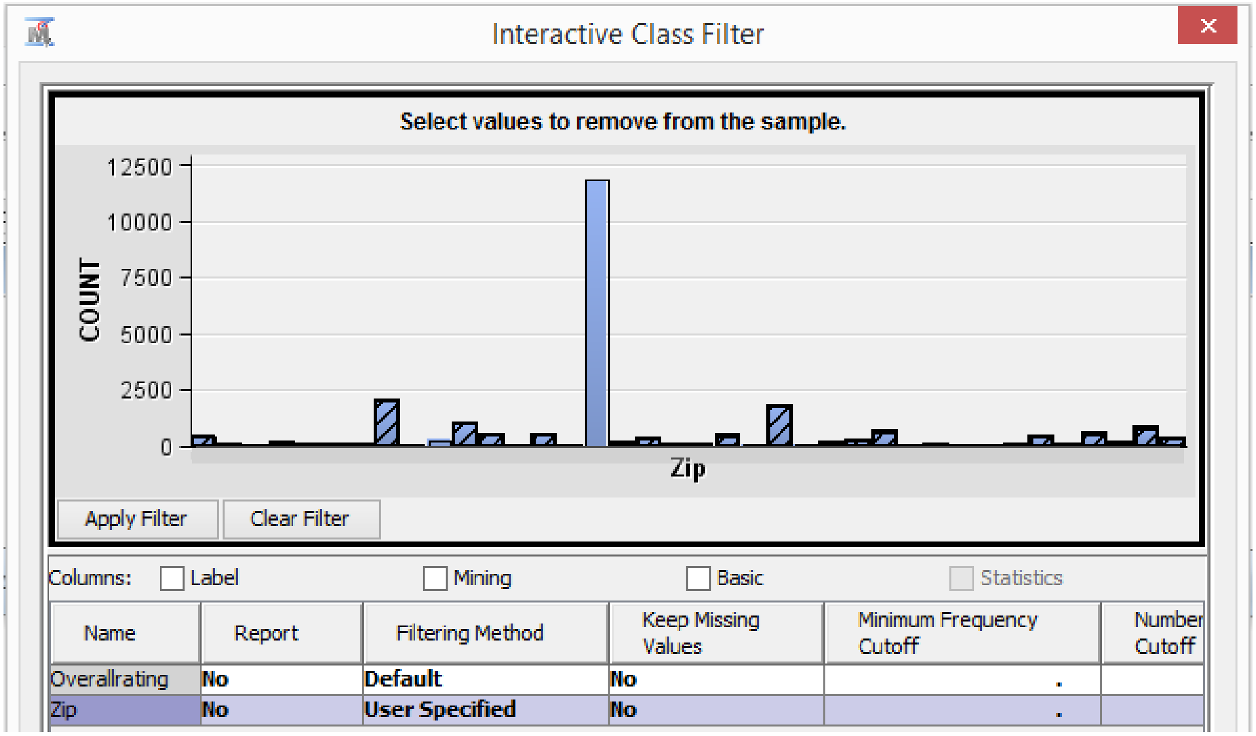
The below map of Las Vegas indicates a number of hotels geographically located in different areas by Zip code. Most significantly, we observed that approximately one third of the hotels in the city are located nearby Paradise (Zip 89109).



Proximity to a number of famous attractions is possibly one of the reason for many players establishing hotels in the above mentioned area. It is challenging for the hotels in this area to compete with each other given the location advantage neutralized for each of the hotel.

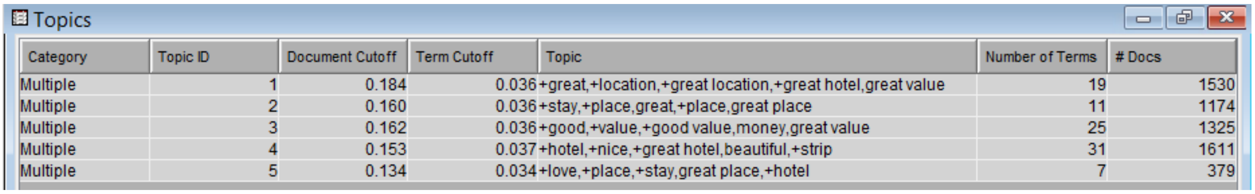
Let us understand in detail on what is perceived as good customer experience and what drives customers to choose a hotel in this particular area by analyzing the reviews provided on these hotels. Data is further decomposed to create a process for analyzing only these reviews and it reveals that the following text topics are most prominent in this area. The below process indicates usage of Filter node to select only those reviews that belongs to hotels in the particular zip code by using user specified criteria in the interactive filter view.





| Name | Report | Filtering Method | Keep Missing Values | Minimum Frequency Cutoff | Number Cutoff |
|---|---|---|---|---|---|
| Overallrating | No | Default | No | . | |
| Zip | No | User Specified | No | . | |

Selecting optimum topics in the text topic node by trying with different values results in the following list of most sought after topics of discussions by customers in the given area.

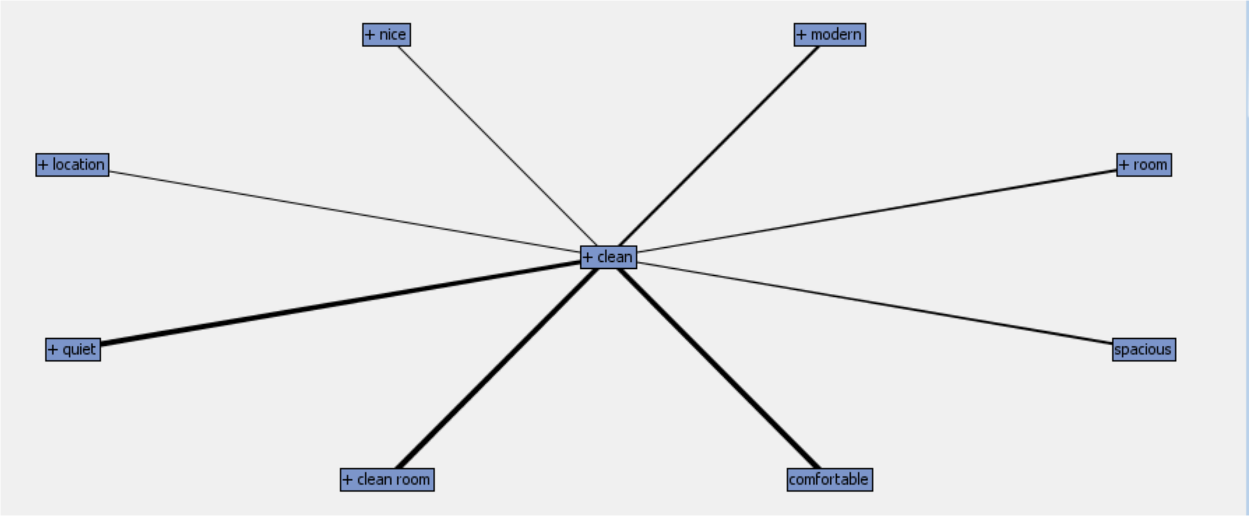


| Category | Topic ID | Document Cutoff | Term Cutoff | Topic | Number of Terms | # Docs |
|---|---|---|---|---|---|---|
| Multiple | 1 | 0.184 | 0.036 | +great,+location,+great location,+great hotel,great value | 19 | 1530 |
| Multiple | 2 | 0.160 | 0.036 | +stay,+place,great,+place,great place | 11 | 1174 |
| Multiple | 3 | 0.162 | 0.036 | +good,+value,+good value,money,great value | 25 | 1325 |
| Multiple | 4 | 0.153 | 0.037 | +hotel,+nice,+great hotel,beautiful,+strip | 31 | 1611 |
| Multiple | 5 | 0.134 | 0.034 | +love,+place,+stay,great place,+hotel | 7 | 379 |

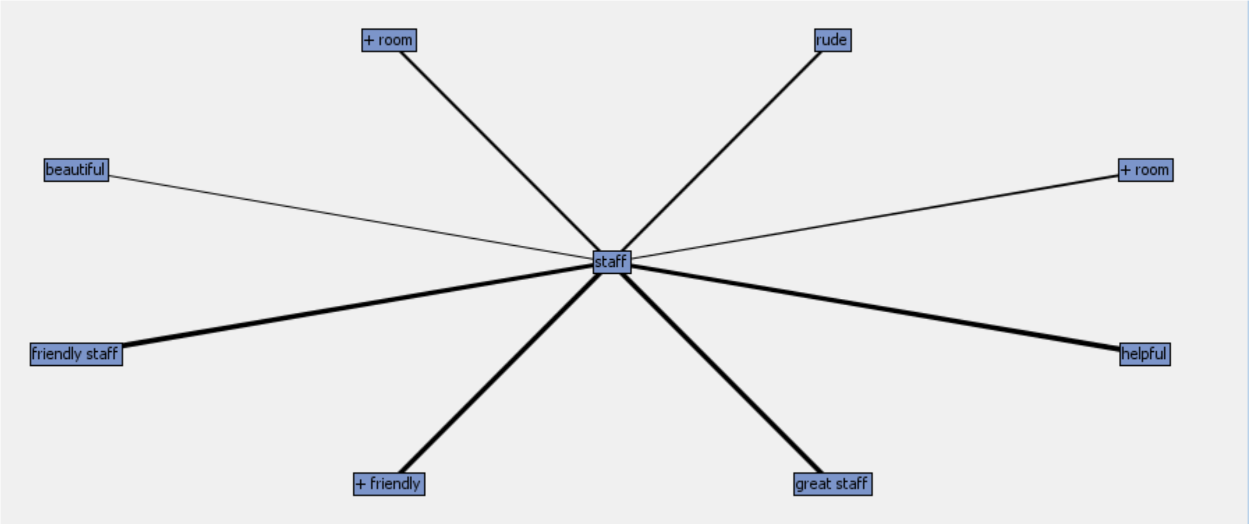
The text topics provides an indication of what sells good and what interests customers that drives them to choose a particular hotel. As we can see below, topics of interest in the best hotels in the selected area revolves around staying experience, service, price, value for money, ambience and location as we would expected it obviour given its promixity of area to a number of happening places.

### Services that interest customer in the best hotels in the chosen area

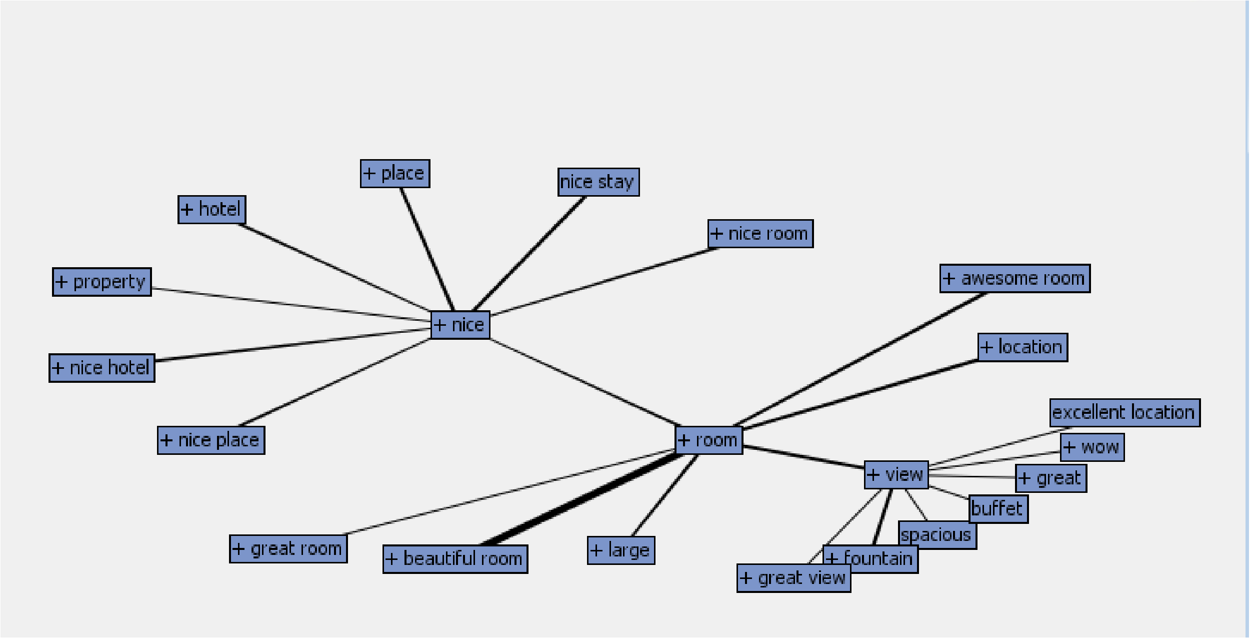
In order to specifically derive what are the most significant factors that drives customer to provide positive feedback on these hotels, we can evaluate the concept linking of top adjectives and nouns from the reviews provided on the best hotels of selected area. The below indicates that customers view on cleanliness on these hotels most significantly linked to room being clean, comfortability of surroundings and quite place.



Hotel management may be time and again interested to understand the view of customers on hotel staff and to take necessary actions if required to improve the quality of service provided by staff through staff training and improvement programs. The below concept linking indicates that staff in the best hotels in this area are friendly and helpful most often.

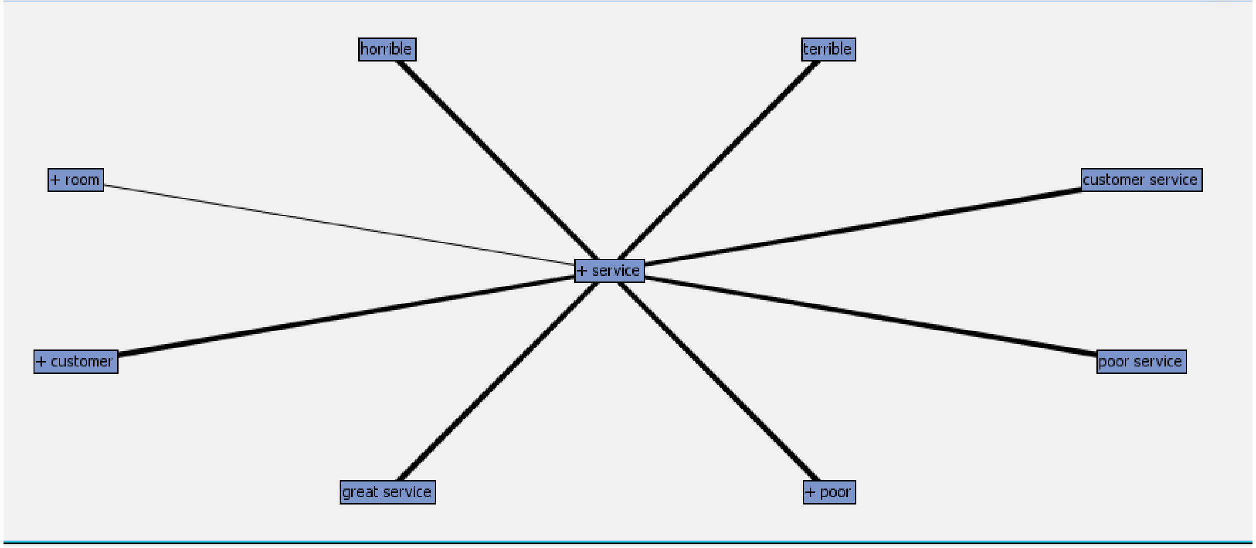


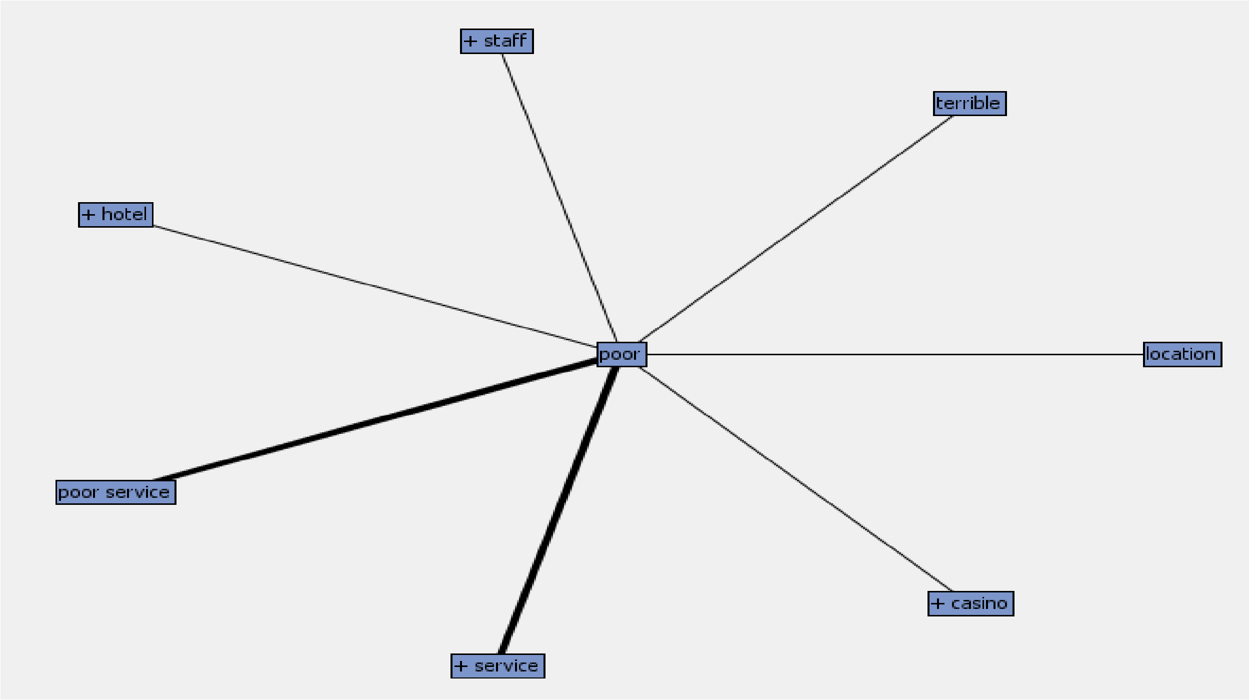
To further specifically understand how a particular customer service is functioning, the concept linking can be extended to further granularity as below. The positive feeling of customers about room, specifically indicates that customers are interested in the view of the room the view pertains to fountain, spacious and possibly great view of surroundings.

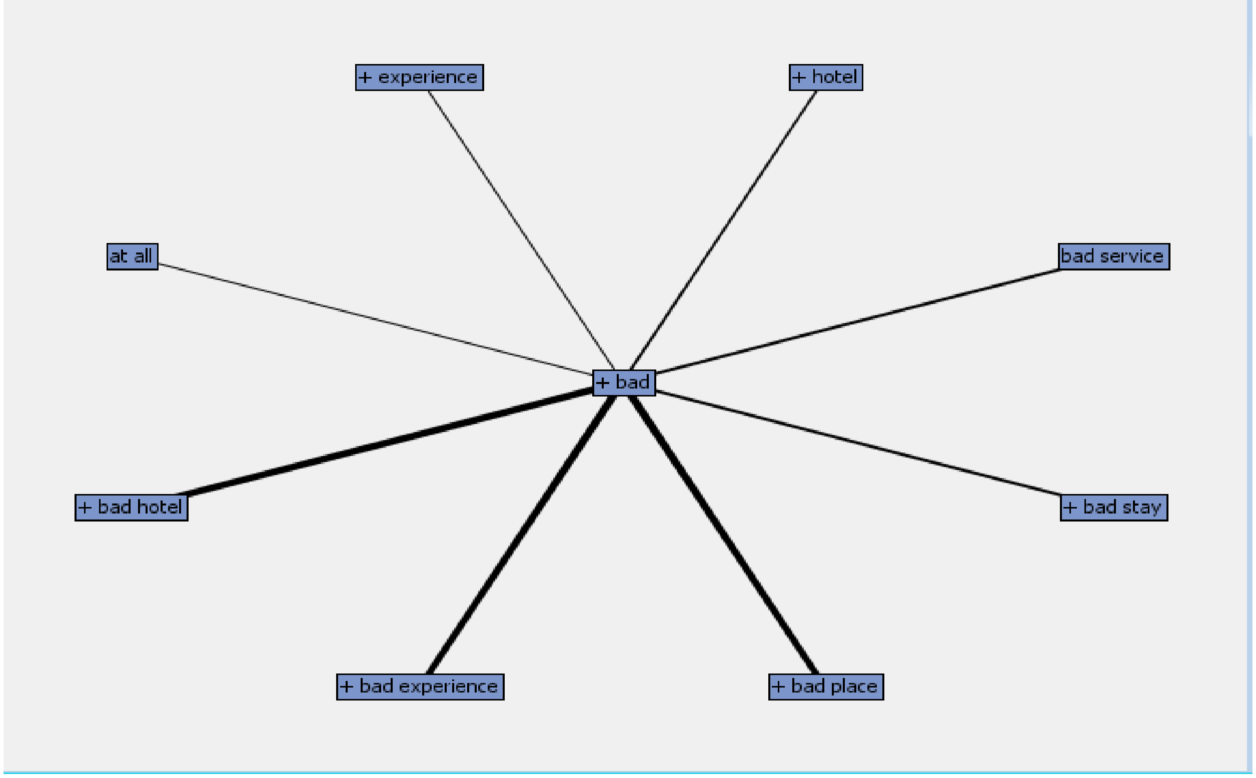


### Services that cause customer dissatisfaction in the chosen area

Choosing top adjectives and nouns from the poorly rated hotels reveals patterns on factors causing customer dissatisfaction and provides an opportunity for improvement initiatives for the hotel management.



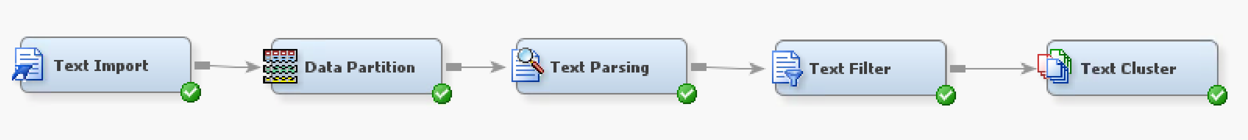

+ staff
terrible
+ hotel
poor
location
poor service
+ casino
+ service


+ experience
+ hotel
at all
bad service
+ bad
+ bad hotel
+ bad stay
+ bad experience
+ bad place

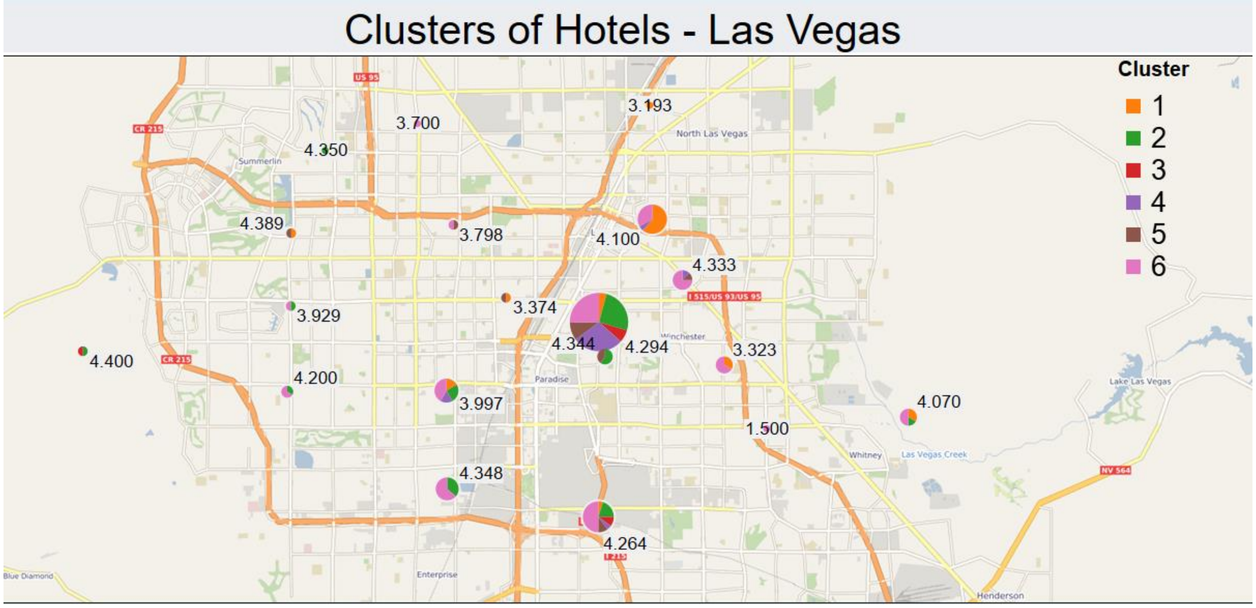
## 4.2 Text Clustering

Text clustering is a topic of interest to hospitality service providers in order to understand the cluster of hotels they belong to, their relative ranking in the market and who should they aim to compete with in order to maximize revenue and occupancy of the hotel

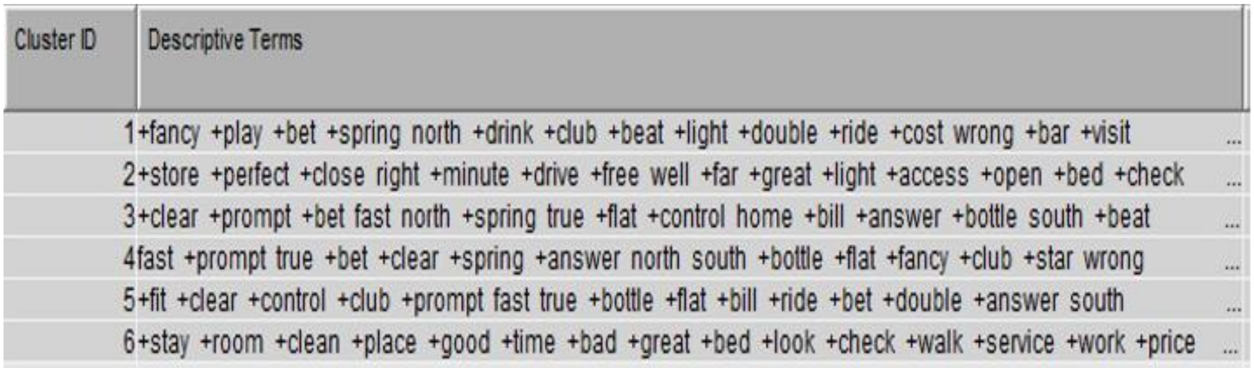
A text cluster process is created for all the hotels present in Las Vegas to understand the market dynamics. A data partition is created with 70% Training and 30% Validation in order to verify the optimum number of text clusters. Various combinations of text clusters have been tried before arriving at optimum clusters.



The below map indicates the clusters distributed across different geographies by Zip Code within Las Vegas based on the hotel reviews and the rating of the best cluster in a particular Zip area is indicated as a label.

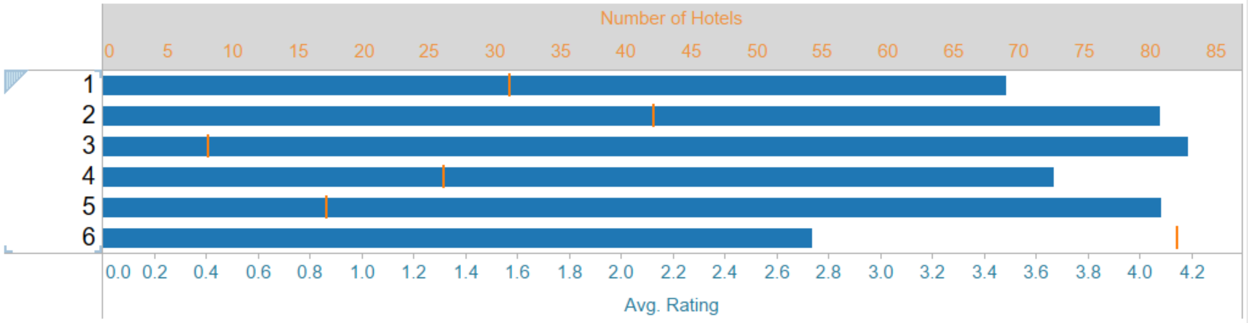


The clusters obtained are describing the following terms.



| Cluster ID | Descriptive Terms |
|---|---|
| 1 | +fancy +play +bet +spring north +drink +club +beat +light +double +ride +cost wrong +bar +visit ... |
| 2 | +store +perfect +close right +minute +drive +free well +far +great +light +access +open +bed +check ... |
| 3 | +clear +prompt +bet fast north +spring true +flat +control home +bill +answer +bottle south +beat ... |
| 4 | fast +prompt true +bet +clear +spring +answer north south +bottle +flat +fancy +club +star wrong ... |
| 5 | +fit +clear +control +club +prompt fast true +bottle +flat +bill +ride +bet +double +answer south ... |
| 6 | +stay +room +clean +place +good +time +bad +great +bed +look +check +walk +service +work +price ... |

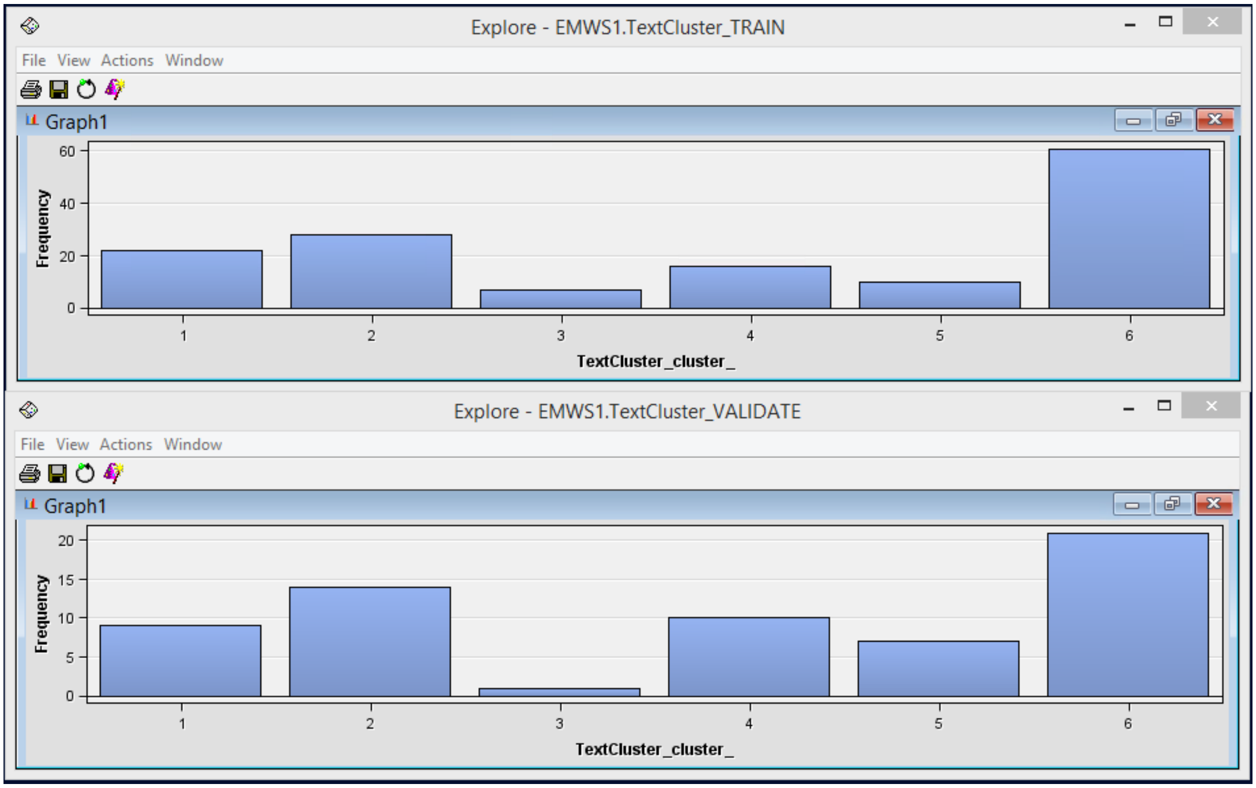
The Clusters with number of hotels and their average rating for each cluster is as shown below.



**Key Observations:**

- Cluster#6 contains terms that are related to Stay, room, price, cleanliness, look which are basic services expected in majority of the hotels.
- Cluster#6 has the most number of hotels and distributed in different areas but average rating is much less than many other hotels in the city.
- Clusters#2, #3 and #5 are relatively less in number of hotels but have much higher rating than other clusters.
- Predominantly clusters 2, 3 and 5 are discussing about easy access, drive, club, play and bar.
- Hotels in Cluster#6 can look to adopt to clusters 2,3 and 5 to improve their market share and relative rating.

To evaluate if the clusters generated by the above methodology is optimum, data is partitioned into 70% Training and 30% Validation and the cluster frequencies are roughly close for both training and validation data sets. This is an indication that for the given data, Los Vegas hotels can be distributed to 6 clusters and the players can concentrate on improving their share by making effort to move into better cluster than they are in.



## 4.3Text Analytics at Hotel Level

We analyzed the individual hotel reviews so that for hotel management could use them for their improvement.

Our team chose two hotels to explore in depth all the reviews.

**Shaftesbury Premier Paddington in London, UK**

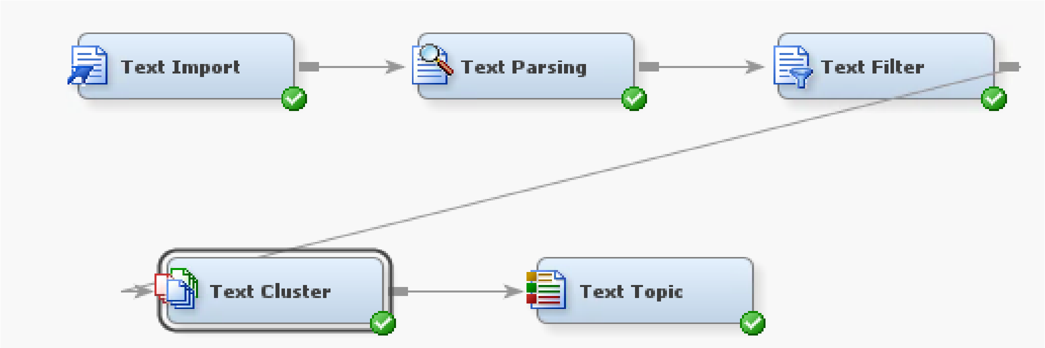
- This is three-star hotel.
- The data contain reviews of period from June 2004 to November 2009

**Kimpton Hotel in New York City, US**

- This is a four-star hotel
- The data contains December 2004 to November 2009

We applied the process flow listed below for the individual hotel Reviews.

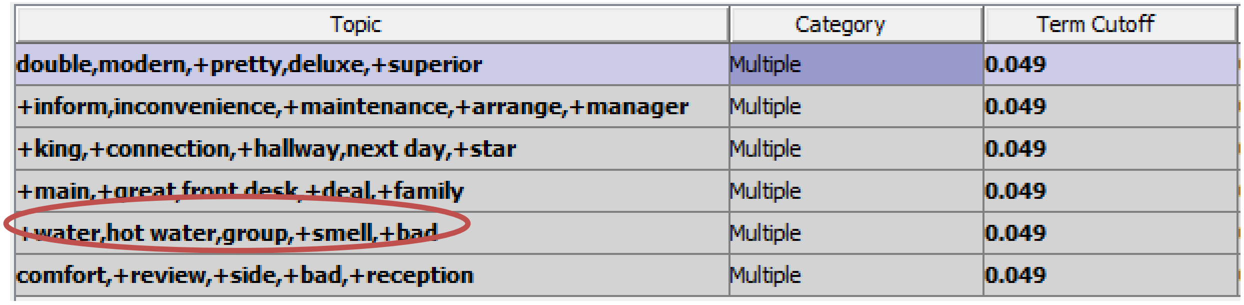
We have set term weight to IDF because of less number of documents and text Clustering with Max SVD dimensions 200 and multi-term topics to 5.



The reviews of the hotel of Shaftesbury reveal that there are both positive and negative reviews, and of 70 Park Avenue Kimpton reveals that the reviews are more positive, and there was more information about Amenities.

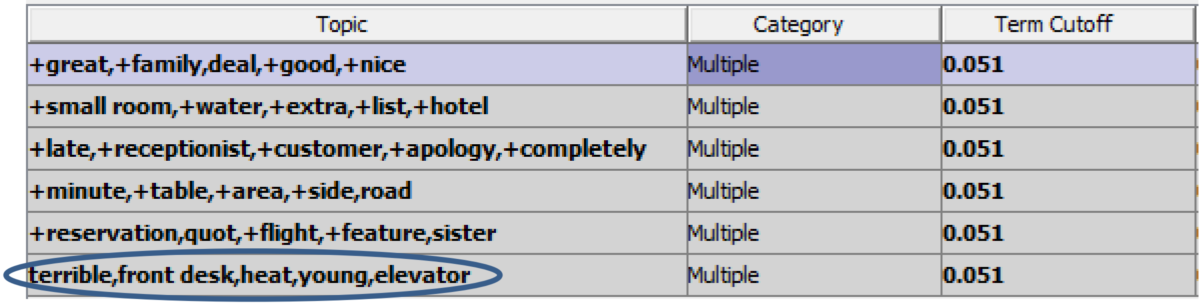
The first approach of analysis of the data set is to find seasonal variations of the reviews.To check if there are any seasonal factors affecting the customer reviews, we separated the data into monthly bases, April to September as Summer Reviews and reviews from October to March into Winter Reviews.

We have run the topic model listed earlier and derived topics for both seasons.



| Topic | Category | Term Cutoff |
|---|---|---|
| double,modern,+pretty,deluxe,+superior | Multiple | 0.049 |
| +inform,inconvenience,+maintenance,+arrange,+manager | Multiple | 0.049 |
| +king,+connection,+hallway,next day,+star | Multiple | 0.049 |
| +main,+great,front desk,+deal,+family | Multiple | 0.049 |
| +water,hot water,group,+smell,+bad | Multiple | 0.049 |
| comfort,+review,+side,+bad,+reception | Multiple | 0.049 |

a) **Shaftesbury Premier Paddington in London, UK**, **Summer Season**



| Topic | Category | Term Cutoff |
|---|---|---|
| +great,+family,deal,+good,+nice | Multiple | 0.051 |
| +small room,+water,+extra,+list,+hotel | Multiple | 0.051 |
| +late,+receptionist,+customer,+apology,+completely | Multiple | 0.051 |
| +minute,+table,+area,+side,road | Multiple | 0.051 |
| +reservation,quot,+flight,+feature,sister | Multiple | 0.051 |
| terrible,front desk,heat,young,elevator | Multiple | 0.051 |

**b) Shaftesbury Premier Paddington in London, UK, Winter Season**

The topics highlighted above shows that there are problems with water in summer and heat in winter. Close look at full reviews with **heat** term in topic viewer indicates that customers are facing issues with heater in winter season. So, the hotel management could use this information to improve the service as fixing the heater system before the winter season.

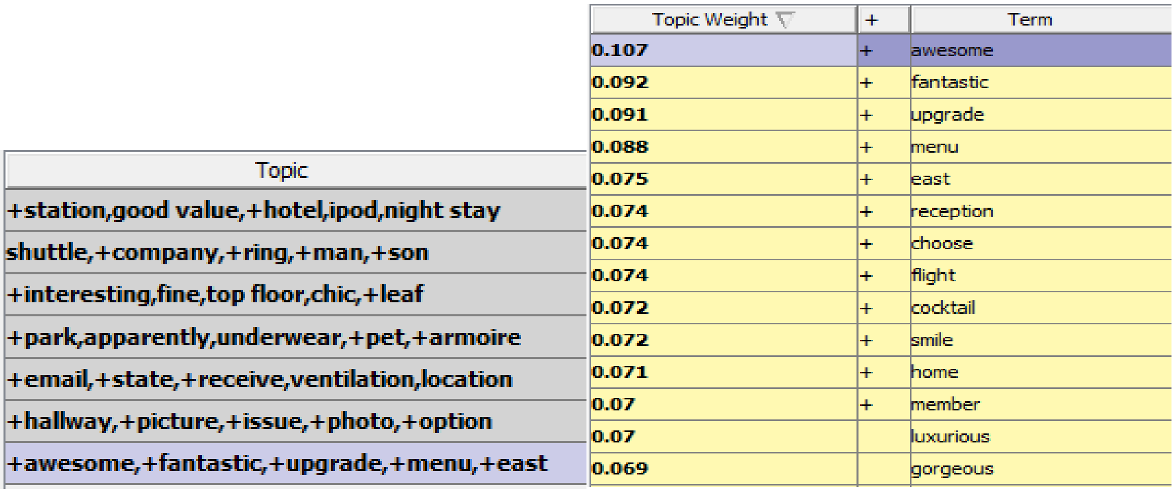
So, the hotel management should offer their service improvements considering seasonal changes.

Regarding hot water issues, we observed that no matter what the season was, customers are complaining that the water is either too hot or extremely cold.

So, the hotel manager should perform maintenance at regular intervals to keep water temperatures in control.

Another approach to look for feedbacks from customers for specific area in hotels. For example, if the hotel renovated bar in the hotel, text mining can be used to analyze the reviews for that specific area. To do this, we used reviews of 70 Park Avenue Kimpton Hotel. It is possible to use User Topic in the Text Topic node to set our own topics to see some specific areas. We wanted to see the reviews of a bar of the hotel.

From the data of 70 Park Avenue Kimpton Hotel that contain reviews from the period of 2008 and 2009, we used the same process as we used in the previous example. In the Text Topic node, the topics from the reviews of the hotel is shown below. Here, we didn't set the multi-terms as 5.



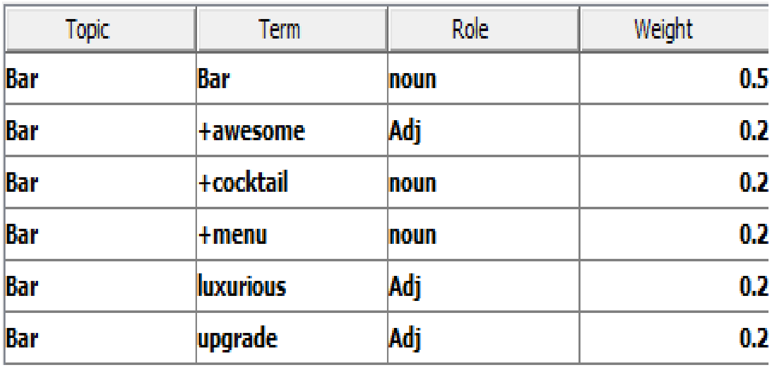
| Topic |
|---|
| +station,good value,+hotel,ipod,night stay |
| shuttle,+company,+ring,+man,+son |
| +interesting,fine,top floor,chic,+leaf |
| +park,apparently,underwear,+pet,+armoire |
| +email,+state,+receive,ventilation,location |
| +hallway,+picture,+issue,+photo,+option |
| +awesome,+fantastic,+upgrade,+menu,+east |

| Topic Weight | + | Term |
|---|---|---|
| 0.107 | + | awesome |
| 0.092 | + | fantastic |
| 0.091 | + | upgrade |
| 0.088 | + | menu |
| 0.075 | + | east |
| 0.074 | + | reception |
| 0.074 | + | choose |
| 0.074 | + | flight |
| 0.072 | + | cocktail |
| 0.072 | + | smile |
| 0.071 | + | home |
| 0.07 | + | member |
| 0.07 |  | luxurious |
| 0.069 |  | gorgeous |

**c) Topic Viewer of reviews of 70 Park Avenue a Kimpton Hotel in New York City, US**

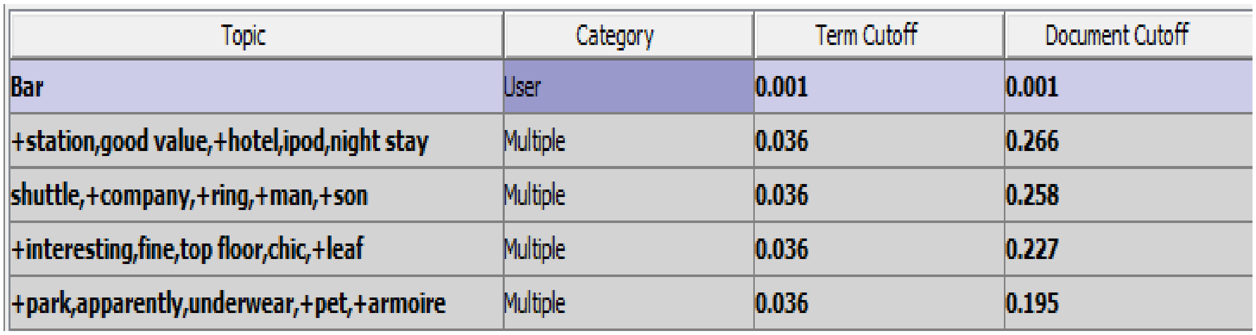
Among the topics, we thought that the last topic might be useful to set a user topic of "Bar."

We thought we could use terms "awesome," "cocktail," "menu," "luxurious," and "upgrade" to set up our own topic. In the User topic, we set up those terms for the topic.



| Topic | Term | Role | Weight |
|---|---|---|---|
| Bar | Bar | noun | 0.5 |
| Bar | +awesome | Adj | 0.2 |
| Bar | +cocktail | noun | 0.2 |
| Bar | +menu | noun | 0.2 |
| Bar | luxurious | Adj | 0.2 |
| Bar | upgrade | Adj | 0.2 |

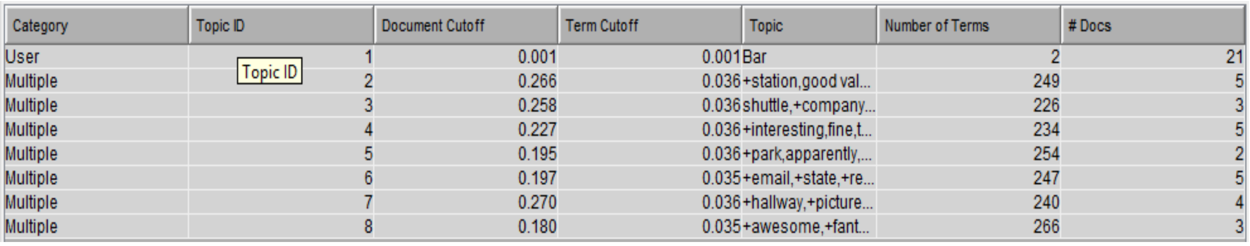
In the topic viewer, we could see "Bar" topic as shown below.



| Topic | Category | Term Cutoff | Document Cutoff |
|---|---|---|---|
| Bar | User | 0.001 | 0.001 |
| +station,good value,+hotel,ipod,night stay | Multiple | 0.036 | 0.266 |
| shuttle,+company,+ring,+man,+son | Multiple | 0.036 | 0.258 |
| +interesting,fine,top floor,chic,+leaf | Multiple | 0.036 | 0.227 |
| +park,apparently,underwear,+pet,+armoire | Multiple | 0.036 | 0.195 |

The User topic, "Bar," is assigned as a new topic.

The results of the Text Topic node are shown below.

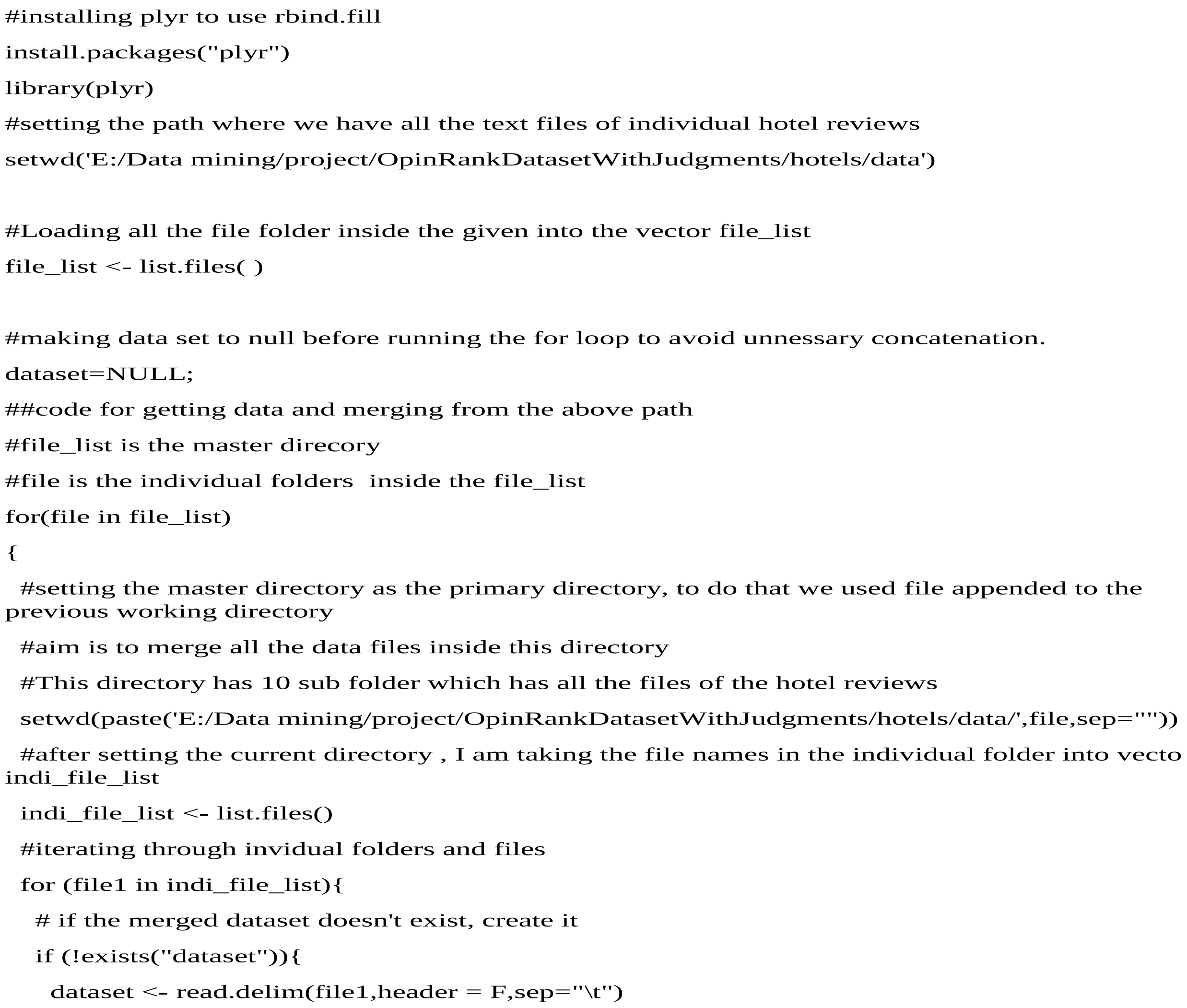


| Category | Topic ID | Document Cutoff | Term Cutoff | Topic | Number of Terms | # Docs |
|---|---|---|---|---|---|---|
| User | 1 | 0.001 | 0.001 | Bar | 2 | 21 |
| Multiple | 2 | 0.266 | 0.036 | +station,good val... | 249 | 5 |
| Multiple | 3 | 0.258 | 0.036 | shuttle,+company... | 226 | 3 |
| Multiple | 4 | 0.227 | 0.036 | +interesting,fine,t... | 234 | 5 |
| Multiple | 5 | 0.195 | 0.036 | +park,apparently,... | 254 | 2 |
| Multiple | 6 | 0.197 | 0.035 | +email,+state,+re... | 247 | 5 |
| Multiple | 7 | 0.270 | 0.036 | +hallway,+picture... | 240 | 4 |
| Multiple | 8 | 0.180 | 0.035 | +awesome,+fant... | 266 | 3 |

Here, since we didn't have a domain knowledge of the Bar of the hotel, and it's a hypothetical situation, we didn't explore the all the terms related to "Bar," the user topic didn't get high importance for each observation. However, hotel managers definitely can use his/her domain knowledge to set the User topic as needed and get some more useful insights from it. We are suggesting new application of text mining of the hotel reviews.

# 5 APPENDIX

To view the entire code, double click on the open document text provided below:



```
#installing plyr to use rbind.fill
install.packages("plyr")
library(plyr)
#setting the path where we have all the text files of individual hotel reviews
setwd('E:/Data mining/project/OpinRankDatasetWithJudgments/hotels/data')

#Loading all the file folder inside the given into the vector file_list
file_list <- list.files( )

#making data set to null before running the for loop to avoid unnessary concatenation.
dataset=NULL;
##code for getting data and merging from the above path
#file_list is the master direcory
#file is the individual folders  inside the file_list
for(file in file_list)
{
  #setting the master directory as the primary directory, to do that we used file appended to the
previous working directory
  #aim is to merge all the data files inside this directory
  #This directory has 10 sub folder which has all the files of the hotel reviews
  setwd(paste('E:/Data mining/project/OpinRankDatasetWithJudgments/hotels/data/',file,sep=""))
  #after setting the current directory , I am taking the file names in the individual folder into vecto
indi_file_list
  indi_file_list <- list.files()
  #iterating through invidual folders and files
  for (file1 in indi_file_list){
    # if the merged dataset doesn't exist, create it
    if (!exists("dataset")){
      dataset <- read.delim(file1,header = F,sep="\t")
```

# 6 REFERENCES

- N-Gram-Based Text Categorization By William B. Cavnar and John M. Trenkle - http://odur.let.rug.nl/~vannoord/TextCat/textcat.pdf
- LDA Visual from CRAN.
- Opinion Based Entity Ranking By Kavita Ganesan