



# Hotel Reviews

Done by:  
Amjad Althinyan  
Eman Alshehri

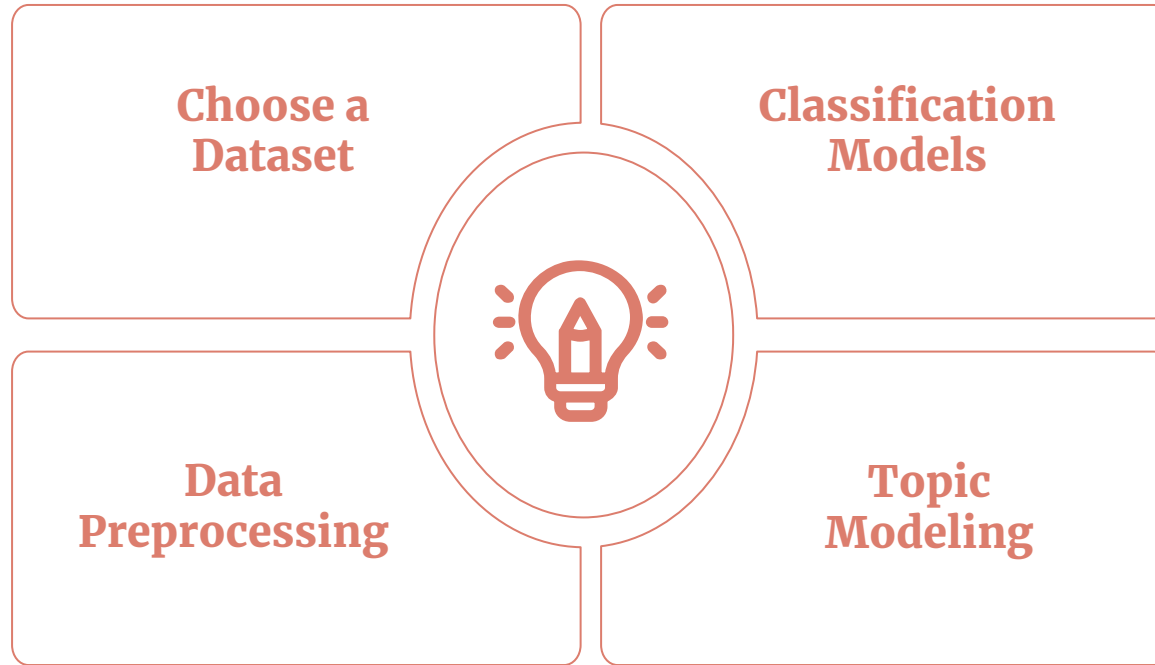
## Overview

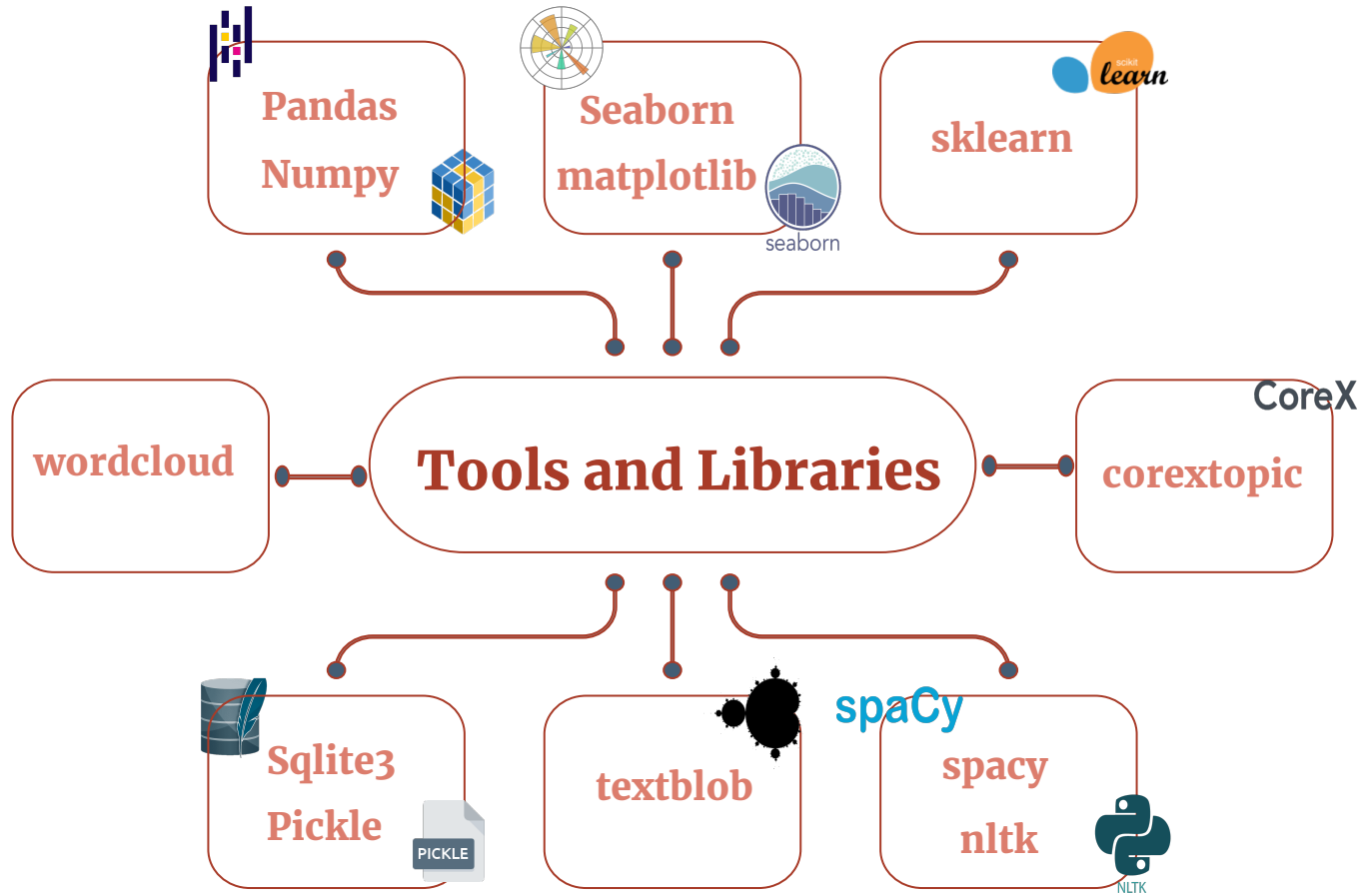
**In this project, we are working on a dataset that consists of text about the hotel reviews. Our observation is a customer's review.**

## Goal

**Building NLP model which is unsupervised learning that focuses on finding meaningful topics on Hotel reviews.**

# Methodology





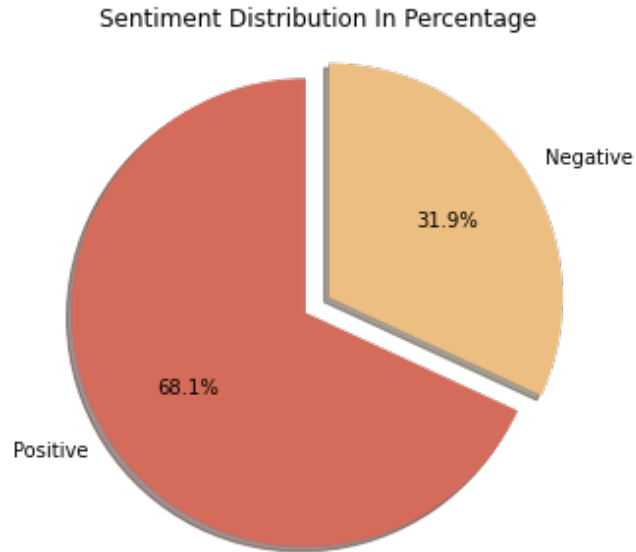
# Dataset

**38,933** documents

**5** terms

User_ID	Description	Browser_Used	Device_Used	Is_Response
---------	-------------	--------------	-------------	-------------

# Exploratory Data Analysis (EDA)



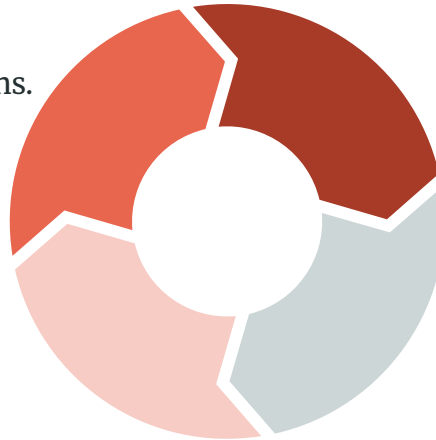
# Data Preprocessing

## Data Cleaning

- Remove Chinese letters.
- Remove spaces and punctuations.
- Remove repeated letters.
- Remove numbers.
- Remove empty tokens.
- Remove stop words.

## Stemming & Lemmatization

- Stemming and lemmatization the review words.



## Delete Meaningless Words

- Remove the meaningless words

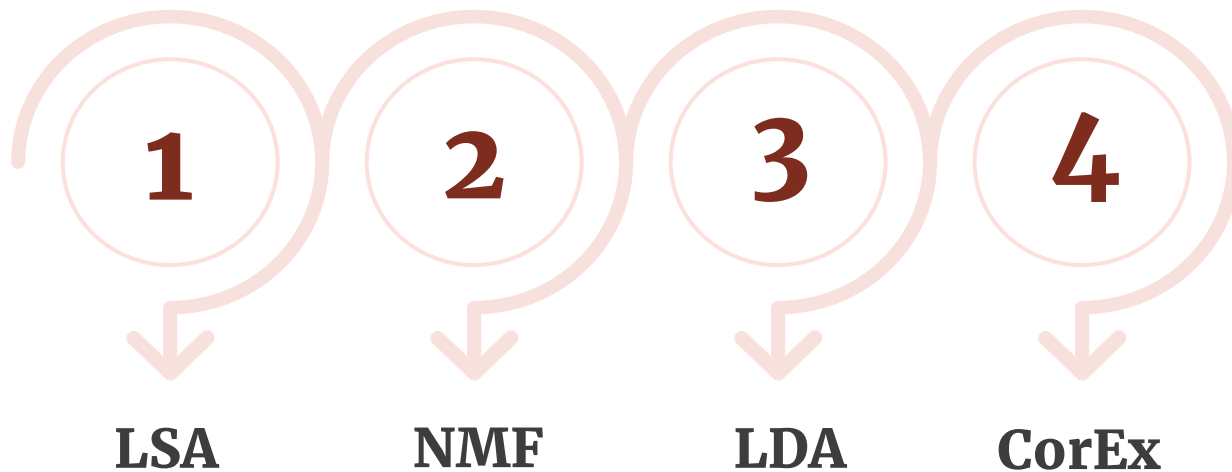
## Vectorization

- Count Vectorizer.
- TF-IDF Vectorizer.

## Spelling Correction

- correcting the words in reviews.

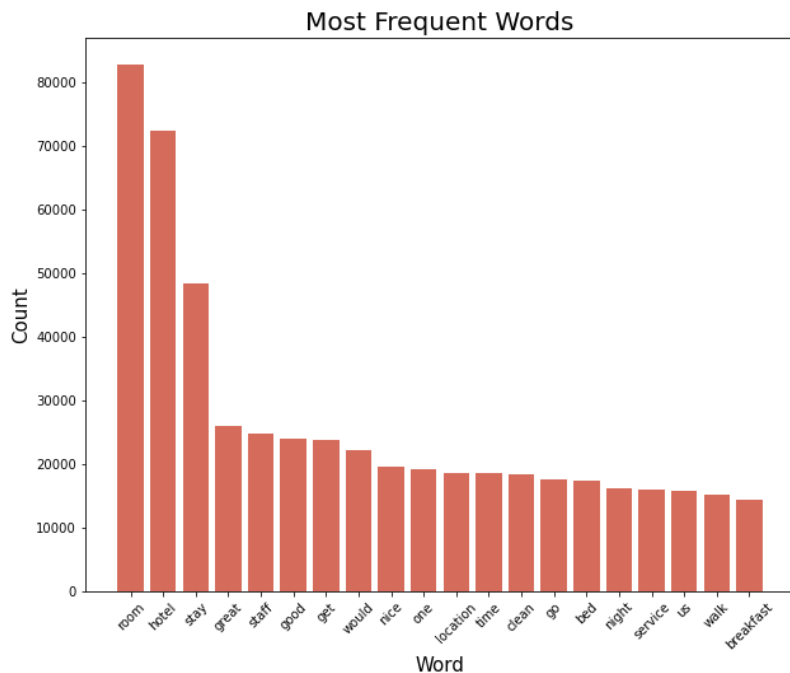
# Topic Modeling Algorithms



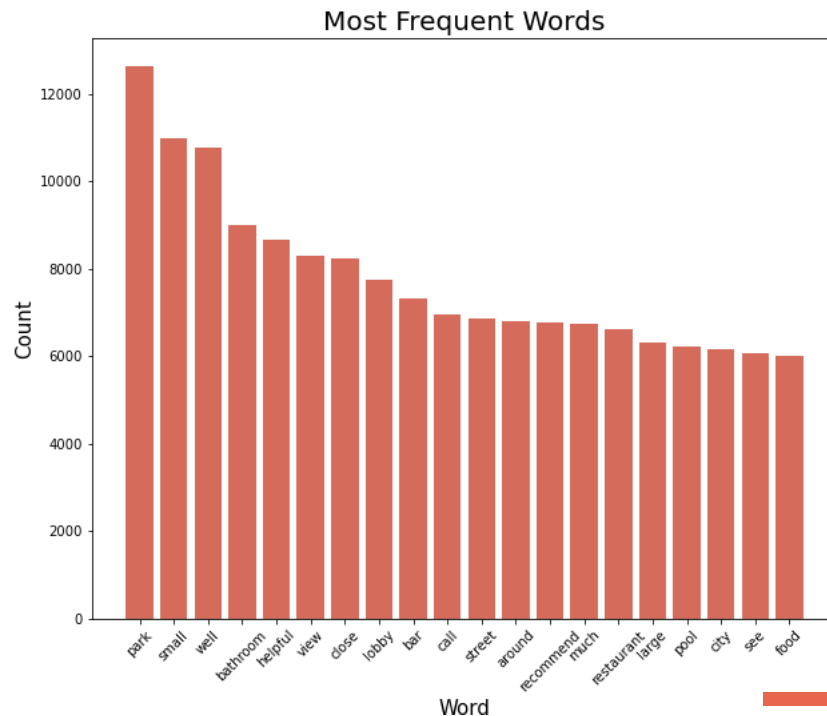


# Delete Meaningless Words

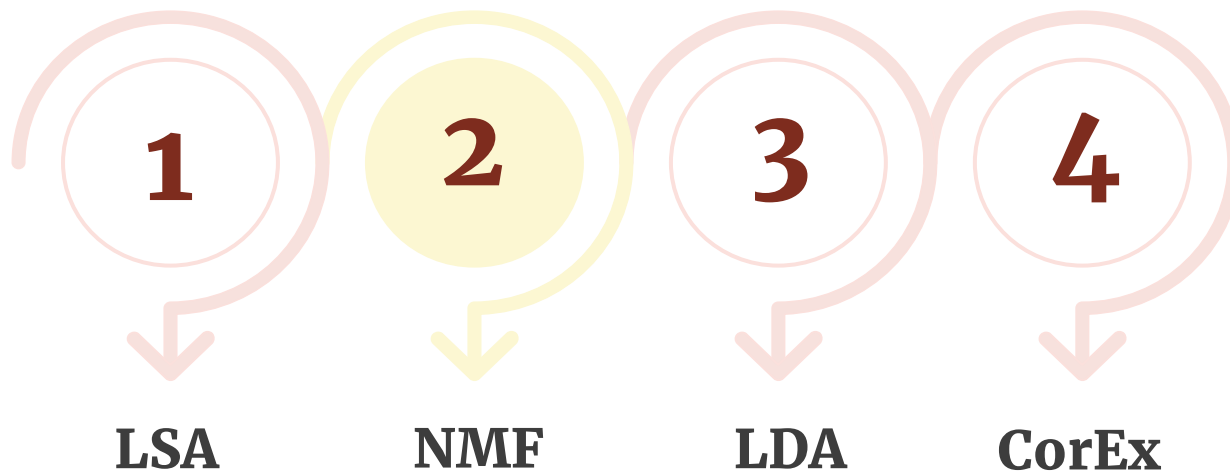
First iteration



Fifth iteration



# Topic Modeling Algorithms

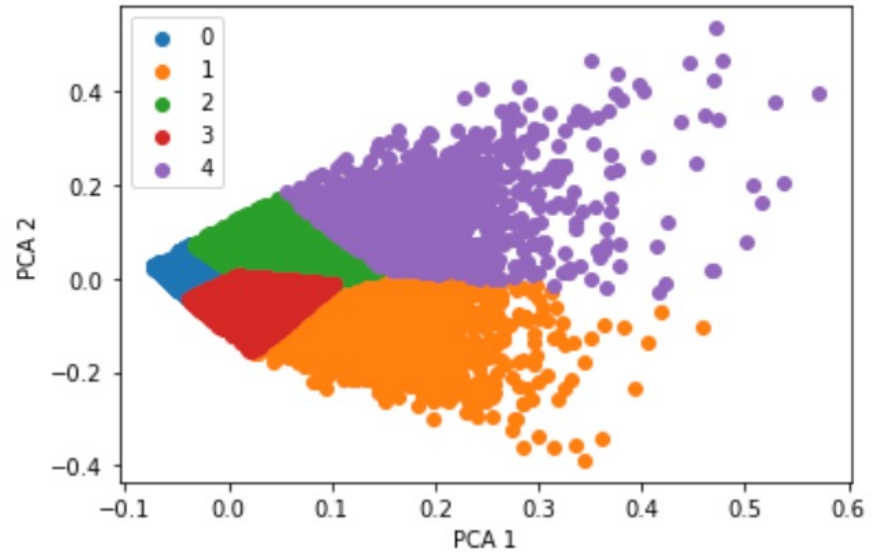
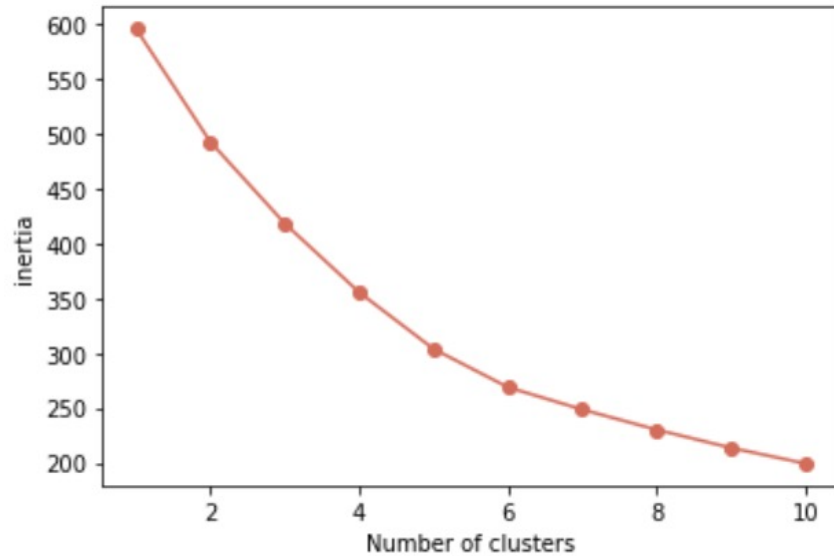


The Best Algorithm is **NMF** with 5 topics

Procedures, Parking, General Atmosphere, Room Facilities, Resort Hotel.

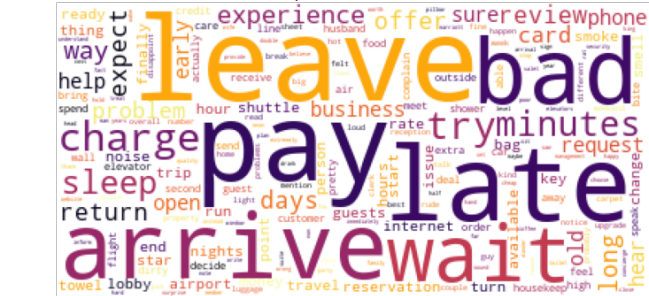
# Clustering & PCA

The Elbow Method

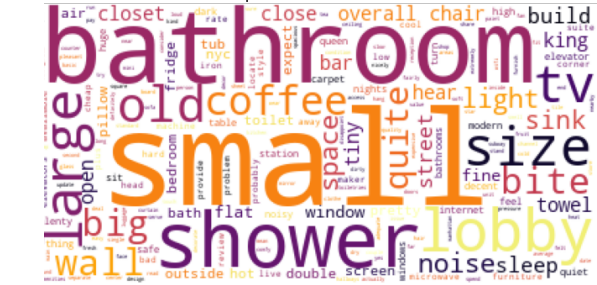


100%

Topic: Procedures



Topic: Room Facilities



Topic: General Atmosphere



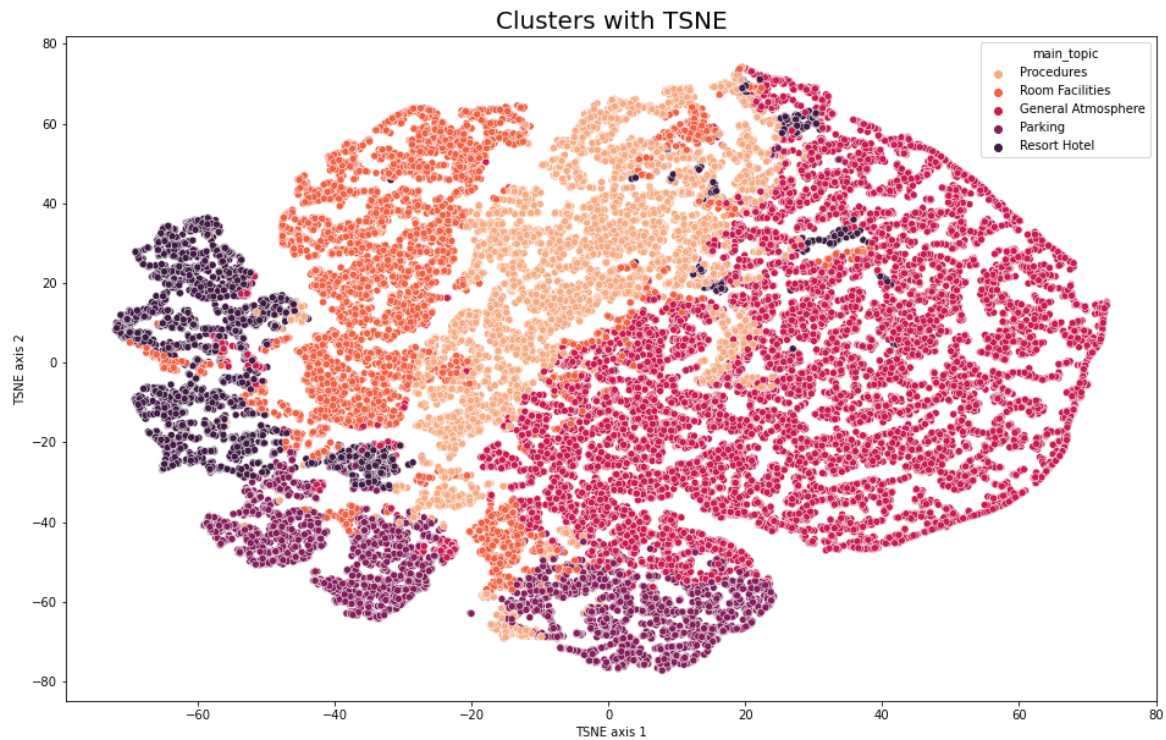
Topic: Resort Hotel



Topic: Parking

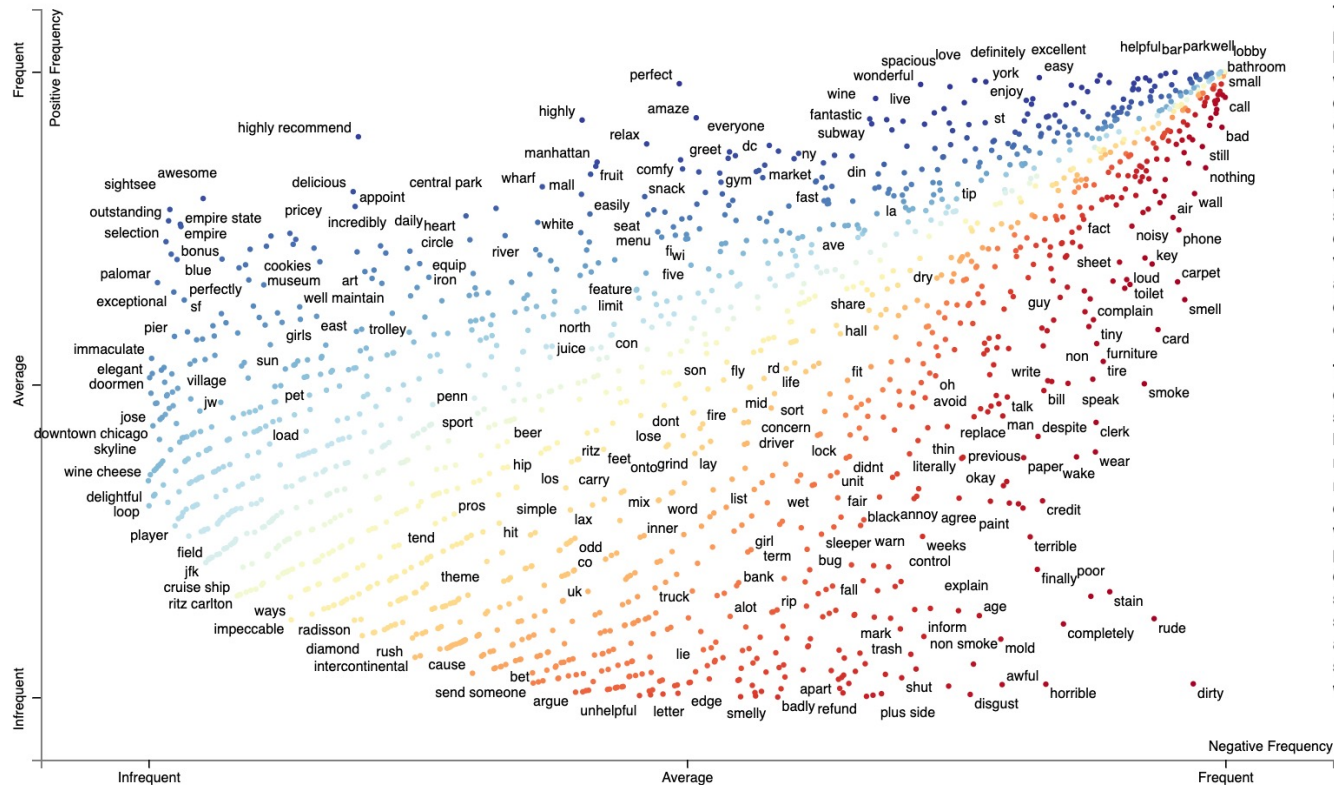


# TSNE





# Scatter Text



## Top Positive Characteristic

perfect  
love  
wonderful  
definitely  
enjoy  
spacious  
excellent  
fantastic  
square  
quiet  
visit  
always  
distance  
everything

housekeep  
sightsee  
overprice  
disappoint  
amaze  
doorman  
accommodating  
kimpton  
adjoin  
renovate  
affinia  
situate  
valet  
doorman  
concierge

## Top Negative

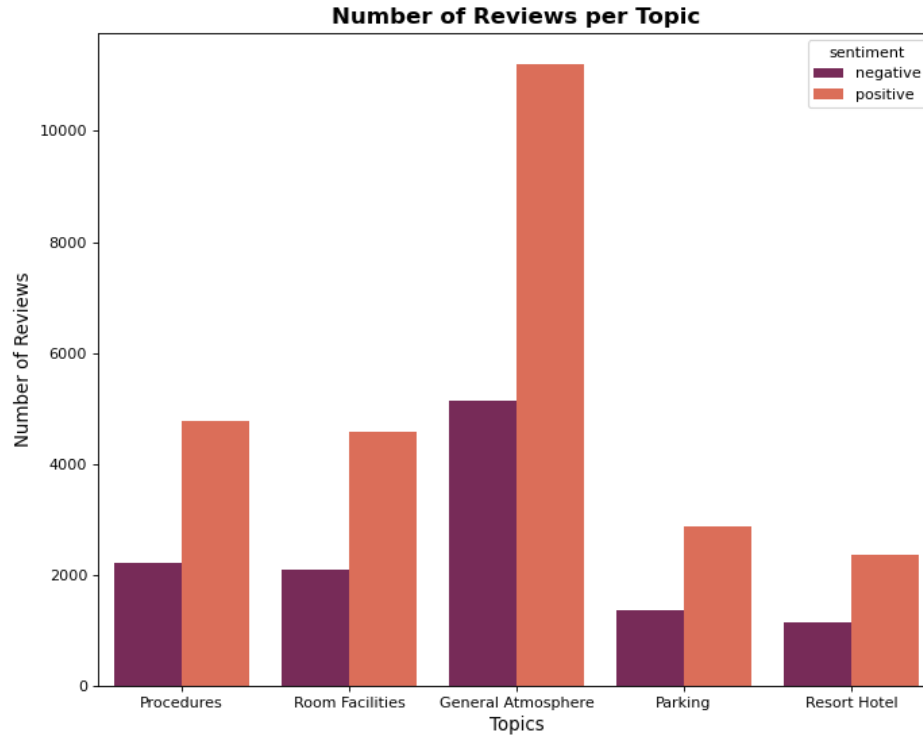
dirty  
smell  
bad  
nothing  
rude  
carpet  
wall  
phone  
card  
still  
smoke  
air  
stain  
wait

gaslamp  
comfy  
touristy  
lobby  
bellman  
soundproof  
elevators  
hallways  
riverwalk  
spotless  
palomar  
bite  
cramp  
courteous

Positive document count: 1,346; word count: 60,778

Negative document count: 654; word count: 38,039

# Sentiment Reviews per Topic



# Classification Models

	Training	Validation
Logistic Regression	0.9685	0.9668
Random Forest Classifier	1.000	0.9820
Bernoulli NB	0.4843	0.4973
Multinomial NB	0.4313	0.4409
Gaussian NB	0.7999	0.8057



# Classification Models

	Training	Validation
Logistic Regression	0.9685	0.9668
Random Forest Classifier	1.000	0.9820
Bernoulli NB	0.4843	0.4973
Multinomial NB	0.4313	0.4409
Gaussian NB	0.7999	0.8057

Random Forest  
Classifier is  
**Best Model**

# Selected Models

	Training		Testing	
	Accuracy	Loss	Accuracy	Loss
Random Forest Classifier	1.000	0.000	0.9842	0.0158

**THANK YOU!**

