Social Media Analytics for Business

Using Python, Tableau

Submitted by

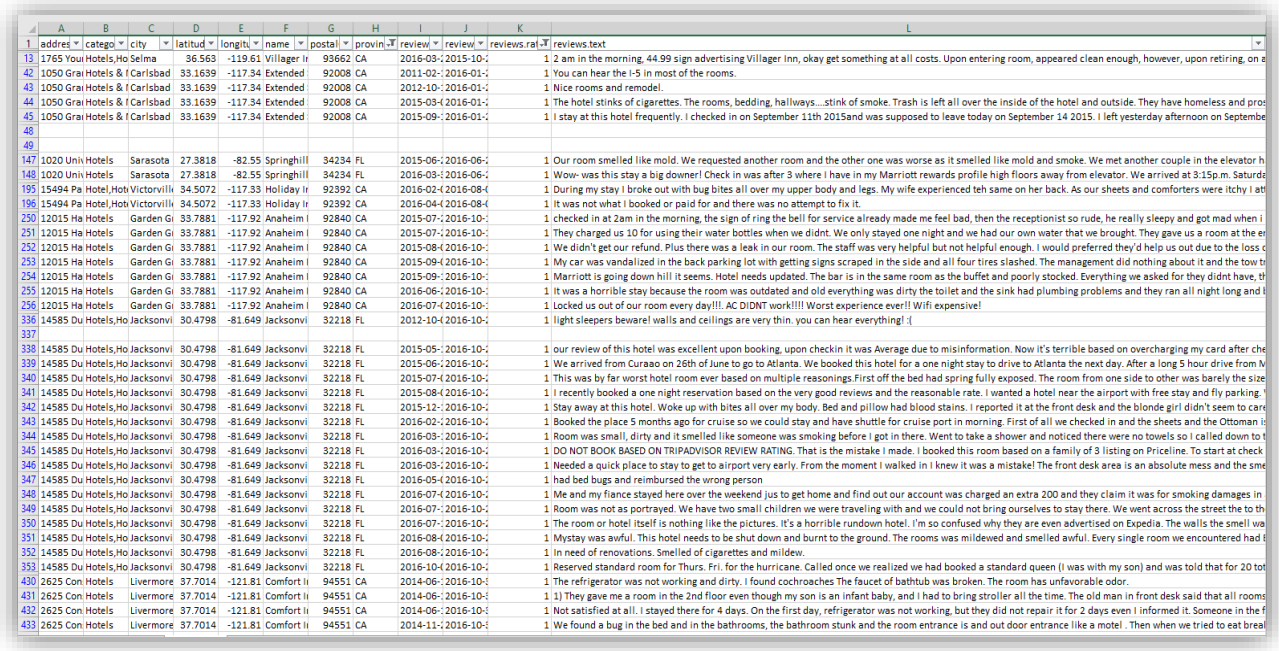
Venkata Sumanth Soma

Introduction:

The data which we have selected is from Data World

Dataset : https://data.world/datafiniti/hotel-reviews.

The data is related to Hotel reviews, ratings, name, location for all the hotels that are in the United States of America. The raw data consists of 35913 rows. In the process of data cleaning we have removed all the empty rows and selecting the review rating between 1 and 5. After the conversion we have deleted all the blank spaces using excel. Thus, resulting the data in 11443 rows.



	A	B	C	D	E	F	G	H	I	J	K	L
1	addres	catego	city	latitud	longitu	name	postal	provin	review	review	reviews.rat	reviews.text
13	1765 You	Hotels,Ho	Selma	36.563	-119.61	Villager Ir	93662	CA	2016-03-(	2015-10-(	1	2 am in the morning, 44.99 sign advertising Villager Inn, okay get something at all costs. Upon entering room, appeared clean enough, however, upon retiring, on a
42	1050 Gra	Hotels & (	Carlsbad	33.1639	-117.34	Extended	92008	CA	2011-02-(	2016-01-(	1	You can hear the I-5 in most of the rooms.
43	1050 Gra	Hotels & (	Carlsbad	33.1639	-117.34	Extended	92008	CA	2012-10-(	2016-01-(	1	Nice rooms and remodel.
44	1050 Gra	Hotels & (	Carlsbad	33.1639	-117.34	Extended	92008	CA	2015-03-(	2016-01-(	1	The hotel stinks of cigarettes. The rooms, bedding, hallways....stink of smoke. Trash is left all over the inside of the hotel and outside. They have homeless and pros
45	1050 Gra	Hotels & (	Carlsbad	33.1639	-117.34	Extended	92008	CA	2015-09-(	2016-01-(	1	I stay at this hotel frequently. I checked in on September 11th 2015and was supposed to leave today on September 14 2015. I left yesterday afternoon on Septembe
48												
49												
147	1020 Uni	Hotels	Sarasota	27.3818	-82.55	Springhill	34234	FL	2015-06-(	2016-06-(	1	Our room smelled like mold. We requested another room and the other one was worse as it smelled like mold and smoke. We met another couple in the elevator h
148	1020 Uni	Hotels	Sarasota	27.3818	-82.55	Springhill	34234	FL	2016-03-(	2016-06-(	1	Wow- was this stay a big downer! Check in was after 3 where I have in my Marriott rewards profile high floors away from elevator. We arrived at 3:15p.m. Saturda
195	15494 Pa	Hotel,Hot	Victorville	34.5072	-117.33	Holiday Ir	92392	CA	2016-02-(	2016-08-(	1	During my stay I broke out with bug bites all over my upper body and legs. My wife experienced teh same on her back. As our sheets and comforters were itchy I att
196	15494 Pa	Hotel,Hot	Victorville	34.5072	-117.33	Holiday Ir	92392	CA	2016-04-(	2016-08-(	1	It was not what I booked or paid for and there was no attempt to fix it.
250	12015 Ha	Hotels	Garden Gi	33.7881	-117.92	Anaheim I	92840	CA	2015-07-(	2016-10-(	1	checked in at 2am in the morning, the sign of ring the bell for service already made me feel bad, then the receptionist so rude, he really sleepy and got mad when i
251	12015 Ha	Hotels	Garden Gi	33.7881	-117.92	Anaheim I	92840	CA	2015-07-(	2016-10-(	1	They charged us 10 for using their water bottles when we didnt. We only stayed one night and we had our own water that we brought. They gave us a room at the er
252	12015 Ha	Hotels	Garden Gi	33.7881	-117.92	Anaheim I	92840	CA	2015-08-(	2016-10-(	1	We didn't get our refund. Plus there was a leak in our room. The staff was very helpful but not helpful enough. I would preferred they'd help us out due to the loss c
253	12015 Ha	Hotels	Garden Gi	33.7881	-117.92	Anaheim I	92840	CA	2015-09-(	2016-10-(	1	My car was vandalized in the back parking lot with getting signs scraped in the side and all four tires slashed. The management did nothing about it and the tow tr
254	12015 Ha	Hotels	Garden Gi	33.7881	-117.92	Anaheim I	92840	CA	2015-09-(	2016-10-(	1	Marriott is going down hill it seems. Hotel needs updated. The bar is in the same room as the buffet and poorly stocked. Everything we asked for they didnt have, th
255	12015 Ha	Hotels	Garden Gi	33.7881	-117.92	Anaheim I	92840	CA	2016-06-(	2016-10-(	1	It was a horrible stay because the room was outdated and old everything was dirty the toilet and the sink had plumbing problems and they ran all night long and l
256	12015 Ha	Hotels	Garden Gi	33.7881	-117.92	Anaheim I	92840	CA	2016-07-(	2016-10-(	1	Locked us out of our room every day!!!. AC DIDNT work!!!! Worst experience ever!! Wifi expensive!
336	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2012-10-(	2016-10-(	1	light sleepers beware! walls and ceilings are very thin. you can hear everything! :(
337												
338	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2015-05-(	2016-10-(	1	our review of this hotel was excellent upon booking, upon checkin it was Average due to misinformation. Now it's terrible based on overcharging my card after che
339	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2015-06-(	2016-10-(	1	We arrived from Curaao on 26th of June to go to Atlanta. We booked this hotel for a one night stay to drive to Atlanta the next day. After a long 5 hour drive from M
340	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2015-07-(	2016-10-(	1	This was by far worst hotel room ever based on multiple reasonings.First off the bed had spring fully exposed. The room from one side to other was barely the size
341	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2015-08-(	2016-10-(	1	I recently booked a one night reservation based on the very good reviews and the reasonable rate. I wanted a hotel near the airport with free stay and fly parking.
342	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2015-12-(	2016-10-(	1	Stay away at this hotel. Woke up with bites all over my body. Bed and pillow had blood stains. I reported it at the front desk and the blonde girl didn't seem to care
343	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-02-(	2016-10-(	1	Booked the place 5 months ago for cruise so we could stay and have shuttle for cruise port in morning. First of all we checked in and the sheets and the Ottoman is
344	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-03-(	2016-10-(	1	Room was small, dirty and it smelled like someone was smoking before I got in there. Went to take a shower and noticed there were no towels so I called down to t
345	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-03-(	2016-10-(	1	DO NOT BOOK BASED ON TRIPADVISOR REVIEW RATING. That is the mistake I made. I booked this room based on a family of 3 listing on Priceline. To start at check
346	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-03-(	2016-10-(	1	Needed a quick place to stay to get to airport very early. From the moment I walked in I knew it was a mistake! The front desk area is an absolute mess and the sme
347	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-05-(	2016-10-(	1	had bed bugs and reimbursed the wrong person
348	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-07-(	2016-10-(	1	Me and my fiance stayed here over the weekend jus to get home and find out our account was charged an extra 200 and they claim it was for smoking damages in
349	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-07-(	2016-10-(	1	Room was not as portrayed. We have two small children we were traveling with and we could not bring ourselves to stay there. We went across the street the to th
350	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-07-(	2016-10-(	1	The room or hotel itself is nothing like the pictures. It's a horrible rundown hotel. I'm so confused why they are even advertised on Expedia. The walls the smell wa
351	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-08-(	2016-10-(	1	Mystay was awful. This hotel needs to be shut down and burnt to the ground. The rooms was mildewed and smelled awful. Every single room we encountered had f
352	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-08-(	2016-10-(	1	In need of renovations. Smelled of cigarettes and mildew.
353	14585 Du	Hotels,Ho	Jacksonvi	30.4798	-81.649	Jacksonvi	32218	FL	2016-10-(	2016-10-(	1	Reserved standard room for Thurs. Fri. for the hurricane. Called once we realized we had booked a standard queen (I was with my son) and was told that for 20 tot
430	2625 Con	Hotels	Livermore	37.7014	-121.81	Comfort Ir	94551	CA	2014-06-(	2016-10-(	1	The refrigerator was not working and dirty. I found cochroaches The faucet of bathtub was broken. The room has unfavorable odor.
431	2625 Con	Hotels	Livermore	37.7014	-121.81	Comfort Ir	94551	CA	2014-06-(	2016-10-(	1	1) They gave me a room in the 2nd floor even though my son is an infant baby, and I had to bring stroller all the time. The old man in front desk said that all rooms
432	2625 Con	Hotels	Livermore	37.7014	-121.81	Comfort Ir	94551	CA	2014-06-(	2016-10-(	1	Not satisfied at all. I stayed there for 4 days. On the first day, refrigerator was not working, but they did not repair it for 2 days even I informed it. Someone in the f
433	2625 Con	Hotels	Livermore	37.7014	-121.81	Comfort Ir	94551	CA	2014-11-(	2016-10-(	1	We found a bug in the bed and in the bathrooms, the bathroom stunk and the room entrance is and out door entrance like a motel . Then when we tried to eat breal

Figure 1

Summary:

Then Dataset contains information about list of hotels, motels with location and reviews from the customers, main idea in analyzing the dataset is to find out the different aspects to improve business and customer satisfaction. For the analysis we use Sentiment Analysis and Topic modeling techniques.

We as a company **"Reborn Limited"** going forward to improve business of hotels/motels who are struggling to gain competitive advantage.

Data Visualization:

As part of this project the first activity that we are going to perform is Data visualization and here we used **"Tableau"** as Data visualization tool, the main idea in performing data visualization is to find the trends and gain knowledge what exactly data wants to talk about. We have performed data visualization for the hotel reviews datasets by taking Province(state) and count of review ratings to find out the list of states having review rating between 0 and 3.5.Please find the analysis in the below screenshot.

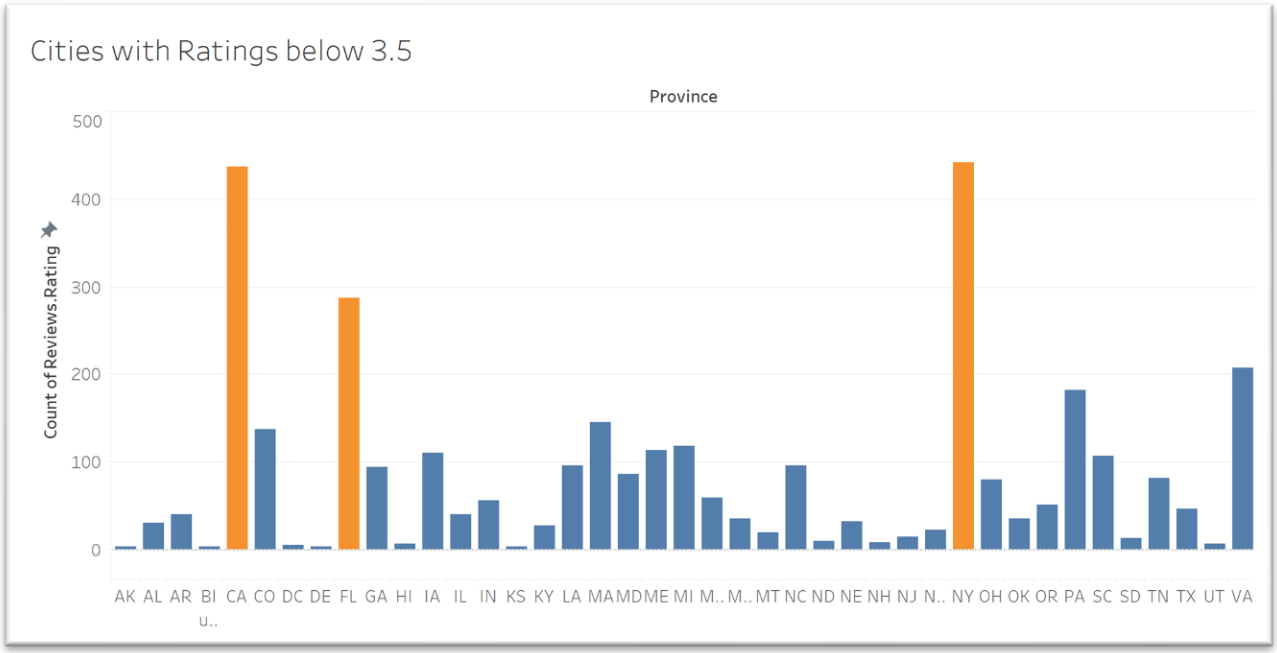


Figure 2

After performing the task, we came across the list of states with the desired output. California, Florida, New York are the states we are going to focus from the dataset for the further analysis.

Main idea in focusing is to find out if these three states have more negative or positive sentiment from the sentiment analysis.

Sentiment Analysis:

As we see we got three states that have the highest negative reviews from the data visualization. The states are California, Florida, New York. Now we are going to calculate the sentient values for each review in the dataset. The data was reduced to 1100 rows after the visualization. Then the data is separated in the respective states.

To calculate the sentiment analysis values, we upload the datasets into PowerBI. The main reason to calculate sentiment analysis is a procedure used to determine if a chunk of text is positive, negative, or neutral. In-text analytics, natural language processing (NLP) and machine learning (ML) techniques are combined to assign sentiment scores to the topics, categories ,or entities within a phrase. Sentiment analysis mainly focuses on opinions that express or imply positive or negative sentiments, also called positive or negative opinions in everyday language.

In the sentiment analysis, the text reviews in English will be converted to numerical values that lie between the values 0 and 1. Considering values between 0 and 0.5 are negative reviews, 0.5 as neutral review, values between 0.5 and 1 are positive reviews.

Here are the screenshots of the PowerBI and the percentages of reviews of each state.

California sentiment analysis values

id	text	Query1.documents.score
2	You can hear the I-5 in most of the rooms.	0.5
3	Nice rooms and remodel.	0.954107881
4	The hotel stinks of cigarettes. The rooms, bedding, hallway	0.02177152
5	I stay at this hotel frequently. I checked in on September 1	0.5
6	During my stay I broke out with bug bites all over my uppe	0.025831014
7	It was not what I booked or paid for and there was no atte	0.158797979
8	checked in at 2am in the morning, the sign of ring the bell	0.020077974
9	They charged us 10 for using their water bottles when we	0.5
10	We didn't get our refund. Plus there was a leak in our roo	0.080751836
11	My car was vandalized in the back parking lot with getting	0.145006657
12	Marriott is going down hill it seems. Hotel needs updated.	0.00475198
13	It was a horrible stay because the room was outdated and	0.006505489
14	Locked us out of our room every day!!!. AC DIDNT work!!!	0.012242407
15	The refrigerator was not working and dirty. I found cochro	0.002046525
16	1) They gave me a room in the 2nd floor even though my s	0.006786644
17	Not satisfied at all. I stayed there for 4 days. On the first da	0.02127409
18	We found a bug in the bed and in the bathrooms, the bath	0.197116971
19	Within minutes of getting into our room we found animal	0.154398769
20	Don't stay here.	0.216487437
21	This place is pretty shady.. in my room now and a lil bit sca	0.020521551
22	This place is minimal but just fine!! It's by everything!! Arb	0.839372039
23	You might as well put a sleeping bag in the middle of the h	0.093276471
24	Walk in to room and smells like smoke yet there's a sign st	0.084033877
25	Don't stay here. It was horrible. Everything was old. The w	0.5

Figure 3

negative reviews percentage = 249/437	neutral reviews = 30/437	positive reviews = 158/437
57.11	6.8	36.09

Figure 4

Florida sentiment analysis values

id	text	Query1.documents.score
2	Wow- was this stay a big downer! Check in was after 3 whe	0.175834298
3	light sleepers beware! walls and ceilings are very thin. you	0.098032415
4	our review of this hotel was excellent upon booking, upon	0.085700989
5	We arrived from Curaao on 26th of June to go to Atlanta. \	0.5
6	This was by far worst hotel room ever based on multiple re	0.140480042
7	I recently booked a one night reservation based on the ver	0.230632693
8	Stay away at this hotel. Woke up with bites all over my bo	0.012183577
9	Booked the place 5 months ago for cruise so we could stay	0.134000629
10	Room was small, dirty and it smelled like someone was sm	0.272049546
11	DO NOT BOOK BASED ON TRIPADVISOR REVIEW RATING. T	0.28738001
12	Needed a quick place to stay to get to airport very early. Fr	0.018645972
13	had bed bugs and reimbursed the wrong person	0.055129051
14	Me and my fiance stayed here over the weekend jus to get	0.085361838
15	Room was not as portrayed. We have two small children w	0.788487792
16	The room or hotel itself is nothing like the pictures. It's a h	0.002921939
17	Mystay was awful. This hotel needs to be shut down and b	0.001330554
18	In need of renovations. Smelled of cigarettes and mildew.	0.099950135
19	Reserved standard room for Thurs. Fri. for the hurricane. C	0.156131983
20	I made the reservation and when I was there doing the che	0.114194244
21	Very bad, rooms were NOT clean at all, for the price i paid	0.097188562
22	We called ahead to request an early check-in as we had so	0.5
23	The front desk was great but the cleaning crew was horrib	0.040712267
24	When we arrived Florida, we were supposed to stay at the	0.235128701
25	The room was very spacious but needs some updating and	0.883118391
26	We ended up going to a different hotel, that is how dirty th	0.075453609
27	i found roaches in the room the a/c leaking inside the roon	0.059053749

Figure 5

positive =82/280	neutral= 19/280	negative= 179/280
29.28	6.79	63.93

Figure 6

NEW YORK SENTIMENT ANALYSIS VALUES

id	text	Query1.documents.score
2	Kitchen clean. common room very small. Couch chair very	0.775883615
3	PLEASE READ VERY DIRTY POOR HOUSEKEEPING SMELLS	0.111201644
4	Bed was not clean, dirt and lots of hair in between shhets.	0.010977983
5	The hotel was adequate, but the room smelled like urine,	0.114179522
6	Issues with telephone, AC, smell. No action taken care des	0.111573547
7	Our first room was freezing despite the heat being on full	0.190731257
8	Finding bedbugs crawling all over the pillows and then all	0.035039812
9	remote missing in room..clerk could not give us another o	0.024232149
10	Hotel has a very strange odor. Rooms are quite dated and	0.001420021
11	we stayed here because the price was right and it advertis	0.227721691
12	I was traveling with my grandchildren, so I choose this hot	0.057146877
13	Fire alarm went off 4 times in one night, 10:00PM until ab	0.090979964
14	To Start the visit with this hotel, we checked in and were d	0.021210968
15	This is a rundown hotel with a sad excuse for a breakfast, f	0.022505224
16	DO NOT GO THERE.THIS PLACE IS OLD AND TIRED. I KNEW	0.128755569
17	This Hotel is VERY rundown. Room was ok for a 1 night sta	0.021590948
18	Upon entrance to our room we were greeted by the smell	0.191323876
19	I wrote it initially	0.788033247
20	We are not picky people, having traveled repeatedly to thi	0.023329586
21	reeks of smoke throughout the entire hotel except for the	0.242364675
22	Not a good experience. The heater was old and beat up;	0.150637627
23	Not a five star hotel but cheap for a weekend getaway	0.720814764
24	the only reason i use this hotel is the location too bad they	0.763093472
25	This hotel was cheaply priced, Ideal for a 1 night stop-over	0.038160533

Figure 7

positive = 106/416	neutral = 24/416	negative =286/416
25.48	5.76	68.76

Figure 8

As we are focusing more on negative reviews, the number of negative reviews is more than the positive reviews. The reviews from California have 57% of negative reviews. The reviews from Florida have 64% of negative reviews. The reviews from New York have 69% of negative reviews.

Topic Modeling:

The next model we are going to perform as part of this project to find out the reasons for getting more negative reviews. In sentiment analysis we have seen that in every state that we performed analysis we got an average of 60% negative sentiment reviews and by executing this model we are going to find out what are reasons by scanning a set of document(here its review document) detecting words and phrase patters.

First steps in starting topic modeling require a set of prerequisites which is preprocessing of data for the model to work effectively. Preprocessing include removing special characters, blank spaces, numbers, Tokenization, Stemming and lemmatization. For this process to occur we have used python 3 programming using jupyter Notebook platform.

Second Step is running LDA model which is one of the NLP on the clean processed data and detecting phrase patterns and words and the frequency of words, the result can be plotted using visualization library of LDA that we have which is ***"pyLDAvis"*** .

Please find the complete code in the end of the document and the analysis is as below in Figure 9 & 10

```
#LDA
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = 5, id2word=dictionary, passes=40)
ldamodel.save('model_combined.gensim')
topics = ldamodel.print_topics(num_words=10)
for topic in topics:
    print(topic)
```

```
(0, '0.012*"room" + 0.011*"good" + 0.008*"stay" + 0.008*"hotel" + 0.007*"breakfast" + 0.007*"locat" + 0.006*"work" + 0.006*"price" + 0.006*"conveni" + 0.006*"nice"')
(1, '0.051*"room" + 0.027*"hotel" + 0.020*"stay" + 0.016*"night" + 0.010*"smell" + 0.010*"door" + 0.010*"clean" + 0.008*"breakfast" + 0.007*"need" + 0.007*"like"')
(2, '0.024*"room" + 0.014*"hotel" + 0.013*"check" + 0.011*"staff" + 0.009*"like" + 0.008*"smell" + 0.008*"help" + 0.008*"would" + 0.007*"stay" + 0.007*"need"')
(3, '0.025*"hotel" + 0.023*"stay" + 0.017*"room" + 0.011*"check" + 0.008*"place" + 0.008*"cruis" + 0.008*"book" + 0.007*"would" + 0.007*"breakfast" + 0.006*"rate"')
(4, '0.034*"room" + 0.031*"hotel" + 0.020*"clean" + 0.018*"staff" + 0.015*"good" + 0.015*"breakfast" + 0.013*"stay" + 0.011*"locat" + 0.010*"nice" + 0.010*"friendli"')
```

Figure 9: LDA model

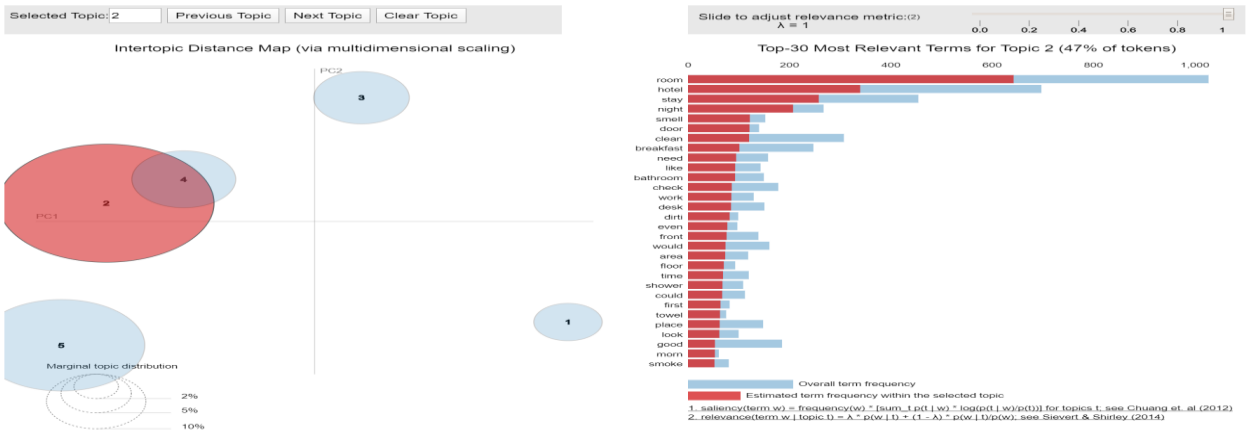


Figure 10: pyLDAvis

Conclusion:

From the above analysis we performed it is clearly seen that every step performed makes lot of sense in executing order as mentioned below:

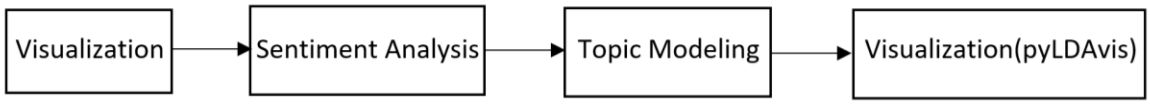


Figure 11: Execution order Map

Step1: To find the trends and understanding the hotel reviews data (Performed using Tableau).

Step2: To find Sentiment Analysis from each review if its Positive/Neutral/Negative.

Step3: To detect the phrase pattern and words in the review corpus.

Step4: Analyzing all the activities performed above and plot to make it easy to understand to all the audience.

From all the tasks performed our analysis comes to one point in business point of view which is that in states like CA,FL,NY there are more reviews compared to others and there are more negative reviews and what does there negative reviews talk about and analysis talks about that room are not clean and smell so bad that is impacting customer satisfaction from the Topic 1& 2 in topic modeling it is clearly seen that smell, dirty are the words focused more and impacting the business.

As a company “Reborn Limited” we are here to give our recommendation to Hotels in CA, FL and NY to make sure rooms are clean and aromatic prior to customers check-in into rooms. From the analysis we also observed that hotel are providing clean breakfast and staff are being very helpful but not keeping rooms clean which makes them smell bad when customers step into the rooms makes a very bad impression for the customer which is the main reason for the hotel in CA, NY and FL to get lots of negative review and this will help improve their business overall and gain competitive advantage.

Topic Modeling Python Code:

```
In [1]:
import pandas as pd
import numpy as np
import nltk
import string
from nltk.corpus import stopwords
from nltk.tokenize import RegexpTokenizer
from nltk.stem import WordNetLemmatizer
from nltk.stem.porter import PorterStemmer
```

```
In [2]:
reviews_df=pd.read_csv('CAFLNY.csv',error_bad_lines=False)
reviews_df.head(10)
```

Out[2]:

	Reviews
0	2 am in the morning, 44.99 sign advertising Vi...
1	You can hear the I-5 in most of the rooms.
2	Nice rooms and remodel.
3	The hotel stinks of cigarettes. The rooms, bed...
4	I stay at this hotel frequently. I checked in ...
5	chambre pas propre cafetiÃ¨re salle .....pas r...
6	Il semble que la femme de mÃ©nage utilise le m...
7	Our room smelled like mold. We requested anoth...
8	Wow- was this stay a big downer! Check in was ...
9	During my stay I broke out with bug bites all ...

```
In [3]:
reviews_df = reviews_df.astype(str)
#reviews_df[=reviews_df['reviews.text']
reviews_df['Token_Reviews']=reviews_df
reviews_df['Token_Reviews'].head(3)
```

```
Out[3]: 0    2 am in the morning, 44.99 sign advertising Vi...
1            You can hear the I-5 in most of the rooms.
2                                  Nice rooms and remodel.
Name: Token_Reviews, dtype: object
```

Remove Special Characters and Numbers:

```
In [4]:
reviews_df['Token_Reviews'] = reviews_df['Token_Reviews'].str.replace("[^a-zA-Z#]"," ")
reviews_df['Token_Reviews'].head(10)
```

```
Out[4]: 0      am in the morning        sign advertising Vi...
1           You can hear the I   in most of the rooms
2                              Nice rooms and remodel
3    The hotel stinks of cigarettes  The rooms  bed...
4    I stay at this hotel frequently  I checked in ...
5    chambre pas propre cafeti  re salle      pas r...
6    Il semble que la femme de m  nage utilise le m...
7    Our room smelled like mold  We requested anoth...
8    Wow  was this stay a big downer  Check in was ...
9    During my stay I broke out with bug bites all ...
Name: Token_Reviews, dtype: object
```

Tokenization:

```
In [5]:
tokenizer = RegexpTokenizer(r'\w+')
```

```
In [6]:
reviews_df['Token_Reviews'] = reviews_df['Token_Reviews'].apply(lambda x: tokenizer.tokenize(x.lower()))
```

```
In [7]:
reviews_df['Token_Reviews'].head(10)
```

```
Out[7]: 0    [am, in, the, morning, sign, advertising, vill...
1    [you, can, hear, the, i, in, most, of, the, ro...
2                        [nice, rooms, and, remodel]
3    [the, hotel, stinks, of, cigarettes, the, room...
4    [i, stay, at, this, hotel, frequently, i, chec...
5    [chambre, pas, propre, cafeti, re, salle, pas,...
6    [il, semble, que, la, femme, de, m, nage, util...
7    [our, room, smelled, like, mold, we, requested...
8    [wow, was, this, stay, a, big, downer, check, ...
9    [during, my, stay, i, broke, out, with, bug, b...
Name: Token_Reviews, dtype: object
```

Remove Stop and Short(Words length less than 3) words:

```
In [8]:
def remove_stopwords(text):
    words = [w for w in text if w not in stopwords.words('english')]
    return words
```

In [9]:
```
reviews_df['Token_Reviews'] = reviews_df['Token_Reviews'].apply(lambda x:remove_stopwords(x))
reviews_df['Token_Reviews'].head(10)
```

Out[9]:
```
0    [morning, sign, advertising, villager, inn, ok...
1                                        [hear, rooms]
2                              [nice, rooms, remodel]
3    [hotel, stinks, cigarettes, rooms, bedding, ha...
4    [stay, hotel, frequently, checked, september, ...
5    [chambre, pas, propre, cafeti, salle, pas, rec...
6    [il, semble, que, la, femme, de, nage, utilise...
7    [room, smelled, like, mold, requested, another...
8    [wow, stay, big, downer, check, marriott, rewa...
9    [stay, broke, bug, bites, upper, body, legs, w...
Name: Token_Reviews, dtype: object
```

In [10]:
```
def remove_shortwords(text):
    name = [w for w in text if len(w)>3]
    return name
```

In [11]:
```
reviews_df['Token_Reviews'] = reviews_df['Token_Reviews'].apply(lambda x:remove_shortwords(x))
reviews_df['Token_Reviews'].head(10)
```

Out[11]:
```
0    [morning, sign, advertising, villager, okay, s...
1                                        [hear, rooms]
2                              [nice, rooms, remodel]
3    [hotel, stinks, cigarettes, rooms, bedding, ha...
4    [stay, hotel, frequently, checked, september, ...
5    [chambre, propre, cafeti, salle, recommandable...
6    [semble, femme, nage, utilise, torchon, pour, ...
7    [room, smelled, like, mold, requested, another...
8    [stay, downer, check, marriott, rewards, profi...
9    [stay, broke, bites, upper, body, legs, wife, ...
Name: Token_Reviews, dtype: object
```

Stemming(Getting words closer to root word):

In [12]:
```
stemmer = PorterStemmer()
```

In [13]:
```
def word_stemmer(text):
    stem_text = [stemmer.stem(i) for i in text]
    return stem_text
```

In [14]:
```
reviews_df['Token_Reviews'] = reviews_df['Token_Reviews'].apply(lambda x:word_stemmer(x))
reviews_df['Token_Reviews'].head(10)
```

Out[14]:
```
0    [morn, sign, advertis, villag, okay, someth, c...
1                                        [hear, room]
2                               [nice, room, remodel]
3    [hotel, stink, cigarett, room, bed, hallway, s...
4    [stay, hotel, frequent, check, septemb, suppos...
5    [chambr, propr, cafeti, sall, recommand, pour,...
6    [sembl, femm, nage, utilis, torchon, pour, tou...
7    [room, smell, like, mold, request, anoth, room...
8    [stay, downer, check, marriott, reward, profil...
9    [stay, broke, bite, upper, bodi, leg, wife, ex...
Name: Token_Reviews, dtype: object
```

Lemmatization(Getting correct english words from inbuild dictionary):

In [15]:
```
lemmatizer = WordNetLemmatizer()
```

In [16]:
```
def word_lemmatizer(text):
    lem_text=[lemmatizer.lemmatize(i) for i in text]
    return lem_text
```

In [17]:
```
reviews_df['Token_Reviews'] = reviews_df['Token_Reviews'].apply(lambda x:word_lemmatizer(x))
reviews_df['Token_Reviews'].head(10)
```

Out[17]:
```
0    [morn, sign, advertis, villag, okay, someth, c...
1                                        [hear, room]
2                               [nice, room, remodel]
3    [hotel, stink, cigarett, room, bed, hallway, s...
4    [stay, hotel, frequent, check, septemb, suppos...
5    [chambr, propr, cafeti, sall, recommand, pour,...
6    [sembl, femm, nage, utilis, torchon, pour, tou...
7    [room, smell, like, mold, request, anoth, room...
8    [stay, downer, check, marriott, reward, profil...
9    [stay, broke, bite, upper, bodi, leg, wife, ex...
Name: Token_Reviews, dtype: object
```

Apply LDA Model on Clean Data from above Steps:

In [18]:
```
#LDA
import gensim
import pyLDAvis.gensim
from gensim import corpora, models, similarities
```

In [19]:
```
#Create a Gensim dictionary from the tokenized data
tokenized = reviews_df['Token_Reviews']
#Creating term dictionary of corpus, where each unique term is assigned an index.
dictionary = corpora.Dictionary(tokenized)
#Filter terms which occurs in less than 1 review and more than 80% of the reviews.
dictionary.filter_extremes(no_below=1, no_above=0.8)
#convert the dictionary to a bag of words corpus
corpus = [dictionary.doc2bow(tokens) for tokens in tokenized]
print(corpus[:1])
```

```
[[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1), (16, 1), (17, 1), (18, 1), (19,
1), (20, 2), (21, 1), (22, 1), (23, 1), (24, 1), (25, 1), (26, 1), (27, 1), (28, 2), (29, 1), (30, 1), (31, 1)]]
```

In [20]:
```
[[(dictionary[id], freq) for id, freq in cp] for cp in corpus[:1]]
```

Out[20]:
```
[[('advertis', 1),
  ('appear', 1),
  ('barrel', 1),
  ('clean', 1),
  ('constant', 1),
  ('cost', 1),
  ('enough', 1),
  ('enter', 1),
  ('felt', 1),
  ('hard', 1),
  ('heard', 1),
  ('howev', 1),
  ('knew', 1),
  ('mattress', 1),
  ('morn', 1),
  ('must', 1),
```

LDA Result:

In [21]:
```
#LDA
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = 5, id2word=dictionary, passes=40)
ldamodel.save('model_combined.gensim')
topics = ldamodel.print_topics(num_words=10)
for topic in topics:
    print(topic)
```

```
(0, '0.012*"room" + 0.011*"good" + 0.008*"stay" + 0.008*"hotel" + 0.007*"breakfast" + 0.007*"locat" + 0.006*"work" + 0.006*"price" + 0.006*"conveni" + 0.006*"nice"')
(1, '0.051*"room" + 0.027*"hotel" + 0.020*"stay" + 0.016*"night" + 0.010*"smell" + 0.010*"door" + 0.010*"clean" + 0.008*"breakfast" + 0.007*"need" + 0.007*"like"')
(2, '0.024*"room" + 0.014*"hotel" + 0.013*"check" + 0.011*"staff" + 0.009*"like" + 0.008*"smell" + 0.008*"help" + 0.008*"would" + 0.007*"stay" + 0.007*"need"')
(3, '0.025*"hotel" + 0.023*"stay" + 0.017*"room" + 0.011*"check" + 0.008*"place" + 0.008*"cruis" + 0.008*"book" + 0.007*"would" + 0.007*"breakfast" + 0.006*"rate"')
(4, '0.034*"room" + 0.031*"hotel" + 0.020*"clean" + 0.018*"staff" + 0.015*"good" + 0.015*"breakfast" + 0.013*"stay" + 0.011*"locat" + 0.010*"nice" + 0.010*"friendli"')
```

Visualization for LDA:

In [23]:
```
#visualizing topics
lda_viz = gensim.models.ldamodel.LdaModel.load('model_combined.gensim')
lda_display = pyLDAvis.gensim.prepare(lda_viz, corpus, dictionary, sort_topics=False)
pyLDAvis.display(lda_display)
```