Hotel Reviews

Done by:

Amjad Althinyyan

Eman Alshehri

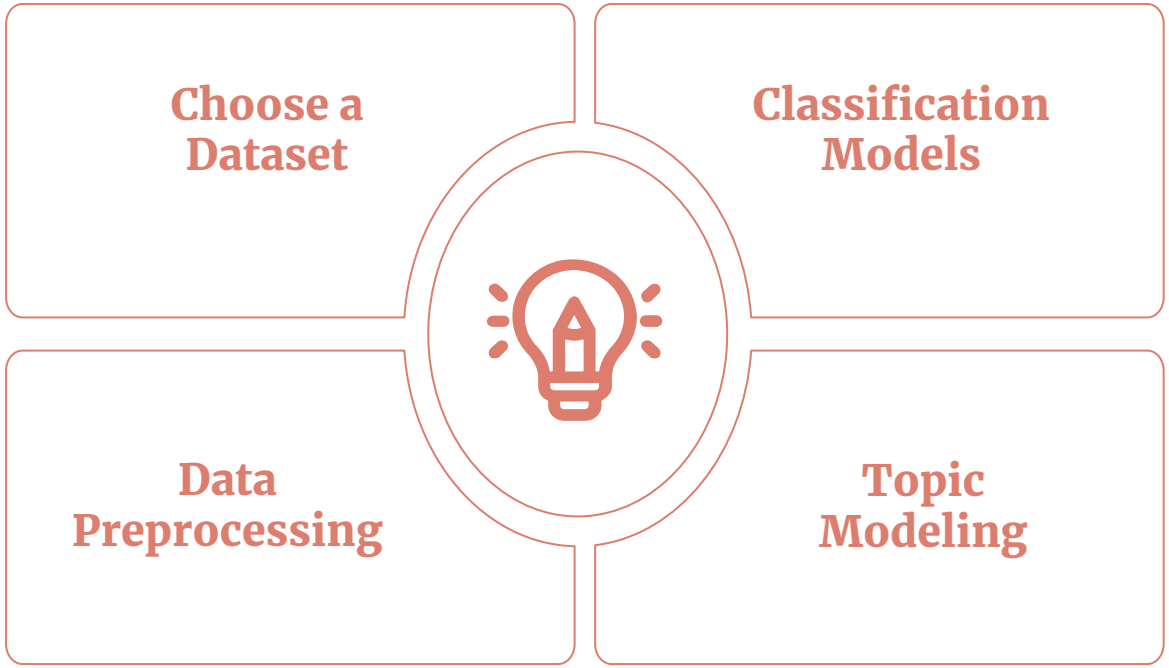
Overview

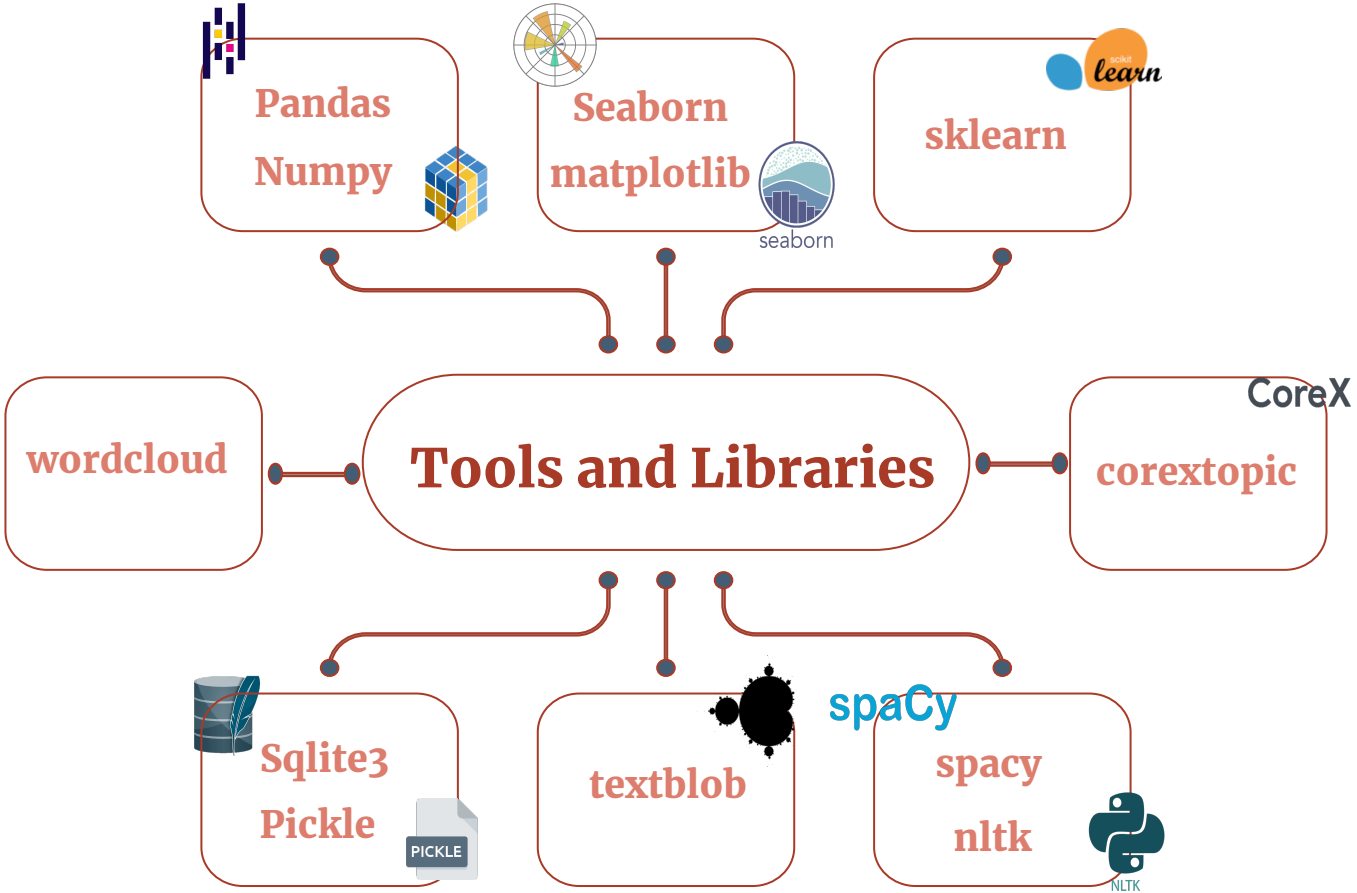
In this project, we are working on a dataset that consists of text about the hotel reviews. Our observation is a customer's review.

Goal

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

Methodology




Pandas
Numpy
Seaborn
matplotlib
seaborn
sklearn
wordcloud
Tools and Libraries
CoreX
corextopic
Sqlite3
Pickle
PICKLE
textblob
spaCy
spacy
nltk
NLTK

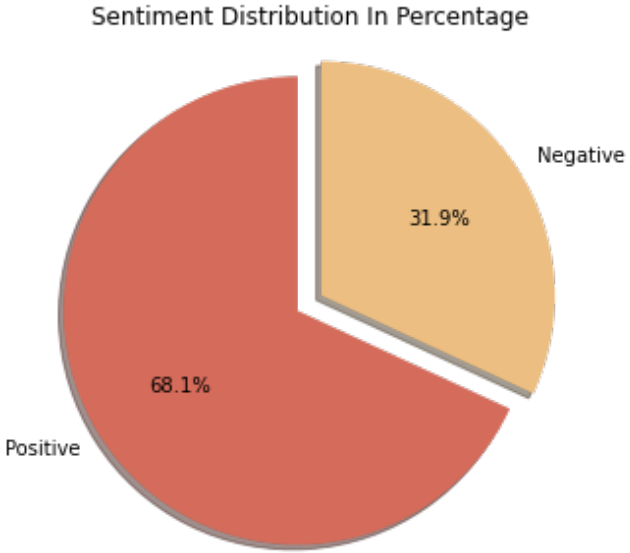
Dataset

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

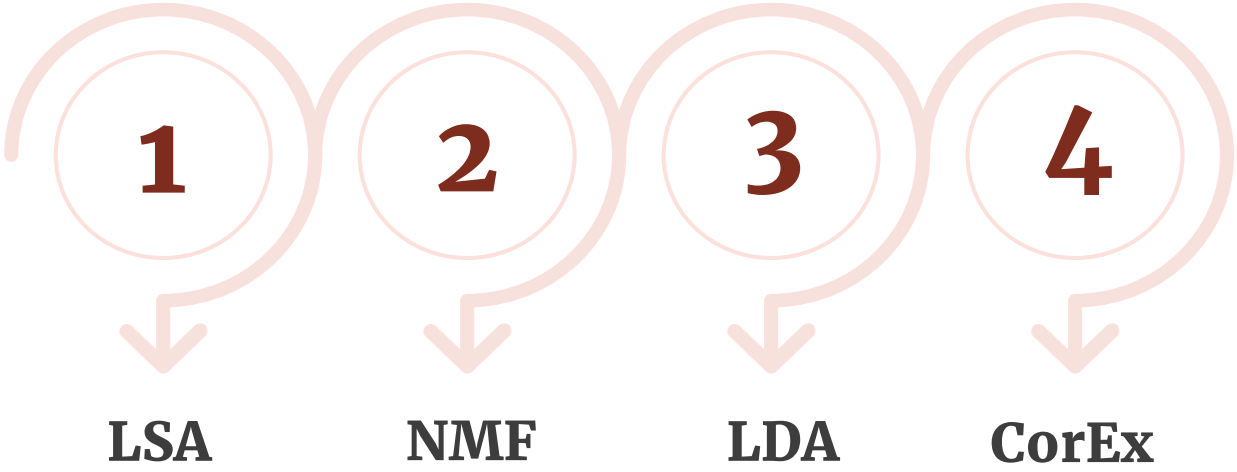
Vectorization

- Count Vectorizer.
- TF-IDF Vectorizer.

Spelling Correction

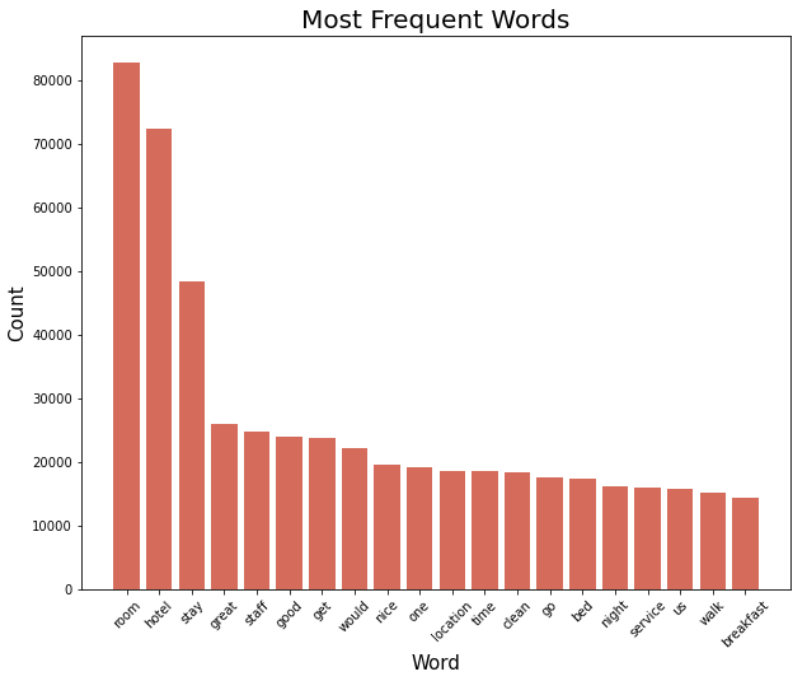
- correcting the words in reviews.

Topic Modeling Algorithms

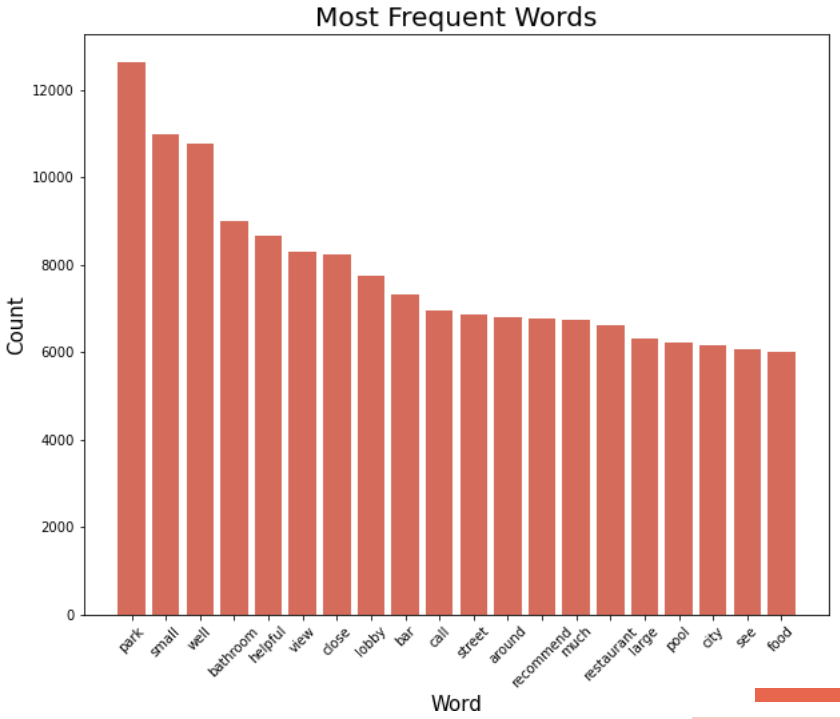


Delete Meaningless Words

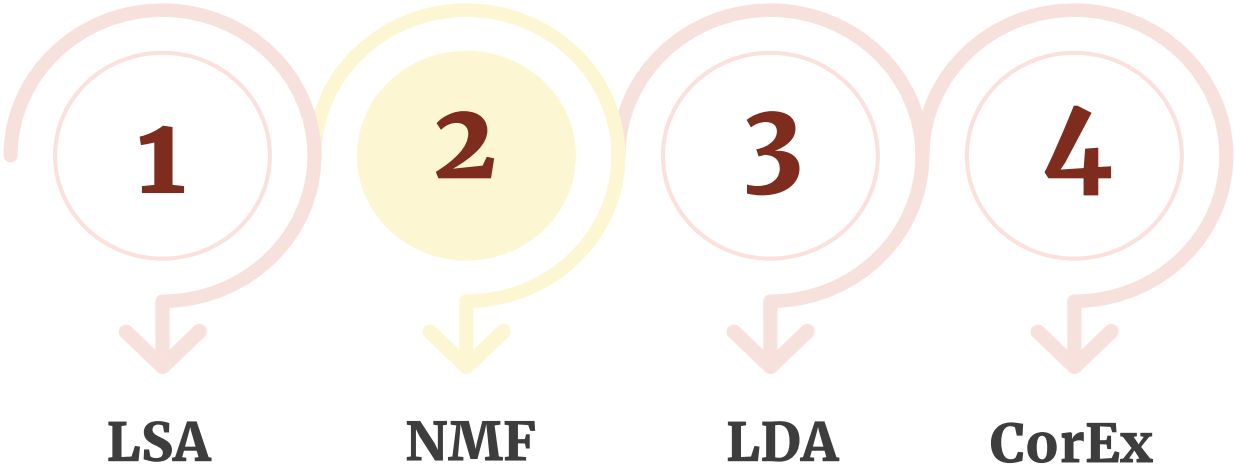
First iteration



Fifth iteration



Topic Modeling Algorithms



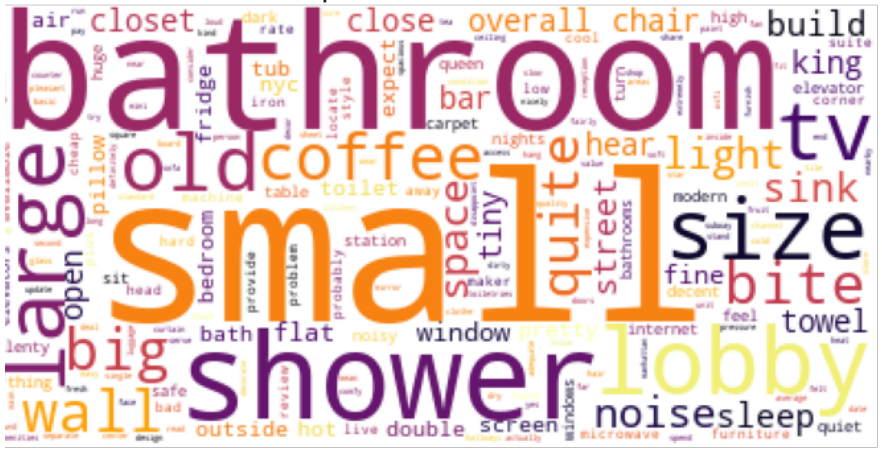
The Best Algorithm is NMF with 5 topics

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

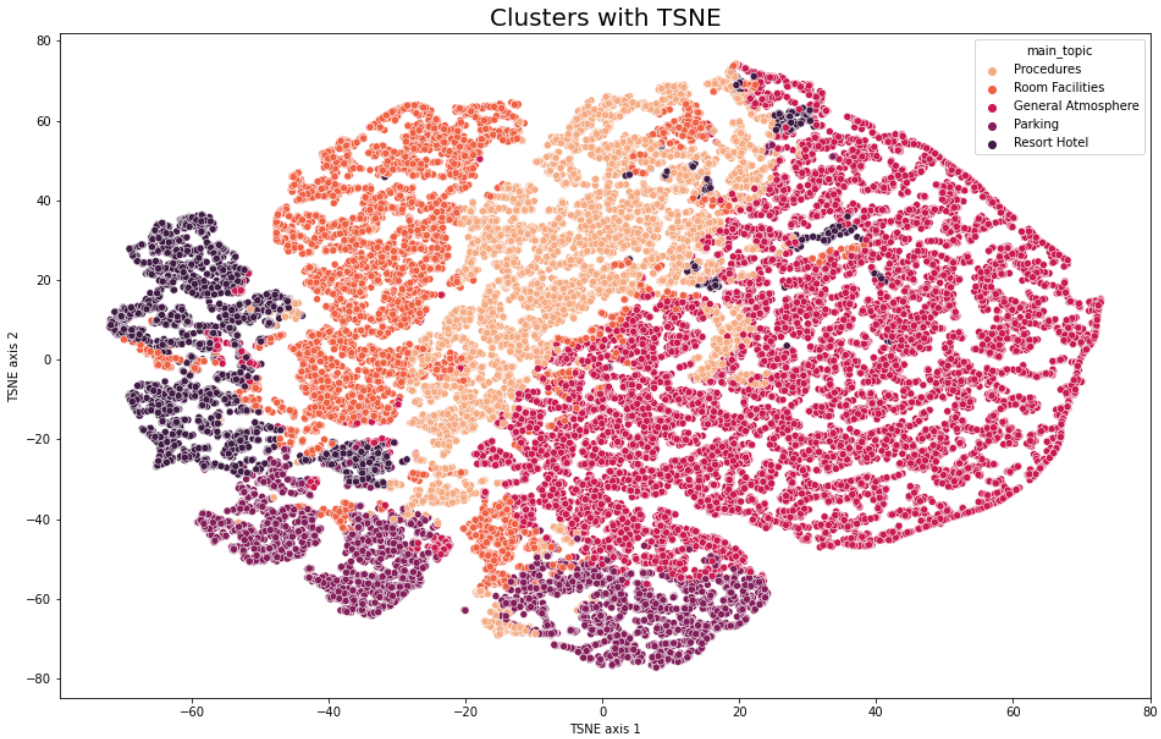
WordCloud

Topic: General Atmosphere

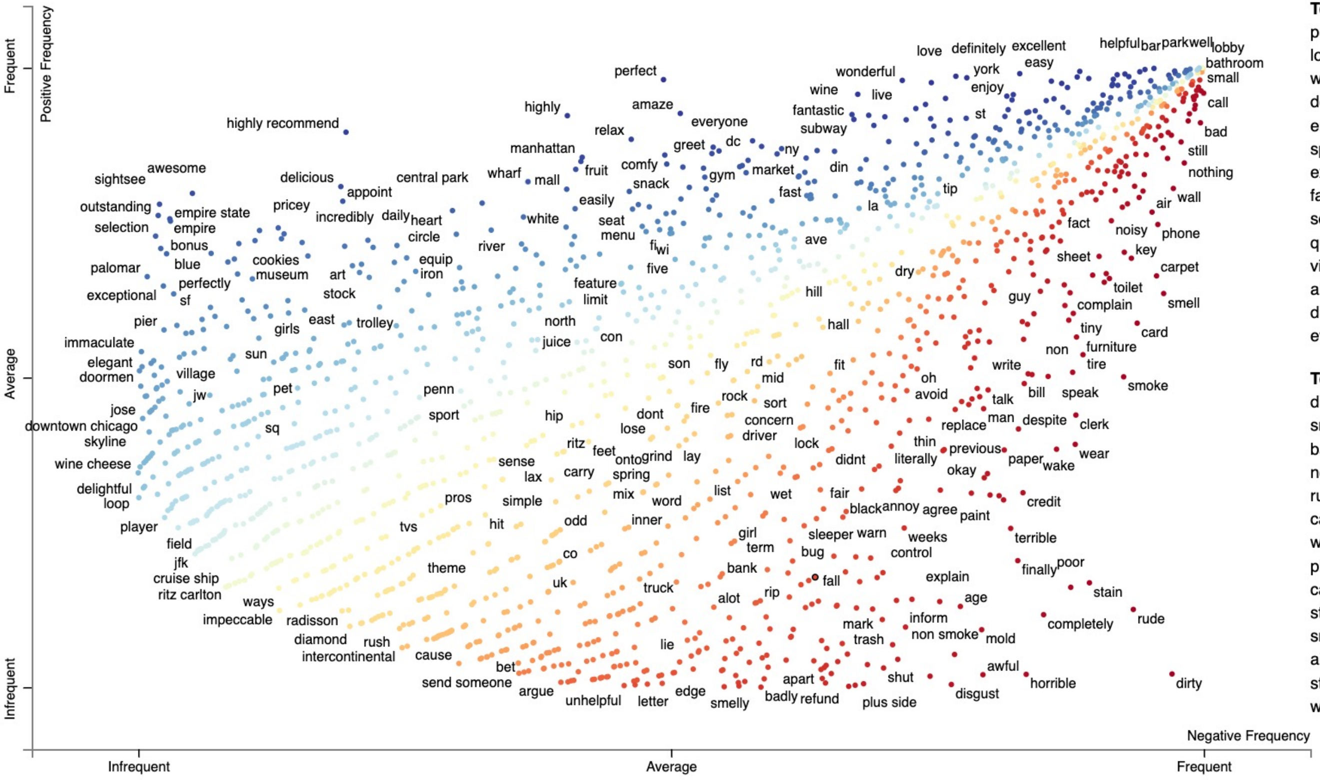
Topic: Room Facilities



TSNE


Clusters with TSNE
main_topic
Procedures
Room Facilities
General Atmosphere
Parking
Resort Hotel
TSNE axis 2
TSNE axis 1
80
60
40
20
0
−20
−40
−60
−80
−60
−40
−20
0
20
40
60
80

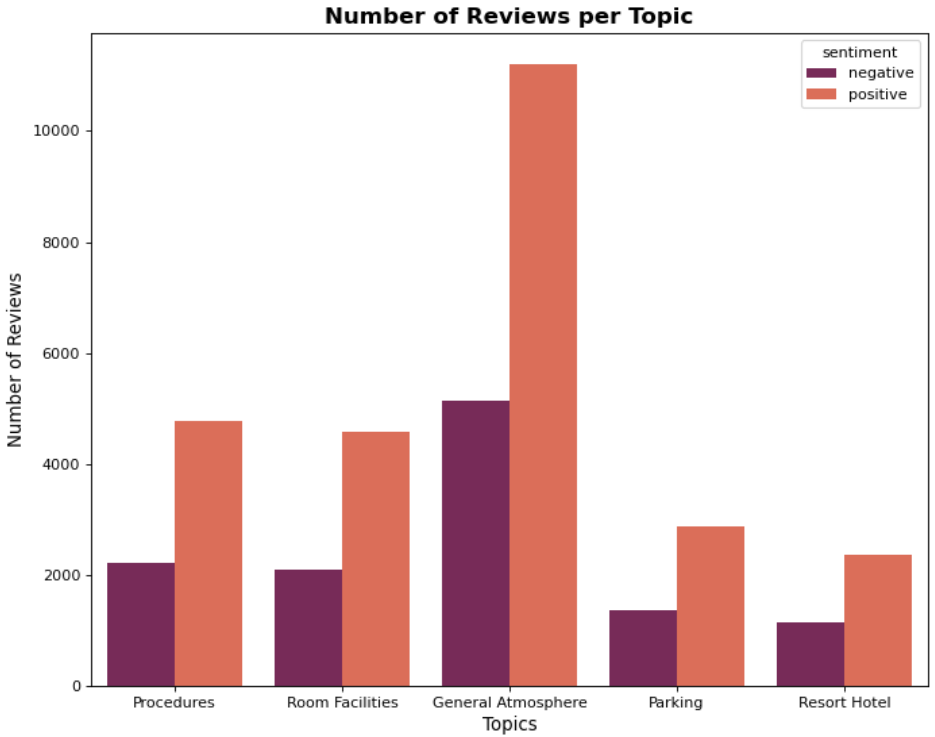
Scatter Text



Top Positive	Characteristic
perfect	housekeep
love	sightsee
wonderful	overprice
definitely	disappoint
enjoy	amaze
spacious	doormen
excellent	accomodating
fantastic	kimpton
square	adjoin
quiet	renovate
visit	affinia
always	situate
distance	valet
everything	doorman
	concierge
Top Negative	clothe
dirty	gaslamp
smell	comfy
bad	touristy
nothing	lobby
rude	bellman
carpet	soundproof
wall	elevators
phone	hallways
card	riverwalk
still	spotless
smoke	palomar
air	bite
stain	cramp
wait	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
Random Forest Classifier	1.000	0.9842

THANK YOU!