Hotel Review (NLP)

TABLE OF CONTENTS

1. INTRODUCTION
2. PREPROCESSING
3. TOPIC MODELING
4. MODELS
5. CONCLUSION

Introduction

- Client reviews are a crucial part of improving hotels and customers experience and travel plan.

- Customers reviews plays an important part for hotel improvement and in order to improve their services.

- The purpose is to build a NLP model that predicts whether the review is positive or negative.

- The goal of this project is to help hotels acknowledge the majority of reviews and determine their customers feedback to improve hotel services and customers experience.

Tools

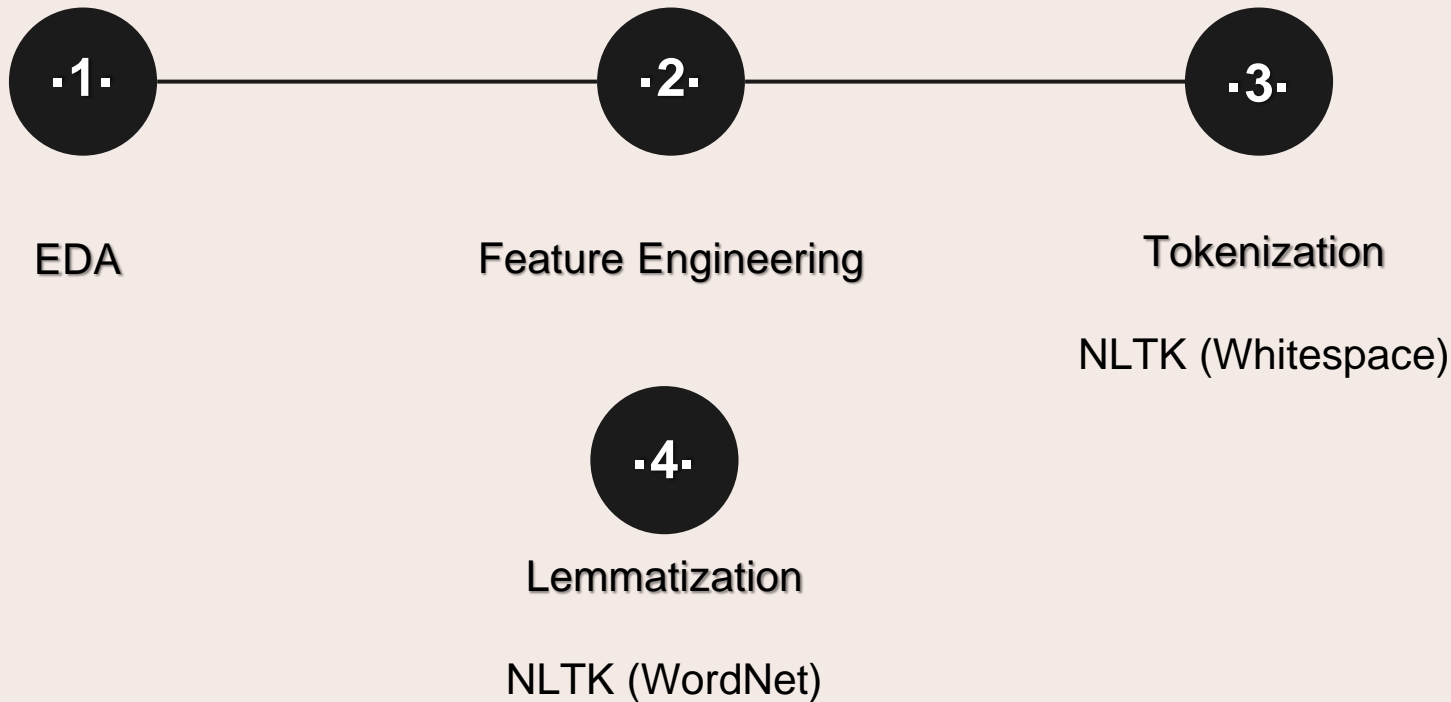
Jupyter Notebook

Pandas

Sklearn

WordCloud

NLTK

Pickle

Data Preprocessing

·1· EDA

·2· Feature Engineering

·3· Tokenization

NLTK (Whitespace)

·4· Lemmatization

NLTK (WordNet)

DATASET

- Scraped data from Booking.com
- More than 500k hotel reviews
- 50000 Sample using pickle

WordCloud

Positive Reviews

Negative Reviews
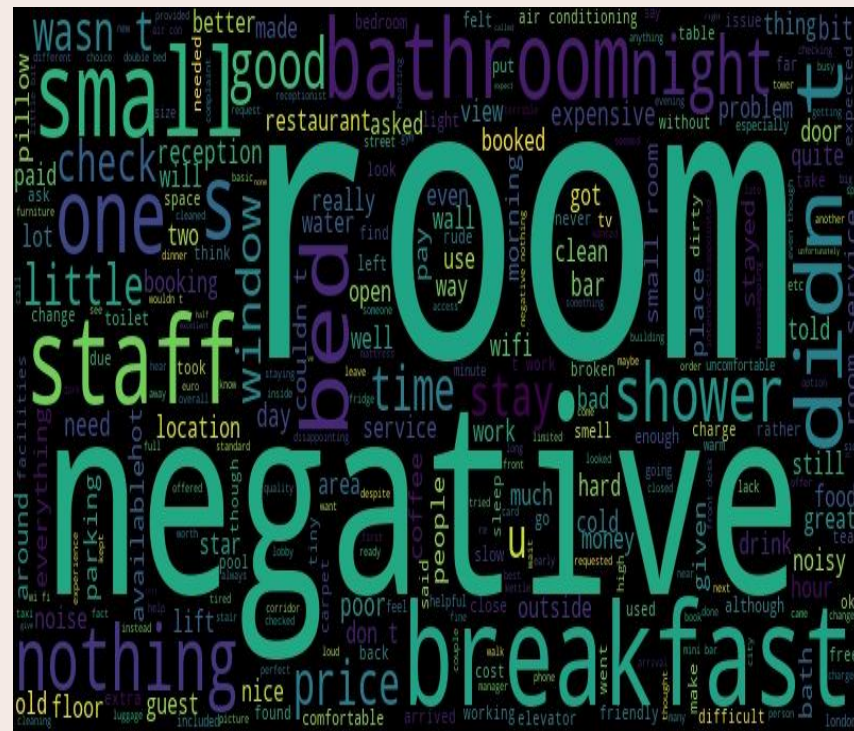
Topic modeling
NMF (Positive)

Topic 1	Topic 2	Topic 3	Topic 4
Staff	**Room**	**Location**	**Transportation**
staff, friendly, helpful, excellent, lovely, reception, extremely, amazing, polite, nice, welcoming, pleasant, restaurant, fantastic, bar, attentive, perfect, professional, super, facility	room, comfortable, clean, bed, nice, excellent, lovely, spacious, quiet, bathroom, modern, view, size, big, shower, station, close, large, facility, perfect	location, excellent, perfect, central, fantastic, facility, convenient, comfort, value, station, cleanliness, nice, close, price, near, amazing, ideal, best, tube	positive, think, attitude, ended, relocating, helpful, nice, excellent, actually, cleanliness, reception, point, experience, doomed, regina, classy, porter, busicuts, environment, brittania

Topic modeling
LSA (Positive)

Topic 1	Topic 2	Topic-3
Staff	Room	Location
room, staff, location, helpful, clean, nice, comfortable, excellent, bed, lovely	room, bed, comfortable, clean, nice, bathroom, spacious, view, big, small	location, room, station, perfect, close, central, walk, excellent, near, tube

Topic modeling
CorEx (Positive and Negative)

Topic 1	Topic 2	Topic 3
Location	Room	Facilities
station, close, walk, metro, walking, tube, distance, near, walking distance, easy	room, comfortable, lean, bed, rooms, spacious, positive, bathroom, nice, staff	bar, coffee, tea, reception, water, restaurant, pool, complimentary, park, arrival

Data Split

60%
Train

20%
Test

20%
Validation

Models

	Train	Validation
Logistic Regression (CV)	0.966	0.9263
Naïve Bayes (CV)	0.8897	0.8735
Decision Tree (CV)	0.9943	0.8716
Logistic Regression (TF-IDF)	0.9503	0.9398
Naïve Bayes (TF-IDF)	0.8860	0.8723
Decision Tree (TF-IDF)	0.9931	0.8844

Conclusions

- The best model is Logistic Regression (TF-IDF).

- Highest validation accuracy (0.9398).

THANKS